Institut des ACTUAIRES

ENSAE IP PARIS

Mémoire présenté devant l'ENSAE Paris
pour l'obtention du diplôme de la filière Actuariat
et l'admission à l'Institut des Actuaires
le 10/03/2021

Par : **Ratiba MIKOU**

Titre : **Estimation de la charge ultime et mesure de la volatilité à travers les méthodes de machine learning pour la garantie RC corporelle automobile**

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

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

Membres présents du jury de la filière

Quentin GUIBERT

Membres présents du jury de l'Institut des Actuaires

GERMAIN Michel

Stéphanie FOATA p/o Chatelain

Arnaud LACOURE p/o C. Hillairet

Entreprise : AXA France
Nom : Said CHAFNI
Signature :

Directeur du mémoire en entreprise :

Nom : Said CHAFNI
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

Secrétariat :

Bibliothèque :

Ecole Nationale de la Statistique et de l'Administration Economique (ENSAE)
5, avenue Henry Le Chatelier - 91120 PALAISEAU, FRANCE

Résumé

Le provisionnement constitue une préoccupation majeure pour les assureurs compte tenu des enjeux économiques et prudentiels sous-jacents. La réglementation en vigueur, au travers notamment de la directive Solvabilité II, exige un provisionnement au plus juste. Cette problématique se révèle d'autant plus complexe lorsqu'il s'agit de garanties à développement long, à l'instar de la garantie RC corporelle automobile traitée dans le cadre de ce mémoire.

Dans ce contexte et face à l'émergence d'initiatives actuarielles basées sur un usage approfondi des données, notamment au travers du *machine learning*, ce mémoire vise à confronter le modèle mis en place par AXA France pour le provisionnement de cette garantie. En effet, ce modèle se base sur les méthodes agrégées qui n'exploitent qu'une infime partie des données disponibles et dont les hypothèses peuvent ne pas être pleinement satisfaites.

Par conséquent, nous nous attachons à construire un modèle à partir de données individuelles et visant à estimer les provisions pour sinistres à payer (PSAP) à la deuxième année d'inventaire, permettant ainsi de confronter les résultats fournis par le modèle AXA France. Le biais induit par la censure à droite des données individuelles dont nous disposons est pour sa part corrigé par l'usage d'une méthode de pondération des observations, dite méthode *IPCW*, appliquée aux modèles utilisés, en l'occurrence les arbres *CART* et les algorithmes de forêts aléatoires et *XGBoost*.

Parallèlement, l'estimation des erreurs de prédiction à l'ultime revêt un intérêt particulier. Ainsi, la méthode de *bootstrap* est appliquée aux modèles retenus afin d'évaluer cette erreur et de la comparer aux résultats obtenus par le modèle AXA France qui repose essentiellement sur le modèle de Mack. En définitive, cette analyse permet de confronter les deux approches et d'estimer le gain en précision apporté par le modèle proposé.

Mots clés : Provisionnement, Mack, Merz & Wüthrich, Censure, *IPCW*, Arbres de régression *CART*, Forêts aléatoires, *XGBoost*.

Abstract

The reserving is a major concern for insurers on account of the economic and prudential stakes. The regulation in force, in particular pursuant to directive Solvency II, demands reserving as accurate as possible. This issue has proven to be a more difficult problem to address when it comes to developing long-term claims, such as automobile bodily liability warranty covered in this brief.

In this context, and faced with the emergence of actuarial projects focusing on the in-depth use of data, notably machine learning approaches, this dissertation attempts to put the models established by AXA France for the provisioning of this type of warranty under scrutiny. In fact, the latter is based on traditional methods that only explore a small portion of the data available and whose assumptions might not be fully met.

Hence, we will mainly focus on building and training a model on two years collected data with the goal of estimating the *Provisions for Claims Payment (PCP)*, thereby allowing to perform a comparison between our results and the results computed by the model of AXA France. Right censored data induces a bias which will be corrected by applying weights on the observations using IPCW method during the training of our chosen models. These are CART trees, Random Forest and XGBoost.

Simultaneously, estimating the prediction error is central to assessing the performance of our solution. Moreover, *bootstrap* will be applied to the selected models with a view to evaluate this error and compare it with the results obtained by AXA France approach, which rests basically on Mack's model. Ultimately, this analysis provides us with a comparison of both approaches and an estimation of the accuracy increase brought by the suggested model.

Key words : Solvency II, Mack, Merz & Wüthrich, censorship, IPCW, regression trees CART, Random Forest, XGBoost.

Note de synthèse

Contexte et problématique

Le provisionnement, correspondant à l'évaluation des provisions nécessaires au règlement et à la gestion des sinistres, est au coeur du métier des assureurs compte tenu des forts enjeux économiques et prudentiels sous-jacents. Par conséquent, ces derniers sont dans une quête perpétuelle d'amélioration de leurs modèles de provisionnement afin d'aboutir à une estimation plus précise. Ce gain en précision induit mécaniquement une meilleure stabilité des estimations de charges ultimes et ainsi des provisions pour sinistres à payer (PSAP).

Nous limitons ainsi les phénomènes de sur-provisionnement, de sous-provisionnement et les effets négatifs qui en découlent, à savoir :

- dans le cadre du sous-provisionnement, outre le risque évident d'insolvabilité et de non indemnisation des assurés, l'effet négatif s'illustre par les malis obtenus lors des ré-estimations les années suivantes ;
- dans le cadre du sur-provisionnement, outre l'immobilisation inutile des ressources pour les assurés et le manque à gagner pour l'assureur, l'effet négatif s'illustre par la taxation des bonis lors des ré-estimations les années suivantes.

Cette problématique se révèle d'autant plus complexe et importante s'agissant de garanties à développement long, à l'instar de la garantie RC corporelle automobile traitée dans le cadre de ce mémoire. Cette garantie permet la prise en charge par l'assureur de l'indemnisation de l'ensemble des victimes de dommages corporels dont l'assuré est déclaré responsable, à hauteur du préjudice subi. La charge des sinistres relatifs à cette garantie évolue au cours du temps sous l'influence de plusieurs facteurs, dont la stabilisation de l'état de la victime, du mode d'indemnisation en capital ou en rente ou encore de la décision du tribunal en cas de contentieux.

Notre objectif dans le cadre de ce mémoire est de confronter le modèle mis en place par AXA France pour l'estimation des PSAP de la garantie RC corporelle automobile. En effet, le modèle actuellement utilisé se base sur les méthodes de provisionnement agrégées qui ne permettent pas de mettre à profit l'importante quantité de données dont dispose l'entreprise. Ainsi, il s'agirait d'identifier un modèle de *machine learning* qui permette de pleinement tirer profit de la quantité et de la fine granularité des données possédées pour aboutir à des estimations de provisions plus précises.

Modèle AXA France - méthodes agrégées

Il existe une multitude de méthodes de provisionnement agrégées qu'il est possible de répartir en deux catégories : d'une part les méthodes déterministes, permettant l'estimation du montant des provisions, et d'autre part les méthodes stochastiques permettant, outre l'estimation du montant des provisions, l'estimation de l'erreur de prédiction à différents horizons.

Dans le cadre de ce mémoire, nous avons sélectionné trois méthodes largement plébiscitées dont nous présentons le fondement théorique. Il s'agit de la méthode déterministe Chain-Ladder, ainsi que des

méthodes stochastiques de Mack et de Merz & Wüthrich. Ces dernières fournissent des estimations de provisions équivalentes à Chain-Ladder tout en estimant respectivement les erreurs de prédiction à l'ultime et à horizon un an.

Pour commencer, l'estimation du montant des provisions au travers de la méthode de Chain-Ladder revient à supposer que les cadences de charge observées dans le passé se reproduiront dans le futur. D'autre part, le principal atout des méthodes de Mack et de Merz & Wüthrich réside dans leur capacité à évaluer l'incertitude des provisions à travers des formules fermées. La méthode de Merz & Wüthrich se distingue néanmoins par sa capacité à les évaluer à horizon un an, conformément à la directive « Solvabilité II ».

Dans le cadre des méthodes agrégées, les données se présentent majoritairement sous la forme de triangles de liquidation. Les lignes correspondent aux exercices de rattachement des sinistres tandis que les colonnes correspondent aux périodes de développement. Dans notre cas, nous retenons la charge D/D nette de recours avec comme exercice de rattachement les années de survenance des sinistres de 1999 à 2014 et à minima 16 années de développement. Aussi, ces sinistres issus du portefeuille AXA France portent sur la garantie RC corporelle automobile pour les particuliers et les professionnels sur l'ensemble des réseaux de distribution, en l'occurrence sur les réseaux des agents, des courtiers et des salariés.

Le modèle de provisionnement retenu par AXA France se base sur les méthodes agrégées présentées ci-dessus. De plus, afin d'homogénéiser les populations de sinistres et ainsi satisfaire au mieux les hypothèses émises par les méthodes de Chain-Ladder et de Mack, Axa France estime sa charge ultime à partir de trois triangles distincts : attritionnels, graves et très graves.

Ainsi l'estimation de la charge ultime peut se résumer comme suit :

- **Triangle 12 (charges D/D $\leq$ 150*k*€) :** application de la méthode de Chain-Ladder.
- **Triangle 3 (charges D/D $\in$] 150k€ ; 750k€]) :** application de la méthode de Chain-Ladder.
- **Triangle 45 (charges D/D $>$ 750*k*€) :** application de la méthode de Mack dans un premier temps, pour pouvoir dans un second temps retenir le quantile à 75% d'une distribution de provisions supposée log-normale d'espérance $\mu = ln(\hat{R}) - \frac{\sigma^2}{2}$ et de variance $\sigma^2 = ln(1 + \frac{\hat{MSEP}(\hat{R})}{\hat{R}^2})$. Cette démarche revient à adopter une marge de prudence supplémentaire à la charge estimée par la méthode de Chain-Ladder. Ceci paraît nécessaire du fait du développement particulièrement long de cette dernière catégorie de sinistres qui peut provenir de leur atypisme, leur gravité, l'estimation de la durée de vie s'il s'agit d'une rente ou encore l'attente d'une décision judiciaire.

Il faut toutefois noter que l'estimation de la charge lors de la première année d'inventaire ne repose pas sur la méthode par tranche de coûts, décrite ci-dessus, mais plutôt sur une méthode d'ajustement fréquence/coût moyen. L'objectif de cette démarche est de parer aux fluctuations de la charge D/D observées la première année et qui peuvent fortement varier d'une survenance à une autre.

Modèle proposé - algorithmes de *machine learning*

Dans l'optique de contrer le manque d'informations au sein des triangles de liquidation, nous nous attachons à collecter un maximum de données afin de construire une base de données individuelle. Nous intégrons au sein de cette dernière toute variable jugée pertinente pour l'estimation de la variable d'intérêt qui, dans notre cas, correspond à la charge ultime du sinistre nette de recours.

La construction de cette base de données nécessite l'agrégation de trois bases de données distinctes, à savoir la base des sinistres, la base des contrats et la base des fiches victimes. Nous obtenons ainsi 88 616 sinistres survenus entre 2010 et 2014, représentant une moyenne annuelle de plus de 17 000 sinistres. Ces derniers concernent également les particuliers et professionnels sur l'ensemble des réseaux de distribution, en l'occurrence le réseau des agents, celui des salariés et enfin celui des courtiers. D'autre part, la variable d'intérêt correspond à une vision arrêtée à fin 2019 tandis que les variables explicatives sont arrêtées à la seconde année d'inventaire.

Il est important de noter que la variable d'intérêt fait l'objet d'une censure à droite. En effet compte tenu de la présence de sinistres non clos à fin 2019, nous ne disposons pas de la version définitive de la charge ultime de ces derniers. Cette variable est donc censurée en raison de la vision arrêtée de la base de données à fin 2019 et de l'impossibilité d'attendre la clôture de l'ensemble des sinistres, contrairement à certaines branches à développement court. En effet, la branche « RC corporelle automobile » est une branche particulièrement longue dont l'atteinte d'un taux de clôture à 100% peut nécessiter plus de trente années pour une survenance donnée.

Il est certain que cette censure n'est pas sans impact sur la façon d'entraîner nos modèles de *machine learning*. Afin de corriger le biais induit par cette dernière, nous pondérons les données à l'aide de poids *IPCW*, calculés au travers de l'estimateur de Kaplan-Meier. Ce mécanisme permet de compenser l'effet de censure en attribuant aux sinistres clos des poids plus importants à mesure que la charge est élevée. Les sinistres non clos, quant à eux, se voient attribuer une pondération nulle. Ces derniers ne sont pour autant pas ignorés car ils contribuent au calcul des poids attribués aux sinistres clos.

En amont de la présentation de la méthodologie adoptée et des résultats associés, il semble nécessaire d'introduire brièvement les modèles de *machine learning* testés.

Pour commencer, les arbres de décision CART constituent une méthode efficace d'exploration des données. Il s'agit d'une méthode itérative, dite de partitionnement récursif des données. Les solutions obtenues sont simples à interpréter car basées sur une séquence récursive de règles de division binaire. Néanmoins, les méthodes CART sont relativement instables et irrégulières, permettant ainsi l'essor des méthodes dites ensemblistes.

Par ailleurs, les forêts aléatoires représentent une amélioration du *bagging* appliquée aux arbres de décision CART au travers d'un aléa additionnel. En effet, à l'image de ce dernier, les arbres sont construits sur la base d'échantillons *bootstrap* tandis que les noeuds sont eux définis à l'aide d'un sous-ensemble de variables tirées aléatoirement qui varie tout au long de la création de l'arbre. Cette composante aléatoire supplémentaire permet de rendre les arbres davantage indépendants.

Enfin, le modèle *XGBoost* est une version améliorée du *boosting*. Le *boosting* est comparable au *bagging* dans la mesure où il a également vocation à créer un ensemble de modèles dont les résultats sont consolidés par pondération. Il se distingue néanmoins dans sa manière de créer les modèles. En l'occurrence chaque modèle, à l'exception du premier, prend en compte les résultats du modèle précédent en donnant plus de poids aux observations incorrectement prédites. Par conséquent, l'effort est concentré autour des observations complexes et la consolidation finale réduit le risque de sur-ajustement. Le *XGBoost* consiste pour sa part en l'ajout d'un terme de régularisation à la fonction perte permettant ainsi d'éviter le surapprentissage en réduisant l'impact des ajustements successifs.

Méthodologie et résultats

En préambule, il est nécessaire de préciser que les applications numériques sont effectuées sans tenir compte des sinistres tardifs, à la fois pour les méthodes agrégées et les méthodes de *machine learning*. Ainsi, nous nous attachons à estimer les provisions *IBNER* et les erreurs de prédiction associées.

Modèle AXA France

Pour commencer, il s'agit de mettre en place le modèle AXA France pour servir de base de référence au modèle de *machine learning* que nous prévoyons de créer. S'agissant de l'estimation de la charge ultime et des erreurs de prédiction pour les sinistres survenus en 2014, nous effectuons les étapes suivantes :

- en prenant un triangle vu à fin 2014, c'est-à-dire, avec une seule année d'inventaire, nous appliquons la modèle d'ajustement en retenons pour hypothèses une évolution de coût moyen annuelle de 5% et les évolutions de fréquences passées pour obtenir une charge ultime estimée à 304 M€ ;
- nous appliquons la méthode par tranche de coûts sur les triangles à des visions différentes. Les résultats obtenus des années d'inventaires 2015 et 2019 sont affichés en table 1.

Survenance 2014 - Vision fin 12/2015	Charge D/D	Charge ultime	Charge ultime avec TF*	Réserves	Réserves avec TF*	MSEP Mack	MSEP M&W *(observables approchées)*
T12	104 725	95 862	95 862	-8 863	-8 863	5 999	4 442
T3	35 259	42 789	42 789	7 530	7 530	5 680	3 707
T45	90 579	161 869	176 781	71 290	86 202	29 735	12 291
Total	**230 563**	**300 520**	**315 432**	**69 957**	**84 869**		

**Tail factor*

Survenance 2014 - Vision fin 12/2019	Charge D/D	Charge ultime	Charge ultime avec TF*	Réserves	Réserves avec TF*	MSEP Mack	MSEP M&W *(observables approchées)*
T12	90 189	89 290	89 290	-899	-899	1 772	1 069
T3	60 296	60 867	60 867	571	571	3 606	1 373
T45	121 825	168 551	178 304	46 726	56 479	19 422	10 115
Total	**272 310**	**318 708**	**328 461**	**46 398**	**56 151**		

**Tail factor*

TABLE 1 – Estimations de la charge ultime par la méthode par tranche de coûts, estimations des erreurs de prédictions à l'ultime (méthode de Mack) et à horizon un an (méthode de M&W) pour la survenance 2014 à partir des trois triangles à dates d'inventaire 2015 et 2019 (montants en k€)

La table 1 illustre l'importante évolution de l'estimation de la charge, passant de 300 M€ en 2015 à 319 M€ en 2019 hors marge de prudence. Ceci est également le reflet du fait que les triangles (notamment T3 et T45) ne satisfont pas pleinement les hypothèses émises par le modèle de Mack.

Enfin, les estimations des erreurs de prédictions, aussi bien à l'ultime qu'à horizon un an, sont obtenues pour chaque triangle. Dans notre cas de figure, la variance totale est composée d'une variance intra-triangles et d'une variance inter-triangles. Nous tenons compte uniquement de la variance intra-triangles. Ainsi, la somme des erreurs de prédiction estimées est retenue comme référence de comparaison par rapport aux erreurs obtenues par les modèles de *machine learning*. L'erreur de prédiction à l'ultime en 2015 est estimée à 41 M€ versus 20 M€ à horizon un an.

Modèle de *machine learning*

La mise en place du modèle de *machine learning*, pour sa part, nécessite un travail préalable de construction et d'analyse de la base de données consistant en un enchaînement d'étapes.

- **Construction et analyse exploratoire de la base de données**
  Cette analyse illustre notamment l'importante dispersion de la variable « charge D/D nette de recours vue à fin 2019 », évoluant entre -53 977€ et 13 033 905€, avec une moyenne de 12 242€. Cette dispersion est engendrée par la présence de sinistres graves, plus de trois quart des sinistres ayant une charge inférieure à 1 668€.
  Nous avons également pu identifier des corrélations entre la charge et certaines variables explicatives, en l'occurrence le mode d'indemnisation du sinistre, le nombre de dommages corporels, la présence d'un contentieux, le niveau de gravité du sinistre.

- **Définition de deux échantillons adéquats pour l'apprentissage et le test**
  L'échantillon d'apprentissage se compose des sinistres survenus entre 2010 et 2013 tandis que les sinistres survenus en 2014 composent l'échantillon de test.

- **Calcul des poids *IPCW* à partir de l'estimateur de Kaplan-Meier**
  Le taux global de censure des données est de 4,5% en volume et de 42,8% en charge, soit une charge D/D nette de recours de 545 M€ à fin 2019. Pour parer au biais induit par cette censure, nous calculons les poids *IPCW* au niveau de la base d'apprentissage.

- **Calibrage et entraînement des modèles sur l'échantillon d'apprentissage**
  Les modèles sont calibrés par validation croisée et entraînés sur l'échantillon d'apprentissage avec prise en compte des poids calculés. Nous appliquons ensuite les modèles obtenus sur l'échantillon de test.

- **Mesure et comparaison de la performance de prédiction des modèles testés**
  La comparaison de la performance des trois modèles sur les sinistres survenus en 2014 et clos à fin 2019 nous permet de constater que le modèle *XGBoost* obtient les meilleurs résultats. En effet, le modèle ne surestime que de $1,5\%$ la charge réelle. Ces estimations constituent un tour de force dans la mesure où elles ne reposent que sur un faible historique de données, en l'occurrence deux années d'inventaire.

- **Application du *bootstrap* pour l'estimation de la distribution empirique de la charge ultime des sinistres survenus en 2014**
  Cette ultime étape nous permet d'estimer les erreurs de prédiction des provisions afin de les comparer à celles des méthodes agrégées. Nous avons pour cela recours à la méthode de *bootstrap* avec 10 000 itérations pour chaque modèle, aboutissant à une distribution de la charge ultime, présentée en figure 1, dont l'erreur de prédiction dispose d'un niveau de confiance de 99,5%.

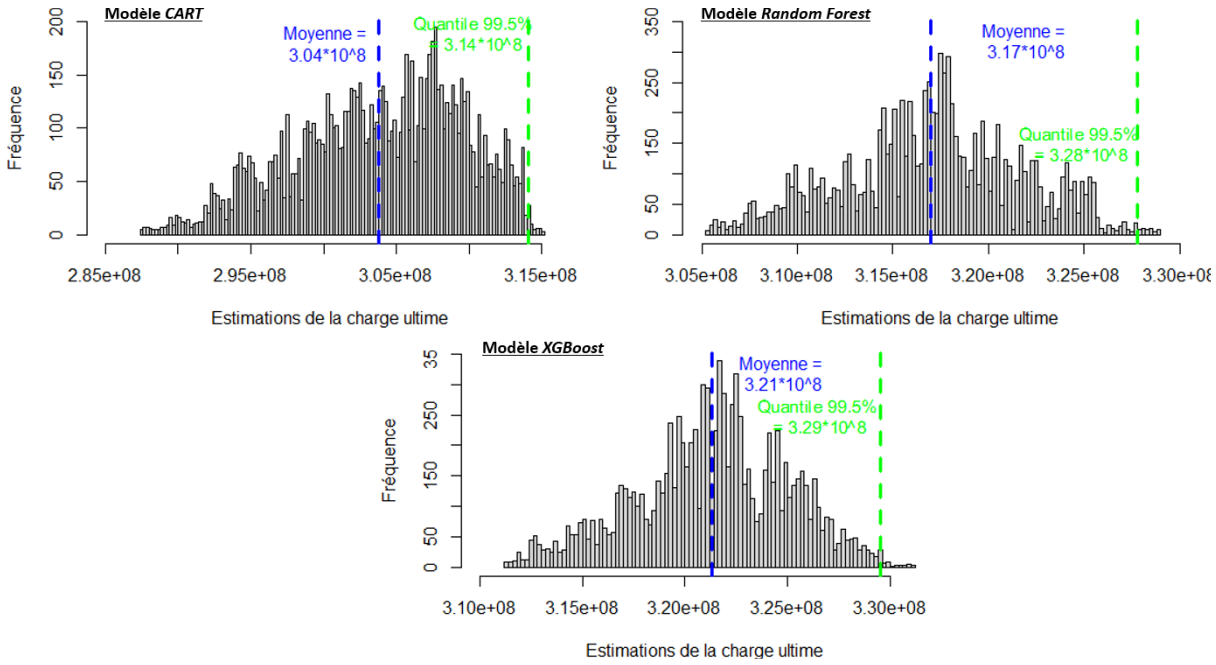


FIGURE 1 – Distribution des estimations de la charge ultime de la survenance 2014 obtenue par la méthode *bootstrap - 10 000 itérations*

La simulation de la distribution empirique de la charge ultime pour chaque modèle nous permet d'apprécier le gain en performance pour les modèles ensemblistes testés, notamment celui du *XGBoost*. En effet, ce dernier se distingue par une meilleure anticipation de la charge ultime, en particulier pour les sinistres graves : tandis que le modèle AXA France n'estime que 300 M€ en 2015 avant d'aboutir à 319 M€ en 2019, le modèle *XGBoost* obtient 321 M€ dès 2015, autrement dit avec 4 années d'avance.

Aussi, nous notons un gain conséquent en précision, de l'ordre de 37 M€ : tandis que le modèle AXA France obtient une erreur de prédiction à l'ultime estimée à 41 M€ en 2015, le modèle *XGBoost* n'est lui qu'à 4 M€ à la même période. Ce gain entraînera notamment la diminution du *SCR* de réserves.

Par ailleurs, l'analyse de l'importance des variables du *XGBoost* met en exergue que celles à plus forte contribution sont les variables montants. Il s'agit des charges D/D, des règlements et des évaluations de règlements à différentes vision jusqu'à la fin de la deuxième année d'inventaire. Ce constat nous mène vers la conclusion que les variables montants, tenant compte implicitement des circonstances du sinistre, l'emportent sur les variables liées au contexte propre du sinistre.

Conclusion

En conclusion, nous retenons le modèle *XGBoost* compte tenu de ses performances et des résultats obtenus. En effet, il permet d'obtenir une estimation à 321 M€ dès 2015 tandis que le modèle AXA France n'aboutit à une estimation équivalente qu'en 2019. De plus, le gain en précision est estimé à 37 M€, allant d'une erreur de prédiction de 41 M€ pour le modèle AXA France à 4 M€ pour le modèle *XGBoost* en 2015. Par conséquent, nous constatons que ce modèle nous permet d'anticiper l'estimation de la charge ultime tout en limitant la volatilité associée.

Dans la continuité des travaux menés dans le cadre de ce mémoire, d'autres axes d'étude peuvent être abordés. Il serait ainsi pertinent de tester le modèle retenu sur différentes survenances afin de conforter les performances obtenues. Aussi, nous pourrions quantifier l'impact du gain en volatilité sur la valeur du *SCR* de réserves en utilisant le modèle interne AXA France. Il serait également intéressant d'expérimenter ce modèle sur les sinistres corporels hors automobile. Enfin, dès lors qu'un historique conséquent sera disponible, il pourrait s'agir d'inclure le taux d'AIPP en tant que variable explicative.

S'il est aujourd'hui périlleux de s'avancer sur un éventuel remplacement à court terme des méthodes agrégées utilisées par AXA France en faveur du modèle *XGBoost* retenu, compte tenu de leur simplicité d'interprétation et de compréhension par les équipes de contrôle interne et externe, il ne fait nul doute que ce dernier serait à minima amené à être utilisé à titre de comparaison pour challenger le résultat des méthodes actuelles, en remontant notamment des points d'alerte. Pour autant, en cas de performances particulièrement attractives dans la durée, il n'est pas à exclure que ce modèle puisse un jour remplacer le modèle actuel.

Executive summary

Context and problematic

Reserving which refers to the appraisal required for the settlement and administration of claims, is at the core of the insurance, given the significant economic and prudential stakes at issue. Thus, insurers are constantly seeking to enhance their reserving models with the aim to increase accuracy. This gain in accuracy leads to a better stability of the ultimate cost estimations and hence of the Claim Payment Provisions (PSAP in french).

We are thereby limiting the phenomena of over-reserving, under-reserving and the side effects arising, namely :

- in the context of under-reserving, in conjugation with the apparent risk of the insured's insolvency and non-indemnification, the malis obtained during the re-estimations in the subsequent years exemplifies the aftereffect.
- within the framework of over-reserving, in addition to the useless immobilisation of the resources for the insureds and the shortfall for the insurer, the negative effect is illustrated by the taxation of bonis during the re-estimation throughout the following years.

This problematic reveals itself to be more complex and important in the context of long-term development guarantees, similarly to automobile bodily liability warranty covered in this brief. This warranty enables the insurer to assess the compensation of all victims of bodily injury for which the insured is found guilty, equal to the loss sustained. The cost of this coverage's claims varies over time under the influence of several factors, including the stabilization of the victim's condition, the mode of indemnity in capital or in annuity or court's decision in case of litigation.

The aim of this dissertation is to call into question the model implemented by AXA France for PSAP's estimation for the automobile bodily injury coverage. In fact, the current model rests on aggregated approaches which don't take advantage of the huge amount of data available to the company. Hence, it would consist in finding the machine learning model which makes the most of the amount and fine granularity of data at hand, to come up with more precise reserving estimates.

AXA France model - aggregated methods

A variety of aggregated strategies exists which can be put into two categories : on one hand, deterministic methods which are able to estimate the amount of reservings, on the other hand, stochastic techniques which can provide a prediction of error at different horizons besides an estimation of the amount of reservings.

In this dissertation, three widely used methods have been selected and for which we explain the theoretical basis. These are the deterministic Chain-Ladder method, and the stochastic methods of Mack and Merz & Wüthrich. The latter provides an estimation of the provisions equivalent to the Chain-Ladder, in addition to the prediction errors at the ultimate and at the one-year horizon, respectively.

To start with, Chain-Ladder approach assumes that past observed cost rates will be reproduced in the future for the estimation of the amount of reservings. On the other hand, the fundamental strength of Mack and Merz & Wüthrich methods lies in their ability to evaluate the uncertainty of the previsions

through closed formulas. Although, Merz & Wüthrich distinguishes itself from the other one by its capacity to evaluate them over one-year horizon, in accordance with "Solvency II" directive.

Within the framework of aggregated methods, the data is represented in terms of run-off triangles. Rows correspond to the exercises of reattachment of claims and columns correspond to periods of development. For our cause, we take the years 1999 to 2014 as the claim incidence years and at least 16 years of development. These claims from the AXA France portfolio relate to personal and professional motor liability coverage for all distribution networks, i.e. agents, brokers and employees.

The reserving model produced by AXA France is based on the approaches explained above. Furthermore, in an effort to harmonise the population of claims and thus to meet as far as possible the assumptions of Chain-Ladder and Mack models, the estimation of the ultimate cost by AXA France is based upon three distinct triangles : attritional, severe and drastic.

Thus, the estimation of the ultimate cost can be summed up as follows :

- **Triangle 12 (cost D/D $\leq$ 150*k*€) :** application of Chain-Ladder method.
- **Triangle 3 (cost D/D $\in$] 150k€ ; 750k€]) :** application of Chain-Ladder method.
- **Triangle 45 (charges D/D $>$ 750*k*€) :** as a first step, we apply Mack method, to be able to preserve, as a second step, the 75 % quantile of a log-normal distribution of reservings with expectation $\mu = ln(\hat{R}) - \frac{\sigma^2}{2}$ and variance $\sigma^2 = ln(1 + \frac{\hat{MSEP}(\hat{R})}{\hat{R}^2})$. This approach entails to add a prudence adjustment cost to the cost estimated by Chain-Ladder method. This is required due to the long development of this type of claims which may owing to their atypicality, their severeness, the estimated life expectancy in the case of an annuity, or the waiting for court's decision.

It should however be noted that cost slices method, described above, is not used to estimate the cost in the first year inventory, but rather a cost-adjustment mechanism based on frequency/average cost is used. The goal of this approach is to mitigate the alterations of the cost D/D observed during the first year and which can vary substantially from one occurrence to another.

Proposed model - machine learning algorithm

With a view of handling the lack of information within the run-off triangles, we strive to acquire as much data as possible in order to create one data set. Any variable deemed relevant to the estimation of the target, which is, in this context, the net ultimate cost, will be added.

Three different data sets are aggregated to create the final data set, specifically the contracts data set, the claims data set and the victim files data set. As a result, we collected 88 616 claims occurring between 2010 and 2014, for a yearly average of over 17 000 claims. The latter are about the distribution channels, such as agent' network, employees' network and finally brokers' network. On the other hand, the target variable relates to a vision set at the end of 2019, while the explanatory variables are set at the second year of the inventory.

It is worth mentioning that our target variable is right-censored. As a matter of fact, there are still pending claims in 2019, thereby we do not have the ultimate cost of the latter. Due to limiting our data set up until 2019 and the inability to wait for the closing out of all these claims, contrary to some short development branches, we work with a right-censored target. Actually, the branch «bodily automobile RC » is a long one by nature, thus waiting for the closing of all the claims may require more than thirty years.

It goes without saying that this censoring impacts the training of our machine learning models. To tackle this issue, *IPCW* weights computed using Kaplan-Meier estimator are assigned to all samples correcting the bias induced by our censored data. This censorship effect is thus compensated by assigning greater weights to closed out claims as the cost increases. Still pending claims, in turn, are assigned a zero weighting. Nevertheless, the latter are not ignored since they contribute to the computation of the weights of closed claims.

Prior to the explanation of the methodology adopted and the corresponding results, we aim to give a quick overview of the machine learning models that were tested.

For openers, Classification and Regression Trees or CART for short are a fruitful data mining strategy. It consists in repeatedly partitioning the data into multiple sub-regions. The technical term for this approach is recursive partitioning. They are characterized by their interpretability since the end model is a recursive sequence of binary division rules. Nonetheless, these methods are inherently irregular and unstable, thus allowing the rise of so-called ensemble methods.

Furthermore, random forest is an enhanced version of bagging with added randomness applied to CART trees. Indeed, the trees in this model are constructed using procedure of CART on sub-samples and the nodes are developed by picking randomly a sub set of variables which varies throughout the process. Thanks to the randomness added to the whole process, uncorrelated and independent trees are created.

Ultimately, *XGBoost* is an improved version of *boosting*. *Bagging* and *boosting* are similar in that they combine predictions from different models created to compute the results. But they differ in the technique used to carry out the construction of the models. In this case, each model, except for the first one, takes into account the results obtained by the previous one assigning more weights to wrongly predicted observations. Hence, more focus is placed on complex observations while reducing overfitting. *XGBoost* is a variant of *gradient boosting* model with a regularisation term added to the loss function resulting in a new cost function. This aspect prevents the model from overfitting by reducing the impact of successive adjustments.

Results and methods

As a prologue, it is essential to state that *machine learning* and the aggregated methods are performed without taking into account late claims. Thus, we concentrate on estimating the *IBNER* reserves and the associated prediction errors.

AXA France model

To begin with, we established the AXA France model to serve as a baseline for the *machine learning* model that we intend to develop. For the claims that happened in 2014, we perform the following processes to estimate the ultimate net cost and forecast errors :

- Taking a triangle set at the end of 2014, i.e., with only one year of inventory, we apply the adjustment model assuming an average annual cost change of 5% and changes in frequencies passed to obtain an ultimate charge estimated at 304 M€ ;

- We apply cost slices method on the triangles to different visions. The results obtained from the 2015 to 2019 inventory years are displayed in table 2.

Loss date 2014 - Vision to 12/2015	Cost D/D	Ultimate cost	Ultimate cost with TF*	Reserves	Reserves with with TF*	MSEP Mack	MSEP M&W *(approximated observables)*
T12	104 725	95 862	95 862	-8 863	-8 863	5 999	4 442
T3	35 259	42 789	42 789	7 530	7 530	5 680	3 707
T45	90 579	161 869	176 781	71 290	86 202	29 735	12 291
Total	**230 563**	**300 520**	**315 432**	**69 957**	**84 869**		

**Tail factor*

Loss date 2014 - Vision to 12/2019	Cost D/D	Ultimate cost	Ultimate cost with TF*	Reserves	Reserves with with TF*	MSEP Mack	MSEP M&W *(approximated observables)*
T12	90 189	89 290	89 290	-899	-899	1 772	1 069
T3	60 296	60 867	60 867	571	571	3 606	1 373
T45	121 825	168 551	178 304	46 726	56 479	19 422	10 115
Total	**272 310**	**318 708**	**328 461**	**46 398**	**56 151**		

**Tail factor*

TABLE 2 – Estimates of ultimate cost using cost slices method, estimates of prediction errors at the ultimate (Mack method) and one-year horizon (M&W method) for the 2014 occurrence from the three triangles with inventory dates of 2015 and 2019 (amounts in k€)

Table 2 illustrates the significant rise of the cost estimate, going from 300 M€ in 2015 to 319 M€ in 2019 excluding the prudence adjustment costs. This also reflects the fact that the triangles (notably T3 and T45) do not fully satisfy the assumptions made by Mack's model.

Finally, estimates of the prediction errors, both at the ultimate and at the one-year horizon, are obtained for each triangle. The total variance is composed of an intra-triangle variance and an inter-triangle variance. In our case we only take into account the intra-triangle variance. As a result, the total amount of the estimated prediction errors is used as a comparison reference to the errors obtained by the *machine learning* models. The ultimate prediction error in 2015 is estimated at 41 M€ versus 20 M€ at the one-year horizon.

Machine learning model

A preliminary work, which consists of data set creation and exploratory analysis, must be conducted following the sequence of steps set out below.

- **Construction and exploratory analysis of the data set**
  This data inspection shows the wide dispersion of the net cost variable D/D set at the end of 2019, with a range of -53 977€ to 13 033 905€ and average of 12 242 €. The presence of severe claims, with more than three quarters of the claims having an expense of less than 1 668 €, causes this dispersion. In addition, this dispersion allowed us to find a correlation between the cost and some explanatory variables, in this case the method of compensation for the loss, the number of bodily injuries, the presence of litigation, the level of severity of the loss.

- **Splitting the data set into training and test set**
  The training set consists of claims that occurred between 2010 and 2013 while the test set is consti-

tuted by claims that occurred in 2014.

- **Computation of IPCW weights using Kaplam-Meier estimator**
  Censored data has a volume rate of 4,5 % and a cost rate of 42,8 %, for a net D/D recourse cost of 545 M€ at the end of 2019. Using the training set, we correct the bias induced by the censored data by computing the IPCW weights.

- **Hyperparameter tuning and training the model in the training set**
  Cross validation is used for hyperparameter tuning and on the other hand, the model is trained in the weighted training set. Then we apply the fine-tuned model to the test set.

- **Measurement and comparison of the prediction performance of the tested models**
  The results of the comparison carried out confirms that *XGBoost* outperforms all the models applied to claims occurred in 2014 and closed at the end of 2019. In point of fact, the model overestimates the actual cost by only $1,5\%$. This estimation highlights the use of only a small proportion of the historical data (two years of inventory).

- **Estimating the empirical distribution of the ultimate cost occurred in 2014 using *bootstrap***
  In this last step, we compute the prediction errors to compare them to those of the aggregated methods. For each model, we use the bootstrap method with 10 000 iterations. Resulting in a distribution of the ultimate cost with a confidence level of 99,5 % in the forecast error, as shown in figure 2.

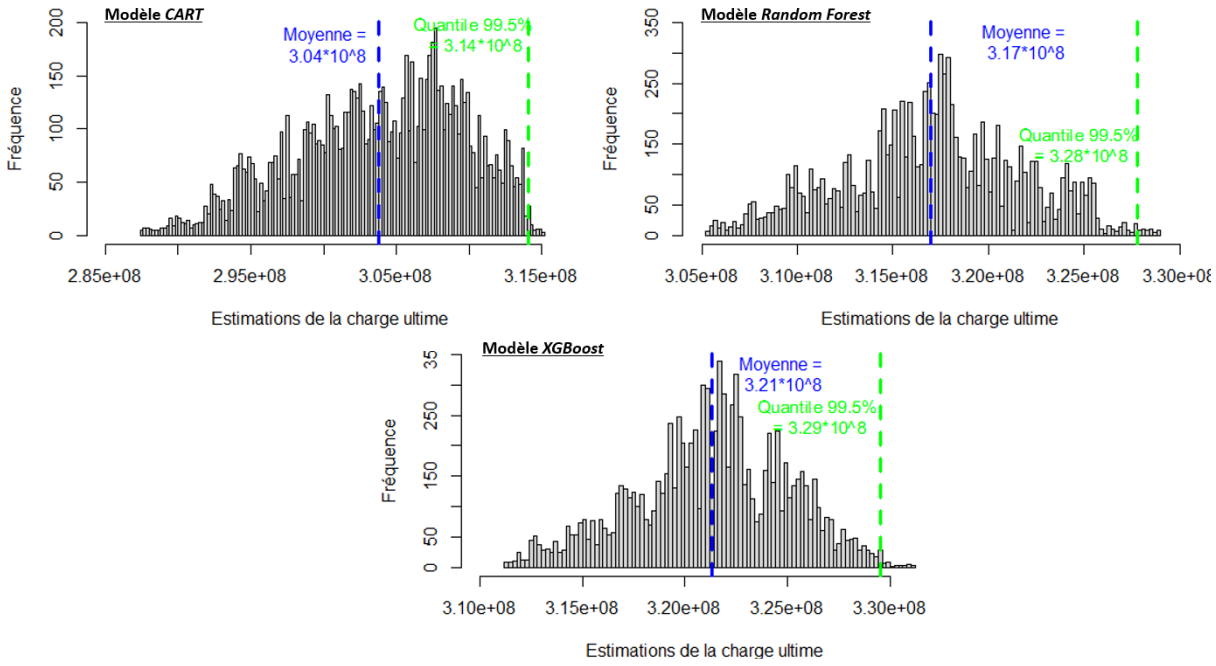


FIGURE 2 – Distribution of ultimate costs estimations with loss date 2014 obtained by *bootstrap - 10 000 iterations*

The accuracy increase brought by ensemble models is possibly appreciated by simulating the empirical distribution of the ultimate cost for each of them, especially the one of the *XGBoost*. As a matter of fact, the latter distinguishes by a more accurate prediction of the ultimate cost, especially for the severe claims : while the model implemented by AXA France only estimates 300 M€ in 2015 before reaching 319 M€ in 2019, *XGBoost* estimates 321 M€ since 2015, in other words 4 years ahead of time.

Moreover, we observe a remarkable gain in precision, about 37 M€ : in 2015, Mack's model had a prediction error of 41 M€, while *XGBoost* obtained just 4 M€ for the same year. Hence, the *SCR* reserve will be reduced as a result of this gain.

Furthermore, the feature importance computed by *XGBoost* underlines that the variables with the greatest contribution are the amount variables. These are the D/D costs, the settlement and settlement assessments set at various views until the end of the second inventory year. This leads us to the conclusion that the amount variables that consider the claim's circumstances, come out ahead over the variables related only to the specific claim's context.

Conclusion

In conclusion, we retain the *XGBoost* model given its performance and the results obtained. Indeed, it allows us to obtain an estimate of 321 M€ as of 2015, whereas the AXA France model only produces an equivalent estimate in 2019. Moreover, the gain in accuracy is estimated at 37 M€, going from a prediction error of 41 M€ for the AXA France model to 4 M€ for the *XGBoost* model in 2015. As a result, this model enables us to better anticipate the estimation of the ultimate cost while limiting the associated volatility.

To pursue the study carried out within this brief, it would be relevant to test the chosen model on different loss dates in order to confirm the performance obtained. We could also quantify the impact of the gain in volatility of *SCR* of reserves by using the internal AXA France model. In the meantime, it would be interesting to test this model on non-automobile bodily injury claims. Finally, as soon as a substantial history is available, we can include the AIPP rate as an explanatory variable.

As the existing model is easy to be interpreted and comprehensible by AXA France internal and external control teams, it would be difficult to see a rapid demise of the aggregate methods to XGBoost model. However, there is no doubt that this model will be at least used to challenge the results of the current methods raising warning points. Nevertheless, in case of a constant great performance over time, the move to this new model would be no longer excluded by AXA France directors.

Remerciements

Je tiens tout d'abord à remercier Pierre-Louis Blanc, Véronique Brignolas et Boris Mihatsch pour leur confiance et leur soutien au sein d'AXA France, et sans qui cette aventure n'aurait pas été possible.

Je souhaite remercier tout particulièrement Said Chafni et Quentin Guibert, respectivement tuteurs entreprise et académique, pour leur soutien, leurs conseils et leur disponibilité tout au long de la réalisation de ce mémoire.

Je tiens à remercier Clémentine Vie ainsi que l'ensemble de mes collègues au sein d'AXA France pour leur soutien et leurs chaleureux encouragements.

J'adresse par ailleurs mes remerciements à Caroline Hillairet et Wissal Sabbagh, pour leur suivi et leur accompagnement personnalisé tout au long de la formation à l'ENSAE.

Je voudrais plus généralement exprimer ma profonde reconnaissance au corps professoral de l'ENSAE pour la qualité de son enseignement et ce, malgré une période profondément bouleversée par la crise sanitaire.

Enfin, je ne saurais conclure sans un mot pour ma famille et mes proches, qui m'ont continuellement encouragée et soutenue durant ces mois de travail.

Table des matières

Introduction

Le secteur de l'assurance, secteur clé de l'économie, est soumis à des contraintes prudentielles importantes. Ces contraintes visent à protéger les droits des assurés en évitant la défaillance des assureurs. Cela se traduit, entre autre, par la constitution de provisions pour sinistres à payer (PSAP) relatifs aux sinistres, déclarés ou non, survenus lors des années passées et n'ayant pas fait l'objet d'une indemnisation complète à la fin de l'exercice.

Ces provisions constituent une part importante du passif dans le bilan des assureurs, il est donc nécessaire de les estimer au plus juste :

- un sous-provisionnement, outre le risque évident d'insolvabilité et de non indemnisation des assurés, gonfle superficiellement les bénéfices de l'année et se répercute sur les années suivantes au travers de malis lors des ré-estimations ;
- un sur-provisionnement quant à lui, outre son inutilité pour les assurés, entraine une taxation sur les bonis, lors des ré-estimations, ainsi qu'une immobilisation inutile des ressources pouvant faire l'objet de bénéfices en cas de placement.

Afin d'estimer ces provisions, les assureurs plébiscitent généralement les méthodes agrégées et en particulier les méthodes de Chain-Ladder, de Mack et de Merz & Wüthrich qui sont utilisées respectivement pour l'estimation de la charge ultime, l'estimation de l'erreur de prédiction de la charge à l'ultime et l'estimation de l'erreur de prédiction à horizon un an. Le succès de ces méthodes repose essentiellement sur leur simplicité. Néanmoins ces dernières présentent plusieurs inconvénients, notamment la perte d'information liée à l'agrégation des données ou encore l'utilisation d'hypothèses qui ne sont pas toujours satisfaites du fait, par exemple, d'un changement dans la réglementation ou dans le mode de gestion.

Dans un contexte marqué par l'essor des technologies de gestion de la donnée, les assureurs investissent de plus en plus dans la collecte de données et leur exploitation. Les méthodes de *machine learning* pour l'estimation des provisions, encore rarement utilisées dans la pratique car jusqu'alors peu matures et nécessitant une quantité considérable de données, semblent désormais pertinentes, car à la portée des assureurs et permettraient d'améliorer l'estimation de la charge ultime et de minimiser la volatilité des estimations de provisions.

Des publications et mémoires existants abordent la problématique du provisionnement et exposent des approches et applications différentes. Nous nous attachons dans le cadre de ce mémoire, dans la continuité du mémoire de (Servel, 2020), à confronter le modèle retenu pour l'estimation de la charge ultime, en introduisant également les erreurs de prédiction à l'ultime et à horizon un an. Pour ce faire, nous testons de nouveaux modèles et de nouvelles variables sur un plus large périmètre incluant, outre le réseau des agents, le réseau des courtiers et des salariés.

Après avoir, dans le chapitre 1, remis en contexte le sujet en introduisant les concepts clés en assurance, les notions liées au provisionnement et fourni un aperçu synthétique de la directive Solvabilité II, nous établissons dans le chapitre 2 le périmètre de l'étude au sein de la garantie RC corporelle en automobile et décrivons les spécificités des sinistres corporels. Par la suite, il s'agit dans le chapitre 3 de présenter le modèle adopté par AXA France ainsi que les trois principales méthodes agrégées utilisées pour les estimations. Nous y exposons les résultats associés à l'application de chacune de ces méthodes. Pour finir, nous introduisons dans le chapitre 4 la théorie des différents modèles de *machine learning* utilisés, nous présentons les résultats de leur application et les comparons aux résultats obtenus suite à l'application des méthodes agrégées.

1 Mise en contexte

1.1 Le secteur de l'assurance

1.1.1 Généralités

L'assurance est définie comme étant une opération d'échange à travers laquelle l'assureur s'engage à assurer une prestation au profit de l'assuré lors de la survenance d'un risque en contrepartie du paiement d'une prime ou d'une cotisation. Cet accord se matérialise sous forme d'un contrat d'assurance dont le risque encouru par l'assuré en constitue l'objet. L'assurance est un secteur d'activité au **cycle de production inversé**, ainsi le flux de la ressource (les primes) précède le flux de la dépense (les sinistres). De ce fait, les assureurs sont tenus de constituer des provisions pour pouvoir faire face à leurs engagements.

Il convient de distinguer deux formes principales d'assurance, en l'occurrence l'assurance vie et l'assurance non-vie. Les principales différences entre ces deux grandes familles sont :

- **le fait générateur d'un sinistre :** en assurance vie, il est en lien avec la survie ou le décès de l'assuré, alors qu'en non-vie, il est en lien avec un événement extérieur et indépendant de l'assuré ;
- **le montant d'indemnisation en cas de survenance d'un sinistre :** en assurance vie, le montant est connu à l'avance contrairement à l'assurance non-vie.

1.1.2 L'assurance non-vie

Les assurances non-vie, aussi appelées assurances IARD (Incendie-Accidents-Risques Divers), couvrent les branches précisées dans l'article $R321-1$ du Code des Assurances réparties en deux grandes familles :

- **les assurances de biens et de responsabilité,** ont pour rôle de protéger le patrimoine de l'assuré. Les assurances de biens permettent de couvrir les biens matériels que détient l'assuré ainsi que les préjudices qui s'en suivent. Les assurances de responsabilité, quant à elles, permettent de couvrir les dommages matériels ou corporels causés involontairement à un tiers ;
- **les assurances individuelles accident et les assurances de santé**, font partie des assurances de personnes et ont pour objectif de garantir la personne et non son patrimoine.

1.1.3 Quelques chiffres

En assurance non-vie

La figure 1.1 montre qu'en 2020, les cotisations en assurance non-vie représentent 42% des cotisations totales au sein du marché français, soit un montant total de **84,4 Mds€** : 59,2 Mds€ en assurance de biens et responsabilité et 25,2 Mds€ en assurance maladie et accidents corporels. Cette année a été marquée par un ralentissement des cotisations accompagné d'une hausse des prestations pour atteindre un montant total de **62,7 Mds€** : 42,9 Mds€ en assurance de biens et responsabilité et 19,8 Mds€ en

assurance maladie et accidents corporels. Toutefois, la solvabilité des sociétés non-vie reste solide avec un ratio de solvabilité à 265% en 2020.

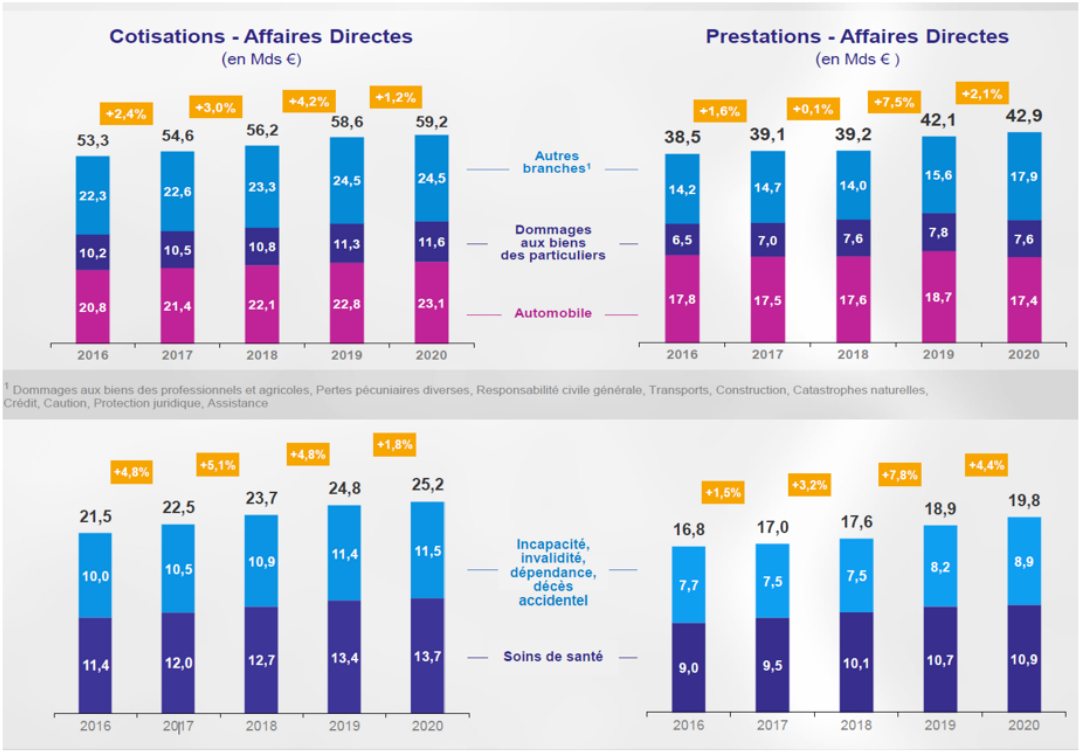


FIGURE 1.1 – Evolution des cotisations et des prestations en assurance non-vie - Particuliers et professionnels - (FRANCE ASSUREURS, 2021b)

Zoom sur la garantie RC corporelle automobile

Cette partie a pour rôle de présenter quelques chiffres clés relatifs à la garantie RC corporelle automobile, traitée dans le cadre de ce mémoire. Il faut tout d'abord noter que les cotisations de la branche automobile des particuliers et professionnels en 2020 représentent $27,4\%$ de l'ensemble des cotisations en assurance non-vie, soit l'équivalent de $23,1$ Mds€ (cf. figure 1.1). Les cotisations en responsabilité civile en 2020 représentent pour leur part 40% de l'ensemble des cotisations, soit l'équivalent de 8,4 Mds€ tel qu'indiqué en figure 1.2. Ces indicateurs demeurent relativement stables tout au long de la période d'analyse allant de 2016 à 2020.

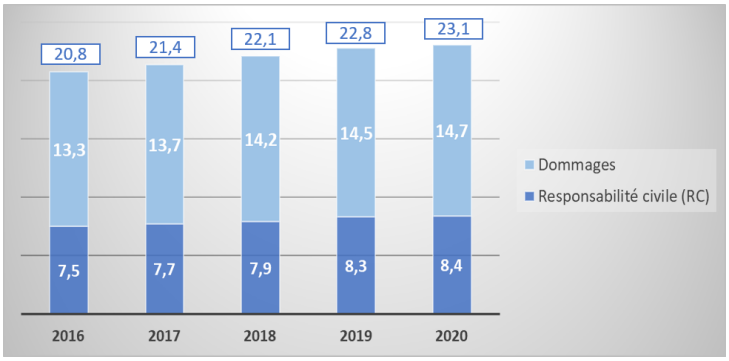


FIGURE 1.2 – Evolution des cotisations en assurance automobile selon le type de garantie - Particuliers et professionnels - (FRANCE ASSUREURS, 2021a)

Au travers des évolutions de fréquences des différentes garanties de la branche automobile communiquées par France Assureurs, présentées en figure 1.3 et en annexe A.1, il est possible de constater une baisse de la sinistralité. Cette baisse s'explique notamment par les nombreuses initiatives de prévention et de sécurité routière ainsi que par le développement des voitures équipées de systèmes *ADAS* d'aide à la conduite.

Ainsi, il est possible de noter une évolution de fréquence de -1,2% en 2019 vs. 2018 au niveau de la RC corporelle. Cette baisse s'accentue en 2020 pour atteindre -26,8% en raison du faible usage des véhicules induit par les confinements liés à la pandémie de Covid-19 (cf. annexe A.1).

Néanmoins, malgré une sinistralité à la baisse, le coût moyen des sinistres est quant à lui en hausse constante : il augmente depuis 2010 de 5,7% par an en moyenne pour les accidents corporels et de 3,2% pour les accidents matériels.

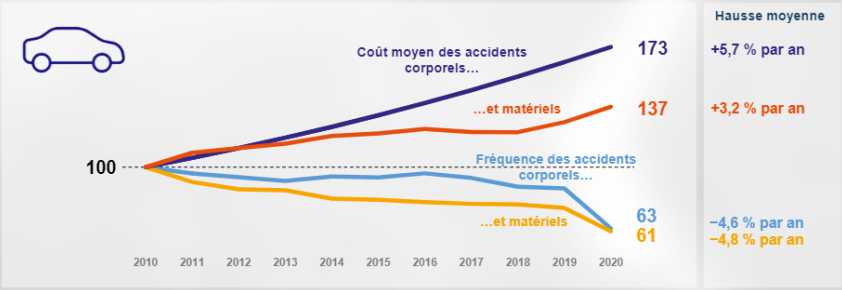


FIGURE 1.3 – Évolution des coûts moyens et des fréquences des accidents matériels et corporels - Base 100 en 2010 - (FRANCE ASSUREURS, 2021b)

En ce qui concerne AXA France, la sinistralité des particuliers et professionnels évolue conformément aux tendances du marché. Ainsi, nous notons une baisse de fréquence de -1,8% en 2019 vs. 2018 parallèlement à une hausse du coût moyen des sinistres clos de 6,8% et une évolution du coût moyen final prévisible à 3,3% sur la même période.

Enfin, il est nécessaire de souligner que malgré un faible volume de sinistres en RC corporelle automobile au sein d'AXA France, de l'ordre de 2% en survenance 2019, cette branche concentre une part importante de la charge. En effet, comme précisé en figure 1.4, cette dernière représente 20% de la charge D/D, ce qui justifie l'intérêt qui lui est porté dans le cadre de ce mémoire, ainsi qu'à son provisionnement.

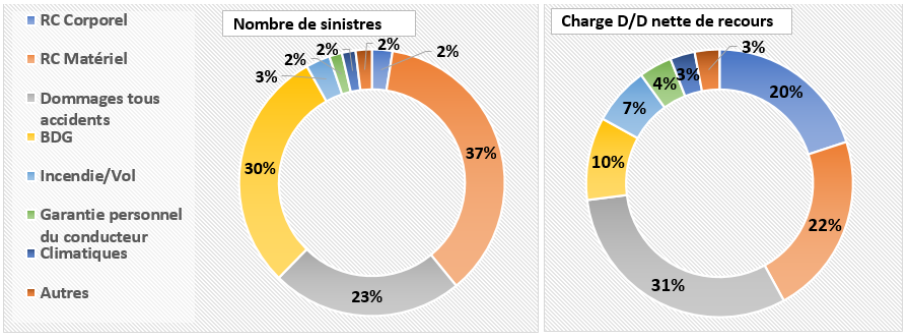


FIGURE 1.4 – Répartition du nombre de sinistres et de la charge D/D selon les garanties en assurance automobile - survenance 2019 vision à fin 2019 (chiffres AXA France PP)

1.2 Le provisionnement en non-vie

Après avoir brièvement rappelé en quoi consiste l'activité de l'assurance non-vie et avant d'entamer la phase de modélisation, il est important de bien cerner la problématique du provisionnement dans ce domaine d'activité ainsi que les contraintes et enjeux qui y sont rattachés. Ce second chapitre aura donc pour mission de mettre en lumière, dans un premier temps, le rôle fondamental des provisions ainsi que l'importance de les prédire de façon la plus juste possible. Dans un second temps, il s'agira de présenter une vision globale du processus de gestion des sinistres en non-vie et la constitution des provisions. Dans un troisième temps, l'objectif sera de définir les principaux types de provisions techniques pour enfin aboutir dans un quatrième et dernier temps, à la description du cadre réglementaire et comptable dans lequel évoluent les provisions.

1.2.1 Objectifs et enjeux du provisionnement

L'évaluation des provisions nécessaires au règlement et gestion des sinistres est une des étapes clés au sein d'une compagnie d'assurance, notamment de par la nature de son activité caractérisée par l'inversion du cycle de production. En effet, les assureurs définissent les primes et s'engagent à assurer le règlement si présence de sinistres. Ces provisions représentent donc une dette des assureurs envers leurs assurés. Il est donc crucial de s'assurer de maîtriser au mieux le calcul des provisions et de tenter de coller au mieux à la réalité. Il faut également noter que ceci est en adéquation avec les exigences des normes comptables et prudentielles qui sont de plus en plus orientées vers la transparence et l'amélioration des pratiques de provisionnement.

Le sous-provisionnement peut avoir de très lourdes conséquences sur la santé financière d'une entreprise, voire même entraîner sa ruine dans les cas les plus extrêmes. S'il s'agit d'une compagnie de taille importante, sa faillite aurait des conséquences sur l'économie à part entière. Sous oublier que l'incapacité d'une compagnie d'assurance à indemniser ses assurés entraînerait de la défiance et également pourrait les impacter financièrement.

Quant au sur-provisionnement, ce ne serait pas la solution de facilité en soit. Sur-provisionner entraînerait à une utilisation sous-optimale des ressources de la compagnie d'assurance. En effet, l'excédent des sommes mobilisées entant que provisions pourrait être investi et assurer un meilleur rendement. De plus, des provisions trop élevées auront pour conséquence une baisse du résultat de l'assureur accompagnée d'une baisse de la charge fiscale. Ceci a pour conséquence également d'entraîner des taxes sur les bonis. Ce dernier point sera abordé plus en détail dans la suite de ce chapitre.

En définitive, il faut retenir que les provisions, évaluées à leur juste valeur, sont censées refléter le niveau de risque de l'assureur outre toute pratique de « gestion de résultat ».

1.2.2 Généralités sur la gestion de sinistres

Le cycle de vie d'un sinistre

Toujours dans une logique d'appréhender l'intérêt du provisionnement, il est indispensable de bien comprendre comment s'articule la vie d'un sinistre et de bien connaître certaines notions en lien avec la sinistralité et présentées en figure 1.5.

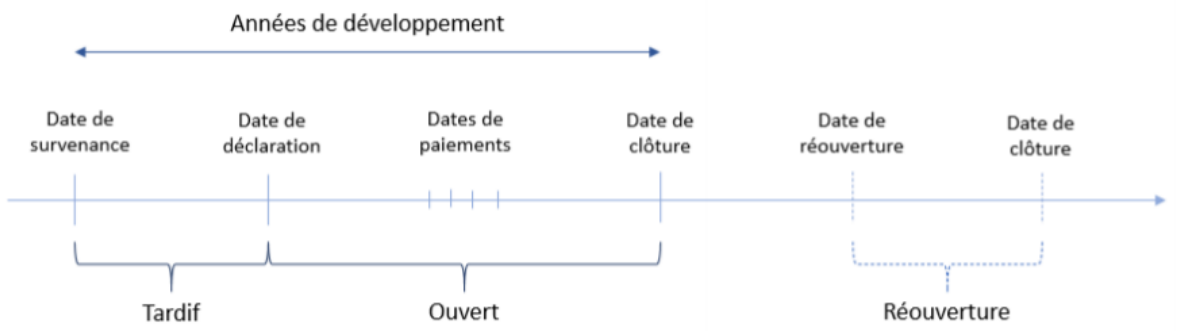


FIGURE 1.5 – Description du déroulement d'un sinistre

Il faut tout d'abord noter que généralement, en assurance non-vie, le coût final d'un sinistre n'est pas connu dès son ouverture. Pour cette raison, les compagnies d'assurances ont l'obligation d'assurer des provisions qui doivent correspondre au mieux aux montants dûs aux assurés au moment du calcul de ces dernières. En effet, les sinistres sont parfois amenés à évoluer dans le temps, ce qui entraîne l'évolution de la charge estimée à la hausse ou à la baisse, en fonction des nouvelles estimations, des nouvelles informations rajoutées aux dossiers comme par exemple des états de santé qui se détériorent pour les sinistres corporels ou encore des preuves de fraude.

Il y a également un second élément à prendre en compte : les **tardifs**. Il s'agit de sinistres survenus mais pas encore notifiés à l'assureur. En effet, lors de la survenance d'un sinistre, l'assuré dispose d'une période légale pour le déclarer à sa compagnie d'assurance. En fonction de la nature du sinistre, les délais invoqués peuvent varier mais doivent toujours être respectés, sous peine de voir l'indemnisation prévue être annulée ou minorée. Bien qu'une compagnie d'assurance ne dispose pas d'informations précises sur le nombre et coût des tardifs au moment du calcul des provisions, elle doit tout de même en tenir compte lors du provisionnement.

Un troisième point à évoquer est la possibilité de présence de **recours**. Ainsi, un assureur peut émettre un recours s'il a indemnisé son assuré et si ce dernier est non responsable ou à moitié responsable et en subir un si c'est plutôt le tiers qui répond à ces critères. Ces derniers génèrent à leur tour une nouvelle source d'entrée ou sortie de flux. Il y a plusieurs conventions entre assureurs qui définissent quel assureur doit payer et quel montant faut-il régler, notamment nous y trouverons la convention IRSA [1] pour les sinistres en automobile matériel et la convention IRCA [2] pour les sinistres en automobile corporel. L'objectif principal de ces conventions est de faciliter l'indemnisation des assurés. Il y a également le droit pénal qui peut intervenir dans le cadre des recours. Tout dépend des situations.

Le quatrième et dernier point à mentionner est la possibilité d'avoir des **réouvertures** de sinistres au préalable clôturés. Ceci peut par exemple arriver si un nouvel élément ou une nouvelle conséquence du sinistre vient se rajouter au dossier, comme lors d'une procédure judiciaire pour statuer sur la responsabilité d'un sinistre ou encore des aggravations médicales pour des sinistres corporels. Ceci entraînera alors de nouveaux règlements et une ré-estimation de la charge du sinistre. Ces réouvertures restent toutefois marginales. Néanmoins, ceci n'empêche pas la nécessité de tenir compte de cet élément lors de

1. Indemnisation et de Recours entre Sociétés d'Assurance.
2. Indemnisation et de Recours Corporel Automobile.

l'estimation des réserves.

Enfin, il est évident que le provisionnement représente un enjeu important au sein d'une compagnie d'assurance où il est nécessaire de considérer toutes les inconnues citées dans les paragraphes ci-dessus.

Ainsi, l'estimation des provisions permet d'effectuer un bilan de la sinistralité à une date bien précise ce qui permet in fine d'estimer le reste à charge de la compagnie.

Avant d'introduire les différentes notions qui vont suivre, il est utile de bien différencier entre une année de rattachement et une année comptable.

- **Année de rattachement :** réfère à l'année à laquelle est relié un sinistre. Il est possible que l'année de rattachement soit l'année de survenance du sinistre, l'année de sa déclaration, l'année de souscription du contrat qui couvre ce sinistre ou encore l'année de début de travaux dans certains cas. Souvent en France, il s'agit de l'année de survenance du sinistre hormis quelques cas spécifiques. En assurance construction (l'assurance décennale par exemple), l'année de rattachement est l'année de souscription du contrat. La notion d'année de rattachement est essentielle car elle permet de construire des indicateurs dont l'objectif est de mettre en regard le coût de la sinistralité avec les primes acquises tel que le « *Loss Ratio* » (il s'agit d'un indicateur qui met en rapport le coût des sinistres avérés et évalués ainsi que les provisions divisés par le montant total des primes acquises).

- **Année comptable :** comme précisé précédemment, le paiement d'un sinistre peut s'échelonner sur plusieurs années suite aux différentes réévaluations qui peuvent avoir lieu au cours du temps. Ainsi, une année comptable correspond à l'année à laquelle a été effectuée une transaction comme un paiement ou un encaissement ou encore les variations de provisions.

Les triangles de liquidation

Également appelés triangles en « *Run off* », les triangles de liquidation illustrés sur la figure 1.6 permettent de visualiser la dynamique des sinistres étudiés. En effet, ils permettent de synthétiser l'évolution de cadences de la sinistralité entre plusieurs années de survenance par exemple. Cette cadence peut porter sur les règlements effectués, la charge dossier/dossier, etc.

Les méthodes de provisionnement agrégées ont principalement recours à des données présentées sous forme de triangles. L'idée étant d'avoir des données de sinistralité agrégées avec deux axes d'analyse : l'année de rattachement (très souvent l'année de survenance) et la période de développement (semestre / année ou autre). La finalité est d'avoir une vision de la sinistralité passée pour pouvoir estimer la sinistralité future par exercice de survenance (dans la majorité des cas).

Il est possible de présenter ces triangles de plusieurs façons et dans différents sens. Nous optons pour la représentation indiquée dans la figure 1.6 :

- en colonnes, nous avons les années de survenance ;
- en lignes, nous avons les années de développement ;
- en diagonales, nous avons par exemple la charge totale des différentes années de survenance présentes dans le triangle pour un exercice donné.

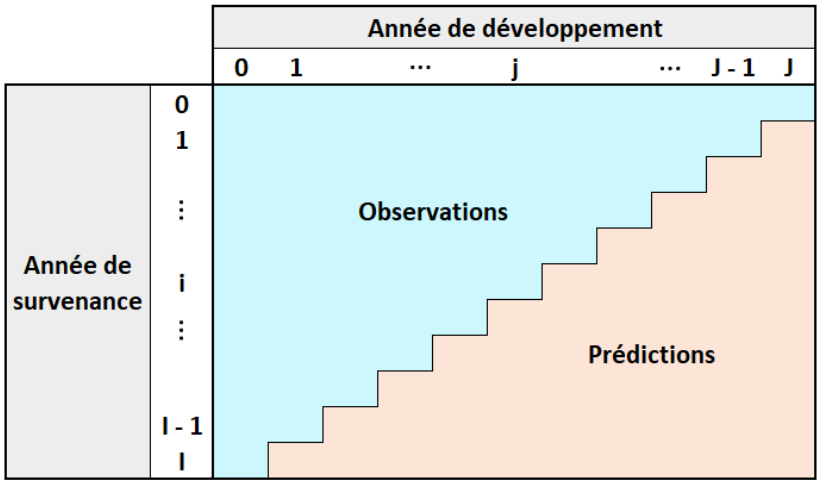


FIGURE 1.6 – Exemple de présentation d'un triangle de liquidation

Si la variable étudiée dans le triangle est directement la charge, le triangle supérieur contiendra la charge cumulée agrégée par date de survenance des sinistres et le triangle inférieur contiendra les estimations des charges cumulées pour les années de développement futures. A titre de précision, intuitivement, nous pourrions imaginer que la charge agrégée cumulée soit toujours croissante, cependant il est possible mais peu fréquent que ça ne soit pas le cas si par exemple l'assureur a été trop prudent dans son provisionnement et qu'il a revu à la baisse ses provisions dossier/dossier.

Que les méthodes de projection soient déterministes ou stochastiques, celles-ci supposent une certaine homogénéité du risque et du développement des sinistres. Nous ne pouvons donc pas construire des triangles à partir de données complètement différentes, par exemple mélanger des branches à développement différent. Ceci impliquerait la combinaison de risques complètement hétérogènes et biaiserait l'information. Si, au contraire, nous nous plaçons sur une granularité très fine, les risques seraient d'avoir des données volatiles et/ou que la loi des grands nombres ne soit plus applicable.

Enfin, il est important de souligner que malgré la liberté de l'actuaire à choisir le périmètre à étudier, il reste tout de même contraint par la data dont il dispose. Ainsi, il est possible de se retrouver avec des groupes de risques hétérogènes sans possibilité d'effectuer tous les retraitements nécessaires. Dans la partie application de ce mémoire, nous vérifierons la véracité ou pas des hypothèses émises par les différentes méthodes de provisionnement agrégées.

La charge ultime

Pour une branche donnée, la **charge ultime** (*ultimate*) correspond au coût final prévisible des sinistres pour un exercice donné. Dans le cadre des sinistres clos, il s'agit de la somme des règlements effectués par l'assureur. Tel qu'illustré sur la figure 1.7, elle peut être décomposée en quatre éléments dont la description détaillée figure en section 1.2.3 :

a. les paiements déjà effectués par l'assureur (*payments*)

b. les provisions dossier/dossier (*File/File reserves*)

c. les provisions pour les sinistres survenus, déclarés mais sous-évalués (*IBNER*)

d. les provisions pour les sinistres tardifs (*IBNYR*).

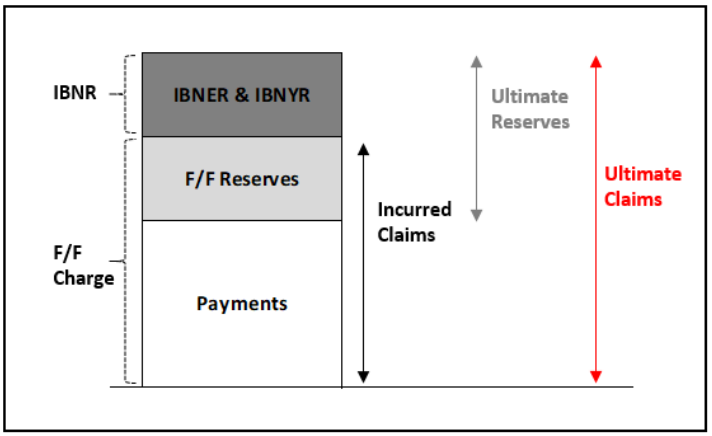


FIGURE 1.7 – Décomposition de la charge ultime (estimée)

1.2.3 Provisions techniques

Comme évoqué précédemment, l'activité d'assurance se caractérise par l'inversion du cycle de production. Il en résulte un traitement comptable particulier comparé aux pratiques de la comptabilité générale. En effet, un assureur ne doit pas uniquement se contenter de chiffrer les dettes qu'il a pour sure mais également ses engagements probables qu'il devra honorer vis-à-vis de ses assurés et qui se matérialisent sous forme de **provisions techniques**.

En effet, les provisions techniques ont pour rôle de refléter la capacité de la compagnie à assurer le règlement de différents sinistres notamment ceux dont les conséquences sont différées dans le temps et donc qui impliquent des règlements (ou encaissements de recours) durant des exercices comptables futurs, différents de l'exercice où ont été encaissées les primes des contrats couvrant ces sinistres. Il faut également mentionner que généralement la sinistralité d'un contrat est affectée à l'exercice comptable d'encaissement de la prime de ce contrat, ce qui permet à l'assureur de déduire son résultat annuel. A noter que ces provisions constituent une partie importante du passif du bilan.

Au final, il est intéressant de préciser que les méthodologies de calcul des provisions techniques sont différentes selon le type de contrats sous-jacents, c'est-à-dire s'il s'agit de contrats en assurance vie ou plutôt en non-vie. De par le contexte de ce mémoire, nous ferons un focus particulier sur les provisions techniques en non-vie. L'article R331-6 du Code des Assurances énonce les différentes provisions techniques à constituer. Nous en définirons quelques unes plus en détail ci-dessous.

Charge dossier/dossier

Les paiements et réserves dossier/dossier effectués au niveau des sinistres ouverts au préalable sont regroupés pour former la charge dossier/dossier. A l'ouverture d'un sinistre, il est provisionné par un forfait calculé par l'assureur selon différents critères (garantie sinistrée, responsabilité de l'assuré, coûts moyens précédents, etc.). Durant la vie du sinistre, sa provision est réévaluée par le gestionnaire. La fréquence de réévaluation des dossiers, à la hausse ou à la baisse, dépend du dossier en question et du processus instauré par l'assureur. Au fur et à mesure que sont effectués les règlements, la provision dossier/dossier du sinistre se voit diminuée de ces montants. Une fois le sinistre clos, sa provision est entièrement soldée et devient nulle.

Provisions *IBNER (Incurred But Not Enough Reserved)*

Ces provisions sont destinées à combler le manque potentiel de réserves dossier/dossier allouées au provisionnement des sinistres survenus et déclarés à la date de clôture des états financiers. Les provisions *IBNER* peuvent également être négatives si les réserves dossier/dossier sont trop précautionneuses. Elles sont calculées à un niveau agrégé et sont impactées par différents changements tels que l'évolution du processus de gestion des sinistres, les variations des taux techniques des rentes, etc.

Provisions *IBNYR (Incurred But Not Yet Reported)*

L'estimation de ces provisions consiste à anticiper au mieux le coût final des sinistres survenus mais non encore déclarés à l'assureur lors de la clôture des états financiers. Prenons l'exemple d'un sinistre survenu le 31 décembre n'ayant pas pu être déclaré avant le 1er janvier. La compagnie d'assurance se doit tout de même de provisionner ce sinistre l'année de sa survenance et non de sa déclaration. Le calcul des *IBNYR* peut s'appuyer sur la méthode fréquence/sévérité.

Provision pour Sinistres A Payer (PSAP)

La PSAP représente la part la plus importante de l'ensemble des provisions techniques, soit 85% en moyenne. L'article R331-6 définit la PSAP comme « la valeur estimative des dépenses en principal et en frais, tant internes qu'externes, nécessaires au règlement de tous les sinistres survenus et non payés, y compris les capitaux constitutifs des rentes non encore mises à la charge de l'entreprise ».

Cette provision concerne deux catégories de sinistres :

- les **ouverts (*RBNS*[3])**, sinistres ouverts dont le règlement n'a pas été complètement finalisé à la date de clôture des états financiers ;
- les **tardifs (*IBNYR*)**, sinistres survenus mais non déclarés à la date de clôture des états financiers.

Ainsi, la PSAP est la somme des trois provisions suivantes : les réserves dossier/dossier et les provisions des *IBNER* qui concernent les *RBNS* et enfin les provisions des *IBNYR*.

La PSAP doit être calculée exercice par exercice et figure au passif du bilan. Elle doit être évaluée **brute de réassurance**. En contre-partie, la provision pour sinistres réassurés apparaît à l'actif du bilan.

La PSAP doit être définie **brute de recours** et doit inclure les **charges externes individualisables** telles que les frais d'expertise, d'huissiers, etc.
Les recours, quant à eux, sont estimés séparément. Il s'agit d'estimer le montant qui reste à percevoir par l'assureur suite à ses différentes émissions de recours et suite aux multiples actions menées dans le cadre des recours auprès des compagnies adverses et/ou des tiers. Pour obtenir la **PSAP nette de recours**, il suffit d'effectuer une simple opération de soustraction.

Réglementairement, la PSAP doit être ventilée entre les catégories ministérielles 20 à 39 de l'assurance non-vie (cf. annexe A.2).

Provision pour Primes Non Acquises (PPNA)

La PPNA est définie par l'article R331-6 de la sorte « provision [...] destinée à constater, pour l'ensemble des contrats en cours, la part des primes émises et des primes restant à émettre se rapportant

3. *Reported But Not Settled.*

à la période comprise entre la date de l'inventaire et la date de la prochaine échéance de prime ou, à défaut, du terme du contrat ».
Ainsi, la PPNA représente la part de primes que l'assureur doit détenir pour faire face au risques à venir. La PPNA se calcule en soustrayant les primes acquises des primes émises. Il s'agit d'un poste important si la compagnie a émis plusieurs contrats pluriannuels. A l'opposé, le poste est négligeable si par exemple l'assureur a émis des contrats annuels qui se renouvellent le 1er janvier (avec une date d'inventaire au 31 décembre). Ce concept correspond au concept comptable de produits constatés d'avance.

Provision pour Risques En Cours (PREC)

La PREC est à son tour définie par l'article R331-6 « provision [...] destinée à couvrir, pour l'ensemble des contrats en cours, la charge des sinistres et des frais afférents aux contrats, pour la période s'écoulant entre la date de l'inventaire et la date de la première échéance de prime pouvant donner lieu à révision de la prime par l'assureur ou, à défaut, entre la date de l'inventaire et le terme du contrat, pour la part de ce coût qui n'est pas couverte par la provision pour primes non acquises ».
Cette provision a été établie pour compléter la PPNA si cette dernière s'avère insuffisante, autrement dit pour compenser le déficit en cas de sous-tarification.

D'autres provisions techniques non-vie sont exigées et définies dans l'article R331-6 mais ne seront pas détaillées dans ce mémoire.

1.2.4 Cadre réglementaire & comptable

Avant d'aborder tous les aspects réglementaires et comptables auxquels sont soumis les organismes d'assurance, il est utile d'avoir une vision globale de comment s'articulent un compte de résultat et un bilan au niveau d'une compagnie d'assurance notamment en assurance non-vie.
En effet, il est notable que la comptabilité s'est naturellement adaptée à ce secteur d'activité pour traduire ce critère primordial d'inversion du cycle de production. D'une part, le compte de résultat est présenté est en fonction du cycle d'exploitation : acquisition, administration et règlement du sinistre. Et d'autre part, le bilan affiche bien le niveau des engagements à l'égard des assurés qui sont donc des créanciers privilégiés. Ces deux derniers seront présentées dans ce qui suit sous la norme française.

Compte de résultat

Le compte de résultat expose les flux de l'exercice et contient, comme illustré sur la figure 1.8 :

- un compte de résultat technique : composé d'une partie « vie » et/ou d'une partie « non-vie » ;
- un compte de résultat non technique.

Le compte technique a pour rôle d'afficher le résultat en lien avec l'activité même de l'assurance. Le non technique, quant à lui, permet de présenter les éléments non liés à l'activité d'assurance.

I. Compte de résultat technique non-vie	
1. Primes acquises	
1a Primes	+
1b Variation des provisions pour primes non acquises	+/-
2. Produits des placements alloués du compte non technique	+
3. Autres produits techniques	+
4. Charges sinistres	
4a Prestations et frais payés	-
4b Charges des provisions pour sinistres à payer	+/-
5. Charges des autres provisions techniques	+/-
6. Participation aux résultats	-
7. Frais d'acquisition et d'administration	
7a Frais d'acquisition	-
7b Frais d'administration	-
7c Commissions reçues des assureurs et des garants en substitution	
8. Autres charges techniques	-
9. Variation de la provision pour égalisation	+/-
I. Résultat technique de l'assurance non-vie	

III. Compte de résultat non-technique	
1. Résultat technique de l'assurance non-vie	
2. Résultat technique de l'assurance vie	
3. Produits des placements	+
4. Produits des placements alloués du compte technique vie	+
5. Charges des placements	-
6. Produits des placements transférés au compte technique non-vie	-
7. Autres produits non techniques	+
8. Autres charges non techniques	-
9. Résultat exceptionnel	+/-
10. Participation des salariés	-
11. Impôt sur les bénéfices	-
12. Résultat de l'exercice	

FIGURE 1.8 – Compte de résultat technique non-vie et compte de résultat non technique

Comme indiqué sur la figure ci-dessus, les **variations de provisions** présentes dans le compte de résultat technique non-vie sont encadrées. Une variation de provision est calculée en soustrayant la provision de clôture à celle d'ouverture. La provision d'ouverture réfère à l'estimation de la provision lors de l'exercice comptable précédent et la provision de clôture correspond à l'estimation obtenue lors de la clôture des états financiers.

Au niveau des « charges des provisions pour sinistres à payer » se trouve également les variations de PSAP (tient compte des RBNS et des IBNYR).

Au niveau des « charges des autres provisions techniques » figurent plusieurs variations de provisions notamment la variation de la PM [4] des rentes et la PRC [5].
La **provision mathématique des rentes** est égale à la valeur actuelle des engagements de l'entreprise en ce qui concerne les rentes et accessoires de rentes mis à sa charge.
La **provision pour risques croissants** est une provision pour les opérations d'assurance contre les risques de maladie et d'invalidité et est égale à la différence des valeurs actuelles des engagements pris par l'assureur et par les assurés.

La **provision pour égalisation** est destinée à faire face aux charges exceptionnelles afférentes aux opérations garantissant certains risques tels que : les catastrophes naturelles, le risque atomique, la RC pollution, les risques spatiaux, l'assurance-crédit, l'assurance groupe dommages corporels.

4. provision mathématique
5. provision pour risques croissants

En définitive, il faut noter que les estimations des différentes provisions à la clôture de l'exercice peuvent fortement impacter le résultat technique de la compagnie d'assurance. Certaines de ces provisions notamment la PSAP vont subir la conjoncture de la sinistralité, c'est-à-dire qu'elle sera mécaniquement à la hausse en cas de sinistralité plus grave, de sur-fréquence, etc. D'autre part, certaines de ces provisions ont pour rôle de lisser ce résultat dans le sens où elles seront à la hausse les années favorables (année à faible sinistralité par exemple) puis à la baisse les années défavorables. La provision pour égalisation est une des provisions soumis à cette logique.

Bilan

Le bilan d'une entreprise rend compte de sa situation patrimoniale et fournit de façon détaillée l'état de sa solvabilité à une date d'inventaire donnée.

Il est essentiel de commencer par noter que le bilan d'une compagnie d'assurance diffère de celui d'une compagnie classique. Cette différence émane du critère d'inversion du cycle de production propre au secteur assurantiel. Ainsi, le bilan d'une compagnie classique se lit de gauche à droite contrairement à celui d'une compagnie d'assurance. Dans le premier cas d'une compagnie classique, la colonne gauche représente les actifs de cette entreprise, autrement dit sa richesse, et celle de droite les ressources déployées pour assurer le financement de ces actifs. Quant au second cas d'une compagnie d'assurance, la colonne de droite représente les engagements pris envers les assurés, des engagements couverts par les actifs présents dans la colonne de gauche.

En effet, comme mentionné précédemment, un assureur encaisse de l'argent (les primes) avant d'en décaisser et dispose de ce fait d'un *cash flow* positif combiné à une prise d'engagements envers les assurés suite à ces encaissements (les contrats souscrits). Il en découle donc la nécessité d'évaluer ces engagements futurs se matérialisant sous forme de provisions techniques au passif du bilan. Quant au *cash flow* généré, il amène les assureurs à disposer de fonds à investir, qui le sont notamment sous forme de placements financiers (obligations, actions ...) et de biens immobiliers et qui apparaissent à l'actif du bilan. Ainsi, le bilan d'une compagnie d'assurance peut être présenté de façon très simpliste en figure 1.9.

<table>
<tr><th>Actif</th><th>Passif</th></tr>
<tr><td rowspan="2">Placements</td><td>Fonds propres</td></tr>
<tr><td>Provisions techniques</td></tr>
</table>

Figure 1.9 – Bilan simplifié d'une compagnie d'assurance

Il ressort du bilan simplifié ci-dessus que la différence entre les deux composites « placements » et « provisions techniques » est égale aux fonds propres inscrits au passif du bilan. Ces derniers sont constitués à partir des capitaux initiaux investis par les actionnaires et des réserves accumulées (les réserves accumulées se composent des résultats des exercices précédents non distribués aux actionnaires et du résultat courant).

Ces capitaux propres permettent de refléter la situation nette de l'entreprise :

- une situation nette positive : traduction de supériorité des biens et créances de l'entreprise à ses engagements et dettes ;
- une situation nette négative : traduction d'insolvabilité de la compagnie. Il s'avère, dans ce cas de figure, nécessaire de faire appel aux actionnaires pour un apport de liquidités et/ou à des fonds de garantis dans le but d'assurer ses engagements et éviter la faillite.

Ainsi, les fonds propres ont un rôle majeur au sein de la compagnie car garantissent sa solvabilité future et lui permettent de se prémunir des aléas de l'année à venir.

En définitive, il est important de souligner que la valorisation des différents éléments présents au niveau du bilan d'une compagnie d'assurance varie en fonction du référentiel réglementaire. Dans la partie qui va suivre, nous allons nous intéresser aux particularités du régime prudentiel «Solvabilité II», notamment à l'évaluation des provisions techniques.

Solvabilité II

Cette partie sera dédiée à présenter de façon synthétique la directive « Solvabilité II ». Des approches plus détaillées figurent dans certains mémoires, en l'occurrence (Barruel et Bougnon, 2016) et (Wu, 2016).

La directive « Solvabilité II » a pris effet le 1er janvier 2016 au sein des différents pays de l'Union Européenne avec l'objectif de pallier aux critiques et limitations de la directive anciennement en vigueur « Solvabilité I ». En effet, « Solvabilité I » était considérée comme une directive trop simpliste qui, d'une part ne prenait pas en compte l'effet de diversification des risques pris par les assureurs et d'autre part n'assurait pas une harmonisation européenne suffisante, ce qui constituait en soit un frein au développement d'un marché européen unique dans le secteur de l'assurance. La directive « Solvabilité II » a donc pour but d'harmoniser le marché de l'assurance en Europe tout en s'assurant de la solidité des assureurs d'un point de vue solvabilité. Elle a ainsi imposé une nouvelle façon de valorisation du bilan et a également imposé l'instauration d'un système global de gestion des risques.

Tel que présenté sur la figure 1.10, la directive repose sur trois piliers. Le premier pilier regroupe les exigences quantitatives. Le deuxième pilier porte sur les exigences qualitatives, notamment en termes de maîtrise des risques (financiers, techniques et opérationnels) et de gourvenance. Le troisième et dernier pilier traite l'aspect communication et diffusion d'informations auprès du public, principalement les actionnaires et analystes, et également auprès des autorités de contrôle dans un souci de plus de transparence et homogénéisation d'informations et publications au niveau européen. Une synthèses des spécificités de chacun de ces trois piliers est illustré en figure 1.10.

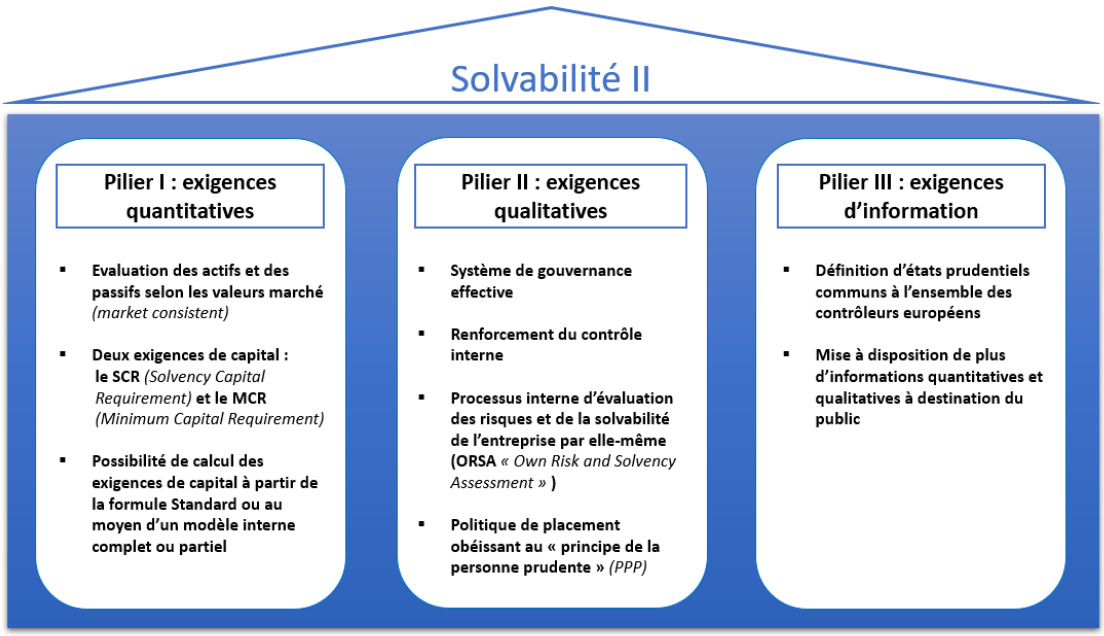


Figure 1.10 – Les trois piliers de la directive « Solvabilité II »

Zoom sur le Pilier I : les exigences en fonds propres

Comme mentionné ci-dessus, le premier pilier présente deux exigences de capital :

- **Le SCR *(Solvency Capital Requirement)* :** il s'agit du niveau de capital requis pour assurer la continuité de l'activité et limiter la probabilité de ruine de la compagnie à moins de 0,5% par an, c'est-à-dire le capital nécessaire pour pouvoir faire face à un choc extrême bicentenaire. Le SCR peut également être défini comme étant la VaR[6] à 99,5% des fonds propres.

  Les organismes soumis à la directive « Solvabilité II » ont le choix de calculer le SCR en se basant soit sur la formule Standard, soit sur un modèle interne qui doit au préalable être approuvé par l'autorité de contrôle. Ce modèle peut être complet ou partiel (combinaison entre modèle interne et formule Standard selon les branches et/ou les risques).

  Le processus de calcul du SCR « Formule Standard » repose sur une approche modulaire, illustrée sur la figure 1.11 où le SCR final résulte de l'agrégation de plusieurs modules de SCR évalués séparément. Ces modules, comme présenté dans le graphique ci-dessous, sont à leur tour composés de sous-modules et sont agrégés en se basant sur des matrices de corrélation renseignées par la directive dans le but de tenir compte de l'effet de diversification. Il faut noter que chacun des modules de SCR (marché, santé, contrepartie, actifs incorporels vie et non-vie) a pour objectif de tenir compte et mesurer les risques encourus par la compagnie selon ses propres activités.

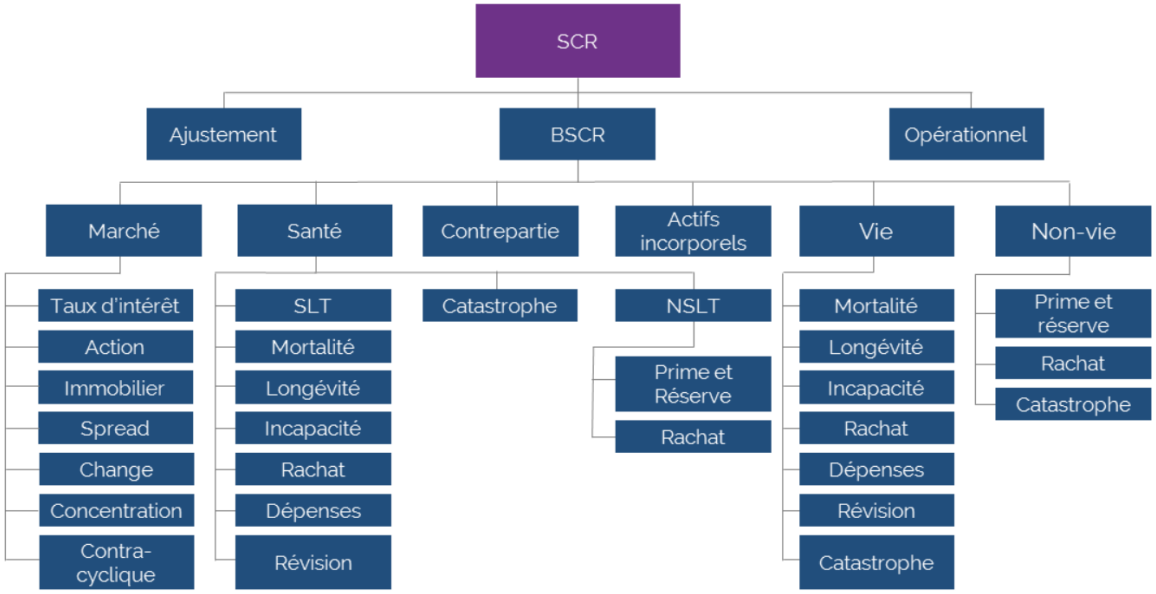


FIGURE 1.11 – Modules et sous-modules de risques pour le calcul du SCR en « Formule Standard »

6. *VaR : Value at Risk*

- **Le MCR *(Minimum Capital Requirement)* :** correspond au niveau minimum en dessous duquel l'entreprise opérerait dans des conditions très risquées où les intérêts des assurés seraient menacés si cette dernière était autorisée à poursuivre son activité. Il faut ainsi mentionner qu'en cas de transgression de ce seuil, ceci amène automatiquement les autorités de contrôle à exiger un plan de redressement dans des délais très courts ou encore un retrait d'agrément.

Zoom sur le Pilier I : le bilan prudentiel sous « Solvabilité II »

Toujours dans un souci d'une meilleure visualisation et prise en compte des risques auxquels sont exposés chacune des entreprises, la directive « Solvabilité II » a proposé une vision plus économique du bilan comparé à l'ancienne directive « Solvabilité I » qui avait plutôt une approche comptable.

La figure 1.12 synthétise le passage d'un bilan comptable vers un bilan prudentiel.

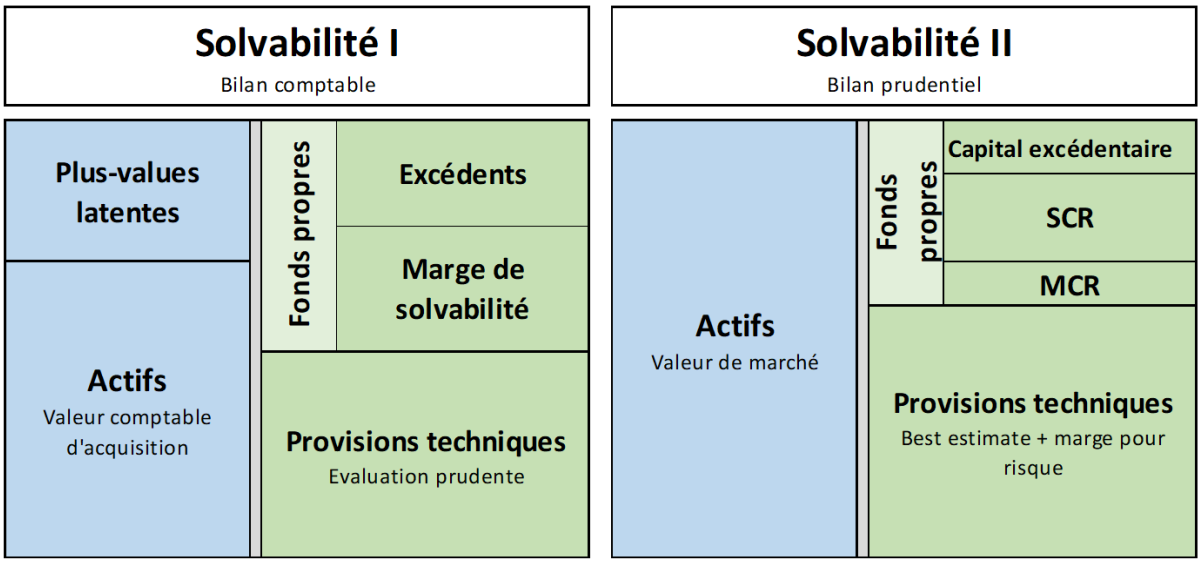


FIGURE 1.12 – Du bilan comptable sous « Solvabilité I » vers un bilan prudentiel sous « Solvabilité II »

Ainsi, de façon brève, les spécificités propres au bilan prudentiel peuvent être énumérées de la manière suivante :

- **La valorisation d'actifs :** s'il s'agit d'instruments financiers dont les cotations sont disponibles sur le marché, ces derniers sont comptabilisés en valeurs de marché (le *mark-to-market*). En cas d'absence de cotations, il est possible d'adopter d'autres méthodes, la méthode du «*mark-to-model*» par exemple. Cette dernière est principalement utilisée pour valoriser des produits de gré à gré qui suite à leurs spécificités ne permettent pas d'obtenir des valorisations *mark-to-market* et repose sur la mise en oeuvre de modèles mathématiques alimentés par plusieurs données disponibles.

- **L'estimation des provisions techniques** est égale à la somme des deux éléments suivants :

  □ ***Best Estimate (BE)* :** est calculé à partir de l'actualisation des *cash-flows* probables en lien avec les engagements propres à chaque compagnie (primes, frais généraux, prestations ...). Cette valeur est actualisée en se référant aux courbes de taux sans risque jugées les plus

pertinentes pour tenir compte de la valeur de l'argent dans le temps ainsi qu'en considérant un certain nombre d'hypothèses. Cette mesure est calculée brute de réassurance et concerne un périmètre de contrats existants lors du calcul.

Il faut noter que l'évaluation du « ***Best Estimate* non-vie** » sous Solvabilité II nécessite le calcul du « ***Best Estimate* de primes** » (également appelé *BE* des provisions pour primes) ainsi que le « ***Best Estimate* de réserves** » (également appelé *BE* des provisions pour sinistres).
Ces deux provisions sont étroitement liées étant donné que toutes deux sont destinées à couvrir les flux de trésorerie attendus, entrants et sortants, pendant la durée de vie des obligations d'assurance et de réassurance. La différence majeure qui réside entre ces deux notions est que la provision pour sinistres couvre les sinistres survenus avant la date d'évaluation (déclarés ou pas), tandis que la provision pour primes porte sur l'exposition future (sinistres qui surviendront après la date d'évaluation relatifs à des contrats existants à cette date).

□ ***Risk Margin (RM)* :** est calculée de telle sorte que la valeur des provisions techniques soit équivalente au montant que les entreprises d'assurance et de réassurance demanderaient pour reprendre et honorer les engagements d'assurance et de réassurance [7].

Il est également intéressant de mentionner que les entreprises d'assurance et de réassurance calculent la marge de risque en déterminant le coût que représente la mobilisation d'un montant de fonds propres éligibles égal au capital de solvabilité requis nécessaire pour faire face aux engagements d'assurance et de réassurance pendant toute la durée de ceux-ci [8].

Pour finir, il faut noter que globalement le marge pour risque est calculée en actualisant le coût annuel d'immobilisation du SCR. Ce dernier est estimé à 6% par an sur la durée de vie résiduelle des engagements utilisés pour le calcul du *BE*. Ainsi, la marge pour risque se calcule à partir de la formule 1.1.

$$RM = CoC \times \sum_{t \geq 0} \frac{\mathbb{E}[SCR_{RU}(t)]}{(1 + r_{t+1})^{t+1}}. \tag{1.1}$$

où :
— CoC réfère au taux de coût du capital, fixé à 6%
— SCR_{RU}(t) désigne le SCR de l'année t tel que calculé pour l'entreprise de référence
— r_t désigne le taux sans risque à maturité t (courbe fournie par l'EIOPA [9]).

- **Les fonds propres prudentiels :** diffèrent des fonds propres comptables et tiennent compte du profil de risque propre à l'organisme (comme expliqué ci-dessus lors de la présentation du calcul du SCR à partir la formule Standard). Les fonds propres sont classés en trois catégories (Tiers 1, Tiers 2 et Tiers 3) qui sont définis à partir de plusieurs critères en fonction de leur disponibilité, de leur degré de subordination et de leur durée ou permanence.

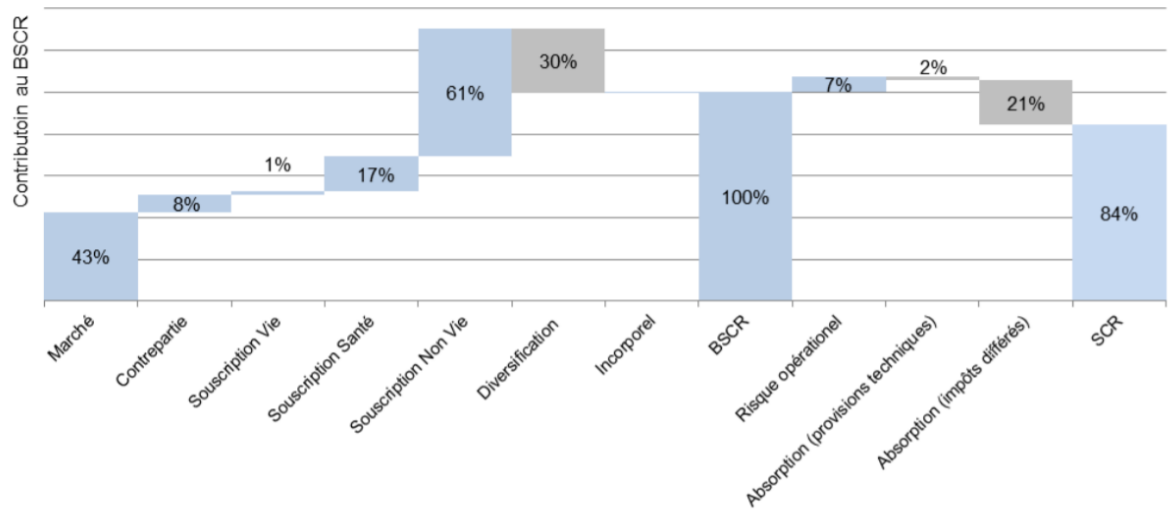
En guise de conclusion de cette sous-partie et suite à toutes les précisions de valorisations des éléments présents au niveau du bilan prudentiel, il est notable qu'une des ambitions de la réforme «Solvabilité II»

7. Définie par l'article 77-3 de la directive « Solvabilité II ».
8. Définie par l'article 77-4 de la directive « Solvabilité II ».
9. *EIOPA : European Insurance and Occupational Pensions Authority*

était le passage d'un mode de quantification où la prudence est implicite vers plutôt de l'explicite.

Le risque de réserve

Le SCR de souscription non-vie est un module central pour les organismes classés en IARD, un point mis en exergue par la décomposition du SCR présentée en figure 1.13 (61% du BSCR [10]). En effet, il représente l'exigence en capital nécessaire pour couvrir les risques sous-jacents aux activités d'assurance non-vie, notamment le risque de prime et de provision qui représente 77% du module, suivi par le risque de catastrophe qui représente environ 20% du module avant diversification [11].



Note de lecture : *Tous les éléments de ce graphique sont exprimés en pourcentage du BSCR. Le BSCR est égal à la somme des charges des différents risques diminué de l'effet de diversification. Puis le SCR final est égal au BSCR, augmenté du besoin en capital lié au risque opérationnel et diminué des ajustements liés à la capacité d'absorption des pertes par les impôts différés et par les provisions techniques (autrement dit la participation aux bénéfices).*

FIGURE 1.13 – Décomposition du SCR des organismes non-vie (ACPR, 2015) [12]

Dans le cadre particulier de ce mémoire, nous allons nous focaliser sur l'utilisation de techniques qui permettent de réduire l'incertitude sur le niveau de la charge ultime ce qui constitue une perspective pour affiner le niveau d'exigence en capital au titre du risque de réserve sous « Solvabilité II ».

Le risque de réserve a pour but de valoriser l'incertitude à un an du *Best Estimate* des provisions techniques, autrement dit du risque de fluctuation de cette estimation de provisions entre l'année d'inventaire et l'année qui suit. Ainsi, il permet de se prémunir contre deux types d'erreurs : l'erreur d'estimation et l'erreur du modèle. L'erreur d'estimation réside dans l'écart existant entre la variable estimée et l'estimateur retenu, puis l'erreur du modèle qui réside dans la variabilité réelle des montants de provisions.

10. *BSCR : basic SCR.*
11. Ces pourcentages proviennent de l'analyse effectuée par l'ACPR en 2014 qui concerne l'exercice de préparation au passage à « Solvabilité II ».
12. Organismes ayant participé à l'exercice de préparation au passage à « Solvabilité II » et représentant 89% de l'activité non-vie en France. Cet exercice a été effectué par l'ACPR en 2014.

1.3 Brève revue bibliographique

Méthodes agrégées déterministes

La méthode de provisionnement la plus fréquemment employée est celle de Chain-Ladder, principalement en raison de sa simplicité. Un certain nombre de méthodes, appelées « méthodes déterministes », en découlent. C'est notamment le cas de la méthode de London-Chain introduite par Benjamin et Eagles (1986), autorisant une ordonnée à l'origine non nulle, ou encore une autre méthode particulièrement appréciée de Bornhuetter et Ferguson (1972), introduisant une information exogène pour l'estimation de la charge ultime, dont il existe nombre de variantes. Davantage de méthodes déterministes découlent du modèle de Chain-Ladder dont celle de London-Pivot présentée par Straub (1988) ou encore celle de Munich Chain Ladder introduite par Quarg et Mack (2004).
Par ailleurs, le modèle de Taylor (1977), sur les pas du modèle proposé par Verbeek (1972), permet quant à lui la prise en compte des effets calendaires notamment l'inflation. Cette méthode est une méthode dite de séparation.

Méthodes agrégées stochastiques

Le provisionnement a pour finalité, outre l'estimation des provisions, de mesurer l'incertitude liée à cette estimation. Cette mesure, non prise en charge au travers des méthodes déterministes, est rendue possible au travers des méthodes dites « stochastiques ». Certaines de ces méthodes ont été élaborées pour aboutir à un résultat similaire à celui de Chain-Ladder. C'est le cas des modèles de Mack (1993a) et de Merz et Wüthrich (2008) qui permet d'estimer les erreurs de prédiction respectivement à l'ultime et à horizon un an (modèles développés au chapitre 3). De plus, la méthode *bootstrap*, d'abord introduite par Efron et Tibshirani (1993) et reprise par England et Verrall (1999), permet de simuler des triangles de données dont les caractéristiques sont semblables aux triangles de paiements incrémentaux et de charges afin d'aboutir à la distribution des pertes ultimes. Enfin, d'autres méthodes stochastiques existent dont notamment Murphy (1994) ou Hertig (1985).

Modèles individuels

Comme le mentionnent England et Verrall (2002) : "[...] il faut garder à l'esprit que les techniques traditionnelles ont été développées avant l'avènement des ordinateurs, en utilisant des méthodes qui pourraient être évaluées à l'aide d'un crayon et d'un papier. Avec l'augmentation continue de la puissance des ordinateurs, il faut se demander s'il ne serait pas préférable d'utiliser les observations individuelles plutôt que des données agrégées. Les bases de données d'observations individuelles étant régulièrement utilisées à des fins de tarification, il semble tout à fait faisable de les utiliser à des fins de provisionnement".

En effet, tous les modèles introduits précédemment sont basés sur des données agrégées se matérialisant sous la forme d'un triangle de liquidation, représentant ainsi leur force mais probablement aussi leur faiblesse. Cette faiblesse se caractérise principalement par une conséquente perte d'informations individuelles résultant de l'agrégation des données et qui auraient pu permettre d'améliorer la prédiction. En complément, nous pouvons ajouter à cette faiblesse la sensibilité aux valeurs extrêmes.

En parallèle de l'introduction du modèle de (Mack, 1993a), un nouveau cadre probabiliste approprié au provisionnement individuel a vu le jour au travers de plusieurs publications notamment (Arjas, 1989) et (Norberg, 1993). Ce dernier suggère l'usage de processus de Poisson marqués. Ces modèles permettent de modéliser la survenance des sinistres en se basant sur des processus de Poisson en temps

continu. Au sein de cette approche, les modélisations du nombre de sinistres et de la gravité sont dissociées. Ainsi, un vecteur de variables aléatoires est affecté à chaque sinistre survenu, modélisant les différentes étapes du cycle de vie de ce dernier. Il en découle notamment les modèles développés par Larsen (2007) et Haastrup et Arjas (1996).

Par la suite, un grand nombre de modèles ont vu le jour, notamment (Zhao et al., 2009), (Zhao et Zhou, 2010) et (Antonio et Plat, 2014). Ces derniers ont chacun modélisé le développement des sinistres individuels en temps continu. D'autre part, des modélisations en temps discret existent comme celle proposée par (Pigeon et al., 2013).

Enfin, face à l'engouement pour les modèles de *machine learning*, les publications actuarielles se sont tournées vers l'exploration des possibilités offertes par ces modèles d'apprentissage. En effet, l'usage de ces derniers est possible pour un large champs d'application notamment celui du provisionnement. Wüthrich (2016) propose ainsi de modéliser des provisions pour sinistres individuels à l'aide d'arbres de régression. Lopez et al. (2016), quant à eux, proposent une adaptation des arbres de CART aux données censurées. Ils retiennent deux cas d'usage, dont l'estimation de la charge ultime d'une branche à développement long où la base d'apprentissage comporte des sinistres en cours.

Il existe ainsi plusieurs travaux estimant l'apport des modèles de *machine learning* aux problématiques actuarielles, notamment Duval et Pigeon (2019) qui proposent un modèle de provisionnement basé sur l'usage de l'algorithme *XGBoost*.

Pour finir, il est également possible de faire appel aux méthodes de *machine learning* dans le cadre du provisionnement avec l'usage de triangles « *run-off* », ce qui est notamment le cas de (Kuo, 2019).

1.4 Problématique et objectifs du mémoire

Ce premier chapitre a en outre permis de mettre en évidence le fort enjeu que représente le provisionnement pour les assureurs. Cet enjeu s'avère d'autant plus important s'agissant de garanties à développement long pour des sinistres graves ou très graves, en l'occurrence ici la garantie RC corporelle automobile traitée dans le cadre de ce mémoire.

Étant donné cet enjeu, nous souhaitons challenger le modèle mis en place par AXA France pour l'estimation des PSAP de la garantie RC corporelle automobile. En effet, le modèle actuellement utilisé se base sur les méthodes de provisionnement agrégées qui ne permettent pas de mettre à profit l'importante quantité de données dont dispose l'entreprise. Ainsi, il s'agirait d'identifier un modèle de *machine learning* qui permette de pleinement tirer profit de la quantité et de la fine granularité des données possédées pour aboutir à des estimations de provisions plus précises.

Par ailleurs, ce mémoire s'inscrit dans la continuité du mémoire (Servel, 2020) dont l'objectif était principalement l'estimation de la charge ultime, auquel nous adjoindrons l'estimation des erreurs de prédiction.

A la différence de l'approche explicative employée dans le mémoire cité précédemment, car intégrant une variable observée à la clôture du sinistre, nous souhaitons opter pour une approche prédictive, basée uniquement sur des données disponibles de l'ouverture des sinistres jusqu'à la date du second inventaire. Dès lors, nous nous plaçons dans des conditions semblables à celles en vigueur lors de l'usage des méthodes agrégées.

Par conséquent, nous serons en mesure de soutenir la comparaison avec le modèle utilisé par AXA France, notamment en ce qui concerne les erreurs d'estimation de la charge à l'ultime.

Toutefois, en amont de l'utilisation des modèles de *machine learning* et afin de mieux appréhender le cadre de leur utilisation, nous nous attacherons dans les deux chapitres à venir à introduire dans un premier temps la garantie RC corporelle ainsi que les concepts clés qui lui sont associés avant de présenter le modèle AXA France et les méthodes agrégées dont il découle.

Enfin, il est important de préciser qu'au niveau des dossiers en hypothèse de rente, la charge D/D est provisionnée par les gestionnaires sinistres avec un taux d'actualisation à $3,5\%$ et la table TD 88-90 pour toutes les survenances et tous les inventaires. Lorsqu'un tel dossier est jugé ou transigé pour une indemnisation sous forme de rente, un paiement équivalent à cette provision est alors réglé et clôturé en gestion IARD et dans le triangle. Les dossiers indemnisés sous forme de rente sont gérés dans un service dédié. Ainsi, les données des triangles étudiés sont immunisées des impacts de taux et de table de mortalité. Par ailleurs, il existe une provision complémentaire au niveau des comptes AXA destinée à couvrir les impacts de taux et de table de mortalité. Cette provision complémentaire et son évolution ne font pas l'objet d'analyse dans le présent mémoire et n'impactent pas les résultats des méthodes étudiées.

2 Périmètre de l'étude

Après avoir présenté le contexte de ce mémoire et quelques notions phares, dont les provisions techniques et un aperçu de la directive Solvabilité II, il apparaît nécessaire de revenir sur les concepts spécifiques aux sinistres corporels et indispensables à la compréhension de cette étude. L'objectif de ce chapitre est donc d'introduire ces concepts en revenant à leurs fondements historiques avant de présenter la base de données individuelles des sinistres corporels utilisée dans le cadre de ce mémoire.

2.1 La responsabilité civile

2.1.1 Généralités

La responsabilité civile est définie comme l'obligation légale de réparer les dommages causés à autrui, qu'ils soient corporels ou matériels. Cette obligation légale est encadrée par les articles de loi du code civil :

- **Les articles 1382 et 1383** (Code Civil, 1804) précisent la responsabilité individuelle et unipersonnelle, volontaire ou involontaire, d'un individu au regard des préjudices qu'il fait subir à autrui.
  Art. 1382 du Code Civil - *"Tout fait quelconque de l'homme, qui cause à autrui un dommage, oblige celui par la faute duquel il est arrivé à le réparer."*
  Art. 1383 du Code Civil - *"Chacun est responsable du dommage qu'il a causé non seulement par son fait, mais encore par sa négligence ou par son imprudence."*

- **Les articles 1384, 1385 et 1386** étendent la notion de responsabilité à tout préjudice causé par une personne, un animal ou un objet dont le principal intéressé est responsable. La responsabilité civile d'un parent est donc engagée dans le cas d'un dommage causé par son enfant, de même qu'un employeur vis-à-vis de son employé, d'un propriétaire vis-à-vis de son animal de compagnie ou encore d'un propriétaire de bâtiment dont la ruine occasionne un dommage qui lui est imputable au titre d'un défaut d'entretien.

Dans certains cas, le montant du préjudice peut dépasser la capacité de paiement de l'individu. Une assurance Responsabilité Civile - parfois obligatoire - peut être souscrite afin de prendre en charge tout ou partie des dommages causés à un tiers.

Cette garantie est notamment inclue dans l'assurance MRH (Multirisque Habitation) pour l'ensemble des dommages causés par un élément du logement (fuite d'eau, incendie, chute d'arbre...) ainsi que dans l'assurance automobile pour l'ensemble des dommages causés par le véhicule (dégâts matériels, blessures, décès...). Aussi, l'article L211-1 du code des assurances (Code des Assurances, 2007) précise l'obligation d'une assurance minimale, communément appelée assurance "au tiers", pour les propriétaires de véhicule pour la couverture des dommages causés à autrui.

Les dommages couverts par une assurance peuvent être matériels ou corporels. Un dommage matériel représente une atteinte au patrimoine de la victime tandis qu'un dommage corporel représente une atteinte à son intégrité physique.

2.1.2 Spécificités de la responsabilité civile corporelle

Selon l'arrêté du 27 mars 2007 (Texte de loi, 2007), un accident corporel de la circulation routière implique au moins une victime, un véhicule et survient sur une voie ouverte à la circulation publique. On distingue les victimes (décédées ou ayant nécessité des soins médicaux) des indemnes (personnes impliquées mais non victimes). Les victimes sont réparties en 3 catégories :

- **Les personnes tuées à trente jours :** personnes décédées jusqu'à trente jours suite à l'accident ;
- **Les personnes blessées et hospitalisées :** personnes ayant nécessité une hospitalisation de plus de 24 heures suite à l'accident ;
- **Les personnes blessées légèrement :** personnes ayant nécessité des soins médicaux et n'étant pas inclus dans les deux catégories précédentes.

A titre de précision, les termes "dommage" et "préjudice" sont parfois utilisés sans distinction par abus de langage (Lingibé, 2019). En réalité, le dommage représente l'atteinte matérielle ou corporelle subie par la victime tandis que le préjudice représente la traduction juridique de cette atteinte. Pour cette raison, un dommage corporel peut induire des préjudices de différentes natures (souffrances physiques ou psychiques, perte de revenus...).

Les différents postes de préjudice seront introduits ultérieurement dans le cadre de ce mémoire, notamment selon la Nomenclature Dintilhac.

2.2 Histoire et concepts de l'indemnisation des sinistres corporels

2.2.1 Concepts clés

Indemnisation en capital/rente

L'indemnisation des victimes peut se faire sous forme de capital ou de rente.

Le capital représente le versement, en une fois, de l'indemnité à laquelle la victime a droit.

La rente implique le versement d'un certain montant (arrérage) à intervalle régulier (annuel ou mensuel) pendant une durée définie (rente temporaire) ou à vie (rente viagère).

Le montant de la rente est calculé sur la base d'une table de mortalité et des espérances de vie inhérentes.

Le choix du mode d'indemnisation impacte l'assureur a plusieurs égards. Ci-dessous sont détaillés les avantages pour l'assureur de chacun des modes :

- **Capital** :
  + Clôture rapide des dossiers
  + Gestion simplifiée (ex. pas de suivi de rente et pas de revalorisation)
- **Rente** :
  + Réduction du coût du sinistre dans le cas d'un décès prématuré de la victime
  + Pas de barème de capitalisation à négocier.

Notion d'Atteinte à l'Intégrité Physique et Psychique (AIPP)

L'Atteinte à l'Intégrité Physique et Psychique, aussi appelée Déficit Fonctionnel Permanent (DFP), est la "réduction définitive du potentiel physique, psychosensoriel ou intellectuel résultant d'une atteinte

à l'intégrité anatomophysiologique" (Index Assurance, 2020). Cette dernière a été définie par la Confédération européenne d'experts en évaluation et réparation du dommage corporel (CEREDOC).

Cette notion sert à évaluer le degré d'atteinte des victimes. On parle alors de taux d'AIPP, représentant l'évaluation donnée en pourcentage par le médecin expert à la consolidation de la victime suite à l'accident.

Éléments clés de la loi Badinter

- Généralisation du droit à l'indemnisation pour l'ensemble des victimes de dommages corporels, responsables ou non, causés par un accident de la route impliquant un véhicule terrestre avec moteur (VTM). Deux cas particuliers subsistent : le conducteur responsable de l'accident ou la victime qui provoque a elle seule l'accident par une faute inexcusable ;
- Généralisation du droit à l'indexation des rentes, prévu par la loi du 27 décembre 1974, à l'ensemble des rentes allouées, par décision judiciaire ou par transaction, et à l'ensemble des victimes ;
- L'assureur a un devoir d'information vis-à-vis de la victime concernant ses droits ;
- L'assureur est tenu de respecter des délais légaux après l'accident : au maximum six semaines pour la fourniture du questionnaire Badinter aux victimes et au maximum huit mois pour la proposition d'une offre d'indemnité ;
- La victime est en droit de demander réparation à l'assureur en cas d'aggravation du dommage subi ;
- Le bénéficiaire d'une rente peut, s'il le souhaite et que sa situation le justifie, demander une indemnisation par capital plutôt qu'en rente pour tout ou partie du reste à payer. La table de conversion pour le calcul du capital étant elle fixée par décret.

Éléments clés de la convention IRCA

Les prérequis à l'application de la convention IRCA sont :

- Un accident localisé en France, aux DOM ou à Monaco ;
- Au moins une victime blessée dont le taux d'AIPP ne dépasse pas 5% ;
- Deux véhicules impliqués, immatriculés en France et ayant des assureurs adhérents à la convention.

Une fois ces prérequis vérifiés, la convention IRCA fonctionne de sorte que :

- L'assureur de la victime non responsable indemnise son assuré et émet des recours envers les assureurs des tiers responsables ;
- Un recours est émis pour chaque victime dont le montant dépend du taux d'AIPP. En cas de taux d'AIPP nul, le montant du recours est forfaitaire mais dans le cas contraire, le recours est effectué au coût réel encadré ;
- Un recours est également émis pour chaque tiers responsable. En cas de responsabilité partagée, le montant du recours est proportionnel au taux de responsabilité de chacun des tiers ;
- La victime conserve le droit, si elle le souhaite, d'appliquer le droit commun plutôt que la convention IRCA.

Postes de préjudices - Nomenclature Dintilhac

Les recommandations pour l'usage de la nomenclature Dintilhac sont les suivants :

- Une indemnisation qui s'effectue pour chaque poste de préjudice en identifiant ceux devant faire l'objet d'un recours subrogatoire par les organismes tiers payeurs ;

- Un recours subrogatoire s'effectue poste par poste et ne peut excéder l'assiette du poste de préjudice ayant fait l'objet d'une indemnisation ;
- La date de consolidation est définie comme étant "le moment où les lésions se fixent et prennent un caractère permanent, tel qu'un traitement n'est plus nécessaire, si ce n'est pour éviter une aggravation, et qu'il est possible d'apprécier un certain degré d'incapacité permanente réalisant un préjudice définitif" ;
- Il est entendu que demeurent certains préjudices pour lesquels aucune date de consolidation ne saurait être identifiée de par leur caractère récurrent et permanent, tels que les victimes des contaminations à l'amiante.

La nomenclature Dintilhac, basée sur la division tripartite des postes de préjudice, peut se résumer au travers de la table 2.1.

	Préjudices des victimes directes		Préjudices des victimes indirectes	
	Préjudices temporaires	Préjudices permanents	En cas de survie de la victime directe	En cas de décès de la victime directe
Préjudices patrimoniaux	Dépenses de santé actuelles	Dépenses de santé futures	Perte de revenus des proches	Pertes de revenus des proches
	Pertes de gains professionnels actuels	Pertes de gains professionnels futurs	Frais divers des proches	Frais divers des proches
		Frais de logement adapté		Frais d'obsèques
		Frais de véhicule adapté		
	Frais divers	Assistance par tierce personne		
		Incidence professionnelle		
		Préjudice scolaire, universitaire ou de formation		
Préjudices extra-patrimoniaux	Déficit fonctionnel temporaire	Déficit fonctionnel permanent	Préjudice d'affection	Préjudice d'affection
	Préjudice esthétique temporaire	Préjudice esthétique permanent	Préjudices extra-patrimoniaux exceptionnels	Préjudice d'accompagnement
		Préjudice sexuel		
		Préjudice d'établissement		
		Préjudice d'agrément		
	Souffrances endurées	Préjudices permanents exceptionnels		
	Préjudices liés à des pathologies évolutives			

TABLE 2.1 – Présentation de la nomenclature Dintilhac détaillant les postes de préjudices

2.2.2 Chronologie et évolution de l'indemnisation des sinistres corporels

Afin de mieux appréhender les notions et le fonctionnement de l'indemnisation des sinistres corporels, et notamment en responsabilité civile, nous nous proposons de les remettre sur une échelle temporelle pour en comprendre les fondements et les évolutions successives.

Cette section a donc pour but de présenter les dates clés correspondant à la création de concepts, à leur première mise en application ou encore aux différentes réglementations y afférant.

1693 - Première table de mortalité par E. Halley

L'astronome anglais E. Halley (1656 – 1742) est à l'origine de la plus ancienne table de mortalité de l'histoire, datant de 1693 (DUPÂQUIER, 1976).

En partant des travaux de ses prédécesseurs qui ont initié un tableau de survie (Graunt-1662) et la notion d'espérance de vie (Huygens-1669 & Leibniz-1680), il analyse pour la première fois des données d'observation en récoltant les bulletins de décès des cinq dernières années de la ville de Breslau (actuellement Wroclaw) (Dupâquier, 1996).

La création de sa table de mortalité lui permet d'aboutir à deux conclusions majeures. Tout d'abord, il est en mesure d'obtenir la répartition par âge de la population et donc d'identifier la proportion d'hommes en âge de porter les armes. Ensuite, il obtient le taux de mortalité par âge lui permettant d'ajuster le tarif des assurances sur la vie et la valeur des rentes perçues en fonction de l'âge.

Jusqu'ici uniquement gérés par capital, c'est à ce moment que naît le versement en rentes tel que nous le connaissons aujourd'hui et la dualité des modes d'indemnisation en capital ou en rente.

Loi du 27 février 1958 - Obligation d'assurance responsabilité civile pour les détenteurs de véhicules

C'est au travers de cette loi que l'assurance de la responsabilité civile est devenue obligatoire pour "toute personne physique ou morale autre que l'Etat dont la responsabilité peut être engagée en raison des dommages corporels ou matériels causés à des tiers par un véhicule terrestre à moteur, ainsi que par ses remorques ou semi-remorques... ".

Un véhicule terrestre à moteur (VTM) est alors défini comme étant "tout véhicule automoteur destiné à circuler sur le sol et qui peut être actionné par une force mécanique sans être lié à une voie ferrée, ainsi que toute remorque, même non attelée."(Rose, s. d.).

Cette définition s'étend donc progressivement, au travers de la jurisprudence, aux engins de chantiers, engins agricoles, chariots élévateurs, tondeuses auto-portées, et même aux fauteuils roulants électriques et scooters électriques de plus de 6km/h.

C'est par ailleurs le 17 mars 2011 que la deuxième chambre civile de la Cour de Cassation a déclaré que les engins de déplacement personnel (EDP), tel que les trottinettes électriques roulant à plus de 6km/h, étaient également considérées comme véhicules terrestres à moteur et donc soumis aux mêmes lois.

Loi du 27 décembre 1974 - Obligation d'indexation des rentes

Il nécessaire dans un premier temps de savoir que l'indexation d'une rente consiste à en faire varier la valeur en fonction d'un indice dans le temps, appelé "coefficient de revalorisation".

Les premières notions d'indexation apparaissent à la suite de la première guerre mondiale (Malignac, 1978). Jusqu'alors, la stabilité monétaire n'avait pas nécessité de tels mécanismes, et les fondements juridiques avaient implicitement considéré cette stabilité comme constante.

A compter de 1918, la notion d'indexation apparaît, faisant l'objet de jugements au cas par cas, mais est continuellement contestée allant jusqu'à aboutir à l'ordonnance du 30 juin 1945, en pleine seconde guerre mondiale, qui suspend "nonobstant toutes stipulations contraires l'application des clauses

contractuelles qui prévoient la détermination d'un prix au moyen de formules à variation automatique.".

Malgré tout, les indexations réapparaissent progressivement dans les lois, notamment pour les loyers (Loi du 1er septembre 1948) et les baux ruraux ou commerciaux (décret du 30 septembre 1953).

L'indexation des rentes quant à elle, est systématiquement refusée par la Cour de Cassation jusqu'au 6 novembre 1974, date à laquelle cette dernière révise sa doctrine et octroie pour la première fois une rente indexée pour une victime d'accident grave.

Le gouvernement s'empare immédiatement du sujet à la suite de ce jugement pour cadrer et homogénéiser ces indexations de rentes, donnant naissance à la loi du 27 décembre 1974 (Texte de loi, 1974) qui rend obligatoire l'indexation des rentes versées soit aux victimes directes d'un accident automobile dont le taux d'atteinte à l'intégrité physique et psychique (AIPP) est supérieur à 75% soit aux personnes dont les victimes directes décédées avaient la charge. L'article de loi précise également que ces revalorisations ne peuvent excéder 8 fois le salaire moyen défini par la sécurité sociale et que ces dernières seront prises en charge par un fond de majoration des rentes qui est créé à cette occasion.

Loi Badinter du 5 juillet 1985 - Indemnisation des victimes d'accident de la route

Jusqu'en 1982, l'indemnisation des victimes d'accident de la route était sujette à controverse. En premier lieu, la responsabilité de l'accident était difficile à établir clairement, notamment en présence d'événements pouvant être qualifiés de "force majeure". De ce fait, la responsabilité du conducteur et la faute potentielle de la victime étaient évaluées au cas par cas par le tribunal, si bien que la jurisprudence a tâtonné, et ses décisions ont paru inégales, injustes et dépourvues de logique d'ensemble.

L'arrêt Desmares, rendu en 1982 par la Cour de Cassation française crée une polémique et pousse le ministre de la Justice, Robert Badinter, à déposer un projet de loi qui deviendra la loi dite "Badinter" du 5 juillet 1985 (Texte de loi, 1985).

Cette loi constitue une avancée majeure dans l'homogénéisation des indemnisations de victimes d'accident de la route tout en renforçant le droit de ces victimes à une indemnisation.

1er avril 2002 - Mise en application de la convention IRCA

Jusqu'en 1968 et malgré la pluralité d'assureurs impliqués dans chaque sinistre, la coordination entre ces derniers se fait sans réel cadre établi.

Au fil du temps, un certain nombre de conventions ont vu le jour, orientés dans un premier temps sur les sinistres matériels au travers de la convention d'indemnisation directe des assurés (IDA) en 1968 devenue ensuite la convention inter-sociétés de règlement des sinistres automobile (IRSA) en 1974. Ce n'est qu'en 1977 que la première convention appliquée aux sinistres corporels voit le jour sous le nom de convention d'indemnisation directe des accidents corporels (IDAC).

A la suite de la loi Badinter, il apparaît nécessaire d'identifier un correspondant unique pour représenter la pluralité des assureurs impliqués dans un accident ce qui donne lieu en 1986 à la convention d'Indemnisation pour Compte d'Autrui (ICA). Cette dernière est finalement remplacée en 2002 par la convention d'indemnisation et de recours corporel automobile (IRCA), toujours en vigueur à ce jour (Index Assurance, 2021).

Loi du 1er août 2003 - Prise en charge de l'indexation obligatoire des rentes par le FGAO

C'est en 1951 que sont lancés les Fonds de Garantie Automobile (FGA) dont le but est d'indemniser les victimes d'accident de la route dont les responsables sont non identifiés ou non assurés et non solvables.

La loi de sécurité financière du 1er août 2003 renomme ces fonds en Fonds de Garantie des Assurances Obligatoires de dommages (FGAO) et leur attribue une prérogative supplémentaire. Si leur mission principale reste identique, il leur incombe désormais également la prise en charge de l'indexation obligatoire des rentes.

Rapport du 28 octobre 2005 - Séparation des postes de préjudice - Dintilhac

Le ministre de la Justice, Dominique Perben, lors d'une réunion plénière le 19 décembre 2002 confie deux objectifs principaux au Conseil National de l'Aide aux Victimes (CNAV) présidé par Yvonne Lambert-Faivre dont celui de définir plus clairement les différents postes de préjudices nécessaires à une indemnisation plus claire et plus juste des victimes.

Le projet de nomenclature obtenu par le CNAV, suite à son analyse, et présenté dans son rapport de Juin 2003 (Lambert-Faivre, 2003) repose sur une triple distinction :

- La distinction entre les préjudices de la victime directe et les préjudices des victimes par ricochet ;
- La distinction entre les préjudices économiques patrimoniaux et les préjudices non-économiques personnels ;
- La distinction entre les préjudices temporaires et les préjudices permanents.

Par la suite et en s'inspirant des travaux du CNAV, un groupe de travail mené par le président de la deuxième chambre civile de la Cour de Cassation, Jean-Pierre Dintilhac, remet le 28 octobre 2005 au Garde des Sceaux un rapport proposant une nomenclature des préjudices corporels, dite nomenclature "Dintilhac" (Dintilhac, 2005). Le groupe de travail a assorti sa proposition de nomenclature d'une série de recommandations permettant d'en assurer une application concrète.

Loi du 21 décembre 2006 - Obligation de l'usage d'une méthodologie pour la détermination des préjudices

Suite à la constitution de la nomenclature Dintilhac, la loi du 21 décembre 2006 impacte l'indemnisation des sinistres corporels en imposant un recours poste par poste des tiers payeurs, comme recommandé par le groupe de travail présidé par M. Dintilhac, et un droit de préférence pour la victime.

Cela implique que les sommes dépensées par un tiers payeur ne peuvent être prélevées sur l'ensemble de l'indemnisation accordée par le tribunal mais uniquement sur l'assiette correspondante au poste de préjudice couvert par le tiers payeur, comme défini au sein de la nomenclature Dintilhac.

Loi du 29 décembre 2012 - Prise en charge de l'indexation obligatoire des rentes par les assureurs

Une fois la prise en charge des indexations obligatoires transférée au FGAO, ce dernier a vu ses finances se détériorer progressivement, de sorte que les années 2008 à 2012 ont toutes été déficitaires jusqu'à aboutir à près de 500 M€ de déficit au 31 décembre 2012 selon la commission des finances du Sénat.

Cette commission atteste en conclusion de ses analyses pour le projet de loi de finances rectificative pour 2012 que le FGAO n'est plus en mesure de financer la revalorisation des rentes des victimes d'accidents de la circulation (Conseil de normalisation des comptes publics, 2014).

Par conséquent, la loi de finances rectificative du 29 décembre 2012 limite la prise en charge des indexations de rentes du FGAO aux seuls accidents survenus avant le 1er janvier 2013 et transfère l'indexation obligatoire de l'ensemble des accidents amenés à survenir à partir de cette date aux assureurs.

2.3 Base de données individuelles

Après avoir défini le contexte au sein duquel évoluent les sinistres corporels, cette partie sera dédiée à la présentation ainsi qu'à l'analyse des données individuelles qui seront à posteriori exploitées en vue d'un gain en précision des estimations de charges ultimes.

2.3.1 Description de la base de données

Pour rappel, la base de données qui sera exploitée dans ce mémoire présente plusieurs similitudes avec celle du mémoire (Servel, 2020). Néanmoins, elle couvre un périmètre plus important incluant, outre le réseau des agents, celui des courtiers et des salariés et s'étend sur des survenances plus importantes allant de la survenance de 2010 à celle de 2014.

Cette base de données a été construite de telle manière que chaque ligne corresponde à un sinistre. Ainsi, pour une fréquence moyenne de 17723 sinistres par an au niveau de la garantie RC corporelle automobile, cette dernière contient 88616 sinistres.

Toujours dans un objectif de collecte maximale d'informations, cette base de données a été construite à partir de trois bases de données différentes.

La base des sinistres

Cette base de données comporte toutes les informations se rapportant à un sinistre. Un travail conséquent a été réalisé afin d'avoir au niveau d'une unique ligne le déroulement chronologique de ce dernier avec une vision arrêtée à la fin de chaque trimestre et également une vision à l'ouverture du sinistre.
Nous y trouverons les dates clés d'un sinistre : la date de survenance, la date d'ouverture, la date de réouverture et la date de clôture si ce dernier est clos. Ils y figurent également six montants fondamentaux définis ci-dessous.

- **Le montant de règlements en principal :** correspond aux montants réglés à l'assuré dans le cadre de son indemnisation. Il s'agit de règlements bruts de recours.
- **Le montant de règlements en frais :** correspond principalement aux honoraires d'expertise des sinistres.
- **Le montant de réserves D/D :** correspond à l'évaluation des règlements restants à effectuer à l'assuré. A l'ouverture du sinistre, ce dernier est provisionné par un montant forfaitaire calculé au préalable par l'équipe "Actuariat". Ce montant est amené à évoluer dans le temps en fonction des évaluations du gestionnaire ainsi que des règlements effectués à l'assuré.
- **Le montant de recours encaissés :** correspond au montant encaissé suite à l'émission de recours auprès de la compagnie du tiers dans le cadre de la convention IRCA ou celle du droit commun comme mentionné dans la partie précédente.

- **Le montant des estimations de recours à encaisser :** correspond au montant estimé restant à encaisser par AXA. Ce montant peut être à son tour forfaitaire ou au coût réel. Il est estimé par le gestionnaire ou l'expert et amené à évoluer tout au long de la gestion du recours.
- **La charge nette de recours :** correspond à la somme des encaissements de recours et des estimations de recours à encaisser, déduction faite des règlements effectués et restants à effectuer. Ceci permet de réellement valoriser le coût de revient du sinistre qui peut être négatif ou positif.

Outre les variables mentionnées ci-dessus, cette base dispose également d'informations potentiellement intéressantes pour notre problématique. Nous trouverons notamment :

- **La cause détaillée du sinistre :** précise s'il s'agit d'un accident entre deux véhicules, un accident de parking, un accident en chaîne ou d'un carambolage de trois véhicules ou plus.
- **Litige :** cette variable indique si le sinistre est en contentieux ou pas.
- **Le taux de responsabilité de l'assuré :** indique la part de responsabilité de l'assuré (0%, 50% ou 100%).
- **Le code département du lieu de survenance du sinistre.**

La base des contrats

Cette base de données met à disposition des informations en lien avec le souscripteur du contrat. Ces informations sont collectées à la souscription, également utilisées pour le calcul de la prime, et sont mises à jour suite à tout changement signalé par l'assuré ou identifié suite à un sinistre.

Nous pouvons répartir les variables retenues en deux catégories comme mentionnées ci-dessous.

- **Les variables en lien avec l'assuré :** son âge, son coefficient bonus-malus et le réseau de distribution du contrat.
- **Les variables en lien avec le véhicule de l'assuré :** type du véhicule (2 roues/4 roues) et son groupe SRA reflétant sa puissance.

  SRA (Sécurité et Réparation Automobile)
  Il s'agit d'une association créé en 1977 et ayant pour adhérents l'ensemble des assureurs automobiles en France. Cette dernière fournit un fichier recensant l'ensemble des informations sur les véhicules 2, 3 et 4 roues de moins de 3,5 T adressés au marché français, permettant ainsi aux assureurs de calculer les primes de leurs contrats.

La base des fiches victimes

Cette base de données, contenant initialement les informations propres à chaque victime, a été retraitée afin d'avoir un sinistre par ligne étant donné qu'au niveau d'un sinistre, il est possible d'avoir plusieurs victimes. Parmi les variables dont nous disposons, nous pouvons citer les variables ci-dessous.

- **Le nombre de victimes corporelles**
- **Le nombre de victimes graves**
- **Le nombre de victimes décédées**
- **Le nombre de cyclistes/piétons/passagers et conducteurs**
- **L'âge de chaque victime à la survenance du sinistre**
- **Le choix du mode d'indemnisation :** capital ou rente

- **Le niveau de préjudice subis par chaque victime.**

Un dictionnaire dédié à la description des variables est présent en annexe A.3.

Il était souhaitable d'inclure dans notre base de données la variable « taux d'AIPP » (Atteinte à l'Intégrité Physique et Psychique - notion développée en 2.2.1), une variable à fort potentiel discriminant, toutefois nous ne disposons pas de cette dernière au niveau des survenances étudiées.

En définitive, nous avons cherché à vérifier la cohérence entre la base de données individuelles et les triangles agrégés présentés et utilisés par la suite en section 3.5 (la vision à fin 2015 est disponible en annexe A.7). Pour ce faire, nous avons agrégé chaque variable charge D/D nette de recours vu à la fin de chaque année par année de survenance pour s'assurer que nous retombions bien sur les montants présents au niveau des triangles. Néanmoins, disposant uniquement au niveau des bases individuelles des survenances 2010 à 2014, nous étions limités en termes d'analyses notamment pour expérimenter de nouveaux regroupements plus homogènes au niveau des triangles agrégés dans le but de satisfaire au mieux les hypothèses émises par les méthodes agrégées. A titre d'exemple, nous aurions voulu utiliser des méthodes de classification non supervisée (comme la méthode *kmeans)*) afin d'obtenir potentiellement des triangles plus homogènes.

2.3.2 Retraitement des données

Il faut également noter que pour pouvoir exploiter les données de façon optimale au sein des modèles, il est souvent nécessaire de procéder aux retraitements cités ci-dessous.

Le traitement des valeurs manquantes. Globalement, quand il est estimé utile de garder la variable et que l'absence d'informations est jugée tolérable, nous avons remplacé les valeurs manquantes par la valeur médiane au niveau des variables numériques et par la modalité la plus fréquente au niveau des catégorielles.

Le traitement de variables catégorielles. Si la variable dispose d'une modalité à très faible fréquence sans que ceci présente un intérêt particulier par rapport à la variable que nous souhaitons prédire, cette modalité est remplacée par la modalité la plus fréquente.

Les deux retraitements décrits ci-dessus restent mineurs et n'impactent pas la variable charge ultime que nous souhaitons prédire.

La création de nouvelles variables potentiellement pertinentes. Plusieurs variables binaires et catégorielles ont été créées à partir des variables numériques existantes en vue de les tester au sein des différents modèles. Pour les variables binaires, nous pouvons prendre pour exemple la variable « Top_Deces » indiquant la présence ou non à minima d'un décès au sein du sinistre, « Top_Grave » indiquant la présence ou non à minima d'une victime gravement blessée (cf. figure 2.6), etc.

Pour les variables catégorielles, nous pouvons prendre pour exemple la variable « top_S1 » à partir de la charge vue à la deuxième année d'inventaire. Les bornes retenues pour définir les différentes modalités de cette dernière ont été choisies en créant un arbre binaire CART avec comme variable cible la charge vue à fin 2019 et comme unique variable explicative la variable charge vue à la deuxième année d'inventaire.

L'idée est d'obtenir des groupes à la fois homogènes et discriminants de la variable cible. Cette variable est représentée en fonction de la charge en figure 2.7.

2.3.3 Analyse de la variable d'intérêt et présentation de la notion de censure à droite

Une fois la constitution et les retraitements de la base de données clarifiés, nous allons nous attarder sur l'analyse de ces variables. Dans un premier temps, nous allons commencer par nous focaliser sur la variable d'intérêt représentée par la charge nette du sinistre (nette de recours) dont la distribution vue à fin 2019 pour les survenances de 2010 à 2013 est représentée en figure 2.1. En figure 2.2 est disponible une vision identique avec un *focus* uniquement sur les sinistres dont la charge est comprise entre -5000€ et 20000€. Une vision par année de survenance est disponible en annexe A.4.

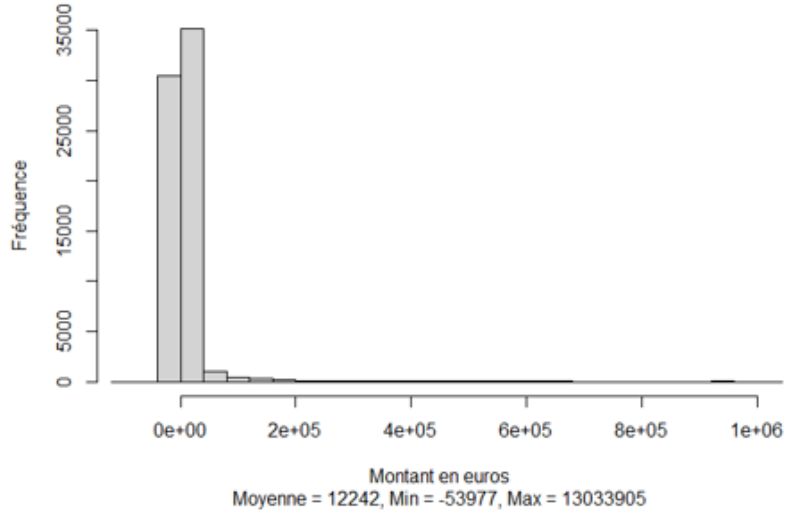


FIGURE 2.1 – Distribution de la charge D/D nette de recours (survenances 2010 à 2013) - vision à fin 12/2019

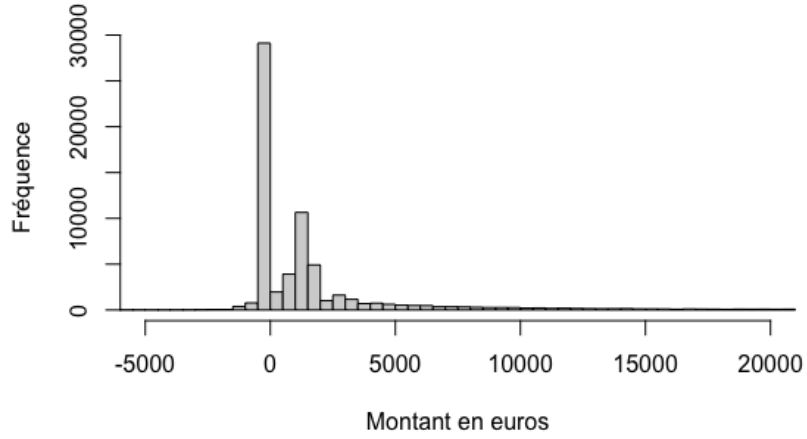


FIGURE 2.2 – Distribution de la charge D/D nette de recours (survenances 2010 à 2013) - vision à fin 12/2019 - *Zoom* sur la charge $\in [-5000€, 20000€]$

La visualisation de la distribution de cette variable, au niveau de la figure 2.1, affiche l'importante dispersion de la charge D/D nette de recours d'un sinistre atteignant, dans notre cas de figure, un minimum de -53 977€ et un maximum de 13 033 905€ avec une moyenne de 12 242€. Néanmoins, il faut tout de même noter que trois quart de ces sinistres ont une charge inférieure à 1 668€ avec un pic au niveau de la valeur nulle correspondant aux dossiers clos sans suite (28 233 sinistres).

Il est également possible d'identifier un second pic autour de 1 500€, ce qui correspond approximativement au montant forfaitaire fixé par la convention IRCA, convention présentée brièvement ainsi que ses modalités d'application en 2.2.1. Ce forfait peut varier d'une année à une autre mais conserve tout de même le même ordre de grandeur. A titre d'exemple, il avait une valeur de 1 490€ pour les survenances 2010 et 2011.

Quant à la présence de certaines valeurs négatives, ceci relève principalement de sinistres dont l'indemnisation (ou réserve D/D ou la somme des deux) effectuée envers l'assuré est inférieure aux encaissements perçus ou évalués par l'assureur dans le cadre d'un recours (1 958 sinistres).

Une fois imprégnés des particularités de la variable cible « charge ultime d'un sinistre nette de recours » que nous souhaitons prédire par la suite, il est important de noter que cette dernière fait l'objet d'une censure à droite. En effet, comme schématisé au niveau de la figure 2.3, de par la présence de sinistres toujours en cours à fin 2019, nous ne disposons pas de la version définitive de la charge ultime de ces sinistres. Cette variable est donc censurée suite au choix d'arrêter la vision de la base de données à fin 2019 et à l'impossibilité d'attendre la clôture de tous les sinistres comme il est possible de faire pour certaines branches. En effet, la branche « RC corporelle automobile » est une branche particulièrement longue dont l'atteinte d'un taux de clôture à 100% peut prendre une trentaine d'années pour une survenance donnée. Le formalisme mathématique lié à ce phénomène est présenté en sous-section 4.1.1.

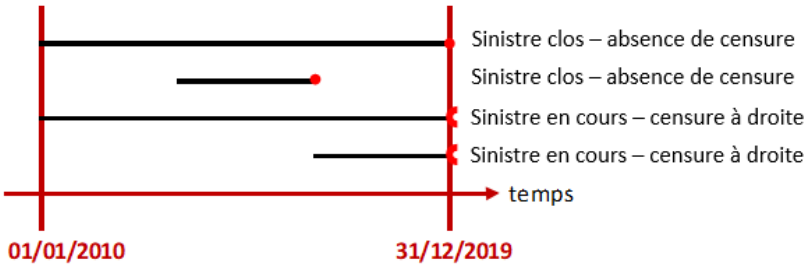


FIGURE 2.3 – Représentation de données censurées à droite

Il est certain que cette censure n'est pas sans impact sur la façon d'entraîner nos modèles de *machine learning* comme présenté à posteriori en sous-section 4.1.2. Il est ainsi intéressant de quantifier cette censure aussi bien en volume qu'en charge comme réalisé au niveau du tableau 2.2. Nous remarquons tout d'abord une décroissance de la censure par année de survenance, ce qui paraît logique. De plus, nous notons un taux de censure global de $4,5\%$ en volume, soit 3 064 sinistres toujours en cours et un taux de censure en charge de $42,8\%$, soit une charge D/D nette de recours de 545 M€ à fin 2019.

Année de survenance	Taux de censure à droite en volume	Taux de censure à droite en charge D/D nette de recours
2010	4,0%	40,3%
2011	4,2%	41,2%
2012	4,9%	43,6%
2013	4,9%	45,2%
Total	**4,5%**	**42,8%**

TABLE 2.2 – Taux de censure à droite en volume et en charge D/D nette de recours par année de survenance et au total - vision à fin 12/2019

Enfin, en visualisant la distribution de la charge D/D selon le statut du sinistre (clos ou en cours), nous pouvons clairement noter qu'en moyenne la charge des sinistres en cours est plus importante que celle des clos, soit une charge moyenne de 64 977€ pour les sinistres en cours *vs.* 4 317€ pour les sinistres clos. Une des explications principales justifiant une telle différence est que généralement un sinistre grave à moyennement grave nécessite un temps de stabilisation de l'état de santé de l'assuré donc dure dans le temps et implicitement implique des charges importantes. A contrario, un sinistre de gravité mineure implique une indemnisation rapide puis une clôture du dossier. Il est également à noter que l'indemnisation sous forme de rente est un axe d'explication, néanmoins la contribution à cette censure n'est pas prépondérante.

Ce constat est encore une fois un argument majeure pour prendre en compte les sinistres en cours lors de la modélisation de la charge ultime car les négliger impliquerait une sous-estimation de cette dernière.

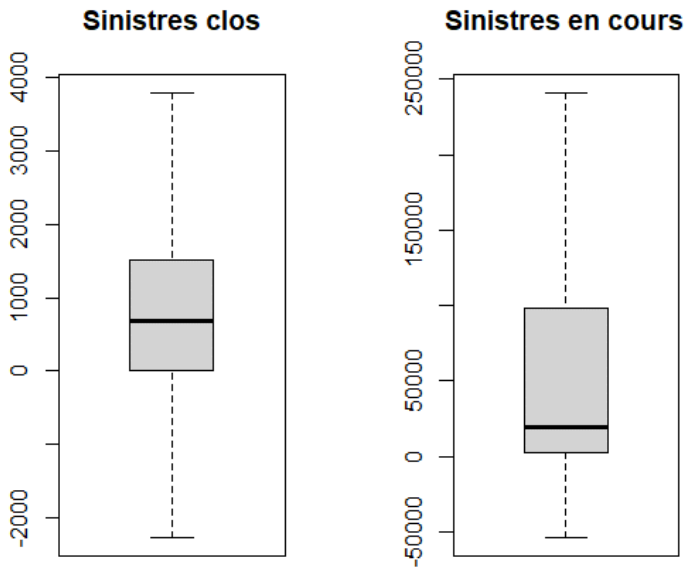


FIGURE 2.4 – Boxplots de la charge D/D nette de recours des sinistres clos vs. en cours (survenances 2010 à 2013) - vision à fin 12/2019

2.3.4 Analyse de corrélation

Cette partie s'inscrit dans la continuité de l'analyse des données individuelles, l'objectif est d'identifier dans un premier temps les variables les plus discriminantes, celles qui captent au mieux le degré de gravité du sinistre et in fine son coût final. Ainsi, les variables ayant le plus d'impact au niveau bidimensionnel sont celles présentées dans ce qui suit.

Les figures 2.5 et 2.6 nous permettent de visualiser la distribution de la charge en fonction du nombre de dommages corporels et de la présence à minima d'une victime grave. En effet, nous sommes amenés à penser que plus le nombre de dommages est important, plus le sinistre est onéreux à gravité équivalente. La part des sinistres dont le nombre de dommages excède 6 est de 1% pour une charge moyenne de 204 676€. Quant à la présence à minima d'une victime grave, la proportion de sinistres est plus conséquente atteignant 10% pour une charge moyenne de 65 435€ contre 6 623€ pour les non graves.

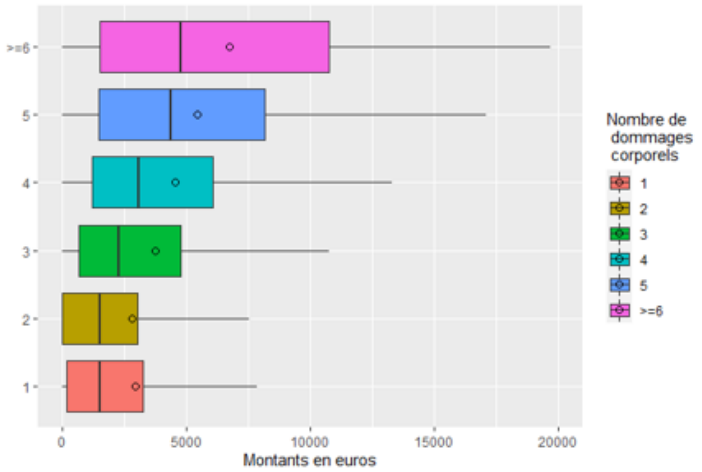


FIGURE 2.5 – Boxplots de la charge D/D nette de recours selon la variable « nombre de dommages corporels » - vision à fin 12/2019

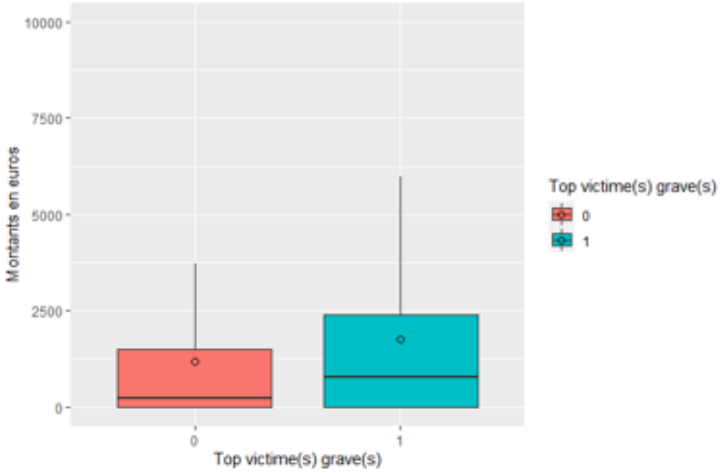


FIGURE 2.6 – Boxplots de la charge D/D nette de recours selon la variable « top victime(s) grave(s) » - vision à fin 12/2019

Comme mentionné précédemment, nous avons créé la variable « top_S1 » à partir de la charge D/D nette de recours observée la deuxième année après l'année de survenance où l'idée est d'appréhender à partir de cette variable le niveau de gravité du sinistre à sa clôture. Les seuils retenus sont : 25k€, 196k€, 618k€ et 1,4M€ et les modalités respectives sont : "NORM", "MOY1", "MOY2", "GRAV1" et "GRAV2" (cf. figure 2.7).

La distribution de la charge en fonction de la variable « litige » en figure 2.7 permet de confirmer que les sinistres en contentieux présentent une charge moyenne nettement plus importante que ceux réglés à l'amiable : 1% des sinistres sont en contentieux pour une charge moyenne de 196 635€ vs. 10 591€ pour les sinistres à l'amiable. Le montant des frais des sinistres en contentieux explique en partie l'importance du coût de revient de ces sinistres. Ces frais sont principalement liés aux honoraires d'avocats, huissiers de justices, expertises imposées par le tribunal, etc. De plus, les dossiers où il y a recours au tribunal concernent majoritairement des dossiers à forts enjeux en terme d'indemnisation. Un dernier point concerne la durée de ces sinistres qui peut prendre plusieurs années suite à l'attente de décision du tribunal : deux tiers des sinistres en contentieux sont toujours en cours.

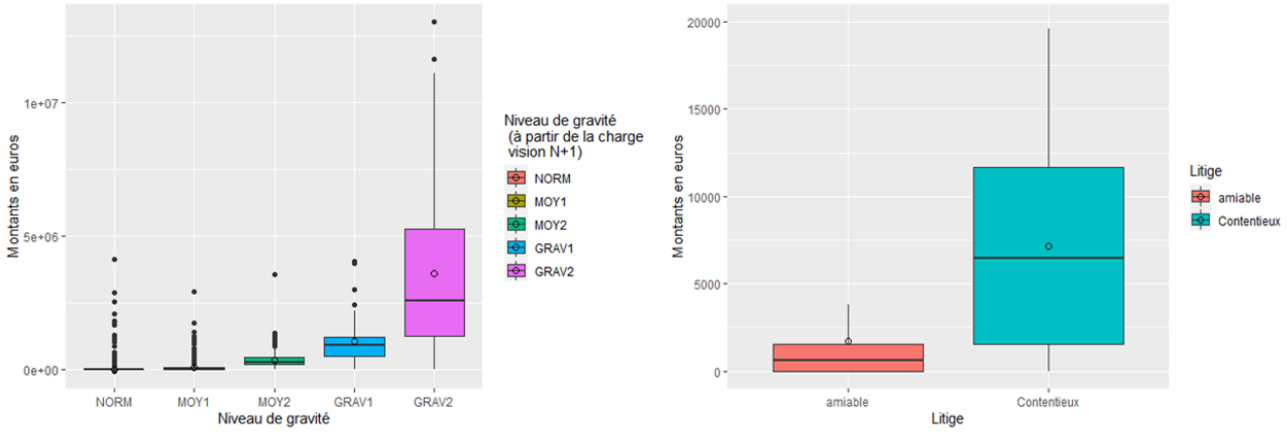


FIGURE 2.7 – Boxplots de la charge D/D nette de recours selon les variables « top_S1 » (niveau de gravité) et « litige » - vision à fin 12/2019

Enfin, la figure 2.8 montre que les sinistres liquidés sous forme de rente ont une charge moyenne nettement supérieure à celle des sinistres liquidés sous forme de capital : une charge moyenne de 1,6M€ pour les sinistres en rente vs. 9 820 pour les sinistres en capital. Néanmoins, l'indemnisation en rentre reste minoritaire et concerne moins de 1% des sinistres.

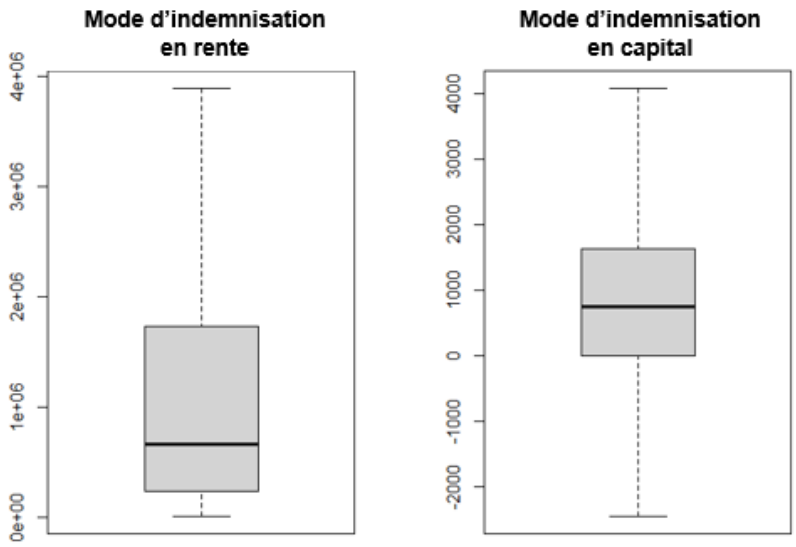


FIGURE 2.8 – *Boxplots* de la charge D/D nette de recours des sinistres indemnisés en rente vs. en capital (survenances 2010 à 2013) - vision à fin 12/2019

Dans un second temps, nous avions pour objectif d'identifier les corrélations qui peuvent exister entre les variables explicatives, notamment les quantitatives. Ceci permettra de mieux comprendre les variables manipulées et d'en supprimer quelques-unes en cas d'importantes corrélations. Nous avons réalisé cela au travers la matrice de corrélation de Spearman dont les résultats sont présentés en annexe A.5. La corrélation de Spearman permet de mesurer le lien entre deux variables et il est recommandé de l'utiliser en présence de variables dont la distribution ne suit pas une loi normale. Toutefois, ces résultats ne sont pas concluants. Nous n'identifions pas d'importantes corrélations en dehors de celles qui existent implicitement entre les différents montants.

Dans un dernier temps, nous avons cherché à explorer les données en multidimensionnel en ayant notamment recours à deux méthodes d'analyse factorielle :

- l'analyse en composantes principales (ACP) : cette méthode, introduite par HOTELLING (1933), porte sur les variables quantitatives uniquement. Elle permet de visualiser les données sur une dimension plus réduite tout en cherchant à conserver au mieux la réalité, autrement dit elle consiste à représenter les données en fonction de composantes principales décorrélées linéairement entre elles ;

- l'analyse factorielle multiple des correspondances (AFCM) : cette méthode est plutôt réservée à l'analyse descriptive de plusieurs variables qualitatives. Pour davantage de détails sur cette méthode, il est possible de se référer à l'ouvrage BRIGITTE (2008).

Les résultats, n'étant pas très probants, ne sont pas présentés dans cette section.

3 Méthodes usuelles de provisionnement

L'objectif de ce chapitre est de se concentrer sur quelques méthodes usuelles pour estimer le montant des provisions et en particulier mesurer la marge d'erreur associée. Dans un premier temps, il est pertinent de rappeler leur fondement théorique puis dans un second temps, de les appliquer au niveau des données présentées dans le chapitre précédent tout en vérifiant les hypothèses sous-jacentes.

Avant d'aborder ces quelques méthodes, il faut mentionner qu'il est possible de distinguer deux grandes familles au sein des méthodes agrégées : les méthodes déterministes et les méthodes stochastiques. Il est possible de se contenter de l'utilisation des méthodes déterministes si l'intérêt est uniquement l'estimation du montant des provisions. Si l'objectif est également d'évaluer et de réduire la volatilité de cette estimation sur différents horizons, il sera essentiel de s'appuyer sur des méthodes plutôt stochastiques. L'ASTIN [1], première section de l'association actuarielle internationale (AAI) dont l'objectif principal est la promotion de la recherche actuarielle, notamment en assurance non-vie, a publié un rapport en 2016 qui permet d'avoir une vision globale sur les méthodes de provisionnement les plus répandues : la méthode Chain-Ladder ressort en tête de liste pour les méthodes déterministes et la méthode de Mack (le modèle stochastique sous-jacent au modèle déterministe de Chain-Ladder) pour les méthodes stochastiques (cf. annexe A.3).

Fort de ce constat, il est intéressant de commencer par présenter la méthode déterministe de Chain-Ladder, puis dans un second temps, de s'attarder sur le modèle de Mack, un modèle stochastique permettant l'estimation de l'erreur de prédiction sur la provision à l'ultime. Enfin, introduire le modèle de Merz-Wüthrich permettra d'expliciter une méthode d'estimation de l'incertitude sur le montant des provisions à horizon un an, une notion particulièrement intéressante suite à la directive «Solvabilité II». Avant d'aborder ces modèles, il serait utile de commencer par définir quelques notations et notions qui seront évoquées par la suite.

3.1 Généralités

3.1.1 Notations

Comme décrit dans le premier chapitre (partie 1.2.2), les données pour la grande majorité utilisées par les méthodes agrégées se présentent sous forme de triangles de liquidation. Les lignes ($i = 0, \ldots, I$) correspondent aux exercices de rattachement des sinistres et les colonnes correspondent aux périodes de

1. ASTIN : *Actuarial Studies in Non-life Insurance*

développement ($j = 0, \ldots, J$). Dans les cas traités, les exercices de rattachement seront des années de survenance et les périodes de développement correspondront à des années de développement.

Il est également utile de définir les deux quantités suivantes :

- $X_{i,j}$ représente le montant incrémental de la variable d'intérêt pour l'année de de rattachement i et l'année de développement j
- $C_{i,j}$ représente le montant cumulé de la variable d'intérêt pour l'année de de rattachement i et l'année de développement j

Avec

$$\forall i \in \{0, \ldots, I\}, \forall j \in \{0, \ldots, J\},\ C_{i,j} = \sum_{t=0}^{j} X_{i,t}\ et\ X_{i,j} = C_{i,j} - C_{i,j-1}.$$

Ainsi, la figure 3.1 représente un triangle à I années de survenance et J années de développement.

Année de survenance	Année de développement						
	0	1	...	j	...	J - 1	J
0	$C_{0,0}$	$C_{0,1}$	...	$C_{0,j}$	...	$C_{0,J-1}$	$C_{0,J}$
1	$C_{1,0}$	$C_{1,1}$	...	$C_{1,3}$	...	$C_{1,J-1}$	
⋮	⋮		⋰	⋰			
i	$C_{i,0}$	⋰	⋰				
⋮	⋮	⋰					
I - 1	$C_{I-1,1}$	$C_{I-1,1}$					
I	$C_{I,0}$						

FIGURE 3.1 – Représentation d'un triangle d'une « variable d'intérêt » en cumulé

3.1.2 Formalisme du problème de provisionnement

Pour poser le problème du provisionnement, nous considérons que $I = J = n$, que les $C_{i,j}$ désignent les paiements cumulés pour l'année de survenance i au bout de j années et les $X_{i,j}$ comme étant les incréments de paiements des sinistres survenus l'année i pour l'année de développement j.

Le problème du provisionnement revient à un problème de prédiction conditionnel aux informations disponibles au moment de l'estimation. Ainsi, l'information disponible à la date n peut être exprimée de la façon suivante

$$\mathcal{D}_n = \{C_{i,j},\ pour\ i + j \leq n\} = \{X_{i,j},\ pour\ i + j \leq n\}.$$

Et avec une vision pour chaque année de survenance i, l'information disponible peut être exprimée de la sorte

$$\mathcal{D}_{i,n-i} = \{C_{i,j},\ pour\ j = 0, \ldots, n - i\} = \{X_{i,j},\ pour\ j = 0, \ldots, n - i\}.$$

L'objectif est d'étudier la loi conditionnelle de la charge ultime $C_{i,\infty}$ sachant $\mathcal{D}_{i,n-i}$ pour chaque année de survenance. Il est possible d'étudier plutôt la loi de $C_{i,n}$ si nous prenons pour hypothèse que les sinistres sont clos au bout de n années.

Au final, nous chercherons à travers les méthodes usuelles que nous présenterons ci-dessous à estimer les éléments suivants :

- **la provision pour sinistre à payer :** il s'agira de prédire dans un premier temps le montant total des sinistres à payer pour chaque année de survenance i, c'est-à-dire

$$\hat{C}_{i,n}^{(n-i)} = \mathbb{E}[C_{i,n} \mid \mathcal{D}_{i,n-i}].$$

  Puis déduire le montant de la provision pour sinistre à payer en retranchant les paiements déjà effectués

$$\hat{R}_i = \hat{C}_{i,n}^{(n-i)} - C_{i,n-i}.$$

- **l'incertitude à horizon ultime :** il s'agit de l'erreur de prédiction associée à la prédiction de la charge ultime, ainsi il faudra calculer les variances pour chaque année de survenance i : $\mathbb{V}[\hat{C}_{i,n}^{(n-i)}]$.

- **l'incertitude à horizon un an :** mesure proposée et requise par la directive « Solvabilité II » qui a pour but de quantifier le changement d'estimation de la charge ultime effectuée l'année suivant l'année d'estimation, suite à la détention d'informations supplémentaires, comparé à l'estimation faite à la date d'aujourd'hui. Cette notion est désignée par le terme CDR [2] et sera développée plus en détail par la suite.

3.2 Méthode de Chain-Ladder déterministe

Dans un premier temps, il est utile et nécessaire de présenter la méthode de Chain-Ladder. Probablement encore, la méthode la plus répandue parmi toutes les autres méthodes de provisionnement, de part sa simplicité tant bien d'un point de vue compréhension et communication que mise en place.

3.2.1 Principe

Estimer le montant des provisions en utilisant la méthode de Chain-Ladder revient à supposer que les cadences de paiement observées dans le passé se reproduiront dans le futur. Autrement dit, ceci revient à prendre pour hypothèse forte que les ratios $\frac{C_{i,j+1}}{C_{i,j}}$ sont indépendants de l'année d'origine i. Ainsi, cette méthode repose sur les deux hypothèses suivantes :

- **H1 :** les montants cumulés d'années de survenance différentes sont des variables aléatoires indépendantes, c'est-à-dire que : $\forall i \neq k, C_{i,j} \perp C_{k,j}$.

- **H2 :** $(C_{i,j})_j$ est une chaîne de Markov et il existe $f_0, \ldots, f_{J-1} > 0$ tels que pour tout i, j

$$\mathbb{E}[C_{i,j} \mid C_{i,0}, \ldots, C_{i,j-1}] = \mathbb{E}[C_{i,j} \mid C_{i,j-1}] = f_{j-1} C_{i,j-1}.$$

2. CDR : *Claims Development Result*

Pour tout j $\in \{0, \ldots, J-1\}$, le facteur de développement de l'année j peut être estimé par la formule 3.1.

$$\hat{f}_j^{CL} = \frac{\sum_{i=0}^{I-j-1} C_{i,j+1}}{\sum_{i=0}^{I-j-1} C_{i,j}}. \tag{3.1}$$

Il s'agit d'un estimateur sans biais du facteur de développement f_j conditionnement ou pas à l'information passée.

L'estimation du montant des réserves en utilisant Chain-Ladder se résume autour des étapes qui suivent :

1. Calcul des différents estimateurs des facteurs de développement : $\hat{f}_0^{CL}, \ldots, \hat{f}_{J-1}^{CL}$.
2. Calcul de la charge ultime pour chaque année de survenance à partir de la relation 3.2.

$$\hat{C}_{i,J}^{CL} = \hat{f}_{J-1}^{CL} \times \ldots \times \hat{f}_{I-i}^{CL} \times C_{i,I-i}. \tag{3.2}$$

3. Calcul du montant des provisions par exercice de survenance, puis du montant total selon la formule 3.3.

$$\hat{R}^{CL} = \sum_{i=0}^{I} \hat{R}_i^{CL} = \sum_{i=0}^{I} (\hat{C}_{i,J}^{CL} - C_{i,J-i}). \tag{3.3}$$

Il est essentiel de garder à l'esprit l'importance des choix effectués lors de l'estimation des facteurs de développement pour in fine estimer les charges ultimes. Ce choix doit être, en théorie, en phase avec la méthode de Chain-Ladder qui part du principe que le développement de la charge d'une année à l'autre est indépendant de l'année de survenance. Ainsi, pour vérifier cette hypothèse, il est par moment nécessaire d'exclure certains facteurs estimés aberrants : facteurs pouvant , par exemple, résulter d'une année où la gestion de sinistres a été au ralenti, donc des règlements plus en retard comparé à la normale, ou un changement d'outil de gestion donc des clôtures de sinistres plus tardives, etc.

Toujours, dans la perspective d'avoir l'estimation la plus juste possible, il est également important de tenir compte dans la projection de la sinistralité si changements opérés récemment dont les conséquences ne sont pas encore visibles, tels que des changements dans la gestion, dans le portefeuille, etc. De ce fait, pour opérer le choix de calcul de ces facteurs de développement, il est judicieux d'échanger avec le métier et d'avoir une connaissance des changements opérées sur la ligne de business en question.

Enfin, dans certains cas pratiques, il s'avère nécessaire de procéder à un lissage des facteurs de développement estimés dans le but de limiter et/ou de supprimer les irrégularités présentes dans le triangle de liquidation. Ces irrégularités peuvent provenir de données manquantes ou de valeurs aberrantes. Il est également possible de recourir au lissage afin d'estimer un facteur de queue, notamment dans le cas des branches à développement long où il est compliqué d'avoir un triangle de liquidation avec une ou plusieurs années de survenance dont le développement est complet. Nous pouvons prendre pour exemple la branche « Responsabilité civile » où les sinistres corporels peuvent prendre plusieurs années pour être clôturés. Dans ce cas, il est indispensable d'estimer un facteur de queue pour pouvoir estimer le développement de la sinistralité au-delà du triangle de liquidation. Une des méthodes utilisées pour cette estimation est le lissage.

3.2.2 Avantages et inconvénients

La méthode de Chain-Ladder présente l'avantage d'être une méthode simple à mettre en place qui permet de communiquer relativement rapidement sur la valeur des provisions à détenir. Toutefois, ce mo-

dèle présente comme premier inconvénient les hypothèses sur lesquelles il repose. Il s'agit d'hypothèses fortes qui ne sont pas forcément vérifiées en pratique. En effet, le modèle de Chain-Ladder suppose que le schéma de développement est identique quelque soit l'année de survenance. Ainsi, tout changement au niveau réglementaire, au niveau de la politique de gestion des sinistres, ou encore à d'autres niveaux, peut impacter les cadences de règlement d'une ou plusieurs années de survenance et donc annuler cette hypothèse. Cette méthode suppose également que l'évolution de la charge de la sinistralité est uniquement dépendante de la durée de développement des sinistres, ce qui n'est pas forcément toujours le cas. D'autre part, nous pouvons également reprocher à cette méthode d'être très sensible aux variations des données observées, notamment la présence de sinistres graves ou atypiques.

Pour finir, cette méthode, étant déterministe, elle permet d'estimer le montant des provisions *« Best Estimate »* sans pour autant permettre d'estimer la volatilité des réserves et donc déterminer l'erreur de prédiction à l'ultime. Néanmoins, il est évident que les survenances les plus récentes présentent une incertitude non négligeable. Si nous prenons pour exemple l'année de survenance la plus récente, sa charge ultime est égale au produit de tous les facteurs de développement estimés multiplié par le montant des règlements effectués la première année. Ainsi, il suffit d'un ralentissement de cadence de règlements courant cette première année pour sous-estimer la charge ultime de cette année de survenance et au contraire un risque de surestimer la charge en cas d'accélération de la cadence.

3.3 Méthode de Mack

Après avoir présenté la méthode déterministe la plus pratiquée dans sa catégorie pour estimer le montant *« Best Estimate »* des provisions pour sinistre à payer et également cité ses avantages et inconvénients, nous allons nous attarder sur une des méthodes de provisionnement stochastiques les plus répandues, une méthode développée par (Mack, 1993a).

En général, comme mentionné précédemment, les méthodes stochastiques présentent un intérêt supplémentaire à celui de l'estimation du montant des provisions qui est la possibilité d'évaluer l'incertitude des prédictions. Le modèle de Mack propose une formule fermée permettant de quantifier cette incertitude. Avant de présenter le modèle dans les détails, nous allons commencer par définir les composants d'une erreur de prédiction MSEP [3].

3.3.1 Erreur de prédiction (MSEP)

Un des critères permettant de mesurer la précision d'estimation du montant des provisions est l'erreur quadratique moyenne de prédiction. Commençons alors par définir cette mesure.
Soit $\hat{\theta}$ l'estimateur de θ, la MSEP de décompose alors selon la formule 3.4.

$$MSEP_{|D}(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \theta)^2 \mid D] = \underbrace{(\hat{\theta} - \mathbb{E}[\theta \mid D])^2}_{Erreur\ d'estimation} + \underbrace{VAR(\theta \mid D)}_{Erreur\ de\ processus} . \tag{3.4}$$

Le premier terme désigne l'erreur d'estimation des paramètres. En général cette erreur diminue d'autant plus que nous disposons d'observations. Quant au second terme, il désigne la variance conditionnelle du processus, c'est-à-dire la variance qui est à l'intérieur du modèle stochastique. Le premier terme est le seul paramètre sur lequel nous avons la main, contrairement au second terme qui est purement aléatoire et ne peut être minimisé.

3. MSEP : *Mean Square Error of Prediction*

3.3.2 Principe

Le modèle de Mack est un modèle non-paramétrique. Il est le pendant stochastique du modèle déterministe de Chain-Ladder. En effet, le montant des provisions estimé est identique à celui obtenu par la méthode de Chain-Ladder avec l'avantage d'obtenir l'estimation de l'erreur de prédiction. Autrement dit, Mack, au travers de son approche probabiliste, a permis de mesurer la volatilité des réserves déterministes de Chain-Ladder.

Le modèle de Mack réitère les deux hypothèses **H1** et **H2** considérées par Chain-Ladder (cf. sous-section 3.2.1) et ajoute une troisième hypothèse sur les moments du second ordre afin d'estimer l'erreur de prédiction sur la provision « à l'ultime ». Cette troisième hypothèse s'exprime de la façon suivante :

- **H3 :** il existe $\sigma_0^2, \dots, \sigma_{J-1}^2 > 0$ tels que pour tout i, j

$$VAR[C_{i,j} \mid C_{i,0}, \dots, C_{i,j-1}] = VAR[C_{i,j} \mid C_{i,j-1}] = \sigma_{j-1}^2 C_{i,j-1}.$$

Pour rappel, les estimateurs de Chain-Ladder (3.2) sont sans biais et non corrélés. De plus, la troisième hypothèse assure à ces estimateurs d'être de variance minimale comparé à toutes les combinaisons linéaires sans biais des facteurs de développement individuels.

Comme indiqué précédemment, le modèle de Mack fournit une formule fermée permettant d'estimer l'erreur quadratique moyenne de prédiction (MSEP). Ceci permet d'évaluer l'incertitude d'estimation des provisions pour chaque année de survenance ainsi que pour toutes les années de survenance.

Tout d'abord, nous avons l'estimateur de la variance conditionnelle calculé par la formule suivante

$$\hat{\sigma}_j^2 = \begin{cases} \frac{1}{I-j-1} \sum_{i=0}^{I-j-1} C_{i,j}(f_{i,j} - \hat{f}_j^{CL})^2, & \textit{pour } j \leq J-2 \\ min(\frac{\hat{\sigma}_{J-2}^4}{\hat{\sigma}_{J-3}^2}, min(\hat{\sigma}_{J-3}^2, \hat{\sigma}_{J-2}^2)), & \textit{pour } j = J-1. \end{cases}$$

Étant donné que

$$C_{i,J} - \hat{C}_{i,J} = C_{i,J} - C_{i,J-i+1} - (\hat{C}_{i,J} - C_{i,J-i+1}) = R_i - \hat{R}_i.$$

nous pouvons en déduire que

$$MSEP(\hat{C}_{i,J}) = MSEP(\hat{R}_i).$$

D'autre part, étant donné que $\hat{C}_{i,J}^{CL}$ est D_I-mesurable, nous pouvons obtenir une décomposition de la MSEP conditionnelle pour chaque année de survenance exprimée de la façon suivante

$$MSEP_{|D_I}(\hat{C}_{i,J}^{CL}) = (\hat{C}_{i,J}^{CL} - \mathbb{E}[C_{i,J} \mid D_I])^2 + VAR(\hat{C}_{i,J}^{CL} \mid D_I)$$

où D_I désigne l'information disponible au moment de l'estimation.

Ainsi, sous les hypothèses du modèle de Mack, nous avons l'estimation suivante pour chaque année de survenance i selon la formule 3.5.

$$\widehat{MSEP}_{|D_I}(\hat{R}_i^{CL}) = \widehat{MSEP}_{|D_I}(\hat{C}_{i,J}^{CL}) = (\hat{C}_{i,J}^{CL})^2 \sum_{j=I-i}^{J-1} \frac{\hat{\sigma}_j^2}{(\hat{f}_j^{CL})^2} \left(\frac{1}{\sum_{l=0}^{I-j-1} C_{l,j}} + \frac{1}{\hat{C}_{i,J}^{CL}} \right). \tag{3.5}$$

Le rapport $\frac{1}{\sum_{l=0}^{I-j-1} C_{l,j}}$ représente la j$^{\text{ème}}$ contribution à l'erreur d'estimation et le rapport $\frac{1}{\hat{C}_{i,j}^{CL}}$ la j$^{\text{ème}}$ contribution à l'erreur du processus.

En définitive, l'estimation de la MSEP du montant de provision total est définie par la formule 3.6.

$$\widehat{MSEP}_{|D_I}(\hat{R}^{CL}) = \sum_i \widehat{MSEP}_{|D_I}(\hat{C}_{i,J}^{CL}) + 2\sum_{i<m} \hat{C}_{i,J}^{CL}\hat{C}_{m,J}^{CL} \sum_{j=I-i}^{J-1} \frac{\hat{\sigma}_j^2}{(\hat{f}_j^{CL})^2 \sum_{l=0}^{I-j-1} C_{l,j}}. \tag{3.6}$$

3.3.3 Vérification des hypothèses du modèle

Lors de l'application du modèle, il est nécessaire de s'assurer que les hypothèses sous-jacentes sont bien vérifiées au niveau du triangle en question. Les différentes méthodes et tests, exposées au sein de (Mack, 1993b), permettent de confirmer ou d'infirmer les trois hypothèses émises par le modèle de Mack. Une brève description de ceux-ci est présentée ci-dessous.

Vérification de l'hypothèse H1

L'hypothèse d'indépendance entre les années de survenance reste une des plus fondamentales parmi les trois hypothèses retenues. Elle peut être annulée pour plusieurs raisons dont notamment la présence d'effets calendaires. Ces effets peuvent résulter de changements aussi bien internes, tel qu'un changement de gestion de la sinistralité, qu'externes, tel qu'un changement de législation. La présence d'un effet calendaire impacte une des diagonales du triangle, une diagonale dont les éléments contribuent à l'estimation des facteurs de développement individuels adjacents. Ces derniers seront donc soit surestimés, et les facteurs suivants sous-estimés, soit l'inverse. Fort de ce constat, Thomas Mack a proposé le test d'indépendance résumé dans les étapes suivantes.

1. Calculer et ordonner les facteurs de développement individuels pour tout $0 \leq j \leq I-1$, $F_j = \{ \frac{C_{i,j+1}}{C_{i,j}}, \forall\ 0 \leq i \leq I-j-1 \}$, autrement dit tous les facteurs de passage de l'année de développement j à l'année $j+1$.
2. Diviser les facteurs de développement individuels présents dans F_j en deux sous-groupes : les inférieurs à la médiane de F_j, noté LF_j, et ceux qui lui sont supérieurs noté SF_j. Les facteurs de développements égaux à la médiane sont supprimés.
   Ceci est appliqué pour tout $0 \leq j \leq I-1$, ainsi $L = LF_0 + \cdots + LF_{I-2}$ et $S = SF_0 + \cdots + SF_{I-2}$.
3. Considérer $Z_j = min(L_j, S_j)$ où L_j correspond au nombre de petits facteurs ($\in L$) présents dans la diagonale j et S_j le nombre de grands facteurs ($\in S$) présents dans la diagonale j.
   Sous l'hypothèse nulle d'absence d'effets calendaires, chaque facteur de développement a autant de chance d'appartenir au groupe L qu'au groupe S. Ceci amène à considérer que Z_j suit une loi binomiale de paramètres $\frac{1}{2}$ et $N_j = L_j + S_j$.
4. Effectuer le test au niveau de la variable globale $Z = \sum_{j=1}^{I} Z_j$ dans le but de contourner la problématique de cumul des probabilités d'erreurs. Cette variable, étant la somme de variables aléatoires (presque) non corrélées sous l'hypothèse d'absence d'effets calendaires, a pour espérance $\mathbb{E}(Z) = \sum_{j=1}^{I} \mathbb{E}(Z_j)$ et pour variance $VAR(Z) = \sum_{j=1}^{I} VAR(Z_j)$. Il s'ensuit de considérer sous l'hypothèse nulle du test que Z suit asymptotiquement une loi normale. Ainsi, il est possible de définir l'intervalle de confiance suivant $IC_{1-\alpha} = \left[\mathbb{E}(Z) - q_{1-\frac{\alpha}{2}}\sqrt{VAR(Z)}\ ,\ \mathbb{E}(Z) + q_{1-\frac{\alpha}{2}}\sqrt{VAR(Z)}\right]$ où

$q_{1-\frac{\alpha}{2}}$ correspond au quantile d'ordre $1-\frac{\alpha}{2}$ d'une loi normale centrée réduite. Pour rappel, l'hypothèse nulle n'est pas rejetée si $Z \in IC_{1-\alpha}$ et elle est rejetée si $Z \notin IC_{1-\alpha}$.

Vérification de l'hypothèse H2

Cette hypothèse implique que les facteurs de développement successifs $\frac{C_{i,j}}{C_{i,j-1}}$ et $\frac{C_{i,j+1}}{C_{i,j}}$ sont non corrélés. Afin de vérifier cette dernière, Thomas Mack a proposé le test du coefficient de corrélation de rang de Spearman. Le choix de ce test provient du constat que les facteurs de développement adjacents satisfont rarement les deux hypothèses généralement requises par le test usuel de non-corrélation qui se résument en paires identiquement distribuées et de distribution normale. Le test retenu, étant non paramétrique, semble donc approprié pour ce cas de figure.

Ainsi, la marche à suivre pour effectuer le test proposé peut être résumé comme il s'ensuit ci-dessous.

1. Ordonner les facteurs de développement individuels de deux années de développement successives $j-1$ et j de façon croissante. A noter que le dernier facteur de développement de l'année j est supprimé étant donné qu'il ne possède pas d'équivalent l'année $j-1$.
   Ainsi, soient $r_{i,j} \in [1, I-j]$ les rangs des facteurs $\frac{C_{i,j+1}}{C_{i,j}}$ pour tout $0 \leq i \leq I-j-1$ et $s_{i,j} \in [1, I-j]$ les rangs des facteurs $\frac{C_{i,j}}{C_{i,j-1}}$ pour tout $0 \leq i \leq I-j-1$. Le coefficient de corrélation de rang de Spearman T_j est calculé à partir de la formule suivante

$$T_j = 1 - 6. \sum_{i=0}^{I-j-1} \frac{(r_{i,j} - s_{i,j})^2}{((I-j)^3 - I + j)}.$$

   Si T_j est proche de 0, ceci implique que les facteurs de développement compris entre les années $j-1$ et j et ceux entre j et $j+1$ sont non corrélés. D'autres valeurs de T_j impliquent que ces derniers sont corrélés.

2. Calculer la statistique de test sans tenir compte de chaque paire d'années de développement séparément dans le but d'éviter l'accumulation des erreurs de probabilités. Ainsi, la statistique de test est calculée à partir de la formule suivante

$$T = \sum_{j=1}^{I-2} \frac{I-j-1}{(I-2)(I-3)/2}.T_j.$$

   Nous avons $\mathbb{E}(T) = 0$ car sous l'hypothèse nulle de ce test, l'espérance de T_j est nulle pour tout $1 \leq j \leq I-2$ et également en utilisant la non corrélation des T_j sous cette hypothèse nulle, nous obtenons la variance $VAR(T) = \frac{1}{(I-2)(I-3)/2}$.
   Pour finir, sous l'hypothèse nulle, nous pouvons supposer que la statistique de test T suit asymptotiquement une distribution normale. Étant donné que le test est effectué à titre approximatif, l'intervalle de confiance retenu est à 50%. Ainsi, l'hypothèse nulle de non corrélation n'est pas rejetée si la statistique de test satisfait la condition suivante

$$\frac{-0.67}{\sqrt{\frac{(I-2)(I-3)}{2}}} \leq T \leq \frac{0.67}{\sqrt{\frac{(I-2)(I-3)}{2}}}.$$

Une méthode graphique, en complément du test présenté ci-dessus, permet de vérifier l'hypothèse $H2$. Il suffit de représenter les couples $(C_{i,j}\ , C_{i,j+1})_{0 \leq i \leq I-j}$ pour tout $0 \leq j \leq J-2$. L'hypothèse est considérée satisfaite en cas d'alignement des points obtenus sur la droite passant par l'origine avec une pente égale au facteur de développement estimé $\hat{f}_j^{CL}$.

Vérification de l'hypothèse H3

Cette hypothèse peut être vérifiée graphiquement en représentant les résidus $(\frac{C_{i,j+1}-C_{i,j}.\hat{f}_j^{CL}}{\sqrt{C_{i,j}}})_{0\leq i\leq n-j}$ en fonction des $(C_{i,j})_{0\leq i\leq n-j}$. Cette dernière est alors validée si les résidus ne présentent aucune tendance particulière mais plutôt ont une distribution aléatoire. Ce test doit être effectué pour chaque année de développement, à condition d'avoir tout de même un nombre de points suffisants.

3.3.4 Avantages et inconvénients

L'application du modèle de Mack implique une estimation de provisions identique à celle de Chain-Ladder tant bien en méthodologie de calcul qu'en résultat, ainsi les mêmes avantages et inconvénients, cités dans la partie précédente, sont toujours valables pour le modèle de Mack.

Ce modèle présente un avantage supplémentaire étant donné qu'il permet d'évaluer l'incertitude sur la provision estimée à l'ultime grâce à la formule fermée proposée qui permet d'estimer l'erreur de prédiction. Néanmoins, étant donné que le modèle de Mack est non-paramétrique avec des hypothèses uniquement sur les deux premiers moments, il ne fournit pas toute la distribution des réserves. Ainsi, en se contentant du modèle de Mack sans hypothèse supplémentaire, il est impossible de calculer une VAR à $99,5\%$ par exemple. Le recours à cette méthode peut être accompagné d'une hypothèse supplémentaire sur la distribution des réserves afin d'obtenir différents quantiles des provisions à l'ultime. A titre d'exemple, il est possible de considérer une loi normale, une loi log-normale ou encore d'autres lois.

3.4 Méthode de Merz & Wüthrich

3.4.1 Volatilité à l'ultime vs. volatilité à horizon un an

A des fins comptables, il n'est nécessaire de connaître que l'estimation de la charge ultime. Toutefois, pour des raisons de solvabilité, il est également nécessaire de connaître la qualité de cette estimation de charge ultime. Pour mesurer cette dernière, il faut s'intéresser aux moments d'ordre 2. En étudiant le modèle de Mack, nous avons pu obtenir une formule permettant l'évaluation de cette incertitude totale dans l'évolution des sinistres jusqu'à ce que ces derniers soient définitivement réglés. L'étude de cette incertitude totale est une vision à long terme. Cependant, dans le cadre de la directive «Solvabilité II», nous devons nous attarder sur une deuxième vision tout aussi importante : la vision à court terme. Il est ainsi très utile de pouvoir mesurer la volatilité de nos provisions pour sinistres sur un horizon d'un an pour diverses raisons que nous citons ci-dessous :

- si le comportement à court terme n'est pas adéquat, l'entreprise ne peut tout simplement pas atteindre le « long terme » car elle sera déclarée insolvable avant de pouvoir atteindre le long terme.

- Une vision à court terme est pertinente pour les décisions de gestion, car des mesures doivent être prises régulièrement. Il s'agit par exemple des clôtures financières, de la tarification des produits d'assurance, des ajustements de primes, etc.

- Elle est communiquée à travers les états financiers et les rapports annuels et reflète la performance de l'entreprise à court terme, sa solidité financière ainsi que sa réputation sur le marché de l'assurance. Ceci est à fort intérêt et grande importance pour les régulateurs, les clients, les

investisseurs, les agences de notation, les marchés boursiers, etc.

Pour toutes ces raisons, il paraît pertinent de savoir quantifier de combien pourrait évoluer l'estimation de la charge ultime dans un an comparée à celle d'aujourd'hui. Formulé autrement, ceci revient à évaluer l'évolution possible des provisions présentes dans le passif entre l'année d'estimation et l'année qui suit. En effet, de ce besoin naît la nécessité d'estimer la volatilité du montant des réserves à horizon un. A cette fin, (**merzWuthrich2008**) ont développé un modèle qui fournit une formule fermée permettant l'estimation du risque de réserve à horizon un an. Avant de présenter ce modèle, nous allons définir la notion de CDR plus en détail.

3.4.2 Claims Development Result (CDR)

Comme précisé précédemment, les assureurs sont dans l'obligation de se prémunir contre d'éventuelles insuffisances de provisions pour sinistres au niveau du bilan et du compte de résultats à horizon d'un an. Une des conséquences de cette exigence est la nécessité d'étudier le CDR. Ce dernier est défini comme étant la différence entre deux estimations successives :

- l'estimation du montant de la charge ultime effectuée l'année I ;
- la ré-estimation du montant de la charge ultime de la même population de sinistres effectuée l'année $I + 1$.

L'objectif final est de pouvoir estimer la volatilité du CDR. En effet, en pratique, chaque année est mise à jour l'estimation de la charge ultime comme illustré en figure 3.2. Si l'estimation est revue à la hausse, ceci entraînera un « mali » et il faudra augmenter la provision afin de pouvoir payer les sinistres, si elle est plutôt revue à la baisse, nous parlerons de « boni ». C'est l'incertitude autour de ces bonis/malis que souhaitent les assureurs mesurer afin de pouvoir les couvrir.

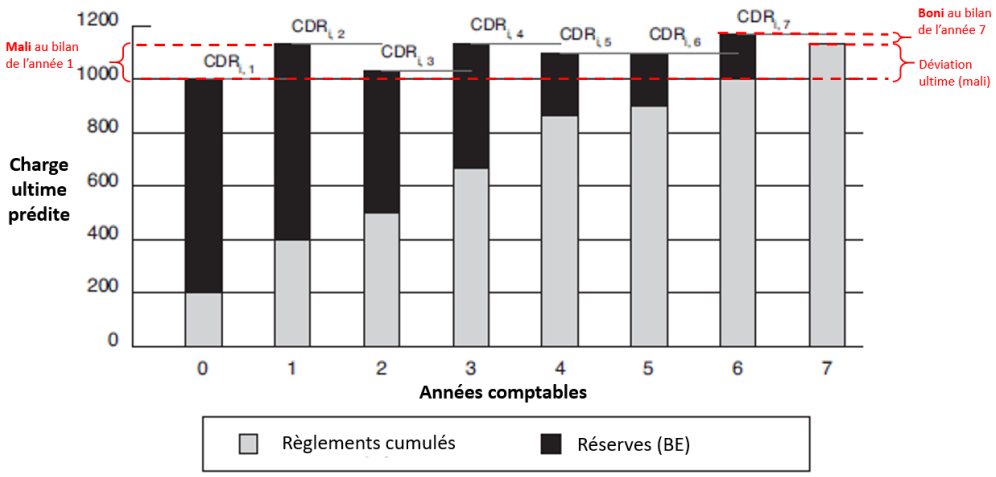


FIGURE 3.2 – Évolution de l'estimation de la charge ultime au cours des années

Il faudra tout d'abord commencer par faire la distinction entre le « CDR réel » qui a pour critère de ne pas être observable à la fin de l'année I et le « CDR observable », qui quant à lui, représente la position figurant dans le compte de résultat au 31 décembre de l'année I. Ce poste est dans le budget énoncé prédit par 0. Nous allons par la suite mesurer la qualité de prédiction, ce qui détermine les exigences de

solvabilité (vision prospective).

Avant de définir ces deux notions, commençons par définir et illustrer ci-dessous :

- l'information disponible l'année I, que nous désignerons par D_I

$$D_I = \{C_{i,j};\ i+j \leq I \ et \ i \leq I\}. \tag{3.7}$$

- et l'information disponible l'année suivante i.e $I+1$ que nous désignerons par D_{I+1}, tel que

$$D_{I+1} = \{C_{i,j};\ i+j \leq I+1 \ et \ i \leq I\} = D_I \cup \{C_{i,I-i+1};\ i \leq I\}. \tag{3.8}$$

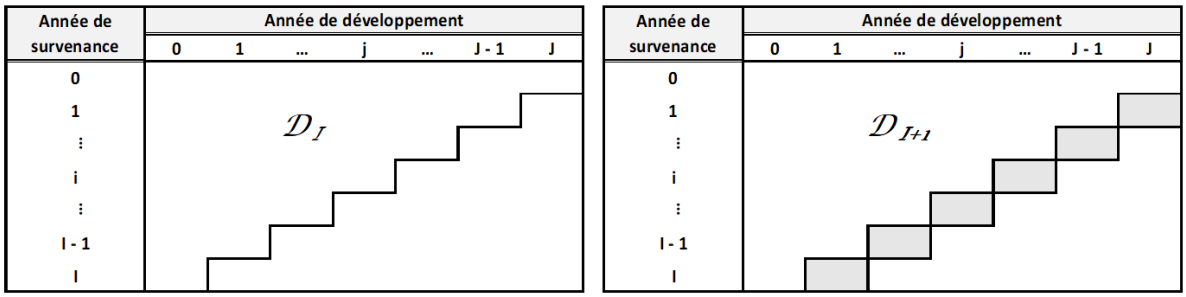


FIGURE 3.3 – Représentation des triangles de liquidation en t $= I$ et en t$= I+1$)

A présent, définissions le « CDR réel » puis dans un second temps le « CDR observable » :

- **Le CDR réel :** pour l'année comptable $]I, I+1]$, pour toute année de survenance $i \in \{0, \ldots, I\}$,

$$\begin{aligned} CDR_i(I+1) &= \mathbb{E}[C_{i,J} \mid D_I] - \mathbb{E}[C_{i,J} \mid D_{I+1}] \\ &= \mathbb{E}[R_i^I \mid D_I] - (X_{i,J-i+1} + \mathbb{E}[R_i^{I+1} \mid D_{I+1}]) \end{aligned}$$

  Avec : $R_i^I = C_{i,J} - C_{i,J-i}$, $R_i^{I+1} = C_{i,J} - C_{i,J-i+1}$ et $X_{i,J-i+1} = C_{i,J-i+1} - C_{i,J-i}$.

  Il faut noter que le CDR réel n'est pas calculable avant la fin de l'année $I+1$. Ainsi, son estimation nécessite le recours à estimer les deux termes $\mathbb{E}[C_{i,J} \mid D_I]$ et $\mathbb{E}[C_{i,J} \mid D_{I+1}]$ en utilisant les estimateurs de Chain-Ladder. Ce qui introduit le notion de « CDR observable ».
- **Le CDR observable :** pour l'année comptable $]I, I+1]$, pour toute année de survenance $i \in \{0, \ldots, I\}$,

$$\begin{aligned} \hat{CDR}_i(I+1) &= \hat{C}_{i,J}^I - \hat{C}_{i,J}^{I+1} \\ &= \hat{R}_i^{D_I} - (X_{i,I-i+1} + \hat{R}_i^{D_{I+1}}) \end{aligned}$$

  Avec : $\hat{R}_i^{D_I} = \hat{C}_{i,J}^I - C_{i,I-i}$ et $\hat{R}_i^{D_{I+1}} = \hat{C}_{i,J}^{I+1} - C_{i,I-i+1}$.
  NB : $\hat{R}_i^{D_{I+1}}$ est un estimateur de $\mathbb{E}[R_i^{I+1} \mid D_{I+1}]$ et désigne le montant des réserves en temps $I+1$ vu en temps I.

3.4.3 Vision rétrospective et vision prospective

Il est évident que l'information en $t = I + 1$ est indisponible en $t = I$. Partant de ce constat et en se plaçant dans le cadre du modèle de Mack, Merz & Wüthrich ont proposé deux visions afin de mesurer l'erreur de prédiction du CDR : une vison rétrospective et/ou une vision prospective.

- La **vision rétrospective**, une vision qui permettrait de faire du *backtesting* sur la prédiction effectuée l'année I, en se positionnant en l'année calendaire $I + 1$. Elle s'exprime, pour chaque année de survenance i, par la formule 3.9.

$$MSEP_{CDR_i(I+1)|D_I}(C\hat{D}R_i(I+1)) = \mathbb{E}[(CDR_i(I+1) - C\hat{D}R_i(I+1))^2 \mid D_I]. \tag{3.9}$$

  Cette erreur ne sera pas d'avantage développée dans le cadre de ce mémoire. Nous nous attarderons plutôt sur la vision prospective.

- La **vision prospective**, une vision qui prend tout son sens dans le cadre de Solvabilité II étant donné que l'assureur a l'obligation de détenir un niveau de capital suffisant pour faire face aux fluctuations du CDR observable autour de 0. Ainsi, pour tout année de survenance i, cette erreur s'exprime par la formule 3.10.

$$MSEP_{C\hat{D}R_i(I+1)|D_I}(0) = \mathbb{E}[(C\hat{D}R_i(I+1) - 0)^2 \mid D_I]. \tag{3.10}$$

  Le modèle de Merz & Wüthrich fournit une estimation des deux premiers moments de $C\hat{D}R_i(I+1) \mid D_I$, ce qui permettra d'estimer l'erreur de prédiction pour chaque année de survenance, puis pour toutes les années agrégées.

3.4.4 Principe

A présent, nous allons nous attarder sur le modèle de Merz & Wüthrich, le modèle référence permettant de mesurer l'incertitude d'estimation des provisions pour sinistres à horizon d'un an. En effet, ce dernier propose une formule fermée qui permet d'évaluer le risque de réserve. Le modèle de Merz & Wüthrich est une adaptation du modèle de Mack ce qui implique qu'il est également une approche stochastique du modèle de Chain-Ladder. Ainsi, il paraît cohérent que l'estimation du *«Best Estimate»* est calculée à partir de la méthode de Chain-Ladder.

Pour rappel, les hypothèses du modèle :

- **H1 :** les montants cumulés d'années de survenance différentes sont des variables aléatoires indépendantes, c'est-à-dire que : $\forall i \neq k, C_{i,j} \perp C_{k,j}$.

- **H2 :** $(C_{i,j})_j$ est un processus de Markov et il existe $f_0, \ldots, f_{J-1} > 0$ et $\sigma_0^2, \ldots, \sigma_{J-1}^2 > 0$ tels que pour tout i ≥ 0 et j ≥ 1

$$\mathbb{E}[C_{i,j} \mid C_{i,j-1}] = f_{j-1} C_{i,j-1}.$$

$$VAR[C_{i,j} \mid C_{i,j-1}] = \sigma_{j-1}^2 C_{i,j-1}.$$

Ces hypothèses impliquent que

$$\mathbb{E}[C_{i,J} \mid D_I] = C_{i,I-i} \prod_{j=I-i}^{J-1} f_j \quad et \quad \mathbb{E}[C_{i,J} \mid D_{I+1}] = C_{i,I-i+1} \prod_{j=I-i+1}^{J-1} f_j.$$

Et temps $t = I$, les facteurs de développement sont estimés par la méthode de Chain-Ladder à travers la formule suivante

$$\hat{f}_j^I = \frac{\sum_{i=0}^{I-j-1} C_{i,j+1}}{\sum_{i=0}^{I-j-1} C_{i,j}}.$$

et en $t = I + 1$, nous avons :

$$\hat{f}_j^{I+1} = \frac{\sum_{i=0}^{I-j} C_{i,j+1}}{\sum_{i=0}^{I-j} C_{i,j}}.$$

Au final, nous obtenons les deux estimateurs suivants

$$\forall j \geq I - i, \quad \hat{C}_{i,j}^I = C_{i,I-i} \hat{f}_{I-i}^I \dots \hat{f}_{j-2}^I \hat{f}_{j-1}^I.$$

et

$$\forall j \geq I - i + 1, \quad \hat{C}_{i,j}^{I+1} = C_{i,I-i+1} \hat{f}_{I-i+1}^{I+1} \dots \hat{f}_{j-2}^{I+1} \hat{f}_{j-1}^{I+1}.$$

Après ces rappels, nous allons nous focaliser sur comment le modèle de Merz & Wütrich estime la déviation des CDR observables à la fin de l'année calendaire de 0 (la vision prospective).

Étant donné que $\mathbb{E}[C_{i,J} \mid D_I]$ est une martingale, nous pouvons en déduire que

$$\mathbb{E}[CDR_i(I+1) \mid D_I] = 0.$$

Ainsi :

$$\begin{aligned}
MSEP_{CDR_i(I+1)|D_I}(0) &= VAR(CDR_i(I+1) \mid D_I) \\
&= VAR(\mathbb{E}[C_{i,J} \mid D_I] - \mathbb{E}[C_{i,J} \mid D_{I+1}] \mid D_I) \\
&= VAR(\mathbb{E}[C_{i,J} \mid D_{I+1}] \mid D_I) \\
&= VAR(C_{i,I-i+1} \mid D_I) \prod_{j=I-i+1}^{J-1} f_j^2 \\
&= C_{i,I-i}.\sigma_{I-i}^2 \prod_{j=I-i+1}^{J-1} f_j^2 \\
&= \mathbb{E}[C_{i,J} \mid D_I]^2 \frac{\sigma_{I-i}^2 \mid f_{I-i}^2}{C_{i,I-i}}. \qquad (3.11)
\end{aligned}$$

Cette quantité correspond à la variance conditionnelle du CDR (conditionnelle à D_I). Cette variance peut donc être estimée par la formule 3.12.

$$\widehat{VAR}(CDR_i(I+1) \mid D_I) = (\hat{C}_{i,J}^{CL(I)})^2 \frac{\hat{\sigma}_{I-i}^2 \mid (\hat{f}_{I-i}^{CL(I)})^2}{C_{i,I-i}}. \qquad (3.12)$$

D'autre part, en introduisant le CDR observable $\widehat{CDR}_i(I+1)$, nous retrouvons sa variance autour de son meilleur estimateur de valeur nulle, ce qui implique que

$$VAR(\widehat{CDR}_i(I+1) \mid D_I) = MSEP_{\widehat{CDR}_i(I+1)|D_I}(0).$$

Le résultat à retenir est l'estimation de la qualité de prédiction pour chaque année de survenance i du $\widehat{CDR}_i(I+1)$ par 0, explicité ci-dessous

$$\widehat{MSEP}_{\widehat{CDR}_i(I+1)|D_I}(0) = (\hat{C}_{i,J}^{CL(I)})^2(\widehat{\Delta}_{i,J}^I + \widehat{\Gamma}_{i,J}^I). \tag{3.13}$$

où

$$\widehat{\Delta}_{i,J}^I = \frac{\hat{\sigma}_{I-i}^2 \mid (\hat{f}_{I-i}^{CL(I)})^2}{S_{I-i}^I} + \sum_{j=I-i+1}^{J-1} \left(\frac{C_{I-j,j}}{S_j^{I+1}}\right)^2 \frac{\hat{\sigma}_j^2 \mid (\hat{f}_j^{CL(I)})^2}{S_j^I}$$

$$\widehat{\Gamma}_{i,J}^I = \frac{\hat{\sigma}_{I-i}^2 \mid (\hat{f}_{I-i}^{CL(I)})^2}{C_{i,I-i}} + \sum_{j=I-i+1}^{J-1} \left(\frac{C_{I-j,j}}{S_j^{I+1}}\right)^2 \frac{\hat{\sigma}_j^2 \mid (\hat{f}_j^{CL(I)})^2}{C_{I-j,j}}$$

$$S_j^I = \sum_{i=0}^{I-j-1} C_{i,j} \quad et \quad S_j^{I+1} = \sum_{i=0}^{I-j} C_{i,j}.$$

NB : la formule de $\widehat{\Gamma}_{i,J}^I$ est une approximation de sa formule initiale dans un but de simplification. Pour d'avantage de détails, se référer à (Merz et Wüthrich, 2008).

Au final, nous avons pour tout année de survenance i,

$$\widehat{MSEP}_{\widehat{CDR}_i(I+1)|D_I}(0) = (\hat{C}_{i,J}^{CL(I)})^2\left(\frac{\hat{\sigma}_{I-i}^2 \mid (\hat{f}_{I-i}^{CL(I)})^2}{C_{i,I-i}} + \frac{\hat{\sigma}_{I-i}^2 \mid (\hat{f}_{I-i}^{CL(I)})^2}{S_{I-i}^I} + \sum_{j=I-i+1}^{J-1} \frac{C_{I-j,j}}{S_j^{I+1}} \frac{\hat{\sigma}_j^2 \mid (\hat{f}_j^{CL(I)})^2}{S_j^I}\right). \tag{3.14}$$

En définitive, dans le but de quantifier l'incertitude de prédiction à horizon un an dans sa totalité, Merz & Wüthrich fournissent une formule fermée estimant l'erreur quadratique moyenne de prédiction (MSEP) conditionnelle aux informations contenues dans D_I de la $\sum_i CDR_i(I+1)$. Nous pouvons noter dans la formule ci-dessous la prise en compte de corrélations entre les différentes années de survenance. En effet, ceci résulte du fait que les facteurs de développement estimés à partir de la méthode de Chain-Ladder sont appliqués sur les différentes années de survenance en ayant une base d'apprentissage commune, ce qui implique l'existence de ces corrélations.

$$\widehat{MSEP}_{\sum_{i=1}^I \widehat{CDR}_i(I+1)|D_I}(0) = \sum_{i=1}^I \widehat{MSEP}_{\widehat{CDR}_i(I+1)|D_I}(0) + 2\sum_{k>i>0} \widehat{C}_{i,J}^{CL(I)} \widehat{C}_{k,J}^{CL(I)} [\widehat{\Xi}_{i,J}^I + \widehat{\Lambda}_{i,J}^I]. \tag{3.15}$$

où

$$\widehat{\Xi}_{i,J}^I = \sum_{j=I-i+1}^{J-1} \left(\frac{C_{I-j,j}}{S_j^{I+1}}\right)^2 \frac{\hat{\sigma}_j^2 \mid (\hat{f}_j^{CL(I)})^2}{C_{I-j,j}} + \frac{(\widehat{\sigma}_{I-i})^2 \mid (\widehat{f}_{I-i}^{CL(I)})^2}{S_{I-i}^{I+1}}.$$

$$\widehat{\Lambda}_{i,J}^I = \frac{C_{i,I-i}}{S_{I-i}^{I+1}} \frac{\hat{\sigma}_{I-i}^2 \mid (\hat{f}_{I-i}^{CL(I)})^2}{S_{I-i}^I} + \sum_{j=I-i+1}^{J-1} \left(\frac{C_{I-j,j}}{S_j^{I+1}}\right)^2 \frac{\hat{\sigma}_j^2 \mid (\hat{f}_j^{CL(I)})^2}{S_j^I}.$$

Au final, nous avons le résultat suivant :

$$\widehat{MSEP}_{\sum_{i=1}^{I} \widehat{CDR}_i(I+1)|D_I}(0) = \sum_{i=1}^{I} \widehat{MSEP}_{\widehat{CDR}_i(I+1)|D_I}(0) + 2 \sum_{k>i>0} \widehat{C}_{i,J}^{CL(I)} \widehat{C}_{k,J}^{CL(I)} \left[\frac{\hat{\sigma}_{I-i}^2 \mid (\hat{f}_{I-i}^{CL(I)})^2}{S_{I-i}^I} + \sum_{j=I-i+1}^{J-1} \frac{C_{I-j,j}}{S_j^{I+1}} \frac{\hat{\sigma}_j^2 \mid (\hat{f}_j^{CL(I)})^2}{S_j^I} \right]. \quad (3.16)$$

3.4.5 Comparaison avec la formule de Mack

Dans cette section, nous allons comparer la formule permettant d'évaluer l'incertitude d'estimation des provisions à l'ultime que propose le modèle Mack à celle à horizon d'un an proposée par le modèle de Merz & Wüthrich. Nous commençons par analyser les deux formules proposées pour une année de survenance i, nous pouvons ainsi les décomposer de la façon suivante

$$\widehat{MSEP}_{|D_I}(\hat{R}_i^{CL(I)}) = (\hat{C}_{i,J}^{CL(I)})^2 \sum_{j=I-i}^{J-1} \left(\underbrace{\frac{\hat{\sigma}_j^2 \mid (\hat{f}_j^{CL(I)})^2}{\sum_{l=0}^{I-j-1} C_{l,j}}}_{Erreur\ d'estimation} + \underbrace{\frac{\hat{\sigma}_j^2 \mid (\hat{f}_j^{CL(I)})^2}{\hat{C}_{i,J}^{CL(I)}}}_{Erreur\ de\ processus} \right). \quad (3.17)$$

$$\widehat{MSEP}_{\widehat{CDR}_i(I+1)|D_I}(0) = (\hat{C}_{i,J}^{CL(I)})^2 \left(\underbrace{\frac{\hat{\sigma}_{I-i}^2 \mid (\hat{f}_{I-i}^{CL(I)})^2}{C_{i,I-i}} + \frac{\hat{\sigma}_{I-i}^2 \mid (\hat{f}_{I-i}^{CL(I)})^2}{S_{I-i}^I}}_{Erreurs\ d'estimation\ et\ de\ processus\ au\ niveau\ de\ la\ 1^{\text{er}}e\ diagonale} + \underbrace{\sum_{j=I-i+1}^{J-1} \frac{C_{I-j,j}}{S_j^{I+1}} \frac{\hat{\sigma}_j^2 \mid (\hat{f}_j^{CL(I)})^2}{S_j^I}}_{Erreur\ d'estimation\ sur\ les\ diagonales\ suivant\ la\ 1^{\text{er}}e\ diagonale} \right) \quad (3.18)$$

A partir des décompositions présentées ci-dessus, il est intéressant de souligner que les deux premiers termes de la formule de Merz & Wüthrich réfèrent aux erreurs d'estimation et de processus au niveau de la première diagonale. L'estimation de ces erreurs est analogue à celle fournie par le modèle de Mack. Quant au dernier terme de la formule de Merz & Wüthrich, il porte uniquement sur l'erreur d'estimation des diagonales qui suivent, contrairement à la formule de Mack qui tient compte des deux erreurs à la fois.

Nous pouvons de plus noter que l'estimation de l'erreur de prédiction au niveau de la formule de Merz & Wüthrich est principalement concentrée au niveau de la première diagonale étant donné que le troisième terme de cette formule reste négligeable (le rapport $\frac{C_{I-j,j}}{S_j^{I+1}} \leq 1,\ \forall j$).

En ce qui concerne les formules pour toutes les années de survenance agrégées, nous restons sur des constats similaires. Nous pourrons ainsi retenir que l'estimation de la MSEP du « CDR observable »

fournie par Merz & Wüthrich tient principalement compte des erreurs d'estimation et de processus au niveau de la première diagonale.

3.4.6 Avantages et inconvénients

Il faut tout d'abord commencer par souligner que ce modèle se basant sur le modèle Mack, hérite de ses avantages et inconvénients. Ce qui implique notamment la nécessité que la triangle utilisé comme base d'apprentissage respecte les hypothèses émises par ce modèle.

Par ailleurs, un avantage majeur du modèle de Merz & Wüthrich repose sur sa capacité à évaluer l'incertitude des provisions estimées à horizon d'un an à partir d'une formule fermée. Ceci permet une rapidité de calcul qui constitue un atout majeur dans le cadre de la directive «Solvabilité II» car cette dernière nécessite des remontées de résultats fréquentes.

En revanche, ce modèle ne prend pas en compte les queues de développement. Il est donc primordial de n'appliquer la formule qu'une fois la première année de survenance close, à savoir lorsqu'il n'y a plus de sinistre significatif rattaché à cette première année de survenance. Cette contrainte peut être problématique dans le cas de branches à développement long.

Pour finir, cette méthode ne permet pas de recueillir la distribution complète car seuls les deux premiers moments de la distribution des provisions sont obtenues à l'aide d'une formule fermée.

3.5 Application numérique

Pour commencer, il est nécessaire de préciser que tout au long des applications numériques que nous effectuerons, nous ne tiendrons pas compte des tardifs aussi bien à travers les méthodes agrégées que les modèles de *machine learning*. Nous nous attacherons ainsi à l'estimation des provisions *IBNER* et des erreurs de prédictions qui y sont rattachées. A noter également que toutes les applications numériques sont effectuées à l'aide du langage (R Core Team, 2022).

Afin d'appliquer les méthodes de provisionnement présentées dans ce chapitre, nous disposons de différents triangles agrégés se rapportant à la garantie RC corporelle automobile du portefeuille AXA France des particuliers et professionnels et sur l'ensemble des réseaux de distribution, en l'occurrence sur les réseaux des agents, des courtiers et des salariés. Pour rappel, les triangles portent sur la charge D/D nette de recours et concernent les survenances 1999 à 2014 le long de 16 années de développement.

Avant d'entamer cette phase d'application, nous nous attardons sur le choix du triangle de projection : un triangle net ou brut de recours, un triangle de règlements cumulés ou plutôt de charges D/D cumulées ? Suite à une analyse des différents triangles disponibles, nous notons qu'il est préférable de traiter les charges D/D nettes de recours comme illustré sur la figure 3.4. Ce choix se justifie principalement par le caractère long de cette branche et donc une possibilité d'étalement des règlements dans le temps notamment pour les sinistres graves et les indemnisations sous forme de rente.

Année de survenance	Année de développement															
	N	N+1	N+2	N+3	N+4	N+5	N+6	N+7	N+8	N+9	N+10	N+11	N+12	N+13	N+14	N+15
1999	192 856	224 945	236 012	234 018	236 972	233 681	234 580	233 296	236 041	242 232	244 414	243 278	242 778	242 742	242 718	243 459
2000	208 199	239 351	228 910	230 239	238 233	240 909	244 563	252 175	253 201	259 973	262 146	262 240	263 734	267 319	269 353	
2001	167 064	204 045	206 805	220 592	228 388	227 387	228 935	232 400	233 824	235 603	236 499	240 005	240 490	240 144		
2002	173 857	200 102	212 973	220 252	228 305	225 209	224 394	230 204	233 345	238 153	243 481	247 475	246 324			
2003	157 265	178 667	197 149	205 298	208 905	204 011	203 816	205 351	204 462	204 945	205 589	213 722				
2004	165 265	186 579	197 963	206 777	205 878	212 029	216 468	217 287	221 708	219 440	218 235					
2005	157 866	186 458	204 554	206 893	209 587	211 471	213 321	217 488	218 963	213 678						
2006	131 104	176 759	191 207	197 163	206 145	209 990	221 402	220 342	218 512							
2007	129 976	173 707	197 684	198 575	201 189	209 188	212 186	212 747								
2008	127 304	175 425	190 442	202 836	213 396	217 006	203 748									
2009	131 139	196 596	219 166	244 337	245 179	241 903										
2010	116 593	187 198	221 301	231 415	219 247											
2011	133 207	223 538	247 837	264 037												
2012	162 195	211 839	208 702													
2013	209 385	214 669														
2014	230 890															

Figure 3.4 – Triangle de charges D/D cumulées net de recours (en k€) - vision fin 12/2014

Afin d'homogénéiser les populations de sinistres et ainsi satisfaire au mieux les hypothèses émises par les méthodes de Chain-Ladder et de Mack, Axa France estime sa charge ultime à partir de trois triangles distincts : attritionnels, graves et très graves (cf. annexe A.7).

Ainsi l'estimation de la charge ultime, détaillée par ailleurs dans le mémoire de (Servel, 2020), peut se résumer comme suit :

- **Triangle 12 (charges D/D $\leq 150k$€) :** application de la méthode de Chain-Ladder.
- **Triangle 3 (charges D/D $\in$] 150k€ ; 750k€]) :** application de la méthode de Chain-Ladder.
- **Triangle 45 (charges D/D $> 750k$€) :** application de la méthode de Mack dans un premier temps, pour pouvoir dans un second temps retenir le quantile à 75% d'une distribution de provisions supposée log-normale d'espérance $\mu = ln(\hat{R}) - \frac{\sigma^2}{2}$ et de variance $\sigma^2 = ln(1 + \frac{\hat{MSEP}(\hat{R})}{\hat{R}^2})$. Cette démarche revient à adopter une marge de prudence supplémentaire à la charge estimée par la méthode de Chain-Ladder. Ceci paraît nécessaire du fait du développement particulièrement long de cette dernière catégorie de sinistres qui peut provenir de leur atypisme, leur gravité, l'es-

timation de la durée de vie s'il s'agit d'une rente ou encore l'attente d'une décision judiciaire.

Il faut toutefois noter que l'estimation de la charge lors de la première année d'inventaire ne repose pas sur la méthode décrite ci-dessus mais plutôt sur une méthode d'ajustement fréquence/coût moyen. L'objectif de cette démarche est de parer aux fluctuations de la charge D/D observées la première année et qui peuvent fortement varier d'une survenance à une autre.

En effet, l'analyse du triangle global (cf. figure A.6) et des triangles des différentes tranches de coûts figurant en annexe A.7 permet d'effectuer deux constats.

Tout d'abord, il est intéressant de noter que l'écart de la charge D/D totale entre l'année N et N+1 est beaucoup moins important en 2014 que lors des années précédentes. Ceci justifie donc bien le recours à la méthode d'ajustement à la première date d'inventaire. Cette variation peut se justifier par les importants efforts, matériels et humains, consentis par AXA France pour améliorer la précision des provisions réalisées et plus particulièrement sur les sinistres graves et très graves.

Par ailleurs au même moment, sur le triangle des attritionnels, il est possible d'observer une baisse plus prononcée de la charge D/D entre l'année N et N+1 en 2014 en comparaison aux années précédentes. Ceci s'explique principalement par un rythme plus important de clôture impliquant une accélération des évaluations de règlements nulles.

3.5.1 Estimation de la charge ultime pour la survenance 2014

En appliquant le modèle AXA France décrit ci-dessous, nous allons estimer la charge ultime pour l'année de survenance 2014. En disposant dans un premier temps uniquement de la vision à fin décembre 2014, nous commençons par appliquer la méthode d'ajustement sur un historique de cinq ans (survenances 2009 à 2013). Pour une année de survenance donnée, la charge ultime estimée par la méthode par tranche de coûts est projetée à 2014 en tenant compte des évolutions de fréquences et en actualisant selon une hypothèse d'évolution de coût moyen (+5% par an dans ce cas de figure). La charge ultime de la survenance 2014 obtenue est de **304 M€**, moyenne des cinq charges projetées. Le tableau 3.1 récapitule les calculs et hypothèses retenues.

	Année de survenance					
	2009	**2010**	**2011**	**2012**	**2013**	**2014**
Charge ultime*	272 754	254 193	333 010	262 408	293 067	
Evolution de fréquence	*3,8%*	*-4,3%*	*-6,4%*	*-5,3%*	*-3,1%*	*5,2%*
Evolution du coût moyen	*5%*	*5%*	*5%*	*5%*	*5%*	*5%*
Charge ultime - surv. 2014	**312 675**	**267 266**	**348 565**	**279 467**	**313 846**	**304 364**

**Estimation par la méthode par tranches de coûts*

TABLE 3.1 – Méthode d'ajustement : estimation de la charge ultime pour la survenance 2014 - vision à fin 12/2014 (montants en k€)

A présent, en se plaçant à fin décembre 2015, autrement dit au niveau de la deuxième année d'inventaire, la charge ultime est estimée à partir de la méthode par tranche de coûts. Nous obtenons une charge totale de **315 M€**, soit un montant de réserves à détenir de **85 M€**. Il faut également mentionner que la marge de prudence adoptée au niveau de l'estimation de la charge ultime des très graves a impliqué l'ajout d'une provision de **14 M€**. Le tableau 3.2 fournit une vision détaillée des résultats obtenus selon

les différents triangles.

Survenance 2014 - Vision fin 12/2015	Charge D/D	Charge ultime	Charge ultime avec TF*	Réserves	Réserves avec TF*
T12	104 725	95 862	95 862	-8 863	-8 863
T3	35 259	42 789	42 789	7 530	7 530
T45	90 579	161 869	176 781	71 290	86 202
Total	**230 563**	**300 520**	**315 432**	**69 957**	**84 869**

** Tail factor*

TABLE 3.2 – Méthode par tranche de coûts : estimation de la charge ultime pour la survenance 2014 - vision à fin 12/2015 (montants en k€)

Vérification des hypothèses du modèle

Lors de l'application d'une des méthodes de Chain-Ladder, Mack, Merz & Wüthrich ou encore toute autre méthode faisant appel à des hypothèses, il est nécessaire de s'interroger sur la validité de ces dernières au niveau des données traitées et de les vérifier. Dans notre cas de figure, nous allons vérifier les hypothèses retenues par le modèle de Mack pour les trois triangles présents en annexe A.7 (les triangles T12, T3 et T45). Pour rappel, ces hypothèses sont identiques à celles de Merz & Wüthrich et incluent celles de Chain-Ladder. La section 3.3.3 décrit l'exhaustivité des tests et graphiques présentées ci-dessous.

Vérification de l'hypothèse H1

	T12	T3	T45
Statistique de test Z	40	51	41
IC 95%	[40.7 , 53.8]	[41.2 , 54.6]	[40.6 , 54.2]

TABLE 3.3 – Résultats de tests de non corrélation entre les années de survenance au niveau des trois triangles

Nous constatons sur la figure 3.3 que globalement les statistiques de test sont incluses dans les intervalles de confiances, les hypothèses de non corrélation ne sont donc pas rejetées.

Vérification de l'hypothèse H2

	T12	T3	T45
Statistique de test T	0.17	-0.11	0.26
IC 50%	[-0.07 , 0.07]	[-0.07 , 0.07]	[-0.07 , 0.07]

TABLE 3.4 – Résultats de tests des effets calendaires au niveau des trois triangles

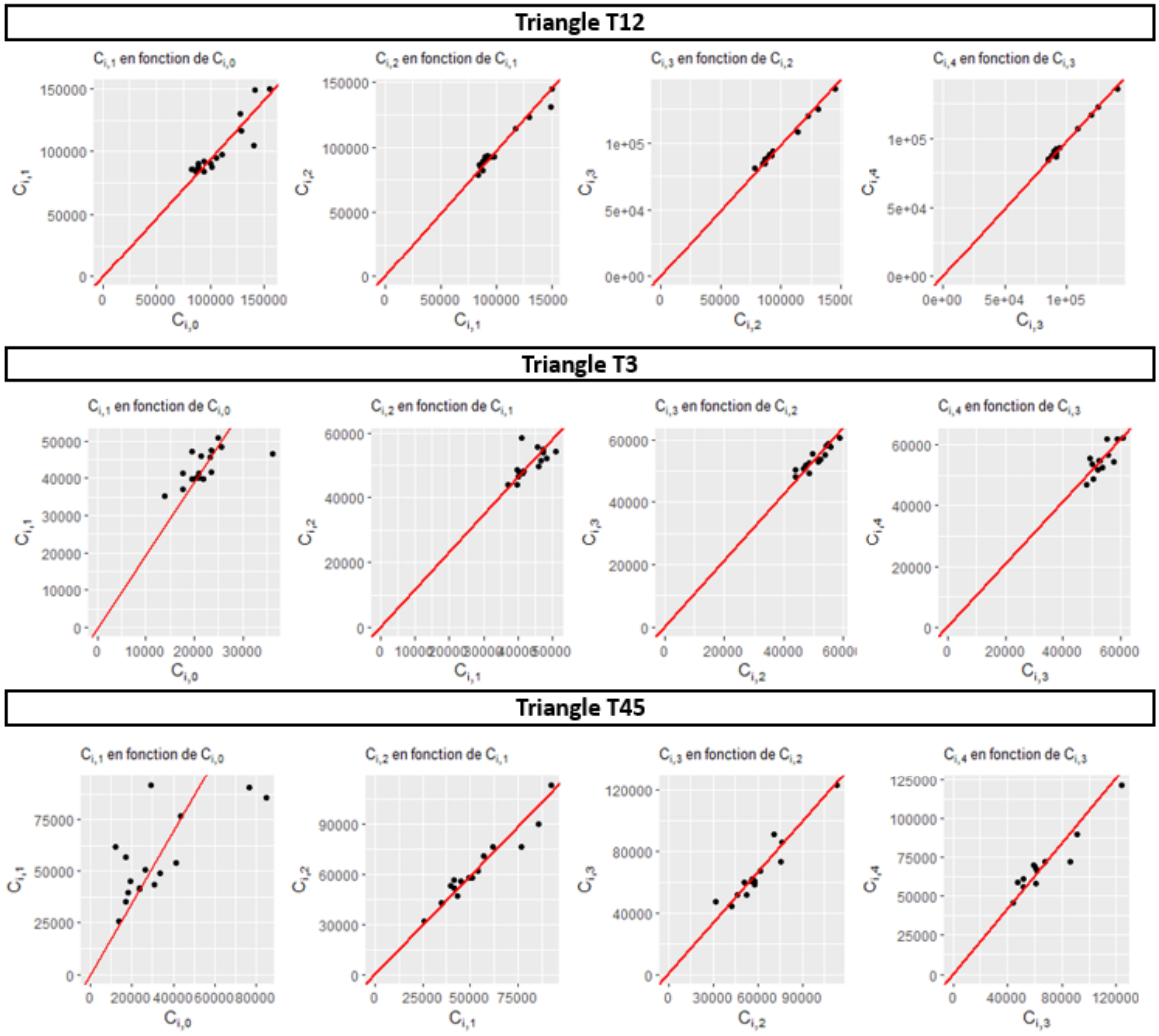


FIGURE 3.5 – Représentations des $C_{i,j+1}$ en fonction des $C_{i,j}$ pour les quatre premières années de développement - Triangles de charges D/D cumulées nettes de recours

Les résultats de tests des effets calendaires présentés en figure 3.4 impliquent le rejet des hypothèses d'absence d'effets calendaires étant donné que les statistiques de test sont en dehors de l'intervalle de confiance de 50%. D'autre part, la visualisation graphique des couples ($C_{i,j}$, $C_{i,j+1}$) illustrée en figure 3.5 montre un alignement des points obtenus sur la droite (droite passant par l'origine dont le la pente est égale au facteur de développement estimé) à partir de la deuxième année de développement notamment pour le triangles des attritionnels. Néanmoins, cet alignement reste imparfait et inexistant pour la première année de développement ce qui justifie bien, encore une fois, le recours à la méthode « as-if » lors de la première année de développement.

Vérification de l'hypothèse H3

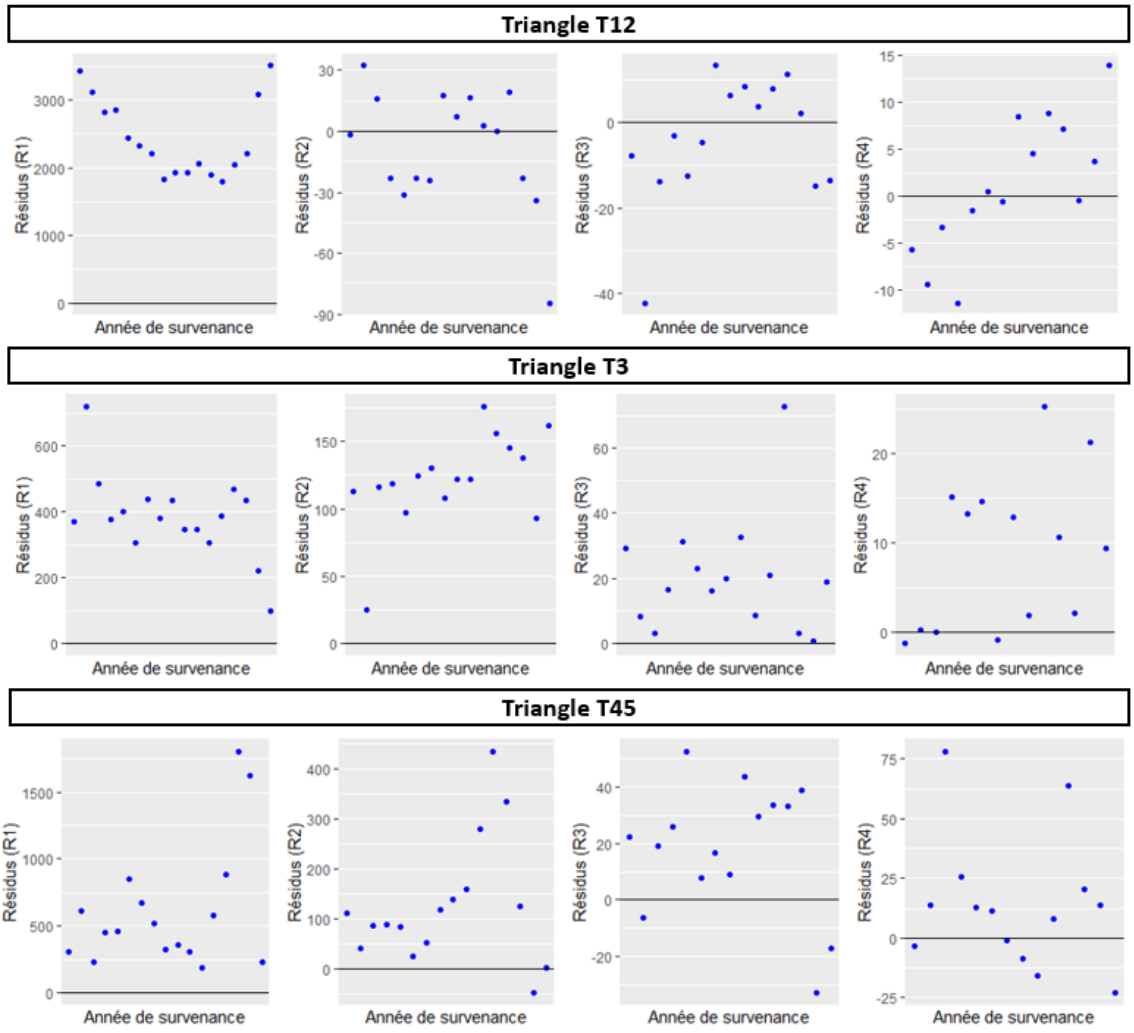


FIGURE 3.6 – Représentations des résidus standardisés pour les quatre premières années de développement - Triangles de charges D/D cumulées nettes de recours

Les graphiques de résidus standardisés présentés en figure 3.6 pour les quatre premières années de développement révèlent un comportement aléatoire de ces derniers au niveau de certaines années et de certains triangles uniquement. Ainsi, ce comportement aléatoire ne se révèle pas unanime. Nous ne pouvons pas donc affirmer que l'hypothèse H3 est satisfaite.

Pour récapituler, la mise en oeuvre des tests des hypothèses du modèle de Mack montre qu'aucun des trois triangles contribuant à l'estimation de la charge ultime totale ne satisfait pleinement les hypothèses sous-jacentes à ce modèle, notamment les hypothèses H2 et H3. Ceci pourrait remettre en question la validité des résultats obtenus, néanmoins nous retenons tout de même les estimations obtenues et les prenons pour références étant donné qu'il s'agit du modèle utilisé par AXA France et que nous souhaitons challenger les résultats de ce dernier.

3.5.2 Estimations des erreurs de prédiction à l'ultime et à horizon un an

Survenance 2014 - Vision fin 12/2015	Charge D/D	Charge ultime	Charge ultime avec TF*	Réserves	Réserves avec TF*	MSEP Mack	MSEP M&W *(observables approchées)*
T12	104 725	95 862	95 862	-8 863	-8 863	5 999	4 442
T3	35 259	42 789	42 789	7 530	7 530	5 680	3 707
T45	90 579	161 869	176 781	71 290	86 202	29 735	12 291
Total	**230 563**	**300 520**	**315 432**	**69 957**	**84 869**		

**Tail factor*

TABLE 3.5 – MSEP de Mack (à l'ultime) versus M&W (à horizon un an) des différents triangles pour l'estimation de la charge ultime de la survenance 2014 - vision à fin 12/2015 (montants en k€)

La table 3.5 permet de visualiser, outre les estimations de la charge ultime et celles des réserves pour la survenance de 2014, les estimations des erreurs de prédiction aussi bien à l'ultime obtenues à travers la méthode de Mack qu'à horizon un an obtenues à travers la méthode de Merz & Wüthrich. Ces estimations sont obtenues pour chaque triangle. Dans notre cas de figure, la variance totale est composée d'une variance intra-triangles et d'une variance inter-triangles. Nous tiendrons compte uniquement de la variance intra-triangles qui est un minorant de la variance totale. Ainsi, la somme des erreurs de prédiction estimées sera retenue comme référence de comparaison par rapport aux erreurs obtenues par les modèles de *machine learning* introduits ultérieurement. L'erreur de prédiction à l'ultime au niveau du triangle totale est estimée à **41 M€** versus **20 M€** à horizon un an.

Année de survenance	Triangle T12			Triangle T3			Triangle T45		
	MSEP Mack	MSEP M&W (observables approchées)	MSEP M&W (observables exactes)	MSEP Mack	MSEP M&W (observables approchées)	MSEP M&W (observables exactes)	MSEP Mack	MSEP M&W (observables approchées)	MSEP M&W (observables exactes)
1999	-	-	-	-	-	-	-	-	-
2000	148	148	153	1 118	1 118	1 272	1	1	1
2001	267	236	236	1 561	1 241	1 241	21	21	21
2002	276	143	143	1 672	1 010	1 010	1 423	1 423	1 423
2003	371	287	287	1 805	872	872	2 142	1 746	1 746
2004	417	202	202	1 827	826	826	3 639	2 951	2 951
2005	447	167	167	2 373	1 366	1 366	5 811	4 838	4 838
2006	483	228	228	2 262	960	960	7 197	3 289	3 289
2007	741	562	562	2 597	1 304	1 304	8 613	5 355	5 355
2008	985	647	647	2 669	936	936	7 960	2 219	2 219
2009	1 576	1 223	1 223	3 136	1 420	1 420	10 950	4 386	4 387
2010	1 783	990	990	3 379	1 368	1 368	13 036	8 204	8 205
2011	2 594	1 919	1 919	3 337	1 282	1 282	19 074	7 617	7 618
2012	2 972	1 599	1 599	4 637	3 221	3 222	18 152	10 827	10 828
2013	3 610	2 065	2 065	4 760	2 123	2 123	24 718	12 590	12 592
2014	5 999	4 442	4 442	5 680	3 707	3 708	29 735	12 291	12 293
2015	13 209	11 027	11 028	6 332	5 071	5 071	34 930	31 785	31 792
Total	**16 802**	**13 091**	**13 092**	**21 875**	**14 278**	**14 291**	**78 409**	**48 973**	**48 981**

TABLE 3.6 – MSEP de Mack (à l'ultime) versus M&W (à horizon un an) des différents triangles toutes survenances - vision à fin 12/2015 (montants en k€)

Cette deuxième vision des estimations présentée en table 3.6 des erreurs de prédiction à l'ultime (obtenues à partir des formules de Mack) versus celles à horizon un an (obtenues à partir des formules proposées par Merz & Wüthrich) pour chaque année de survenance puis en vision agrégée est intéressante car elle met en évidence l'intérêt d'adopter une vision sur le court terme, une vision sur laquelle repose l'exigence en capital imposé par la directive Solvabilité II. En effet, les estimations des erreurs à horizon un représentent en moyenne 2/3 des estimations des erreurs à l'ultime : une part de 78% au niveau du

triangle des attritionnels, 65% au niveau du triangle des graves et 62% au niveau du triangle des très graves. Ce constat est en phase avec les conclusions de l'étude menée par une AISAM-ACME [4] auprès de différentes compagnies (AISAM-ACME, 2007).

Un second constat peut être établi celui de la croissance des estimations des erreurs de prédiction en fonction de la récence de l'année de survenance aussi bien à l'ultime qu'à horizon un an. Ce constat est tout à fait cohérent, plus l'année de survenance est récente donc disposant de peu d'années de développement, plus cette dernière portera de l'incertitude sur son développement et sera source d'un capital plus important à détenir comparé aux années de survenances antérieures, c'est notamment le cas au niveau de la dernière année de survenance (survenance 2015).

3.5.3 Estimations de la charge ultime, des erreurs de prédiction à l'ultime et à horizon un an pour la survenance 2014 à différentes dates d'inventaire

	Vision à fin 12/2015					Vision à fin 12/2016					Vision à fin 12/2017				
Survenance 2014	Charge D/D	Charge ultime	Charge ultime avec TF*	Réserves	Réserves avec TF*	Charge D/D	Charge ultime	Charge ultime avec TF*	Réserves	Réserves avec TF*	Charge D/D	Charge ultime	Charge ultime avec TF*	Réserves	Réserves avec TF*
T12	104 725	95 862	95 862	-8 863	-8 863	90 513	85 321	85 321	-5 192	-5 192	89 706	85 774	85 774	-3 932	-3 932
T3	35 259	42 789	42 789	7 530	7 530	40 590	44 150	44 150	3 560	3 560	51 538	53 242	53 242	1 704	1 704
T45	90 579	161 869	176 781	71 290	86 202	109 534	173 492	187 332	63 958	77 798	107 557	150 156	160 103	42 599	52 546
Total	**230 563**	**300 520**	**315 432**	**69 957**	**84 869**	**240 637**	**302 963**	**316 803**	**62 326**	**76 166**	**248 801**	**289 172**	**299 119**	**40 371**	**50 318**

*Tail factor

	Vision à fin 12/2018					Vision à fin 12/2019				
Survenance 2014	Charge D/D	Charge ultime	Charge ultime avec TF*	Réserves	Réserves avec TF*	Charge D/D	Charge ultime	Charge ultime avec TF*	Réserves	Réserves avec TF*
T12	89 638	87 257	87 257	-2 381	-2 381	90 189	89 290	89 290	-899	-899
T3	59 263	61 705	61 705	2 442	2 442	60 296	60 867	60 867	571	571
T45	113 623	154 497	163 683	40 874	50 060	121 825	168 551	178 304	46 726	56 479
Total	**262 524**	**303 459**	**312 645**	**40 935**	**50 121**	**272 310**	**318 708**	**328 461**	**46 398**	**56 151**

*Tail factor

	Vision à fin 12/2015		Vision à fin 12/2016		Vision à fin 12/2017		Vision à fin 12/2018		Vision à fin 12/2019	
Survenance 2014	MSEP Mack	MSEP M&W (observables approchées)	MSEP Mack	MSEP M&W (observables approchées)	MSEP Mack	MSEP M&W (observables approchées)	MSEP Mack	MSEP M&W (observables approchées)	MSEP Mack	MSEP M&W (observables approchées)
T12	5 999	4 442	3 671	2 087	3 044	1 725	2 674	2 001	1 772	1 069
T3	5 680	3 707	4 631	2 300	4 402	2 933	4 310	2 011	3 606	1 373
T45	29 735	12 291	28 321	14 483	22 030	12 493	19 462	8 237	19 422	10 115

TABLE 3.7 – Estimations de la charge ultime par la méthode par tranche de coûts, estimations des erreurs de prédictions à l'ultime (méthode de Mack) et à horizon un an (méthode de M&W) pour la survenance 2014 à partir des trois triangles à différentes dates d'inventaire (montants en k€)

Cette dernière partie recense les résultats obtenus et présentés en table 3.7 composés de l'estimation de la charge ultime pour la survenance 2014 ainsi que des estimations des erreurs de prédiction à l'ultime et à horizon un an pour la date d'inventaire à fin 12/2015 mais également pour les exercices suivants jusqu'à fin 12/2019. Un des objectifs de cet exercice est d'observer l'évolution de ces estimations au fur et à mesure du développement de cette survenance aussi bien avec inclusion d'une marge de prudence au niveau des sinistres très graves que sans.

Il est tout d'abord possible de noter les fluctuations au niveau de l'estimation de la charge ultime notamment à fin 12/2017 pour un montant de **289 M€** versus **319 M€** à fin 12/2019 (299 M€ vs. 328 M€ en incluant une marge de prudence), fluctuations principalement observées au niveau des sinistres dont la charge D/D est supérieure à 150k€. Ceci est également le reflet du fait que les triangles

4. AISAM-ACME : Association Internationale des Sociétés d'Assurance Mutuelle - *Association of European Cooperative and Mutual Insurers*

(notamment T3 et T45) ne satisfaisaient pas pleinement les hypothèses émises par le modèle de Mack.

Un second constat, préétabli, concerne la décroissance des estimations des erreurs de prédiction au fur et à mesure du développement de la survenance avec une part d'incertitude principalement concentrée au niveau des sinistres très graves dont la charge D/D excède 750k€. Le long développement de ces sinistres est souvent dû à l'attente de stabilisation de l'état d'une ou de plusieurs victimes pour pouvoir statuer sur le coût définitif du sinistre ou encore des sinistres en contentieux sujets à de longs procès judiciaires ou encore l'indemnisation sous forme de rentes.

4 Provisionnement ligne à ligne

Après avoir introduit les fondements théoriques des méthodes agrégées ainsi que l'application du modèle AXA France, nous pouvons dès lors nous attacher à estimer la charge ultime et l'erreur de prédiction associée au travers de méthodes de *machine learning* appliquées à la base de données individuelles présentée en section 2.3.

Avant toute chose, il est nécessaire de s'attarder sur une particularité de cette base de données, à savoir la présence de censure à droite. Il s'agira donc de définir cette notion de censure avant d'exposer les problématiques mathématiques sous-jacentes pour enfin identifier une méthode permettant de corriger le biais induit par cette dernière. En l'occurrence dans notre cas, nous utiliserons la méthode dite *IPCW* (*inverse-probability-of-censoring weighting*).

Enfin, après avoir présenté les modèles de *machine learning* et leur mise en oeuvre, nous les comparerons sur la base de leurs résultats, dans un premier temps entre eux puis avec le modèle AXA France.

4.1 Apprentissage à partir de données censurées

Le phénomène de censure des données affecte une large variété de problématiques, dont celle du provisionnement de sinistres à développement long (Lopez et al., 2019) et (Lopez et al., 2016) que nous traitons dans le cadre de ce mémoire.

Nous pouvons entre autres citer, sans être exhaustifs, les problématiques liées à la construction de tables de mortalité (Guibert et Planchet, 2017), à la mesure du risque de crédit au travers des durées d'attente avant la survenance d'une défaillance de paiement (Hainaut et Robert, 2014) ou encore à l'étude de résiliation des contrats d'assurance (Milhaud, 2013).

4.1.1 Schéma mathématique d'observation des données

Après avoir illustré en section 2.3.4 la notion de censure à droite à laquelle nous sommes confrontés, nous allons désormais présenter le problème mathématique sous-jacent tel qu'exposé par (Lopez et al., 2016).

Considérons le vecteur aléatoire (M, T, X) où :
- $M \in \mathbb{R}$, correspond au montant de charge ultime nette de recours connu à la clôture du sinistre ;
- $T \in \mathbb{R}^+$, correspond à la durée de vie du sinistre, à savoir la durée séparant la date d'ouverture du sinistre de sa date de clôture ;

— $X \in \mathbb{R}^d$, correspond aux variables explicatives pouvant impacter la variable M ou la variable T ou les deux à la fois.

Introduisons à présent la variable de censure $C \in \mathbb{R}^+$ telle que celle-ci correspond à la durée séparant la date d'ouverture du sinistre de la date de fin d'observation de ce dernier. Dans notre cas de figure, l'arrêt d'observation du sinistre peut avoir lieu en raison de sa clôture ou de son annulation mais également en cas de dépassement de la date de fin d'observation de la base de données, en l'occurrence ici au 31 décembre 2019.
De part la présence de cette censure à droite, les variables (M, T) ne sont pas toujours observées directement.

Nous supposons que les variables T et C sont continues et nous considérons que les variables explicatives X sont pleinement observées, c'est-à-dire non sujettes à la censure. Ainsi, notre base de données individuelles peut être modélisée par un set de n réalisations indépendantes et identiquement distribuées $(N_i, Y_i, \delta_i, X_i)_{1 \leq i \leq n}$, telles que pour tout $i \in \{1, \ldots, n\}$:

$$\begin{aligned} Y_i &= min(T_i, C_i) \\ \delta_i &= \mathbb{1}_{T_i \leq C_i} \\ N_i &= \delta_i M_i \\ X_i &= (X_i^{(1)}, \ldots, X_i^{(d)}) \end{aligned}$$

Comme mentionné par (Lopez et al., 2016), sous les deux hypothèses (1) et (2)

1. C et (M, T) sont indépendants
2. $\mathbb{P}(T \leq C \mid M, T, X) = \mathbb{P}(T \leq C \mid T)$

nous avons pour toute fonction $\psi \in L^1$ la relation (4.1).

$$\mathbb{E}\Big[\frac{\delta.\psi(N, Y, X)}{S_C(Y^-)}\Big] = \mathbb{E}[\psi(M, T, X)] \tag{4.1}$$

où $S_C(t) = \mathbb{P}(C > t)$. La fonction de survie S_C, souvent inconnue, peut être estimée, en émettant en plus des hypothèses (1) et (2) deux hypothèses supplémentaires d'indépendance entre T et C et que $\mathbb{P}(T = C) = 0$ pour les variables continues, par l'estimateur consistant de Kaplan-Meier (Kaplan et Meier, 1958) défini en 4.1.2.

4.1.2 L'estimateur de Kaplan-Meier et la méthode IPCW

Cette partie s'inscrit dans la continuité de la partie précédente avec comme objectif d'introduire l'estimateur de Kaplan-Meier en présence de censure à droite. Il s'agit d'un estimateur non paramétrique. Ayant également pour appellation l'estimateur « produit-limite », il est considéré comme étant l'outil statistique de base utilisé pour estimer la fonction de survie d'une variable censurée à droite T.

Ainsi, l'estimateur de Kaplan-Meier de la fonction de survie T (en absence d'*ex-aequo*) est calculé à partir de la formule (4.2).

$$\hat{S_T}(t) = \prod_{Y_i \leq t} \left(1 - \frac{\delta_i}{\sum_{j=1}^{n} \mathbb{1}_{Y_i \leq Y_j}}\right). \tag{4.2}$$

En s'attardant sur la formule (4.2), nous pouvons constater que la fonction $\hat{S}_T$ est une fonction décroissante et constante par morceaux où le nombre de sauts correspond au nombre d'observations non

censurées (où $\delta_i = 1$).

Il est à noter qu'en absence de censure, l'estimateur de Kaplan-Meier d'une fonction de survie se résume à la fonction de survie empirique.

Après avoir brièvement introduit l'estimateur de Kaplan-Meier en présence de censure à droite, nous allons dès à présent nous attarder sur l'interprétation de cet estimateur afin d'établir le lien avec la notion des poids *IPCW*.

Ainsi comme démontré au niveau de (Le Faou, 2019), l'estimateur de Kaplan-Meier peut être exprimé à travers la formule (4.3).

$$\hat{S_T}(t) = \frac{1}{n}\sum_{i=1}^{n} \frac{\delta_i}{\hat{S}_C(Y_i)}.\mathbb{1}_{Y_i>t}. \tag{4.3}$$

En faisant le parallèle entre l'estimateur de Kaplan-Meier et la fonction de répartition empirique, nous pouvons dire qu'en absence de censure, chaque observation se voit attribuer la pondération $\frac{1}{n}$ alors qu'en présence de censure, le poids attribué à chaque observation a pour valeur $\hat{w}_i = \frac{\delta_i}{n.\hat{S}_C(Y_i)}$. De part la formule de $\hat{w}_i$, nous remarquons que les observations censurées se voient attribuer un poids nul. Ceci n'implique pas pour autant que ces dernières sont ignorées compte tenu de leur contribution dans l'estimation de la fonction de survie S_C et donc implicitement leur contribution à l'estimation des poids w_i.

Il est également important de noter qu'étant donné que l'estimateur $\hat{S}_C$ est une fonction décroissante, les poids $\hat{w}_i$ sont croissants en fonction de Y_i. Ainsi, plus l'observation Y_i a une valeur importante, plus le poids qu'il lui est attribué est important. Ce phénomène permet de compenser l'effet de la censure qui a pour principalement pour conséquence dans notre cas de figure de limiter les sinistres dont la charge nette est importante en raison de la corrélation positive existant entre la durée de vie du sinistre T et sa charge ultime M.

Ainsi, la description des poids introduits par l'estimateur de Kaplan-Meier combinée à la formule (4.1) permettent de justifier que le recours à ces poids, également appelés poids *IPCW*, et aux observations $(N_i, Y_i, \delta_i, X_i)_{1\leq i\leq n}$ permet d'estimer la charge ultime $(M_i)_{1\leq i\leq n}$.

Fort de ces constats, l'approche que nous souhaitons mettre en place pour corriger le biais introduit par la censure, similaire à l'approche présentée par (Lopez et al., 2016), repose sur les poids *IPCW (Inverse Probability of Censoring Weighting)* introduits la première fois pour traiter les données censurées par (van der Laan et Robins, 2003).

Il est intéressant, à titre de comparaison, de confronter cette approche à celle proposée par (Molinaro et al., 2004). En effet, celles-ci se distinguent en deux points clés. Pour commencer, l'approche de (Molinaro et al., 2004) nécessite la modélisation de la distribution conditionnelle de la censure tandis que la nôtre introduit des pondérations basées sur l'estimateur de Kaplan-Meier de la fonction de survie appliquée à la variable C. Par ailleurs, notre approche ne s'attache pas uniquement à la variable de durée, notamment la durée de vie d'un contrat ou l'espérance de vie d'un assuré, mais davantage à la charge ultime qui n'est connue qu'une fois le sinistre clos.

Enfin, la mise en oeuvre de la méthode *IPCW* en utilisant l'estimateur de Kaplan-Meier au sein des modèles de *machine learning*, de façon similaire à (Vock et al., 2016), est décrite ci-dessous.

1. Estimer la fonction de survie S_C de la variable de temps censurée C en utilisant l'estimateur de Kaplan-Meier à partir de la base d'apprentissage des modèles. Nous pouvons observer cette fonction au niveau de la figure 4.1.

2. Calculer les poids *IPCW* qui consistent à attribuer à chaque observation non censurée (ici les sinistres clos) l'inverse de sa probabilité d'être censurée ($\frac{1}{\hat{S}_C(Y_i)}$) et à chaque observation censurée un poids nul. Cette étape fait référence à la description des poids de Kaplan-Meier présentée ci-dessus.

3. Entraîner les modèles de *machine learning* sur une version pondérée de la base d'apprentissage où chaque sinistre présent dans cette base est pondéré par le poids qui lui est attribué.

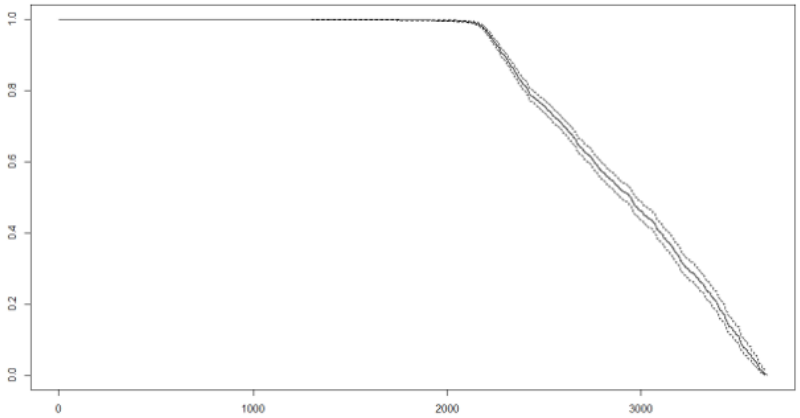


FIGURE 4.1 – Fonction de survie de la variable de temps censurée C - estimateur de Kaplan-Meier

Pour calculer les poids *IPCW* avec l'estimateur de Kaplan-Meier, nous avons eu recours à la fonction *ipcw*() du *package* R *pec* et obtenons les poids exposés au niveau du tableau 4.1.

Le poids maximal de 11.48 s'interprète comme si le sinistre auquel est attribué ce poids est présent 11.48 fois au sein de la base d'apprentissage lors de l'entraînement des modèles.

Min.	1er quantile	Médiane	Moyenne	3ème quantile	Max.
0,00	1,00	1,00	0,99	1,00	11,48

TABLE 4.1 – Récapitulatif des poids *IPCW* calculés au niveau de la base d'apprentissage.

4.2 Provisionnement ligne à ligne par des méthodes de *machine learning*

Cette section aura pour but de décrire la méthodologie employée pour la construction du modèle d'apprentissage statistique ainsi que d'introduire quelques notions communes aux différents modèles de *machine learning* existants.

4.2.1 Méthodologie

Avant d'introduire la méthodologie utilisée pour la construction du modèle d'apprentissage statistique, il est nécessaire d'identifier le type d'apprentissage à mettre en oeuvre. Deux types d'apprentissage existent selon la présence ou l'absence de la variable cible Y ou de la forme à reconnaître.

En cas de présence de la variable Y, il s'agit d'un problème de modélisation, aussi appelé **apprentissage supervisé**. L'objectif poursuivi est de trouver une fonction f qui, appliquée à un ensemble de variables explicatives X, permet de reproduire Y. L'efficacité de cette fonction sera quant à elle étudiée selon un ou plusieurs critères définis.

A l'inverse, lorsque la variable cible Y n'est pas disponible, il s'agit d'un problème de classification ou *clustering*, aussi appelé **apprentissage non-supervisé**. L'objectif alors poursuivi est de regrouper les observations en classes homogènes, les plus dissociées les unes des autres.

Dans notre cas, il s'agit d'un apprentissage dit supervisé où nous disposons d'une base de données contenant à la fois les variables explicatives X ainsi que la variable cible Y. Cette dernière représente le coût final du sinistre lorsqu'il est clos à fin 12/2019, autrement elle reflète la charge D/D nette de recours du sinistre.

Nous pouvons à présent énumérer les principales étapes par lesquelles nous sommes passés pour définir le modèle le plus pertinent à retenir qui peuvent être résumées de la façon suivante.

1. **Construction de la base de données** (voir section 2.3)
   Cette première étape consiste en la création d'une base de données de telle sorte que chaque ligne représente un sinistre et regroupe à la fois les informations sur le sinistre, sur le contrat et enfin sur la ou les victimes.

2. **Exploration et retraitement/recodage des données** (voir section 2.3)
   Il s'agit d'une étape clé destinée à mieux appréhender les données disponibles. Cette étape s'effectue, sans s'y limiter, au travers de l'identification des corrélations existantes entre les différentes variables explicatives mais également entre ces variables et la variable cible, à la fois de manière bidimensionnelle et multidimensionnelle. Cette étape a également pour but le retraitement des données afin que l'exploitation des variables au travers des modèles soit optimale et aboutisse au modèle le plus performant.

3. **Définition de deux échantillons adéquats pour l'apprentissage et le test**
   Il est ici important de définir une base d'apprentissage représentative et couvrant l'ensemble des

effets souhaités afin que la généralisation du modèle soit optimale et pertinente. Il faut ainsi tenir compte de toute particularité pouvant altérer la performance du modèle entraîné, notamment les effets de saisonnalité. Dans le cadre de ce mémoire, les survenances de 2010 à 2013 représenteront l'échantillon d'apprentissage tandis que la survenance 2014 correspondra à l'échantillon de test.

4. **Calibrage et entraînement des modèles sur l'échantillon d'apprentissage**
   La plupart des algorithmes de *machine learning* nécessitent d'être calibrés, à savoir d'optimiser un certain nombre de paramètres avant de les entraîner sur la totalité de l'échantillon d'apprentissage. Nous effectuons ce calibrage à l'aide de la validation croisée (voir sous-section 4.2.2) au sein de l'échantillon d'apprentissage en ayant défini au préalable la liste des paramètres à ajuster et leurs valeurs candidates.

5. **Mesure et comparaison de la performance de prédiction des modèles testés**
   Comme évoqué dans la partie 2.3.3, les données dont nous disposons sont censurées à droite, à savoir que nous ignorons le développement ultérieur des sinistres encore ouverts à fin 2019. Par conséquent, nous emploierons deux approches complémentaires. D'une part, l'analyse sur les sinistres survenus en 2014 et clos durant la période d'observation nous permettra de comparer les modèles testés entre eux. D'autre part, l'analyse de l'ensemble des sinistres survenus en 2014, clos ou encore ouverts à fin 2019, permettra de comparer l'estimation de la charge ultime obtenue au travers des modèles testés à celle obtenue par le modèle mis en place par AXA France (cf. section 3.5).

6. **Estimation de la variance**
   Cette ultime étape nous permet d'estimer les erreurs de prédiction des provisions afin de les comparer à celles des méthodes agrégées. Nous aurons pour cela recours à la méthode de *bootstrap* avec 10 000 itérations pour chaque modèle, aboutissant à une distribution de la charge ultime dont l'erreur de prédiction dispose d'un niveau de confiance de $99,5\%$. Le choix du modèle déprendra également de sa performance en terme de minimisation de cette variance.

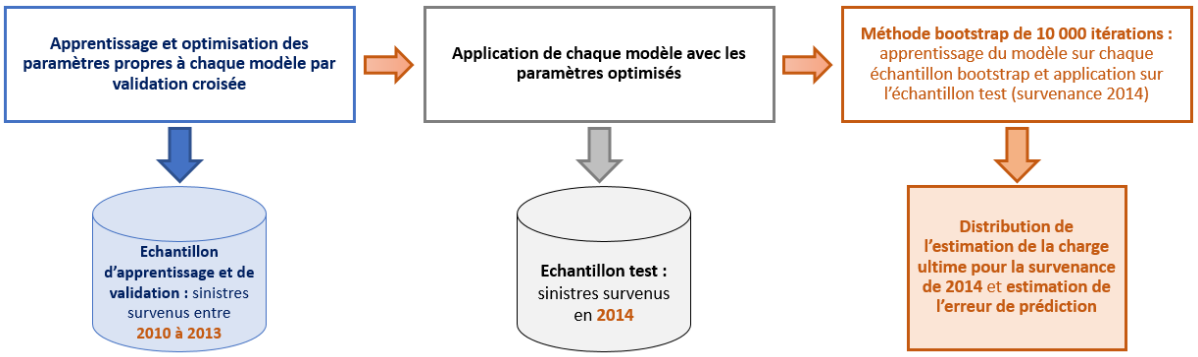


FIGURE 4.2 – Récapitulatif des étapes pour estimation de la charge ultime de la survenance de 2014 et de sa variance à travers les méthodes de *machine learning*

4.2.2 Quelques notions

La capacité de prédiction sur des données de test indépendantes d'un modèle d'apprentissage détermine sa performance de généralisation. Il est primordial d'évaluer cette performance dans la pratique, car

non seulement elle détermine le choix du modèle, mais elle mesure également la qualité du modèle retenu.

Les principales méthodes d'évaluation de la performance sont décrites et illustrées dans ce chapitre. Tout d'abord, nous commençons par expliquer le compromis biais-variance.

Compromis biais-variance

Considérons un *dataset* $D = \{(x_1, y_1), \cdots, (x_n, y_n)\}$ défini i.i.d et tiré d'une distribution $P(X, Y)$. L'objectif est de prédire Y sachant X. Le processus consiste à choisir un algorithme de *machine learning* $\mathcal{A}$ et à apprendre une hypothèse $h_D = \mathcal{A}(\mathcal{D})$ dans la base d'apprentissage. Afin d'évaluer la qualité de la performance de l'hypothèse apprise, nous définissons l'erreur de généralisation (également appelée erreur de test) comme suit

$$\mathbb{E}_{(x,y)\sim P}[(h_D(x) - y)^2].$$

Nous choisissons l'erreur quadratique moyenne en raison de ses bonnes propriétés mathématiques, mais toute autre fonction de perte peut être utilisée. Il est à noter que h_D est une variable aléatoire car elle est fonction de D qui est elle-même une variable aléatoire tirée de P^n. Ainsi, son espérance peut être définie et calculée comme suit

$$\overline{h} = \mathbb{E}_{D\sim P^n}[h_D].$$

Nous souhaitons évaluer la qualité de l'algorithme de *machine learning* $\mathcal{A}$ par rapport à une distribution de données $P(X, Y)$. Pour ce faire, il est nécessaire de calculer l'erreur de test pour $\mathcal{A}$. Ce calcul est possible grâce au fait que h_D est une variable aléatoire, comme mentionné dans le paragraphe précédent. En prenant également l'espérance de D, nous définissons

$$E_{\substack{(x,y)\sim P \\ D\sim P^n}}[(h_D(x) - y)^2].$$

Ce terme peut être décomposé en biais et en variance comme suit :

$$\begin{aligned} E_{x,y,D}[(h_D(x) - y)^2] &= E_{x,y,D}[[(h_D(x) - \overline{h}(x)) + (\overline{h}(x) - y)]^2] \\ &= E_{x,D}[(h_D(x) - \overline{h}(x))^2] + 2E_{x,y,D}[(h_D(x) - \overline{h}(x))(\overline{h}(x) - y)] + E_{x,y}[(\overline{h}(x) - y)^2] \end{aligned}$$

Il est aisé de voir que,

$$\begin{aligned} E_{x,y,D}[(h_D(x) - \overline{h}(x))(\overline{h}(x) - y)] &= E_{x,y}[E_D[h_D(x) - \overline{h}(x)](\overline{h}(x) - y)] \\ &= E_{x,y}[(\overline{h}(x) - \overline{h}(x))(\overline{h}(x) - y)] \\ &= 0 \end{aligned}$$

Par conséquent, il nous reste,

$$E_{x,y,D}[(h_D(x) - y)^2] = E_{x,D}[(h_D(x) - \overline{h}(x))^2] + E_{x,y}[(\overline{h}(x) - y)^2]$$

où le premier terme désigne la variance et où le second peut être décomposé comme suit,

$$\begin{aligned} E_{x,y}[(\overline{h}(x) - y)^2] &= E_{x,y}[((\overline{h}(x) - \overline{y}) + (\overline{y} - y))^2] \\ &= E_{x,y}[(\overline{h}(x) - \overline{y})^2] + E_{x,y}[(\overline{y} - y)^2] + 2E_{x,y}[(\overline{h}(x) - \overline{y})(\overline{y} - y)] \end{aligned}$$

où le dernier terme de l'équation ci-dessus est 0. La décomposition finale est

$$E_{x,y,D}[(h_D(x) - y)^2] = E_{x,D}[(h_D(x) - \overline{h}(x))^2] + E_{x,y}[(\overline{h}(x) - \overline{y})^2] + E_{x,y}[(\overline{y} - y)^2].$$

où le premier terme est la variance, la moyenne du carré des écarts à la moyenne des valeurs de la distribution, le second terme est le biais au carré, l'écart entre la moyenne de notre estimation et la moyenne réelle, et le troisième terme est le bruit irréductible.

Conceptuellement, le biais quantifie l'erreur inhérente au modèle et ce, même s'il est entraîné à l'aide d'une quantité infinie de données. Cette erreur inhérente est introduite en raison des hypothèses formulées pour un modèle appliqué à un problème complexe. Par conséquent, les modèles émettant moins d'hypothèses possèdent un biais plus faible.

D'autre part, la variance nous permet de capter les changements liés à l'apprentissage sur un nouveau *dataset* et le niveau de surapprentissage de notre modèle sur une base d'apprentissage particulière. Ainsi, les modèles comportant moins d'hypothèses tendent à capter tous les modèles complexes sous-jacents de la base d'apprentissage résultant sur une variance élevée.

Le biais et la variance constituent la partie réductible de l'erreur généralisée. Il est évident que l'objectif principal est de réduire autant que possible ces deux valeurs. Cependant, ils évoluent inversement l'un à l'autre, ce qui implique de trouver le bon compromis, tel qu'illustré en figure 4.3.

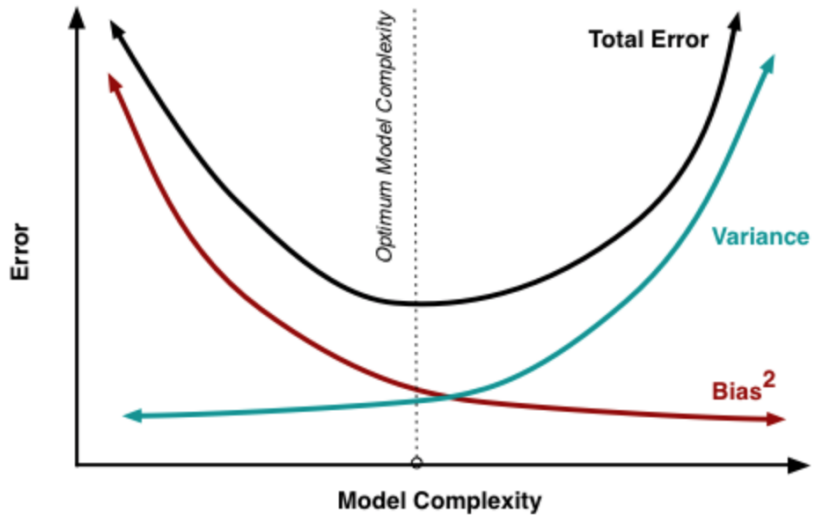


FIGURE 4.3 – Compromis biais-variance.

À mesure que la complexité augmente, la variance augmente également tandis que le biais diminue. Il est crucial de comprendre ce compromis afin de comprendre le comportement du modèle de prédiction. Le scénario parfait correspond au point de complexité au niveau duquel l'augmentation du biais équivaut à la réduction de la variance.

Validation croisée

L'une des techniques les plus simples pour estimer l'erreur de prédiction est la validation croisée. Ainsi, l'erreur de prédiction moyenne $\mathbb{E}[L(y, \hat{f}(x))]$ peut être facilement calculée en utilisant cette méthode, où L correspond à la fonction perte.

Le recours à un échantillon test différent de l'échantillon d'apprentissage servant à entraîner le modèle est une autre alternative possible pour estimer cette erreur. Néanmoins, ceci peut s'avérer contraignant si nous disposons d'un nombre limité de données. La validation croisée est ainsi une des méthodes permettant de répondre à cette problématique.

Cette dernière s'effectue via la séparation du jeu de données en K parties égales, appelés *folds*.

La figure 4.4 illustre le scénario où $K = 5$.

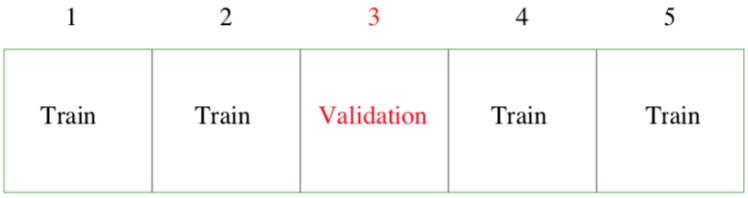


FIGURE 4.4 – Exemple de validation croisée

Alternativement, chacune des K parties est utilisée en tant que jeu de test, le reste servant de base d'entraînement. Cette étape est reproduite de sorte que chaque partie ait servi de jeu de test une fois, et de base d'entraînement $K - 1$ fois. L'erreur de prédiction est enfin calculée sur l'ensemble des jeux de test ainsi constitués.

Pour plus de détails, soit $k : 1, \ldots, n \to 1, \ldots, K$ une fonction d'indexation qui attribue aléatoirement à chaque observation sa partition correspondante. On considère $\hat{f}^{-k}$ la fonction ajustée calculée sans la $k^{\text{ème}}$ partie des données. Enfin, on définit l'estimation de l'erreur de prédiction comme suit

$$CV(\hat{f}) = \frac{1}{n} \sum_{i=1}^{n} L(y_i, \hat{f}^{-k}(x_i)).$$

Métriques d'erreur de régression

Dans cette partie, nous définirons les métriques utilisées pour évaluer les performances des modèles testés ultérieurement. Étant donné que nous sommes dans un cadre de régression, nous avons retenu les métriques définies ci-dessous.

RMSE *(Root mean square error)*

La racine de l'erreur quadratique moyenne est une métrique fréquemment utilisée pour mesurer la différence entre les valeurs observées et celles prédites par le modèle. Elle est égale à la racine carrée du moment d'ordre 2 des résidus. Elle est calculée à partir de la formule ci-dessous

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

où y_i est la valeur réelle, $\hat{y_i}$ la valeur prédite par le modèle et n la taille de l'échantillon.

MAE *(Mean absolute error)*

L'erreur absolue moyenne mesure l'amplitude moyenne des écarts entre les observations et les valeurs prédites sans tenir compte de leur direction. Elle est calculée à partir de la formule ci-dessous

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

où y_i est la valeur réelle, $\hat{y_i}$ la valeur prédite par le modèle et n la taille de l'échantillon.

Ces deux métriques (MAE et RMSE) ont pour similitudes d'être toutes deux positives et indifférentes à la direction des erreurs. Il s'agit de scores orientés négativement, c'est-à-dire que plus les valeurs sont faibles, meilleurs sont les modèles.
Quant aux différences, la RMSE peut s'avérer plus intéressante dans le cas où les erreurs importantes sont fortement indésirables. En effet, de part l'élévation des erreurs au carré avant d'être moyenner accorde plus de poids aux importantes erreurs comparé à la MAE.

4.3 Estimation de la charge ultime par les méthodes de *machine learning*

4.3.1 Arbres de régression CART

Les arbres de décision constituent une méthode efficace d'exploration des données. Il s'agit d'une méthode itérative, dite de partitionnement récursif des données.

Il existe plusieurs types d'arbres de décision. Dans cette partie, nous ferons le focus sur les arbres de classification et de régression, introduits par Breiman et al. (1984). Également connus sous le nom de CART, ces arbres de décision se présentent uniquement sous forme d'arbres binaires.
Par ailleurs et contrairement à d'autres méthodes, la méthode CART présente l'avantage de traiter à la fois des variables numériques et catégorielles aussi bien pour les variables explicatives que pour la variable d'intérêt.

Nous pouvons résumer la construction d'un arbre binaire en une création de noeuds successifs selon des règles définies.

→ Le noeud initial, aussi appelé racine, représente l'échantillon dans son ensemble.
→ Chaque noeud, autre que la racine, est constitué d'un sous-ensemble de l'échantillon du noeud parent.
→ La séparation de l'échantillon entre les deux noeuds qui en découlent s'effectue par le choix d'une division en deux sous-partitions ainsi que le choix de la variable explicative sur la base de laquelle cette division s'effectue.
Cette division est représentée par une valeur seuil de la variable explicative lorsque cette dernière est quantitative, ou une séparation en deux groupes de modalités dans le cas d'une variable qualitative.
→ Cette procédure est ensuite appliquée à l'ensemble de l'arbre par itération.

Étant donné que notre variable d'intérêt est une variable quantitative, nous approfondirons la réflexion autour des arbres binaires de régression CART.

A titre de précision, la séparation du noeud parent en deux noeuds fils, indépendamment des partitionnements à venir, s'effectue via l'identification d'une variable et de sa valeur seuil permettant la séparation en deux noeuds N_G et N_D tout en minimisant le critère d'hétérogénéité intrinsèque retenu. Dans notre cas, il s'agit de minimiser la somme S de leurs variances D_{K_G} et D_{K_D} tel que

$$S = D_{K_G} + D_{K_D}$$

où

$$D_{K_G} = \sum_{i \in N_G} (y_i - c_G)^2$$

et

$$D_{K_D} = \sum_{i \in N_D} (y_i - c_D)^2$$

et où c_G et c_D, les constantes associées aux noeuds, représentent les moyennes des observations contenues respectivement dans les noeuds N_G et N_D.

Cette séparation s'effectue de manière récursive tant que les critères d'arrêt ne sont pas obtenus. Ces derniers sont généralement relatifs au faible nombre d'observations dans le noeud ou à la forte homogénéité du noeud, c'est à dire qu'aucune variable ne permet de le dissocier en deux ensembles suffisamment hétérogènes. Une fois la croissance de l'arbre interrompue, nous obtenons des noeuds terminaux, aussi appelés feuilles.

Le but étant de chercher un modèle parcimonieux, il est nécessaire de suivre une stratégie de recherche d'arbre optimal. Un arbre trop détaillé, associé à une sur-paramétrisation, est instable et donc probablement plus défaillant pour la prévision d'autres observations.

La procédure de sélection d'un **arbre optimal** suit les étapes suivantes :

1. construire l'arbre maximal A_{max},
2. ordonner les sous-arbres selon une séquence emboîtée suivant la décroissance du critère de pénalisation,
3. puis à sélectionner le sous-arbre optimal ; il s'agit de la procédure d'**élagage**.

Construction de la séquence de sous-arbres

Nous définissons K_A, le nombre de noeuds terminaux de l'arbre A, dont la valeur reflète la complexité de l'arbre.

La mesure de la qualité d'ajustement d'un arbre A se fait au travers de la formule qui suit (Wikistat, 2016b) :

$$D(A) = \sum_{k=1}^{K_A} D_k(A)$$

où D_k correspond, tel que mentionné précédemment, à l'hétérogénéité de la feuille k. Outre la variance, ce critère d'hétérogénéité peut également correspondre à d'autres critères à titre d'exemple à la déviance.

Ainsi, la mise en place d'une séquence d'arbres emboîtés consiste à introduire une pénalisation γ de la complexité de l'arbre tel que

$$C(A) = D(A) + \gamma K_A.$$

En l'absence de pénalisation ($\gamma = 0$), l'arbre qui minimise ce critère est l'arbre maximal A_{max} ayant K_A noeuds terminaux.

En introduisant la pénalisation ($\gamma > 0$), il s'agit pour chaque division générant des noeuds terminaux, de comparer l'amélioration apportée à D à la complexité induite via γ. Dès lors que l'amélioration apportée à D est inférieure à γ, les deux feuilles sont considérées comme superflues et regroupées dans le noeud père qui devient à son tour terminal. Il en résulte que le nombre de noeuds terminaux passe de K_A à $K_A - 1$.

Nous reproduisons ces étapes pour l'ensemble des feuilles jusqu'à ce que les améliorations apportées à D par les divisions soient toutes supérieures à γ. Nous aboutissons dès lors à une séquence d'arbres emboîtés tel que

$$A_K \supset A_{K-1} \supset ...A_2 \supset A_1$$

où A_1, le noeud racine, représente l'échantillon dans son ensemble.

Les arbres binaires ne sont pas nécessairement les plus performants en terme de prédiction mais présentent l'avantage d'être schématiques, interprétables et facilement assimilés par des interlocuteurs métier, ce qui en fait aujourd'hui une méthode encore largement plébiscitée.

Après avoir introduit les arbres CART, nous allons à présent procéder à l'application de ce modèle sur nos données. Pour ce faire, nous utilisons le *package "rpart"* (Therneau et Atkinson, 2019) implémenté dans R. Afin d'éviter le phénomène de surapprentissage décrit précédemment, nous cherchons à obtenir un arbre optimal en ayant recours à la procédure d'élagage décrite ci-dessus. Pour ce faire, nous optimisons le paramètre de complexité, également appelé critère de pénalisation, par validation croisée à 5 blocs (par défaut fixé à 10 blocs).

La figure 4.5 nous permet de visualiser l'amélioration de la performance du modèle au fur et à mesure de l'augmentation du nombre de feuilles suivie d'une dégradation de cette dernière suite au surapprentissage. Nous retenons ainsi le paramètre de complexité qui minimise l'erreur estimée. Ainsi l'arbre optimal dans notre cas de figure est détenteur de 8 noeuds terminaux.

Il est également important de noter que le vecteur de poids estimé antérieurement par la fonction *ipcw()* dans R dont le *summary* est présenté au tableau 4.1 est bien pris en compte lors du calibrage du modèle grâce à l'argument *"weights"* de la fonction *rpart()*.

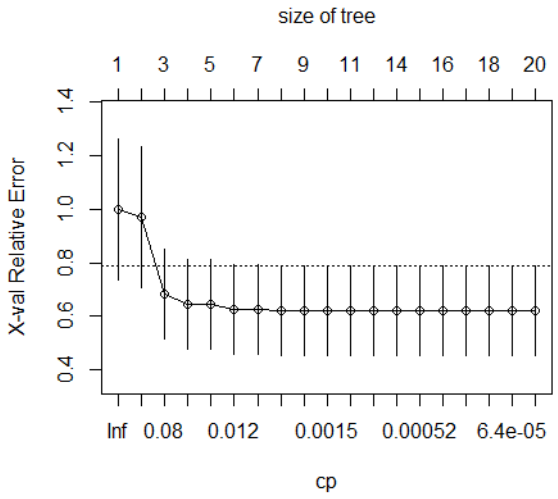


Figure 4.5 – Optimisation du paramètre de complexité *(cp)* par validation croisée

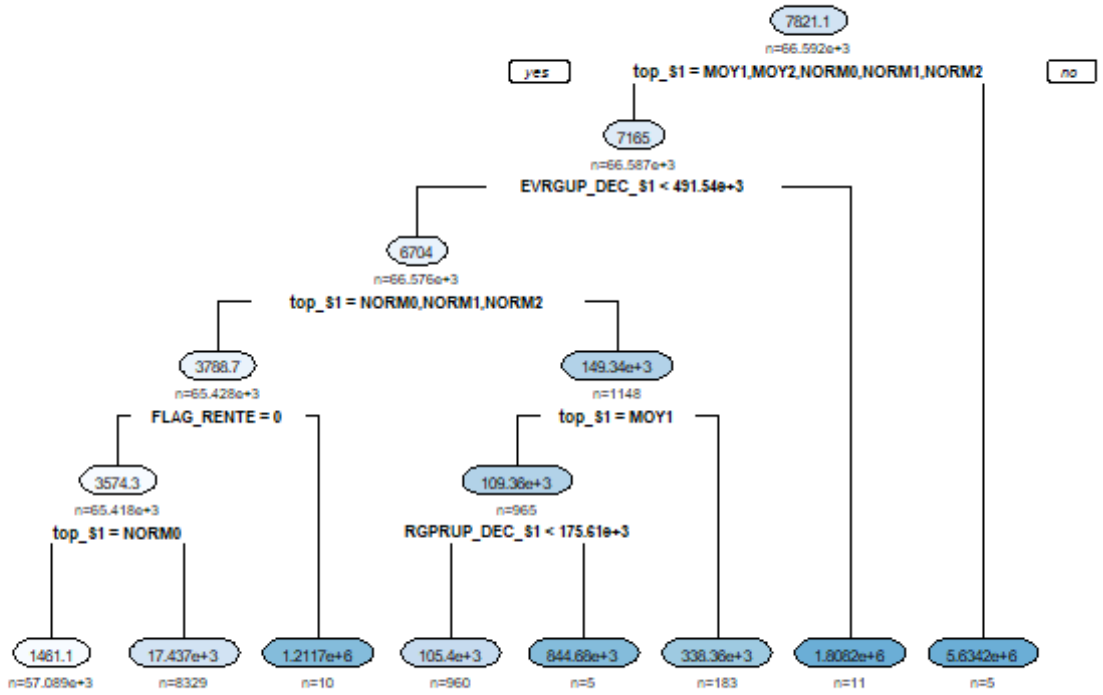


FIGURE 4.6 – Arbre binaire de régression de la charge ultime nette de recours - arbre élagué par validation croisée

En amont d'une analyse de l'arbre élagué obtenu en figure 4.6, nous présentons un exemple de lecture de ce dernier. Nous pouvons dire que si un sinistre a une valeur au niveau de la variable « top_S1 » différente de ("NORM1", "NORM2", "MOY1", "MOY2"), la charge ultime nette de recours de ce dernier sera estimée à 5,6 M€.

En visualisant cet arbre, un premier constat est la mise en exergue de quatre variables discriminant la charge ultime : « top_S1 » reflétant le niveau de gravité à la deuxième année d'inventaire (à partir de la charge D/D), « FLAG_RENTE » indiquant si l'indemnisation du sinistre est sous forme de rente ou pas, « RGPRUP_DEC_S1 » précisant le montant réglé en principal à la deuxième année d'inventaire et « EVRGUP_DEC_S1 » indiquant le montant des évaluations de règlement à la deuxième année d'inventaire.

Au final, le modèle obtenu aura permis de distinguer 8 profils de sinistres différents. Celui-ci présente l'avantage d'être simple à assimiler et rapide à mettre en place. Néanmoins, les arbres de régression présentent, de façon générale, l'inconvénient d'être fortement sensible aux fluctuations de l'échantillon d'apprentissage. Ce phénomène peut s'avérer réellement contraignant. Dans notre cas de figure, ceci présente la limitation de l'estimation d'un sinistre grave au montant de 5,6M€, montant qui peut être dépassé par certains sinistres. Cet aspect est une des raisons à l'origine du succès des méthodes ensemblistes. Deux algorithmes faisant appel à ces méthodes sont présentés et appliqués par la suite.

4.3.2 Forêts aléatoires

Avant de présenter le modèle des forêts aléatoires, il est nécessaire au préalable d'introduire les notions clés qui y sont liées à savoir le *bootstrap* et le *bagging*.

Bootstrap

Le *bootstrap* est une méthode d'inférence statistique basée sur la multiple réplication des données issues de l'échantillon initial. En effet, chaque échantillon *bootstrap* est construit en tirant n observations avec remise depuis l'échantillon initial de taille n. Par conséquent certaines observations ont été tirées plusieurs fois tandis que d'autres ne l'ont pas été. Ces observations non tirées sont appelées observations *Out-Of-Bag* et servent de base de validation.

Bagging

Ce mot-valise introduit par Breiman (1996) et formé par la fusion des mots *Bootstrap* et *Aggregation* est une méthode d'apprentissage ensembliste basée sur l'agrégation de modèles indépendants. Cette agrégation réalisée au travers d'une moyenne des prévisions en cas de régression ou d'un vote à la majorité en cas de classification, permet ainsi d'améliorer les performances de prédiction en aboutissant à une plus faible variance de prévision en comparaison de celle obtenue unitairement par chacun des modèles.

Les forêts aléatoires, ou *Random Forest*, ont pour leur part été introduites par Breiman (2001) et représentent une amélioration du *bagging* appliquée aux arbres de décision CART via l'intégration d'un aléa additionnel. En effet, si à l'image de ce dernier les arbres sont construits sur la base d'échantillons *bootstrap*, les noeuds sont eux définis à l'aide d'un sous-ensemble de variables tirées aléatoirement qui varie tout au long de la création de l'arbre. Cette composante aléatoire supplémentaire permet de rendre les arbres davantage indépendants.

Algorithme 1 Forêts Aléatoires

Soit $x_i = \{x_{i1}, \ldots, x_{im}\}$ les valeurs des m variables explicatives de l'observation à prévoir i
tel que $i \notin \{1, \ldots, n\}$
$z = \{(x_1, y_1), ..., (x_n, y_n)\}$ l'échantillon en entrée
B le nombre d'échantillons *bootstrap*
Pour b partant de 1 à B, **faire** :
Tirer un échantillon *bootstrap* z_b
Estimer un arbre sur z_b :
Pour chaque noeud, **faire** :
Tirer aléatoirement p variables parmi les m existantes
Identifier la variable la plus optimale parmi celles tirées
Créer les deux noeuds fils
Fin pour
Fin pour
Si Problème de régression **alors**
Calculer l'estimation moyenne $\hat{f}_B(x_i) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}_{z_b}(x_i)$
Fin si
Si Problème de classification **alors**
Prendre le résultat du vote à la majorité
Fin si

Par ailleurs, il est possible d'optimiser les paramètres du modèle au travers de l'erreur *Out-Of-Bag*.

Erreur *Out-Of-Bag (OOB)*

L'erreur *Out-Of-Bag* est une méthode de mesure de l'erreur de prédiction dans le cas du *bagging*, ici en l'occurrence dans le cas des forêts aléatoires.
Elle consiste dans un premier temps à calculer l'erreur de prédiction de chaque arbre sur l'ensemble de ses observations *Out-Of-Bag*. Ces erreurs sont ensuite regroupées par observation et leur moyenne représente l'erreur *Out-Of-Bag*.

Enfin, Breiman (2001) a démontré que le taux d'erreur converge à mesure que le nombre d'arbres *bootstrap* augmente. Ainsi, le *finetuning* du nombre d'arbres *bootstrap* revêt un intérêt limité à la différence de leur profondeur qu'il est important de contrôler pour augmenter la vitesse d'entraînement et éviter le surapprentissage.

Pour mettre en oeuvre cet algorithme sur nos données tout en prenant en compte les pondérations estimées par la fonction *ipcw()*, nécessaires pour la prise en compte du biais induit par la censure, nous avons recours au *package "ranger"* (Wright et Ziegler, 2017) présent dans R. Ce *package* est une implémentation plus rapide des forêts aléatoires (Breiman, 2001), particulièrement adaptée aux données de grandes dimensions. Le recours à ce *package* est également justifié par la présence de l'argument *case.weights* qui permet d'inclure les pondérations évoquées ci-dessus.

Pour ce faire, nous avons cherché à optimiser certains hyper-paramètres du modèle, notamment le *mtry* référant au nombre de variables explicatives tirées pour la construction de chaque noeud et le *min.node.size* indiquant le nombre minimal d'observations au sein des noeuds terminaux tout en fixant le nombre d'arbres à sa valeur par défaut, c'est-à-dire à 500 arbres. Cette optimisation a été réalisée par validation croisée à 5 blocs et le critère de division choisi est celui de la variance. La figure 4.7 expose le résultat de cette optimisation. Ainsi, nous retenons les valeurs des paramètres minimisant l'erreur quadratique moyenne : *mtry* à 10 et *min.node.size* à 1.
L'interprétation et la mesure de performance du modèle sont adressées dans la suite du mémoire.

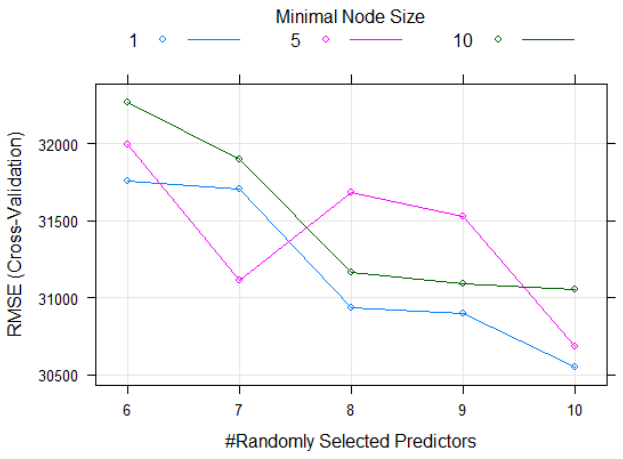


Figure 4.7 – Optimisation des hyper-paramètres *mtry* et *min.node.size* par validation croisée

4.3.3 *Gradient Boosting*

Principe du *boosting*

Pour commencer, étant donné la nature de la variable d'intérêt dans notre cas de figure, nous présenterons les algorithmes de cette section pour une régression.

Le *boosting* est comparable au *bagging* dans la mesure où il a également vocation à créer un ensemble de modèles dont les résultats sont consolidés par pondération. Il se distingue néanmoins dans sa manière de créer les modèles. En l'occurrence chaque modèle, à l'exception du premier, prend en compte les résultats du modèle précédent en donnant plus de poids aux observations incorrectement prédites. Par conséquent, l'effort est concentré autour des observations complexes tandis que la consolidation finale réduit le risque de sur-ajustement.

La construction d'un algorithme de *boosting* s'effectue à l'aide de paramètres qui le caractérisent :

- la fonction perte pour le calcul de l'erreur de prédiction ;
- la manière de pondérer, à savoir de donner davantage de poids aux observations dont l'erreur de prédiction est la plus importante ;
- la manière de consolider les résultats unitaires de chaque modèle en les pondérant.

En 1999, Breiman (1999) propose, dans son rapport technique, de reconsidérer le *boosting* en tant qu'algorithme d'optimisation car il permet, en effet, d'approximer la fonction f à l'aide d'un modèle additif basé sur des itérations successives

$$\hat{f}(x) = \sum_{m=1}^{M} c_m \delta(x; \gamma_m)$$

où c_m correspond au poids et δ au modèle basé sur les données x et le paramètre γ_m.

En effet, cela correspond à chaque étape à la résolution de l'équation suivante

$$(c_m, \gamma_m) = \underset{(c,\gamma)}{\arg\min} \sum_{i=1}^{n} l(y_i, \hat{f}_{m-1}(x_i) + c\delta(x_i; \gamma))$$

où l correspond à la fonction de perte.

Par conséquent, nous obtenons la formule suivante que nous appliquons successivement afin d'obtenir les améliorations des modèles

$$\hat{f}_m(x) = \hat{f}_{m-1}(x) + c_m \delta(x; \gamma_m).$$

Gradient Boosting

Comme relaté au sein de Wikistat (2016a), Friedman (2002) est le premier à introduire en 2002 la notion de *Gradient Boosting* qui correspond à un algorithme reposant sur une fonction perte l différentiable et supposée convexe. L'ensemble des algorithmes de ce type sont regroupés sous l'appellation *Gradient Boosting Models (GBM).*

Ces algorithmes ont la particularité de créer successivement des modèles qui sont systématiquement plus proches de la meilleure solution que ne l'était le modèle précédent.

Afin de garantir la convergence du modèle, les étapes successives suivent la direction du gradient de la fonction perte l. Par ailleurs, nous évitons le surapprentissage en réalisant une approximation du gradient à l'aide d'un arbre de régression.

La formule évolutive du *boosting* présentée précédemment prend la forme d'une descente de gradient tel que

$$\hat{f}_m = \hat{f}_{m-1} - \gamma_m \sum_{i=1}^{n} \nabla_{\hat{f}_{m-1}} l(y_i, \hat{f}_{m-1}(x_i)).$$

Par conséquent, nous pouvons réduire le problème à l'identification du meilleur pas de descente γ tel que

$$min_\gamma \sum_{i=1}^{n} \left[l\left(y_i, \hat{f}_{m-1}(x_i) - \gamma \frac{\partial l(y_i, \hat{f}_{m-1}(x_i))}{\partial \hat{f}_{m-1}(x_i)} \right) \right].$$

Algorithme 2 Gradient Boosting

Soit $\hat{f}_0(x) = \arg\min_\gamma \sum_{i=1}^{n} l(y_i, \gamma)$

Pour m partant de 1 à M, **faire** :

Calculer les pseudo-résidus $r_{im} = -\left[\frac{\partial l(y_i, \hat{f}(x_i))}{\partial \hat{f}(x_i)}\right]_{\hat{f}(x)=\hat{f}_{m-1}(x)}$ pour $i = 1, ..., n$

Ajuster un algorithme de faible performance $\gamma_m(x)$ aux pseudo-résidus, à savoir l'entraîner sur la base d'apprentissage $(x_i, r_{im})_{i=1}^{n}$

Calculer le paramètre γ_m en résolvant l'équation d'optimisation

$\gamma_m = \arg\min_\gamma \sum_{i=1}^{n} l(y_i, \hat{f}_{m-1}(x_i) + \gamma \delta_m(x_i))$

Mettre à jour le modèle $\hat{f}_1(x) = \hat{f}_0(x) + \gamma_0 \delta_0(x)[x, y - \hat{f}_0(x)]$

Fin pour

Le résultat correspond à $\hat{f}_M(x)$

Une variante du *Gradient Boosting*, le *Gradient Boosting* stochastique, permet de pallier à la problématique du surapprentissage en rendant les modèles davantage indépendants au travers d'un sous-échantillonage effectué à chaque étape.

Une autre méthode pour pallier au même problème consiste en l'ajout d'un coefficient de rétrécissement η compris entre 0 et 1. Ce dernier réduit la vitesse de convergence en limitant l'impact des nouveaux modèles dans l'agrégation.

Enfin, pour s'assurer d'obtenir un modèle ayant la plus faible erreur de prédiction, il est conseillé d'optimiser les trois paramètres suivants :

- le nombre d'arbres ;
- la profondeur maximale souhaitée de ces arbres ;
- le coefficient de rétrécissement.

Extreme Gradient Boosting (XGBoost)

En 2016, CHEN et GUESTRIN (2016) proposent à leur tour une version améliorée du *boosting*, l'*Extreme Gradient Boosting* aussi appelé *XGBoost*. Ce dernier consiste en l'ajout d'un terme de régularisation à la fonction perte, définie convexe et différentiable, aboutissant à une nouvelle fonction nommée L, aussi appelée fonction objectif

$$L(f) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{m=1}^{M} \Omega(\delta_m)$$

où Ω correspond à

$$\Omega(\delta) = \alpha|\delta| + \frac{1}{2}||w||^2$$

où $|\delta|$ correspond au nombre de noeuds terminaux de l'arbre et w à leurs valeurs.

Ce terme de régularisation permet d'éviter le surapprentissage en réduisant l'impact des ajustements successifs. Il est composé d'une pénalisation de type L_1 ainsi que d'une pénalisation de type L_2.

L'approximation du gradient peut se faire au travers d'un développement de Taylor à l'ordre 2. Dans ce cas, il s'agit de calculer la somme des dérivées première et seconde pour chaque noeud terminal. Cela nous permet notamment de paralléliser l'algorithme pour obtenir un entraînement plus rapide.

L'algorithme *XGBoost* permet également le traitement des données manquantes ou incomplètes en calculant le gradient uniquement sur les données présentes.

Pour finir et en complément des paramètres conseillés pour le *Gradient Boosting*, il est recommandé d'optimiser également les paramètres suivants pour le *XGBoost* :

- valeur minimale de réduction de la fonction perte l nécessaire à la création d'un nouveau noeud ;
- fraction d'observations utilisées par l'algorithme ;
- fraction de variables à utiliser pour la construction de chaque arbre ;
- coefficient de pénalisation L1 ;
- coefficient de pénalisation L2.

Pour appliquer le modèle *XGBoost* sur nos données, nous avons fait appel au *package "XGBoost"*(CHEN et al., 2021) implémenté dans R. Ce dernier a pour avantage l'optimisation du temps de calcul grâce à la possibilité de parallélisation exploitant ainsi tous les coeurs disponibles au niveau de notre machine. Ceci s'avère particulièrement utile en phase d'optimisation des hyper-paramètres. Il faut savoir que le paramétrage du *XGBoost* impacte fortement sa performance, d'où la nécessité d'y consacrer un intérêt particulier.

Pour ce faire, nous avons fixé quelques hyper-paramètres et en avons optimisé d'autres par validation croisée à 5 blocs. Nous avons défini les hyper-paramètres suivants :

- *subsample* permet de définir le rapport entre les sous-échantillons servant de bases d'apprentissages pour définir les arbres et l'échantillon d'apprentissage initial. Nous l'avons fixé à 0,6 (valeur par défaut à 0,5). Ceci permet également de réduire le temps de calcul car les données à analyser

sont moins conséquentes.

- *gamma* correspond à un des paramètres de régularisation pour l'étape d'élagage des arbres. Ce paramètre précise ainsi la réduction de perte minimale nécessaire pour créer une partition supplémentaire. Nous avons conservé sa valeur par défaut à 0.

Concernant les hyper-paramètres optimisés, nous avons défini une grille de valeurs permettant ainsi la construction des modèles selon la matrice des valeurs possibles. Nous obtenons ainsi les résultats présentés en figure 4.8. Nous avons ainsi optimisé les paramètres suivants :

- *eta* (valeurs testées 0,1 et 0,2 - valeur par défaut à 0,3) contrôle le taux d'apprentissage du modèle. Ce paramètre permet de réduire les poids des caractéristiques afin d'éviter le phénomène de surapprentissage. Ainsi, plus la valeur de ce paramètre est faible, plus le modèle est robuste face au surapprentissage.

- *colsample_ bytree* (valeurs testées 0,5/0,6 et 0,7 - valeur par défaut à 1) correspond à la proportion de variables tirées aléatoirement afin de construire chacun des arbres.

- *max_ depth* (valeurs testées 6 et 7 - valeur par défaut à 6) correspond à la profondeur maximale de chaque arbre.

Au vue des résultats obtenus en figure 4.8, les valeurs retenues sont : *eta* à 0,2 ; *max_ depth* à 7 et le *colsample_ bytree* à 0,7.

Une fois ces paramètres définis, nous avons également cherché à optimiser le paramètre *nrounds* correspondant au nombre d'arbres maximal. Les résultats de cette optimisation sont affichés en annexe A.9. Ainsi, afin d'éviter le surapprentissage du modèle, nous fixons sa valeur à 30. Ceci permet également de réduire le temps de calcul.

Il ne faut pas omettre que l'application de ce modèle prend, à son tour, en compte les pondérations estimées antérieurement en utilisant l'argument *weight*.

De même que pour les forêts aléatoires, l'interprétation et la mesure de performance de ce modèle sont adressées dans ce qui suit.

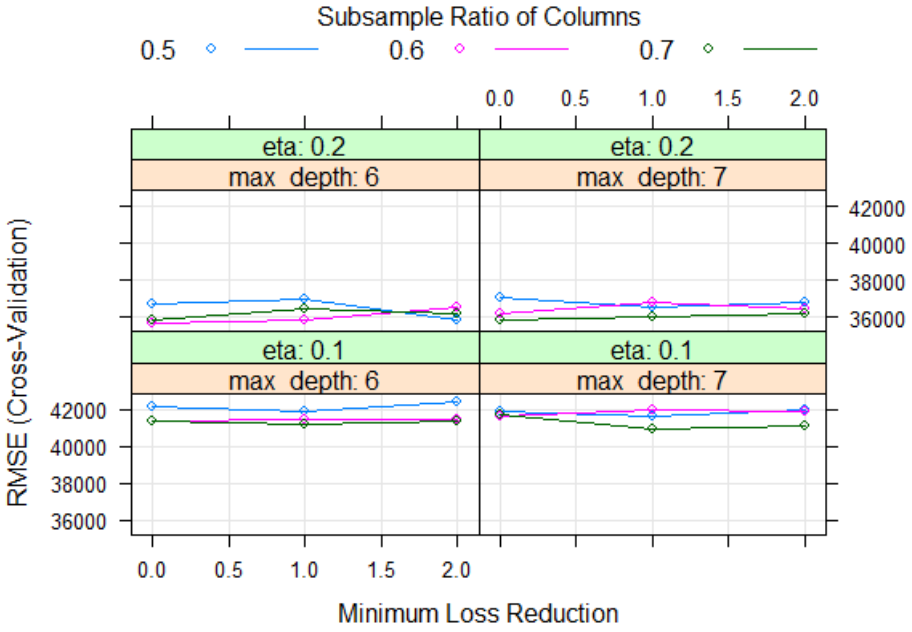


FIGURE 4.8 – Optimisation des hyper-paramètres du modèle *XGBoost* (*eta*, *max_ depth*, *colsample_ bytree*) par validation croisée

4.3.4 Importance des variables

La critique qui peut être émise concernant les forêts aléatoires et le *XGBoost* et les modèles construits par agrégation de façon générale est qu'il s'agit de boîtes noires. Toujours est-il qu'il est possible d'obtenir des informations au travers ces modèles grâce aux calculs de l'importance de chaque variable explicative. Il est ainsi possible de déduire la contribution de chaque variable dans la construction du modèle. Ceci présente un fort intérêt, notamment en présence d'un nombre important de variables comme notre cas de figure.

Ci-dessous une présentation du calcul de l'importance des variables aussi bien au niveau des forêts aléatoires qu'en *XGBoost* selon les méthodes de calcul que nous avons choisies et dont les résultats sont exposés en figure 4.9.

Random Forest

La mesure de l'importance des variables pour l'algorithme *Random Forest*, comme défini par ZHU et al. (2015), s'effectue comme suit. La première étape consiste à entraîner le modèle sur les données en retenant l'erreur *out-of-bag* associée.

Afin de mesurer l'importance de la $i^{\text{ème}}$ variable, à la suite de la première itération, il s'agit de permuter les valeurs de la $i^{\text{ème}}$ variable entre les différentes observations avant de calculer et de retenir l'erreur *out-of-bag* nouvellement obtenue.

Ainsi, l'importance de la $i^{\text{ème}}$ variable est mesurée en prenant la moyenne des différences d'erreur *out-of-bag* avant et après la permutation sur l'ensemble des arbres. Plus cette valeur est grande, plus la

variable est importante.

XGBoost

La mesure de l'importante des variables pour l'algorithme *XGBoost*, telle que définie par Breiman et al. (1984), s'effectue différemment. Il s'agit dans un premier temps de mesurer l'importance de la $l^{\text{ème}}$ variable sur chacun des arbres selon la formule suivante

$$\mathcal{I}_l^2(T) = \sum_{t=1}^{J-1} \hat{i}_t^2 I(v(t) = l).$$

où T correspond à l'arbre et $J-1$ au nombre de noeuds internes dans cet arbre.

En effet, pour mesurer l'importance de la $l^{\text{ème}}$ variable, il est nécessaire de calculer la somme des améliorations estimées $\hat{i}^2$ pour chaque noeud dont la variable l a servi de séparateur.

Par la suite et dans le cas de modèles additifs, il s'agit de calculer la moyenne de $\mathcal{I}_l^2(T)$ pour l'ensemble des M arbres

$$\mathcal{I}_l^2 = \frac{1}{M} \sum_{m=1}^{M} \mathcal{I}_l^2(T_m).$$

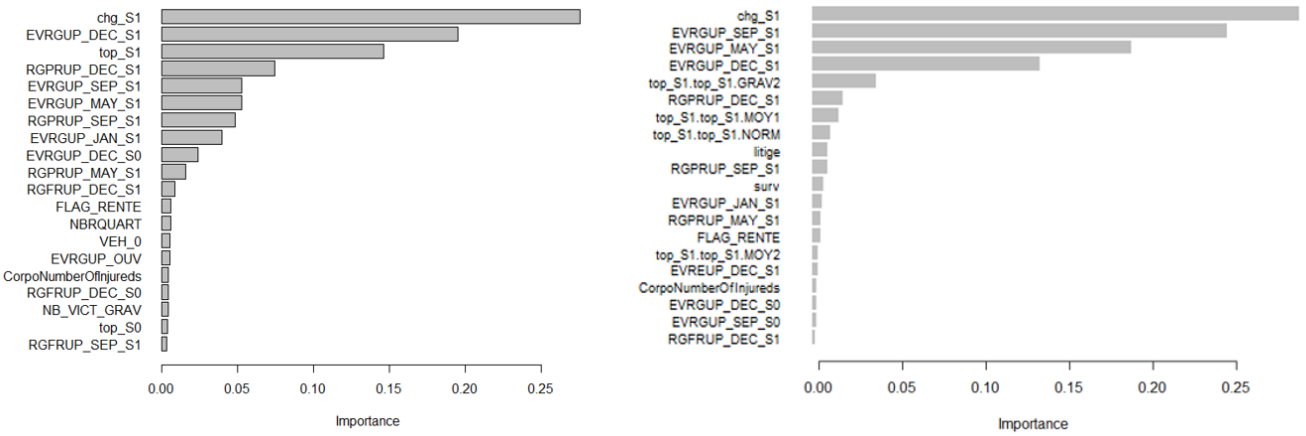


Figure 4.9 – Classement décroissant des importances de variables explicatives - Forêts aléatoires (librairie *"ranger"*) (à gauche) vs. *XGBoost* (à droite)

La figure 4.9 montrent que les deux modèles font ressortir globalement les mêmes variables. Il s'agit principalement de variables portant sur les montants : la charge D/D nette de recours, les règlements en principal et les évaluations de règlements vus à des trimestres différents de la deuxième année d'inventaire. La variable « top_S1 » ressort également (introduite précédemment en section 2.3). De façon beaucoup plus atténuée, nous retrouvons au niveau des forêts aléatoires les variables « FLAG_RENTE » et le taux de responsabilité « NBRQUART » et au niveau du *XGBoost* les variables « litige » et année de survenance.

Ce constat nous mène vers la conclusion que les variables montants, tenant compte implicitement des circonstances du sinistre aussi bien en règlement qu'en évaluations de règlement, l'emportent sur les variables liées au contexte propre au sinistre.

4.3.5 Récapitulatif des résultats

Après avoir calibré les trois modèles présentés ci-dessus sur la base d'apprentissage, constituée des sinistres survenus entre 2010 et 2013, nous les avons appliqués sur la base test constituée des sinistres survenus en 2014.

L'objectif de cette partie est d'opérer une première comparaison de la performance prédictive des trois modèles construits. Néanmoins, disposant de données censurées à droite, notre base de comparaison est ainsi limitée aux sinistres clos non sujets à la censure. Cette base a pour charge ultime 144 M€, soit 53% de la charge D/D nette de recours de la totalité des sinistres survenus en 2014 vus à fin 2019 (un taux de censure en volume de 6%).

Pour ce faire, nous nous sommes basés sur les deux métriques présentées en sous-section 4.2.2 : l'erreur quadratique moyenne *(RMSE)* et l'erreur absolue moyenne *(MAE)*. Nous avons également comparé la charge totale prédite à la charge totale réelle. Ceci permet de définir si le modèle sous-estime ou au contraire surestime les réserves sur le périmètre en question.

Le tableau 4.2 récapitule les résultats obtenus. A titre d'exemple, le modèle CART surestime la charge réelle de 5,1% en estimant la charge totale à 151 M€. Quant aux deux métriques de performances, nous obtenons une RMSE de 59 235 € et une MAE de 918 €. Au vue des différents résultats obtenus, nous constatons que le modèle le plus performant d'un point de vue prédictif est le *XGBoost*. Ainsi, en utilisant ce modèle, nous aurions eu un boni de 2 M€.

Néanmoins, il faut garder à l'esprit que ce constat autour de la performance du modèle ne peut être généralisé sur toute la base de survenance 2014 étant donné qu'un modèle peut s'avérer performant pour l'estimation d'une catégorie de sinistres et moins pour une autre et vice-versa.

	Erreur absolue moyenne - *MAE (en €)*	**Erreur quadratique moyenne** - *RMSE (en €)*	**Montant de la charge totale estimée** *(en €)*	**Taux d'évolution la charge totale estimée par rapport à la charge totale observée**
Arbres de régression CART	918	59 235	151 612 311	5,1%
Forêts aléatoires	502	47 958	149 519 613	3,6%
Extrem Gradient Boosting	348	38 015	146 488 810	1,5%

TABLE 4.2 – Récapitulatif des métriques et des estimations de la charge ultime des trois modèles appliqués à la survenance 2014 - (mesures sur les sinistres clos uniquement)

4.4 Estimation de la variance

4.4.1 Méthode du *bootstrap*

La méthode du *bootstrap* (Efron, 1982), dont le concept a été introduit brièvement en sous-section 4.3.2, est une méthode de rééchantillonage où l'idée est que la distribution empirique $\hat{F}$ se substitue à la distribution de probabilité inconnue F d'où provient l'échantillon d'apprentissage. Suite au calcul de l'estimateur étudié (dans notre cas de figure la charge ultime) à partir des différents échantillons *bootstrap*, il en résulte une distribution simulée de cet estimateur qui converge asymptotiquement vers la loi de ce dernier (sous les hypothèses de présence d'un échantillon indépendant de même loi et que l'estimateur est indépendant de l'ordre des observations). Il est ainsi possible, à travers cette approximation, d'estimer le biais, la variance et l'intervalle de confiance de l'estimateur sans émettre d'hypothèse sur sa loi réelle.

A préciser que dans cas de figure, nous effectuons du *bootstrap* dit « non-paramétrique » étant donné que $\hat{F}$ est une estimation non paramétrique de F.

4.4.2 Récapitulatif des résultats

Après avoir calibré les trois modèles de *machine learning* et comparé leur performance de prédiction sur la base des sinistres survenus en 2014 et clos à fin 2019, nous avons souhaité estimer la distribution de la charge ultime de l'ensemble des sinistres survenus en 2014. Pour ce faire, nous avons eu recours à la méthode du *bootstrap* qui peut être résumée dans l'algorithme 3.

Algorithme 3

Soient x l'échantillon d'apprentissage composé des sinistres survenus entre 2010 et 2013 de taille n, p le vecteur de pondérations estimées par l'estimateur de Kaplan-Meier à travers la fonction *ipcw()* et y l'échantillon test composé des sinistres survenus en 2014 (clos ou en cours à fin 2019).
Pour b partant de 1 à 10000, **faire** :
Sélectionner un échantillon $x^{*b} = (x_1^{*b}, \ldots, x_n^{*b})$ à l'aide d'un tirage avec remise à partir de l'échantillon initial x. Le vecteur p^{*b} contient les poids correspondants aux observations de l'échantillon x^{*b}.
Calibrer le modèle f^{*b} sur l'échantillon x^{*b} en incluant les pondérations p^{*b}.
Appliquer f^{*b} sur l'échantillon test y et obtenir l'estimation moyenne de la charge ultime $\hat{\theta}^{*b}$.
Fin pour
Obtention de la distribution empirique de la charge ultime de la survenance 2014.

Cet algorithme a été appliqué au niveau des trois modèles afin d'obtenir les distributions de charge ultime exposées en figure 4.10.

Dans un premier temps, il est possible d'observer au sein de cette figure davantage de similitudes entre les deux distributions du modèle *XGBoost* et celui des forêts aléatoires qu'avec le modèle *CART* avec des bornes plus larges au niveau de ce dernier.

Dans un second temps, il faut noter que la charge ultime est estimée en moyennant les 10 000 estimations obtenues. Cet exercice permet ainsi de fournir une estimation plus aboutie en comparaison de celle basée uniquement sur l'échantillon d'apprentissage initial. En effet, le fait d'entraîner les modèles

sur différents échantillons permet de tenir compte de différents scénarios et ainsi d'évaluer la volatilité que peut avoir la charge ultime.

La charge moyenne de la distribution *bootstrap* est de 304 M€ pour *CART*, de 317 M€ pour les forêts aléatoires et de 321 M€ pour le *XGBoost*. A titre de comparaison, nous obtenons une charge de 300 M€ à l'aide du modèle AXA France (sans facteur de queue), à vision similaire (c'est-à-dire à la deuxième année d'inventaire).

L'exercice d'estimation de la charge ultime à l'aide du modèle AXA France pour les sinistres survenus en 2014 a été réitéré chaque année jusqu'à fin 2019 et ses résultats ont été synthétisés dans le tableau 3.7. Ainsi, nous pouvons constater que la charge ultime des sinistres survenus en 2014 est estimée à fin 2019 à 319 M€. L'important delta de 19 M€ existant entre les deux estimations du modèle AXA France à des années différentes 2015 vs. 2019 nous amène à déduire que les deux modèles des forêts aléatoires et du *XGBoost* anticipent plus rapidement le niveau de charge atteint par les méthodes agrégées quatre années plus tard.

Cette anticipation proviendrait principalement d'une meilleure prise en compte des sinistres graves. Néanmoins, le modèle CART reste à niveau d'estimation similaire à celui obtenu à fin 2015 avec une charge à 304 M€. Cette sous-performance pourrait notamment s'expliquer par la forte dépendance des arbres de régression à l'échantillon d'apprentissage qui pourrait, tel que présenté en sous-section 4.3.1, mener vers une sous-estimation des sinistres graves.

Dans un troisième temps, le tableau 4.3 récapitule les estimations de charge ultime évoquées ci-dessus ainsi que l'erreur standard de chacune de ces estimations. Il s'agit d'un des avantages majeurs du recours à la méthode du *bootstrap*, à savoir de pouvoir quantifier cette valeur afin d'aboutir à une mesure de la volatilité de nos estimations. En se basant sur ce critère, il apparaît à nouveau que les modèles des forêts aléatoires et *XGBoost* possèdent des performances comparables.

En se comparant encore une fois aux erreurs de prédiction à l'ultime estimées par le modèle de Mack, nous pouvons affirmer que les modèles de *machine learning* permettent de gagner en précision avec des erreurs de prédiction de 3,7 k€ pour le *XGBoost* et 4,6 k€ pour les forêts aléatoires contre 41 k€ en 2015 et 25 k€ en 2019 pour le modèle AXA France. A partir des résultats exposés, il est évident que l'utilisation du modèle *XGBoost* permettrait d'aboutir à un *SCR* de réserve plus faible que celui obtenu au travers du modèle AXA France. Néanmoins, n'ayant pas accès au modèle interne AXA France, il était impossible de quantifier ce gain avec précision ou encore de mesurer sa sensibilité à la variation des différentes volatilités obtenues.

Dans un quatrième temps, il faut noter qu'il est possible d'adopter une stratégie de provisionnement plus précautionneuse. En effet, il suffit de considérer un quantile de la distribution supérieur à la moyenne pour l'estimation de la charge ultime. Ceci permet d'intégrer une marge de prudence supplémentaire, à l'image d'AXA France qui l'applique pour ses sinistres les plus graves. Néanmoins, ce parallèle peut s'avérer trompeur en raison du périmètre d'application de cette marge qui concerne l'ensemble des sinistres dans notre cas contrairement à AXA France.

En effet, nous obtenons à l'aide du modèle AXA France une charge ultime de 315 M€ à fin 2015, dont 15 M€ de marge de prudence, et 328 M€ à fin 2019, dont 10 M€ de marge de prudence contre 324 M€ pour le modèle *XGBoost*, dont 3 M€ de marge de prudence avec un quantile à 75%. Cela illustre

la performance de ce dernier qui, malgré une marge de prudence relativement faible, obtient un écart de seulement 4 M€ avec l'estimation AXA France en 2019 contre 13 M€ d'écart entre ces mêmes estimations en 2015 et 2019, et ce malgré une marge de prudence élevée en 2015 à 15 M€.

Dans un dernier temps, au niveau de la figure 4.10, nous notons qu'il est possible d'avoir une première approximation du *SCR* de réserve grâce au quantile à 99,5% de la distribution. Ce dernier, dans le cas du modèle *XGBoost*, s'élève à 10 M€. Ainsi, détenir des réserves de 84 M€ (charge D/D de 230 M€ à fin 2015) permettrait de faire face à un choc extrême bicentenaire.

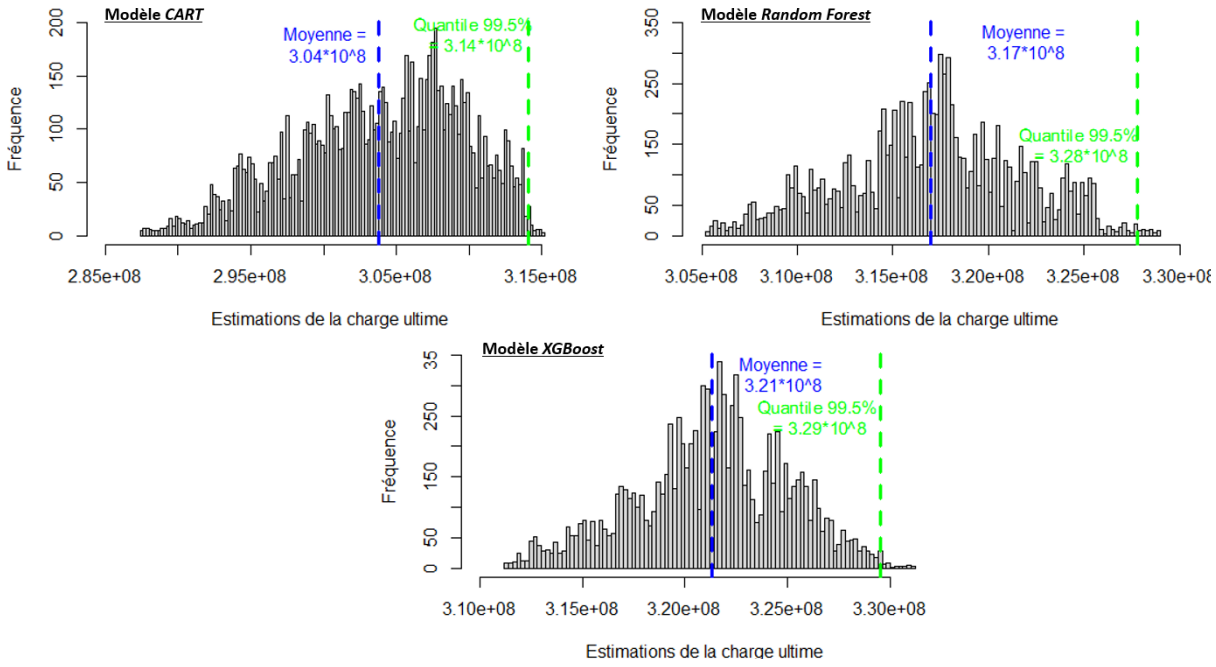


FIGURE 4.10 – Distribution des estimations de la charge ultime de la survenance 2014 obtenue par la méthode *bootstrap - 10 000 itérations*

Survenance 2014	**Moyenne des estimations de la charge ultime** *(en k€)*	**Erreur standard de la moyenne des estimations de la charge ultime** *à 99.5% (en k€)*
Arbres de régression CART	303 779	5 779
Forêts aléatoires	316 962	4 674
Extreme Gradient Boosting	321 283	3 738

TABLE 4.3 – Récapitulatif des estimations de charge ultime obtenues par la méthode *bootstrap* et de l'erreur standard de ces estimations pour les trois modèles de *machine learning*

4.5 Limites

A partir des résultats obtenus, nous avons pu démontrer que le modèle retenu, basé sur l'algorithme *XGBoost*, ouvre des perspectives prometteuses menant vers un provisionnement plus précis de la garantie RC corporelle automobile dès la deuxième année d'inventaire. Néanmoins, afin de conforter le choix de ce modèle, nous proposons les analyses complémentaires suivantes qu'il serait pertinent d'effectuer.

Pour commencer, notre base de données individuelle se limite aux seuls sinistres survenus entre 2010 et 2014, restreignant dès lors les analyses que nous sommes en mesure de réaliser. Ainsi, il serait intéressant d'évaluer l'évolution de la précision du modèle induite par l'apprentissage sur davantage d'années de survenance. Aussi, afin de s'assurer de sa capacité de généralisation, il serait pertinent de tester le modèle sur d'autres années de survenance dont le taux de censure est faible et de vérifier qu'il fournit des résultats satisfaisants.

Par ailleurs, nous limitons dans notre cas les variables explicatives servant à l'apprentissage du modèle aux deux premières années d'inventaire seulement. Ce choix est essentiellement motivé par l'intérêt d'obtenir une estimation précise le plus tôt possible. Il serait intéressant d'analyser le gain en précision induit par l'élargissement des variables explicatives aux trois premières années d'inventaire. La marge de progression ainsi mesurée permettrait d'identifier l'intérêt d'un tel changement.

Afin de mieux comprendre la fluctuation des estimations de la méthode AXA France, il aurait été intéressant de disposer des bases individuelles correspondantes aux triangles agrégés. Une meilleure compréhension de ces données accompagnée du retraitement des triangles permettrait d'identifier ou d'exclure des causes de fluctuations afin de gagner, à terme, en stabilité de prédiction de la charge ultime.

Ensuite, pour corriger le biais induit par la censure, nous avons eu recours aux poids *IPCW* calculés à partir de l'estimateur de Kaplan-Meier. Ce calcul s'effectue au travers de la modélisation de la variable durée de vie du sinistre, en raison de la corrélation positive considérée entre cette dernière et la charge ultime. Étant donné que la variable d'intérêt est elle même censurée, il serait intéressant de calculer les poids *IPCW* au travers de la modélisation de cette même variable, en ayant potentiellement recours à un autre estimateur, en l'occurrence le modèle de Cox.

Enfin, nous pourrions quantifier l'impact du gain en volatilité sur la valeur du SCR de réserves en utilisant le modèle interne AXA France. Ceci permettrait de valoriser le gain réalisé au travers de l'usage du modèle de *machine learning*.

Conclusion

Évoluant dans le cadre de la branche responsabilité civile corporelle automobile, ce mémoire vise à confronter le modèle de provisionnement mis en place par AXA France, basé sur les méthodes agrégées, en proposant un nouveau modèle adoptant une approche d'exploitation des données individuelles au travers du *machine learning*. Il a en outre pour objectif de fournir une estimation précise de la charge ultime des sinistres tout en minimisant la volatilité de cette estimation.

Pour commencer, nous nous sommes attachés à analyser le modèle mis en place par AXA France afin d'en comprendre le fonctionnement. En effet, ce modèle se base sur les méthodes agrégées, en l'occurrence Mack et Chain Ladder, et segmente les sinistres en différentes tranches de coûts. Ce fractionnement a pour but d'obtenir des ensembles de données davantage homogènes permettant ainsi l'estimation de la charge ultime. Néanmoins, la volatilité de ces estimations s'avère importante, en particulier sur les sinistres graves. Ainsi, l'estimation de la charge ultime (hors marge de prudence) des sinistres survenus en 2014 a progressé de 29 M€, passant de 289 M€ en 2017 à 318 M€ en 2019.

Afin d'améliorer l'estimation des provisions mais également de s'abstraire du risque de non-satisfaction des hypothèses liées aux méthodes agrégées, nous optons pour des modèles de *machine learning*. Nous commençons par en sélectionner trois, basés respectivement sur les algorithmes de CART, des forêts aléatoires et de *XGBoost*. Nous les calibrons ensuite sur les sinistres survenus entre 2010 et 2013 pour enfin les tester sur les sinistres survenus en 2014. De plus, il est à noter que nous pondérons les données à l'aide de poids *IPCW*, calculés au travers de l'estimateur de Kaplan-Meier, afin de corriger le biais induit par la censure de nos données.

Dans un premier temps, la comparaison de la performance des trois modèles sur les sinistres survenus en 2014 et clos à fin 2019 nous permet de constater que le modèle *XGBoost* obtient les meilleurs résultats. En effet, le modèle ne surestime que de $1,5\%$ la charge réelle. Ces estimations constituent un tour de force dans la mesure où elles ne reposent que sur un faible historique de données, en l'occurrence deux années d'inventaire.

Par la suite, la simulation de la distribution empirique de la charge ultime pour chaque modèle, au travers de la méthode *bootstrap* à 10000 itérations, nous permet d'apprécier le gain en performance pour les modèles ensemblistes testés, notamment celui du *XGBoost*. En effet, ce dernier se distingue par une meilleure anticipation de la charge ultime, en particulier pour les sinistres graves : tandis que le modèle AXA France n'estime que 300 M€ en 2015 avant d'aboutir à 319 M€ en 2019, le modèle *XGBoost* obtient 321 M€ dès 2015, autrement dit avec 4 années d'avance.

Aussi, nous notons un gain conséquent en précision, de l'ordre de 37 M€ : tandis que le modèle de Mack obtient une erreur de prédiction estimée à 41 M€ en 2015, le modèle *XGBoost* n'est lui qu'à 4 M€ à la même période.

Dans la continuité des travaux menés dans le cadre de ce mémoire, d'autres axes d'étude peuvent être abordés. Il serait ainsi pertinent de tester le modèle retenu sur différentes survenances afin de conforter les performances obtenues. Aussi, nous pourrions quantifier l'impact du gain en volatilité sur la valeur du *SCR* en utilisant le modèle interne AXA France. Il serait également intéressant d'expérimenter ce modèle sur les sinistres corporels hors automobile. Enfin, dès lors qu'un historique conséquent sera disponible, il pourrait s'agir d'inclure le taux d'AIPP en tant que variable explicative.

S'il est aujourd'hui périlleux de s'avancer sur un éventuel remplacement à court terme des méthodes agrégées utilisées par AXA France en faveur du modèle retenu, compte tenu de leur simplicité d'interprétation et de compréhension par les équipes de contrôle interne et externe, il ne fait nul doute que ce dernier serait à minima amené à être utilisé à titre de comparaison pour challenger le résultat des méthodes actuelles, en remontant notamment des points d'alerte. Pour autant, en cas de performances particulièrement attractives dans la durée, il n'est pas à exclure que ce modèle puisse un jour remplacer le modèle actuel.

Bibliographie

ACPR (2015). Analyse de l'exercice 2014 de préparation à Solvabilité II. fr. Analyses et synthèses. URL : https://acpr.banque-france.fr/sites/default/files/medias/documents/20150223-analyse-de-l-exercice-2014-de-preparation-a-solvabilite-ii_01.pdf.

AISAM-ACME (2007). Reserve risk and risk margin assessment under Solvency II. en. Study on non-life long tail liabilities. URL : https://silo.tips/download/aisam-acme-study-on-non-life-long-tail-liabilities.

Antonio, K. et Plat, R. (2014). Micro-level stochastic loss reserving for general insurance. *Scandinavian Actuarial Journal* 2014.7, p. 649-669.

Arjas, E. (1989). The Claims Reserving Problem in Non-Life Insurance: Some Structural Ideas. *ASTIN Bulletin* 19.2, 139–152.

Barruel, G. et Bougnon, N. (2016). Pilotage du risque de souscription non vie sous Solvabilité II. Mémoire d'actuariat. Paris : CEA.

Baudry, M. et Robert, C. Y. (2019). A machine learning approach for individual claims reserving in insurance. *Applied Stochastic Models in Business and Industry* 35.5, p. 1127-1155.

Benjamin, S. et Eagles, L. (1986). Reserves in Lloyd's and the London market. en. *Journal of the Institute of Actuaries*, p. 197-256.

Bornhuetter, R. et Ferguson, R. (1972). The Actuary and IBNR. en. *Proceedings of the Casualty Actuarial Society* 59, p. 181-195.

Breiman, L. (1996). Bagging predictors. en. *Machine Learning* 24, p. 123-140.

Breiman, L. (1999). Prediction Games and Arcing Algorithms. *Neural Computation* 11.7, p. 1493-1517.

Breiman, L. (2001). Random Forests. en. *Machine Learning* 45, p. 5-32.

Breiman, L., Friedman, J. H., Olshen, R. A. et Stone, C. J. (1984). Classification and Regression Trees. Wadsworth Brooks.

Brigitte, E. (2008). Analyses factorielles simples et multiples : objectifs, méthodes et interprétation / Brigitte Escoffier,... Jérôme Pagès,... fre. 4e édition. Sciences Sup Mathématiques. Dunod.

CCR Re (2019). L'indemnisation des préjudices corporels graves en RC Automobile - France 2019. fr. Livre blanc. URL : https://www.ccr-re.com/documents/20123/54390/Livre%2Bblanc%2BRC%2Bauto%2BFR%2B-%2BCCR%2BRe%2B-%2BWEB.pdf.

Chen, T. et Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *CoRR* abs/1603.02754. arXiv : 1603.02754.

Chen, T. et al. (2021). xgboost: Extreme Gradient Boosting. R package version 1.5.0.2. URL : https://CRAN.R-project.org/package=xgboost.

Code Civil (1804). Articles 1382 à 1383. https://www.legifrance.gouv.fr/codes/id/LEGISCTA000006136352/1804-02-19.

Code des Assurances (2007). Article L211-1. https://www.legifrance.gouv.fr/codes/article_lc/LEGIARTI000017735447/.

Conseil de normalisation des comptes publics (2014). Avis préalable afférent au projet d'arrêté portant création d'une comptabilité auxiliaire du FGAO. https://www.economie.gouv.fr/files/files/directions_services/cnocp/avis/avis_prealable/2014/Avis_prealable_arrete_FGAO_ss_sign_16_janvier_2014.pdf.

Dintilhac, J.-P. (2005). Rapport du groupe de travail chargé d'élaborer une nomenclature des préjudices corporels. fr. Report. Cour de Cassation.

Dupâquier, J. (1976). La table de mortalité d'E. Halley présentée et commentée. *Annales de Démographie Historique*, p. 485-503.

Dupâquier, J. (1996). L'invention de la table de mortalité. De Graunt à Wargentin. *Population*, p. 497-498.

Duval, F. et Pigeon, M. (2019). Individual Loss Reserving Using a Gradient Boosting-Based Approach. *Risks* 7.3.

Efron, B. (1982). The Jackknife, the Bootstrap and Other Resampling Plans. Society for Industrial et Applied Mathematics.

Efron, B. et Tibshirani, R. (1993). An Introduction to the Bootstrap. Chapman et Hall, New York.

England, P. et Verrall, R. (2002). Stochastic Claims Reserving in General Insurance. en. *British Actuarial Journal* 8, p. 443-518.

England, P. et Verrall, R. (1999). Analytic and bootstrap estimates of prediction errors in claims reserving. en. *Insurance: Mathematics and Economics* 25.3, p. 281-293.

France Assureurs (2021a). L'assurance française : Données clés 2020. fr. Rapport annuel. URL : https://www.franceassureurs.fr/wp-content/uploads/VF-Donnees-cles-2020.pdf.

France Assureurs (2021b). Les assureurs, acteurs de la relance durable. fr. Dossier de presse. URL : https://www.franceassureurs.fr/wp-content/uploads/2021/11/dossier-conference-presse-20210324.pdf.

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis* 38, p. 367-378.

Guibert, Q. et Planchet, F. (2017). Utilisation des estimateurs de Kaplan-Meier par génération et de Hoëm pour la construction de tables de mortalité prospectives. *Bulletin trimestriel de l'Institut des actuaires français* 17.33, p. 5-24.

Haastrup, S. et Arjas, E. (1996). Claims Reserving in Continuous Time; A Nonparametric Bayesian Approach. *ASTIN Bulletin* 26.2, 139–164.

Hainaut, D. et Robert, C. Y. (2014). Credit risk valuation with rating transitions and partial information. *International Journal of Theoretical and Applied Finance* 17.07.

Hertig, J. (1985). A Statistical Approach to IBNR-Reserves in Marine Reinsurance. *ASTIN Bulletin* 15.2, 171–183.

Hotelling, H. (1933). Analysis of a complex statistical variables into principal components. en. *Journal of Educational Psychology*, p. 417-441.

Index Assurance (2020). Seuil d'AIPP : définition. fr. Index Assurance [en ligne]. url : `https://www.index-assurance.fr/dictionnaire/seuil-aipp` (visité le 11/08/2021).

Index Assurance (2021). Convention IRCA. fr. Index Assurance [en ligne]. url : `https://www.index-assurance.fr/pratique/sinistre/convention-irca` (visité le 12/10/2021).

Jamal, S. (2018). Machine Learning and Traditional Methods Synergy in Non-Life Reserving. en. Report. ASTIN.

Kaplan, E. L. et Meier, P. (1958). Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association* 53.282, p. 457-481.

Kuo, K. (2019). DeepTriangle: A Deep Learning Approach to Loss Reserving. *Risks* 7.3.

Lambert-Faivre, Y. (2003). Rapport sur l'indemnisation du dommage corporel. fr. Report. Conseil National d'Aide aux Victimes (CNAV).

Larsen, C. R. (2007). An Individual Claims Reserving Model. *ASTIN Bulletin* 37.1, 113–132.

Le Faou, Y. (2019). Contributions à la modélisation des données de durée en présence de censure : application à l'étude des résiliations de contrats d'assurance santé. Theses. Sorbonne Université. url : `https://tel.archives-ouvertes.fr/tel-03017164`.

Lingibé, P. (2019). Dommage corporel : comment est-il indemnisé ? fr. Village de la justice [en ligne]. url : `https://www.village-justice.com/articles/dommage-corporel-comment-est-indemnise,30361.html` (visité le 19/07/2021).

Lopez, O., Milhaud, X. et Thérond, P.-E. (2016). Tree-based censored regression with applications in insurance. *Electronic Journal of Statistics* 10.2, p. 2685-2716.

Lopez, O., Milhaud, X. et Thérond, P.-E. (2019). A tree-based algorithm adapted to microlevel reserving and long development claims. *ASTIN Bulletin* 49.3, p. 741-762.

Mack, T. (1993a). Distribution-free calculation of the standard error of Chain-Ladder reserve estimates. en. *ASTIN Bulletin* 23.2, p. 213-225.

Mack, T. (1993b). Measuring the Variability of Chain Ladder Reserve Estimates. en. *CAS Prize Paper Competition*, p. 101-182.

Malignac, G. (1978). La réglementation des indexations. *Journal de la société statistique de Paris* 119.2, p. 140-149.

Merz, M. et Wüthrich, M. V. (2008). Stochastic Claims Reserving Methods in Insurance. The Wiley Finance Series. Wiley.

Milhaud, X. (2013). Exogenous and endogenous risk factors management to predict surrender behaviours. *ASTIN Bulletin* 43.3, p. 373-398.

Molinaro, A., Dudoit, S. et Laan, M. (2004). Tree-based Multivariate Regression and Density Estimation with Right-Censored Data. *Journal of Multivariate Analysis* 90.1, p. 154-177.

Murphy, D. (1994). Unbiased loss development factors. en. *Proceedings of the Casualty Actuarial Society*, p. 154-222.

Norberg, R. (1993). Prediction of Outstanding Liabilities in Non-Life Insurance. *ASTIN Bulletin* 23.1, 95–115.

Pigeon, M., Antonio, K. et Denuit, M. (2013). Individual Loss Reserving with the multivariate skew normal framework. *ASTIN Bulletin* 43.3, 399–428.

Quarg, G. et Mack, T. (2004). Munich chain ladder. en. *Blätter der DGVFM* 26, p. 597-630.

R Core Team (2022). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. URL : https://www.R-project.org/.

Rose, N. (s. d.). L'histoire de l'assurance automobile obligatoire en france. fr. Ifpass [en ligne]. URL : https://www.ifpass.fr/leblogdesexperts/lhistoire-de-lassurance-automobile-obligatoire-en-france (visité le 15/09/2021).

Servel, M. (2020). Une approche individuelle du provisionnement des sinistres corporels automobiles. Mémoire d'actuariat. Paris : ISFA.

Straub, E. (1988). Non-Life Insurance Mathematics. Association of Swiss Actuaries. Springer Verlag.

Taylor, G. C. (1977). Separation of Inflation and other Effects from the Distribution of Non-Life Insurance Claim Delays. *ASTIN Bulletin* 9.1-2, p. 219-230.

Texte de loi (1974). Loi n° 74-1118 du 27 décembre 1974 relative à la revalorisation de certaines rentes allouées en réparation du préjudice causé par un véhicule terrestre à moteur. https://www.legifrance.gouv.fr/loda/id/JORFTEXT000000333535/.

Texte de loi (1985). Loi n° 85-677 du 5 juillet 1985 tendant à l'amélioration de la situation des victimes d'accidents de la circulation et à l'accélération des procédures d'indemnisation. https://www.legifrance.gouv.fr/loda/id/LEGITEXT000006068902/.

Texte de loi (2007). Arrêté du 27 mars 2007 relatif aux conditions d'élaboration des statistiques relatives aux accidents corporels de la circulation. https://www.legifrance.gouv.fr/loda/id/LEGITEXT000006055936/.

Therneau, T. et Atkinson, B. (2019). rpat : Recursive Partitioning and Regression Trees. R package version 4.1-15. URL : https://CRAN.R-project.org/package=rpart.

Van der Laan, M. et Robins, J. (2003). Unified Methods for Censored Longitudinal Data and Causality. Springer series in statistics. 3Island Press.

Verbeek, H. G. (1972). An approach to the analysis of claims experience in motor liability excess of loss reinsurance. *ASTIN Bulletin* 6.3, p. 195-202.

Vock, D. M., Wolfson, J., Bandyopadhyay, S., Adomavicius, G., Johnson, P. E., Vazquez-Benitez, G. et O'Connor, P. J. (2016). Adapting machine learning techniques to censored time-to-event health record data: A general-purpose approach using inverse probability of censoring weighting. *Journal of Biomedical Informatics* 61, p. 119-131.

Wikistat (2016a). Agrégation de modèles. [En ligne; Page disponible le 13-octobre-2021]. URL : https://www.math.univ-toulouse.fr/~besse/Wikistat/pdf/st-m-app-agreg.pdf.

Wikistat (2016b). Arbres binaires de décision. [En ligne; Page disponible le 28-septembre-2021]. URL : https://www.math.univ-toulouse.fr/~besse/Wikistat/pdf/st-m-app-cart.pdf.

Wright, M. N. et Ziegler, A. (2017). ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software* 77.1, p. 1-17.

Wu, C. (2016). Solvabilité II : Impact et analyse de sensibilité sur le besoin en fonds propres des produits d'assurance décès. Mémoire d'actuariat. Paris : ISUP.

Wüthrich, M. V. (2016). Machine Learning in Individual Claims Reserving. 16-67.

Zhao, X. B. et Zhou, X. (2010). Applying copula models to individual claim loss reserving methods. English. *Insurance: Mathematics and Economics* 46.2, p. 290-299.

Zhao, X. B., Zhou, X. et Wang, J. L. (2009). Semiparametric model for prediction of individual claim loss reserving. *Insurance: Mathematics and Economics* 45.1, p. 1-8.

Zhu, R., Zeng, D. et Kosorok, M. (avr. 2015). Reinforcement Learning Trees. *Journal of the American Statistical Association* 110, p. 0-0.

A Annexes

A.1 Évolution des variations de fréquences des sinistres par garantie en assurance automobile

Variations	2016	2017	2018	2019	2020
Responsabilité civile (RC)	**-0,3%**	**-1,8%**	**-4,3%**	**-1,5%**	**-25,8%**
dont RC matériel	-0,6%	-1,7%	-4,2%	-1,6%	-25,7%
dont RC corporel	2,7%	-2,9%	-5,7%	-1,2%	-26,8%
Dommages tous accidents	**0,7%**	**-1,1%**	**-1,8%**	**-1,0%**	**-22,8%**
Bris de glace	**-3,9%**	**-1,0%**	**2,2%**	**-6,3%**	**-13,2%**
Vol	**-9,5%**	**-7,8%**	**-8,1%**	**-4,2%**	**-15,3%**

TABLE A.1 – Évolution des variations de fréquences des sinistres par garantie en assurance automobile - Particuliers et professionnels - (FRANCE ASSUREURS, 2021a)

A.2 Liste des catégories ministérielles en assurance non-vie

— 20 - dommages corporels (contrats individuels), y compris garanties accessoires
— 21 - dommages collectifs (contrats collectifs), y compris garanties accessoires
— 22 - automobile (responsabilité civile), segmentée en RC corporels et RC matériels (art. R 331-26)
— 23 - automobile (dommages)
— 24 - dommages aux biens des particuliers
— 25 - dommages aux biens professionnels
— 26 - dommages aux biens agricoles
— 27 - catastrophes naturelles
— 28 - responsabilité civile générale
— 29 - protection juridique
— 30 - assistance
— 31 - pertes pécuniaires diverses
— 34 - transports
— 35 - assurance construction (dommages)
— 36 - assurance construction (responsabilité civile)
— 37 - crédit
— 38 - caution
— 39 - acceptations en réassurance (non-vie)

A.3 Dictionnaire des variables

Variable	Type	Description
DTSURV	Date	Date de survenance
DTOUV	Date	Date d'ouverture
DTCLOT	Date	Date de clôture
DTREOUV	Date	Date de réouverture
surv	Numérique	Année de survenance
RGPRUP_OUV	Numérique	Montant de règlements en principal à l'ouverture de la garantie
RGFRUP_OUV	Numérique	Montant de règlements en frais à l'ouverture de la garantie
EVRGUP_OUV	Numérique	Montant de réserves D/D à l'ouverture de la garantie
RECEUP_OUV	Numérique	Montant des recours encaissés à l'ouverture de la garantie
EVREUP_OUV	Numérique	Montant des estimations de recours à encaisser à l'ouverture de la garantie
reg_ouv	Numérique	Somme de RGPRUP_OUV, RGFRUP_OUV et EVRGUP_OUV
res_ouv	Numérique	Somme de RECEUP_OUV et EVREUP_OUV
chg_ouv	Numérique	Montant de la charge nette de recours à l'ouverture de la garantie
top_ouv	Caractère	Classe de la variable chg_ouv ()
XXXXX_&MOIS._S&N.	NA	Evaluation de la variable XXXXX pour la vision de mois &MOIS. et d'année &N.
chg	Numérique	Montant de la charge nette de recours à fin 12/2019
LossCauseDetail	Numérique	Cause détaillée du sinistre
litige	Booléen	Presence d'un litige
NBRQUART	Numérique	Taux de responsabilité de l'assuré
dep	Caractère	Département de survenance du sinistre
duree	Caractère	Variable définissant si le sinistre est clos en 2 ans ou plus
ETATUP	Caractère	Etat de la garantie à fin 12/2019 (clos avec suite/ clos sans suite/en cours)
FLAG_RENTE	Caractère	Sinistre liquidé en rente : vrai ou faux
PASSAGE_RENTE	Booléen	Date de passage en rente
CorpoNumberOfInjureds	Numérique	Nombre de victimes corporelles
NoLumpSumInSubrogationFlagText	Booléen	Montants évalués au forfait
nvict_&N.	Numérique	Numéro de la &N. ème victime
lesion_&N.	Numérique	Lésion de la &N. ème victime
blesgrav_&N.	Numérique	Gravité des bléssures de la &N. ème victime
qualite_&N.	Caractère	Qualité de l'usager de la route (exemple : piéton)
contex_&N.	Caractère	Contexte

FIGURE A.1 – Dictionnaire des variables

A.4 Distribution de la charge D/D nette de recours par année de survenance

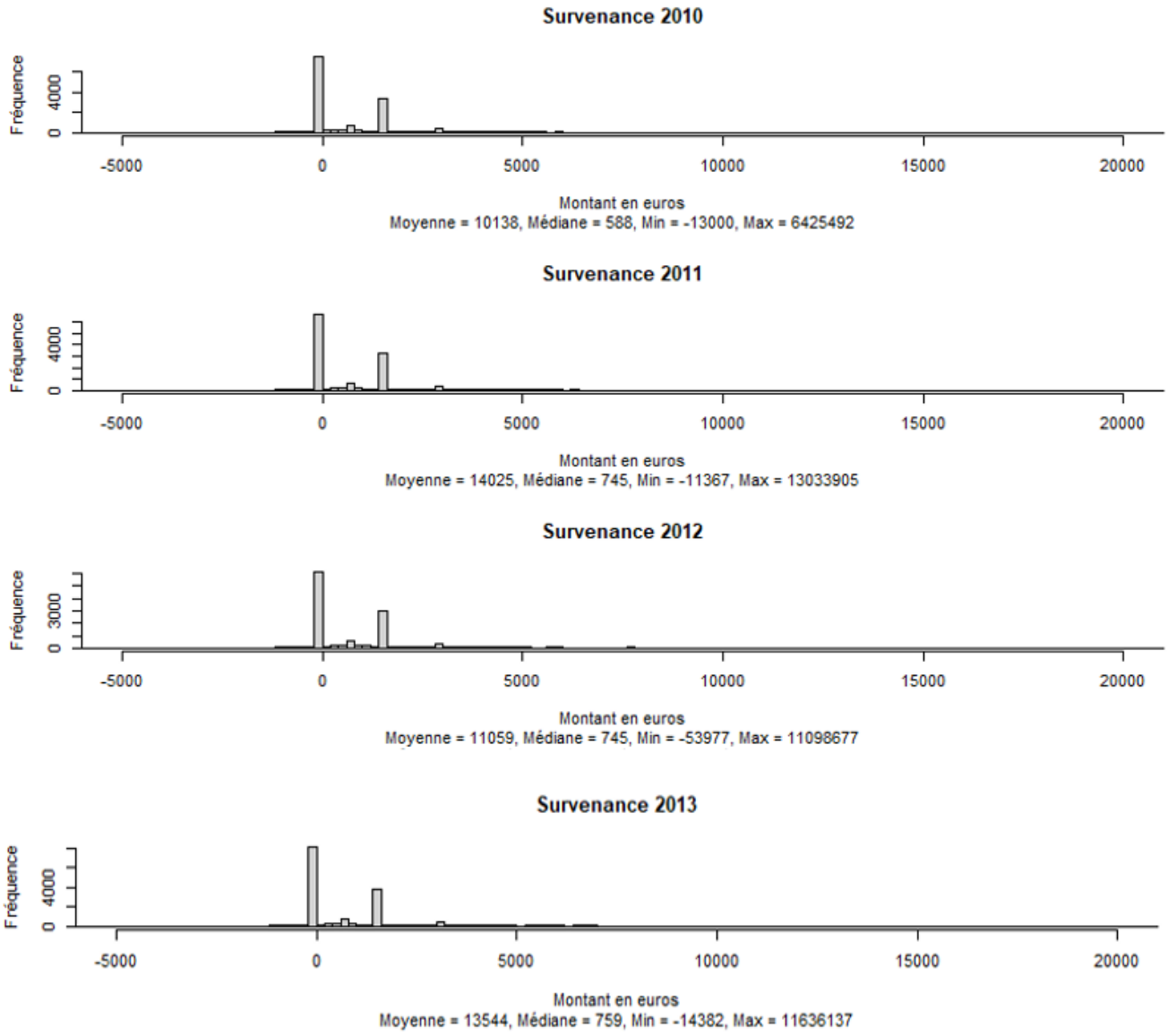


FIGURE A.2 – Distribution de la charge D/D nette de recours par année de survenance - vision à fin 12/2019 - *Zoom* sur la charge $\in [-5000€, 20000€]$

A.5 Matrice de corrélation de spearman

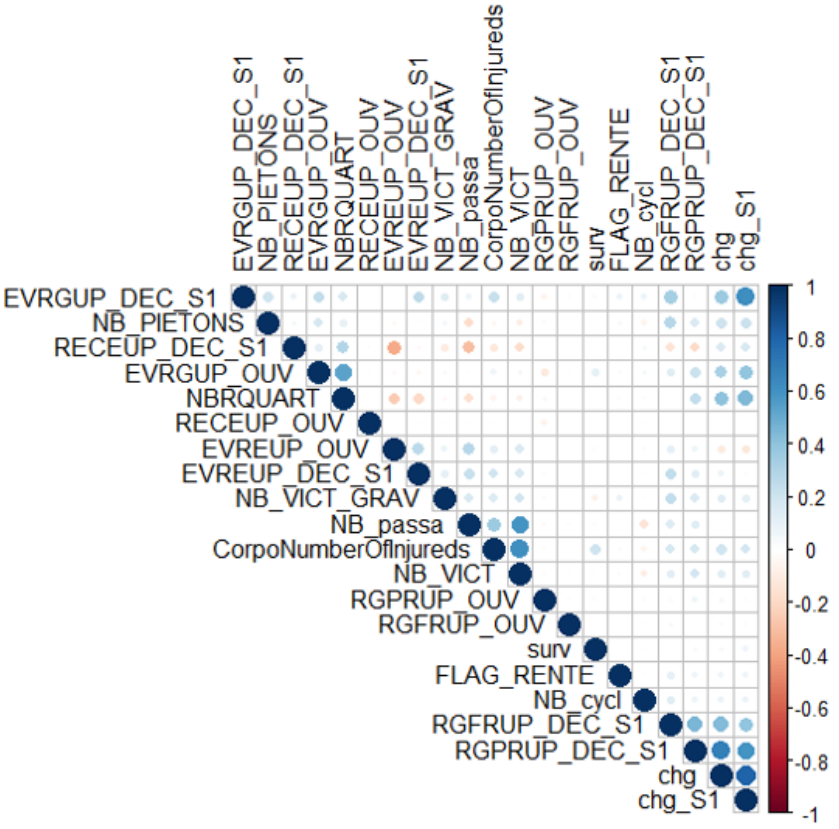


FIGURE A.3 – Matrice de corrélation de Spearman

A.6 Panorama des principales méthodes de provisionnement à l'échelle mondiale

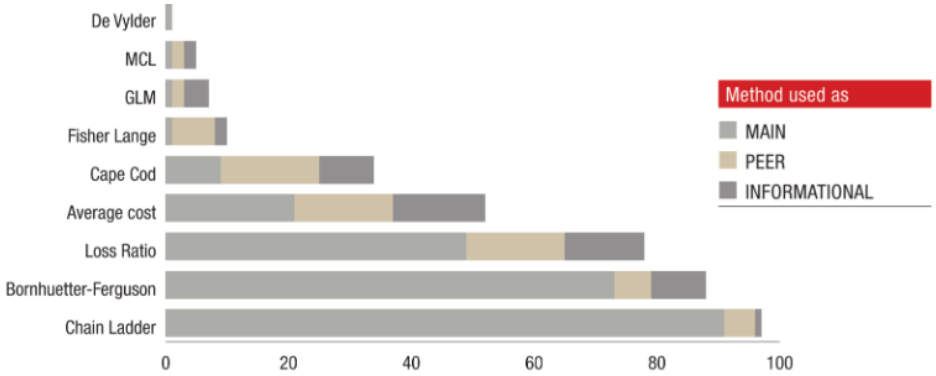


FIGURE A.4 – Panorama des principales méthodes de provisionnement déterministes utilisées *(source : rapport (JAMAL, 2018))*

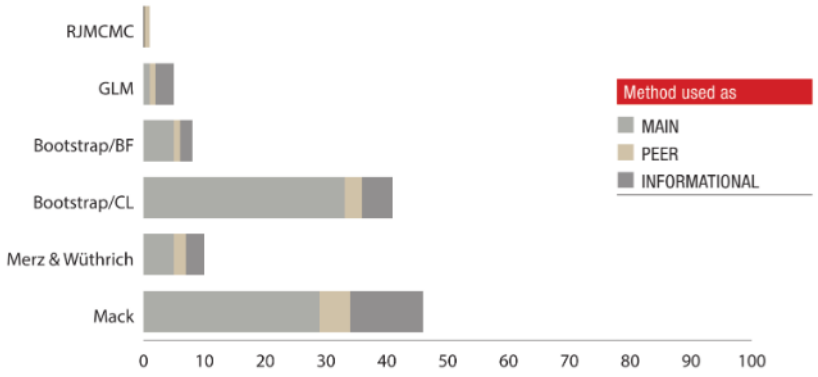


FIGURE A.5 – Panorama des principales méthodes de provisionnement stochastiques utilisées *(source : rapport (JAMAL, 2018))*

A.7 Triangles de charges D/D cumulées nettes de recours par tranche de coûts (en k€)

Année de survenance	Année de développement															
	N	N+1	N+2	N+3	N+4	N+5	N+6	N+7	N+8	N+9	N+10	N+11	N+12	N+13	N+14	N+15
1999	192 856	224 945	236 012	234 018	236 972	233 681	234 580	233 296	236 041	242 232	244 414	243 278	242 778	242 742	242 718	243 459
2000	208 199	239 351	228 910	230 239	238 233	240 909	244 563	252 175	253 201	259 973	262 146	262 240	263 734	267 319	269 353	269 720
2001	167 064	204 045	206 805	220 592	228 388	227 387	228 935	232 400	233 824	235 603	236 499	240 005	240 490	240 144	239 686	
2002	173 857	200 102	212 973	220 252	228 305	225 209	224 394	230 204	233 345	238 153	243 481	247 475	246 324	246 900		
2003	157 265	178 667	197 149	205 298	208 905	204 011	203 816	205 351	204 462	204 945	205 589	213 722	218 105			
2004	165 265	186 579	197 963	206 777	205 878	212 029	216 468	217 287	221 708	219 440	218 235	219 264				
2005	157 866	186 458	204 554	206 893	209 587	211 471	213 321	217 488	218 963	213 678	216 379					
2006	131 104	176 759	191 207	197 163	206 145	209 990	221 402	220 342	218 512	216 666						
2007	129 976	173 707	197 684	198 575	201 189	209 188	212 186	212 747	214 337							
2008	127 304	175 425	190 442	202 836	213 396	217 006	203 748	202 498								
2009	131 139	196 596	219 166	244 337	245 179	241 903	241 673									
2010	116 593	187 198	221 301	231 415	219 247	223 701										
2011	133 207	223 538	247 837	264 037	263 141											
2012	162 195	211 839	208 702	212 529												
2013	209 385	214 669	219 826													
2014	230 890	230 563														

FIGURE A.6 – T0 : Triangle de charges D/D (vision fin 12/2015)

Année de survenance	Année de développement															
	N	N+1	N+2	N+3	N+4	N+5	N+6	N+7	N+8	N+9	N+10	N+11	N+12	N+13	N+14	N+15
1999	155 059	149 895	144 732	140 265	135 625	133 655	132 384	128 384	129 069	128 950	128 886	128 990	128 843	129 378	129 636	129 403
2000	141 338	149 292	130 886	125 413	122 589	119 595	117 384	117 635	116 915	117 006	116 882	116 745	116 906	116 770	116 874	116 933
2001	127 839	129 786	122 946	119 835	116 774	110 215	108 729	109 134	109 421	108 953	108 985	109 181	108 851	108 924	108 946	
2002	129 232	117 024	114 313	108 663	107 094	103 223	102 392	101 154	101 329	100 900	100 513	100 503	100 447	100 331		
2003	111 461	97 676	92 412	90 500	87 811	83 868	83 626	84 208	83 279	84 304	84 566	84 784	84 581			
2004	106 063	95 376	92 558	91 249	86 676	86 071	85 332	85 840	86 444	86 550	86 343	86 396				
2005	100 701	90 111	92 812	91 186	89 427	88 803	88 756	88 136	88 962	89 484	89 316					
2006	83 528	86 124	86 718	87 852	87 785	87 229	86 855	86 559	86 231	85 812						
2007	88 580	88 123	89 403	89 363	90 905	90 625	89 399	88 516	88 700							
2008	88 460	90 810	90 608	91 819	92 084	91 481	91 683	90 956								
2009	94 388	92 395	93 494	94 189	93 304	92 623	94 456									
2010	87 001	84 452	86 534	85 018	84 556	85 942										
2011	82 454	85 606	85 019	84 761	84 810											
2012	93 863	84 062	78 565	81 217												
2013	101 303	87 512	82 233													
2014	140 550	104 725														

FIGURE A.7 – T12 : Triangle de charges D/D $\leq$ 150 k€ (vision fin 12/2015)

Année de survenance	Année de développement															
	N	N+1	N+2	N+3	N+4	N+5	N+6	N+7	N+8	N+9	N+10	N+11	N+12	N+13	N+14	N+15
1999	20 504	39 990	48 398	49 384	55 417	55 700	55 226	58 397	57 558	57 824	58 406	58 677	58 467	58 412	57 202	57 156
2000	36 097	46 612	51 404	52 811	54 711	56 411	56 707	57 651	57 601	58 349	57 922	56 454	55 950	56 825	57 162	55 761
2001	25 615	48 324	52 148	53 502	52 326	52 254	53 373	54 030	53 746	55 519	54 875	53 503	54 437	54 322	54 332	
2002	20 874	41 398	47 405	51 955	51 860	53 354	52 370	52 814	52 326	51 605	50 259	49 694	49 743	49 228		
2003	21 812	39 723	48 522	52 707	54 818	56 230	54 140	54 002	54 551	52 711	53 430	50 765	50 358			
2004	17 678	37 107	43 925	48 132	46 716	46 567	47 101	47 449	48 129	47 562	47 624	47 455				
2005	23 595	47 349	53 960	55 157	61 986	60 088	59 806	61 119	59 993	59 062	58 579					
2006	20 978	39 930	46 531	50 520	48 628	49 832	48 858	49 854	50 304	50 290						
2007	23 359	45 712	55 613	57 515	54 306	54 131	55 377	54 123	53 559							
2008	19 448	39 595	43 873	50 311	53 768	53 452	53 238	53 666								
2009	19 517	47 224	54 822	58 741	61 785	61 826	60 907									
2010	17 623	41 152	58 554	60 607	62 287	63 277										
2011	21 330	45 936	49 564	55 601	56 702											
2012	25 005	50 805	54 295	57 894												
2013	23 519	41 557	48 104													
2014	13 946	35 259														

FIGURE A.8 – T3 : Triangle de charges D/D comprises entre 150 k€ et 750 k€ (vision fin 12/2015)

Année de survenance	Année de développement															
	N	N+1	N+2	N+3	N+4	N+5	N+6	N+7	N+8	N+9	N+10	N+11	N+12	N+13	N+14	N+15
1999	17 293	35 061	42 881	44 368	45 929	44 326	46 970	46 515	49 415	55 459	57 123	55 612	55 469	54 951	55 879	56 900
2000	30 763	43 447	46 620	52 015	60 933	64 904	70 472	76 890	78 686	84 617	87 342	89 041	90 878	93 724	95 317	97 027
2001	13 610	25 935	31 712	47 255	59 289	64 919	66 834	69 236	70 657	71 130	72 639	77 321	77 202	76 897	76 407	
2002	23 751	41 680	51 255	59 634	69 351	68 632	69 633	76 236	79 690	85 648	92 709	97 277	96 134	97 341		
2003	23 992	41 268	56 216	62 092	66 276	63 913	66 050	67 141	66 632	67 931	67 592	78 173	83 166			
2004	41 523	54 097	61 481	67 397	72 486	79 392	84 035	83 998	87 135	85 327	84 268	85 413				
2005	33 570	48 998	57 782	60 550	58 174	62 580	64 759	68 233	70 009	65 132	68 484					
2006	26 598	50 705	57 957	58 791	69 732	72 929	85 689	83 929	81 976	80 565						
2007	18 036	39 873	52 668	51 696	55 978	64 432	67 411	70 108	72 079							
2008	19 396	45 020	55 960	60 707	67 543	72 072	58 828	57 876								
2009	17 235	56 978	70 850	91 407	90 090	87 454	86 310									
2010	11 969	61 593	76 212	85 790	72 404	74 482										
2011	29 423	91 996	113 254	123 674	121 629											
2012	43 327	76 972	75 842	73 418												
2013	84 564	85 599	89 490													
2014	76 394	90 579														

FIGURE A.9 – T45 : Triangle de charges D/D $\geq$ 750 k€ (vision fin 12/2015)

A.8 Test d'hypothèses du modèle de Mack

Triangle T12							
j	S_j	L_j	Z_j	n	m	E_Zj	Var_Zj
2	1	1	1	2	0	0.5	0.3
3	1	2	1	3	1	0.8	0.2
4	2	2	2	4	1	1.3	0.4
5	4	1	1	5	2	1.6	0.4
6	4	1	1	5	2	1.6	0.4
7	7	0	0	7	3	2.4	0.6
8	7	0	0	7	3	2.4	0.6
9	4	5	4	9	4	3.3	0.7
10	6	3	3	9	4	3.3	0.7
11	3	8	3	11	5	4.1	0.9
12	4	7	4	11	5	4.1	0.9
13	4	8	4	12	5	4.6	1.2
14	3	11	3	14	6	5.5	1.3
15	6	9	6	15	7	5.9	1.3
16	7	8	7	15	7	5.9	1.3
E_Z = 47.3				Var_Z = 11.1			

Triangle T3							
j	S_j	L_j	Z_j	n	m	E_Zj	Var_Zj
2	1	1	1	2	0	0.5	0.3
3	1	2	1	3	1	0.8	0.2
4	4	0	0	4	1	1.3	0.4
5	3	2	2	5	2	1.6	0.4
6	4	2	2	6	2	2.1	0.6
7	2	5	2	7	3	2.4	0.6
8	3	5	3	8	3	2.9	0.8
9	3	6	3	9	4	3.3	0.7
10	5	5	5	10	4	3.8	1.0
11	4	6	4	10	4	3.8	1.0
12	7	4	4	11	5	4.1	0.9
13	5	6	5	11	5	4.1	0.9
14	6	8	6	14	6	5.5	1.3
15	6	9	6	15	7	5.9	1.3
16	8	7	7	15	7	5.9	1.3
E_Z = 47.9				Var_Z = 11.7			

Triangle T45							
j	S_j	L_j	Z_j	n	m	E_Zj	Var_Zj
2	0	2	0	2	0	0.5	0.3
3	2	0	0	2	0	0.5	0.3
4	3	1	1	4	1	1.3	0.4
5	4	1	1	5	2	1.6	0.4
6	2	4	2	6	2	2.1	0.6
7	1	6	1	7	3	2.4	0.6
8	3	5	3	8	3	2.9	0.8
9	5	4	4	9	4	3.3	0.7
10	6	3	3	9	4	3.3	0.7
11	5	5	5	10	4	3.8	1.0
12	5	7	5	12	5	4.6	1.2
13	7	5	5	12	5	4.6	1.2
14	1	13	1	14	6	5.5	1.3
15	5	7	5	12	5	4.6	1.2
16	11	5	5	16	7	6.4	1.5
E_Z = 47.4				Var_Z = 12.1			

TABLE A.2 – Test des effets calendaires - Méthode de calcul détaillée

A.9 Optimisation de l'hyper-paramètre *nrounds* - *XGBoost*

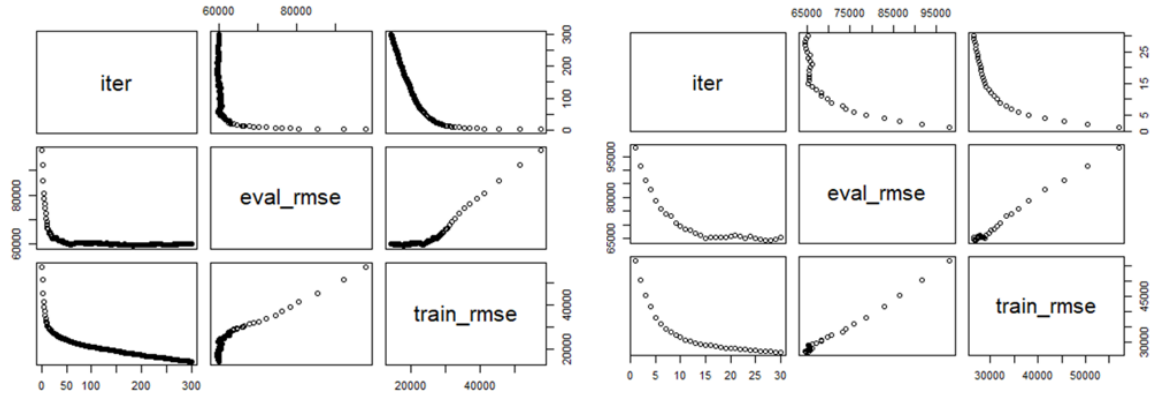


FIGURE A.10 – Optimisation de l'hyper-paramètre *nrounds* - *XGBoost*