IRM | institut du risk management BY Institut.

Mémoire présenté pour la validation de la Formation
« Certificat d'Expertise Actuarielle »
de l'Institut du Risk Management
et l'admission à l'Institut des actuaires
le

Par : LEJEUNE Marie

Titre : Evaluation de la charge ultime aviation et maritime par des méthodes de Machine Learning: Impact des données sur les résultats observés.

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

Membres présents du jury de l'Institut des actuaires :

Membres présents du jury de l'Institut du Risk Management :

Secrétariat :

Bibliothèque :

Entreprise : La Réunion Aérienne

Nom : FOATA Stéphanie

Signature et Cachet :

Directeur de mémoire en entreprise :

Nom : FOATA Stéphanie

Signature :

Invité :

Nom :

Signature :

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

Signature du responsable entreprise

Signature(s) du candidat(s)

Résumé

En assurance, la charge ultime, calculée lors du provisionnement, est une notion fondamentale pour l'entreprise. Celle-ci doit être calculée avec soin, puisque le *Best Estimate*, part importante lors du calcul du risque de réserve sous solvabilité 2 en découle. Cette charge ultime, est généralement obtenue par des méthodes actuarielles déterministes, du type *Chain Ladder*. Cette méthode est particulièrement adaptée aux assurances de masses, telles que l'automobile. Cependant, les limites du modèle sont atteintes lorsque les données sont moins nombreuses, plus volatiles et irrégulières dans le temps, telles que les sinistres aviation et maritime.

Aujourd'hui, les entreprises d'assurance ont à disposition de nombreuses données concernant les sinistres et les assurés, sur un historique de plusieurs années. Cette grande volumétrie, octroie de nouveaux outils aux actuaires et leur permet de se tourner vers de nouvelles méthodes pour la tarification, mais également en provisionnement. De ce fait, de nouvelles méthodes de provisionnement ont vu le jour ces dernière années, dont de nombreuses basées sur des méthodes ligne à ligne grâce au *Machine Learning*.

De ce fait, dans ce mémoire, nous allons appliquer des méthodes de provisionnement ligne à ligne grâce à plusieurs modèles de partitionnement récursifs basés sur la méthode CART (*Classification And Regression Trees*) définie par Breiman L. et al. (1984), ainsi qu'un modèle de régression pénalisée, *Elastic Net,* définie par Zou H. et Hastie T. (2004), afin de prédire la charge ultime d'un exercice de souscription, à partir des informations disponibles à sa deuxième année de développement. Pour tester la capacité de prédiction de ces modèles, ces derniers sont appliqués à deux portefeuilles de la même *Line of Business*, Marine Aviation Transport, une branche contenant des portefeuilles tous deux plus ou moins volatils et plus ou moins sinistrés, sur lesquels la méthode du *Chain Ladder* nous montre ses limites. La notion de censure, très importante lors de l'application de ces méthodes est abordée, grâce à l'introductions des poids IPCW (*Inverse Probability of Censoring Weighting*).

A l'issue de ces tests, les deux prédictions obtenues sont confrontées à leur résultat *Chain Ladder* respectifs, afin d'identifier les modèles les plus précis, et les raisons des différences en termes de qualité de résultats.

Mots clés : Charge ultime, *Best Estimate*, Censure, IPCW, Kaplan Meier, CART, Forêts aléatoires, *Boosting* de gradient, *XGBoost*, *Elastic Net*, Aviation, Maritime.

Abstract

In insurance, the ultimate loss cost, calculated during reserving, is a fundamental concept for the company. It must be calculated with care, since the Best Estimate, an important part when calculating the reserving risk under solvency 2 results from it. This ultimate charge is generally obtained by deterministic actuarial methods, like Chain Ladder. This method is particularly suitable for mass insurance, such as auto. However, the limits of the model are reached when the data are fewer, more volatile and irregular over time, such as aviation and marine claims.

Today, insurance companies have access to a lot of data concerning claims and the insured, over a history of several years. This large volume provides new tools to actuaries and allows them use new methods for pricing, but also for reserving. As a result, new reserving methods have emerged in recent years, many of which are based on line-by-line methods using Machine Learning.

Therefore, in this dissertation, we will apply line-by-line reserving methods using several recursive partitioning models based on the CART (Classification And Regression Trees) method defined by Breiman L. et al. (1984), as well as a penalized regression model, Elastic Net, defined by Zou H. and Hastie T. (2004), in order to predict the ultimate loss cost of an underwriting year, from the information available at its second year of development. To test the level of prediction of these models, they are applied to two portfolios of the same Line of Business, Marine Aviation Transport, a branch containing portfolios both more or less volatile and with more or less claims, on which the method of Chain Ladder shows us its limits. The notion of censorship, which is very important when applying these methods, is addressed, thanks to the introduction of IPCW weights (Inverse Probability of Censoring Weighting).

At the end of these tests, the two predictions obtained are compared with their respective Chain Ladder results, in order to identify the most accurate models, and the reasons for the differences in terms of quality of results are explained.

Key words: Ultimate loss cost, Best Estimate, Censorship, IPCW, Kaplan Meier, CART, Random Forest, Gradient boosting, XGBoost, Elastic Net, Aviation, Maritime.

Note de synthèse

1 CONTEXTE ET OBJECTIF

En assurance, la charge ultime ainsi que le *Best Estimate* (BE), représentant respectivement le montant de charge sinistre totale obtenue, calculée lors du provisionnement en assurance non vie, et la somme des flux futurs probabilisés et actualisés, sont toutes deux des données cruciales lors du calcul du *Solvency Capital Requirement* non vie, ou SCR non vie. Le SCR représente le capital cible nécessaire à l'entreprise afin d'assurer sa solvabilité, compte tenu de son activité.

En effet, le BE fait partie intégrante des provisions techniques, ces dernières étant particulièrement importantes dans la règlementation prudentielle de Solvabilité 2. Les provisions techniques doivent être objectives, prudentes et fiables. Le *Best Estimate* et la charge ultime se doivent donc d'être calculées avec soin.

Ce provisionnement, effectué une à plusieurs fois dans l'année, est fréquemment appliqué de manière conventionnelle, à l'aide de la méthode du *Chain Ladder*, une méthode actuarielle déterministe, rapide à appliquer et simple à mettre en place au sein des entreprises, reposant sur l'agrégation de la sinistralité antérieure en triangles de liquidation, ou triangles de *run-off*. Cependant, cette méthode n'est pas toujours adéquate, celle-ci reposant sur deux hypothèses peu fréquemment vérifiées.

En plus de ces hypothèses, le portefeuille étudié doit contenir des données fiables, nombreuses, le passé doit être régulier, et étudié sur des branches peu volatiles. Une variation de portefeuille au cours du temps implique que les ratios de développement obtenus grâce à la sinistralité passée, ne sont plus représentatifs du portefeuille actuel et donc du développement de la sinistralité future. Les portefeuilles aviation et maritime, qui seront étudiés tout au long de ce mémoire, sont typiques d'une sinistralité faible en termes de nombres, mais volatile. Cependant, le portefeuille maritime contient plus de données, et malgré une volatilité élevée, celle-ci reste moins forte que sur le portefeuille aviation.

Par conséquent, notre objectif dans ce mémoire sera d'exposer la méthode du *Chain Ladder* puis la *challenger* sur l'exercice de souscription 2020 en aviation ainsi qu'en maritime, par des méthodes de *Machine Learning,* permettant le provisionnement ligne à ligne. Ces méthodes permettent de s'affranchir des hypothèses liées au *Chain Ladder*, et d'ajouter des informations qualitatives et quantitatives concernant les sinistres et les assurés. Les différences et similitudes entre les bases étudiées seront décrites, puis les méthodes CART (*Classification And Regression Tree*), forêts aléatoires, *boosting* de gradient*, XGBoost*, ainsi que la méthode *Elastic Net*, seront appliquées. Les deux bases utilisées nous permettrons en plus de vérifier la fiabilité des modèles, d'apprécier les différences de résultats selon le type de base étudiée.

Le but de la méthode CART est de créer un modèle, permettant de prédire la valeur d'une variable cible, à partir de plusieurs variables d'entrées. L'arbre est construit par répartition récursive. Cette méthode est simple et rapide en termes d'exécution mais peu fiable.

Pour les forêts aléatoires, le principe se base sur le modèle précédent. En effet, N arbres de régression sont calibrés sur des sous-échantillons des données. La prédiction se fait par vote majoritaire sur les résultats des N arbres.

La méthode du *Boosting* de gradient, est dérivée du *Boosting*, c'est une méthode permettant de créer un « prédicteur fort », à partir de « prédicteurs faibles ». Dans notre cas, nos prédicteurs faibles seront les arbres de régression.

XGBoost est une méthode reprenant la structure générale du *Gradient boosting*, en ajoutant plus de paramètres, permettant une meilleure estimation des résultats, ainsi qu'une diminution du temps de calcul.

Pour terminer, la régression *Elastic Net*, est une régression régularisée grâce aux pénalités de Ridge et LASSO.

Dans la partie suivante, les méthodes de provisionnement usuelles vont être exposées ainsi que leurs résultats en aviation et en maritime. Ces charges ultimes sur l'exercice de souscription 2020 seront les bases de comparaison pour les méthodes ligne à ligne.

2 PROVISIONNEMENT USUEL PAR CHAIN LADDER

La méthode du *Chain Ladder* repose sur la triangulation des sinistres passés sur un minimum de 10 ans, dans le cas de la *Line of Business* Marine Aviation Transport, la triangulation est faite par exercice de souscription. Elle est de la forme suivante (Figure 1) :

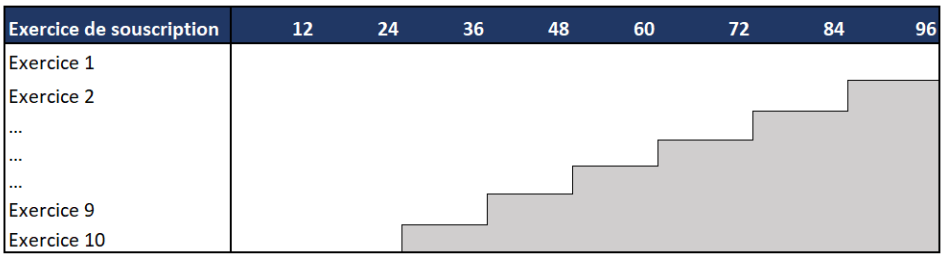


Figure 1-Schéma d'un triangle de liquidation.

Dans notre cas, chaque triangle étudié représente la charge totale cumulée (paiement + provisions dossier à dossier) observée tous les 12 mois, jusqu'au 31/12/2021, de tous les sinistres agrégés par exercice de souscription, la triangulation se fait sur 15 ans en aviation, 10 ans en maritime. Le but de la méthode étant de compléter cette triangulation grâce à des ratios de développement calculés sur la portion de triangle connue, jusqu'à l'obtention d'une charge totale sinistre appelée charge ultime.

2.1 METHODE ET RESULTATS AVIATION

Concernant l'aviation, trois triangles de liquidation sont créés. Les triangles devant être homogènes, ceux-ci sont séparés selon le risque touché :

- Responsabilité civile des passagers ou (RC)
- Responsabilité civile professionnelle ou produit (RC Produit)
- Corps

La projection n'est réalisée que sur les sinistres attritionnels, les sinistres dits exceptionnels, suivi par les gestionnaires et faisant appels à la réassurance non proportionnelle, sont évalués à dire d'expert. Dans ce

cas, la charge actuelle évaluée par les gestionnaires sinistres est jugée comme étant la meilleure estimation possible. Ces sinistres sont donc retirés de la triangulation, et ajoutés à la charge ultime, sans être projetés.

Une fois ces triangles de sinistres attritionnels créés, chacun doit vérifier les deux hypothèses d'application. Dans le cas de l'aviation, la première hypothèse, imposant l'indépendance des ratios de développement à l'année d'origine est vérifiée sur chacun des triangles. Cependant, la deuxième hypothèse imposant la linéarité des couples $(C_{i,j}, C_{i,j+1})$ par rapport à la droite des coefficients de développements, passant par l'origine, malgré le retraitement des coefficients de développement, reste invalide sur chacun des triangles. Toutefois, bien que la deuxième hypothèse ne soit pas validée, le modèle est appliqué.

Les triangulations Corps et Responsabilité Civile sont projetés grâce à un historique de cinq ans, avec des ratios de développement pondérés, donnant plus de poids aux données les plus récentes. La triangulation Responsabilité Civile professionnelle, particulièrement touchée par les variations de portefeuille, est projetée grâce à des ratios de développement calculés par une moyenne simple, hors exercices de souscription 2009 à 2014, ceux-ci n'étant plus représentatifs du portefeuille actuel.

L'exercice de souscription 2020, particulièrement touché par la pandémie, n'a pas pu être estimé par la méthode du *Chain Ladder* ou la méthode de *Boernhutter Fergusson.* Il a donc exceptionnellement été décidé, d'évaluer la charge ultime de l'exercice 2020, par une moyenne entre le ratio sinistres sur primes cible, obtenu grâce aux évaluations du *Business Plan,* et des résultats de la méthode *Chain Ladder*.

Une fois les charges ultimes des trois triangles établis, la charge ultime globale de 2020, est représenté par la somme des charges ultimes des trois triangles.

Finalement, nous observons les résultats suivants (Tableau 1) sur l'exercice de souscription 2020 :

Données en $	Charge	Charge ultime	IBNR	Evolution de la charge
Corps	77 676 270	81 043 720	3 367 450	4%
RC	22 822 382	70 799 487	47 977 105	210%
RC Produit	12 001 288	56 258 937	44 257 649	369%
Total	112 499 940	208 102 144	95 602 204	85%

Tableau 1-Résultats de la méthode Chain Ladder en aviation.

L'évolution globale de la charge sinistre est très élevée, entre la deuxième année de développement (31/12/2021) et la charge ultime. En effet, celle-ci évolue de 85%. Sur 2020, aucun sinistre exceptionnel n'a été écarté de la projection. Dans ce cas, le tableau ci-dessus est le résultat final de l'exercice de souscription 2020. Celui-ci sera la base de comparaison par rapport aux méthodes lignes à lignes.

2.2 METHODE ET RESULTATS MARITIME

Les données maritimes, sont-elles aussi séparées en plusieurs triangles. Selon les données fournies, trois triangles ont pu être recréés, Corps maritime (CMM), Dommages aux biens (IAR) et Pêche et Autre Corps (PAC). Des seuils ont été attribués à chacun de ces triangles, afin de séparer les sinistres attritionnels des sinistres larges. Les seuils sont présentés ci-dessous (Tableau 2) :

Activité	seuils
CMM	1 000 000
IAR	1 000 000
PAC	800 000

Tableau 2-Seuils entre attritionnels et larges sur chaque activité

Les hypothèses d'application, tout comme en aviation, ne sont pas validées sur les trois triangles.

Par la suite, la méthode du *Chain Ladder* est appliquée avec une moyenne pondérée sur l'historique complet disponible, soit sur 10 ans. Les charges ultimes de chacun des triangles ont été sommées, afin d'obtenir la charge globale. Les résultats suivants représentent la charge ultime totale projetée, sur les sinistres attritionnels de l'exercice de souscription 2020 (Tableau 3) :

Données en €	Charge	Charge ultime	IBNR	Evolution de la charge
CMM	841 926	861 752	19 826	2,30%
IAR	34 007	31 328	- 2 679	-8,55%
PAC	32 622 551	31 929 689	- 692 862	-2,17%
Total	33 498 484	32 822 769	- 675 714	-2,06%

Tableau 3- Résultats de la méthode Chain Ladder en maritime

Contrairement à l'aviation, la charge ultime obtenue, varie peu par rapport à la charge actuelle. Comme l'illustre le tableau ci-dessus, celle-ci est inférieure de 2.06% à la charge au 31/12/2021.

Maintenant que les sinistres attritionnels ont été évalués, il faut évaluer les sinistres larges, qui sont eux calculés selon une méthode de Cout Fréquence.

La fréquence des sinistres larges est projetée grâce à la méthode du Chain Ladder. La sinistralité finale de l'exercice de souscription 2020 est évaluée comme étant de 9.78 sinistres. Cependant, ce nombre est révisé comme étant de 9 sinistres, le nombre de sinistres évoluant très peu à partir de la deuxième année de développement.

Concernant la sévérité, les sinistres larges et clos de l'exercice de souscription 2020 n'évoluerons plus, par conséquent, ils représentent actuellement deux sinistres, pour une charge totale de 2.6 M€. Pour les sept sinistres restants, leur montant à la fin de la deuxième année de développement sont comparés aux sinistres larges passés et aujourd'hui clos, d'un montant comparable en deuxième année de développement. Ceci permet de ne pas surévaluer la sinistralité de l'année à estimer, en incluant des sinistres larges, possédant un montant de charge totale bien supérieur à ce qui est aujourd'hui déclaré sur l'exercice 2020.

La sinistralité moyenne calculée revient à 1.6M€. Au total, il y a donc sept sinistres non clos, pour 1.6M€, soit 11.3M€.

Le tableau suivant (Tableau 4) reprend les résultats en maritime :

Données en €	Charge	Charge ultime	IBNR	Evolution de la charge
Total attritionnels	33 498 484	32 822 769	- 675 714	-2,06%
Larges projetés	11 504 897	11 304 555	- 200 343	-1,77%
Larges clos	2 601 897	2 601 897	-	0,00%
Total	47 605 278	46 729 221	- 876 057	-1,87%

Tableau 4-Résultats de la méthode du Chain Ladder sur sinistres attritionnels et coût fréquence sur les sinistres larges

La charge ultime finale des sinistres attritionnels et larges de l'exercice de souscription 2020 est inférieure de 1.87% à la charge au 31/12/2021.

Maintenant que nous avons détaillés les résultats aviation et maritime de la méthode usuelle, les données disponibles pour la création des bases et l'analyse des différences et similitudes entre les deux bases vont pouvoir être abordés.

3 ANALYSE DES DONNEES : DIFFERENCES ET SIMILITUDES ENTRE AVIATION ET MARITIME

De nombreuses données sont disponibles au sein d'une entreprise d'assurance. Celles-ci se trouvent sous différentes formes, structurées ou non, et sont utilisées par différents départements. Il est capital de regrouper ces dernières afin de les retravailler pour que celles-ci soient exploitables par la suite dans les modèles. Cette étape représente la majeure partie du travail lors d'une étude produite par ce type de méthode. Dans cette partie, les données utilisées vont être exposées, puis les différences entre les bases aviation et maritime vont être étudiées.

3.1 CREATION DES BASES DE DONNEES AVIATION ET MARITIME

Qu'il s'agisse de l'aviation ou du maritime, les données proviennent de bases dites structurées, celles-ci sont exploitables directement, et utilisées tout au long de l'année. Elles reprennent les montants de sinistres, en provisions, paiements, charge totale, recours, les montants de primes associés, ou encore la part signée. En plus de ces informations, des données spécifiques aux assurés, sont disponibles, par des bases non structurées, ces informations sont ajoutées aux données structurées, afin d'apporter autant d'éléments que possibles à propos de l'assuré et du sinistre. Aux données structurées viennent s'ajouter le type de sinistre, la présence de contentieux ou non, le nombre de blessés, la valeur moyenne de la flotte, le type de navire, le nombre de moteurs, s'il s'agit d'une perte totale ou non etc…

Une fois toutes les informations collectées, il est important de retraiter les données manquantes ou aberrantes. Il existe différents types de données manquantes, celles-ci sont traités différemment selon les cas.

Dans ce mémoire, les données manquantes ont été remplacées par la médiane, ou par imputation partielle. La méthode des k plus proches voisins, mise en œuvre par Fix E. & Hodges J. (1951), appartenant à la catégorie de la classification ou de la régression supervisée, a également été appliquée.

Une fois ces bases complétées, il est possible d'analyser les premières données, celle-ci permettant d'observer les différences et similitudes entre les deux jeux de données, et également les interactions entre les variables et la variable cible, cette dernière étant la charge ultime.

3.2 ANALYSE DES DONNEES

Avant tout, la notion de censure doit être évoquée de manière à saisir certaines informations analysées dans la suite. Les données disponibles dans les bases sont dites censurées. Il n'est donc pas possible dans certains cas, d'observer la charge ultime d'un sinistre, ce dernier étant « en cours » au moment de l'observation au 31/12/2021. Cette information de censure, a une importance toute particulière dans l'application des méthodes de *Machine Learning* utilisées, et celles-ci devront tenir compte de la censure, par des poids IPCW (*Inverse Probability of Censoring Weighting)* qui seront calculés pour chacun sinistre, afin de rééquilibrer les données.

Globalement, la base maritime contient 87.7% de sinistres en plus que l'aviation. Ces sinistres observés, qui seront utilisés lors du provisionnement ligne à ligne sont également pour la plupart non censurés (Tableau 5) :

	Maritime				Aviation			
Exercice de souscription	**Censure**	**Clos**	**% censure**	**% PSAP en cours**	**Censure**	**Clos**	**% censure**	**% PSAP en cours**
2010	58	3501	1,6%	**-1%**				
2011	109	4067	2,6%	**22%**				
2012	67	3233	2,0%	**0%**				
2013	53	2693	1,9%	**5%**	274	1161	19,1%	**12%**
2014	24	2319	1,0%	**-1%**	275	1085	20,2%	**31%**
2015	31	2126	1,4%	**0%**	493	1352	26,7%	**9%**
2016	55	1747	3,1%	**43%**	808	1698	32,2%	**22%**
2017	91	1648	5,2%	**35%**	1128	1406	44,5%	**16%**
2018	187	1562	10,7%	**19%**	1307	1001	56,6%	**34%**
2019	338	1474	18,7%	**34%**	1144	514	69,0%	**37%**
2020	626	1230	33,7%	**55%**	893	199	81,8%	**54%**

Tableau 5- Comparaison des pourcentages de censure et des provisions dossier à dossier par exercice de souscription en maritime et aviation

Nous notons un taux de censure de 20% à 70% sur l'aviation contre 1.6% à 18% en maritime. Le pourcentage de Provisions dossier à dossier toujours en cours est également plus élevé en aviation, sur les exercices antérieurs à 2020. Le pourcentage de provisions dossier à dossier en cours en aviation et maritime est le même pour 2020, soit 55%.

Malgré ces importantes différences, des similitudes sont tout de même observables, principalement concernant la durée d'ouverture des sinistres, et le montant de la charge totale. En effet, il est difficile de distinguer une corrélation entre la durée d'un sinistre et sa charge totale. Ceci aura une importance particulière lors de l'application des poids IPCW. Il est tout de même possible de voir un impact de la durée sur la médiane des charges totales des sinistres clos, ainsi qu'un impact des contentieux sur la durée des sinistres, en aviation et en maritime.

Concernant les corrélations entre les variables quantitatives et la charge ultime, il est clair que les corrélations sur les bases maritime et aviation sont similaires.

Pour terminer, sur l'exercice 2020 qui sera prédit, il est important de noter le taux de tardifs habituellement observé à partir de la deuxième année de développement (Tableau 6), ce taux sera reporté sur les résultats 2020 du provisionnement ligne à ligne, ceci permettra de tenir compte de l'impact des tardifs dans le résultat.

	Maritime		Aviation	
Exercices de souscription	**Charge ultime**	**Nombre de dossiers**	**Charge ultime**	**Nombre de dossiers**
2010	2%	5%		
2011	1%	2%		
2012	2%	3%		
2013	3%	5%	6%	9%
2014	3%	5%	4%	7%
2015	4%	5%	3%	6%
2016	1%	4%	3%	6%
2017	1%	3%	1%	6%
2018	2%	2%	4%	2%
2019	0%	2%	0%	0%
Moyenne	**2,0%**	**3,5%**	**3,1%**	**5,1%**

Tableau 6-Comparaison du pourcentage de sinistres déclarés après deux années de développement en maritime et en aviation

Nous notons une moyenne de 2% en termes de montant de charge ultime, sur les sinistres clos des exercice 2010 à 2019 du maritime, et 3% pour l'aviation entre 2013 et 2019.

Maintenant que nous avons étudié et comparé les données des deux bases, il est possible de débuter le provisionnement ligne à ligne. Les poids IPCW vont tout d'abord être développés, et les étapes de la méthode de provisionnement ligne à ligne vont être approfondies. Par la suite, les résultats aviation et maritime seront exposés.

4 PROVISIONNEMENT LIGNE A LIGNE

4.1 METHODOLOGIE

Les poids IPCW sont calculés à partir de l'estimateur de Kaplan Meier. Il permet d'estimer de façon non paramétrique, la fonction de survie, d'après des données de durée de vie. Cet estimateur est couramment utilisé en économie, écologie, ou médecine. Les poids IPCW sont l'inverse de la probabilité d'être censuré à un instant t. Dans ce cas, les sinistres clos se voient attribuer un poids différent de 0, et les sinistres censurés un poids égal à 0.

Si nous incluons cette notion lors de l'application du provisionnement ligne à ligne, alors la méthode s'applique ainsi sur les deux jeux de données :

- Calcul des poids IPCW pour chacun des sinistres aviation et maritime

- Séparation de chacune des bases en deux bases distinctes, l'une d'apprentissage, l'autre, de tests. 80% des sinistres sont utilisés pour l'apprentissage du modèle, et les 20% restants, sont dédiés aux tests.

- Lancement de la phase d'apprentissage, les sinistres censurés, ont un poids de 0, et les sinistres clos, non censurés, se voient attribués un poids différent de 0. Selon la valeur x du poids, le sinistre sera répété x fois lors de l'apprentissage. Ceci permet de donner plus de poids aux informations non censurés, et à durée de vie élevée. Lorsque la durée et le montant de charge sont corrélés positivement, alors les sinistres à valeur élevés, peu représentés dans les bases de données, sont répétés plusieurs fois, afin de corriger cette censure, permettant ainsi une meilleure représentation des sinistres à charge élevée durant l'apprentissage. Lors de cette phase, les paramètres des modèles sont également calibrés par validation croisée afin d'approcher au mieux les valeurs de la base de test.

- Application des modèles calibrés aux données de l'exercice de souscription 2020. A chaque sinistre est attribué une charge sinistre ultime prédite. Toutes ces charges ultimes sont sommées afin d'obtenir la charge ultime de l'exercice 2020.

Les résultats obtenus sont résumés dans le chapitre suivant.

4.2 RESULTATS

Il est important de noter qu'afin d'aisément comparer les modèles, seuls les sinistres non censurés sont pris en compte lors du calcul des erreurs de prédiction.

Concernant l'aviation, si nous comparons les résultats du *Chain Ladder* et ceux obtenus par les cinq méthodes (Figure 2) :

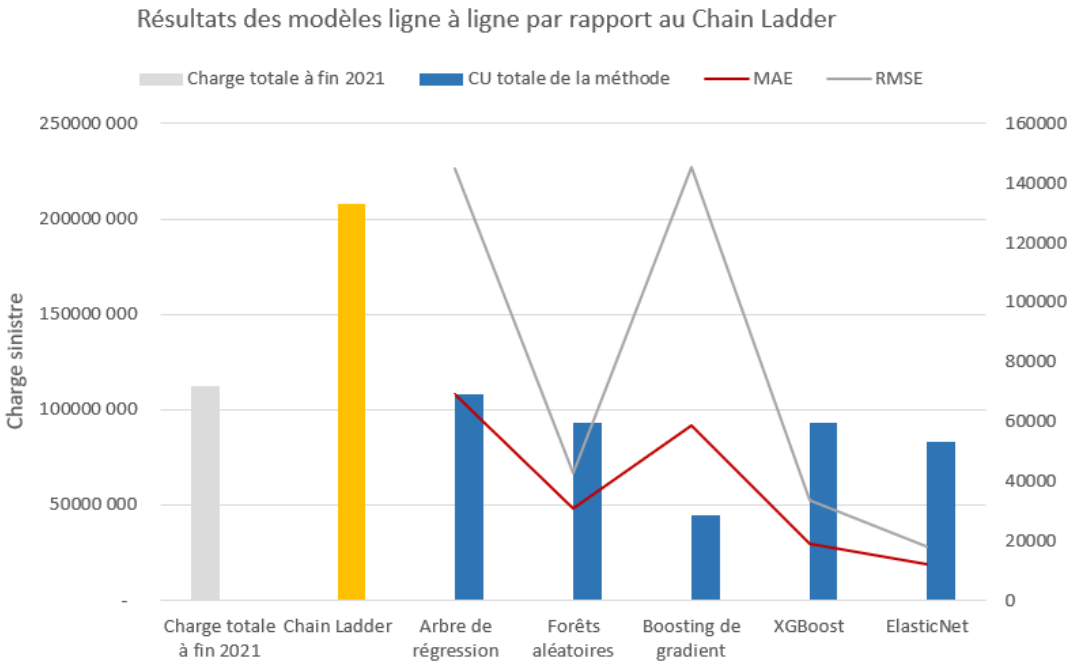


Figure 2-Charges ultimes obtenues par les différentes méthodes, MAE et RMSE (aviation) .

Les erreurs de prédiction présentes dans le graphique précédent, MAE (*Mean Absolute Error*) et RMSE (*Root Mean Square Error*), peuvent être décrites comme suit :

MAE : Le ***Mean Absolute Error***, ou l'erreur absolue moyenne, mesure l'amplitude moyenne des écarts entre les valeurs prédites par le modèle, et les observations. Elle se définie par la formule suivante :

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

RMSE : Le ***Root Mean Square Error***, ou la racine de l'erreur quadratique moyenne mesure la différence entre les valeurs prédites et les valeurs observées. Elle se définie par la formule suivante :

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

Les erreurs de prédiction les plus faibles dans la figure 2, permettant de choisir le modèle le plus pertinent, montrent que les méthodes *Elastic Net* et *XGBoost* sont les plus fiables, malgré une différence entre la charge ultime réelle et la charge ultime prédite élevée. Nous en déduisons ici que la méthode *XGBoost*, donne le meilleur résultat parmi les méthodes de partitionnement récursif.

Si nous retenons maintenant le résultat de la méthode *XGBoost* et que nous le comparons à la méthode du *Chain Ladder*, la méthode de provisionnement ligne à ligne est inférieure au *Chain Ladder*, de 55.20%. Cette différence est due au peu de données disponibles dans la base, ainsi qu'à l'application des poids IPCW, qui ne s'avère pas suffisante pour tenir compte des données censurées et à charge totale élevée. De ce fait, les modèles auront tendance à estimer correctement les sinistres à charge ultime faible voire moyenne, et largement sous-estimer les sinistres à charge ultime élevée, résultant en une charge totale prédite très faible.

Aujourd'hui, au 30/11/2022, soit un an après l'estimation obtenue par la méthode du *Chain Ladder*, la charge totale des sinistres de l'exercice 2020 s'élève à 150,7M$. Cette charge sinistre est déjà bien supérieure aux prédictions obtenues par les modèles de *Machine Learning*, et semble s'approcher des résultats de la méthode du *Chain Ladder*, confirmant l'estimation qui avait été faite l'année dernière, ainsi que le faible niveau de prédiction des modèles sur ce portefeuille.

Concernant le maritime, le graphique suivant reprend les informations de chaque modèle (Figure 3) :

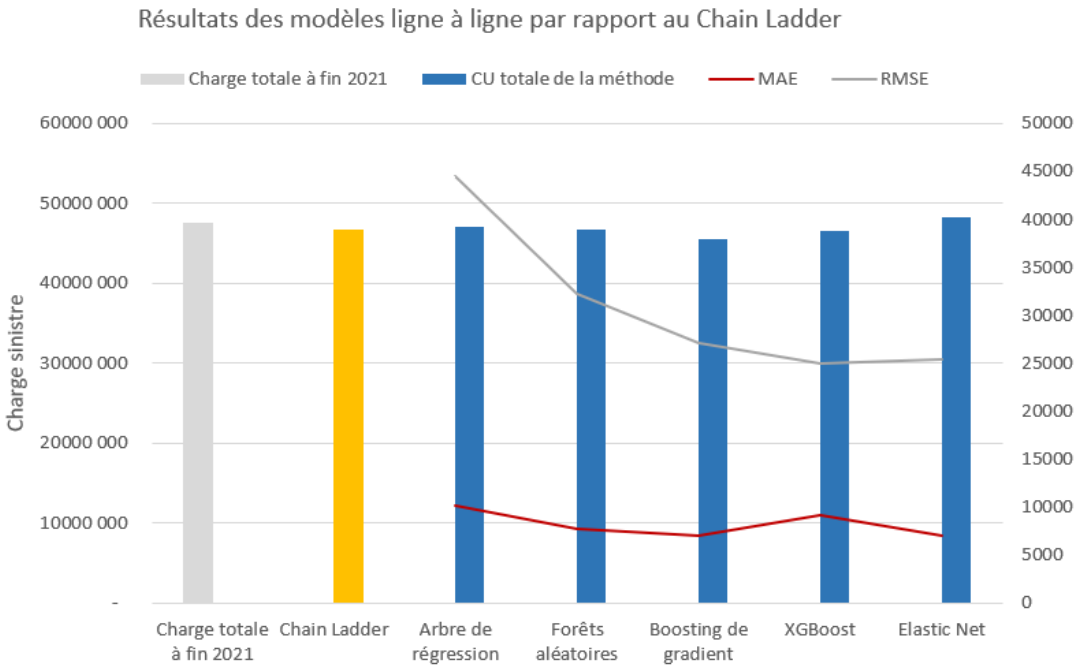


Figure 3- Charges ultimes obtenues par les différentes méthodes, MAE et RMSE (maritime).

Les charges ultimes des modèles avec les erreurs de prédiction les plus faibles, *Elastic Net* et *boosting* de gradient, sont différentes de la méthode *Chain Ladder*, de respectivement 3.27% et -2.64%. Les données des modèles lignes à lignes tiennent compte des 2% de charge ultime attendue des tardifs. Le nombre de sinistres observés étant bien plus élevé comparé à l'aviation, et le taux de censure bien plus bas, la base d'apprentissage utilisée pour l'entrainement des modèles compte bien plus de sinistre. La volatilité est également moins forte. Dans ce cas, la prédiction des sinistres est plus fiable, même si l'application des poids IPCW est peu impactante, tout comme en aviation.

En conclusion, il est effectivement possible d'évaluer la charge ultime des sinistres par des méthodes de *Machine Learning*. Les méthodes *XGBoost*, *Gradient Boosting* et *Elastic Net* se sont montrées plus concluantes que les autres. Il est cependant très important de tenir compte du type de données disponibles, comme peuvent montrer les différences de résultats obtenus.

Un niveau de censure très élevé, comme en aviation, sur une base contenant peu de sinistres, prédit très difficilement les sinistres élevés, souvent censurés et moins nombreux. Pourtant, ceux-ci sont très importants lors de l'évaluation de la charge ultime, et donc du *Best Estimate*. L'utilisation des poids IPCW n'a pas su rééquilibrer les données, les sinistres censurés et élevés n'étant pas assez représentés lors de l'apprentissage. La synthétisation de données par intelligence artificielle serait une étape supplémentaire envisageable avant application de la méthode, afin d'obtenir un nombre bien supérieur d'observations, et donc augmenter le nombre de sinistres larges rencontrés lors de l'apprentissage. Cependant, l'ajout d'informations par de telles techniques a un coût non négligeable.

Dans le cas du maritime, lorsque les données sont peu censurées et plus nombreuses, les résultats sont satisfaisants, les charges prédites obtenues par *Elastic Net* et *Gradient Boosting* étant proches des charges évaluées par *Chain Ladder*. Les données d'apprentissages étaient suffisantes pour évaluer correctement une charge ultime sur l'exercice de souscription 2020. Cependant, un modèle doit être entretenu et entrainé sur de nouvelles données afin de rester fiables sur les années à venir.

Executive summary

1 CONTEXT AND OBJECTIVE

In insurance, the ultimate loss cost, as well as the Best Estimate (BE), respectively refers to the ultimate total amount of claims per underwriting year, calculated during reserving in non life insurance, and the sum of probabilistic and actualised future cash flows. Both are crucial information when calculating the non life Solvency Capital Requirement (SCR). Indeed, solvency 2's prudential regulation covers the importance of technical provisions, which need to be unbiased, cautious, and reliable. The Best Estimate and therefore the ultimate loss cost, represent a non negligible part of these technical provisions, and so must be calculated with caution.

This reserving, done at least every year, is frequently applied conventionally, with the Chain Ladder method, a determinist actuarial method, quick and simple to use in an insurance company, based on the triangulation of the aggregated costs of the past claims, also named run-off triangles. However, this method is not always suitable, for it is based on two important hypotheses rarely proven on the data used.

Moreover, the portfolio has to be composed of numerous and reliable data, the past has to be regular, and studied on branches with low volatility. A variation in the portfolio over time implies that the development ratios obtained thanks to the past claims, do not represent the actual portfolio anymore. The aviation and maritime portfolios which will be studied throughout this dissertation, are typical of a low number of claims, but with high volatility. However, the maritime portfolio contains more data, and a lower volatility than aviation.

Therefore, our objective in this thesis will be to expose the Chain Ladder method and then challenge it for underwriting year 2020 in aviation as well as in maritime, using Machine Learning methods, allowing line-to-line reserving. These methods enable us to break free from the hypotheses of the Chain Ladder and add qualitative and quantitative information concerning the claims and the insured. The differences and similarities between the two data bases will be exposed, and the CART (Classification And Regression Tree), random forest, Gradient Boosting, *XGBoost* as well as the Elastic Net methods will be applied. The two bases used will allow us to check the reliability of the models, and also appreciate the differences in results according to the type of data.

The purpose of the CART method is to create a model, making it possible to predict the value of a target variable, from several input variables. The tree is built by recursive distribution. This method is simple and fast in terms of execution but usually unreliable.

For random forests, the principle is based on the CART method. Indeed, N decision trees are calibrated on sub-samples of the data. The prediction is made by majority vote on the results of the N trees.

The Gradient Boosting method is derived from Boosting, this method creates "strong predictors" from "weak predictors". In our case, our weak predictors will be decision trees.

XGBoost is a method using the general structure of Gradient boosting, adding more parameters, allowing a better estimation of the results, as well as a reduction in the calculation time.

Finally, the Elastic Net regression is a regularized regression, thanks to the Ridge and LASSO penalties.

In the next part, the usual reserving methods will be exposed, as well as their results in aviation and maritime. The two ultimate loss costs of underwriting year 2020 will serve as our basis of comparison for the Machine Learning methods.

2 USUAL RESERVING BY CHAIN LADDER

This method relies on the triangulation of past claims, on a minimum of ten years, in the case of the Marime Aviation Transport Line of Business, the triangulation is done by underwriting year. It has the following shape (Figure 4):

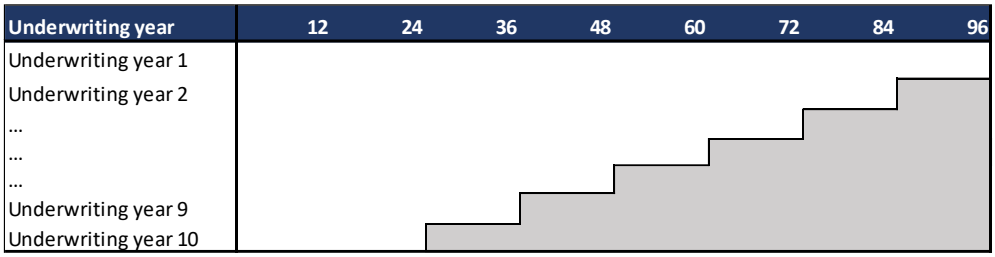


Figure 4-Shape of a run-off triangulation

In our case, each triangle represents the aggregated incurred loss cost (paid + outstanding) observed every 12 months, until 12/31/2021, by underwriting year, it is a fifteen year triangulation for aviation and ten year for maritime. The aim being to complete this triangulation thanks to development ratios, until we obtain an ultimate loss cost.

2.1 AVIATION METHOD AND RESULTS

Concerning aviation, three run-off triangles are created. They are split according to the risk of the claim:

- Liability
- Product liability
- Hull

The Chain Ladder method is only applied on the attritonnal claims, the claims that are exceptional, followed by the claim handlers and mostly needing unproportionnal reinsurance, are already seen as the best estimation possible, therefore they are withdrawn from the run-off triangles and not projected.

Once these triangles have been created, each one of them must validate the two hypotheses seen before. In aviation, the first hypothesis imposing the independence of the development rates and the year of origin, is validated. However, the second one, imposing the linearity of the $(C_{i,j}, C_{i,j+1})$ couples compared to the line drawn by the development rates going through the origin of the plane, despite reprocessing the development rates, stays invalid on each triangle. The method is still applied although the second hypothesis is not valid for our data.

The Hull and Liability run-off triangles are projected with a five-year history, with weighted development rates. The product liability triangle is particularly affected by portfolio variations, and projected thanks to development rates calculated by the average of fifteen full underwriting years except 2009 to 2014, these underwriting years do not represent the actual portfolio.

The 2020 underwriting year, particularly affected by the pandemic, could not be estimated exclusively by Chain Ladder or Boernhutter Fergusson. It has been decided to evaluate the ultimate loss cost of 2020, with an average of a targeted loss ratio, thanks to the business plan, and Chain Ladder.

Once the three ultimate loss costs have been calculated, the aviation 2020 loss cost is the sum of the three results.

Finally, the 2020 underwriting year has the following results (Table 1):

Figures in $	Incurred to date	Ultimate loss	IBNR	Evolution
Hull	77 676 270	81 043 720	3 367 450	4%
Liability	22 822 382	70 799 487	47 977 105	210%
Product liability	12 001 288	56 258 937	44 257 649	369%
Total	112 499 940	208 102 144	95 602 204	85%

Table 1-Chain Ladder results in aviation

The overall evolution is particularly high, between the second development year (12/31/2021), and the ultimate cost. Indeed, it evolves by 85%. No exceptional claim was excluded from the projection. In this case, the table above is the final result of 2020. This will be the basis of comparison with the line-by-line method.

2.2 MARITIME METHOD AND RESULTS

The maritime data are also separated into three triangles, « Corps maritime » (CMM), IAR “dommages aux biens” and “Pêche et Autres Corps” (PAC). Thresholds have been assigned to each of these triangles, in order to separate attritional claims from large claims. The thresholds are presented below (Table 2):

Activities	Thresholds
CMM	1 000 000
IAR	1 000 000
PAC	800 000

Table 2-Thresholds between attritionnals and larges for maritime

The hypotheses are not all validated on these three triangles.

The Chain Ladder method is applied with a weighted average over the full history available, i.e. over ten years. The ultimate cost of each of the triangles were summed to obtain the overall cost.

The following results (Table 3) represent the projected total ultimate charge on attritional claims for underwriting year 2020:

Figures in €	Incurred to date	Ultimate loss	IBNR	Evolution
CMM	841 926	861 752	19 826	2,30%
IAR	34 007	31 328	- 2 679	-8,55%
PAC	32 622 551	31 929 689	- 692 862	-2,17%
Total	33 498 484	32 822 769	- 675 714	-2,06%

Table 3-Chain Ladder results in maritime

Contrary to aviation, the ultimate loss cost obtained is not much different from the actual loss. Only 2.06% lower.

Now that the attritional claims have been assessed, it is necessary to assess the large claims, which are calculated using a Cost Frequency method.

The frequency of large claims is projected using a Chain Ladder method. The final number of claims for 2020 is estimated to be 9.78 claims. However, this number is revised to be 9 claims.

In terms of severity, large and closed claims for the 2020 underwriting year are evaluated as they are, therefore they currently represent two claims, for a total charge of €2.6M. For the seven other claims, their amount at the end of the second year of development is compared to past and closed large claims of a similar amount in the second year of development. This allows us not to overestimate the severity. The average loss amount is €1.6M. In total, there are therefore seven unclosed claims, for €1.6M, so €11.3M for seven claims.

The following table shows the maritime results (Table 4):

Figures in €	Incurred to date	Ultimate loss	IBNR	Evolution
Total attritionnals	33 498 484	32 822 769	- 675 714	-2,06%
Larges claims (Cost and Frequency)	11 504 897	11 304 555	- 200 343	-1,77%
Larges Closed claims	2 601 897	2 601 897	-	0,00%
Total	47 605 278	46 729 221	- 876 057	-1,87%

Table 4- Chain Ladder method's final results in maritime

The ultimate loss cost of 2020 is 1.87% lower than the cost as at 12/31/2021.

Now that we have detailed the aviation and maritime results obtained by the usual method, the data available and the analysis of the differences and similarities between the two databases will be addressed.

3 DATA ANALYSIS: SIMILARITIES AND DIFFERENCES BETWEEN AVIATION AND MARITIME

A lot of data is available within an insurance company. These are found in different forms, structured and non structured, and are used in different departments. It is important to group them and rework them so that they can be used later in the models. This step represents most of the work during a study produced by this type of method. In this part, the available data will be exposed, then the differences between the aviation and maritime bases will be studied.

3.1 CREATION OF AVIATION AND MARITIME DATABASES

Whether in aviation or maritime, the data used comes from structured databases, which can be used directly, and are usually used throughout the year, which includes the amounts of claims, outstanding, payments, incurred, recourse, the amounts of associated premiums, or even the signed line. In addition to this information, data specific to the insured is available through different bases. This information is added to the structured data, in order to provide as many elements as possible on the insured and on the claim.

In addition to the structured data, non-structured data contains information about the type of loss, the presence of litigation or not, the number of injured, the average value of the fleet, the type of vessel, the number of engines, whether it is a total loss or not etc…

Once all the data has been collected, it is important to reprocess any missing or outlying data. There are different types of missing data, these are treated differently depending on the case.

In this dissertation, the missing data were replaced by the median, or by partial imputation. The method of k nearest neighbors, implemented by Fix E. & Hodges J. (1951), belonging to the category of classification or supervised regression, was also applied.

Once these bases are completed, it is possible to analyse them, thus allowing us to observe the differences and similarities between the two data sets, and also the interactions between the variables and the target variable, which is the ultimate loss cost.

3.2 DATA ANALYSIS

The notion of censorship must be mentioned beforehand, in order to grasp certain information analysed below. The data available in the databases is sometimes censored. It is therefore not possible in some cases to observe the ultimate loss cost, because it is still "in progress" at the time of observation. This information, whether it is censored or not, is particularly important in the application of Machine Learning methods, these must take account of the censorship, therefore IPCW weights (Inverse Probability of Censoring Weighting) will be calculated for each claim, in order to balance the data in the face of this censorship.

In general, the maritime base contains 87.7% more claims than aviation. These observed claims, which will be used during line-by-line reserving, are also mostly uncensored (Table 5):

	Maritime				Aviation			
Underwriting year	Censored	Closed	%Censored	% Outstanding	Censored	Closed	%Censored	% Outstanding
2010	58	3501	1,6%	**-1%**				
2011	109	4067	2,6%	**22%**				
2012	67	3233	2,0%	**0%**				
2013	53	2693	1,9%	**5%**	274	1161	19,1%	**12%**
2014	24	2319	1,0%	**-1%**	275	1085	20,2%	**31%**
2015	31	2126	1,4%	**0%**	493	1352	26,7%	**9%**
2016	55	1747	3,1%	**43%**	808	1698	32,2%	**22%**
2017	91	1648	5,2%	**35%**	1128	1406	44,5%	**16%**
2018	187	1562	10,7%	**19%**	1307	1001	56,6%	**34%**
2019	338	1474	18,7%	**34%**	1144	514	69,0%	**37%**
2020	626	1230	33,7%	**55%**	893	199	81,8%	**54%**

Table 5-Percentage of censored claims and outstanding for each underwriting year in aviation and maritime

There is a censorship percentage of 20% to 70% on aviation as against 1.6% to 18% in maritime. The percentage of outstanding in opened claims is also higher in aviation, for years prior to 2020.

Despite these significant differences, similarities are still observable, mainly concerning the duration of the opening of claims, and the incurred. Indeed, it is difficult to distinguish a correlation between the duration of a claim and its total expense. This will be of particular importance when applying IPCW weights. It is still possible to see an impact of duration on the median of the total expenses of closed claims, in aviation and marine, as well as an impact of litigation on the duration of claims. Concerning the correlations between the quantitative variables and the ultimate loss cost, it is clear that the correlations on the maritime and aviation bases are similar.

Finally, for underwriting year 2020 which will be predicted, it is important to note the percentage of belated claims usually observed from the second year of development, this percentage will be reported on the 2020 results of the line-by-line reserving, this will make it possible to consider the impact of belated claims in the results (Table 6):

	Maritime		Aviation	
Underwriting year	**Ultimate loss**	**Number of claims**	**Ultimate loss**	**Number of claims**
2010	2%	5%		
2011	1%	2%		
2012	2%	3%		
2013	3%	5%	6%	9%
2014	3%	5%	4%	7%
2015	4%	5%	3%	6%
2016	1%	4%	3%	6%
2017	1%	3%	1%	6%
2018	2%	2%	4%	2%
2019	0%	2%	0%	0%
Average	**2,0%**	**3,5%**	**3,1%**	**5,1%**

Table 6-Comparison of belated claims after two years between aviation and maritime

We note an average of 2% in terms of ultimate loss cost, on closed claims for the maritime for years 2010 to 2019, and 3% for aviation between 2013 and 2019.

Now that we have studied and compared the data from the two databases, it is possible to start line-by-line reserving. The IPCW weights will first be developed, and the steps of the line-by-line reserving method will be detailed. Subsequently, the aviation and maritime results will be presented.

4 LINE-BY-LINE RESERVING

4.1 METHODOLOGY

The IPCW weights are calculated with the Kaplan Meier estimator. It allows non-parametric estimation of the survival function, based on lifespan data. This estimator is commonly used in economics, ecology, or medicine. The IPCW weights are the inverse of the probability of being censored at time t. Closed claims have a weight different from 0, and censored claims have a weight equal to 0.

If this notion is included when applying line-by-line reserving, then the method applies as follows:

- Calculation of IPCW weights for each of the aviation and maritime claims

- Separation of the total database into two databases, one for training, the other for testing. 80% of claims are used for model training, and the remaining 20% are dedicated to testing.

- During the training phase, censored claims have a weight of 0 and closed, uncensored claims are assigned a weight different from 0. Depending on the value x of the weight, the claim will be repeated x times during the training phase. This method gives more weight to uncensored, high-duration information. When the duration and the incurred amount are positively correlated, then the high value claims, poorly represented in the databases, are repeated several times, in order to correct this censorship. During this phase, the models are also calibrated in order to best approach the values of the test base. This calibration is done by cross-validation.

- Application of calibrated models to the 2020 underwriting year data. Each claim is assigned an ultimate loss cost. All ultimate costs are summed to obtain the ultimate loss cost for underwriting year 2020.

The results for each database are summarized in the following chapter.

4.2 RESULTS

In order to easily compare the models, only uncensored claims are taken into account when calculating prediction errors.

Regarding aviation, if we compare the results of the Chain Ladder and those obtained by the five methods (Figure 5):

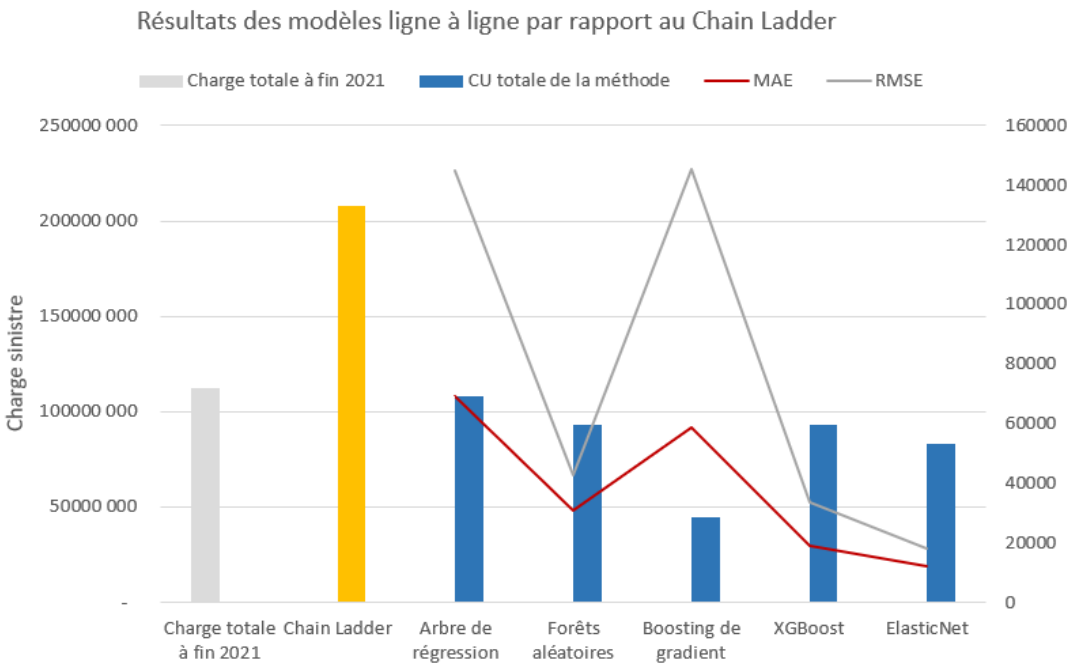


Figure 5-Ultimate loss cost of each method, MAE and RMSE.

The prediction errors present in the figure above, MAE (Mean Absolute Error) and RMSE (Root Mean Square Error), can be described as follows:

MAE: The **Mean Absolute Error**, or the average absolute error, measures the average amplitude of the differences between the values predicted by the model, and the observations. It is defined by the following formula:

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

RMSE: **Root Mean Square Error** measures the difference between predicted values and observed values. It is defined by the following formula:

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

The lowest prediction errors, making it possible to choose the most relevant model, show that the Elastic Net and *XGBoost* methods are the most reliable, despite a high difference between the real ultimate loss cost and the predicted ultimate loss cost. We deduce here that the *XGBoost* method gives the best result among the recursive partitioning methods.

If we now compare the results of the *XGBoost* method to the Chain Ladder method, on all the claims, censored and non-censored, the line-by-line reserving method is inferior to the Chain Ladder, by 55.20%. This difference is due to the limited data available in the database, coupled with the application of IPCW

weights not being sufficient to balance the censored data with very high amounts of incurred. As a result, all models will tend to correctly estimate claims with a low or medium incurred, and greatly underestimate claims with a high incurred.

Today, as of 11/30/2022, i.e. one year after the results obtained by the Chain Ladder method, the total amount of claims for underwriting year 2020 is $150.7 million. This loss load is already much higher than the predictions obtained by Machine Learning models and seems to approach the results of the Chain Ladder method, confirming the estimate that was made last year, and the poor level predictions obtained with Machine Learning.

Concerning the maritime, the following table summarizes the information of each model (Figure 6):

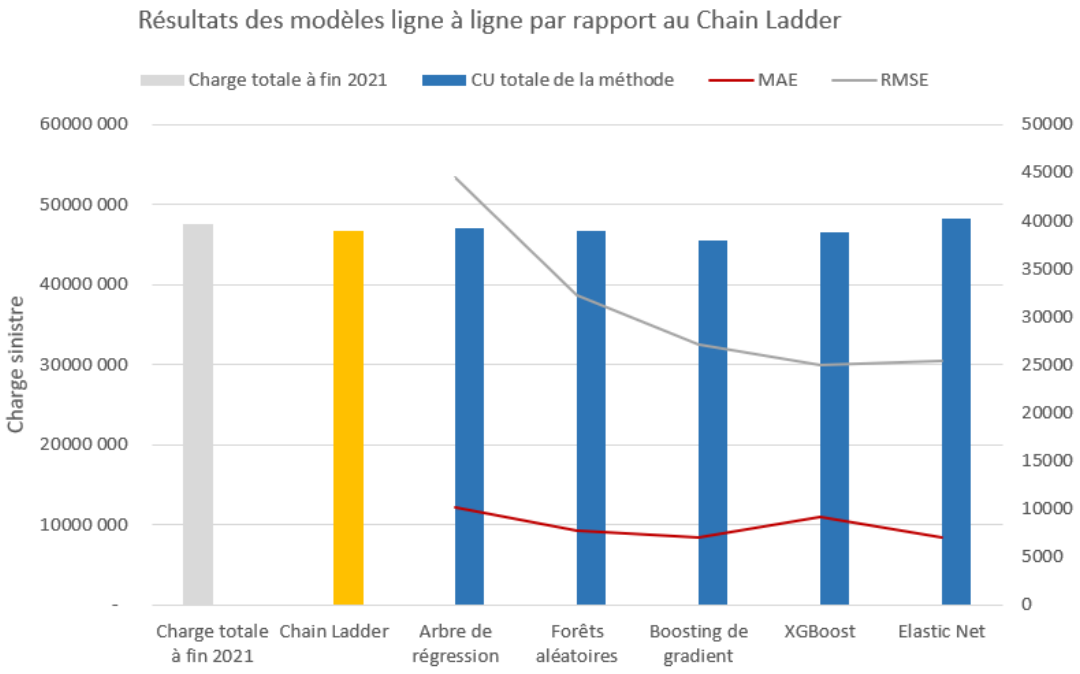


Figure 6-Ultimate loss cost of each method, MAE and RMSE.

The ultimate loss cost of the models with the lowest prediction errors, Elastic Net and Gradient Boosting, are lower than the ultimate loss cost of the Chain Ladder method, by 3.27% and 2.64% respectively. The line-by-line model data takes into account the 2% ultimate load expected from the belated claims. The number of claims observed being much higher compared to aviation, and the rate of censorship much lower, the database used for training the models has many more claims. Volatility is also lower. In this case, the prediction of the highest claims is more reliable, even if the IPCW weights have little impact, just like in aviation.

In conclusion, it is indeed possible to assess the ultimate loss cost using Machine Learning methods. The *XGBoost*, Gradient Boosting and Elastic Net methods have proven to be more successful than the others. However, it is very important to take into account the type of data available, as the differences in results obtained can show.

A very high level of censorship, as in aviation, on a database containing few claims, hardly predicts high claims, which are often censored and fewer in number. However, these are very important when evaluating the ultimate loss cost, and therefore the Best Estimate. The use of IPCW weights was not able to balance the data, censored and high incurred claims not being sufficiently represented during training. The synthesis of data by artificial intelligence would be an additional step that could be considered before

applying the method, in order to obtain a much higher number of observations, and therefore increase the number of large claims encountered during training. However, the addition of information by such techniques has a significant cost.

In the maritime case, when the data is less censored and more numerous, the results are satisfying, the predicted loss costs obtained by Elastic Net and Gradient Boosting being close to the loads evaluated by Chain Ladder. The learning data was sufficient to correctly assess an ultimate loss cost for 2020. However, a model must be maintained and trained on new data in order to remain reliable over the years to come.

Remerciements

Je souhaite remercier Stéphanie Foata et Christophe Graber, qui m'ont fait confiance, et m'ont permis de suivre la formation du CEA. Un merci tout particulier à Stéphanie, qui a relu de nombreuses fois ce mémoire, et a suivi tout au long de cette année son avancée.

Je tiens également à remercier Christophe Deshayes et Paul-André Berger, qui ont pu fournir les informations d'Helvetia, et me guider durant l'étude de ces données.

Je remercie également ma famille, pour le soutien constant qu'ils m'ont apporté durant cette année.

Introduction

La charge ultime, habituellement obtenue par des méthodes de provisionnement classiques, est une donnée importante lors de l'établissement du *Best Estimate*, partie fondamentale lors du calcul du *non life Solvency Capital Requirement* ou SCR non vie. L'établissement d'une charge ultime fiable, est donc primordial pour un actuaire.

La méthode utilisée lors de l'estimation de cette charge ultime est très souvent celle du *Chain Ladder*, méthode déterministe, reposant sur l'évolution des charges sinistres totales observées par exercice de souscription ou année de survenance. Cette méthode, par sa simplicité d'exécution, est fréquemment utilisée dans les entreprises d'assurance. Dans le cadre de ce mémoire, deux portefeuilles, aviation et maritime, sur lesquelles cette méthode est utilisée, seront étudiés.

L'application de la méthode du *Chain Ladder* nécessite des données stables dans le temps, reposant sur de nombreux sinistres, avec une sinistralité passée similaire à la sinistralité présente, et à charge sinistre peu volatile. Cependant, les activités aviation et maritime étudiées ici, sont des activités de niche, générant peu de primes et peu de sinistres. Par conséquent, du fait de leur nature, elles possèdent une sinistralité faible en termes de nombre, ainsi qu'une sinistralité actuelle peu similaire à celle du passée.

En effet, d'une part, d'après Paulin C. (2018) et Broudin R., l'évolution importante du trafic aérien ces vingt dernières années, les navires étant plus grands, et le risque cyber étant également plus élevée, implique que les années passées en aviation et maritime sont peu représentatives du présent. D'autre part, l'arrivée de la pandémie, impacte également fortement les résultats des triangles de liquidation des sinistres sur les exercices de souscription récents, la sinistralité de ces années ayant été fortement réduite par la baisse d'activité sur les deux secteurs.

L'augmentation de la complexité des sinistres aviation et maritime au fil du temps, impacte également leur charge totale, celle-ci en est de plus en plus volatile. En effet, comme l'indique Paulin C., un sinistre aviation peut par exemple aujourd'hui avoir une limite de garantie (Corps + Responsabilité Civile) jusqu'à deux fois la prime mondiale collectée.

Les sinistres enregistrés sur ces deux branches sont donc peu nombreux, volatiles, et l'évolution de la sinistralité passée, est peu représentative de la sinistralité actuelle. L'application de la méthode du *Chain Ladder* ne semble pas adéquate sur ce type de portefeuille.

Fort de ce constat, la méthode du *Chain Ladder* sur les portefeuilles de sinistres aviation et maritime seront challengés par des méthodes de *Machine Learning*, reposant sur les informations quantitatives et qualitatives disponibles, permettant le provisionnement ligne à ligne, telles que les arbres de décisions, introduits par Breiman L. et al. (1984) dans la méthode CART (*Classification And Regression Trees*). Des méthodes se basant sur la méthode CART comme les forêts aléatoires, *gradient boosting* et *XGBoost* seront également testés, ainsi qu'un modèle de régression pénalisé, *Elastic Net* définie par Zou H. et Hastie T. (2004).

Pour ce faire, nous allons aborder dans un premier temps le contexte général de l'assurance non vie en aviation et maritime, ainsi que le cycle de vie d'un sinistre et afin de comprendre l'importance du *Best Estimate* lors de l'estimation des provisions techniques, nous terminerons par une introduction au calcul du SCR non vie afin d'apporter un cadre règlementaire sur ces notions.

Dans un second temps, nous détaillerons l'application de la méthode du *Chain Ladder* sur les deux portefeuilles étudiés, afin d'obtenir les charges ultimes aviation et maritime sur l'exercice de souscription 2020. Ces résultats seront notre base de comparaison avec les modèles ligne à ligne étudiés dans la suite.

Dans un troisième temps, les bases de données disponibles seront détaillées et étudiées, afin d'exposer les différences et similitudes entre les deux jeux de données. Ceci nous permettra par la suite de comprendre les différences de résultat observées, lors du provisionnement ligne à ligne.

Nous aborderons également la notion de censure en assurance non vie, et la notion de poids IPCW (*Inverse Probability of Censoring Weighting*), permettant de tenir compte de cette censure, à partir de l'estimateur de Kaplan Meier, permettant d'associer un poids à chaque sinistre observé.

Pour terminer, les cinq méthodes de provisionnement ligne à ligne utilisées seront détaillées puis appliquées aux données aviation et maritime. Les modèles seront comparés afin d'obtenir la meilleure estimation possible par rapport à la méthode du *Chain Ladder*. Nous verrons finalement si ces méthodes de provisionnement ligne à ligne sont applicables sur nos jeux de données, les différences de résultats entre l'aviation et le maritime seront également expliquées. Les limites des méthodes seront exposées, ainsi que de possibles pistes d'amélioration.

Table des matières

Liste des abréviations

ACPR: Autorité de Contrôle Prudentiel et de Résolution

BE: Best Estimate

CMM: Corps Maritime

IAR: Dommages aux biens

IBNeR: Incurred But Not enough Reported

IBNR: Incurred But Not Reported

IBNyR: Incurred But Not yet Reported

IPCW: Inverse Probability of Censoring Weighting

IUAI: International Union of Aerospace Insurers

PAC: Pêche et Autre Corps

PPNA: Provision pour Prime Non Acquise

RC: Responsabilité Civile

SCR: Solvency Capital Requirement

1 Présentation de l'assurance non vie

1.1 CONTEXTE GENERAL

Le contrat d'assurance est un **contrat aléatoire**, par lequel une partie, un assureur, s'engage envers une autre partie, le souscripteur (représentant une ou plusieurs personnes déterminées), à couvrir moyennant le paiement d'une prime d'assurance ou cotisation, une catégorie de risque déterminés par le contrat. Ce fonctionnement est caractéristique de ce que nous appelons l'inversion du cycle de production. La souscription du contrat peut passer par l'aide de professionnels de l'assurance, comme des intermédiaires, souvent des courtiers ou agents généraux.

L'exécution de la prestation repose sur la réalisation ou non d'un évènement. Cet aléa doit pouvoir être mutualisé, grâce à de nombreuses primes d'assurances, chacune servant à payer les sinistres des autres assurés. Le montant final de l'engagement est incertain, il est donc important **d'estimer au mieux l'engagement futur lié à chacun de ces contrats**, cet exercice correspond à l'établissement des **provisions techniques**. De ce fait, les provisions techniques font partie intégrante de la gestion des sinistres.

D'après Marly, P.-G. (2013), et Carlot, J.-F. (2021), les assurances se classent en deux grandes catégories distinctes :

- Les assurances de dommages, dont nous pouvons distinguer les dommages atteignant l'actif de l'assuré, ou le passif de l'assuré. Respectivement les assurances de biens et pertes pécuniaires ou les assurances de responsabilité dont l'assuré serait redevable.
- Les assurances de personnes, qui regroupent les assurances sur la vie (en cas de décès ou en cas de vie), ou les assurances de dommages corporels, protégeant l'assuré contre des affections physiques possibles.

1.1.1 Assurance non vie

Dans cette section, nous présenterons l'assurance non vie, en détaillant les chiffres de la *Line Of Business* (ou LoB) Marine Aviation Transport, puis les activités aviation et maritime seront étudiées, afin d'apprécier les différences entre certaines branches composant l'assurance non vie.

D'après Marly, P.-G. (2013), l'assurance non vie regroupe différentes opérations d'assurance, tels que les assurances de dommages corporels, les dommages aux biens, les assurances de Responsabilité Civile (RC) ou encore l'assistance et la protection juridique. Plusieurs principes régissent l'assurance non vie. Celle-ci se fait par **répartition**, cela consiste pour l'assureur, à verser aux victimes des sinistres, pour une année donnée, l'ensemble des cotisations versées par tous les assurés sur cette même année. L'assurance dommage se base également sur le **principe indemnitaire**, il est par conséquent impossible pour l'assuré de s'enrichir grâce à ses indemnisations. En assurance de personnes, le **principe forfaitaire** permet quant à lui aux souscripteurs de fixer le montant d'indemnisation en cas de réalisation du risque.

Selon l'Autorité de Contrôle Prudentiel et de Résolution ou ACPR, les primes acquises brutes des affaires directes et sinistres associés, étaient les suivantes de 2018 à 2021 pour les *Line of Business* suivantes (Tableau 7) :

Milliards d'euros	Dommages corporels	Automobile	Dommages aux biens	Responsabilité civile générale	Construction	Catastrophes naturelles	Crédit caution	Transport
Sinistres 2018	43,42	17,34	11,90	1,94	1,56	1,51	0,17	0,25
Primes 2018	53,26	21,87	17,64	2,78	2,06	1,59	0,79	0,61
S/P 2018	82%	79%	67%	70%	76%	95%	21%	42%
Sinistres 2019	44,99	18,37	12,59	2,04	1,47	2,09	0,21	0,30
Primes 2019	54,78	22,46	18,41	2,91	2,17	1,62	0,91	0,50
S/P 2019	82%	82%	68%	70%	68%	129%	23%	59%
Sinistres 2020	46,20	17,20	14,50	1,80	1,80	2,10	0,20	0,30
Primes 2020	55,60	22,50	18,40	2,90	2,20	1,60	0,90	0,50
S/P 2020	83%	76%	79%	62%	82%	131%	22%	60%
Sinistres 2021	46,50	18,50	13,60	1,80	2,50	1,20	0,10	0,40
Primes 2021	55,90	23,60	19,50	3,20	2,40	1,70	1,10	0,60
S/P 2021	83%	78%	70%	56%	104%	71%	9%	67%

Tableau 7-Répartition par branches des primes et sinistres des affaires directes Françaises.

Les Primes proviennent principalement du dommage corporel, automobile et dommages aux biens, soit 92.26% des primes totales sur l'exercice 2020. Le ratio des sinistres sur les primes le plus élevé, correspond quant à lui aux catastrophes naturelles.

Si nous étudions les proportions des primes et des sinistres par branche sur le marché Français en 2020 (Figure 7) :

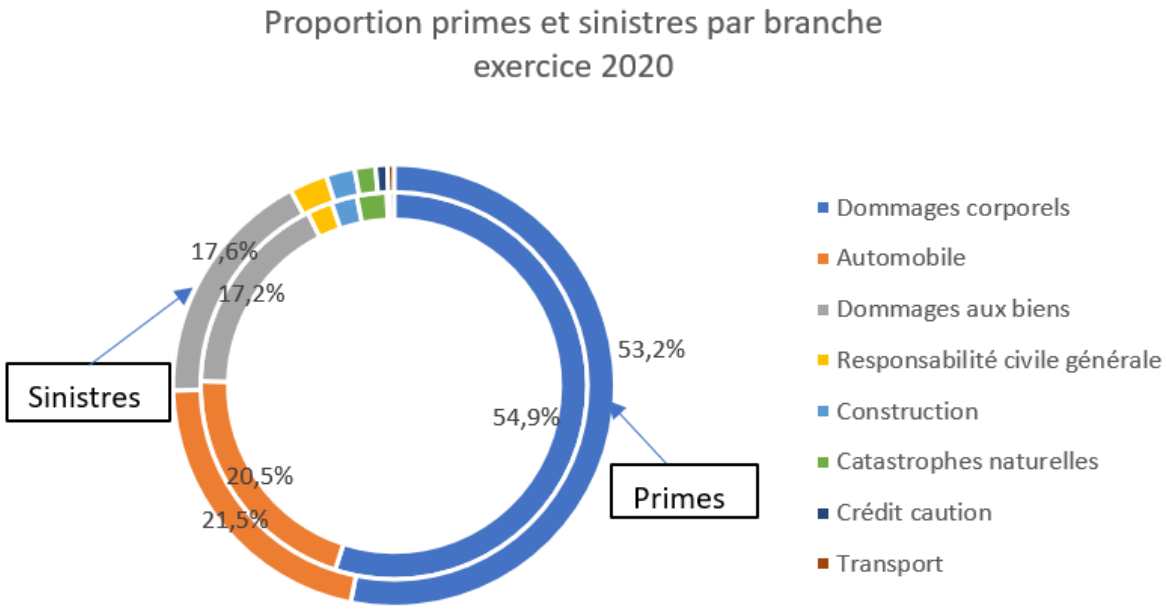


Figure 7-Répartition des primes et sinistres par branches, sur l'exercice 2020.

Nous pouvons noter que les dommages corporels et l'automobile représentent à eux seul 75.4% des sinistres et 74.7% des primes. La sinistralité du transport quant à elle, est peu élevée par rapport aux autres branches, avec une part de primes elle-même faible, soit 0.48% de la prime totale des affaires directes Françaises, sur 2020, résultant en un ratio de sinistres sur primes de 60%. Nous remarquons ici aisément que la branche Marine Aviation Transport ou MAT est une activité de niche.

1.1.2 Assurance aviation et maritime

Nous allons ici étudier plus en détail la répartition des primes aviation et maritime de la branche Marine Aviation Transport, puis les deux entreprises ayant fournis les informations nécessaires à l'élaboration de ce mémoire vont être présentées, un bref historique sur chacune des activités sera également fait.

Si nous nous focalisons principalement sur l'assurance transport, et plus principalement sur les études menées par le marché français de l'assurance maritime transport et aviation ou PARISMAT, la prime globale (affaires directes et **acceptations**) représente 2md€ en 2020. Cette prime regroupe les garanties corps, (maritime, fluvial, lacustre et aviation) qui indemnise l'assuré en cas de de disparition de l'appareil, ou en cas de dégâts subis par l'appareil. Les facultés, qui regroupent les marchandises transportées (maritimes et aviation) ainsi que la Responsabilité Civile qui indemnise les tiers en cas d'accident (maritime, fluviale, lacustre et aviation).

Ce chiffre d'affaires global se scinde en trois grandes familles, détaillées dans la figure 8 ci-dessous :

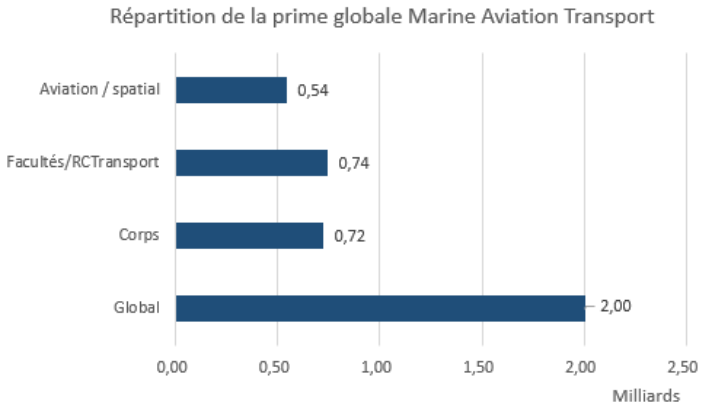


Figure 8-Répartition de la prime de la branche transport selon la ParisMAT.

L'activité aviation et spatial représente la plus petite part de ce marché, avec 0.54Md€ de primes. Le maritime est représenté en partie par les facultés/Responsabilité Civile Transport et le Corps.

Réunion Aérienne et Spatiale ainsi qu'Helvetia sont deux principaux acteurs sur chacune de ces activités. Nous allons les présenter dans les chapitres suivants.

Présentation Réunion Aérienne et Spatiale : La Réunion Aérienne

Depuis plus de 70 ans, La Réunion Aérienne (LRA) est un acteur majeur de l'assurance aviation en France et dans le monde, en assurant et réassurant une clientèle de compagnies aérienne, d'industriels de l'aéronautique, d'aéroports ainsi que l'ensemble des acteurs de l'aviation générale. La souscription de ces affaires se fait par deux plateformes, l'une à Paris, l'autre à Londres.

L'activité de Réunion Aérienne et Spatiale (RA&S) est développée avec l'appui de grands groupes d'assurance européens et chinois, permettant ainsi aux intermédiaires et aux clients d'obtenir un haut niveau de sécurité financière. Réunion Aérienne & Spatiale souscrit pour les deux marques commerciales LRA et LRS (La Réunion Spatiale), et la capacité de souscription est donnée par ses cinq mandants.

L'organisation de l'activité de RA&S peut se résumer par la figure 9 suivante :

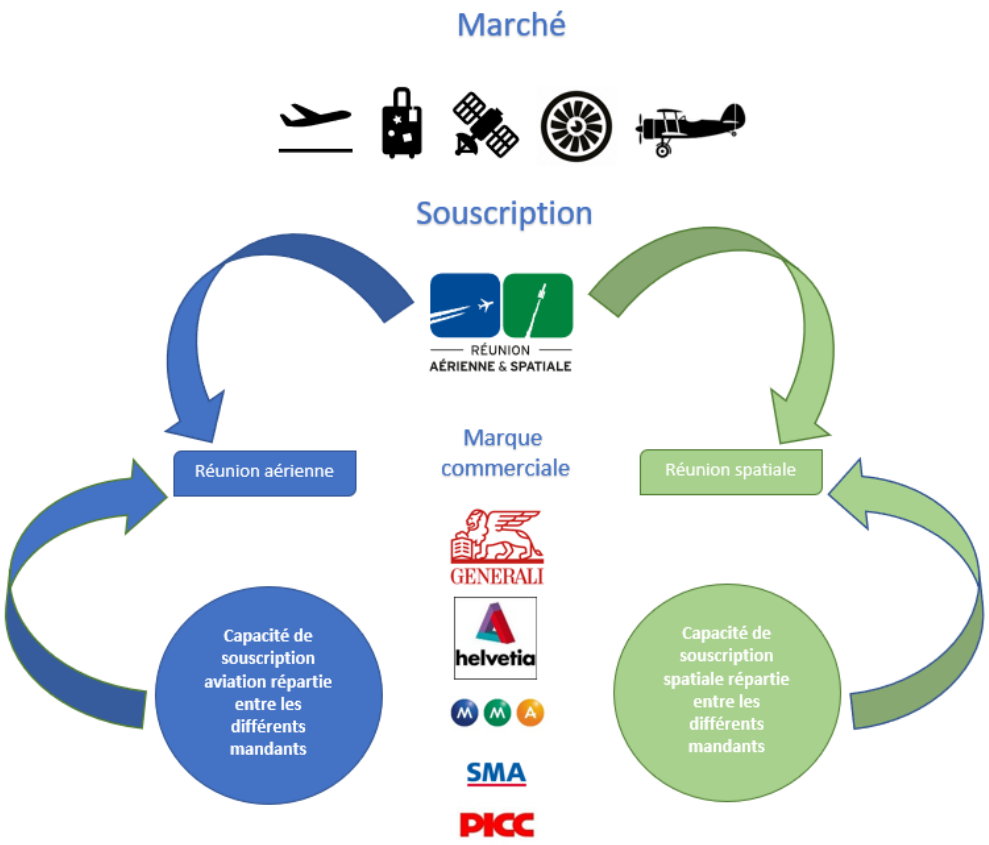


Figure 9-Illustration ddu fonctionnement de LRA.

Présentation de l'assurance aviation

Si nous remontons aux prémices de l'assurance aviation, seul le corps de l'aéronef était couvert, et même cette garantie n'était pas obligatoire. Les autres couvertures comme la Responsabilité Civile des tiers, pouvaient s'ajouter en complément, en faisant appel à d'autres types d'assurances. Ce n'est qu'après le développement plus commercial et de loisir de l'aviation, que les assureurs maritimes proposèrent la première couverture de transport aérien. Plus tard, l'*International Union of Aerospace Insurers* ou IUAI fondée en 1934, proposa de gérer l'entièreté de l'assurance aviation, ils représentent encore aujourd'hui les acteurs de l'assurance et réassurance aviation. De nos jours, l'assurance Responsabilité Civile est obligatoire, l'assurance du corps de l'aéronef, ainsi que les risques Corps de guerre et Responsabilité Civile de guerre sont facultatifs.

Trois grandes activités composent l'assurance aviation :

- **Airlines** : Principalement des compagnies aériennes, possédant de grandes flottes, ou des appareils d'une valeur élevée, effectuant des vols internationaux ou nationaux. (American Airlines/ Lufthansa / Ethiopian Airlines).
- **Aviation générale** : Assurance d'appareils de plus petites tailles, possédés par des amateurs, des collectionneurs, ou encore l'assurance de manifestations aériennes ponctuelles.
- ***Manufacturers*** : Regroupant principalement des constructeurs, ravitailleurs, aéroports ou encore des services de maintenance au sol.

L'assurance des trois activités principales se décline en trois grandes familles de garanties, que nous avons évoquées précédemment :

- **La Responsabilité Civile** (ou RCPax) : Il s'agit de la couverture obligatoire depuis 2004, qui couvre les dommages matériels ou corporels causés aux tiers, ou aux passagers.
- **Le Corps** : C'est une garantie optionnelle, qui couvre les dommages subis par l'aéronef.
- **La Responsabilité Civile Produit** (ou RC professionnelle) : Il s'agit de la garantie qui couvre les dommages causés aux assurés ayant une activité *Manufacturers*.

L'assurance aviation étant une assurance de niche, elle créée peu de primes et peu de sinistres en termes de nombre. Pourtant, ce secteur est particulièrement changeant, celui-ci évoluant en fonction du trafic aérien et des évolutions techniques. Ce développement rapide fut stoppé en 2019, puisque l'activité s'est brutalement interrompue avec la crise sanitaire, ramenant tout le secteur au niveau des années 2000, soit à peine deux milliards de personnes transportées. D'après l'association du transport aérien international (IATA) un retour à une activité normale est estimé autour de 2023, l'activité de 2019 pourrait même être dépassée en 2024.

Les risques couverts sont eux de plus en plus complexes, et extrêmement volatiles puisqu'un assuré peut avoir une limite de garantie maximum (Corps + RC) jusqu'à deux fois supérieure à la prime globale mondiale collectée. D'après Paulin C., souscripteur chez Beazley France « Au cours des 10 dernières années, le marché de l'assurance aviation générale aux US a été touché par un sinistre excédant 50m$ en RC tous les 1,2 ans, et excédant 100m$ tous les 2,2 ans » (cf. Ainonline, 07.05.2018).

Dans ce mémoire, afin de pouvoir comparer la qualité des résultats obtenus sur différents jeux de données, plus ou moins sinistrés, et à développement plus ou moins long, une deuxième base, de sinistres maritimes (LoB MAT), sera étudiée. Celle-ci a été fournie par Helvetia, acteur majeur en assurance maritime.

Présentation de Helvetia

Helvetia France fondé en 1921, est l'un des piliers du Specialty Market d'Helvetia Groupe, lui-même fondé en 1858. Il s'agit du deuxième assureur sur la branche Maritime et Transport en France. Implantés dans plusieurs sites en Outre-Mer, Afrique, et France métropolitaine, leurs activités s'étendent sur plusieurs domaines, dont le Fine art, la construction, les risques techniques et bien d'autres. Depuis 2019 ils ont élargi encore un peu plus leurs activités en ajoutant l'aviation et le spatial, grâce à LRA.

Concernant le maritime, les types d'activité suivantes sont assurées par Helvetia France :

- Armateur
- Fluvial : assurance des péniches et autres bateaux fluviaux (passagers et transport)
- Marchandises transportées
- Pêche : assurance dommage ainsi que responsabilité civile
- Plaisance : navigation de loisir
- Portuaire : chantiers navals, autorités portuaires, manutention portuaire, services maritimes
- Transport et logisticiens

Malgré un portefeuille diversifié, par souci d'homogénéité des données, nous n'étudierons que les activités fluviales, pêche et plaisance dans la suite de ce mémoire.

L'assurance maritime est l'une des plus ancienne créée, elle est aujourd'hui bien différente de ce qu'elle a été autrefois. Une brève présentation de l'assurance aviation et de son évolution est abordée dans la partie suivante.

Présentation de l'assurance maritime

Comme développé par M.Boyer (2008), les assurances elles-mêmes ont été développées par les marins, pour des marins, entre la fin du XIIème siècle et le milieu du XIVème siècle. Non pas pour une question de valeurs de marchandises, mais principalement parce que le capitalisme moderne a commencé dans le commerce maritime, avant qu'il n'apparaisse sur terre. En effet, le nombre de navires était conséquent, les routes terrestres étant peu praticables, et la quantité de marchandises transportées en une fois était bien plus grande que sur terre. De surcroît, si les navires provenaient de Moyen-Orient et du Proche-Orient, ces cargaisons étaient bien plus précieuses, avaient un investissement initial élevé, et les périples étant long, il y avait par conséquent un risque plus élevé de brigandage et piraterie. L'assurance maritime était née et continua à se développer dans les ports de la Méditerranée puis de l'Atlantique.

Aujourd'hui, seule l'assurance responsabilité civile professionnelle est obligatoire. Par souci de simplicité, s'il y a transport maritime, l'assurance maritime va couvrir l'ensemble du trajet, même si une partie est effectué par un autre mode de transport.

Comme Broudin R. directeur mondial de l'indemnisation en assurance maritime chez AGCS le décrit pour Allianz, le marché est volatil, avec une augmentation de la sinistralité ces dernière années. En effet, les navires sont aujourd'hui plus grands, le risque cyber est également plus présent, par l'augmentation du nombre de systèmes de connectivité avec les ports, augmentant ainsi le risque de piratages et par conséquent le risque d'échouements ou de collisions. De plus, des sinistres incendie de grande ampleur sur des porte-conteneurs ont été nombreux ces dernières années et ont marqués le marché, comme les sinistres *Marsk Honam*, *Yantis Express* et *APL Vancouver*. Ces types d'incendies se caractérisent par un fort dommage matériel, de fortes dépenses sur les avaries, ainsi qu'un coût d'opération de sauvetage bien plus élevé, par la taille du navire.

En conclusion les branches aviation et maritime font partie des grands risques, comme définis dans le code des assurances. Toutes deux produisent peu de sinistres et peu de primes, mais les sinistres associés peuvent être très élevés. Ces branches sont fortement sensibles aux variations de portefeuille, ainsi qu'aux changements liés aux nouvelles technologies et aux nouveaux risques tels que les risques cyber.

1.2 PROVISIONNEMENT EN ASSURANCE AVIATION ET MARITIME

Nous allons maintenant aborder le fonctionnement des sinistres en aviation et maritime, en commençant par le cycle de vie standard d'un sinistre, puis nous poursuivrons par la décomposition des provisions de sinistres, pour terminer par Solvabilité 2, afin de comprendre l'importance de ces provisions dans la règlementation.

1.2.1 Généralités sur les sinistres

Cycle de vie d'un sinistre

Contrairement à l'assurance vie, pour laquelle l'incertitude porte principalement sur la date de réalisation du risque, **l'incertitude en assurance non vie porte sur la réalisation du risque ou non**, la date de l'évènement, ainsi que son montant total. Cette incertitude provient de l'inversion du cycle de production, évoqué précédemment. Il est primordial pour l'assureur d'évaluer au mieux les provisions nécessaires au règlement total des sinistres.

Afin de bien appréhender la suite de ce mémoire, il faut comprendre l'articulation générale de la vie d'un sinistre, qui se résume dans la figure 10 ci-dessous :

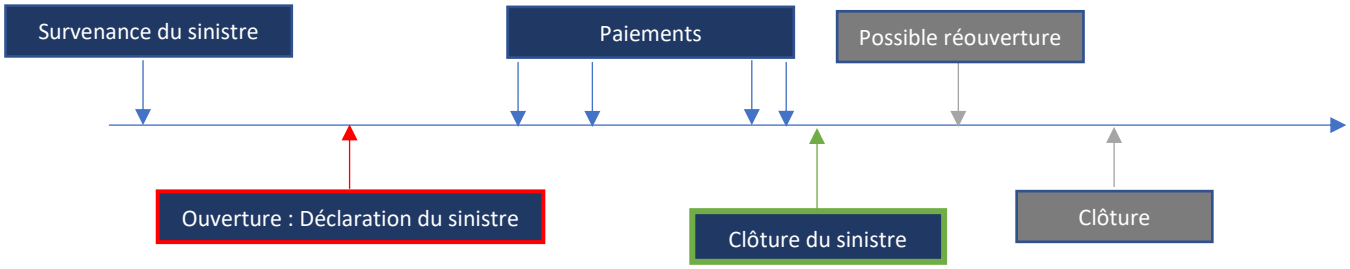


Figure 10-Cycle de vie d'un sinistre.

Comme l'illustre le schéma ci-dessus, une fois le sinistre survenu, celui-ci est déclaré peu de temps après à l'assureur. Celui-ci évalue le montant total probable du sinistre, puis paie au fur et à mesure l'indemnisation à l'assuré, jusqu'à la clôture du sinistre.

Le cout d'un sinistre n'est pas connu à l'avance, par conséquent, la charge totale entrée par le gestionnaire se sépare en deux parties :

- Les paiements : représentent les différents versements ayant eu lieu jusqu'à la date t, souvent observés en cumulés au travers de ce mémoire.
- Les provisions dossier à dossier : établies à dire d'expert, afin d'approcher au mieux le montant total final à payer. Ces provisions sont réestimées autant de fois que nécessaire au cours de la vie du sinistre.

Il est important de noter que les paiements ne sont pas toujours positifs, il arrive de percevoir des recours, et dans ce cas, un remboursement à lieu, d'où les incréments négatifs qui peuvent être observés par moments. Dans la majeure partie des cas, un recours à lieu si l'assuré indemnisé était finalement non responsable, ou partiellement, et dans ce cas une partie de l'indemnité doit être récupérée.

Si la déclaration d'un sinistre a lieu quelques temps après sa survenance, il s'agit d'un sinistre tardif. Ce sinistre a donc eu lieu, mais n'a pas encore été notifié auprès de l'assureur, dans ce cas aucun sinistre n'a été entré dans les systèmes.

Une fois déclaré, un sinistre commencera par une majeure partie des fonds en provisions dossier à dossier. Ces provisions passeront petit à petit en paiements tout au long de la vie du sinistre, jusqu'à sa clôture, lorsque les provisions dossier à dossier atteignent zéro. Chaque sinistre suivra en moyenne cette évolution

(hors recours, baisse inhabituelle des provisions après étude des experts…). Le développement sera différent en termes de durée, en fonction du risque observé, et de la gravité du sinistre.

Comme nous pouvons le voir dans les graphiques ci-dessous, les sinistres seront clos plus ou moins rapidement (Figure 11) :

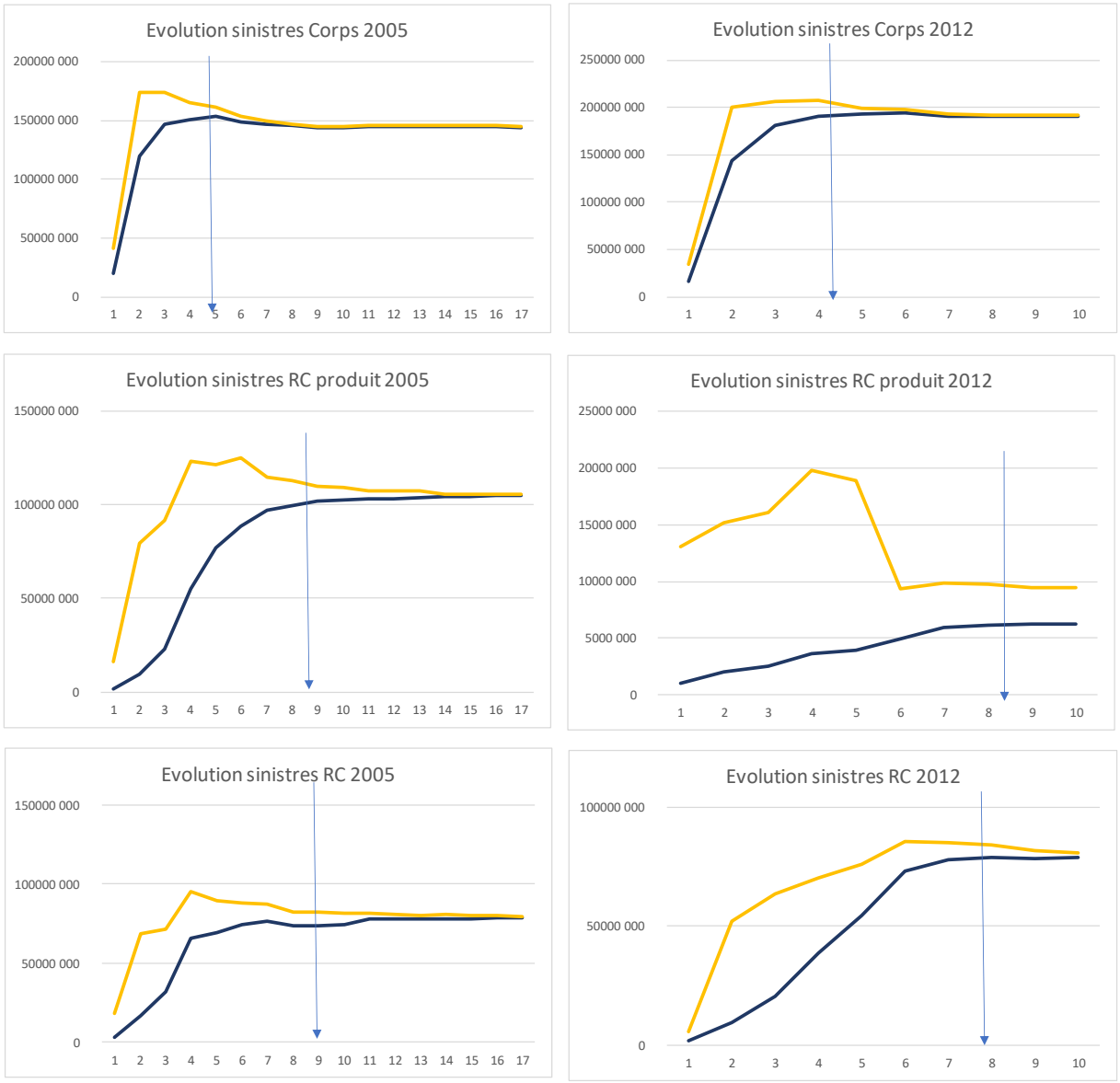


Figure 11- Evolution de la charge sinistre sur les exercices 2005 et 2012 selon le risque observé.

Qu'il s'agisse de l'exercice 2005 ou 2012 ci-dessus, les sinistres corps s'approchent de leur charge ultime en 5 ans, alors que la Responsabilité Civile ou Responsabilité Civile produit prendra environ 9 ou 10 ans.

La sinistralité future ne se traduit pas seulement par les provisions dossier à dossier. Les provisions techniques, abordées dans la partie suivante, viennent ajouter d'autres types de provisions, qu'il est nécessaire d'aborder pour la suite du mémoire.

Décomposition des provisions techniques

Par définition, les provisions techniques représentent les dettes futures probables de l'assureur, à l'égard de ses assurés. Il s'agit de l'ensemble des provisions évaluées et suffisantes pour le règlement intégral des sinistres passés non réglés et sinistres futurs.

Ces provisions se décomposent en plusieurs catégories :

- **Provisions pour Primes Non Acquises ou** PPNA : Représentent la portion des primes nécessaires pour couvrir les risques futurs.
- **Provisions dossier à dossier** : définies au chapitre précédent
- ***IBNeR*** : *Incurred but Not enough Reported*, dans le cas où un sinistre déjà enregistré connaitra une hausse de sa charge totale.
- ***IBNyR*** : *Incurred But Not yet Reported*, dans le cas où un sinistre a eu lieu, mais n'est aujourd'hui pas encore enregistré dans le portefeuille de l'entreprise, il s'agit des tardifs, cités précédemment.

Les *IBNeR* et *IBNyR* rassemblés dans les ***IBNR*** (*Incurred But Not Reported*) sont établis par l'actuaire, une à plusieurs fois dans l'année, pour estimer au mieux l'évolution de la charge totale des sinistres portés par l'entreprise.

Il existe plusieurs méthodes permettant l'estimation des *IBNR*, la plus courante et la plus simple étant la méthode du *Chain Ladder*. Ces méthodes établissent une charge ultime, à partir de laquelle un *Best Estimate* (BE_{LRA}), la meilleure estimation des sinistres futurs, est calculé. Pour LRA, ce montant n'est pas actualisé, chacun des mandants obtient le Best Estimate brut, calculé grâce à la formule ci-dessous, et y appliquent par la suite leurs propres hypothèses d'actualisation :

$$BE_{LRA} = Provisions\ dossier\ à\ dossier + IBNR.$$

Dans le reste du mémoire, lorsque le Best Estimate sera évoqué, il s'agira du BE_{LRA}.

Finalement, les provisions techniques peuvent être résumées par la figure suivante (Figure 12) :

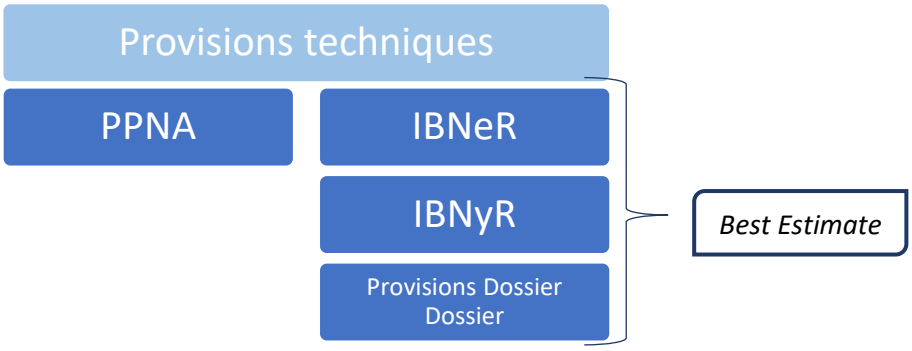


Figure 12-Répartition des provisions techniques et du Best Estimate.

Dans le cas où les provisions sont nettement supérieures à ce qui était nécessaire, il s'agit de **sur-provisionnement** celui-ci aura pour effet de réduire artificiellement le résultat comptable de l'entreprise, et augmenter les taxes sur les bonis. Dans le cas inverse, si les provisions comptabilisées sont inférieures à la sinistralité à venir, il s'agit d'un cas de **sous-provisionnement,** qui entrainera une incapacité de la part de l'entreprise, à indemniser ses bénéficiaires. Ce manque de fiabilité, entrainera un manque de confiance de la part des assurés, et finalement une augmentation non négligeable du risque de ruine.

Ces provisions techniques prennent une part importante dans la règlementation Solvabilité 2, celle-ci sera décrite dans le chapitre suivant afin de mieux appréhender l'importance du *Best Estimate*, et donc de l'estimation de la charge ultime en assurance.

1.2.2 Solvabilité 2 et le risque de réserve

Solvabilité 2 est le régime prudentiel imposé depuis le premier janvier 2016, aux entreprises d'assurance et réassurance européennes.

Certains des grands objectifs de Solvabilité 2 sont l'amélioration de la protection des preneurs d'assurance et des ayants droits, ou encore la promotion d'une meilleure réglementation. Il était aussi primordial d'harmoniser les normes au sein de l'UE, et au sein du secteur financier. Toutes ces exigences peuvent être résumées en trois piliers, que nous présentons ci-dessous.

Importance du Best Estimate dans la règlementation solvabilité 2

Figure 13-Les exigences de capital sous Solvabilité 2.

Comme le résume l'ACPR (Figure 13), **le premier pilier** regroupe les exigences quantitatives, comme les règles de valorisation des actifs et passifs, les exigences de capital, et les méthodes de calcul. C'est dans ce pilier que s'appliquent les modèles standards, internes ou partiels.

Dans le pilier 1, Solvabilité 2 prévoit plusieurs exigences, portant sur les provisions techniques. L'exigence de capital, le MCR *Minimum Capital Requirement*, le SCR *Solvency Capital Requirement*, et les actifs éligibles pour les couvrir. Les provisions techniques doivent être prudentes, fiables et objectives, celles-ci permettent les comparaisons entre les différents assureurs. Le SCR, lui, remplace l'exigence de marge de solvabilité de Solvabilité 1, et peut être définie comme un capital cible, qui doit être couvert par un montant équivalent de fond propre éligible. Il correspond au montant de fond propres que l'entreprise doit détenir pour limiter sa probabilité de ruine à un an, à 0.5%. Ce calcul s'effectue une fois par an, et est recalculé si le profil de risque de l'entreprise varie sensiblement. Ce SCR s'obtient à l'aide d'un modèle standard, interne, ou partiel, sous l'hypothèse d'une continuité d'activité de l'entreprise.

Si nous nous concentrons sur la formule standard, et plus précisément, sur le module « risque de souscription en non-vie », qui reflète le risque portant sur les engagements d'assurance non-vie, nous remarquons que plusieurs sous modules y sont attachés (Figure 14) :

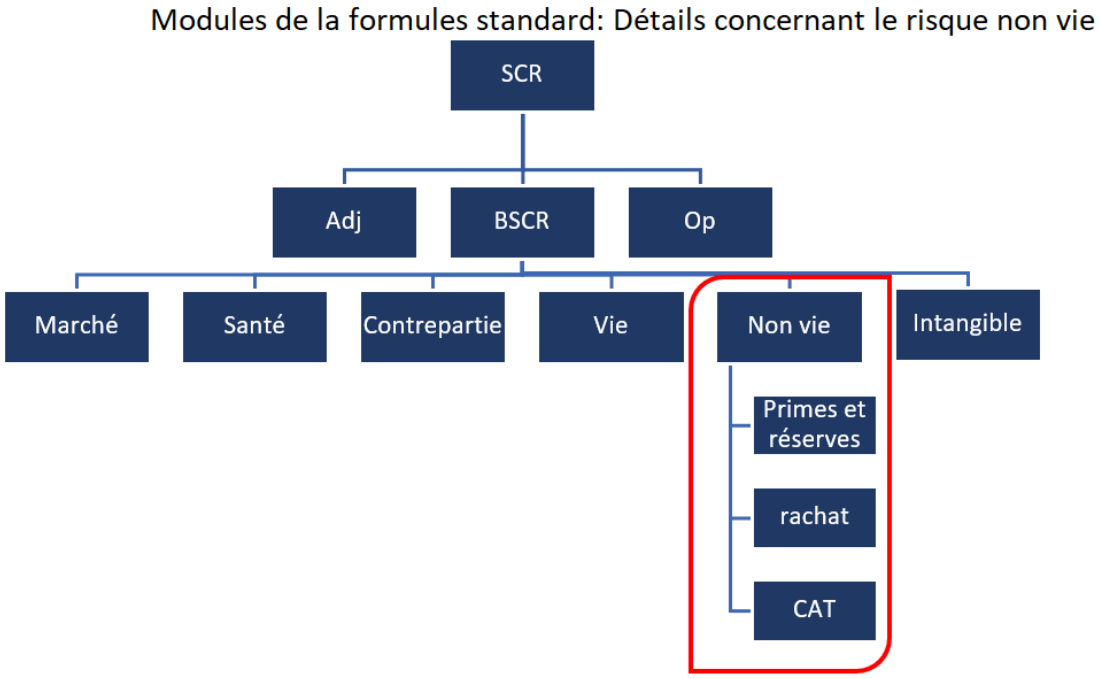


Figure 14-Arbre des modules non vie de solvabilité 2.

Les modules représentés ci-dessus contiennent quatre notions fondamentales :

- Le risque de primes : Risque que les primes perçues soient inférieures aux coûts des sinistres futurs.
- Le risque de réserve : Risque de perte lié à la fréquence et la gravité des évènements assurés.
- Le risque de rachat : Risque de pertes, résultant de la volatilité du taux de cessation, d'échéance, de renouvellement et de rachats de polices.
- Le risque de catastrophe (CAT) : Risque de perte lié aux évènements extrêmes ou exceptionnels.

L'estimation des provisions à l'ultime fait partie intégrante du risque de réserve, et donc du calcul du SCR non vie. Il est par conséquent très important d'utiliser des méthodes qui permettent d'apprécier au mieux le *Best Estimate*, permettant ainsi d'obtenir le SCR le plus juste possible.

1.3 PROBLEMATIQUE ET OBJECTIF

Les données de LRA vont être détaillées dans le chapitre ci-dessous, ainsi que la méthode d'estimation de la charge ultime. La méthode appliquée par Helvetia étant similaire, celle-ci ne sera pas détaillée ici, mais dans la partie liée au provisionnement usuel. Par la suite, les limites de cette méthode seront exposées, afin d'introduire la problématique et les modèles testés dans ce mémoire.

1.3.1 Données et Méthodologies : Limites des modèles en aviation et maritime

RA&S est une agence de souscription, dont la capacité est fournie par différents mandants. Chacun porte une part (clé de souscription du risque) différente des risques, et calcule son SCR non vie en tenant compte des données et résultats de la souscription dans chacun des sous modules. LRA gère le suivi des primes et des sinistres aviation pour les mandants tout au long de l'année, en fournissant de nombreuses informations importantes comme tous types de provisions comptables, le BE_{LRA} ou encore le risque CAT.

Données LRA

Même si LRA ne dispose pas de capitaux propres, et n'a pas à calculer son SCR, de nombreuses informations sont fournies aux mandants, dont les PPNA, le BE_{LRA} et le risque cat.

Sont calculés régulièrement :

- Le risque **de prime**, estimé grâce aux Provisions pour Primes Non Acquises (PPNA), qui représentent la portion des primes émises qui correspondent aux risques non encore encourus. La méthode appliquée est celle du *prorata temporis*.

  Soit e la date d'expiration du contrat, c la date d'inventaire, n le nombre de jours de l'année en cours, et P la prime, nous obtenons : $PPNA = \frac{e-c}{n} P$.

- Le risque **de réserve**, estimé grâce aux études actuarielles de provisionnement, qui se fait sur des triangulations par année de souscription des charges sinistres cumulées. Les méthodes traditionnelles de provisionnement sont appliquées.

- Le risque **de rachat**, qui s'applique principalement pour les contrats pluriannuels, les contrats de LRA étant annuels, le risque est estimé nul. Il ne sera pas abordé dans la suite.

- Le risque **catastrophe (CAT)** est principalement Man Made, et est estimé grâce à deux scénarios internes.

  L'un des scénarios, Tenerife, consiste en la reprise d'un évènement concernant la collision entre deux appareils, KLM 4805 et Pan Am 1736 ayant eu lieu à l'aéroport de Los Rodeos à Tenerife. Ce sinistre est vu « AsIf », en reprenant les deux plus gros engagements du portefeuille actuel, puis en mettant à jour les limites de garantie RC actuelles, selon la nationalité des passagers dans les appareils. Une fois la totalité de l'engagement calculé, la période de retour de l'évènement est calculée, et sa couverture par le programme de réassurance actuel est appliqué, afin d'établir si le programme de réassurance couvre suffisamment le plus gros sinistre probable sur le portefeuille.

En fin d'année, le BE_{LRA} de réserve et le BE de prime sont estimés. Par conséquent, tous les fin novembre, lors du *Fast Close*, un BE_{LRA} est calculé, et donné aux mandants. Cette estimation se base sur des données converties en USD à un taux unique, reprenant la sinistralité du portefeuille triangulée par année de souscription sur un périmètre de 15 ans. Les données sont converties à un taux unique, permettant de ne projeter que la sinistralité, mais pas les variations de taux de change. Une projection par *Chain Ladder* est effectuée, ou par *Bornhutter Ferguson,* méthode déterministe basée sur des ratios sinistres sur primes, pour les exercices les plus récents.

A ces informations s'ajoutent les cadences de développement des primes et des sinistres, ainsi que le ratio des provisions comptables nettes sur provisions comptables brutes afin de permettre aux assureurs d'estimer le BE_{LRA} net à partir du BE_{LRA} brut. Le risque de défaut des réassureurs est lui aussi estimé grâce à la note moyenne des provisions cédées.

Concernant le BE de primes, les PPNA sont fournis, ainsi que la cadence de développement primes. Le ratio des PPNA bruts de réassurance sur PPNA nets de réassurance permettent de passer d'un résultat brut à un résultat net, le risque de contrepartie est encore une fois estimé grâce à la notation moyenne du défaut des réassureurs.

Toutes ces informations sont par la suite regroupées par nos mandants, pour y appliquer leur modèle de calcul de la solvabilité.

De son côté, pour le risque de réserve, Helvetia procède de façon quasi similaire à LRA lors de leur provisionnement en maritime. Les sinistres sont projetés par la méthode du *Chain Ladder*, puis le BE_{LRA} en est déduit sur toutes les années de souscriptions étudiées, soit dix ans dans ce mémoire. L'étude du provisionnement maritime sera détaillée dans la suite de ce mémoire, tout comme pour la base aviation.

La méthode du *Chain Ladder* est utilisée fréquemment lors du provisionnement. Cependant, celle-ci ne doit s'appliquer que sous certaines conditions, qui sont rarement respectées. C'est pourquoi nous allons nous intéresser aux limites de ce modèle.

Limites des méthodes et problématique

Les méthodes de projection du type *Chain Ladder* sont les plus rapides et les plus simples à mettre en œuvre. Toutefois, celles-ci demandent de valider plusieurs hypothèses afin de les appliquer :

- Le portefeuille doit contenir des données fiables
- Les données du portefeuille doivent être nombreuses
- Le passé doit être régulier
- La branche doit être peu volatile
- Le passé, le présent et le futur doivent être structurellement peu différents

Dans le domaine de l'assurance aviation et maritime, les données sinistres sont peu nombreuses par rapport à une activité comme l'automobile, et bien plus volatiles.

En aviation, les évolutions techniques ainsi que l'augmentation du trafic aérien (1.8 milliards de passagers à 4.5 milliards en 2019) impliquent que le passé est peu représentatif du présent, et l'arrivée de la pandémie en 2020, ajoute une grande part d'incertain dans les exercices de souscription les plus récents (principalement 2019 et 2020). Du point de vue de LRA, des modifications de souscription impactent aussi la stabilité du portefeuille, plus particulièrement sur le risque Responsabilité Civile produit de l'activité *Manufacturers*. La part de LRA prise sur chaque risque varie elle aussi d'un exercice à un autre.

En maritime, les évolutions techniques ont également modifié la sinistralité au fil du temps. Des modifications de portefeuille ont également impactées la base maritime qui sera étudiée, augmentant par exemple la sinistralité de l'exercice 2020 par rapport aux exercices de souscription précédents.

Dans ce cas, l'application des méthodes de *Chain Ladder* est peu adaptée à ces données. C'est pourquoi d'autres modèles, permettant de capter les spécificités du portefeuille seront présentés dans le chapitre suivant, puis testés dans ce mémoire. Ces dernières seront testées sur les deux bases, et comparées à leur résultat du provisionnement *Chain Ladder* usuel respectif, sur l'exercice de souscription 2020.

Deux bases sont utilisées afin d'observer les différences de prédictions sur des données différentes en termes de sinistralité. En effet, la base maritime fournie par Helvetia, est plus volumineuse, moins volatile et à durée de vie de sinistre plus courte qu'en aviation.

1.3.2 Modèles testés

Comme le décrivent Dubois D. et al. (2021) les entreprises d'assurance font face à différents défis, l'un d'eux étant la révolution numérique, qui aujourd'hui impacte grandement le secteur. Il est donc important de remettre en cause certaines pratiques métier, afin de les modifier ou de les perfectionner. D'après L'ACPR « La donnée sera au cœur des modèles économiques futurs, que ce soit pour son utilisation commerciale ou, plus naturellement en assurance, pour une tarification plus efficace ».

Les actuaires ont été parmi les premiers à s'intéresser aux *Big Data* et au *Machine Learning*. Le *Big Data* repose sur les 5V, le volume, la variété, la vélocité, la véracité, et la valeur (Figure 15) :

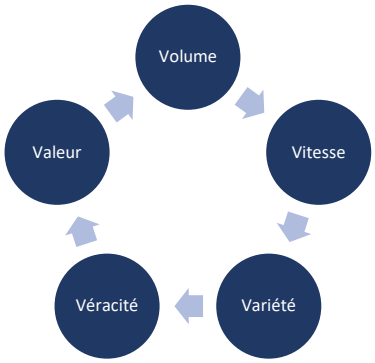


Figure 15-5V

- **Le volume** représente la taille des bases de données,
- **La variété** représente tous les différents formats de données disponibles, (images/ textes/ chiffres…),
- **La vélocité** fait référence à la vitesse de production et circulation des données,
- **La véracité**, représente la quantité de données et la rapidité avec laquelle les données sont générées,
- **La valeur**, qui représente la valeur ajoutée de ces données.

Beaucoup de travaux actuariels tels que le provisionnement, la tarification ou l'estimation du besoin en capital, sont aujourd'hui effectués à l'aide de ces méthodes, ou avec des méthodes semblables. L'actuaire *data scientist* doit être capable d'allier expertise métier, et connaissance en *data* science, afin entre autres, de ne pas remplacer les méthodes actuelles, mais de les perfectionner grâce aux nouvelles données disponibles via le *Machine Learning*.

Dans notre cas, il s'agira d'obtenir la charge ultime de l'exercice de souscription 2020 selon différentes méthodes de *Machine Learning,* en se basant sur les informations disponibles sur les sinistres et les assurés, de 2013 à 2019 pour l'aviation, et 2010 à 2019 pour le maritime. Pour ce faire, nous utiliseront différentes méthodes, afin d'obtenir le résultat le plus fiable possible.

Les cinq méthodes suivantes seront implémentées sur le portefeuille sinistres aviation, et sur le portefeuille sinistre maritime afin d'apprécier la fiabilité des résultats sur deux branches différentes.

Arbres de décision (CART)

Les méthodes dites de partitionnement récursif ou de segmentation, datent des années 1960. Elles sont par la suite formalisées par Breiman en 1984, sous l'acronyme de *CART : Classification And Regression Tree*.

L'algorithme CART s'applique à deux situations distinctes, la prévision de résultats qualitatifs (classification) ou quantitatifs (régression). Le but de la méthode est de créer un modèle, permettant de prédire la valeur d'une variable cible, à partir de plusieurs variables d'entrées. L'arbre est construit par répartition récursive de chaque nœud, en fonction de la valeur de l'attribut testé à chaque itération. Le processus récursif s'arrête quand les éléments d'un nœud ont la même valeur pour la variable cible.

La figure suivante (figure 16) reprend la forme générale d'un arbre de décision :

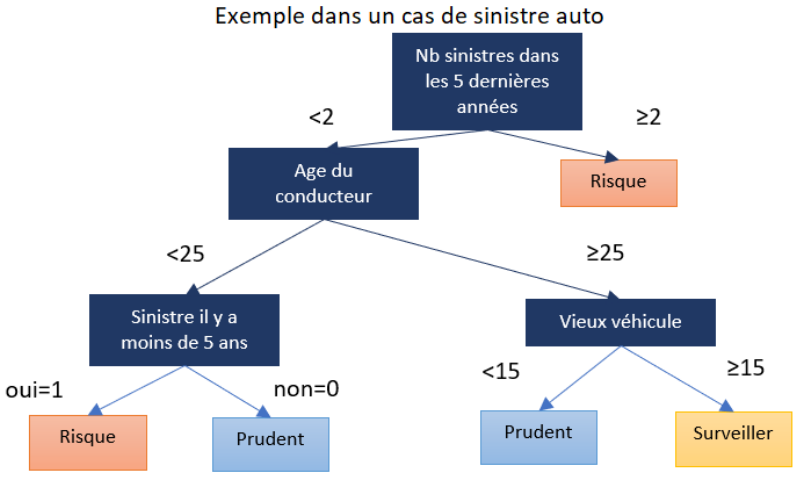


Figure 16-Arbre de classification sur un cas d'assurance auto (Source Dubois D. et al. 2021).

Ces arbres sont simples d'interprétation, et gèrent aisément de grandes quantités de données. L'arbre ci-dessus classe en trois catégories les assurés (Risque/Prudent/Surveiller) selon les informations disponibles dans les bases.

Cette méthode est la plus simple à mettre en œuvre et à comprendre, et constitue une base solide pour les méthodes suivantes construites sur ce type de modèle, telles que les forêts aléatoires.

Forêts aléatoires

Le principe se base sur la méthode CART, où N arbres de décision sont calibrés sur des sous-échantillons créés à partir des données disponibles. La prédiction finale découle d'un vote majoritaire des résultats des N arbres (figure 17).

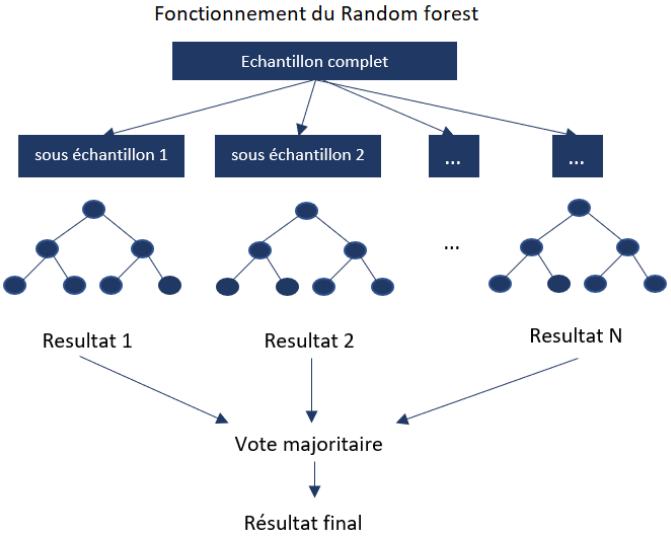


Figure 17-Représentation de la méthode des forêts aléatoires.

De la même manière que pour la méthode CART, les méthodes suivantes permettent de perfectionner ce modèle, afin d'estimer au mieux la prédiction.

Gradient boosting et XGBoost

Les méthodes du *Gradient Boosting* ou boosting de gradient, dérivent du *boosting*, qui est une méthode permettant de créer un « prédicteur fort » ou *strong learner*, à partir de multiples « prédicteur faibles » ou *weak learners.* Dans le cas du *Gradient boosted tree*, nos prédicteurs faibles seront les arbres de décisions.

Le *Gradient Boosting* va se baser sur deux approches fondamentales :

- Combiner de manière séquentielle des prédicteurs faibles, pour en créer un plus fort.
- Entrainer chaque arbre de décision sur une version modifiée de l'échantillon d'entrainement.

Ce deuxième point implique que le nouveau prédicteur se focalisera sur les endroits où le précédent n'avait pas été suffisamment bon. Dans ce cas, plus de poids sera donné aux échantillons de données pour lesquels l'erreur était importante.

XGBoost ou *Extreme Gradient Boosting* est une méthode reprenant la structure générale du *Gradient boosting*, en ajoutant plus de paramètres, permettant une meilleure estimation des résultats, ainsi qu'une diminution du temps de calcul.

Pour terminer, une méthode ne reposant pas sur des arbres de décision, la régression *Elastic Net*, régression régularisée grâce aux pénalités de Ridge et LASSO, sera également appliquée aux jeux de données.

Elastic Net

La méthode *Elastic Net*, introduite par ZOU, H. & HASTIE, T. en 2005, se base sur les modèles de régression simple de la forme :

$$Y = \beta X + \varepsilon.$$

Y représente la variable à expliquer et X les variables explicatives.

La régression *Elastic Net* ajoute une pénalisation de type L_2, ainsi qu'une pénalisation de type L_1.

Nous pouvons écrire L_2 comme :

$$\|\beta\|_2^2 = \sum_{j=1}^{p} \beta_j^2.$$

Et L_1 comme :

$$|\beta| = \sum_{j=1}^{p} |\beta_j|.$$

La régression de type *Elastic Net* va s'écrire :

$$L_{ENet} = argmin_{\widehat{\beta}} \left(\|Y - \beta X\|^2 + \lambda_1 |\beta| + \lambda_2 \|\beta\|_2^2\right).$$

En conclusion de cette première partie, l'aviation et le maritime génèrent une sinistralité plus volatile que les branches d'assurance non vie classiques, mais bien plus réduite en nombre. La souscription a grandement changé au cours du temps, par la modification des risques, et l'arrivée de nouveaux contrats. Par conséquent, la sinistralité passée n'est plus toujours représentative de la sinistralité actuelle. Dans ce cas, l'application de la méthode du Chain Ladder, se basant sur une sinistralité peu volatile, nombreuse et structurellement peu différente du passé, n'est potentiellement pas la méthode adéquate afin d'estimer la charge ultime.

C'est pourquoi des méthodes par partitionnement récursif et un modèle Elastic Net seront testés sur les deux jeux de données, afin de prédire la charge ultime obtenue pour l'exercice de souscription 2020 et ainsi challenger les résultats obtenus par la méthode Chain Ladder.

Deux jeux de données sont testés, afin d'observer les différences de prédictions, sur des types de bases plus ou moins fournies, plus ou moins volatiles, et à durée de vie de sinistre plus ou moins longue.

2 Provisionnement usuel

Il est avant tout important de reprendre succinctement les fondamentaux de la méthode du *Chain Ladder*, vérifier si les hypothèses d'application sont validées ou non sur les deux bases, et développer les deux méthodologies appliquées. Ceci nous permet d'obtenir la charge ultime de l'exercice de souscription 2020 en aviation et en maritime, qui sera notre point de comparaison avec les méthodes lignes à lignes qui seront testées par la suite.

2.1 TRIANGULATION DES DONNEES

Afin de procéder au provisionnement par *Chain Ladder*, les sinistres doivent être triangulés sur un minimum de 10 ans. Les activités aviation et maritime font parties de la branche Marine Aviation Transport, et sont provisionnées en année de souscription. Ces triangles de liquidation sont aussi appelés triangles de *run-off*. Comme développé par DENUIT, M. & CHARPENTIER, A. (2005), chaque ligne du triangle reflète la dynamique d'un exercice de souscription, dans laquelle tous les montants de sinistres de cet exercice sont agrégés. De plus, la triangulation devant être homogène, les sinistres ne sont pas étudiés en un seul triangle, mais en plusieurs triangles de même nature.

De manière formelle, le provisionnement par triangles de liquidations revient à l'établissement d'une prédiction conditionnée aux informations disponibles à un instant t , dans notre cas au 31/12/2021. A une date t, nous pouvons écrire l'information disponible comme étant :

$$D_t = \{C_{i,j}, pour\ i + j \leq t\}$$

Avec $C_{i,j}$ étant les charges cumulées, de l'exercice i.

Si nous appliquons pour la suite les notations suivantes :

- i l'indice des années de souscription
- j l'indice des années de développement
- $C_{i,j}$ les charges agrégées (ou cumulées) des sinistres de l'exercice i à chaque développement j. Si les $X_{i,j}$ sont les valeurs incrémentales du triangle de liquidation, alors $C_{i,j} = \sum_{t=0}^{i} X_{i,t}$.

La figure ci-dessous (figure 18) reprend le format de triangle utilisé, et le tableau 8, les types de triangles créés pour chaque base :

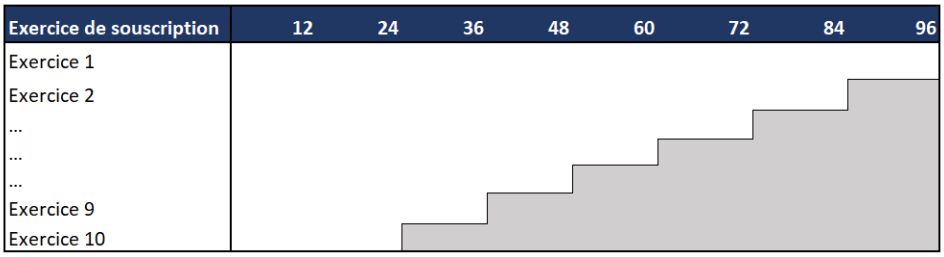


Figure 18-Schéma d'un triangle de liquidation.

Aviation	Maritime
Corps	CMM
RC	FAC
RC Professionnelle	IAR
	PAC
	RC

Tableau 8-Séparation des données en plusieurs triangles de liquidation homogènes.

Ainsi, chaque triangle créé, reprend :

- Les **exercices** en ligne, reflétant les variations de portefeuille liés à la souscription, par l'augmentation ou diminution de certaines parts, ou encore par l'augmentation de certaines activités dans le portefeuille.
- Les **années de développement**, en colonne, reflétant principalement l'évolution des sinistres (plus ou moins long) selon le type de triangle observé.
- Les **années de compte** (ou calendaires) sur les diagonales, qui reflètent les chocs de gestion.

Ces triangles peuvent par la suite être projetés par la méthode du *Chain Ladder* décrites dans le chapitre suivant. Pour l'aviation ainsi que le maritime, il s'agira de compléter la partie inférieure du triangle de liquidation des charges totales, afin de provisionner les exercices jusqu'à leur valeur ultime. La charge ultime obtenue, correspond au montant total (paiements + Provisions dossier à dossier + IBNR) attendu sur un exercice de souscription donné.

Il est possible de décomposer cette charge telle que (figure 19) :

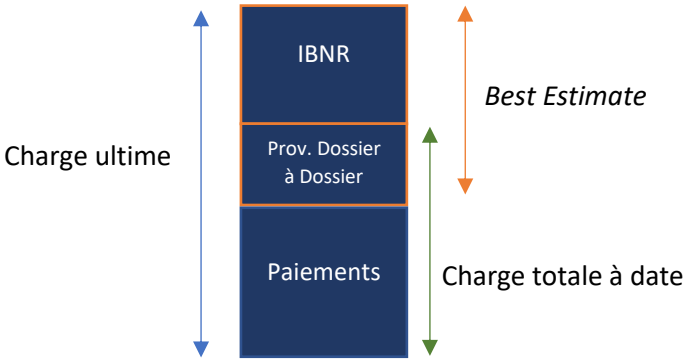


Figure 19-Décomposition de la charge ultime.

Les bases de la triangulation ayant été décrites ci-dessus nous allons maintenant détailler la méthode du *Chain Ladder*, puis l'appliquer aux triangles définis précédemment.

2.2 METHODE DU CHAIN LADDER, APPLICATION EN AVIATION ET MARITIME

La méthode du *Chain Ladder* est utilisée par la majorité des entreprises d'assurance non vie. Il s'agit d'une méthode simple à comprendre, à expliquer, et rapide à mettre en place, et de ce fait, très populaire lors de l'établissement du provisionnement. Introduite dans les années 1930, l'hypothèse principale de ce modèle repose sur le fait que les cadences de développement passées se reproduiront dans le futur.

2.2.1 Méthodologie et hypothèses

Cette méthode est déterministe, ne dépend d'aucune loi de probabilité, et repose sur deux hypothèses, dont une hypothèse forte H_o :

H_0 : Les ratios $f_{i,j} = \frac{C_{i,j+1}}{C_{i,j}}$ des facteurs adjacents sont indépendants de l'année d'origine i.

$H_1 : \forall i = 1, \forall j, E\left[C_{i,j+1} \middle| C_{i,1}, \dots, C_{i,j}\right] = f_j C_{i,j}$

Nous vérifierons la validation de ces deux hypothèses par la suite, sur chaque triangle.

Une fois ces deux hypothèses vérifiées, la méthode se déroule comme suit :

- Des coefficients de passage communs sont calculés sur le triangle étudié, et sont donnés par :

$$\hat{f}_{i,j} = \frac{\sum_{i=0}^{n-j-1} C_{i,j+1}}{\sum_{i=0}^{n-j-1} C_{i,j}} \ (0 \leq j \leq n-1).$$

- La charge ultime S_i pour chaque exercice est calculée par :

$$S_i = C_{i,n-i} \prod_{h=n-i}^{n-1} \hat{f}_h.$$

Le triangle est ainsi complété pour chaque exercice (Figure 20) :

Triangulation de charges sinistres sur 7 ans

Exercices	Dvlp 1	Dvlp 2	Dvlp 3	Dvlp 4	Dvlp 5	Dvlp 6	Dvlp 7	Charge ultime
Exercice 0	C_{01}	C_{02}	C_{03}	C_{04}	C_{05}	C_{06}	C_{07}	S_0
Exercice 1	C_{11}	C_{12}	C_{13}	C_{14}	C_{15}	C_{16}	...	S_1
Exercice 2	C_{21}	C_{22}	C_{23}	C_{24}	C_{25}	...	...	S_2
Exercice 3	C_{31}	C_{32}	C_{33}	C_{34}	...	...	...	S_3
Exercice 4	C_{41}	C_{42}	C_{43}	...	...	...	...	S_4
Exercice 5	C_{51}	C_{52}	$\hat{f}_2 C_{52} = C_{53}$	...	...	...	...	$S_5 = C_{52} \prod_{h=2}^{6} \hat{f}_h$
Exercice 6	C_{61}	$\hat{f}_1 C_{61} = C_{62}$	$\hat{f}_2 C_{62} = C_{63}$	...	...	...	...	$S_6 = C_{61} \prod_{h=1}^{6} \hat{f}_h$

$$\hat{f}_1 = \frac{C_{01} + C_{11} + \cdots + C_{51}}{C_{02} + C_{12} + \cdots + C_{52}}$$

Figure 20-Détails d'un triangle de liquidation et des formules de la méthode Chain Ladder.

- Les provisions pour chaque exercice de souscription sont déduites de ces charges ultimes S_i, par :

$$R_i = S_i - C_{i,n-i} \text{ pour } i = 1, \dots, n.$$

Nous en déduisons évidemment la provision globale comme étant :

$$R = \sum_{i=1}^{n} R_i.$$

Les étapes décrites ci-dessus prouvent la simplicité de la méthode, et sa facilité d'application. Cependant, cette méthode n'est pas toujours applicable, et ses limites vont être décrites dans la partie suivante.

2.2.2 Limites de la méthode

Des ratios de développement individuels $f_{i,j} = \frac{C_{i,j+1}}{C_{i,j}}$ jugés aberrants peuvent être exclus du triangle, sous réserve de justification, afin d'approcher au mieux cette notion d'uniformité dans l'évolution de la sinistralité par exercice de souscription. Cependant, si le portefeuille est soumis à des changements annuels récurrents, comme l'augmentation ou la diminution de la part de l'assureur sur certaines affaires, l'augmentation générale de la couverture sur une activité, ou encore des changements de méthodes de gestion des sinistres, alors l'ajustement des ratios peut ne pas être suffisant pour valider les hypothèses de base de la méthode. De plus, pour les années très récentes, l'incertitude est bien plus forte, particulièrement sur les risques longs, comme le portefeuille aviation. Dans cette méthode, le montant de la charge sinistre finale ne repose finalement que sur la durée de vie du sinistre, mais il est probable que d'autres facteurs importants entrent en compte, et ne soient actuellement pas utilisés.

Quelques variantes de la méthode du *Chain Ladder* ont été créées, afin de palier à quelques-unes de ces limites, par exemple avec la méthode du London *Chain Ladder*, ou tout simplement en ajoutant des pondérations au niveau des ratios de développement individuels, permettant par exemple de donner plus de poids aux exercices récents.

Nous allons maintenant exposer les résultats des hypothèses ainsi que les résultats de l'exercice de souscription 2020 sur chacun des triangles de liquidation.

2.3 APPLICATIONS NUMERIQUES

Les hypothèses énoncées dans la partie précédente vont maintenant être testées sur les deux jeux de données étudiés, afin de vérifier si l'utilisation de la méthode est appropriée ou non, puis la méthode d'application et les résultats seront exposés. Pour plus de clarté, la méthodologie sera expliquée entièrement sur les données maritimes, puis un tableau récapitulatif reprendra les résultats sur les données aviation, les tests aviation se trouvant en annexe A.1.2.

2.3.1 Vérification des hypothèses

Vérification de $\boldsymbol{H_0}$:

Pour l'hypothèse H_0, qui impose que les ratios $f_{i,j} = \frac{C_{i,j+1}}{C_{i,j}}$ soient indépendants de l'année d'origine i, la méthode se décline en plusieurs étapes. Cette hypothèse est vérifiée après retraitement des ratios de développement.

Comme détaillé par PHAN NGOC, H. (2015), une fois les calculs des ratios de développement individuels effectués (Figure 21) :

	1	2	3	4	5	6	7	8	9	10	11
2010	1,9022	1,0460	1,0052	0,8293	1,0002	0,9981	0,9963	0,9918	0,9991	0,9992	1,0010
2011	1,8223	0,9940	0,9479	0,9968	0,9823	0,9921	0,9863	1,0000	0,9933	0,9992	
2012	1,6693	0,9435	1,0003	0,9972	1,0114	0,9971	0,9989	0,9701	0,9895		
2013	1,3598	1,0286	1,0133	0,9921	0,9880	0,9893	0,9970	1,0024			
2014	1,7108	0,9772	0,9720	0,9913	1,0015	0,9803	0,9976				
2015	1,7383	0,9832	0,9881	0,9751	0,9913	0,9973					
2016	1,8347	1,0307	0,9778	1,0116	0,9927						
2017	1,8724	1,0366	0,9816	1,0035							
2018	1,8721	1,0313	0,9900								
2019	1,9491	1,0243									
2020	2										

Figure 21-Ratios de développement d'un triangle de liquidation.

La médiane de chaque colonne est calculée, et chaque ratio de cette colonne est comparé à la médiane puis séparé en deux catégories. L pour les ratios inférieurs à la médiane, ou S pour les ratios supérieurs à cette même médiane. Si le ratio observé est égal à la médiane, alors il n'appartient à aucune des deux catégories, et prend la valeur 0.

Nous obtenons alors la triangulation ci-dessous (Figure 22) :

	1	2	3	4	5	6	7	8	9	10	11	12
1	**S**	S	S	L	S	S	L	L	S	L	0	0
2	0	**L**	L	S	**L**	**L**	**L**	**S**	**0**	**S**	**L**	
3	L	L	S	S	S	S	S	L	L	L		
4	L	S	S	L	L	L	0	S	L			
5	L	L	**L**	L	S	L	**S**	L				
6	L	**L**	**0**	**L**	**L**	**S**	**L**					
7	**S**	**S**	**L**	**S**	**0**	**L**						
8	**S**	**S**	**L**	**S**	**L**							
9	**S**	**S**	**S**	**L**								
10	**S**	L	**L**									
11	L	L										

Figure 22-Séparation des ratios en 2 catégories par rapport à la médiane

Si nous écrivons $D_j = \{C_{j,1}, C_{j-1,2}, \dots, C_{2,j-1}, C_{1,j}\}$ avec $1 \leq j \leq n$ les éléments de la diagonale j, pour chacune de ces diagonales D_j, avec $j = 2, \dots, n-1$ la somme S_j des S présents sur la diagonale est calculée ainsi que la somme L_j des L sur cette même diagonale. Si les années sont similaires, alors S_j et L_j doivent être proches. Dans ce cas, si nous considérons $Z_j = \min(S_j ; L_j)$ alors Z_j doit être proche de $\frac{S_j+L_j}{2}$.

Si nous considérons la variable globale $Z = Z_2 + \cdots + Z_{n-1}$ et supposons que les Z_j ne sont pas corrélés, alors nous pouvons calculer :

$$E(Z) = \sum E(Z_j),$$

$$\text{et } V(Z) = \sum V(Z_j).$$

S'il n'y a pas d'effet calendaire, alors les coefficients ont autant de chance d'être S que L. S_j et L_j suivent donc des lois Binomiales de paramètres $n_j = L_j + S_j$ et $p = \frac{1}{2}$. Par conséquent, il est possible de calculer l'espérance et la variance pour chaque j par les formules suivantes :

$$E(Z_j) = \frac{t}{2} - C_{t-1}^{m} \frac{t}{2^t},$$

$$\text{et } V(Z_j) = \frac{t(t-1)}{2^t} - C_{t-1}^{m} \frac{t(t-1)}{2^t} + E(Z_j) - \left(E(Z_j)\right)^2,$$

$$\text{en posant : } m = \text{ent}(\tfrac{t-1}{2}) \text{ et } t = S_j + L_j.$$

Avec l'exemple précédent, nous obtenons le tableau suivant (tableau 9) :

	j	L	S	Z	t	m	t-1	Cm t-1	E	V
	2	0	1	0	1	0	0	1	0,00	0,00
	3	2	1	1	3	1	2	2	0,75	0,19
	4	4	0	0	4	1	3	3	1,25	0,44
	5	1	4	1	5	2	4	6	1,56	0,37
	6	3	3	3	6	2	5	10	2,06	0,62
	7	5	2	2	7	3	6	20	2,41	0,55
	8	4	3	3	7	3	6	20	2,41	0,55
	9	3	6	3	9	4	8	70	3,27	0,74
	10	5	3	3	8	3	7	35	2,91	0,80
	11	3	6	3	9	4	8	70	3,27	0,74
	12	10	0	0	10	4	9	126	3,77	0,99
Total			**Z**	**19**					**23,65**	**5,99**

Tableau 9-Résultats de l'espérance et de la variance pour chaque diagonale

Dans ce tableau, chaque variable est calculée et les espérances et variances sont sommées afin de calculer un intervalle de confiance pour Z.

Nous calculons l'intervalle de confiance à 95% suivant, en estimant que Z suit approximativement une loi Normale :

$$\left[E(Z) - 2\sqrt{V(Z)}; E(Z) - 2\sqrt{V(Z)}\right].$$

Si Z appartient à cet intervalle, alors nous considérons que nous pouvons rejeter l'hypothèse d'un effet calendaire dans le triangle.

Dans l'exemple, $Z = 19$ et l'intervalle de confiance est de $[18.76\,;28.55]$, Z appartient à l'intervalle de confiance, par conséquent, nous rejetons l'existence d'un effet calendaire sur ce triangle de liquidation.

La vérification de H_0 a été faite sur tous les triangles, aviation et maritime, nous pouvons passer à la deuxième hypothèse dans la partie suivante.

Vérification de H_1

Pour notre deuxième hypothèse, qui estime qu'il existe un paramètre $f_{i,j}$ tel que $C_{i,j+1} = f_{i,j}C_{i,j}$

pour $i = 0, \dots, n-1-j$ il faut vérifier si les $(n-j)$ couples $(C_{i,j}, C_{i,j+1})$ sont alignés par rapport à la droite des coefficients de développements, passant par l'origine.

Par exemple sur le triangle « Pêche et autres corps » (PAC), sur les six premières années de développement (Figure 23) :

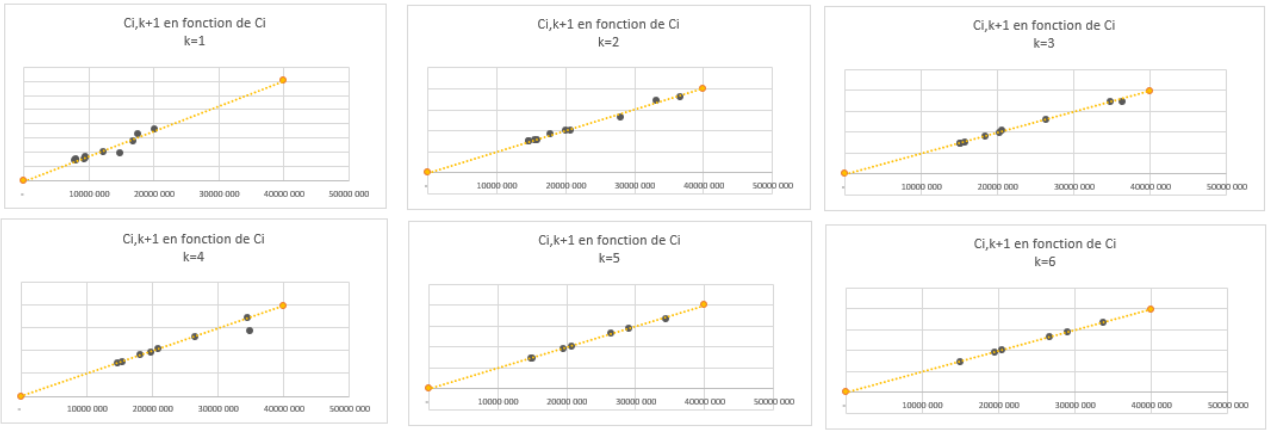


Figure 23-Vérification de l'alignement des couples $(C_{i,j}, C_{i,j+1})$

Nous pouvons remarquer, par les graphiques ci-dessus et ceux disponibles dans les annexes A.1.1, qu'un alignement est bien visible pour les couples $(C_{i,j}\,;C_{i,j+1})$. L'hypothèse est globalement validée sur les triangles PAC et IAR, le triangle CMM (disponible en annexe avec les résultats IAR), ne valide pas cette hypothèse lors de l'application du *Chain Ladder*. Cependant, la triangulation ayant le plus de poids dans les résultats maritime, est la triangulation PAC, qui elle valide les hypothèses.

Résultats

Si nous récapitulons les résultats des deux bases de données dans un même tableau (Tableau 10), nous obtenons :

Base	Triangles de liquidation	H_0	H_1
Aviation	RC	Indépendance	Non linéarité
	RC Produit	Indépendance	Non linéarité
	Corps	Indépendance	Non linéarité
Maritime	CMM	Indépendance	Non linéarité
	IAR	Indépendance	Linéarité
	PAC	Indépendance	Linéaire

Tableau 10-Résultats des hypothèses pour chaque triangle de liquidation utilisé lors du Chain Ladder

D'après les résultats observés, la méthode *Chain Ladder* n'est donc pas à appliquer. Les résultats des triangulations maritime RC et FAC ne sont pas étudiés ici, l'exercice 2020 ne possèdant aucun sinistre sur ces triangles.

Après retraitement des données sur chaque triangle de liquidation, certaines hypothèses sont donc supposées validées lors de l'estimation de la charge ultime, alors que celles-ci ne le sont pas, au moins sur une partie des triangles étudiés. L'application du *Chain Ladder* n'est donc pas correcte. Il est de ce fait judicieux de *challenger* les résultats par de nouveaux modèles, reposant sur moins d'hypothèses lors de leur application.

Dans le chapitre suivant, les méthodes utilisées afin d'obtenir la charge ultime pour l'aviation et le maritime seront décrites, puis les résultats à fin 2021 seront présentés, afin de les comparer par la suite avec les résultats obtenus par nos méthodes de provisionnement ligne à ligne.

2.3.2 Résultats du provisionnement fin 2021 : Exercice de souscription 2020

Aviation

Les données aviation sont séparées par risques (RC / RC Produit / Corps), afin de travailler sur des triangles homogènes. En effet, les sinistres à durée de développement trop différent doivent être évaluées séparément, au risque d'obtenir une estimation globale erronée. Concernant les sinistres dits exceptionnels, touchant les XS, ceux-ci sont retirés de la triangulation. Etant particulièrement suivis, ils sont estimés à dire d'expert, c'est-à-dire que leur charge actuelle, est supposée être la meilleure estimation possible.

Concernant les autres sinistres, dits attritionnels, ceux-ci sont séparés en trois triangulations RC, RC Produit, et Corps. Les triangles RC et Corps sont projetés avec un historique de 5 ans, c'est à dire que seuls les ratios des 5 dernières diagonales des triangles sont utilisées afin de calculer les facteurs de développement. Ces derniers sont aussi pondérés, afin de donner plus de poids aux années récentes.

Sur la garantie RC Produit, la triangulation est très peu homogène, la souscription des *Manufacturers* ayant été ralentie entre les exercices 2009 et 2014, peu de sinistres ont été déclarés, et ces exercices ne sont donc pas représentatifs des années actuelles. Par conséquent, les ratios de ces exercices de souscription sont exclus, afin d'augmenter la stabilité du triangle, et les facteurs de développement sont calculés sur l'historique complet de 15 ans, en utilisant une moyenne simple.

Les années de souscription 2019 et 2020, les plus récentes et les plus touchées par le covid ont été exceptionnellement estimées également par la méthode du S/P cible, la sinistralité ultime est calculée comme une moyenne entre le S/P cible du *Business Plan* et les résultats du *Chain Ladder*.

Pour terminer, les résultats des trois triangles sont sommés, afin d'obtenir un résultat final sur le triangle aviation global.

Comme nous pouvons le voir sur le graphique suivant (Figure 24) :

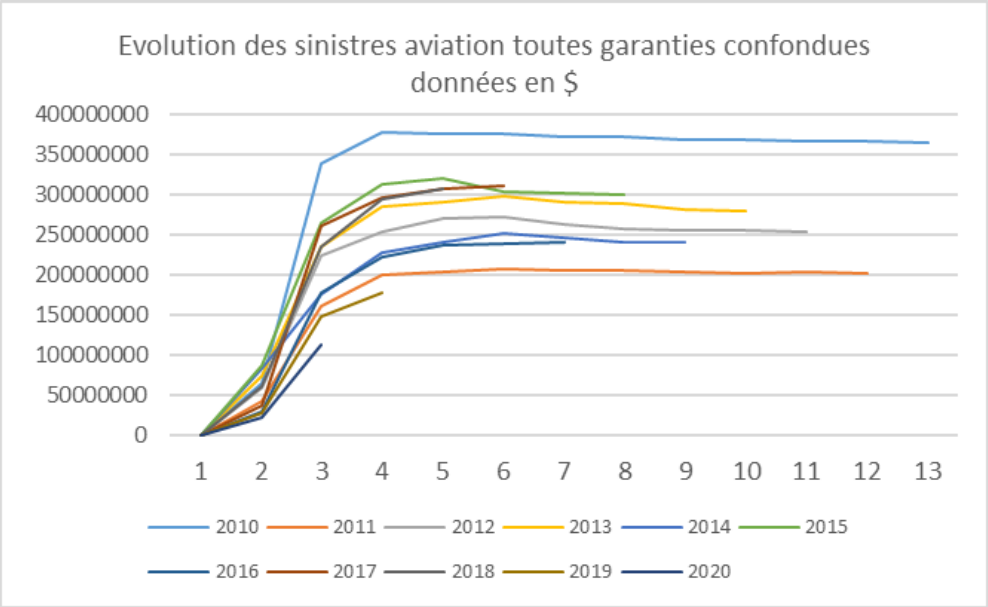


Figure 24-Evolution de la charge sinistres par année de souscription

La sinistralité évolue pendant 8 ans en moyenne, pour finir par se stabiliser à un montant bien supérieur à ce qui était enregistré au terme de deux années de développement.

Finalement, nous obtenons les résultats suivants sur l'exercice de souscription 2020 (Tableau 11) :

Données en $	Charge	Charge ultime	IBNR	Evolution de la charge
Corps	77 676 270	81 043 720	3 367 450	4%
RC	22 822 382	70 799 487	47 977 105	210%
RC Produit	12 001 288	56 258 937	44 257 649	369%
Total	112 499 940	208 102 144	95 602 204	85%

Tableau 11-Résultats de la méthode Chain Ladder en aviation

Deux évolutions distinctes sont visibles sur les résultats de l'exercice 2020. D'un part, la charge ultime Corps évolue de 4% entre la deuxième et la dernière année, d'autre part, les deux RC augmentent de plus de 200%, avec respective 210% pour la RC et 369% pour la RC produit. En effet ces risques ont un développement bien plus long, et une charge ultime bien supérieure au Corps.

Au total, la charge 2020, d'après les méthodes appliquées, va augmenter de 85% afin d'atteindre un montant ultime de 208M$. Aucun sinistre exceptionnel, n'a été reporté au 31/12/2021, les résultats ci-dessus seront donc les résultats définitifs pour la base aviation, et ceux-ci seront notre base de comparaison pour la base aviation lors des tests des modèles lignes à lignes.

Les résultats aviation 2020 ayant été exposés, nous allons maintenant étudier les résultats sur le maritime.

Maritime

Les données maritimes fournies, contiennent plusieurs activités, et s'étalent sur un périmètre allant de l'exercice de souscription 2010 à 2020. Les sinistres étudiés dans cette base sont moins volatils, et ont une durée de vie en moyenne plus courte que l'aviation. La méthode sur le maritime nécessite de séparer les

données selon les activités, chaque activité ayant un seuil spécifique séparant les sinistres attritionnels des sinistres larges, ces seuils, sont présentés dans le tableau suivant (Tableau 12) :

Activité	seuils
CMM	1 000 000
IAR	1 000 000
PAC	800 000

Tableau 12-Seuils entre attritionnels et larges sur chaque activité

Les sinistres attritionnels, donc strictement inférieurs aux seuils définis dans le tableau précédent, seront projetés selon la méthode du *Chain Ladder*. Les sinistres larges quant à eux sont évalués à part, avec une méthode de type coût fréquence.

La méthode du *Chain Ladder* montre une tendance à la baisse ou une certaine stabilité selon les années de souscription, au bout de quelques années sur les trois triangles, comme le montre la figure 25 ci-dessous :

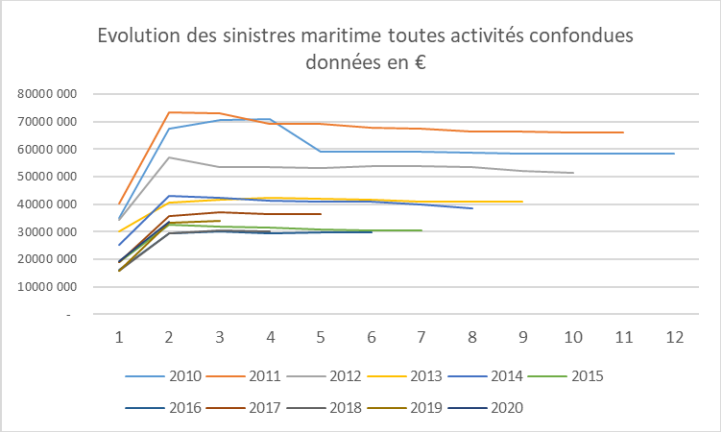


Figure 25-Evolution de la charge sinistres par année de souscription

Nous pouvons aisément remarquer que les sinistres se stabilisent rapidement, soit au bout de 4 ou 5 ans. A la deuxième année de développement, leur charge est déjà au niveau, voire supérieure à la charge ultime.

Finalement, les résultats de chaque triangle donnent sur l'exercice 2020 (Tableau 13) :

Données en €	Charge	Charge ultime	IBNR	Evolution de la charge
CMM	841 926	861 752	19 826	2,30%
IAR	34 007	31 328	- 2 679	-8,55%
PAC	32 622 551	31 929 689	- 692 862	-2,17%
Total	33 498 484	32 822 769	- 675 714	-2,06%

Tableau 13-Résultats de la méthode Chain Ladder en maritime

Nous pouvons remarquer que contrairement à l'aviation, la sinistralité atteint rapidement son montant ultime, et les variations suivantes sont négatives. La triangulation CMM augmente de 2.3% par rapport à la charge actuelle, les triangles IAR et PAC eux diminuent de 8.55% et 2.17%. Au total, la charge maritime diminue de -2.06% afin d'atteindre son montant ultime.

Concernant les sinistres larges, ceux-ci sont estimés par une méthode coût fréquence qui se décompose en deux parties, l'estimation de la fréquence projetée, et l'estimation de la sévérité.

En termes de fréquence, une triangulation est effectuée de 2010 à 2020, afin de projeter le nombre de sinistres larges, par la méthode du *Chain Ladder*.

La triangulation est la suivante (Figure 26) :

	12	24	36	48	60	72	84	96	108	120	132	144
2010	2	11	12	12	8	8	8	8	8	8	8	8
2011	6	12	12	11	11	10	10	10	10	9	9	
2012	10	15	14	14	14	14	14	14	13	13		
2013	8	7	7	6	6	6	6	6	6			
2014	6	7	5	5	5	6	6	6				
2015	3	7	8	9	8	8	8					
2016	2	3	4	5	5	6						
2017	3	3	3	3	3							
2018	2	2	3	3								
2019	4	5	6									
2020	5	9										

Figure 26-Triangulation des nombre de sinistres larges par exercices de souscription

Les exercices 2011 et 2012, particulièrement sinistrés, représentent peu l'évolution moyenne du nombre de sinistres du portefeuille maritime. Si nous ne tenons pas compte de leur évolution, la sinistralité large évolue peu en termes de nombre, à partir de la deuxième année de développement.

La méthode du *Chain Ladder* donne un nombre de sinistre large ultime sur l'exercice de souscription 2020 de 9.78 sinistres, cependant ce nombre peut être révisé afin de conserver 9 sinistres larges en 2020, la fréquence des sinistre évoluant peu après la deuxième année de développement

En termes de sévérité, une sélection des sinistres larges et clos à fin 2021 est effectuée, ces derniers doivent être représentatifs des sinistres de l'exercice estimé.

D'après le tableau ci-dessous, reprenant les montant de charge cumulée et de provisions à fin 2021 (Tableau 14) :

Données en €

Charge cumulée nette de recours 31/12/2021	Provision nette de recours 31/12/2021
1 231 693	1 176 286
3 327 199	3 271 830
1 580 000	1 562 076
1 602 328	- 1 125 454
878 828	-
1 723 069	-
1 294 000	24 000
899 998	71 998
1 569 679	1 539 954

Tableau 14-Charges ultimes et provisions nettes de recours des sinistres larges de l'exercice de souscription 2020

Nous pouvons remarquer que parmi les 9 sinistres larges observés ci-dessus, deux sont déjà clos, avec des provisions nulles. Dans ce cas, leur montant actuel peut être utilisé comme montant ultime. Les sept autres sinistres doivent être estimés par rapport aux sinistres antérieurs clos au 31/12/2021.

Pour ces sept autres sinistres, leur montant au 31/12/2021, soit à la fin de la deuxième année de développement, sont comparés aux sinistres larges passés et clos d'un montant similaire en deuxième année de développement.

Des sinistres à charge totale en deuxième année très élevés ont donc été écartés du calcul du coût moyen, afin de ne pas alourdir la charge sinistre, avec des sinistres larges trop élevés, non représentatifs de ceux de l'exercice 2020.

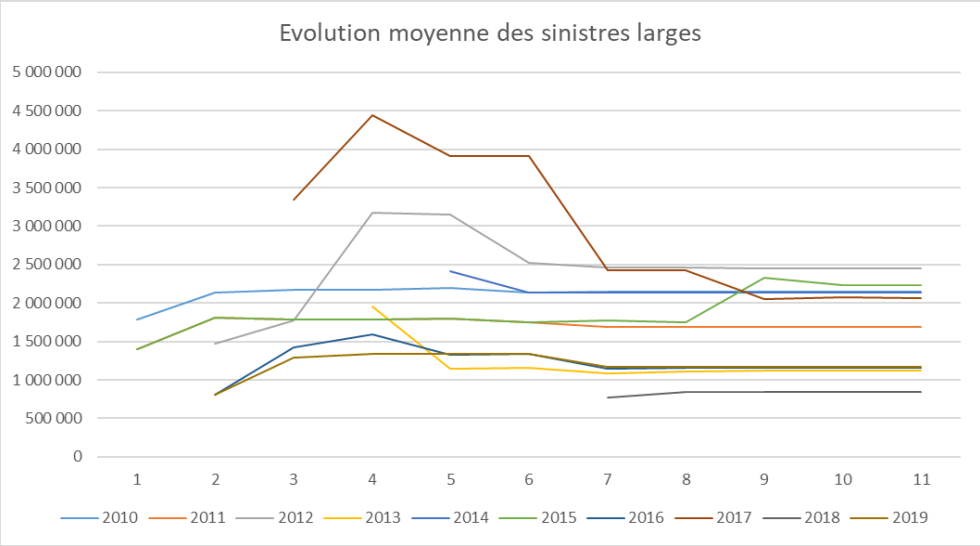


Figure 27-Evolution des sinistres larges par exercice de souscription

Comme le montre la figure ci-dessus (Figure 27), les sinistres larges ont tendance à rester stable ou diminuer, entre leur montant à l'ouverture, et leur montant final.

Finalement, le cout moyen retenu, calculé comme étant la moyenne des sinistres larges et clos des exercices antérieurs, est d'1.6M€, et la fréquence appliquée reste de 7 sinistres pour 2020, soit les neufs sinistres projetés dont sont retirés les deux sinistres clos. La charge des sinistres large est donc de 11.3M€.

Par conséquent, sur l'exercice 2020, la charge sinistre ultime, représentée par les sinistres attritionnels et larges, est la suivante (Tableau 15) :

Données en €	Charge	Charge ultime	IBNR	Evolution de la charge
Total attritionnels	33 498 484	32 822 769	- 675 714	-2,06%
Larges projetés	11 504 897	11 304 555	- 200 343	-1,77%
Larges clos	2 601 897	2 601 897	-	0,00%
Total	47 605 278	46 729 221	- 876 057	-1,87%

Tableau 15-Résultats de la méthode du Chain Ladder sur sinistres attritionnels et coût fréquence sur les sinistres larges

La charge ultime des sinistres projetés est inférieure à la charge actuelle de 1.77%, les sinistres larges et clos n'évoluent plus et restent à un montant total de 2.6M€. Par conséquent, la charge ultime finale, sinistres larges et attritionnels confondus, est inférieure à la charge ultime actuelle, de -1.87%, et représente 46.7M€. Une nouvelle fois, ces résultats seront la base de comparaison avec les résultats des méthodes ligne à ligne.

En conclusion de cette deuxième partie, qu'il s'agisse de l'aviation ou du maritime, les hypothèses du Chain Ladder ne sont pas toutes respectées avant l'application de la méthode, malgré un retraitement des ratios de développement de chaque triangle.

Concernant les résultats de la charge ultime, sur l'exercice de souscription 2020, l'évolution attendue en aviation afin d'atteindre cette charge est bien supérieure à celle attendue en maritime, avec respectivement 85% d'évolution et -1.87%. L'évolution entre les deuxièmes et troisièmes années de développement, étant la principale raison de cette forte variation en aviation, certains exercices voyant leur charge totale doubler voire tripler au cours de l'année.

Finalement, les charges ultimes qui seront utilisées comme point de comparaison avec les méthodes qui seront traités par la suite sont de 208M$ en aviation et 46.7M€ en maritime.

3 Provisionnement ligne à ligne

Un problème d'apprentissage nécessite avant tout d'être compris, et correctement abordé. Il faut donc connaitre les variables disponibles et leur impact sur la valeur finale à prédire. Cela passe par des discussions avec les experts du domaine, et une bonne connaissance du portefeuille de l'entreprise. Ces informations doivent être mises en forme et retraitées intelligemment, lors de la préparation des données. Toutes ces méthodes d'enrichissement sont indispensables pour obtenir finalement des données prédites de qualité.

Cette partie est par conséquent dédiée à l'étude de la création des bases de données et l'analyse des données aviation et maritime, avant d'entrer dans le provisionnement ligne à ligne. Nous allons donc commencer par décrire les bases disponibles, leur création, ainsi que les divers retraitements appliqués. Par la suite nous prendrons le temps d'étudier plus en détail le niveau de censure, et les interactions entre les diverses variables et la charge ultime.

3.1 PERIMETRE ET DONNEES

D'après DUBOIS, D. et al. (2021), les assureurs possèdent, par la nature de leur profession, d'une multitude de données disponibles dans leur entreprise. Chacune de ces bases de données est bâtie dans le but de faciliter les travaux, dans le cadre d'un mode de fonctionnement « traditionnel » de l'actuariat. Il faut donc ajouter à ces informations structurées, des informations internes et externes plus ou moins structurées, pour obtenir un historique avec des caractéristiques détaillées balayant différents aspects du sinistre, et propices à l'élaboration de diverses méthodes de *Machine Learning*.

Dans ce chapitre nous exposerons les différentes données disponibles utilisées pour l'élaboration des deux jeux de données, ainsi que les méthodes utilisées afin de résoudre le problème de données manquantes. Nous commencerons par exposer les données aviation, dans les bases structurées existantes, puis les données internes non structurées. Par la suite les données maritimes seront détaillées, et nous terminerons par les types de données manquantes, et les méthodes utilisées afin d'y remédier.

3.1.1 Bases structurées internes et informations non structurées

Base primes aviation

Deux grands types de bases sont extraites plusieurs fois par an. La « base primes » et la « base sinistres » celle-ci existant sous différentes formes, selon si cette dernière est triangulée ou non, et en norme de conversion économique ou As-if. Les données pertinentes disponibles et méthodes de conversion sont détaillées ci-dessous, la forme globale de la base est disponible dans l'annexe A.2.2.

La base prime reprend toutes les informations disponibles des exercices de souscription 2012 à 2022, séparant les données par numéro de police. L'activité aviation étant internationale, les primes sont perçues sous différentes devises. La base contient les primes en USD converti au taux de l'année en cours. Les primes perçues l'année N, sont converties au taux du 31/12/N-1, les primes perçues les années précédentes, sont converties au taux du 31/12 de l'exercice de souscription.

Plusieurs données peuvent être pertinentes pour la suite de l'étude :

- **Vuau** : Date à laquelle le sinistre est observé
- **Groupement de gestion** : Le groupe auquel l'assuré peut appartenir, par exemple Air France appartient à Air France KLM.
- **Assuré** : Le nom de l'assuré pour lequel la police a été souscrite, dans l'exemple précédent, Air France est l'assuré.
- **Pays de l'assuré** : Code ISO à 2 lettres du pays dans lequel est domicilié l'assuré. Par exemple FR pour Air France
- **Pays d'opération de l'assuré** : Le pays principal où opère l'assuré. Si Air France vole principalement en France, ce sera FR.
- **Regroupement d'activité de l'assuré** : il s'agit des grandes activités du type *Airlines*/Aviation Générale et *Manufacturers*.
- **Activité principale de l'assuré** : Il s'agit d'une sous-catégorie du regroupement d'activité. Par exemple les *Airlines* peuvent être des vols internationaux comme Air France, ou des Cargo comme Fedex.
- **Prime brute à 100% en USD** : il s'agit de la prime totale payée par l'assuré.
- **Prime brute à la part de l'assureur en USD** : Il s'agit d'une portion de prime à 100%, qui reviendra effectivement à l'assureur.
- **Engagement maximum à 100%** de la garantie Corps, et de la garantie RC.

Base sinistres aviation

La base sinistre reprend toutes les informations disponibles sur un historique important. Chaque évènement est séparé en sinistre, et chaque sinistre est séparé en dossier sinistre (Figure 28) :

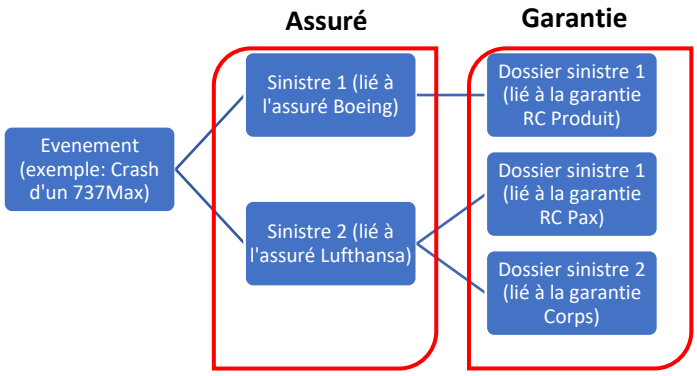


Figure 28-Séparation d'un évènement en sinistre et dossier sinistre

Lors d'un seul évènement, plusieurs assurés peuvent être touchés. Dans ce cas, le numéro de sinistre est appliqué au niveau de l'assuré, et les dossiers sinistres permettent de séparer les différentes garanties touchées, sur un même assuré. L'évènement ci-dessus aura un numéro d'évènement, pour deux numéros de sinistres et trois numéros de dossier sinistre.

Plusieurs bases sinistres existent, selon le taux de change appliqué, et si une triangulation est nécessaire ou non. Concernant le taux de change, tout comme les primes, les sinistres sont eux aussi réglés et provisionnés dans plusieurs devises. Un même dossier sinistre peut contenir deux ou trois devises, surtout lorsque des frais d'avocats entrent en jeu.

La norme de conversion *Asif*, qui est utilisée lors de l'élaboration des différents triangles de liquidation, sera également utilisée dans toute la suite de ce mémoire, et permet de garder un taux de change fixe, afin de ne pas projeter la variation du taux de change dans notre estimation de la charge ultime.

Le taux *AsIf* s'applique selon la méthode suivante :

Pour les **provisions**, celles-ci ne sont pas représentées par des flux, mais à un instant *t*. Le taux appliqué est le dernier taux comptable connu, soit dans notre cas le 31/12/21.

Pour les **paiements** sur un exercice de souscription passé (Figure 29) :

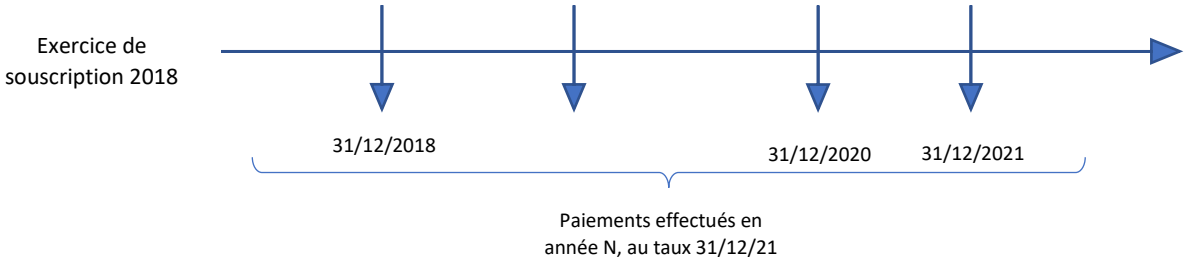


Figure 29-Fonctionnement des paiements en méthode de conversion AsIf

Tous les différents flux de paiements qui ont pu avoir lieu lors de la vie du sinistre, sont convertis un à un à un unique taux, puis cumulés afin d'obtenir la somme des paiements *Asif*.

Les données pertinentes pour les sinistres sont les suivantes :

La **« vu au »** soit la date d'extraction des données, la date de l'évènement ou **« DOL »** pour *date of loss*, la **« date of loss »**, représentant la date d'entrée du sinistre dans le système informatique, le **lieu de l'évènement** et son **pays**. Y sont également disponibles, le **nom de l'assuré** concerné, le **numéro du dossier sinistre**, et la **garantie touchée**. Le montant affecté au dossier sinistre est ensuite exprimé à 100% (Paiements / Frais payés/Provisions dossier à dossier/ Frais provisionnés/Total), puis à la part.

En plus des données usuelles de la base sinistres, de nouvelles données ont pu être ajoutées, concernant le **nombre de blessés de l'équipage, les blessés des personnes transportées, ou au sol**, ainsi que l'existence d'un **recours**, d'une **procédure judiciaire**, et d'une **perte totale**. Ces trois derniers étant des booléens qui renvoient 0 ou 1 pour chaque dossier sinistre.

A ces informations s'ajoutent des informations disponibles dans divers fichiers exploités par la souscription, qui elles ne sont ni structurées ni disponibles mensuellement, il s'agit de données non structurées, celles-ci sont décrites ci-dessous.

Données non structurées en aviation

Les données non structurées sont très importantes, il est indispensable de les obtenir et de simplifier au mieux leur accessibilité. Elles apportent ici des informations nouvelles sur les assurés comme des informations techniques, contractuelles, ou encore commerciales.

Ces données non structurées spécifiques à l'activité *Airlines* sont disponibles dans un document établi par les souscripteurs. Il est complété manuellement par les équipes de souscription tout au long de l'année, et ce depuis 2013, avec les informations disponibles sur les contrats. Ce document contient entre autres pour chaque assuré :

- L'*« Average Fleet Value »* ou **AFV,** la valeur moyenne de la flotte
- Le nombre de passagers ou **PAX**
- Le **nombre de départs (« dep »)**
- Le **nombre de sièges (« seats »)**

Concernant les *Manufacturers*, l'information du chiffre d'affaires de l'assuré a été ajoutée, celle-ci étant utilisée lors de la tarification.

Enfin, de nombreuses données telles que les primes, le chiffre d'affaires, le PAX, les limites, ont été segmentées, afin de permettre aux modèles de mieux appréhender les effets de certaines variables sur la charge ultime.

Le deuxième jeu de données, en maritime, comprend certaines informations similaires à l'aviation, principalement sur les données quantitatives, d'autres données qualitatives propres aux navires et à la souscription maritime sont utilisées afin de décrire au mieux l'assuré sinistré. Ces informations sont détaillées ici.

Données disponibles en maritime

De la même manière, les données de la base maritime nous ont été fournies en regroupant les informations de différentes bases. La base de données finale contient les informations trimestriellement, et débute à l'exercice de souscription 2010, pour se terminer en 2020.

Les données quantitatives concernent principalement le montant des sinistres ainsi que la valeur de la flotte ou de la prime. D'autres informations concernent des informations sur le sinistre :

- Longueur du navire
- Nombre de moteurs
- Type de navire
- Poids du navire
- Type de moteurs
- Présence d'un contentieux ou non
- Année du sinistre…

Une nouvelle fois, d'autres données ont été ajoutées, comme le type de sinistralité (Attritionnelle ou Large selon les seuils donnés en 2.3.2), de la même manière que pour l'aviation, de nombreuses variables ont été segmentées, afin de pouvoir regrouper certaines données similaires, et permettre aux algorithmes d'en déduire plus facilement des effets de groupes.

Contrairement à l'aviation, ces données sont en Euro converties en norme économique, par conséquent, le taux de change n'est pas stable à travers les vues, mais celui-ci, d'après les experts, n'a pas d'impact significatif sur les résultats.

Les deux bases ayant été construites à partir de données existantes, peuvent maintenant être complétées par certaines informations calculées à partir des informations disponibles, il s'agit ici d'enrichissement des données.

Enrichissement des données par de nouvelles variables

L'enrichissement des données se fait par la création de nouvelles informations à partir des données actuelles. Nous ajoutons donc en aviation les ratios sinistres à primes pour chaque assuré, observés sur 5 ans. L'aviation ayant une sinistralité à déroulement long, ce ratio pourrait permettre au modèle d'avoir une information de la sinistralité moyenne passée, liée à l'assuré du dossier sinistre.

La région du sinistre, en aviation ainsi qu'en maritime, est également ajoutée selon le pays de l'évènement. Ceci permet des regroupements par zones, et réduit considérablement le nombre de variables disponibles.

Une fois toutes ces informations regroupées dans une même base de données, reprenant ligne à ligne les dossiers sinistres concernés, de nombreuses données sont « manquantes » ou « non renseignées ». Dans ce cas, divers retraitements sont effectués afin de réduire au maximum leur nombre.

3.1.2 Retraitement des données

Lors de la création d'une base de données, de nombreuses informations peuvent être manquantes ou absentes pour différentes raisons, selon leur nature, leur source ou encore leur forme. Il est possible de distinguer trois grands groupes, détaillés dans l'UE RCP208 du CNAM comme :

a) ***Missing Completely At Random* ou « MCAR »**, lorsque la probabilité que l'information soit absente, est indépendante des autres variables.
b) ***Missing At Random* ou « MAR »**, lorsque la probabilité que la donnée soit manquante, dépend des autres variables observées, mais pas de la variable manquante. Par exemple en aviation, sur l'activité *Airlines*, le montant du chiffre d'affaires de la société ne sera pas précisé, puisque la donnée n'est disponible que s'il s'agit d'un assuré qui a comme activité *Manufacturers*.
c) ***Missing Not At Random* ou « MNAR »**, lorsque la probabilité que la donnée soit manquante, dépend de la valeur non observée. Par exemple dans certains cas, la notion de revenu est plus souvent absente, lorsque la question est posée à une personne à haut revenu. Il est souvent difficile de discerner les données MAR et MNAR dans une base de données.

Ces différentes données manquantes peuvent être gérées de différentes manières.

Il est souvent tentant d'ignorer les valeurs manquantes, et donc de supprimer les observations contenant une ou des données manquantes. Selon la raison de l'absence de valeur (MCAR /MNAR/MAR) cette suppression peut avoir un impact sur les résultats de l'étude, comme dans le cas MCAR, réduisant drastiquement les données disponibles. Dans le cas MAR, cela introduira un biais dans le modèle, puisque certaines observations, ne sont plus présentes dans le modèle. Dans ce cas, il faudra compléter ces données. Pour le cas MNAR, le modèle est lui aussi biaisé, mais il est bien plus difficile d'estimer les valeurs manquantes.

Toutes ces données manquantes peuvent être imputées selon plusieurs méthodes :

a) Dans de nombreux cas sur la base maritime, il y a eu recours à **l'imputation par une valeur unique** avec une médiane, moyenne, ou une valeur par défaut. Il faut toutefois faire attention à cette méthode, puisque la moyenne est fortement sensible aux points extrêmes, la médiane possède une bien meilleure robustesse. Il a également été possible de prendre la valeur avec la plus grande fréquence, lorsqu'il s'agissait de données qualitatives.

b) A partir des $\boldsymbol{k}$ **plus proches voisins,** un algorithme fondamental et simple à mettre en œuvre introduit par Fix E. & Hodges J. (1951), appartenant à la catégorie de la classification ou de la régression supervisée.

Dans notre cas, le but est de prédire le nombre y d'une nouvelle instance x, soit dans ce cas pour les variables suivantes (tableau 16) :

x	y
Primes	CA
AFV	PAX
AFV	Seats
PAX	DEP

Tableau 16- Variables avec données manquantes complétées par la méthode des k plus proches voisins

Les variables x ont été choisies comme étant celles ayant la plus forte corrélation avec la variable cible y. Puis la méthode des k plus proches voisins a été appliquée à chacun des couples.

L'algorithme se déroule en quelques étapes :

Il faut tout d'abord indiquer à l'algorithme un nombre k, généralement obtenu par la méthode de validation croisée détaillée dans le chapitre 4.2.6.

Puis l'algorithme calculera la distance entre le point x, et les autres points connus. Cette distance est la distance Euclidienne ou Manhattan, respectivement $\sum_{i=1}^{n}|x_i - y_i|$ et $\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$. Les k plus petites distances seront retenues.

Pour terminer, la moyenne des valeurs de y de ces points retenus sera calculée, et sera la valeur prédite y de ce nouveau point.

c) Imputation par **une moyenne partielle**, en d'autres termes, les observations complètes sont classifiées, et les observations avec données manquantes appartenant à la même classe sont complétées avec la moyenne des observations complètes de cette classe.

Une fois les bases créées, structurées, complétées, et exempt de données manquantes, il est maintenant possible d'analyser les informations disponibles, les interactions entre certaines variables et la valeur cible, ainsi que les similitudes et différences entre les deux jeux de données. Ici, la valeur cible est la charge vue au 31/12/2021, représentant la charge ultime du sinistre si celui-ci est clos.

3.2 ANALYSE DES DONNEES

Avant de développer l'analyse des données, les notions de censure troncature et durée d'un sinistre doivent être abordées, celles-ci ayant été utilisées dans la suite de l'analyse.

3.2.1 La censure et la troncature des données

Il est impératif de noter que les sinistres, en aviation tout comme en maritime, contiennent une information importante quant à l'état du dossier sinistre observé, leur « position ». Un dossier sinistre peut être ouvert, réouvert, clos ou clos sans suite sur l'intervalle de temps $[I_{début}\,;I_{fin}]$. I_{Fin} dépend simplement dans notre cas, de la date d'extraction des données, mais elle peut dépendre d'autres facteurs comme un arrêt de suivi d'une personne lors d'un suivi médical, ou une date de fin d'une étude statistique. Par conséquent, Comme expliqué par Lefaou Y. (2019) cet intervalle n'englobe pas la totalité de la durée des différents dossiers sinistres. Ces données sont soit incomplètes, et donc sujettes à de la censure, ou sujettes à un biais d'observation, et dans ce cas à de la troncature. En assurance, les données de durée sont tronquées à gauche et censurées à droite, ou seulement censurées à droite.

Si nous posons T la variable aléatoire réelle positive et continue de la durée, associée à chaque sinistre. Alors dans notre cas, T n'est pas tronquée à gauche, mais peut être censurée à droite, si la fin de l'observation a lieu avant la clôture du sinistre. C est la variable aléatoire réelle positive et continue représentant la censure, lorsque T n'est observable que jusqu'à la date D_{fin}. Par ces deux variables, nous pouvons maintenant écrire $Y = \min(T, C)$, et $\delta = \mathbb{1}_{T \leq C}$ représentant la survenance ou non de l'évènement (clôture du sinistre) .

De façon triviale, pour l'intervalle de temps observé $[I_{début}\,;I_{fin}]$ quatre cas sont possibles (Figure 30) :

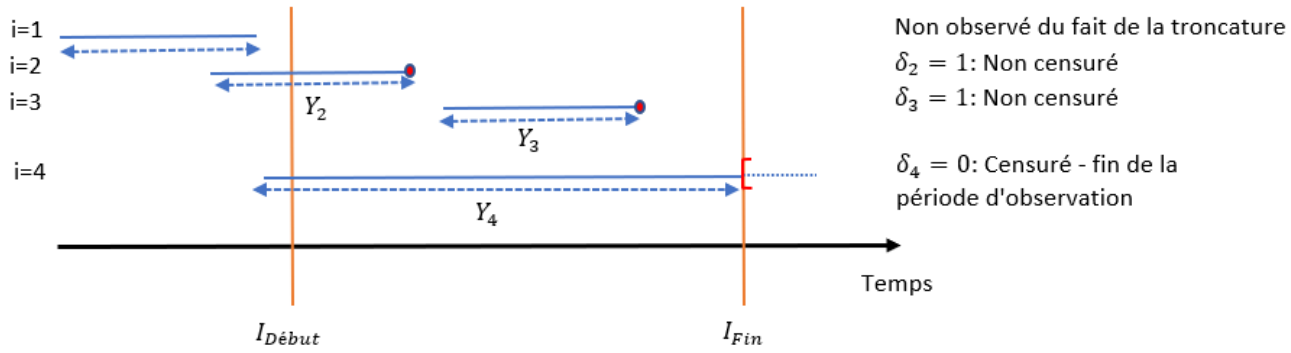


Figure 30-Les différents types de censures et troncatures(Source Lefaou Y.)

Par conséquent, si $\delta = 1$ alors il y a eu clôture du sinistre dans l'intervalle de temps observé, sinon $\delta = 0$. Afin d'étudier les variables de durée, quelques outils mathématiques sont à disposition, si U une durée, nous notons la fonction de répartition comme étant :

$$F_U(t) = P(U \leq t).$$

Dans ce cas, la fonction de survie de U s'écrit :

$$S_U(t) = 1 - F_U(t),$$

La probabilité que la clôture du sinistre ait lieu dans un petit intervalle de temps après t, conditionnellement au fait que U ne se produise pas avant t, est le risque instantané, qui s'écrit :

$$\lambda_U(t) = \lim_{h \to 0} \frac{P(t \leq U < t + h | U \geq t)}{h} = -\frac{\mathrm{S}'_U(\mathrm{t})}{\mathrm{S_U(t)}}.$$

Dans la base aviation ainsi que dans la base maritime, une certaine quantité de dossiers sinistres, principalement en Responsabilité Civile, ne sont pas liquidés. Pour résoudre ce problème, différents outils mathématiques sont aujourd'hui à notre disposition, nous nous intéresseront plus particulièrement à l'estimateur de Kaplan-Meier, et aux poids IPCW, qui seront détaillés dans le chapitre 4.1.1.

La notion de durée ayant été introduite, il est maintenant possible de passer à l'analyse des données aviation, cette analyse a également été effectuée sur les données maritime, et permet d'exposer ensuite les différences et similitudes notables avec les données maritimes. Ceci nous permettra de comprendre les raisons des différences de prédictions obtenues sur les jeux de données.

3.2.2 Analyse de l'Aviation

L'analyse des bases de données permet d'obtenir quelques informations quant aux interactions entre les différentes variables, ainsi que l'impact de certaines variables sur la variable cible.

Dans la suite, seuls les sinistres avec un charge ultime supérieure ou égale à 0 seront sélectionnés, ceux-ci ne représentent que -1640$ sur l'exercice 2020, et sont donc négligeables dans le résultat de l'exercice. Dans cette configuration, la base aviation contient 13 502 sinistres au total (clos ainsi que censurés), en dehors de l'exercice 2020. Leur distribution est la suivante (Figure 31) :

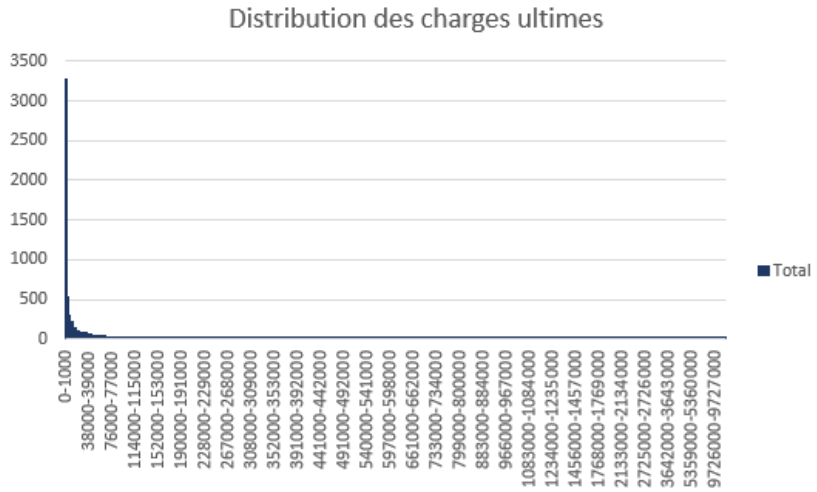


Figure 31-Distrubution de la charge ultime des sinistres sur les exercices de souscription 2013 à 2019

Le montant minimum est de 0$, le montant maximum observé est de 43M$, avec un écart-type de 976k. La distribution est fortement asymétrique, avec une fréquence de sinistres bien plus élevée vers les tranches bases.

Concernant la censure, le graphique suivant reprend le nombre de dossiers clos et censurés, et le taux de provisions dossier à dossier, par exercice de souscription (Figure 32) :

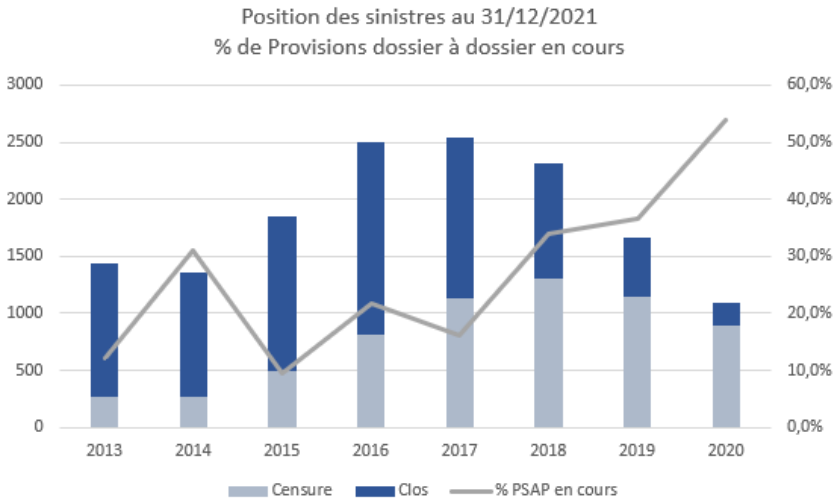


Figure 32- Nombre de sinistres censurés et pourcentages de provisions dossier à dossier en aviation

En termes de montants de provisions dossier à dossier, le pourcentage représente de 9.5% en 2015 à 53% en 2020. Le nombre de sinistres en cours est de 250 environ en 2013, jusqu'à 1400 en 2018.

Tous sinistres confondus (clos et censurés), la durée moyenne d'un sinistre aviation est de 1000 jours. Nous pouvons également noter que, la notion de durée n'est pas corrélée avec la charge ultime des sinistres clos, comme peuvent le prouver les deux graphiques ci-dessous, qui reprennent respectivement la charge ultime par rapport à la durée pour les sinistres inférieurs à 9.5M$ et ceux supérieurs à ce seuil (Figure 33 et 34) :

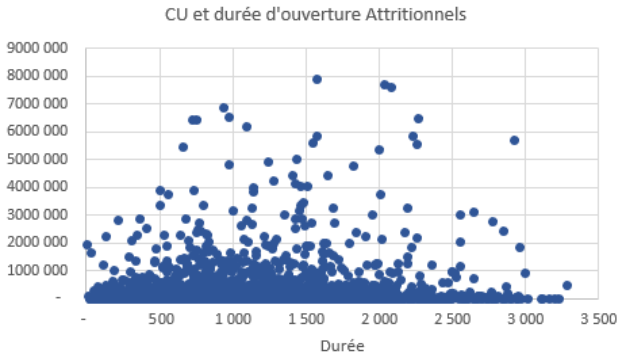


Figure 33- Charges ultimes par rapport à la durée d'ouverture sur les sinistres aviation

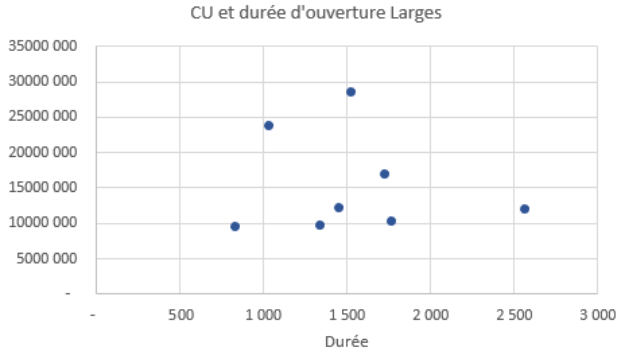


Figure 34-Charges ultimes par rapport à la durée d'ouverture sur les sinistres larges aviation

Cependant, malgré cette observation, comme le montre le graphique suivant (Figure 35) :

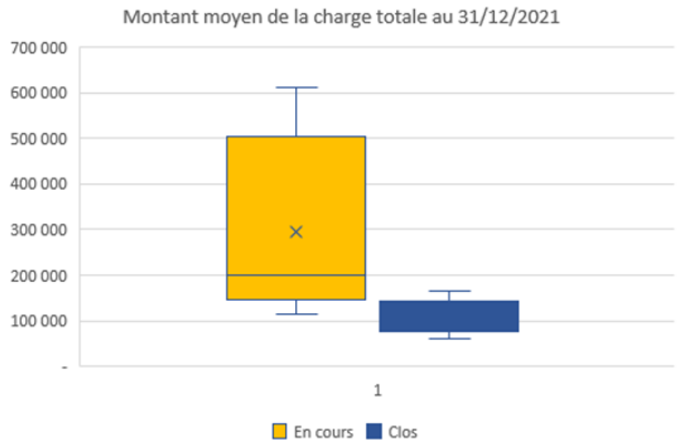


Figure 35-Montant moyen d'un sinistre en cours ou clos

Une charge ultime supérieure pour les sinistres en cours par rapport aux sinistres clos est tout de même notable, si nous comparons les moyennes sur chaque exercice de souscription.

Nous pouvons également noter une forte relation entre la présence de procédure judiciaire, la durée du sinistre ainsi que la charge ultime (Figure 36 et 37) :

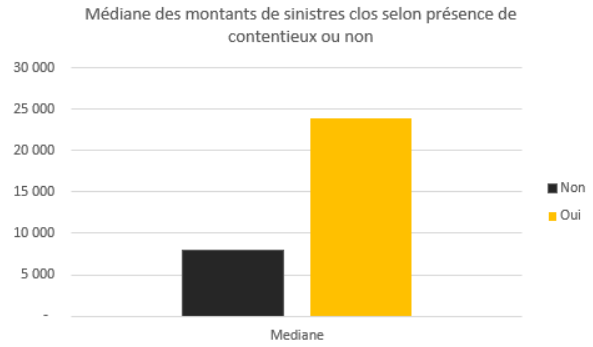


Figure 36-Médiane du montant de sinistre clos avec et sans contentieux

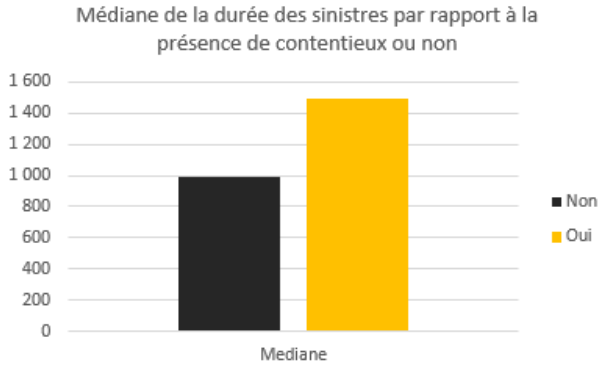


Figure 37-Médiane de la durée d'un sinistre avec et sans contentieux

Sur la figure 36, les sinistres avec contentieux ont une médiane jusqu'à trois fois supérieure à celle des sinistres sans contentieux. De la même manière, Figure 37, la durée d'ouverture des sinistres avec contentieux est également supérieure à celle des sinistres sans contentieux, d'environ 500 jours.

À la suite de ces quelques observations, les données maritimes vont être également étudiées afin de les comparer à celles de l'aviation.

3.2.3 Comparaison avec les données maritimes

Il y a 25 340 sinistres dans la base maritime, sur les exercices de souscription 2010 à 2019 inclus, soit 87.7% de sinistres en plus que la base aviation. De la même manière, seuls les sinistres avec une charge ultime supérieure à 0 sont retenus. Sur l'exercice 2020, aucun sinistre n'est inférieur à 0, il n'y aura donc aucun impact sur les résultats. Le montant maximum de charge ultime est de 10.5M€, et l'écart-type est de 155k€, nous pouvons donc d'ores et déjà remarquer, que malgré une distribution en majeure partie sur la gauche du graphique (disponible en annexe A.2.3) l'écart type est moins élevé que pour la sinistralité aviation.

A l'inverse de l'aviation, la majeure partie des sinistres maritimes sont clos, en effet moins de cent dossiers sont en cours avant 2018, en 2020 un peu plus de 300 dossiers sont en cours. De 2010 à 2019, 4% en moyenne des dossiers sont censurés, avec 33.7% pour l'exercice de souscription 2020.

Le tableau suivant permet de visualiser les différences par exercice de souscription, en termes de nombre de sinistres censurés, ainsi qu'en termes de provisions de sinistres en cours (Tableau 17) :

	Maritime				Aviation			
Exercice de souscription	Censure	Clos	% censure	% Provisions D/D en cours	Censure	Clos	% censure	% Provisions D/D en cours
2010	58	3501	1,6%	**-1%**				
2011	109	4067	2,6%	**22%**				
2012	67	3233	2,0%	**0%**				
2013	53	2693	1,9%	**5%**	274	1161	19,1%	**12%**
2014	24	2319	1,0%	**-1%**	275	1085	20,2%	**31%**
2015	31	2126	1,4%	**0%**	493	1352	26,7%	**9%**
2016	55	1747	3,1%	**43%**	808	1698	32,2%	**22%**
2017	91	1648	5,2%	**35%**	1128	1406	44,5%	**16%**
2018	187	1562	10,7%	**19%**	1307	1001	56,6%	**34%**
2019	338	1474	18,7%	**34%**	1144	514	69,0%	**37%**
2020	626	1230	33,7%	**55%**	893	199	81,8%	**54%**

Tableau 17-Comparaison des pourcentages de censure et des provisions dossier à dossier par exercice de souscription en maritime et aviation

Les exercices de souscription en maritime ont des sinistres majoritairement clos, les provisions sur 2020 sont semblables, peu importe la base observée. Les 22% de provisions dossier à dossier maritime visibles en 2011 proviennent principalement d'un seul sinistre, de charge au 31/12/2021 d'environ 1,8M$, toujours en cours à fin 2021. De la même manière que pour l'aviation, un graphique d'évolution de la clôture des sinistres ainsi que du pourcentage de provisions dossier à dossier en cours est en annexe A.2.3.

Le reste de l'étude est semblable à celle de l'aviation (annexe A.2.3). La durée est très peu corrélée à la charge ultime, malgré une moyenne de charge ultime des sinistres en cours supérieure à celle des sinistres clos, visible sur le diagramme en boîte (annexe A.2.3). La base maritime contient l'information de présence de contentieux sur le sinistre, et cette information a également un impact sur la durée, et la charge ultime des sinistres clos étudiés.

Si nous comparons les corrélations entre certaines variables quantitatives communes entre les deux bases, nous pouvons remarquer certaines similitudes (Figures 38 et 39) :

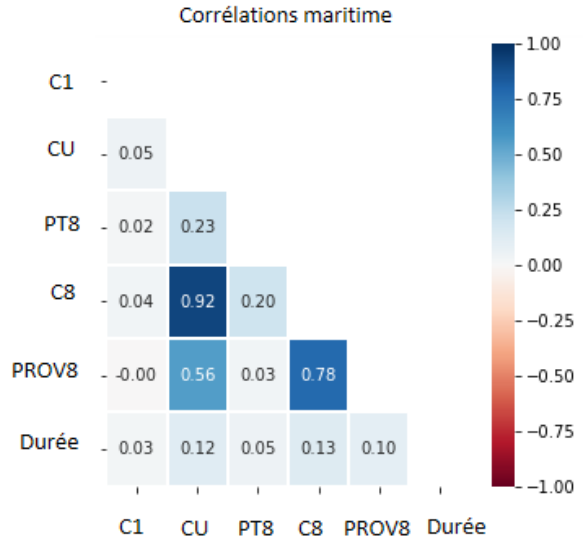


Figure 38-Matrice de corrélation des variables les plus importantes en maritime

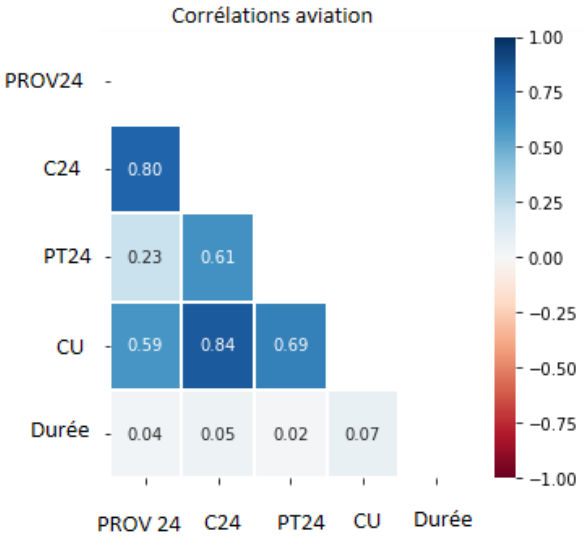


Figure 39-Matrice de corrélation des variables les plus importantes en aviation

Les provisions en année deux (soit les variables *PROV8* en maritime et *PROV24* en aviation) sont corrélées à plus de 50% avec la charge ultime *CU* et environ 80% avec la charge en année 2 (*C8* en maritime et *C24* en aviation). La charge ultime est corrélée à 92% et 84% pour respectivement le maritime et l'aviation. La durée est corrélée à seulement 12% à la charge ultime maritime, mais reste tout de même plus élevée que la corrélation observée en aviation, qui est proche de 0%.

Les corrélations les plus importantes entre les variables explicatives et la charge ultime, qu'il s'agisse de l'aviation ou du maritime, sont sensiblement les mêmes.

Certains sinistres, très tardifs, représentés ici par ceux déclarés deux années après l'exercice de souscription, ne représentent, ni en aviation ni en maritime, une forte quantité de sinistres, mais celle-ci n'est pas pour autant négligeable. Le tableau suivant reprend le pourcentage de charge ultime et nombre de sinistres par exercice de souscription, sur les sinistres clos (Tableau 18) :

	Maritime		Aviation	
Exercices de souscription	**Charge ultime**	**Nombre de dossiers**	**Charge ultime**	**Nombre de dossiers**
2010	2%	5%		
2011	1%	2%		
2012	2%	3%		
2013	3%	5%	6%	9%
2014	3%	5%	4%	7%
2015	4%	5%	3%	6%
2016	1%	4%	3%	6%
2017	1%	3%	1%	6%
2018	2%	2%	4%	2%
2019	0%	2%	0%	0%
Moyenne	**2,0%**	**3,5%**	**3,1%**	**5,1%**

Tableau 18-Comparaison du pourcentage de sinistres déclarés après deux années de développement en maritime et en aviation

Nous notons une moyenne de 2% en termes de charge ultime, sur les sinistres clos des exercice 2010 à 2019 du maritime, et 3% pour l'aviation entre 2013 et 2019.

De ce fait, les sinistres tardifs seront ajoutés à la charge ultime totale calculée par les méthodes lignes à lignes, afin de compenser leur absence lors de la prédiction des résultats.

Il y a de nombreuses similitudes entre les deux types de données. Cependant, les principales différences entre les deux bases proviennent en premier lieu de la différence d'écart-type sur les données, qui montre une variance bien plus élevée en aviation, et en deuxième lieu, à la quantité de censure nettement supérieure en aviation.

Cette dernière différence aura un impact non négligeable sur la suite de l'étude, principalement sur le calcul des poids IPCW, qui va être abordé dans le chapitre suivant.

4 Application

4.1 KAPLAN MEIER ET POIDS IPCW

L'estimateur de Kaplan-Meier, ou produit-limite, permet d'estimer de façon non paramétrique la fonction de survie, d'après des données de durée de vie. Cet estimateur est couramment utilisé en économie, écologie, ou médecine. La courbe obtenue est une série de marches horizontales de grandeur décroissante, qui permet d'approcher la fonction de survie réelle de l'échantillon étudié. Cet estimateur peut prendre en compte certains types de données censurées, en particulier celles qui sont présentes dans notre jeu de données, les censures à droite. Il est à noter que si aucune troncature ou censure est observée, la courbe de Kaplan-Meier s'apparente à une courbe de survie classique.

Dans cette partie nous allons développer la théorie de l'estimateur de Kaplan Meier, ce que représentent les poids IPCW, puis nous terminerons par l'application du calcul de ces poids sur nos sinistres.

4.1.1 Définitions et théories

Kaplan Meier

L'analyse de survie permet de fournir des outils statistiques, afin d'étudier l'évolution d'un évènement au cours du temps. Ces outils permettent d'obtenir une courbe de survie, ou encore de calculer la probabilité de survenue d'un évènement à un instant t. Ne pas tenir compte de ces outils lors de l'analyse d'une base de données censurée, peut générer un biais sur les résultats.

La courbe de survie décrit l'évolution de la survie au cours du temps. Elle représente donc la dynamique de survenue d'un évènement au cours du temps. Couramment utilisée en médecine, elle peut représenter des décès, des sorties de maladie, ou des rechutes. Dans notre cas, la clôture des sinistres.

Afin d'établir un courbe de survie, il est important d'avoir à disposition les données suivantes :

- La date de début d'observation : date de départ de l'étude, identique pour tous les sinistres. Dans notre cas, il s'agit du 01/01/2013 en aviation et du 01/01/2010 en maritime.
- La date de fin d'observation : Date de fin du suivi, dans notre cas le 31/12/2021, pour les deux bases.
- La date de clôture : Date à laquelle le sinistre a été clos, si le sinistre est toujours en cours, alors il s'agit de la date de fin d'observation.

La fonction de survie est monotone, décroissante et continue, et vérifie :

$$S(0) = 1$$

$$\lim_{t \to \infty} S(t) = 0$$

Il existe plusieurs méthodes pour approcher cette courbe, soit les méthodes paramétriques, se basant sur des fonctions connues, telles que la fonction exponentielle, Weibull ou encore Gompertz. Dans ce cas, on estime que la forme de la courbe de survie $S(t)$ est déjà connue. Soit les méthodes non paramétriques, avec entre autres, la méthode de Kaplan Meier.

Si nous reprenons les notations du chapitre 3.2.1:

- $T \rightarrow$ v.a.r continue et supérieur à 0, représentant la durée,
- $C \rightarrow$ v.a.r continue et supérieur à 0, représentant la censure,
- $Y = \min(C; T) \rightarrow$ La durée observée,
- $\delta_i \rightarrow$ Survenance ou non de l'évènement,
  Dans notre cas, $\delta_i = 1$ si clôture du dossier, sinon 0.

L'estimateur de Kaplan-Meier repose sur le fait qu'être en vie à un instant t, c'est être en vie juste avant t, et ne pas mourir en t. Par exemple : être en vie jusqu'au 365ème jour, c'est survivre le 365ème jour sachant que la personne est en vie le 364ème jour. Cet estimateur est particulièrement utilisé par les actuaires, lors l'établissement des tables de mortalité, ou des lois de maintien en invalidité.

En effet, comme détaillé par Foucher, Y. et également Saint Pierre, P. (2021), si $t_1 < t_2$ alors nous pouvons écrire :

$$P(Y > t_2) = P(Y > t_2 | Y > t_1) = P(Y > t_2 | Y > t_1) P(Y > t_1).$$

Lorsque cette formule est généralisée sur tous les couples $(Y_i; \delta_i)$ avec $i = 1 \dots N$, nous obtenons :

$$S_T(Y_N) = \prod_{i=1}^{N} P(T > Y_i | T > Y_{i-1}).$$

La survie à n'importe quel temps de t tel que $t < Y_N$, sera donc définie par :

$$S_T(t) = \prod_{Y_i < t} P(T > Y_i | T > Y_{i-1}).$$

Si nous posons $t_0 = 0$, $n_i = \sum_{j=1}^{n} \mathbb{1}_{Y_j \geq t_i}$ le nombre d'individus exposés au risque de clôture au temps t_i et $d_i = \sum_{j=1}^{n} \delta_j \mathbb{1}_{Y_j = t_i}$ le nombre de clôtures survenues en t_i alors la probabilité d'être clôturé dans l'intervalle de temps $|Y_{i-1}, Y_i|$ s'écrit :

$$P(T \leq Y_i | T > Y_{i-1}) = \frac{d_i}{n_i}.$$

Les probabilités de survie conditionnelles sont définies par :

$$P(T > Y_i | T > Y_{i-1}) = 1 - P(T \leq Y_i | T > Y_{i-1}) = 1 - \frac{d_i}{n_i}.$$

L'estimateur de Kaplan Meier de la survie est finalement :

$$\hat{S}_T(t) = \prod_{Y_i < t} \left(1 - \frac{d_i}{n_i}\right).$$

Comme les temps sont distincts, nous avons :

- $d_i = 0$ en cas de censure en Y_i donc quand $\delta_i = 0$,
- $d_i = 1$ en cas de clôture en Y_i donc quand $\delta_i = 1$.

Nous pouvons également l'écrire :

$$\hat{S}_T(t) = \prod_{Y_i \leq t} (1 - \frac{\delta_i}{\sum_{j=1}^{n} \mathbb{1}_{Y_j \geq Y_i}}).$$

Chaque saut de marche correspond à un $S(t)$ différent, sachant que $S(t)$ prend une valeur différente à chaque observation d'un évènement, alors le nombre de saut de la courbe, correspond au nombre d'évènements (s'il n'y a pas d'ex-aequo), ou inférieur au nombre d'évènement (s'il y a des ex-aequo).

Maintenant que l'estimateur de Kaplan-Meier a été abordé, nous pouvons passer aux détails concernant les poids IPCW.

Sauts de Kaplan Meier et poids IPCW

Comme expliqué par Lefaou Y. (2019), la censure de T par C est symétrique à la censure de C par T, nous pouvons donc écrire l'estimateur de Kaplan Meier de la **censure** C comme étant :

$$\hat{S}_C(t) = \prod_{Y_i \leq t} (1 - \frac{1 - \delta_i}{\sum_{j=1}^{n} \mathbb{1}_{Y_j \geq Y_i}}).$$

Si nous posons $\hat{S}_Y(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{Y_i > t}$ la fonction de **survie empirique de Y**, alors nous pouvons écrire :

$$\hat{S}_C \,.\hat{S}_T = \hat{S}_Y.$$

Et sa dérivée comme étant :

$$d\hat{S}_C \,.\hat{S}_T + \hat{S}_C \,.d\hat{S}_T = d\hat{S}_Y.$$

Si nous nous plaçons en $t = Y_i$ pour une observation non censurée, alors $d\hat{S}_Y(t) = \frac{1}{n}$ et $d\hat{S}_C(t) = 0$.

Nous pouvons donc écrire :

$$d\hat{S}_T(t) = \frac{1}{n\hat{S}_C(t)},$$

Ce qui nous donne l'estimateur de Kaplan Meier de T tel que :

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

Lors d'observations ne contenant aucune censure, un poids égal à $\frac{1}{n}$ est attribué à chaque observation. Dans le cas de la censure à droite, le poids est de la forme :

$$\widehat{W}_i = n^{-1} \frac{\delta_i}{\hat{S}_C(Y_i)}.$$

Ces poids sont donnés aux valeurs non censurées, celles censurées ont un poids nul, mais interviennent tout de même dans le poids des valeurs non censurées, par le biais de l'estimation de $\hat{S}_C(Y_i)$. L'utilisation des W_i vient compenser la censure des données, qui tend à raréfier les observations de grandes valeurs de T. Le

poids associé à une observation non censurée est plus important, si Y_i est plus important (pour rappel $Y_i = \min(C_i\,;T_i)$).

Nous allons maintenant introduire le concept de poids IPCW (*Inverse Probability of Censoring Weighting*), comme détaillé par Lefaou Y. Il est en effet possible d'estimer les poids $W_i = n^{-1}.\frac{\delta_i}{S_C(Y_i)}$, par $\widehat{W}_i$ défini ci-dessus :

Pour tout ψ dans l'intervalle $[0,\tau[$ avec $\tau = \inf\{t \geq 0 : P(C \geq t) = 0\}$, et si nous posons l'indépendance entre T et C représentée par $P(C > T|T) = S_C(T)$.

Alors nous pouvons écrire :

$$E\left[\frac{\delta}{S_C(Y)}.\psi(Y)\right] = E\left[\frac{\mathbb{1}_{C\geq T}}{S_C(T)}.\psi(T)\right] = E\left[\frac{\psi(T)}{S_C(T)}.E[\mathbb{1}_{C\geq T}|T]\right].$$

D'après l'hypothèse ci-dessus, $E[\mathbb{1}_{C\geq T}|T] = P(C \geq T|T) = S_C(T)$, nous pouvons donc écrire l'égalité suivante :

$$E\left[\frac{\psi(T)}{S_C(T)}.E[\mathbb{1}_{C\geq T}|T]\right] = E\left[\frac{\psi(T)}{S_C(T)}.S_C(T)\right] = E[\psi(T)].$$

Par cette égalité, il est possible d'estimer la distribution de T à partir des données censurées, si T et C sont dépendants. A chaque Y_i est alloué un W_i et pour chaque observation non censurée nous pouvons écrire,

$$\frac{\delta_i}{S_C(Y_i)} = \frac{1}{P(\delta_i = 1|T_i)}.$$

W_i est l'inverse de la probabilité d'être observé sachant T_i.

Maintenant que les bases de la méthode ont été exposées, nous pouvons passer à la méthodologie appliquée à nos jeux de données, et les résultats obtenus sur chacun d'eux.

4.1.2 Méthodologie et résultats aviation et maritime

Méthodologie

Si nous reprenons les notations de O.Lopez et al. (2016), et notons le vecteur (M_i, T_i, X_i) représentant notre base de données tel que :

- $T_i \in \mathbb{R}^+ \rightarrow$ Représente la durée de vie du sinistre observé
- $M_i \in \mathbb{R}^p \rightarrow$ Représente la charge ultime à prédire
- $X_i \in \mathrm{X} \subset \mathbb{R}^d \rightarrow$Représente les variables explicatives. Les variables explicatives de notre base de données, ayant un impact sur T ainsi que sur M.
- $C_i \in \mathbb{R}^+ \rightarrow$ La censure des sinistres non clos à la date d'observation

Par cette censure, les variables $(M\,;T)$ ne sont pas directement observables, et les variables observées sont les suivantes :

$$Y_i = \min(C_i; T_i),$$

$$\delta_i = \mathbb{1}_{T_i \leq C_{i,}}$$

$$X_i = \left(X_i^1, \dots, X_i^n\right),$$

$$N_i = \delta M_i.$$

Sous les hypothèses suivantes :

$$H_1 : C\ et\ (M,T)\ sont\ indépendants$$

$$H_2 : P(T \leq C|M,T,X) = P(T \leq C|T)$$

A partir de ces informations, des poids IPCW peuvent être calculés pour chaque observation i.i.d du vecteur $(N_i\,; Y_i, \delta_i, X_i)$.

Dans la pratique, les poids sont calculés avec le logiciel R grâce aux packages *pec* et *survival* et la fonction *prodlim.* Cette phase se déroule en trois étapes :

- Dans un premier temps, les données des bases sont ordonnées par durée Y_i.
- Dans un second temps, la fonction de survie de la variable de censure est calculée grâce à l'estimateur de Kaplan Meier, soit la fonction $\hat{S}_C(Y_i)$.
- Pour terminer, les poids IPCW sont obtenus en calculant $\widehat{W}_i = \frac{\delta_i}{\hat{S}_C(Y_i)}$. Dans le cas où il y a censure, le poids est égal à 0, dans le cas où le sinistre est clos, nous revenons à la définition des poids $\widehat{W}_i$, soit l'inverse de la probabilité de censure.

Dans la partie suivante, les poids et les graphiques de chaque base seront exposés.

Chacun des modèles de *Machine Learning* sera entrainé sur les données pondérées par ces poids. De ce fait, si un poids x est attribué à une observation, alors l'observation sera répétée x fois lors de la phase d'apprentissage, afin de lui donner plus de poids dans l'étude.

Résultats des poids aviation et maritime

En comparant les fonctions de survie de la variable de censure dans les deux graphiques ci-dessous (Figure 40) nous notons ici que les sinistres maritimes sont clos plus rapidement qu'en aviation, ce qui confirme les différences de niveau de censure observées précédemment :

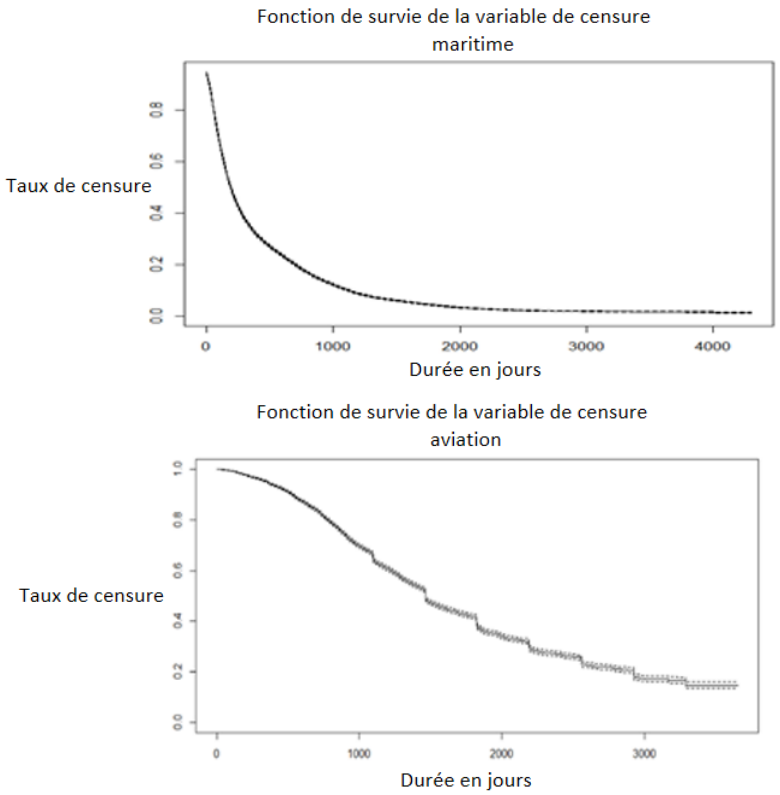


Figure 40-Fonctions de survie de la variable de censure en maritime (premier graphique) et en aviation (deuxième graphique)

Les poids sont répartis entre 0 et une valeur maximum différente selon la branche. Les résultats suivants sont obtenus grâce à la fonction *Summary* sous R (Tableau 19) :

Activité	Minimum	1er quartile	Médiane	Moyenne	3ième quartile	Maximum
Aviation	0	0	1.0096	0.8538	1.2071	11.4857
Maritime	0	1.0001	1.0003	0.9872	1.0044	17.0357

Tableau 19-Résumé des poids IPCW aviation et maritime

Certains sinistres seront donc répétés jusqu'à 17 fois en maritime et 11 fois en aviation, lors de l'apprentissage du modèle, afin de tenir compte leur durée élevée, leur observation étant de ce fait jugée peu fréquente dans la base de données.

En observant plus en détail les poids donnés aux différentes observations par rapport à leur charge ultime (Figures 41 et 42) :

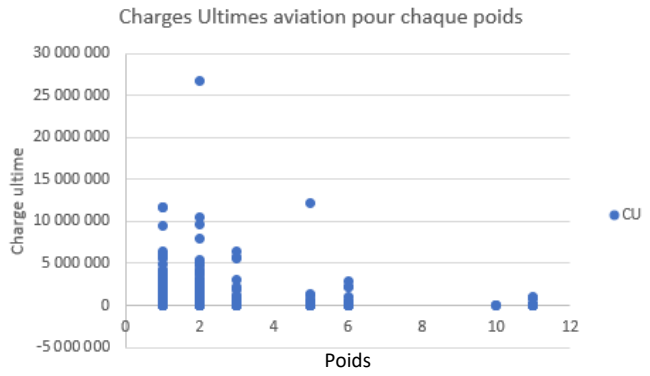


Figure 41-Poids IPCW attribués par rapport aux charges ultimes aviation

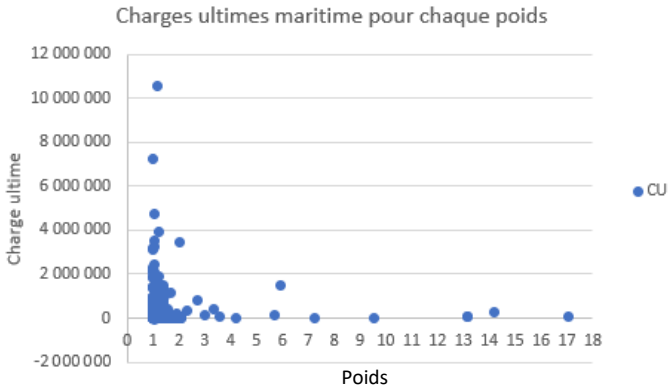


Figure 42-Poids IPCW attribués par rapport aux charges ultimes maritime

Nous pouvons remarquer que les sinistres avec une charge ultime élevée n'ont pas forcément un poids plus élevé. Cela se comprend puisque la durée n'est pas ou très peu corrélée à la charge ultime. Les charges les plus élevées ne seront représentées qu'une à deux fois dans l'apprentissage, permettant difficilement de compenser le peu d'observations disponibles à ce niveau de charge.

4.2 MODELES APPLIQUES

Nous avons étudiées les données, et la création des poids IPCW, les méthodes de *Machine Learning* utilisées pour estimer la charge ultime de l'exercice 2020 vont être maintenant détaillées. Les bases fondamentales pour comprendre le fonctionnement de chaque modèle seront exposées, puis nous expliquerons l'importance de l'ajustement des différentes variables de chaque algorithme, ainsi que la méthode employée, puis nous terminerons par les différents résultats sur les deux bases de données, et les différences observées avec la méthode *Chain Ladder*.

4.2.1 Arbres de décision

Présentation

Les arbres de décisions font partie des méthodes non paramétriques supervisées utilisées pour la régression ou la classification. Un apprentissage supervisé, contrairement au non supervisé, repose sur un ensemble de variables d'entrées x, permettant d'approcher une variable de sortie y.

Comme expliqué par Miclet L. et Cornuéjols A. (2010) « L'apprentissage des arbres de décision procède par une exploration du général au particulier, en commençant par un arbre à un nœud racine correspondant à une partition simple de l'espace X, puis en raffinant progressivement cette partition par ajout successif de nœuds dans l'arbre, ce qui revient à subdiviser itérativement les partitions de l'espace des exemples. »

L'approche est appelée *top-down induction of decision trees*, et commence à partir de l'échantillon complet. Plusieurs découpages en sous échantillon seront fait de manière itérative jusqu'à l'obtention de l'échantillon le plus « pur » possible. Le but étant de pouvoir apprendre correctement sur les données dites d'apprentissages, afin d'éviter le surajustement ou sous-ajustement, afin que le modèle se généralise correctement. Ces notions seront abordées plus précisément dans le chapitre 4.2.6.

De façon schématique :

Dans le cas des arbres de classification, l'algorithme découpe l'espace des entrées $X_d = \{x_1, \dots, x_i, \dots, x_d\}$ en différentes régions, dont les côtés sont des hyperplans perpendiculaires aux axes (Figure 43) :

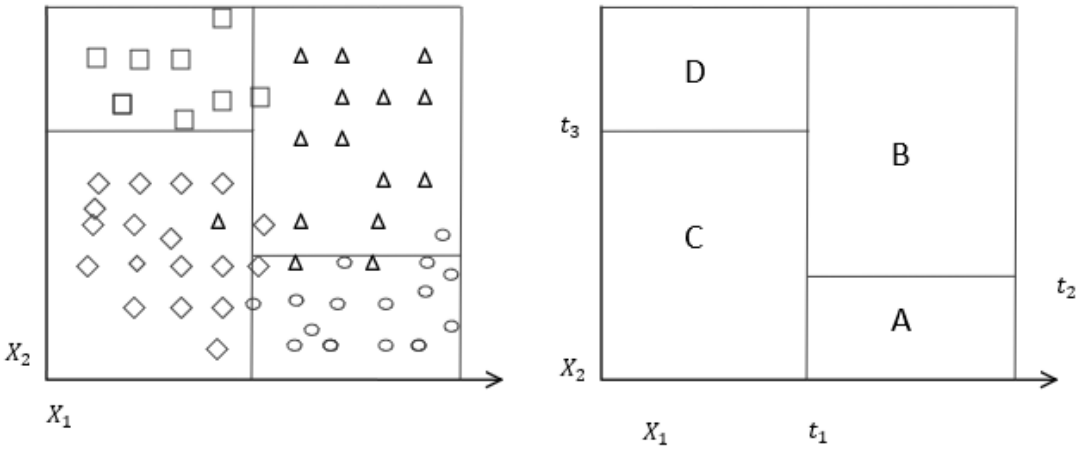


Figure 43-Méthode de partitionnement en classification

Chacune de ces régions ci-dessus comprend une seule et unique classe parmi les valeurs prédites. Plus l'ajustement sera précis, plus ces régions seront précisément découpées.

De cette catégorisation découle un arbre de classification simple (Figure 44) :

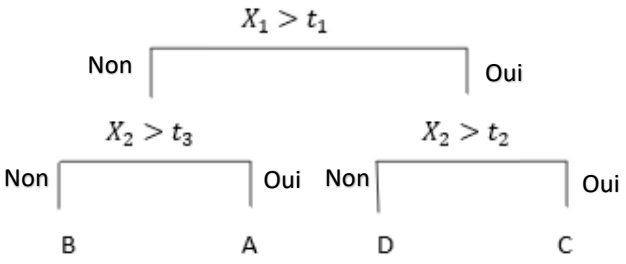


Figure 44-Exemple d'arbre de classification lié au partitionnement de la figure 37

De la même manière que pour la classification mais pour des valeurs continues, le schéma suivant reprend le principe de l'approximation par les arbres de régression (Figure 45) :

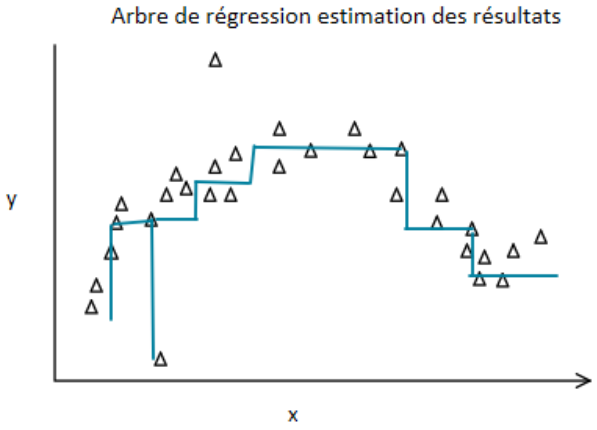


Figure 45-Exemple de résultat par arbre de régression

Ci-dessus, la courbe bleue va approximer les résultats de y, en fonction des informations contenues dans x.

Les arbres de décisions sont en général les premiers arbres étudiés, puisque ceux-ci sont simple à comprendre, et peuvent être facilement visualisés, comme l'illustre l'exemple ci-dessous (Figure 46) :

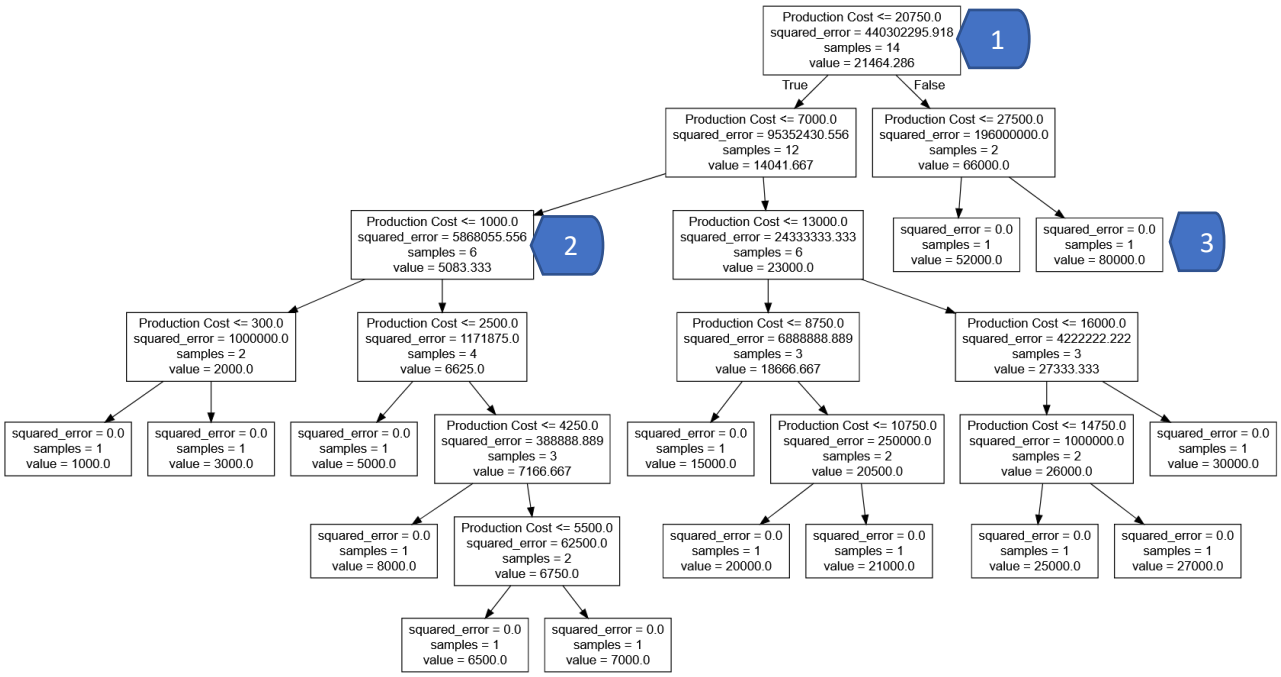


Figure 46-Sortie d'un arbre de régression sous python

Nous pouvons distinguer ci-dessus, plusieurs objets :

- Le nœud initial (1) : Il s'agit de la racine, qui représente l'ensemble de la base étudiée, le départ de l'arbre.
- Les nœuds intermédiaires (2) : chaque nœud est défini par un test construit à partir d'une variable, permettant la création d'autres nœuds.
- Les nœuds terminaux (3) : Ils n'ont pas de descendants, contrairement aux nœuds intermédiaires.

Dans l'exemple précédent, au nœud initial, un premier test est effectué (Production Cost <= 20 750.0). La réponse est « vrai » ou « faux ». Cette première étape permet de séparer la base de données en deux sous-ensembles, les données inférieures ou égales à 20 750.0, et les autres. Sur ces deux nouvelles sous

catégories, une nouvelle question binaire est posée, ainsi de suite jusqu'à l'obtention de tous les nœuds terminaux.

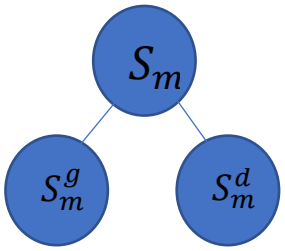
Nous allons maintenant étudier les formules appliquées, ainsi que les limites de cette méthode, avant de passer à la méthode des forêts aléatoires.

Fonctionnement et limites

Pour chaque apprentissage, la base d'étude doit avant tout être séparée en deux, en créant la base d'apprentissage, et la base de test. Sur la base d'apprentissage, qui reprend entre 70% et 80% de la base initiale, le modèle sera entrainé. Puis pour vérifier sa capacité à prédire de nouvelles données, le modèle est appliqué à la base de test, sur laquelle les métriques d'erreur de régression usuelles sont calculées, afin d'observer la capacité du modèle à se généraliser.

Si nous notons S_m le nœud interne, pour $x_i \in \mathbb{R}^n$ avec $i = (1, \dots, l)$ dans la base d'apprentissage et le vecteur $y \in \mathbb{R}^l$ contenant les données cible, un arbre de décision va partitionner les données, pour que les données de valeurs similaires soient groupées.

Si S_m est un nœud interne avec n_m échantillons, il sera séparé sur l'attribut j au seuil a_j, et donnera naissance à deux descendants (Figure 47) :



S_m^d le sous nœud de gauche, qui contient tous les attributs qui ont des valeurs $v_j \leq a_j$ et S_m^g qui contient tous les attributs tels que $v_j > a_j$.

Figure 47-Noeud interne et ses deux descendants

Nous pouvons formaliser avec $\theta = (j, a_m)$ le critère de séparation, $S_m^g(\theta) = \{(x, y) | x_j \leq a_m\}$ et $S_m^d(\theta) = \frac{S_m}{S_m^g(\theta)}$.

Dans le cas de la régression, la méthode *decisiontreeregressor* de la librairie *Sklearn* de python sélectionne θ^*qui minimise l'impureté :

$$\theta^* = argmin_\theta G(S_m; \theta),$$

Avec :

$$G(S_m; \theta) = \frac{n_m^g}{n_m} H\left(S_m^g(\theta)\right) + \frac{n_m^d}{n_m} H\left(S_m^d(\theta)\right),$$

Et :

$$H(S_m) = \frac{1}{n_m} \sum_{y \in S_m} (y - \bar{y}_m)^2.$$

La création des nœuds se poursuit jusqu'à ce que la profondeur maximale soit atteinte, ou si $n_i = 1$ ou encore s'il n'est plus possible de séparer les données. Si aucune variable n'est capable de séparer l'échantillon observé, alors l'arbre est terminé, et ne contient plus que des nœuds terminaux.

Ces modèles comportent plusieurs limites, ils peuvent parfois être complexes, et mal se généraliser sur de nouvelles valeurs. Ce manque de généralisation peut être diminué en fixant la profondeur de l'arbre, afin de minimiser le surajustement par exemple. Ils résistent également assez mal aux variations de données, et peuvent donc être très instables, lorsque le portefeuille étudié n'est plus du tout représenté par le passé. Il est également possible de noter un biais plus ou moins fort, causé par des données déséquilibrées.

Les arbres de décision sont donc une bonne méthode à appliquer pour une première approche en *Machine Learning*, mais il est important de passer rapidement à des méthodes plus fiables, comme celle des forêts aléatoires.

4.2.2 Forêts aléatoires

Présentation

Les forêts aléatoires, comme proposé par Breiman en 2001, se basent sur la technique du *bagging*, contraction de *bootstrap aggregation*, toujours proposée par Breiman en 1996. La méthode du *bagging* repose sur la combinaison d'hypothèses, afin d'obtenir une hypothèse finale. De ce fait, plusieurs arbres de régression définis précédemment non élagués, dont les paramètres ne sont donc pas ajustés, sont entrainés à partir de plusieurs bases d'apprentissage, afin d'obtenir plusieurs résultats. Ces bases sont créées par la méthode du *bootstrap*, qui consiste en de multiples tirages avec remises. Pour chaque tirage, une hypothèse est obtenue, et l'hypothèse finale n'est que la moyenne des hypothèses obtenues sur B tirages. Si nous posons $h_b(x)$ l'hypothèse obtenue pour le tirage b ,le résultat de l'hypothèse finale est le suivant :

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

L'algorithme du *bagging* peut se résumer comme étant :

ALGORITHME BAGGING

Soit B la base d'apprentissage

Pour b allant de 1 à B :

Tirer un échantillon *bootstrap* e_b de la base initiale
Calculer le modèle $h_b(x)$ associé

Fin pour

Agrégation par la moyenne en calculant $H(x) = \frac{1}{B}\sum_{b=1}^{B} h_b(x)$

La moyenne utilisée à la fin de cet algorithme permet de réduire la variance, car pour rappel, pour plusieurs variables $X_1, \dots, X_n$ i.i.d de moyenne μ et variance σ^2, la variance de la moyenne de ces variables est $\frac{\sigma^2}{n}$.

L'algorithme des forêts aléatoire est le plus commun parmi les techniques de bagging. Il ne fait que reprendre le bagging, en y ajoutant un critère de décorrélation des arbres. Nous allons maintenant entrer plus en détail sur le fonctionnement de cet algorithme, ses avantages et ses possibles limites.

Fonctionnement et limites

Cette recherche de décorrélation, c'est-à-dire minimiser la corrélation tout en gardant la variance la plus faible possible, consiste d'après Miclet L. et Cornuéjols A. (2010), en la sélection aléatoire de n variables, à considérer à chaque étape de sélection du meilleur nœud, avec n un sous ensemble des variables explicatives de la base d'apprentissage. Lorsque d est le nombre de variables explicatives alors $n = \sqrt{d}$. Cette décorrélation permet de minimiser au maximum la variance, si chacune des variables a une variance de σ^2 alors la variance finale est $\frac{1}{B}\sigma^2$. Si ces variables sont identiques mais pas nécessairement indépendantes, alors nous avons : $\frac{1-\rho}{B}\sigma^2 + \rho\sigma^2$.

L'algorithme de la forêt aléatoire est évidemment proche de celui du bagging. Il peut s'écrire comme suit :

ALGORITHME FORETS ALEATOIRES

Soit B la base d'apprentissage

Pour b allant de 1 à B :

Tirer avec remise, un échantillon *bootstrap* e_b de la base d'apprentissage
Créer d'un arbre sur l'échantillon e_b :
Pour chaque nœud :
Tirer au hasard une partie des variables explicatives (q parmi les p)
Identifier la variable optimale sur ce sous-ensemble
Créer les deux nœuds suivants
Fin pour
Fin pour

Agrégation par la moyenne en calculant $H(x) = \frac{1}{B}\sum_{b=1}^{B} h_b(x)$

L'algorithme donne généralement de bons résultats, surtout sur des données à grandes dimensions. Il s'agit d'un modèle simple à mettre en œuvre, et qui nécessite peu de paramètres lors de sa mise en place. Cependant, il est moins simple à appréhender et à visualiser que les arbres de décision.

4.2.3 Gradient Boosting

Présentation

Le *boosting* est proche de la méthode du bagging, puisqu'il repose aussi sur plusieurs modèles qui sont consolidés afin d'obtenir la prédiction finale, il fait partie des modèles dits ensemblistes. Cette consolidation se fait cette fois à l'aide de pondérations. Chaque arbre de décision imbriqué apprendra des erreurs du précédent, les *weak learners* pour lesquels leur résultat est légèrement meilleur qu'un résultat aléatoire, et chaque observation incorrectement prédite aura un poids plus élevé dans le prochain modèle. Le *boosting* est un modèle qui se renforce sur les données les plus compliquées à prédire.

Afin d'aisément visualiser la méthode, le schéma suivant reprend le modèle de classification.

Schématiquement (Figure 48) :

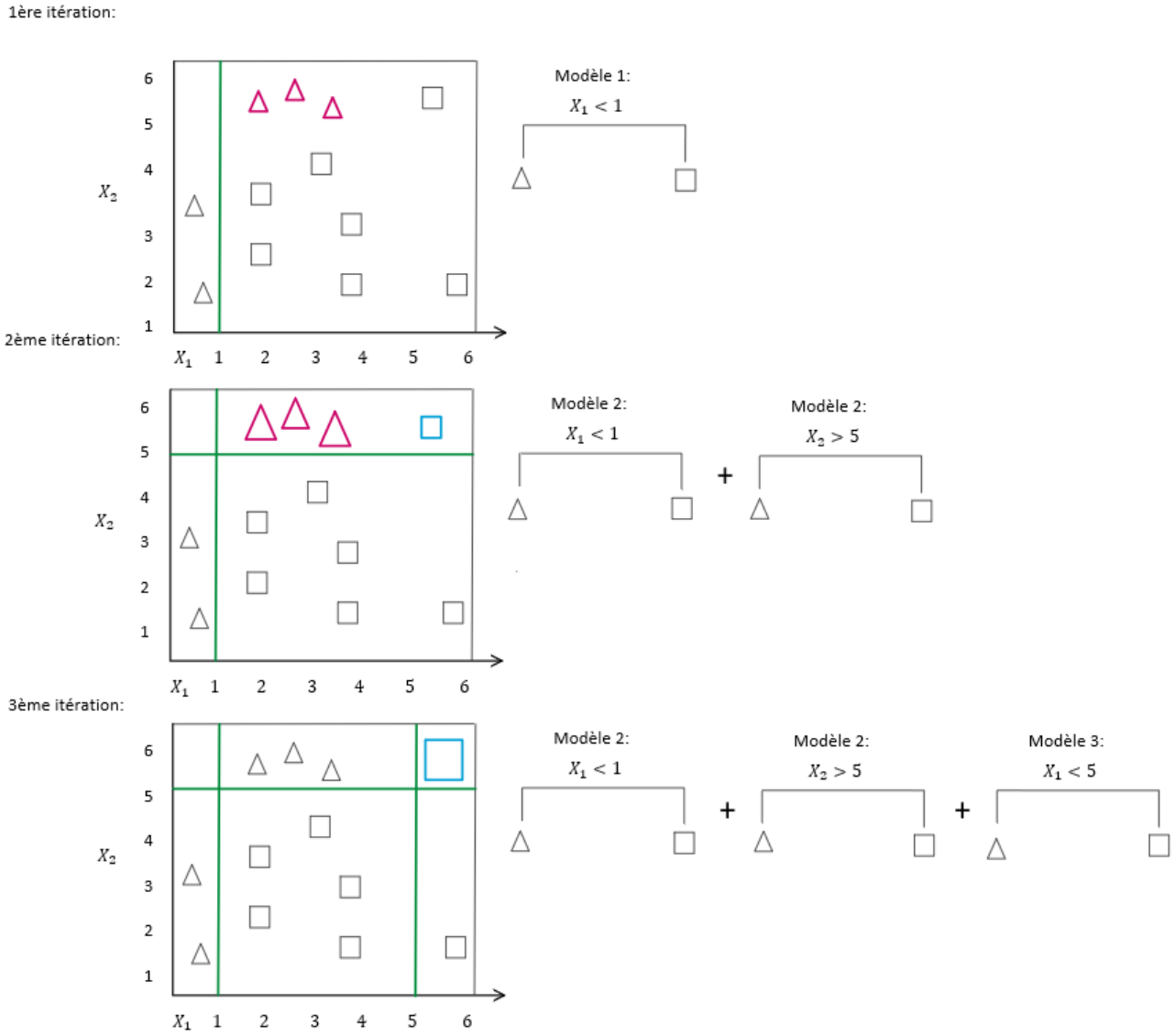


Figure 48-Fonctionnement de la méthode du boosting de gradient (source Zhang Z. et al. (2018))

La première itération permet de scinder une fois les données, de façon succincte. A la deuxième itération, les poids des données mal classées sont augmentés (les triangles rouges), afin de mieux les isoler. Une deuxième séparation est ainsi faite. A la troisième itération, le poids du carré bleu est augmenté, afin de pouvoir le séparer des autres classes.

La paramétrisation de cet algorithme repose principalement sur deux paramètres :

- Le nombre d'estimateurs, qui donne le nombre d'apprenants faibles, soit le nombre d'arbres CART.
- Le *Learning rate* qui régule le risque de surajustement par un paramètre γ.

Fonctionnement et limites

Dans le *Gradient Boosting*, d'après BENTEJAC, C. et al. (2019) et BARRA, V. et al. (2021), la prédiction finale d'une base de données $D = \{x_i, y_i\}$ pour un input précis x_i est noté y_i :

$$y_i = F_M(x_i) = \sum_{m=1}^{M} \rho_m h_m(x_i).$$

Ici, h_m représente les m fonctions, soit les apprenants faibles (*weak learners),* dans ce cas les arbres de décision. M correspond au nombre d'estimateurs, soit le nombre d'arbres CART, et ρ_m le poids de chaque fonction h_m.

Comme nous avons pu le voir dans le schéma précédent, chaque itération vient s'ajouter à la précédente, tel que :

$$F_m(x) = F_{m-1}(x) + \rho_m h_m(x).$$

Le but du *Gradient Boosting* est de trouver une approximation de la fonction $F^*(x)$, en minimisant la fonction de perte $L\big(y, F(x_i)\big)$. L'approximation de $F^*(x)$ est donnée par la formule ci-dessus.

Le premier modèle F_0est une constante, qui peut s'écrire :

$$F_0(x) = argmin_\alpha \sum_{i=1}^{N} L(y_i, \alpha).$$

A chaque itération de l'algorithme, l'arbre donnant h_m est ajouté tel que la somme des pertes soit minimisée :

$$h_m = argmin_h\, L_m = argmin_{h,\rho} \sum_i L\big(y_i, F_{m-1}(x_i) + \rho h(x_i)\big).$$

Cependant, au lieu de résoudre directement ce problème d'optimisation, la méthode de descente de gradient est utilisée. La descente de gradient revient à minimiser une fonction en se déplaçant dans le sens opposé de la direction du gradient.

$$\theta_i = \theta_i - \frac{\rho(\partial J)}{\partial \theta_i}.$$

Nous pouvons donc écrire dans notre problème,

$$\frac{\partial J}{\partial \theta_i} = \frac{\partial \sum_i L\big(y_i, F(x_i)\big)}{\partial F(x_i)} = \frac{\partial L\big(y_i, F(x_i)\big)}{\partial F(x_i)} = F(x_i) - y_i.$$

Nous pouvons interpréter les résidus $y_i - F(x_i)$ comme le négatif du gradient :

$$y_i - F(x_i) = -\frac{\partial L\big(y_i, F(x_i)\big)}{\partial F(x_i)}.$$

Enfin, si nous écrivons l'algorithme :

ALGORITHME *GRADIENT BOOSTING*

Initialisation du modèle avec une constante :

$$F_0(x) = argmin_\alpha \sum_{i=1}^{N} L(y_i, \alpha).$$

Pour m allant de 1 à M :

Calcul des résidus $r_{im} = -\left[\frac{\partial L(y_i, F(x))}{\partial F(x)}\right]_{F(x)=F_{m-1}(x)}$ pour i=1, …, n

Ajustement d'un algorithme, de faible performance $h_m(x)$, soit la création d'un arbre de régression sur la nouvelle base d'apprentissage (x_i, r_{im}) de $i = 1, \dots, n$.

Calcul de h_m en résolvant

$$h_m = argmin_h \sum_i L\big(y_i, F_{m-1}(x_i) + h(x_i)\big).$$

Mise à jour du modèle :

$$F_m(x) = F_{m-1}(x) + h_m.$$

Fin pour

Calcul du résultat final $F_M(x_i) = \sum_{m=1}^{M} h_m(x_i)$.

Le modèle est souvent plus robuste que ceux présentés précédemment, mais reste, encore plus que pour les forêts aléatoires, une boîte noire, par la complexité des arbres créés. Il a aussi tendance à demander de nombreuses itérations et a un temps de calcul conséquent.

4.2.4 XGBoost

XGBoost est un modèle proposé par CHEN et GUESTRIN en 2016. De son nom *Extreme Gradient Boosting*, *XGBoost* est une méthode ensembliste, à l'instar des forêts aléatoires et du *boosting* de gradient. Cet algorithme a pour particularité de pouvoir paralléliser les traitements afin d'optimiser le temps de calcul, et d'ajouter plus de paramètres que le *boosting* de gradient, permettant une meilleure estimation des résultats.

Fonctionnement et limites

Toujours d'après BENTEJAC, C. et al. (2019) cette méthode part du *boosting* de gradient, en ajoutant un terme de régularisation, composé d'une pénalisation L_1et une autre de type L_2, afin de contrôler la complexité du modèle, et éviter le surajustement. Nous écrivons alors une fonction objectif :

$$L = \sum_{i=1}^{N} L\big(y_i, F(x_i)\big) + \sum_{m=1}^{M} \Omega(h_m)\,.$$

Avec $\Omega(h) = \gamma T + \frac{1}{2}\lambda\|\omega\|^2, T$ étant le nombre de nœuds terminaux, et ω la valeur de chaque feuille. Plus γ est élevé, plus les arbres seront simplifiés. Le reste de l'algorithme est similaire à celui du *boosting* de gradient.

XGBoost permet une plus grande flexibilité que les modèles précédents puisqu'il contient de nombreux paramètres afin d'ajuster au mieux les prédictions. Cependant, malgré une optimisation du temps de calcul, le nombre élevé de paramètres peu rapidement poser un problème et augmenter de façon conséquente ce temps de calcul. Son temps de calcul afin d'ajuster les paramètres a été particulièrement long, et ce malgré des bases de données ne dépassant pas 30k lignes. Sa complexité fait aussi de lui une boîte noire, il est donc plus difficile d'expliquer clairement les résultats et de justifier l'application de cette méthode plutôt qu'une autre méthode classique.

4.2.5 Elastic Net

Plusieurs types d'arbres ont été testés. Leur mise en place et leurs différents résultats seront exposés Dans le chapitre 4.3. Cependant, il a été décidé de tester également le modèle *Elastic Net*, afin d'ajouter les résultats d'un modèle linéaire pénalisé, connu pour avoir de bons résultats en termes de régression.

Présentation

D'après ZOU, H. & HASTIE, T. (2005), si nous partons d'un modèle de régression simple, avec des variables X et un bruit ε suivant une loi normale. Nous pouvons écrire :

$$Y = \beta X + \varepsilon.$$

Si β est le vecteur composé des variables à estimer de X. L'estimation de $\hat{\beta}$ doit produire l'erreur résiduelle la plus faible possible. Nous écrivons l'erreur résiduelle telle que :

$$L(\hat{\beta}) = \sum_{i=0}^{n} \left(\left\|y_i - x_i\hat{\beta}\right\|\right)^2 = \left\|Y - X\hat{\beta}\right\|^2.$$

L'estimateur des moindres carrés ordinaires est la solution, et donne :

$$\hat{\beta} = (X^T X)^{-1}(X^T Y).$$

Le biais de $\hat{\beta}$ peut s'écrire $E(\hat{\beta}) - \beta$ et la variance $\sigma^2(X'X)^{-1}$.

Cet estimateur ne fonctionne pas si les $x_{i,j}$ sont corrélés entre eux, donc si X n'est pas de plein rang, et si $p \gg n$. Dans ce cas, il faut régulariser $X'X$ afin de pouvoir l'inverser.

Nous appellons régression de Ridge, la régression qui ajoute une pénalisation de type L_2 afin d'obtenir un compris entre un biais élevé, mais une variance réduite. L_2 est le carré des coefficients, avec une constante λ qui contrôle cette pénalité. Si $\lambda > 0$ nous ajoutons une contrainte au coefficient. La régression de Ridge va s'écrire :

$$L_{ridge} = argmin_{\hat{\beta}} \left(\|Y - \beta X\|^2 + \lambda\|\beta\|_2^2\right).$$

Avec :

$$\|\beta\|_2^2 = \sum_{j=1}^{p} \beta_j^2.$$

Dans ce cas,

$$\hat{\beta} = (X^T X + \lambda I)^{-1}(X^T Y).$$

Son biais devient :

$$-\lambda(X'X + \lambda I)^{-1}\beta.$$

Cette inversion est plus robuste, si les valeurs contenues dans X sont corrélées. Il est important ici de trouver la bonne pénalité λ, différentes méthodes existent, comme la cross validation, la pénalisation (AIC BIC) …

Nous appelons la régression de LASSO, *Least Absolute Shrinkage and Selection Operator* la régression qui ajoute une pénalisation de type L_1. La pénalité L_1 va tronquer les coefficients faibles, et donc les mettre à 0. Elle s'écrit :

$$L_{Lasso} = argmin_{\hat{\beta}} \left(\|Y - \beta X\|^2 + \lambda|\beta|\right).$$

Avec

$$|\beta| = \sum_{j=1}^{p} |\beta_j|.$$

Dans le cas où la prédiction dépend très fortement d'une variable en particulier, la régression de LASSO est biaisée. Dans ce cas, la régression *Elastic Net* est à privilégier.

Fonctionnement et limites

La régression *Elastic Net* combine la régularisation de Ridge et de LASSO. Cette méthode est utile lorsque les données sont corrélées, là où LASSO va utiliser seulement l'une des données, *Elastic Net* va utiliser les deux. La régression de Ridge va aussi apporter de la stabilité, en réduisant la variance. Le biais est aussi mieux géré par cette méthode que pour Ridge et LASSO.

La formule peut s'écrire :

$$L_{ENet} = argmin_{\hat{\beta}} \left(\|Y - \beta X\|^2 + \lambda_1|\beta| + \lambda_2\|\beta\|_2^2\right).$$

Dans notre cas, la régression a lieu sous python, dans ce cas, la formule utilisée est la suivante :

$$L_{ENet} = min_{\hat{\beta}} \frac{1}{2n_{sample}} \|Y - \beta X\|_2^2 + \lambda_1\lambda_2|\beta| + \frac{\lambda_1(1-\lambda_2)}{2}\|\beta\|_2^2.$$

Tout comme le modèle *XGBoost*, le modèle est plus flexible que d'autres modèles linéaires, cependant, cette flexibilité demande également un temps de calcul plus long, qui peut rapidement poser un problème sur les jeux de données bien plus grands.

Maintenant que tous les modèles ont été expliqués, ceux-ci ont été ajustés afin de prédire le plus fidèlement possible les charges ultimes des sinistres de l'exercice de souscription 2020. Par conséquent, des méthodes d'ajustement des paramètres et des métriques de calcul de l'erreur de prédiction ont été utilisées, et vont être détaillées dans le chapitre suivant.

4.2.6 Validation croisée et importance des paramètres

L'objectif premier d'un modèle, n'est pas de s'ajuster parfaitement ligne à ligne à chaque donnée. Il s'agit d'avoir un modèle qui reflète avec précision les régularités observées, et qui se généralise correctement sur un nouveau jeu de données. Il est important avant tout de rappeler ce que sont le biais, la variance, et le dilemme biais/variance.

Dilemme biais variance

Comme il a été dit précédemment, lors de l'application d'un algorithme de *Machine Learning*, il est primordial de séparer la base initiale, en base d'apprentissage, et en base de test. Une fois un modèle calibré sur le jeu de données d'apprentissage, il est testé sur le jeu de données de test, afin de vérifier s'il se généralise correctement. Afin de vérifier la bonne qualité de ce modèle, une erreur est calculée, sur les deux bases.

Comme expliqué par ROUVIERE, L. (2022) si nous considérons une base de données $D = \{(X_1, Y_2), \dots, (X_n, Y_n)\}$, et l'objectif est de prédire Y sachant X, on peut écrire $f_n(x, D_n)$ comme étant un estimateur de la fonction de prévision optimale $f^*(x)$.

- $f_D(x, D_n)$ possède une variance $V[f_n(x, D_n)]$,qui mesure la dispersion des prévisions au point x par rapport à la loi de données D_n. Elle représente également la sensibilité aux variations de l'échantillon, les modèles avec le moins d'hypothèses, vont capter de nombreux modèles sous-jacents, résultant en une forte variance.

- $f_D(x, D_n)$ possède un biais $E[f_n(x, D_n)] - y$, qui mesure l'écart entre la moyenne des prévisions et les valeur réelles y. Il représente aussi l'erreur provenant d'hypothèses erronées dans l'algorithme $f_D(x, D_n)$. Le moins d'hypothèses sont émises à la mise en place du modèle, le plus faible sera le biais.

Le dilemme biais variance est un compromis, permettant d'obtenir la complexité de modèle maximale, pour approcher au mieux la fonction optimale $f^*(x)$.

Nous écrivons habituellement l'erreur de généralisation comme étant l'erreur quadratique moyenne, cependant d'autres fonctions peuvent être utilisées :

$$MSE\big(f_n(x, D_n)\big) = E[(f_n(x, D_n) - y)^2].$$

Si parmi les fonctions de x la solution qui minimise l'erreur de généralisation est $E[x|y]$, il est possible de décomposer cette erreur en biais, variance et irréductible (bruit), nous donnant le compris de biais variance :

$$E[(f_n(x, D_n) - y)^2] = E[y - E[y|x]^2] + \underbrace{E_D[(f(x, D) - E[y|x])^2]}_{\text{Biais}} + \underbrace{E_D[(f(x, D) - E_D[f(x, D)])^2]}_{\text{Variance}}$$

Il est possible d'influer sur le biais et la variance du modèle (Figure 49) :

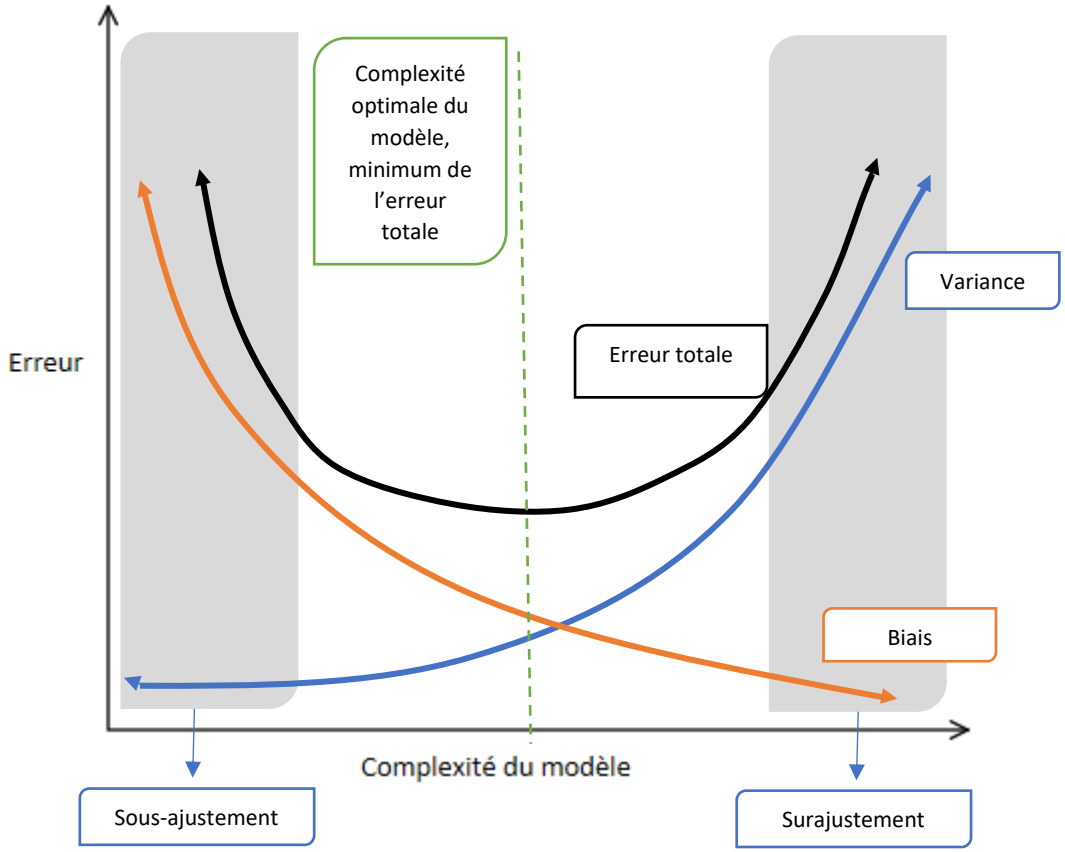


Figure 49-Représentation du dilemme biais variance

Le biais et la variance évoluant dans le sens contraire de l'autre, il est important de trouver le point optimal lors de la mise en place du modèle, celui-ci se trouve au minimum de l'erreur totale.

Afin de mesurer la précision des modèles, les métriques suivantes seront utilisées :

MAE : Le ***Mean Absolute Error***, ou l'erreur absolue moyenne, qui mesure l'amplitude moyenne des écarts entre les valeurs prédites par le modèle, et les observations. Elle est définie par la formule suivante :

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

RMSE : Le ***Root Mean Square Error***, ou la racine de l'erreur quadratique moyenne mesure la différence entre les valeurs prédites et les valeurs observées. Elle est définie par la formule suivante :

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

Grâce aux tests d'erreur sur les bases d'apprentissage et de test, il est possible de déceler un biais ou une variance trop élevée, par exemple (Tableau 20) :

Modèles	Erreur apprentissage (MAE)	Erreur test (MAE)	Résultats
Modèle 1	800	8000	Sur ajustement
Modèle 2	20000	25000	Sous ajustement
Modèle 3	1300	1500	Model optimal

Tableau 20-Exemples de sur ou sous ajustement

Le modèle 1 a une erreur d'apprentissage très faible, mais se généralise difficilement sur l'échantillon de test, avec un MAE de 8000. Il s'agit ici d'une erreur de surajustement. A l'inverse, le modèle 2 a deux erreurs de prédiction élevées, montrant cette fois-ci un problème de sous ajustement. Un modèle optimal comme le modèle 3, donne de bons résultats sur les deux échantillons, apprentissage et test.

Afin d'obtenir le modèle optimal, plusieurs méthodes sont disponibles, l'une des plus courantes étant la validation croisée.

Validation croisée

Lors de l'évaluation d'un modèle, il est possible de séparer la base initiale en trois parties, la base d'apprentissage, de validation et de test. Le modèle est entrainé sur l'apprentissage, validé sur la base de validation, jusqu'à ce que le modèle soit satisfaisant, et enfin testé sur le modèle de test. Cette méthode est fiable, mais demande de nombreuses données dans la base initiale pour pouvoir est scindée en trois parties, ce qui n'est pas toujours le cas. Dans le cas où les données sont peu nombreuses, il est possible d'utiliser la validation croisée. Il s'agit d'une méthode d'évaluation de la performance du modèle, à partir de la base d'apprentissage. Celle-ci est séparée en K parties, aussi appelée *folds.*

Le modèle est entrainé K fois, sur $K-1$ parties utilisées pour l'apprentissage et une permettant la validation du modèle. La méthode est appliquée de façon à ce que chaque partie ayant servi en test, serve une fois en tant que validation (Figure 50) :

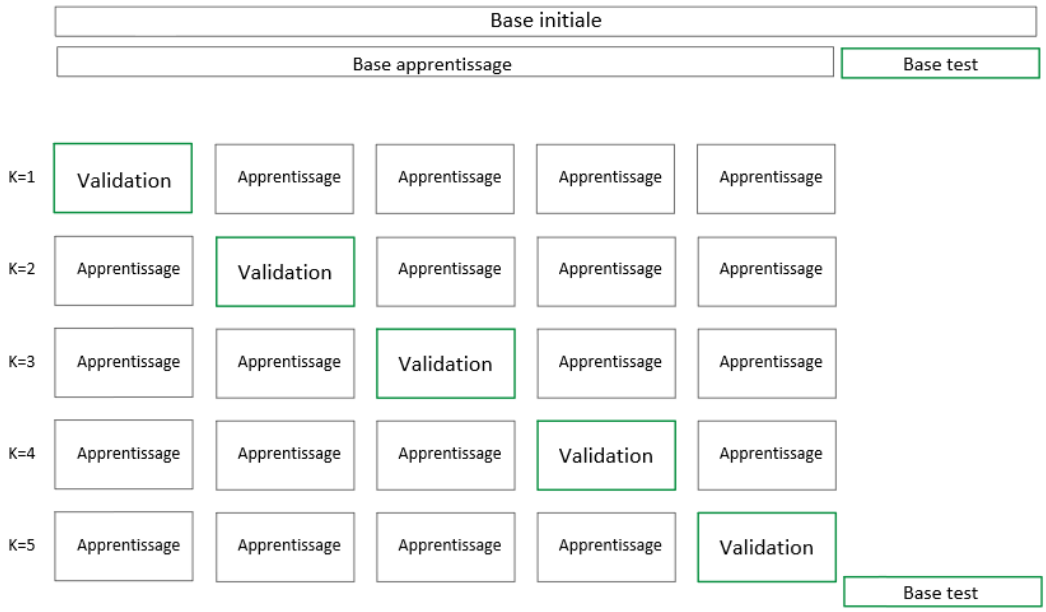


Figure 50-Fonctionnement de la méthode par validation croisée (Source : Librairie Sklearn)

Une erreur de prédiction moyenne est calculée à partir de ces K itérations, chacune d'elle ayant été utilisée pour calculer une erreur de prédiction. Cette étape de validation croisée est utilisée sous python

grâce à la fonction *gridsearchcv,* qui va tester chaque paramètre proposé, appliquer la méthode de validation croisée, et renvoyer les paramètres optimaux selon les résultats des validations croisées.

Importance des paramètres

Selon les modèles appliqués, de nombreux paramètres sont ajustables afin de calibrer le meilleur modèle possible à partir des bases d'apprentissage et de test. Nous allons donc détailler les paramètres qui ont été calibrés sur chacun des modèles.

Pour les arbres de régression, le critère par défaut utilisé pour mesurer la qualité de la séparation du nœud est l'erreur quadratique, cette information n'a pas été modifiée dans les modèles utilisés. D'autres paramètres ont quant à eux été ajustés, comme **le nombre de variables** (*max_features*), qui reprend le nombre n de variables à étudier lors de la création d'un nœud. Dans ce cas, pour un nombre de variables donné n, le modèle prendra aléatoirement n variables à chaque nœud, et séparer le nœud par la variable la plus significative.

La profondeur de l'arbre (*max_depth*) est également très importante, celle-ci représente le nombre de nœuds maximum de l'arbre. Si les nœuds sont trop peu nombreux, il y a sous-ajustement du modèle, ou s'ils sont trop nombreux, il y a sur ajustement, et dans ces deux cas, comme vu précédemment, le modèle ne pourra pas se généraliser correctement et être performant sur un nouveau jeu de données.

Les feuilles d'un arbre de décision ou nœuds terminaux, sont définis par un **nombre d'échantillons minimum** (*min_sample_leaf*) nécessaires afin de considérer un nœud comme un nœud terminal. Ce paramètre est ajustable, et par conséquent, si nous posons *e* le nombre d'échantillons minimum, alors un nœud sera considéré comme étant une feuille, lorsque le nombre d'observations dans ce nœud est de *e*.

De la même manière, le nombre minimum d'échantillon nécessaire pour considérer qu'un nœud peut être séparé en d'autres nœuds intermédiaires est ajustable, avec le paramètre *min_sample_node*.

Lors de l'élaboration du modèle des forêts aléatoires et du *boosting* de gradient, les paramètres précédents sont disponibles, viennent s'ajouter **le nombre d'estimateurs** (*n_estimators*), représentant le nombre d'arbre de régression à construire avant de prendre la moyenne des prédictions de chacun de ces arbres.

De la même manière, le modèle *XGBoost* vient ajouter de nouveaux paramètres comme *subsample* représentant la **part (en %) d'échantillon sélectionné** pour chaque arbre créé.

Pour terminer, **le *Learning_rate*** (ou *eta*) permet d'obtenir un modèle plus robuste en réduisant le poids sur chaque étape du calcul. Les **pénalisations de Ridge L_2** évitant le surajustement et **LASSO L_1** utilisée sur les données à forte dimensionnalité afin d'augmenter la vitesse d'implémentation, sont aussi paramétrables.

Les paramètres et leur ajustement sont une étape fondamentale afin d'obtenir les résultats les plus fiables. Une fois chacun des modèles ajustés par validation croisée, les résultats entre les différents modèles peuvent être comparés, et les meilleurs prédicteurs peuvent être comparés à la méthode *Chain Ladder.* Les résultats sur les deux bases de données vont être détaillés dans le sec suivant.

4.3 RESULTATS AVIATION ET MARITIME SUR L'EXERCICE DE SOUSCRIPTION 2020

Dans cette dernière partie, les résultats de l'aviation et du maritime seront comparés à leur méthode *Chain Ladder* respective, tous les modèles seront également comparés entre eux, grâce à leurs erreurs de prédiction afin d'obtenir les meilleures prédictions possibles sur l'exercice de souscription 2020. Par la suite, nous verrons les raisons pour lesquelles les résultats de l'aviation sont moins concluants que ceux du maritime.

4.3.1 Aviation

Les données aviation ont été prédites sur l'exercice 2020, en tenant compte des données des exercices 2013 à 2019. Un poids calculé par la méthode IPCW a été attribué à chaque observation non censurée, les autres, ayant un poids nul. Chaque modèle va prédire un résultat global attribué à l'exercice de souscription 2020, ces modèles seront comparés entre eux, grâce aux métriques présentées dans le chapitre précédent, *Mean Absolute Error*, et le *Root Mean Square Error*. Le même échantillonnage a été effectué sur la base initiale, à savoir une séparation en 20% test et 80% apprentissage. Aucune modification d'échantillon ne peut être à l'origine des différences de résultats d'un modèle à l'autre.

Importance des variables dans les modèles

Chacun des modèles implémentés permet d'identifier les variables les plus influentes, sous Python, la fonction *feature importance* permet de visualiser l'importance des variables explicatives sur la variable cible du modèle. Il s'agit de calculer un score pour chaque variable, plus ce score est élevé, plus l'impact de la variable est important dans le modèle. L'important ici est de comprendre les relations présentes dans les bases, de rapidement détecter les données les plus pertinentes, et celles ayant un impact presque nul sur la variable cible. Ceci peut aider à simplifier les bases de données, lors de la phase d'apprentissage, pour augmenter la rapidité d'exécution du modèle, ou encore de mieux communiquer sur les données, et leurs impacts les unes sur les autres. Il est important de noter que les résultats obtenus par la fonction *feature importance* ne sont pas semblables suivant les modèles utilisés.

Comme expliqué par GHATTAS B. (2000), l'importance d'une variable X^m d'un arbre A est calculée par :

$$I(X^m) = \sum_{t \in A} \Delta\hat{R}\left(\widehat{d_m}(t), t\right).$$

Avec $\widehat{d_m}(t)$ la division de substitution au nœud t sur la variable m. Il s'agit de la somme des diminutions de la déviance provoquée à chaque nœud t de l'arbre, si nous remplaçons pour chaque nœud la division optimale $d_m(t)$ par $\widehat{d_m}(t)$ sur la variable X^m.

Selon la méthode utilisée, le calcul est modifié, par exemple pour la méthode des forêts aléatoires, chaque sous arbre créé possède son échantillon non utilisé (Out Of Bag ou OOB). Cet échantillon est utilisé afin de calculer l'importance d'une variable. Les valeurs de l'échantillon OOB de cette variable sont aléatoirement mélangées, les autres variables restent fixes. La diminution de la fiabilité de la prédiction sur ces données mélangées est mesurée. Cette méthode est appliquée sur chacun des sous arbres, enfin, la valeur retournée est la moyenne de toutes ces mesures.

Pour la méthode *XGBoost*, celle-ci mesure l'importance de la variable sur chacun des sous arbres, comme étant l'amélioration estimée à chaque nœud dont X_m a servi de séparateur. La valeur donnée par la fonction est la moyenne de tous les résultats, sur chacun de ces arbres.

Si nous comparons les variables d'un arbre de régression simple et celles de la méthode *XGBoost* (Figures 51 et 52) :

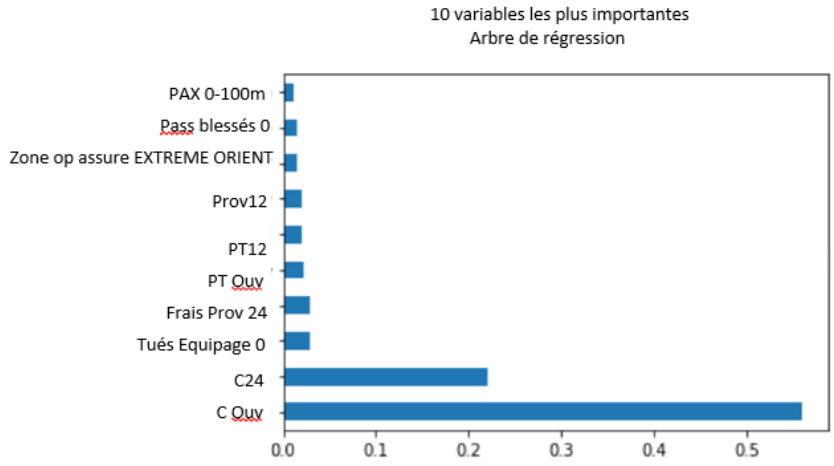


Figure 51-Variables les plus influentes de la méthode CART

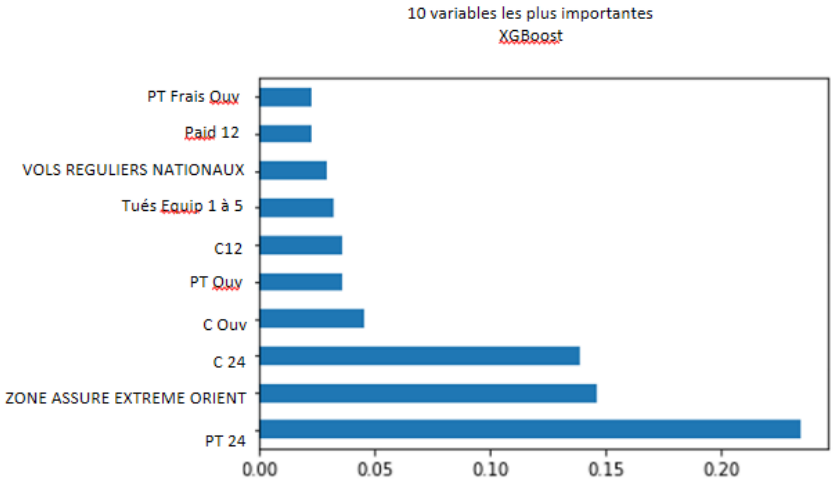


Figure 52-Variables les plus influentes de la méthode XGBoost

Certaines variables considérées comme importantes par l'arbre de régression, ne figurent plus parmi les 10 plus importantes de la méthode *XGBoost*. Nous remarquons bien ici la différence de traitement des informations fournies par la base d'apprentissage. Cependant, les variables quantitatives, comme la charge à l'ouverture ou les paiements au bout de 12 et 24 mois sont toujours présentes dans la liste. Par conséquent, quelques soit le modèle, les charges, paiements, et provisions des années précédentes ont un fort impact sur les modèles.

Résultats

Afin d'aisément comparer les modèles, seuls les sinistres non censurés sont pris en compte lors du calcul des erreurs de prédiction. Les premiers résultats sont disponibles ci-dessous (Tableau 21) :

Méthodes	CU réelle vs CU prédite (sinistres clos à fin 2021)
Arbre de régression	30,22%
Forêts aléatoires	9,44%
Boosting de gradient	-0,10%
XGBoost	-9,39%
ElasticNet	16,25%

Tableau 21-Différence entre charge ultime prédite et charge ultime réelle, sur les sinistres clos pour chaque méthode

Nous pouvons remarquer que certaines méthodes, comme le *boosting* de gradient ainsi que les forêts aléatoires, ou *XGBoost*, sont proches de la charge globale des sinistres clos (9.44% et -0.10%). Cette information n'est cependant pas suffisante, afin de choisir correctement le modèle à utiliser, c'est pourquoi les résultats des erreurs de prédiction seront considérés comme plus pertinents afin de pouvoir comparer correctement les modèles. Le tableau complété ci-dessous nous permet d'apprécier la différence entre la variation charge réelle charge prédite, et l'erreur de prédiction obtenue (Tableau 22) :

Méthodes	MAE	RMSE	CU réelle vs CU prédite (sinistres clos à fin 2021)
Arbre de régression	69 346	144 970	30,22%
Forêts aléatoires	30 796	42 821	9,44%
Boosting de gradient	58 798	145 188	-0,10%
XGBoost	18 730	33 420	-9,39%
ElasticNet	12 292	17 882	16,25%

Tableau 22-Erreurs de prédiction pour chaque méthode

Nous pouvons avant tout noter que les erreurs sont élevées, cependant les méthodes *Elastic Net* et *XGBoost* restent les plus concluantes, malgré une différence de charge ultime réelle et charge ultime prédite élevée. Nous pouvons également noter que par la simplicité de son algorithme, l'arbre de régression donne une erreur de prédiction, et une variation de résultat, bien plus forte que les autres. Nous en déduisons tout de même que la méthode *XGBoost*, donne le meilleur résultat parmi les méthodes de partitionnement récursif.

Pour terminer, si nous comparons les résultats du *Chain Ladder* et ceux obtenus par les cinq méthodes (Figure 53) :

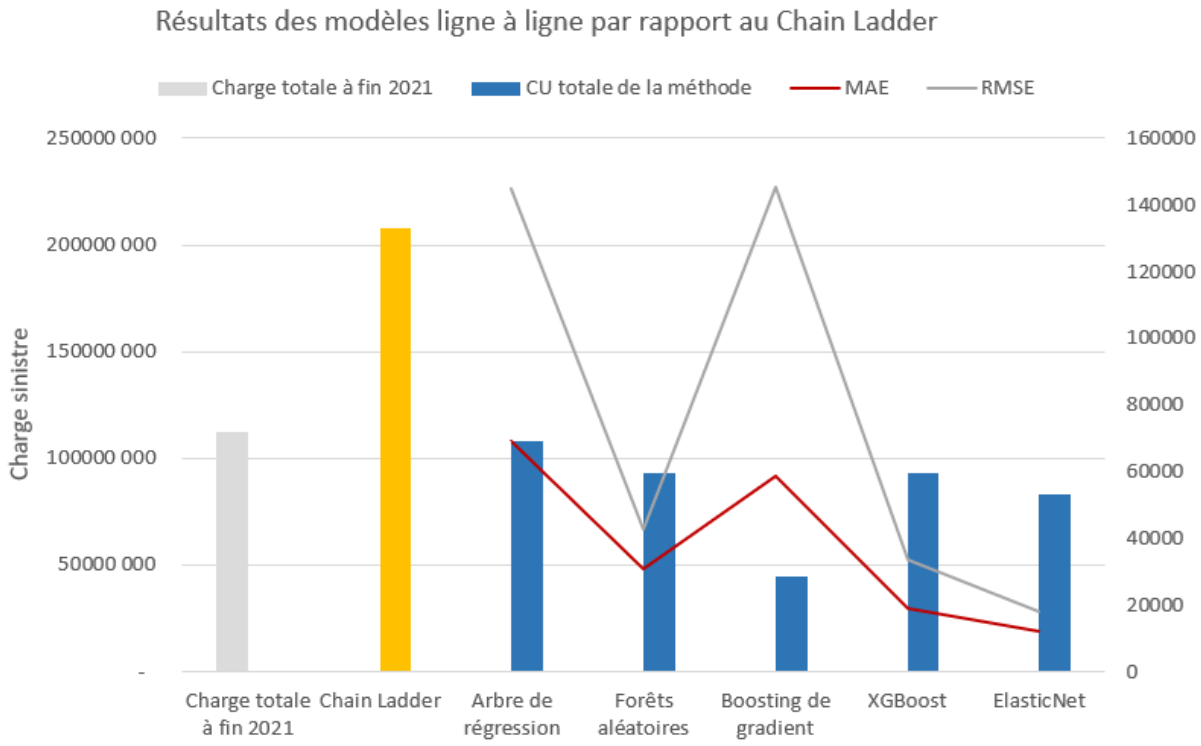


Figure 53- Résultats de la Charge Ultime, MAE et RMSE pour chaque modèle, par rapport au Chain Ladder (aviation).

Nous pouvons noter une très forte différence, la charge ultime des modèles étant bien inférieure à celle du *Chain Ladder*. Les tardifs, représentant seulement en moyenne 3% de la charge ultime, ne permettent pas d'expliquer cet écart. Il est possible que le peu de données, avec seulement 13 502 sinistres, dont 8 101 non censurés, soit l'une des raisons de la mauvaise prédiction. A cela vient s'ajouter la sous-représentation des sinistres à charge ultime très élevée, pour lesquels l'application des poids IPCW n'a pas pu compenser ce manque d'informations. De ce fait, les modèles auront tendance à estimer correctement les sinistres à charge ultime faible voire moyenne, et largement sous-estimer les sinistres à charge ultime élevée.

Le graphique ci-dessus détaille les variations par tranches de charge ultime de la méthode *XGBoost* (Figure 54) :

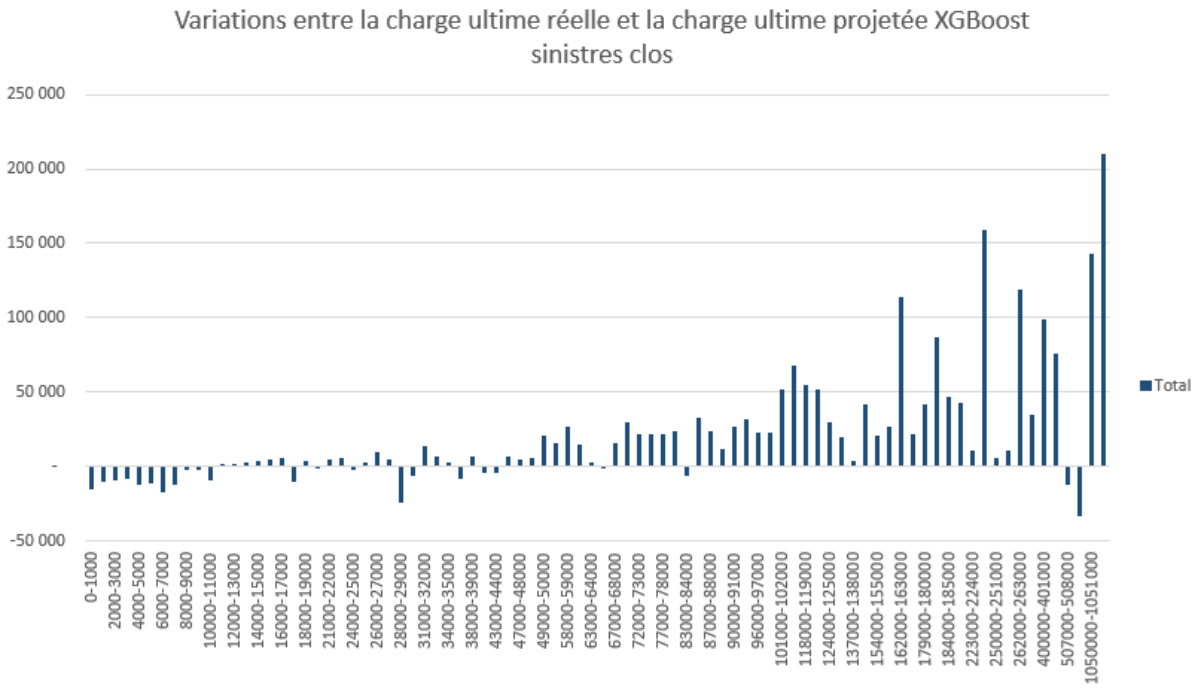


Figure 54-Variations entre charge réelle et charge prédite, sur les sinistres clos, méthode XGBoost

En moyenne, les sinistres supérieurs à 22k$ sont sous-estimés, ceux inférieurs à ce seuil ont des variations proches de 0, et sont donc correctement projetés.

Aujourd'hui, au 30/11/2022, soit un an après l'estimation obtenue par la méthode du *Chain Ladder*, la charge totale des sinistres de l'exercice 2020 s'élève à 150,7M$. Cette charge sinistre est déjà bien supérieure aux prédictions obtenues par les modèles de *Machine Learning*, et semble s'approcher des résultats de la méthode du *Chain Ladder*, confirmant l'estimation qui avait été faite l'année dernière, ainsi que le faible niveau de prédiction des modèles sur ce portefeuille.

4.3.2 Maritime

De la même manière que pour l'aviation, les sinistres de l'exercice de souscription 2020 non censurés sont observés, en termes de variation de montant entre la charge ultime réelle et la charge ultime prédite, ainsi qu'en terme d'erreurs de prédiction (Tableau 23) :

Données en €

Méthodes	MAE	RMSE	CU réelle vs CU prédite (sinistres clos à fin 2021)
Arbre de régression	10 120	44 430	1,95%
Forêts aléatoires	7 692	32 276	2,09%
Boosting de gradient	6 961	27 066	0,87%
XGBoost	9 077	24 917	1,91%
ElasticNet	6 957	25 429	1,56%

Tableau 23- Différence entre charge ultime prédite et charge ultime réelle, sur les sinistres clos pour chaque méthode

Les méthodes *Elastic Net* et *boosting* de gradient se retrouvent avec les erreurs de prédiction les plus faibles, et une variation entre charge ultime réelle et charge projetée inférieure à 2%. Si nous ajoutons à ces données la charge ultime de la méthode *Chain Ladder*, comparée à la charge ultime des méthodes étudiée, avec ajout d'un pourcentage de 2% représentant la portion de tardifs attendus, nous obtenons les résultats suivants (Figure 55) :

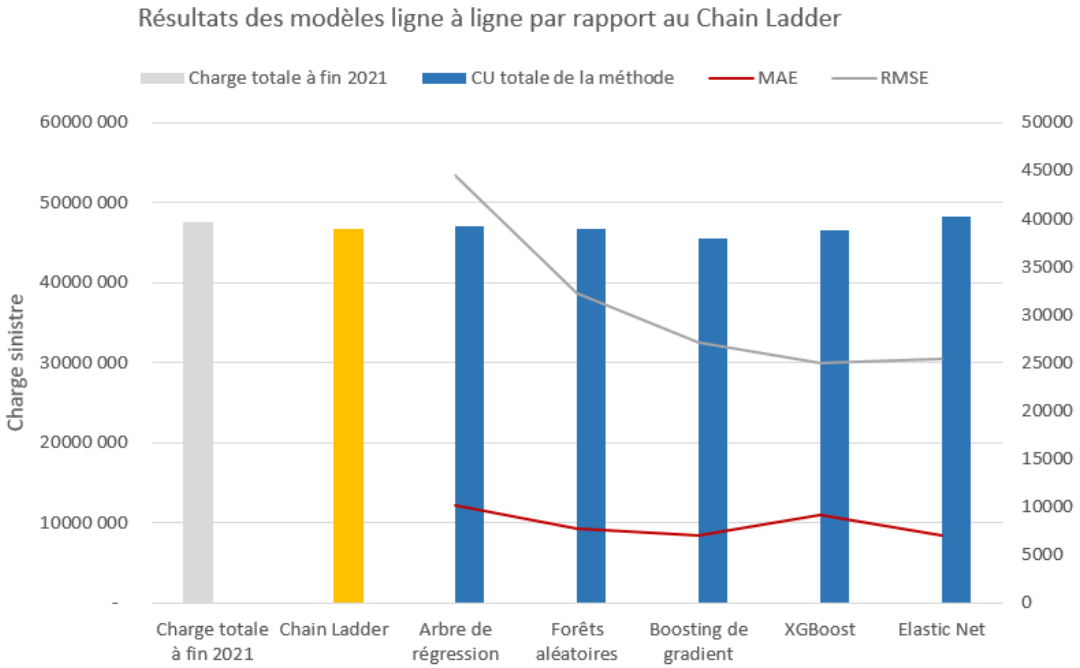


Figure 55-Résultats de la Charge Ultime, MAE et RMSE pour chaque modèle, par rapport au Chain Ladder (maritime).

Les charges ultimes des modèles avec les erreurs de prédiction les plus faibles, *Elastic Net* et *boosting* de gradient, sont différentes de la charge ultime *Chain Ladder*, de respectivement 3.27% et -2.64%. Les méthodes de prédiction par répartition récursives donnent dans ce cas des résultats aussi satisfaisants que la méthode *Elastic Net*.

Les 10 variables les plus importantes d'après le modèle du *boosting* de gradient sont les suivantes (Figure 56) :

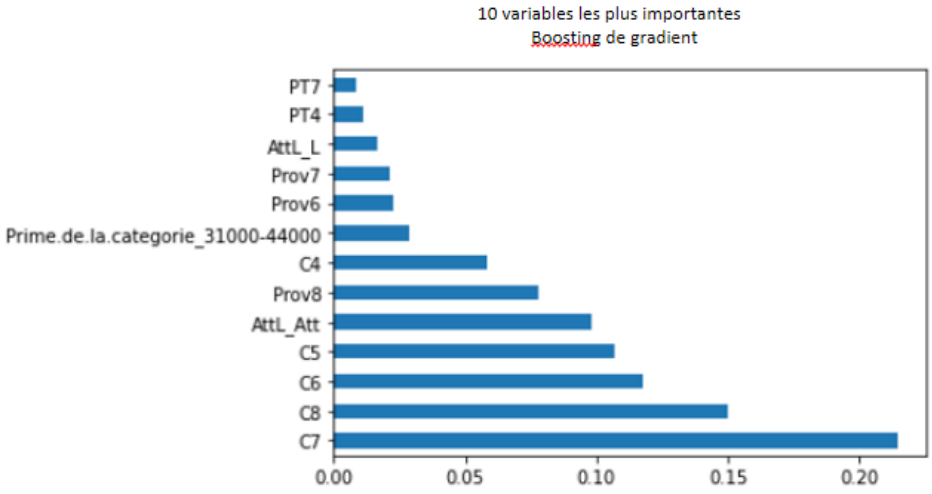


Figure 56- Variables les plus influentes de la méthode Boosting de gradient.

Tout comme l'aviation, les variables les plus importantes du modèle sont principalement les montants de sinistres aux dates antérieures, comme ici les charges trimestrielles, (C4 à C8), ainsi que les provisions trimestrielles (PROV6 à PROV8). Les types de sinistres à la dernière vue, soit au bout de deux années de développement, comme étant attritionnels ou larges, (variable « attL_Att » et « AttL_L ») sont également importants pour l'obtention de la charge ultime projetée.

Si nous observons les sinistres clos regroupés par tranches dans la figure ci-dessous (Figure 57) :

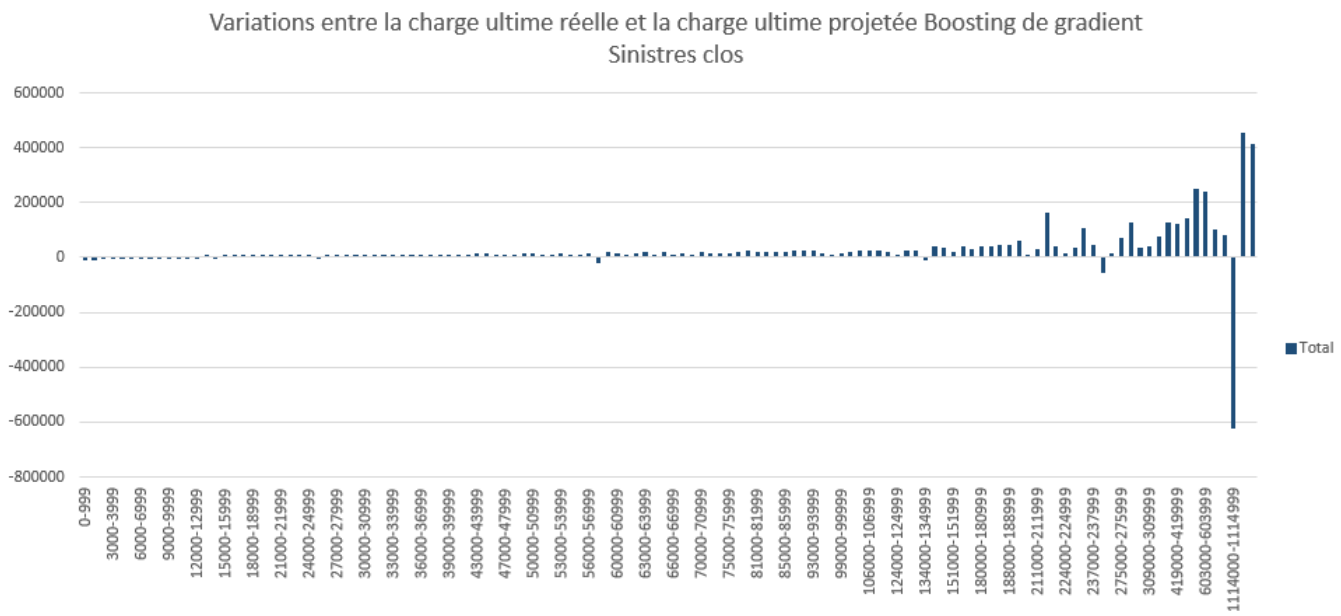


Figure 57-Variations entre charge réelle et charge prédite, sur les sinistres clos, méthode XGBoost

Nous remarquons une sous-estimation modérée des sinistres supérieurs à 275k$, ainsi qu'une variation d'estimation très proche de zéro pour les sinistres inférieurs à ce seuil. Un seul sinistre supérieur à 275k$ montre une charge ultime projetée bien supérieure à la charge réelle, avec une variation de 600k$.

Le nombre de sinistres observés étant bien plus élevé comparé à l'aviation, et le taux de censure bien plus bas, la base d'apprentissage utilisée pour l'entrainement des modèles compte 22 051 sinistres. Les modèles étudiés prédisent tous avec précision la charge ultime de l'exercice de souscription 2020, si nous les comparons à la charge ultime maritime. Cependant, il est important de noter que même si ces modèles s'appliquent correctement à l'année 2020, il n'est pas certain que ces mêmes modèles s'appliquent aussi bien à d'autres exercices. Comme le montre la figure 57, les sinistres très élevés sont plus difficilement estimés, un exercice de souscription avec plus de sinistres larges pourrait se retrouver avec une charge ultime trop faible, par rapport aux résultats de la méthode du *Chain Ladder*.

En conclusion de cette dernière partie, nous pouvons noter d'une part, que les prédictions obtenues sur les données aviation sont très éloignées de l'estimation obtenue par la méthode du Chain Ladder. D'autre part, les prédictions obtenues sur les données maritimes sont bien plus cohérentes par rapport aux résultats du Chain Ladder.

Le type de données étudiées impacte fortement les résultats obtenus par les méthodes de Machine Learning. L'utilisation de ces méthodes sur un portefeuille semblable à celui de l'aviation étudié, avec les modèles testés ici, semble encore difficile aujourd'hui.

Conclusion

Le but de ce mémoire était d'établir s'il était possible d'obtenir une charge ultime fiable, sur des portefeuilles de niche, grâce à des méthodes de *Machine Learning*.

Il est donc possible d'obtenir une charge ultime par provisionnement ligne à ligne, grâce aux méthodes de *Machine Learning* sur ces portefeuilles. Cependant, il y a une différence notable en termes de prédiction des résultats entre l'aviation, et le maritime.

Dans le cas de l'aviation, par son niveau de censure très élevé, les modèles prédisent très difficilement les sinistres extrêmes, souvent censurés et moins nombreux, cependant très importants lors de l'évaluation de la charge ultime, et donc du *Best Estimate*. L'évolution de la sinistralité attritionnelle observée dans la méthode *Chain Ladder* permet de mettre en lumière une évolution de la sinistralité de la deuxième à la troisième année de développement très élevée. Il est par conséquent très difficile de prédire l'évolution de la sinistralité, à partir de la deuxième année de développement. L'utilisation des poids IPCW n'a pas pu rééquilibrer la base, les sinistres censurés et extrêmes n'étant toujours pas assez fréquents lors de l'application des modèles de *Machine Learning*. Le nombre élevé de variables disponibles sur les assurés et les sinistres n'ont également pas pu dégager de tendances fortes impactant les charges ultimes.

Il serait envisageable pour la suite, de passer par la synthétisation de données par des méthodes d'intelligences artificielles. Il s'agit dans ce cas, de créer des données qui restent cohérentes avec les données réelles, qui viennent s'ajouter aux véritables données. Le nombre de sinistres observés pourrait donc être fixé, afin d'obtenir un nombre bien plus conséquent, et les sinistres extrêmes seraient représentés en nombre suffisant, et donc bien mieux prédits. Tout ceci a néanmoins un cout de production élevé.

Dans le cas du maritime, les résultats sont plus satisfaisants, les charges ultimes prédites par les modèles de *Machine Learning* étant toutes très proches de la méthode du *Chain Ladder* (entre 0.5% et 3.2% de différence). Ceci s'explique par le nombre de sinistres plus conséquent qu'en aviation, ainsi qu'une volatilité moins forte. Il faut tout de même nuancer les résultats, les modèles entraînés sur les données maritime fonctionnent sur l'exercice de souscription 2020, ceux-ci ne fonctionneront peut-être plus sur les années suivantes. Si la volatilité augmente trop, les exercices précédents ne représenteront plus aussi bien les années à prédire, et le modèle s'écartera petit à petit de la réalité. Il est donc primordial d'entrainer régulièrement les modèles sur de nouvelles données, afin de tenir compte de toute modification du portefeuille au cours des années.

Bibliographie

ACPR (2020 et 2019), *Chiffres du marché français de la banque et de l'assurance.*

Allianz, *quatre questions au Directeur Mondial Indemnisation Marine, régis Broudin*.

Aviassur, *Historique de l'assurance aviation*.

BARIGOU, F. (2013), Contribution à la catégorisation de textes et à l'extraction d'information, Ph. D. Thesis, Université d'Oran.

BARRA, V., CORNUEJOLS, A. & MICLET, L. (2021), *Apprentissage artificiel : concepts et algorithmes de Bayes et Hume au deep learning* 4ème édition, Eyrolles.

BENTEJAC, C., CSÖRGO, A. & MARTINEZ-MUÑOZ, G. (2019), *A comparative Analysis of XGBoost,* document de travail, Université de Madrid, Espagne.

BOISADAM, S. (2021), *Apport des méthodes de Machine Learning sur la modélisation des taux de Résiliation en Assurance Emprunteur*, Mémoire d'actuariat, Université de Strasbourg.

BOYER, M. (2008), *Une brève histoire des assurances au Moyen âge*, Assurances et gestion des risques, vol.76(3), p. 83-97.

BRADY, N. (2019), *On the Bias-Variance Tradeoff: Textbooks Need an Update,* Mémoire en informatique, Université de Montréal.

BREIMAN, L., FRIEDMAN, J., OLSHEN, R. & STONE, C., (1984), *Classification and regression trees,* Chapman & Hall/CRC.

CARLOT, J. F. (2021), *Droit des assurances 2021-22,* Hachette.

CNAM UE RCP209, *Cours arbres de décision.*
Lien :https://cedric.cnam.fr/vertigo/cours/ml2/coursArbresDecision.html#id4

CNAM UE RCP208, *Cours données manquantes.*
Lien : https://cedric.cnam.fr/vertigo/Cours/ml/coursDonneesManquantes.html

CNAM UE RCP209, *Cours forêts aléatoires.*
Lien :https://cedric.cnam.fr/vertigo/cours/ml2/coursForetsAleatoires.html

DENUIT, M. & CHARPENTIER, A. (2005), *Mathématiques de l'assurance non-vie* vol.2, Economica.

DREYFUSS, M. L. (2015), *Les grand principes de solvabilité 2* 3ième édition, L'argus de l'assurance Editions.

DREYFUSS, M. L. (2020), *La révolution digitale dans l'assurance* 2ième édition, L'argus de l'assurance Editions.

DUBOIS, D., FEDERLE, A. & RANAIVOZANANY, V. (2021), *Appliquer la data science à l'assurance,* L'argus de l'assurance Editions.

DUVAL, F., PIGEON, M. (2019), *Individual Loss Reserving Using a Gradient Boosting-Based Approach*, Risks.

GADAT, S., *Cours : Introduction à l'apprentissage Statistique*, Laboratoire de Statistique de Probabilités UMR 5583 CNRS-UPS université de Toulouse.

GERDS, T. A. (2022), Prediction Error Curves for Risk Prediction Models in Survival Analysis, Package 'pec',

GERON, A. (2019), *Machine Learning avec Scikit-Learn : Mise en œuvre et cas concrets* 2ième édition, Dunod.

GHATTAS, B. (2000), *Importance des variables dans les méthodes CART,* la revue MODULAD, p29-39.

GOUDE, Y., *Cours : Méthodes de Régression Avancées*, Paris-Saclay.

GOUNE, R. (2016), *Calibrage de loi de rachat structurel sur un portefeuille d'épargne : méthode d'apprentissage et approche actuarielle*, Mémoire d'actuaire, ENSAE.

GRAFFEO, N. (2014), *Méthode d'analyse de la survie nette : utilisation des tables de mortalité, test de comparaison et détection d'agrégats spatiaux*, Ph. D. thesis, Aix-Marseille université.

HE, L. (2004), *Méthode de provisionnement et analyse de la solvabilité d'une entreprise d'assurance non-vie*, Mémoire d'actuaire, ENSAE.

JAMAL, S. (2018), *Machine Learning & traditional Methods Synergy in Non-life reserving*, ASTIN non life insurance working party.

LANGEVIN, N. (2019), *Modélisation de la sinistralité tempête, apport de l'Open Data et du Machine Learning*, Mémoire d'actuariat, ENSAE.

LE FAOU, Y. (2019), Contribution à la modélisation des données de durée en présence de censure : application à l'étude des résiliations de contrats d'assurance santé, Ph. D. thesis, Sorbonne université, Paris.

LOPEZ, O., MILHAUD, X. & THEROND P. E. (2016), *Tree-based censored regression with application in insurance*, Electronic Journal of Statistics Vol.10, p.2685-2716.

MARLY, P.G. (2013), *Droit des assurances, DALLOZ.*

MORANI, G. (1999), *Sélection de modèles non linéaires par « leave-one-out » : étude théorique et application des réseaux neurones au procédé de soudage par points*, Ph. D. thesis, Université Pierre et Marie curie.

MÜLLER, A. & GUIDO, S. (2018), *Le Machine Learning avec Python*, Edition First.

NICHIL, G. (2014), *Provisionnement en assurance non-vie pour des contrats à maturité longue et à prime unique : application à la réforme Solvabilité 2*, Ph. D. thesis, Université de Lorraine.

PARISMAT, *Chiffres significatifs en 2020.*

PATRAT, C., LECOEUR, E., NESSI, J. M., NISIPASU, E. & REIZ, O. (2007), *Provisionnement technique en assurance non-vie, perspectives actuarielles modernes*, Economica.

PEDREGOSA et al. (2011), Scikit-learn: Machine Learning in Python

PHAN NGOC, H. (2015), *Provisionnement stochastique adapté aux spécialités de la réassurance,* Mémoire d'actuariat, ISFA.

RAITI, Z. (2017), *Modélisation de la durée de maintien en arrêt de travail d'une population de travailleurs non-salariés,* Mémoire d'actuariat ISUP.

ROKACH, L. & MAIMON, O. (2015), *Data mining with decision trees Theory and applications* 2nd edition, World Scientific.

ROUVIERE, L. (2022), *Cours d'apprentissage supervisé - Machine Learning*, Université Rennes II.

ROUVIERE, L., *Cours d'introduction aux méthodes d'agrégation : boosting, bagging, et forêts aléatoires. Illustrations avec R*, Université Rennes II.

SAINT PIERRE, P. (2021), *Cours d'introduction à l'analyse des durées de survie*, Université Paul Sabatier – Toulouse III.

SAUPIN, G. (2022), *Gradient boosting : Exploitez les arbres de décision pour le Machine Learning*, Epsilon.

TRAORE, K. & VERMET, F., *Méthodes de Chain Ladder et Mack*

WADE, C. (2020), *Hands-on Gradient Boosting with XGBoost and Scikit-learn*, PackT.

WikiStat, Université de Toulouse.
Lien : https://www.math.univ-toulouse.fr/~besse/Wikistat/pdf/st-m-app-linSelect.pdf

WikiStat, Université de Toulouse.
Lien : https://www.math.univ-toulouse.fr/~besse/Wikistat/pdf/st-m-app-idm.pdf

Zang, Z. Mayer,G. , Dauvilliers, Y. & Plazzi, G. (2018), *Exploring the clinical features of narcolepsy type 1 versus narcolepsy type 2 from European Narcolepsy Network database with machine learning,* scientific reports

ZOU, H. & HASTIE, T. (2005), *Regularization and variable selection via the Elastic Net*, J. R. Statist. Soc. B, part 2, p. 301-320.

A. Annexes

A.1. HYPOTHESES *CHAIN LADDER*

A.1.1. Maritime

	j	L	S	Z	t	m	t-1	Cm t-1	E	V
	2	2	0	0	2	0	1	1	0,50	0,25
	3	1	2	1	3	1	2	2	0,75	0,19
	4	0	2	0	2	0	1	1	0,50	0,25
	5	1	2	1	3	1	2	2	0,75	0,19
	6	1	1	1	2	0	1	1	0,50	0,25
	7	4	0	0	4	1	3	3	1,25	0,44
	8	1	1	1	2	0	1	1	0,50	0,25
	9	1	2	1	3	1	2	2	0,75	0,19
	10	1	1	1	2	0	1	1	0,50	0,25
	11	1	1	1	2	0	1	1	0,50	0,25
	12	10	0	0	10	4	9	126	3,77	0,99
Total			**Z**	**7**					**10,27**	**3,49**

IC	6,535393248	14,003669

Z appartient à l'intervalle

Tableau 24-Résultats de l'hypothèse H_0 triangle IAR

	j	L	S	Z	t	m	t-1	Cm t-1	E	V
	2	0	1	0	1	0	0	1	0,00	0,00
	3	1	2	1	3	1	2	2	0,75	0,19
	4	4	0	0	4	1	3	3	1,25	0,44
	5	2	3	2	5	2	4	6	1,56	0,37
	6	3	2	2	5	2	4	6	1,56	0,37
	7	4	2	2	6	2	5	10	2,06	0,62
	8	1	4	1	5	2	4	6	1,56	0,37
	9	2	2	2	4	1	3	3	1,25	0,44
	10	3	3	3	6	2	5	10	2,06	0,62
	11	3	3	3	6	2	5	10	2,06	0,62
	12	10	0	0	10	4	9	126	3,77	0,99
Total			**Z**	**16**					**17,89**	**5,03**

IC	13,41122481	22,377838

Z appartient à l'intervalle

Tableau 25-Résultats de l'hypothèse H_0 triangle CMM

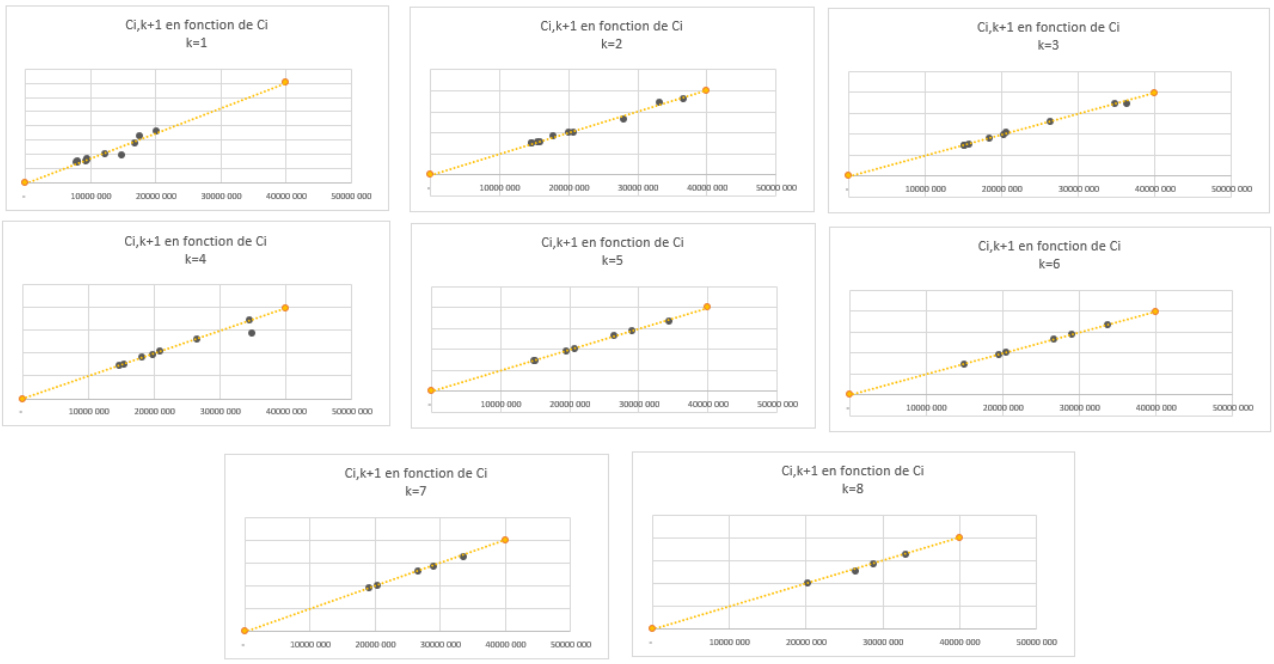


Figure 58-Résultats de l'hypothèse H_1 sur le triangle PAC

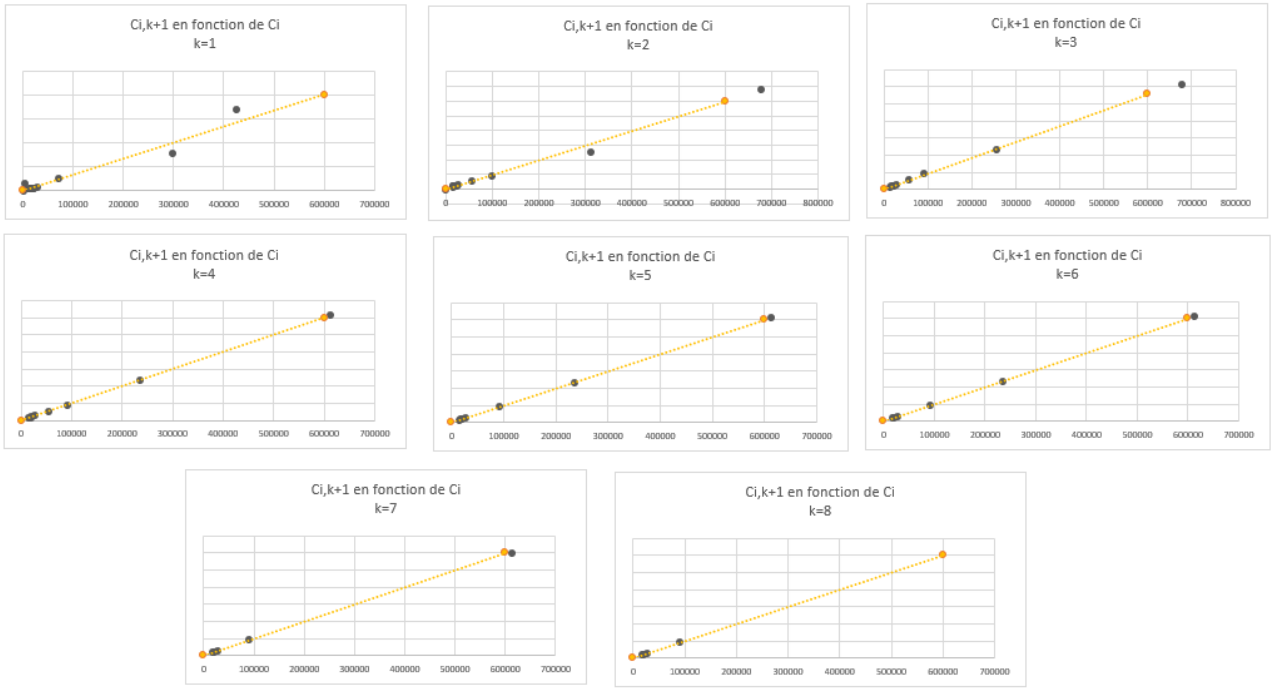


Figure 59- Résultats de l'hypothèse H_1 sur le triangle IAR

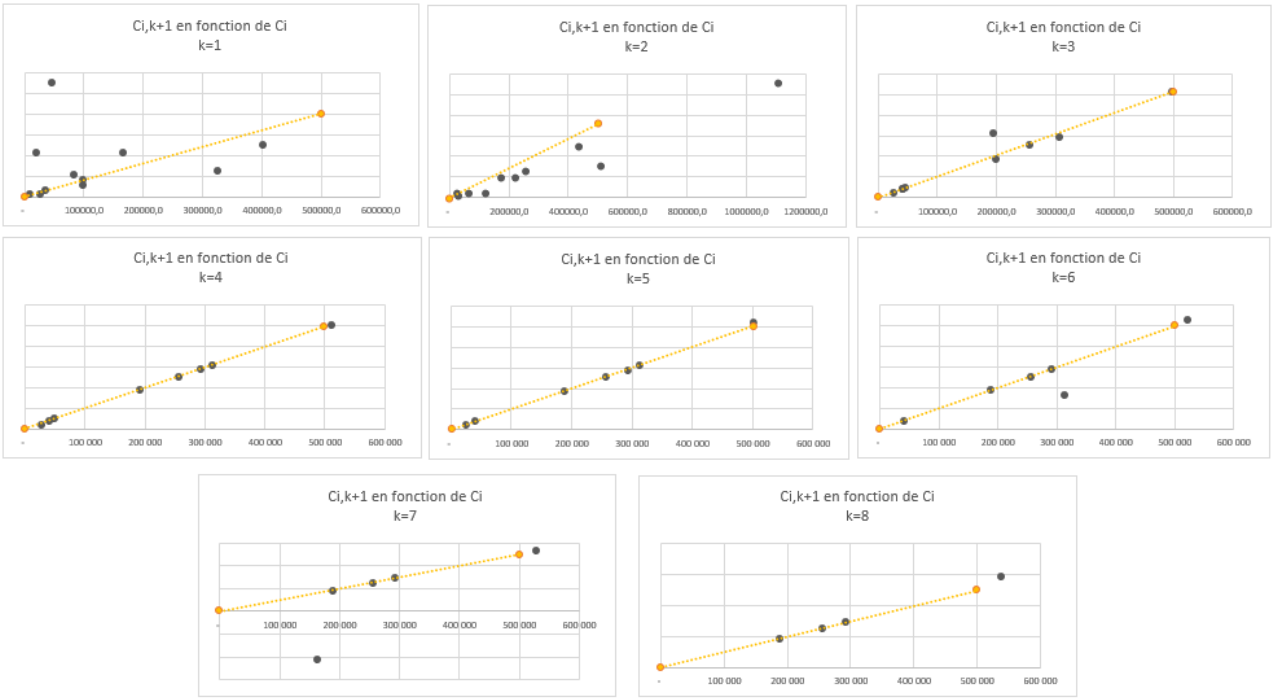


Figure 60- Résultats de l'hypothèse H_1 sur le triangle CMM

A.1.2. Aviation

	j	L	S	Z	t	m	t-1	Cm t-1	E	V
	2	1	1	1	2	0	1	1	0,50	0,25
	3	2	1	1	3	1	2	2	0,75	0,19
	4	2	2	2	4	1	3	3	1,25	0,44
	5	3	0	0	3	1	2	2	0,75	0,19
	6	3	2	2	5	2	4	6	1,56	0,37
	7	2	4	2	6	2	5	10	2,06	0,62
	8	4	4	4	8	3	7	35	2,91	0,80
	9	3	4	3	7	3	6	20	2,41	0,55
	10	1	5	1	6	2	5	10	2,06	0,62
	11	7	3	3	10	4	9	126	3,77	0,99
	12	5	5	5	10	4	9	126	3,77	0,99
	13	4	5	4	9	4	8	70	3,27	0,74
	14	7	3	3	10	4	9	126	3,77	0,99
	15	3	9	3	12	5	11	462	4,65	1,17
	16	6	6	6	12	5	11	462	4,65	1,17
Total				**Z** **40**					**38,12**	**10,06**

IC	31,776653	44,465534

Z appartient à l'intervalle

Tableau 26-Résultats de l'hypothèse H_0 triangle Corps

	j	L	S	Z	t	m	t-1	Cm t-1	E	V
	2	1	1	1	2	0	1	1	0,50	0,25
	3	1	2	1	3	1	2	2	0,75	0,19
	4	3	1	1	4	1	3	3	1,25	0,44
	5	4	1	1	5	2	4	6	1,56	0,37
	6	5	1	1	6	2	5	10	2,06	0,62
	7	4	0	0	4	1	3	3	1,25	0,44
	8	2	2	2	4	1	3	3	1,25	0,44
	9	2	4	2	6	2	5	10	2,06	0,62
	10	3	4	3	7	3	6	20	2,41	0,55
	11	3	6	3	9	4	8	70	3,27	0,74
	12	5	7	5	12	5	11	462	4,65	1,17
	13	4	4	4	8	3	7	35	2,91	0,80
	14	6	5	5	11	5	10	252	4,15	0,92
	15	5	6	5	11	5	10	252	4,15	0,92
	16	4	9	4	13	6	12	924	5,03	1,10
Total			**Z**	**38**					**37,24**	**9,56**

IC	31,05864	43,426712

Z appartient à l'intervalle

Tableau 27- Résultats de l'hypothèse H_0 triangle RC

	j	L	S	Z	t	m	t-1	Cm t-1	E	V
	2	2	0	0	2	0	1	1	0,50	0,25
	3	1	1	1	2	0	1	1	0,50	0,25
	4	2	1	1	3	1	2	2	0,75	0,19
	5	2	0	0	2	0	1	1	0,50	0,25
	6	1	2	1	3	1	2	2	0,75	0,19
	7	1	3	1	4	1	3	3	1,25	0,44
	8	1	2	1	3	1	2	2	0,75	0,19
	9	2	1	1	3	1	2	2	0,75	0,19
	10	1	0	0	1	0	0	1	0,00	0,00
	11	1	3	1	4	1	3	3	1,25	0,44
	12	1	1	1	2	0	1	1	0,50	0,25
	13	2	4	2	6	2	5	10	2,06	0,62
	14	2	5	2	7	3	6	20	2,41	0,55
	15	5	3	3	8	3	7	35	2,91	0,80
	16	6	3	3	9	4	8	70	3,27	0,74
Total			**Z**	**18**					**18,14**	**5,34**

IC	13,523076	22,765987

Z appartient à l'intervalle

Tableau 28- Résultats de l'hypothèse H_0 triangle RC Produit

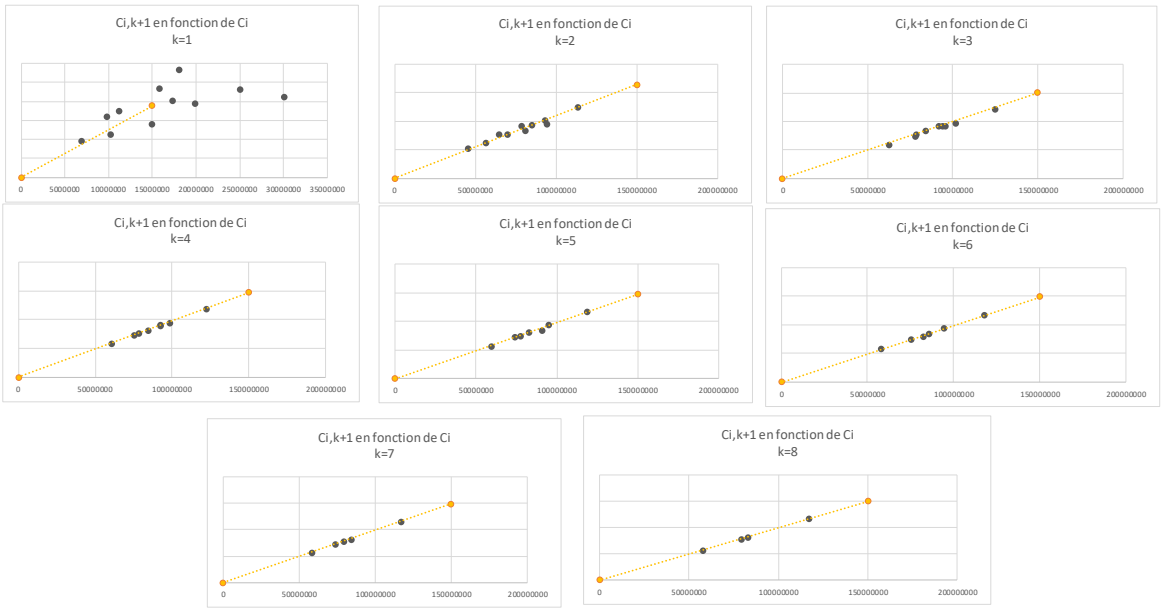


Figure 61- Résultats de l'hypothèse H_1 sur le triangle Corps

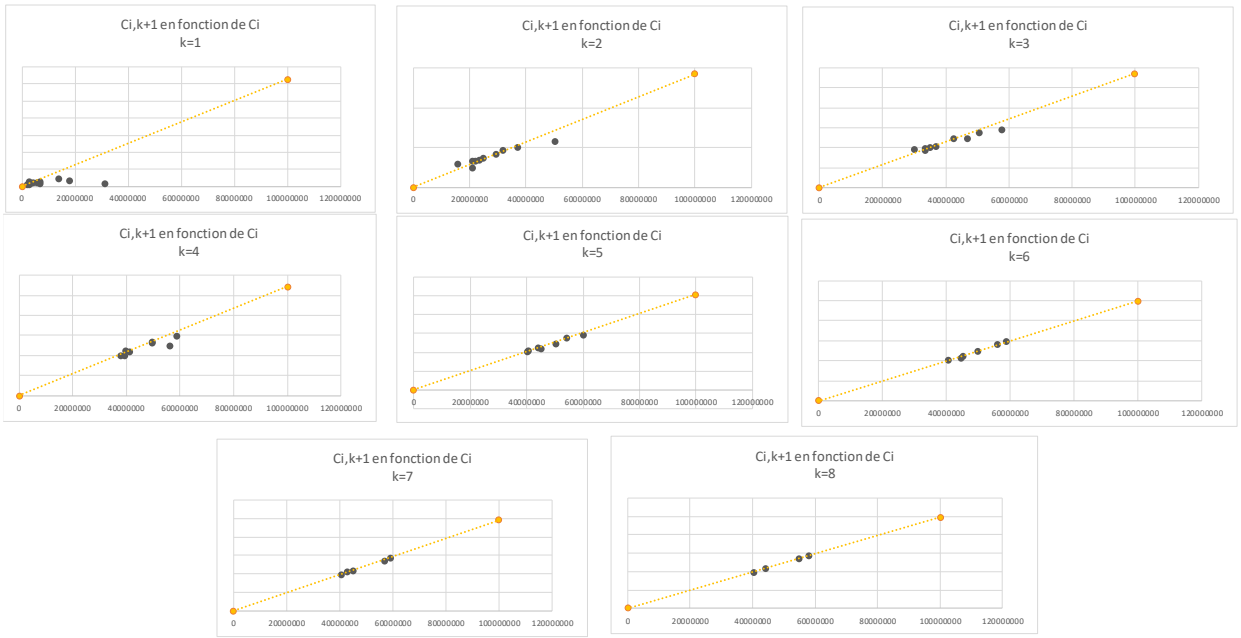


Figure 62-Résultats de l'hypothèse H_1 sur le triangle RC

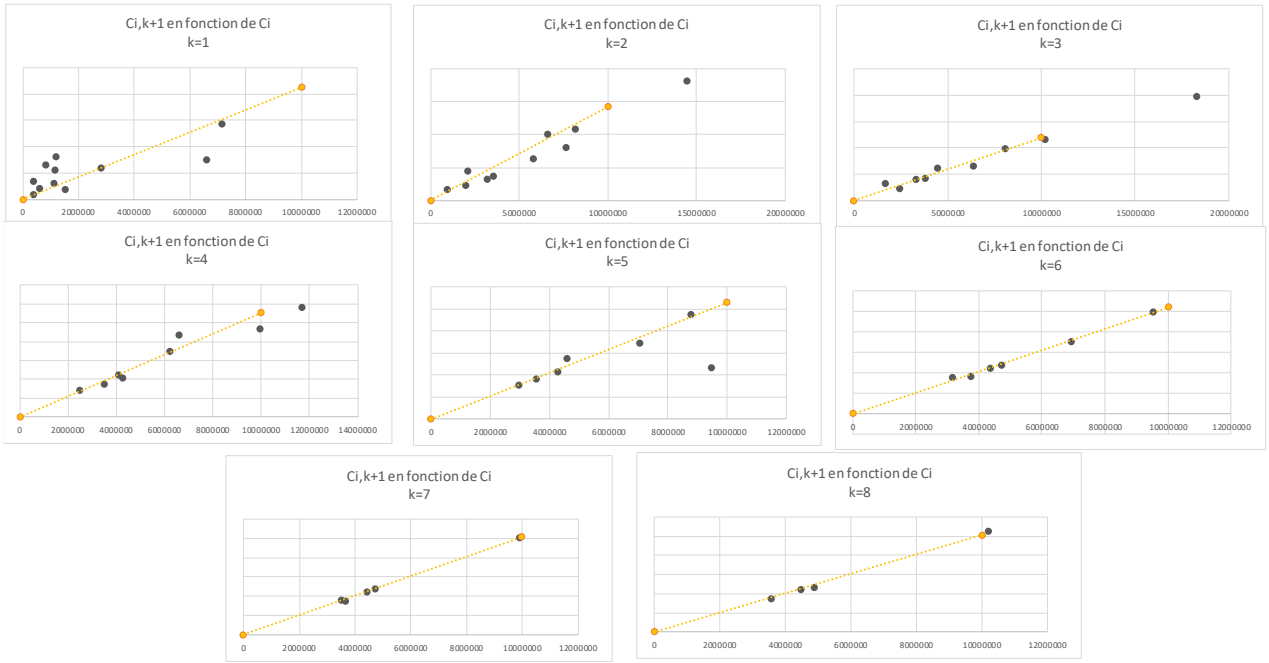


Figure 63-Résultats de l'hypothèse H_1 sur le triangle RC Produit

A.2. ANALYSES DES DONNEES

A.1.1. Bases maritimes

Variables	Commentaires
PT1 PT2 PT3 PT4 PT5 PT6 PT7 PT8	Montants de paiements cumulées nets de recours à chaque trimestre année 1 à 2.
Prov1 Prov2 Prov3 Prov4 Prov5 Prov6 Prov7 Prov8	Montants de provisions cumulées nettes de recours à chaque trimestre année 1 à 2.
C1 C2 C3 C4 C5 C6 C7 C8	Montants des charges totales cumulées nettes de recours à chaque trimestre année 1 à 2.
EXERCICE	Exercice de souscription
COMPAGNIE	
POURCENTAGE	Part assurée
Prime.de.la.categorie	Prime payée
Nombre.de.navire	Nombre de navires dans la flotte
VALEUR_ASSURE	Valeur totale assurée
annee_construction	Année de construction du bâteau sinistré
TONNAGE	Poids du navire sinistré
LONGUEUR	Longueur du navire sinistré
NB_MOTEURS	Nombre de moteurs sur le bateau sinistré
PUISSANCE	Puissance du bateau sinistré
CONTENTIEUX	Si contentieux ou non sur le sinistre
lob_actuariel	Type de sinistres (PAC CMM ….)
Famille	
Garantie	Type de garantie (Dommage RC Rcproduit)
Libelle_Terr	Territoir du sinistre
MARCHE	
MATERIAU	Matériau du navire
AttL	Si le sinistre est attritionnel ou large au bout de 2 années de développement
Censure	Sinistre Clos ou non au 31/12/2021
Durée	jours écoulés entre l'ouverture et la clôture du sinistre (ou la fin d'observation 31/12/2021)
CU	La charge ultime du sinistre, à la dernière vue disponible soit 31/12/2021 (variable cible)

Tableau 29-Variables utilisées en maritime

A.1.2. Bases primes et sinistres aviation

Base primes Aslf

Nom des champs		Nom des champs	
Vu au	DIMENSIONS	Part RA&S (GIE) brute%	FAITS
Exercice-UWYR		Part RA&S (GIE) brute nette de fac rml%	
Agence de souscription		Cumul Police 100% Maître en USD	
Code Groupement de Gestion		Cumul Police Part RA&S (GIE) en USD	
Nom Groupement de Gestion		Cumul/Limite Garantie commerciale RC 100% Maître en USD	
Code Groupement statistique		Cumul/Limite Garantie commerciale RC part RA&S (GIE) en USD	
Nom Groupement statistique		Prime émise brute HT 100% Maître en USD	
Code_ Assuré		Commission Courtage initiale 100% Maître en USD	
Nom Assuré		Prime émise nette de com courtage 100% Maître en USD	
Pays de l'Assuré		Prime estimée brute HT 100% Maître en USD	
PKPAYS du Pays de l'assuré		Prime estimée nette de com courtage 100% Maître en USD	
Pays d'opération de l'Assuré		Prime émise brute HT 100% Police en USD	
PKPAYS du Pays d'opération de l'Assuré		Commission Courtage initiale 100% Police en USD	
Code Activité Principale Assuré		Prime émise nette de com courtage 100% Police en USD	
Nom Activité Principale Assuré		Prime estimée brute HT 100% Police en USD	
Regroupement NSI Activité Assuré		Prime estimée nette de com courtage 100% Police en USD	
Code Activité Police		Prime émise brute HT Part RA&S (GIE) en USD	
Nom Activité Police		Commission Courtage initiale Part RA&S (GIE) en USD	
Regroupement NSI Activité Police		Prime émise nette de com courtage Part RA&S (GIE) en USD	
Code_ Cédante		Coût d'acte Part RA&S (GIE) en USD	
Name Cédante		Taxes collectées Part RA&S (GIE) en USD	
Pays de la Cédante		Taxes déduites Part RA&S (GIE) en USD	
Code_ Intermédiaire		Prime de dépôt Part RA&S (GIE) en USD	
Name Intermédiaire		Commission Courtage initiale de la Prime de dépôt Part RA&S (GIE) en USD	
Pays de l'Intermédiaire		Prime de dépôt Part RA&S (GIE) nette de com courtage en USD	
Code Gestionnaire prime		Prime estimée brute HT Part RA&S (GIE) en USD	
Nom Gestionnaire prime		Prime estimée nette de com courtage Part RA&S (GIE) en USD	
Pays Gestionnaire prime		Paid (Principal+fees) 100% Maître en USD	
Code Apériteur		PSAP (Principal+fees) 100% Maître en USD	
Nom Apériteur		Total Charge Sinistre (Principal+fees) 100% Maître en USD	
Statut de la Police		Paid (Principal+fees) Part RA&S (GIE) en USD	
Code STAMP		PSAP (Principal+fees) Part RA&S (GIE) en USD	
Nom STAMP		Total Charge Sinistre (Principal+fees) Part RA&S (GIE) en USD	
Type de Police			
Produit NSI			
N° Police			
N° police Maître			
Agence police maitre			
Date d'effet de la Police			
Mois effet police			
Date d'expiration de la police			
Date de résiliation			
Type Affaire			
Souscripteur			
Schedule ABC			

Tableau 30-Variables de la base prime aviation

Base sinistres Aslf

Nom des champs		Nom des champs	
Vu au	DIMENSIONS	Paid+fees Group share	FAITS en USD
Mois		PSAP "Principal+fees" Group share	
Exercice-UWYR		Total Group share	
Num Evt			
Nom Evt			
Num Sinistre - NSI			
Num dossier sinistre-NSI			
Risques			
Garantie_Reassurance			
Code_ Assured			
Nom_Assured			
Agence			
Rgpt_Activite			

Tableau 31-Variables de la base sinistres

A.1.3. Analyses maritime

Distribution des montants de sinistres

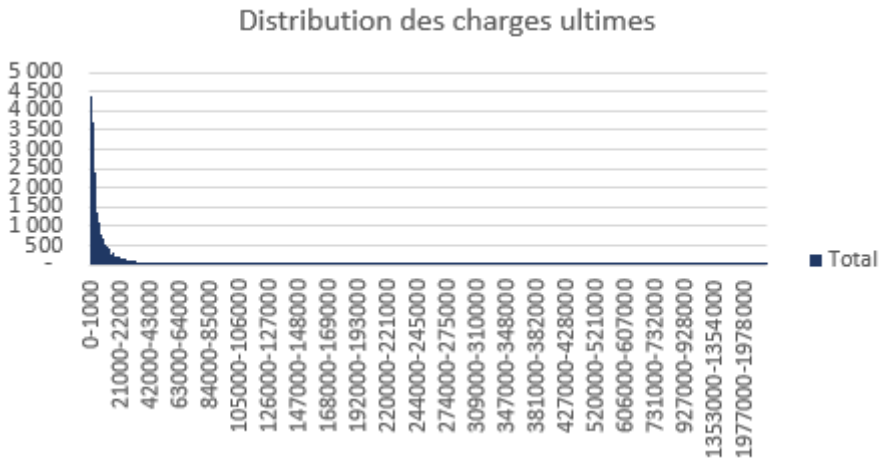


Figure 64-Distribution des charges ultimes exercices 2010 à 2019 en maritime

Evolution du nombre de sinistres

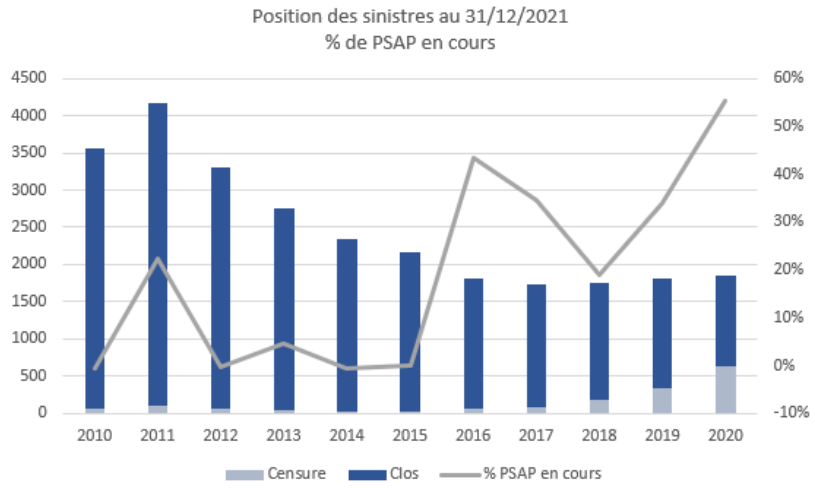


Figure 65-Nombre de sinistres censurés et pourcentages de provisions dossier à dossier en maritime

Durée et charge ultime

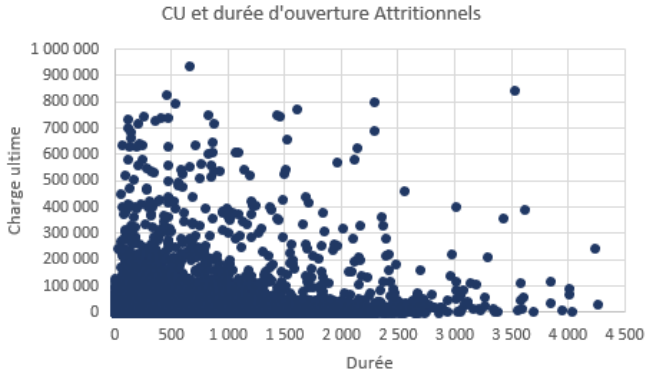


Figure 66-Charges ultimes par rapport à la durée d'ouverture sur les sinistres maritime

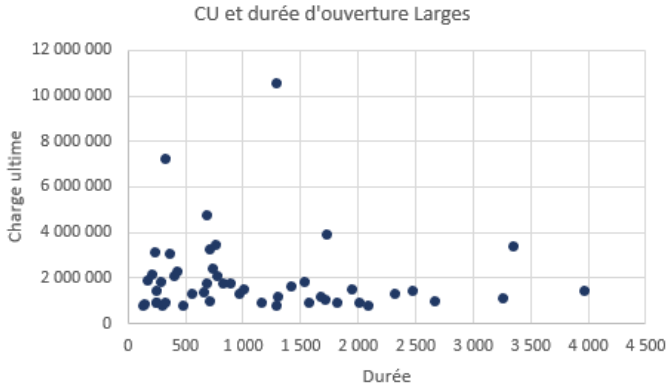


Figure 67-Charges ultimes par rapport à la durée d'ouverture sur les sinistres larges maritime

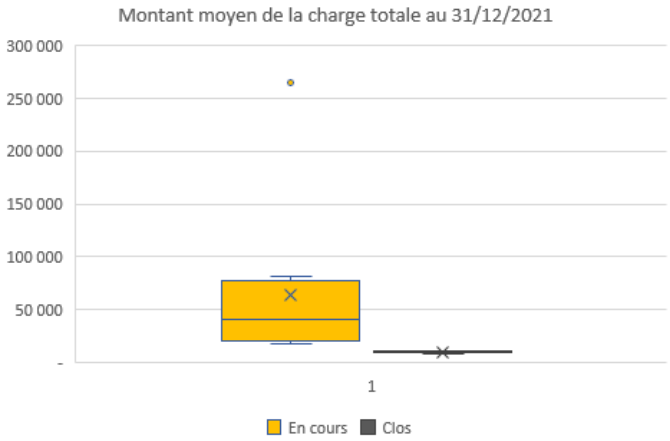


Figure 68-Montant moyen des sinistres en cours et clos en maritime

Durée et contentieux

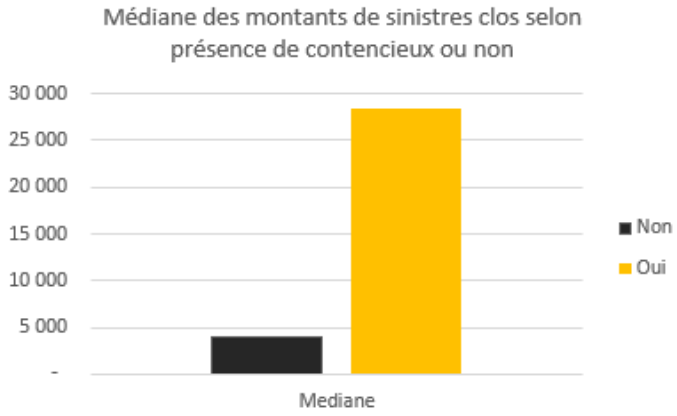


Figure 69-Médiane des montants de sinistres avec contentieux ou non en maritime

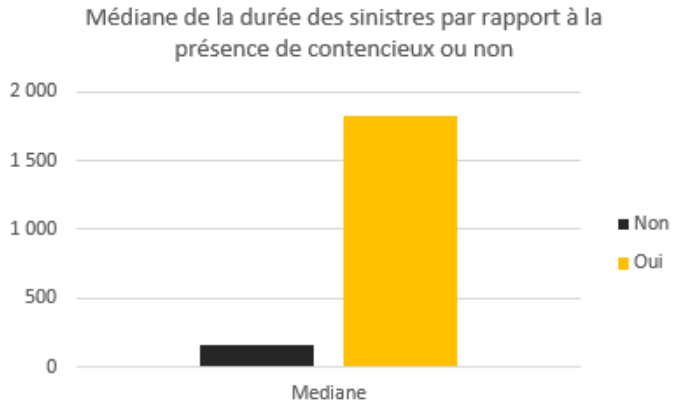


Figure 70- Médiane de la durée de sinistres avec contentieux ou non en maritime