Mémoire présenté devant le jury de l'EURIA en vue de l'obtention du Diplôme d'Actuaire EURIA et de l'admission à l'Institut des Actuaires

le 7 Septembre 2022

Par : Mohammed-Amine SKOUBANI
Titre : Assurance construction : Provisionnement Dommages-ouvrage exploitant les typologies des sinistres et les risques assurés

Confidentialité : Non

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

Membre présent du jury de l'Institut des Actuaires :
Alexandre YOU
Signature :
Hélène GIBELLO (visio)
Signature :
Davy SENGDY (visio)
Signature :

Membres présents du jury de l'EURIA :
Rainer BUCKDAHN
Signature :

Entreprise : Allianz France
Mohamed ZAIMI
Signature :

Directeur de mémoire en entreprise :
Sébastien FARKAS
Nicolas TROUILH
Signature :

Invité :
Bernard BAILLEUL
Khadija CHEHABI
Signature :

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

Signature du responsable entreprise :

Signature du candidat :

EURIA
EURo Institut d'Actuariat

6, avenue le Gorgeu
CS 93837
29238 Brest Cedex 3

T +33 (0)2 98 01 66 55
euria@univ-brest.fr

Résumé

Les maîtres d'œuvre à la construction d'un ouvrage ont l'obligation de souscrire une assurance responsabilité civile décennale (RCD) pour couvrir leurs responsabilités à la manifestation d'un sinistre. Le déclenchement de cette couverture est ainsi soumis aux responsabilités engagées, dont la détermination peut s'avérer complexe. La garantie Dommages-ouvrage (DO), souscrite par le maître d'ouvrage, permet d'indemniser le préjudice dès son observation et sans recherche de responsabilités. Par subrogation au client, l'assureur DO cherchera à son tour à appliquer les garanties RCD, recouvrant ainsi tout ou partie du dédommagement.

Cela étant dit, la nature de cette branche perturbe, pour les assureurs, le suivi de leur niveau de sinistralité et rend difficilement applicable les techniques de provisionnement usuelles en assurance non-vie. Ces difficultés rendent la prise en compte des caractéristiques des ouvrages essentielle pour une meilleure estimation des provisions. La destination des ouvrages, fait partie de ces caractéristiques, et est liée par nature à la sinistralité des ouvrages par la notion d'impropriété à la destination. Elle peut donc dégager une hétérogénéité de la dynamique de provisionnement des sinistres, que ce soit en charge ou en recours.

Nous proposons dans ce mémoire d'actuariat d'étudier la prise en compte de la destination des ouvrages pour enrichir la modélisation des provisions pour la garantie DO. Dans un premier temps, une analyse de la projection sera étudiée selon une vision agrégée des sinistres, classiquement modélisée par les triangles de développement. Par rapport à ce champ, nous proposerons d'étudier la maille d'agrégation qui dégage le plus d'homogénéité en termes de développement. Nous tenterons dans un second temps d'illustrer le potentiel des approches qui incorporent l'information ligne à ligne des sinistres en modélisant l'état à la sortie des sinistres, en quantifiant le coût des sinistres enregistrés non clos après une adaptation du modèle [Lopez 2018] ou encore en considérant des dynamiques individuelles de recours.
On comparera ces différents types de modélisation selon un *backtest*. Une attention particulière sera portée à une idée de modélisation qui combine la projection des triangles en agrégé et la modélisation ligne à ligne.

Mots clefs: Provisionnement non-vie, Dommages-ouvrage, Segmentation, Provisionnement ligne à ligne, Données censurées

Abstract

Project managers in the construction of a structure are required to take out ten-year civil liability insurance (RCD) to cover their responsibilities in the event of a claim. The triggering of this coverage is thus subject to the responsibilities incurred, the determination of which can prove to be complex. The work damage guarantee (DO), taken out by the project owner, makes it possible to repair the damage as soon as it is observed and without seeking liability. By subrogation to the client, the DO insurer will in turn seek to apply the RCD guarantees, thus recovering all or a part of the compensation.

That being said, the nature of this branch makes it difficult for insurers to monitor their level of claims and also makes it difficult to apply the usual reserving techniques in non-life insurance.

These difficulties make it mandatory to take into account the features of the structures for a better estimation of the reserves. The destination of the structures is one of those characteristics, and is inherently linked to the loss experience of the structures by the notion of inappropriateness to the destination. It can therefore reveal a heterogeneity in the patterns of reserving of claims, whether in payements or in recoveries.

We propose in this actuarial dissertation to study the consideration of the destination of the structures to enrich the modeling of the reserves for the DO guarantee.

Initially, an analysis of the projection will be studied according to an aggregate vision of the claims, classically modeled by the triangles of development. In relation to this field, we will propose to study the aggregation classification that releases the most homogeneity in terms of development. We will then attempt to illustrate the potential of approaches that incorporate individual information of claims by modeling the state at the exit of the claims, by quantifying the cost of reported claims not settled after an adaptation of the model [Lopez 2018] or by considering individual dynamics of recoveries.

We will compare these different types of modeling according to a backtest. Particular attention will be paid to a modeling idea that combines the projection of triangles in aggregate form and individual modeling.

Keywords: Non life Reserving, DO, Clustring, Individual reserving, Censored data

Note de synthèse

Contexte

La construction d'un ouvrage nécessite l'intervention de plusieurs acteurs et introduit différents types de risques. Le maître d'ouvrage, s'agissant d'un promoteur immobilier ou d'un particulier constructeur, supervise les travaux réalisés par les maîtres d'œuvre, pouvant être des architectes, un bureau d'étude ou encore des ingénieurs, selon la nature des travaux. Chacun de ces acteurs voit sa responsabilité engagée quant à la livraison des travaux.
Ceci rend l'assurance particulièrement primordiale. Le régime décennal de l'assurance construction apparut en 1978 par la loi Spinetta, instaure une double obligation, d'une part vis-à-vis du maître d'ouvrage en introduisant une assurance dommage (**Dommages-ouvrage**) et d'une autre part vis-à-vis des maîtres d'œuvre, quels que soient leurs rôles, avec une assurance **responsabilité civile décennale** (RCD).

Ces deux assurances sont interreliées : L'assurance dommages-ouvrage s'introduit pour permettre au maître d'ouvrage d'obtenir une indemnisation des sinistres qu'il déclare pour un préfinancement rapide, indépendamment de toute recherche de responsabilité. Une fois l'acteur causant les désordres identifié, l'assureur de la garantie dommages-ouvrage pourra exercer son **recours** envers l'assureur de responsabilité. La loi Spinetta institue un principe de garanties interdépendantes dit à « double détente ».
L'appréciation des dommages dépend de leur degré de gravité. L'article 1792 du code civil spécifie que les désordres manifestés doivent compromettre la solidité de l'ouvrage ou le rendent impropre à sa **destination** en affectant l'un de ses éléments constitutifs ou de ses équipements.
Ainsi, l'appréciation des désordres d'ouvrage à destination habitation sera différente de celle à destination industrielle ou commerciale.
Une autre particularité de ce régime est que ses garanties couvrent les dommages apparus dix ans après la date de fin de réception des travaux. Ceci rend sa maîtrise cruciale, notamment en matière de **provisionnement**.

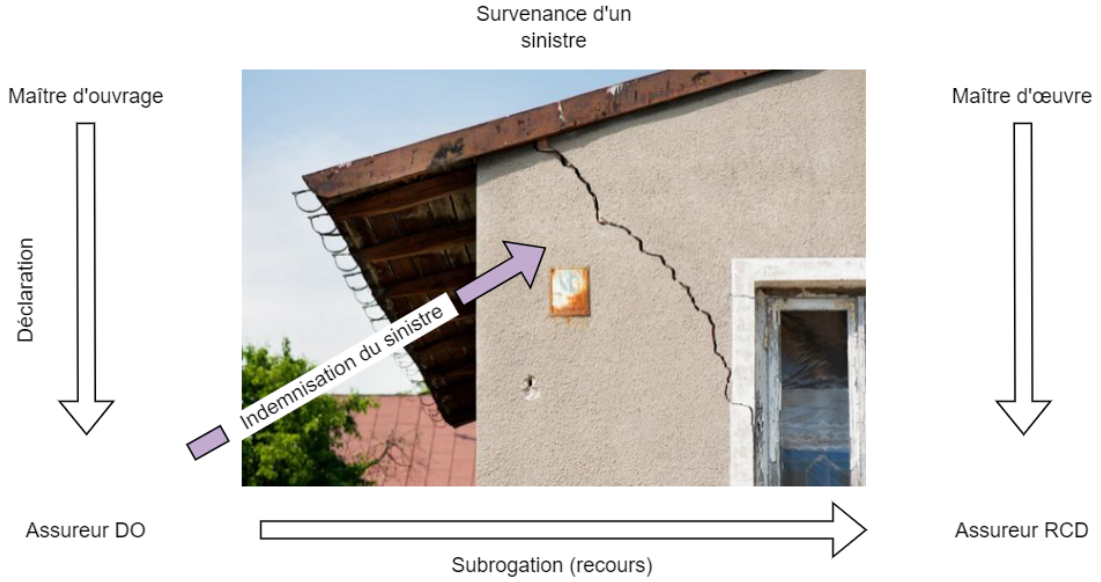

FIGURE 1 – Synthèse des garanties DO et RCD [1]

Les organismes d'assurance proposent ainsi deux types de produits : Une assurance RCD et une assurance DO. Concernant la DO, le tarif proposé est notamment fixé en fonction des préjudices et des recours attendus *a priori*. Une fois commercialisés, l'organisme d'assurance quantifie régulièrement le montant de ses engagements, décomposés de la manière suivante, par exercice de provisionnement :

- **IBNyR** (*Incurred But Not Yet Reported*) : Les engagements relatifs aux sinistres survenus non encore enregistrés
- **IBNeR** (*Incurred But Not Enough Reported*) : Les engagements relatifs au développement de l'état des sinistres enregistrés
- **PSNEM** (Provisions pour sinistres non encore manifestés) : Les engagements relatifs aux sinistres non encore manifestés.

La PSNEM représente une autre spécificité de l'assurance construction venant de la période décennale des garanties qui permet une manifestation ultérieure des sinistres avec une cause génératrice rattachée à la période des travaux de chantiers.

De ce fait, une attention particulière est donnée à la modélisation actuarielle des réserves, qui doivent être d'autant plus ajustées et affinées sur les garanties décennales, notamment la Dommages-ouvrage, qui connaît des particularités en termes de destination des ouvrages. L'intégration de ces spécificités en termes de provisionnement et de modélisation constituera le cœur des travaux de ce mémoire.

1. Source de l'image de désordres : `https://www.lesaffaires.com/mes-finances/immobilier/vice-cache-faut-il-poursuivre-les-anciens-proprietaires`

Méthodologie

Les modèles de provisionnement classiques sont généralement basés sur une vision agrégée des données (triangles) par année d'ouverture de chantiers (**DOC**), de survenance ou d'enregistrement, selon ce qu'on cherche à estimer. Diverses techniques actuarielles permettent d'estimer la partie non connue des triangles. La méthode de **Chain-Ladder**, comme exemple utilisé dans ce mémoire, projette progressivement la dernière vision connue du développement des flux à travers des facteurs de développement. D'autres méthodes permettent une **modélisation individuelle** des sinistres en prenant en compte leurs états à la clôture, leurs durées et leurs coûts.
Cela étant dit, un choix de la **maille** sur laquelle les triangles seront construits est nécessaire dans le sens où il serait plus pertinent de regrouper des typologies particulières de sinistres pour affiner la modélisation.

Après une importante étape de traitement textuel, visant à construire des classes de destination à partir des champs textuels bruts, on propose dans ce mémoire de les introduire dans la modélisation de provisionnement à travers différentes méthodes et sur plusieurs aspects.
La comparaison de ces méthodes se fera à base de ***backtesting***, en comparant les règlements estimés à l'année 2021, en se limitant à la vision des données à fin 2020, avec les règlements observés.

1. Méthode en **2D** (deux dimensions) : Globale
   En vision agrégée, le calcul de réserves le plus naturel est selon un triangle DOC - Développement. Ce triangle permet d'estimer des réserves par DOC relatives à toutes les typologies de sinistres, s'agissant d'aggravations, de sinistres tardifs ou de sinistres non encore survenus.
2. Méthode en 2D : Avec segmentation
   Un premier travail effectué dans ce mémoire serait de sélectionner la maille la plus optimale d'une projection en deux dimensions en testant différents croisements de variables caractérisant les sinistres. La destination, le réseau de distribution et la première évaluation de la charge comparée au **ticket modérateur** (TM), seuil défini par une convention entre les assureurs construction (convention CRAC), sont parmi ses caractéristiques. Ceci revient à créer des triangles par typologies et de projeter chacun à part.
3. Méthode en **3D** (trois dimensions) : Globale
   La vision en deux dimensions (DOC-Développement), utilisée pour la sélection de la maille, ne permet pas d'estimer directement la PSNEM. Une idée intéressante en assurance construction permet d'estimer la PSNEM sur la base d'un triangle DOC - Survenance. En effet, ce triangle change à chaque date d'arrêté avec les aggravations et les sinistres tardifs qui s'ajoutent sur sa partie supérieure supposée connue. Une projection directe à l'ultime n'est pas donc envisageable. L'idée évoquée consiste à vieillir ce triangle à l'ultime en ventilant les IBNR estimées par

année de survenance. Cette méthode (appelée méthode 3D) combine donc trois visions différentes, à savoir l'année d'observation, l'année de survenance et l'année de DOC.
Nous proposons dans ce mémoire de la tester par rapport à l'année d'enregistrement, ce qui permet d'estimer une PSNEM augmentée des IBNyR.

4. Méthode en 3D : Avec segmentation
Bien que la segmentation ait été choisie selon une vision en deux dimensions, elle influera également l'écoulement en trois dimensions.
Nous évaluerons cette méthode sur la maille sélectionnée.

Ce type de modélisation agrégée, interprétable et relativement facile à implémenter, ne prend pas le caractère individuel des sinistres et leurs projections. Leurs caractéristiques deviennent de plus en plus importantes en Dommages-ouvrage vu la nature très hétérogène des sinistres, avec plus de 50% de **sinistres sans suite en principal** et de destinations très variées.

Pour affiner l'estimation des IBNeR, nous proposons de tester une modélisation adaptée à ces spécificités, combinant une estimation de la durée et des coûts des sinistres en prenant en compte leur dépendance, ceci après la construction d'un modèle de prédiction de l'état à la sortie des sinistres (sans ou avec suite en principal).
Aussi, la modélisation individuelle devrait prendre en compte les sinistres censurés. En effet, construire un modèle, que ce soit pour prédire l'état à la clôture des sinistres ou pour estimer leurs coûts, en ne se basant que sur les sinistres clos, risque de biaiser l'estimation. On adaptera tous nos travaux de modélisation individuelle à cette problématique en utilisant la méthode "*Inverse-probability-of-censoring weighting*" (**IPCW**), qui permet de lever ce biais, en pondérant les sinistres clos avec des poids permettant de mieux prendre en compte les sinistres clos de longue durée.
Cette dernière modélisation permet d'estimer les IBNeR.

Nous proposerons avec la modélisation suivante une approche à notre connaissance novatrice, permettant de quantifier les engagements liés aux sinistres IBNyR et des sinistres non encore survenus au moyen d'une méthodologie agrégée 3D utilisant l'estimation ligne à ligne - et non plus 2D - de la charge ultime IBNeR.
Ceci permettra d'estimer une PNSEM augmentée d'IBNyR, vu qu'on projette par année d'enregistrement selon le modèle individuel.

5. Modélisation ligne à ligne + 3D : Globale
L'estimation de la PSNEM pourrait naturellement s'effectuer avec une combinaison adaptée du modèle ligne à ligne et de la méthode 3D, en vieillissant les triangles DOC-Enregistrement avec les IBNeR du modèle individuel. Ceci permet de coupler les bénéfices de la modélisation individuelle avec le caractère robuste et interprétable des projections avec les triangles agrégés.
6. Modélisation ligne à ligne + 3D : Avec segmentation
Également, la segmentation de la méthode en trois dimensions permet de mieux

capter le cadencement des sinistres non encore enregistrés en ventilant les IBNeR sur des triangles qui sont censés être plus homogènes.

Finalement, la maîtrise du provisionnement en Dommages-ouvrage, étant une garantie de préfinancement sans recherche de responsabilité, devrait prendre en compte les recours et leurs dynamiques. Nous proposerons d'estimer donc un taux de recours, sur la maille sélectionnée et au global.
Les taux de recours pourront changer au cours du temps selon les déformations du portefeuille et l'évolution de la jurisprudence . Un calcul d'un taux global ne sera pas spécifique à ces changements vu qu'il faut prévoir une période dépassant la couverture décennale pour qu'une DOC donnée devienne relativement mature. Un calcul en se limitant sur une sélection de DOC, pour chaque segment de destination, serait donc plus pertinent. On utilisera également ici la correction de l'effet de censure par IPCW.

Résultats

Backtest 2021

— Sélection de la maille optimale : Une méthodologie à base d'échantillonnage a été suivie pour sélectionner la maille optimale. La maille sélectionnée est le croisement des variables **TM x Destination**.
— Comparaison des modèles :

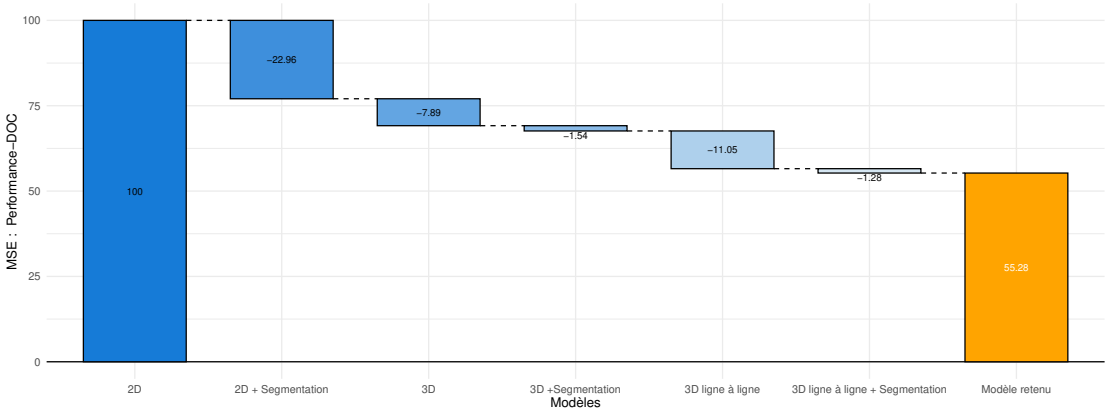

FIGURE 2 – Résultat du *backtesting* : Performance à base de la MSE par DOC rapportée au modèle 2D

La combinaison des modèles qu'on introduit paraît être plus adaptée selon les résultats du *backtesting*. Cette modélisation permet de prendre les spécificités individuelles des sinistres en la combinant avec la robustesse de la projection triangulaire agrégée.

Taux de recours et réserves nettes de recours

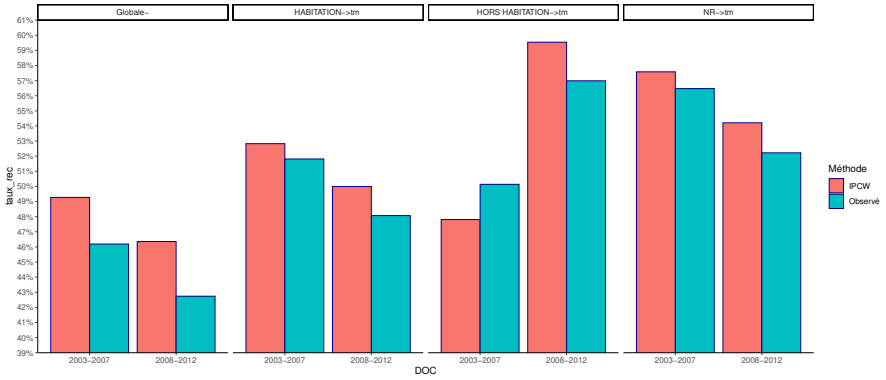

FIGURE 3 – Taux de recours par Strate : Observé-IPCW

Les taux de recours ont été estimés par classe de destination. Ils pourront aider à une meilleure estimation des recours, si l'assureur décide de changer de stratégie de souscription par nature de destination.
En retenant un taux de recours global de 49.2%, les réserves globales nettes de recours, par modèles, sont illustrées ci-dessus :

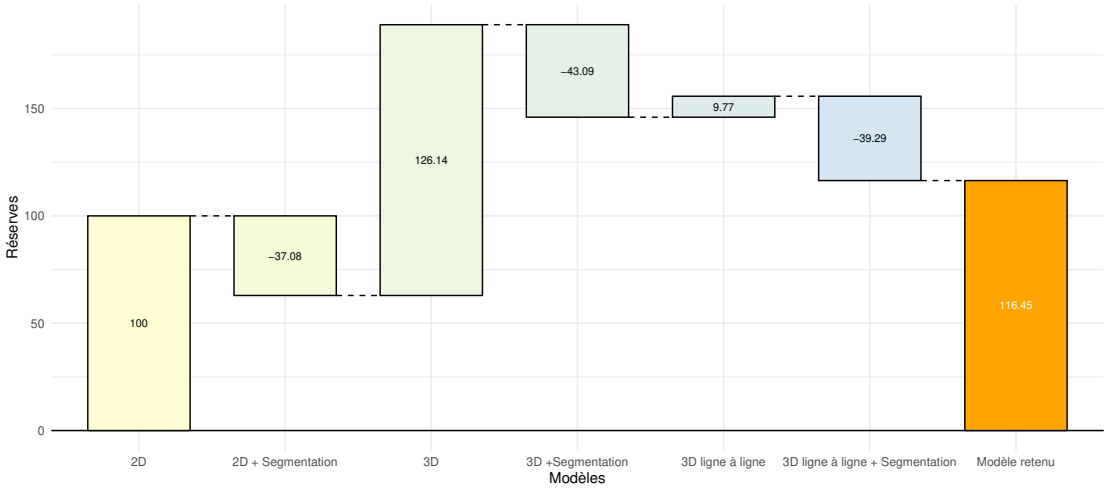

FIGURE 4 – Réserves nettes de recours exprimées relativement au modèle 2D

Ces modèles produisent un niveau de réserve très varié par rapport à la modélisation en deux dimensions.

Conclusion

La démarche proposée dans ce mémoire a permis d'illustrer le potentiel de la prise en compte des typologies des sinistres sur la quantification des provisions liées aux garanties DO. Nous avons également mis en évidence l'impact des choix de modélisation sur les performances et les résultats des méthodes, permettant de suggérer une modélisation optimale et à notre connaissance novatrice, reposant sur la combinaison de triangles agrégés avec quantification des engagements liés aux sinistres IBNeR selon une modélisation individuelle.

Le périmètre de ce mémoire actuariel n'englobe cependant pas l'ensemble des enjeux liés aux calculs des provisions techniques.
Le *backtest* effectué se limite à l'année 2021 et ne permet pas de généraliser nos conclusions sur la pertinence des modèles. Des hypothèses de modélisation ont été également prises, notamment par rapport au retraitement de **l'inflation**, aux choix des lois marginales et du modèle de dépendance malgré la non-validation des tests statistiques. L'étape de qualité de données, relative au retraitement du **champ textuel** de la destination, montre aussi une limite vu qu'on regroupe des mixtes de destination dans la classe des destinations non classifiées, créant une hétérogénéité additionnelle.

Finalement, nos travaux peuvent voir une utilisation directe dans les différentes approches normatives de calculs des réserves, notamment selon la nouvelle norme IFRS17 qui incite à une sélection de maille selon la profitabilité, pouvant par ailleurs avoir une appréciation différente selon la destination des ouvrages. Tout ceci dans l'espérance d'éclairer les risques par rapport à une branche d'assurance difficilement maîtrisable.

Executive Summary

Context

The construction of a structure requires the intervention of several actors and introduces different types of risks. Whoever builds (project owner) risks damage to his work and the builders (project managers), for their part, see their responsibilities engaged.

This makes insurance particularly paramount. The ten-year construction insurance scheme appeared in 1978 through the Spinetta law establishing a double obligation, on the one hand of the client by introducing damage insurance (**Dommages-ouvrage**) and on the other of the contractors, whatever their roles, with **ten-year civil liability** (RCD) insurance.

These two insurances are interrelated : Property damage insurance is introduced to allow the project owner to obtain compensation for the claims he declares for rapid pre-financing, independently of any search for liability. Once the actor causing the disturbances has been identified, the property damage guarantee insurer may exercise its **recoveries** against the liability insurer. The Spinetta law establishes a principle of interdependent guarantees known as "double trigger".

The assessment of the damages is according to their degree of gravity. Article 1792 of the Civil Code specifies that the disorders manifested must compromise the solidity or render it unsuitable for its **destination** by affecting one of its constituent elements or equipment.

Thus, the assessment of the disorders of works intended for housing will be different from those intended for industrial or commercial use.

Another particularity of this scheme is that these guarantees cover damage occurring ten years after the date of completion of acceptance of the work. This makes its control crucial, especially in terms of **reserving**.

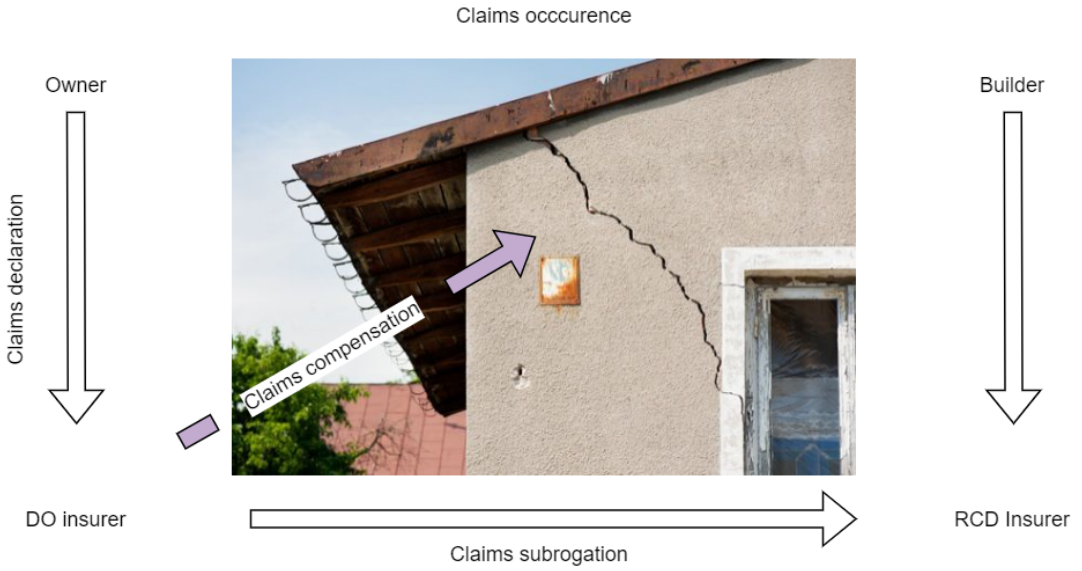

FIGURE 5 – Summary of DO and RCD guarantees [2]

Insurers, to honor their commitments, constitute reserves according to their risk estimates.
They establish, after declaration of a claim, a so-called outstanding reserves, representing the cost of the claim, knowing the information available on the date of the assessment. This estimate is reassessed according to the state of the expertise and the degree of aggravation of the damage.

More generally, insurers estimate different types of reseves :

— **IBNyR** : Related to Incurred But Not Yet Reported claims
— **IBNeR** : Related to Incurred But Not Enough Reported claims
— **PSNEM** : Reseves for claims that have not yet occurred

The PSNEM represents another specificity of construction insurance coming from the ten-year period of guarantees which allows a subsequent manifestation of claims with a generating cause attached to the period of construction works.

As a result, particular attention is given to the actuarial modeling of reserves, which must be all the more adjusted and refined on ten-year guarantees, in particular property damage, which has particularities in terms of the destination of the structures. The integration of these specificities in terms of reserving and modeling will constitute the heart of the work of this dissertation.

2. Source of the image of disorders : `https://www.lesaffaires.com/mes-finances/immobilier/vice-cache-faut-il-poursuivre-les-anciens-proprietaires`

Methodology

After an important step of textual processing, aiming to build destination classes from raw textual fields, we propose in this thesis to include it in the provisioning modeling through different methods and on several aspects.
The comparison of the methods will be based on **backtesting**, by comparing the estimated settlements for the year 2021, limited to the view of the data at the end of 2020, with the observed settlements.

Conventional reserving models are generally based on an aggregate view of data (triangles) by year of opening of construction sites (**DOC**), occurrence or recording, depending on what you are trying to estimate. Various actuarial techniques can estimate the unknown part of the triangles. The **Chain-Ladder** method, as an example used in this thesis, progressively projects the latest known view of flow development through development factors. Other methods allow **individual modeling** of claims by taking into account their status at closing, their durations and their costs.
That being said, a choice of the **classification** on which the triangles will be built is necessary in the sense that it would be more relevant to group particular types of claims together to refine the modelling.
We propose in this thesis to test various types of modeling :

1. Two-dimensional method : Global
   In aggregate view, the most natural calculation of reserves is according to a DOC - Development triangle.
2. Two-dimensional method : With segmentation
   A first work done in this thesis would be to select the most optimal classification of a two-dimensional projection by testing different intersections of variables characterizing claims. The destination, the distribution network and the first evaluation of the load compared to the **moderator ticket** (TM), threshold defined by an agreement between construction insurers (CRAC agreement), are among its characteristics.
3. Method in **three dimensions** : Global
   The two-dimensional vision (DOC-Development), used for clustring procedure, does not allow the PSNEM to be estimated directly. An interesting idea in construction insurance makes it possible to estimate the PSNEM on the basis of a triangle DOC - Occurrence. Indeed, this triangle changes at each closing date with aggravations and late ones which are added to its supposedly known upper part. A direct projection to the ultimate is therefore not possible. The idea mentioned consists in aging this triangle to the ultimate by breaking down the estimated IBNRs by year of occurrence. This method (called 3D method) therefore combines three different views, namely the year of observation, the year of occurrence and the year of DOC.
   We propose in this thesis to test it in relation to the year of recording, which makes it possible to estimate a PSNEM increased by IBNyR

4. Three-dimensional method : With segmentation
   Although the segmentation was chosen according to a two-dimensional view, it will also influence the three-dimensional flow.
   We will evaluate this method on the selected segmentation.

This type of aggregate modelling, interpretable and relatively easy to implement, does not take on the individual nature of claims and their projections. Their characteristics are becoming increasingly important in property damage given the very heterogeneous nature of claims, with more than 50% of **claims without main action** and a wide variety of destinations.

To refine the estimate of IBNeR, we propose to test a model adapted to these specificities, combining an estimate of the duration and costs of claims taking into account their dependence, this after the construction of a prediction model of the status at the end of claims (without or with main action).
Also, individual modeling should take into account censored claims. Indeed, constructing a model, whether to predict the status at the close of claims or to estimate their costs, based only on closed claims, risks biasing the estimate. We will adapt all our individual modeling work to this problem by using the "*Inverse-probability-of-censoring weighting*" (**IPCW**) method, which removes this bias, by weighting closed claims with weights to better take into account long-term closed claims.

This last model allows the IBNeR to be estimated.

We will propose with the following modeling an innovative approach, to our knowledge, which allows an estimation of the costs of claims not yet manifested increased by the costs of late claims :

5. Individuel model + 3D modeling : Global
   The estimation of the PSNEM could naturally be done with an adapted combination of the individual model and the 3D method, by aging the DOC-Reported triangles with the IBNeRs of the individual model. This makes it possible to couple the benefits of individual modeling with the robust and interpretable character of the projections with the aggregated triangles.
6. Individuel model + 3D modeling : With segmentation
   Also, the segmentation of the method in three dimensions makes it possible to better capture the frequency of claims not yet recorded by breaking down the IBNeR into triangles which are supposed to be more homogeneous.

Finally, the control of the reseves in damages-works, being a guarantee of pre-financing without research of responsibility, should take into account the recoveries and their dynamics. We will therefore propose to estimate a rate of use, on the selected grid and overall.
The recovery rates, changing according to the deformations of the portfolio and the evolution of case law, may change over time. A calculation of an overall rate will not be

specific to these changes since it takes a period beyond the ten-year coverage for a given DOC to become relatively mature. A calculation limited to a selection of DOCs, for each destination segment, would therefore be more relevant. We will also use here the correction of the censoring effect by IPCW.

Results

Backtest 2021

— Selection of the optimal segmentation A sampling-based methodology was followed to select the optimal segmentation. The segmentation selected is the intersection of the **TM x Destination** features.

— Model comparison :

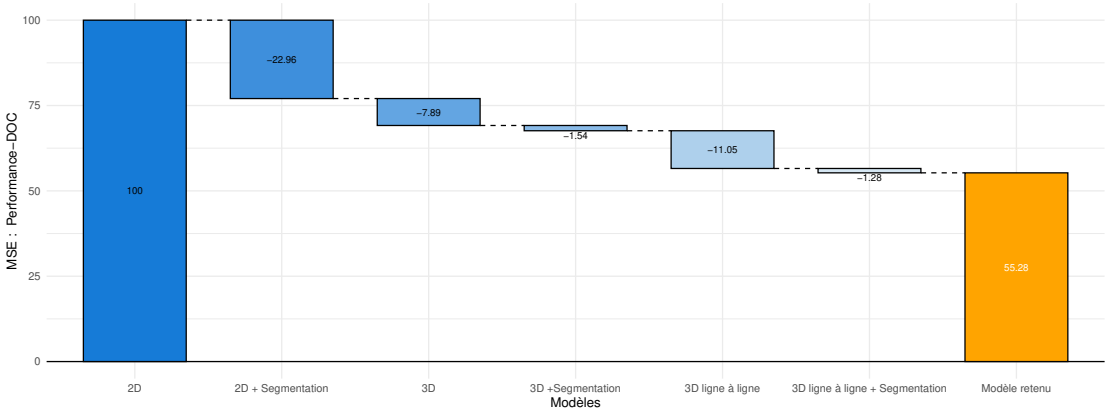

FIGURE 6 – Result of the backtesting : Performance based on the MSE by DOC compared to the 2D model

The combination of models that we introduce seems to be more suitable according to the results of the backtesting. This modeling makes it possible to take the individual specificities of claims by combining it with the robustness of the aggregated triangular projection.

Recovery rate and reserves net from recoveries

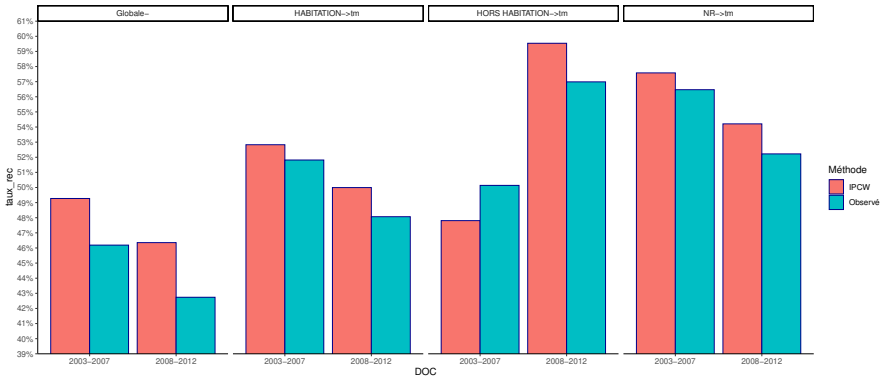

FIGURE 7 – Recovery rate per Strate : Observed-IPCW

Observed rates were estimated by class of destination. They will be able to help in a better estimate of recoveries, if the insurer decides to change underwriting strategy by type of destination.
Assuming an overall recovery rate of 49.2%, the overall reserves net of recoveries, by model, are illustrated above :

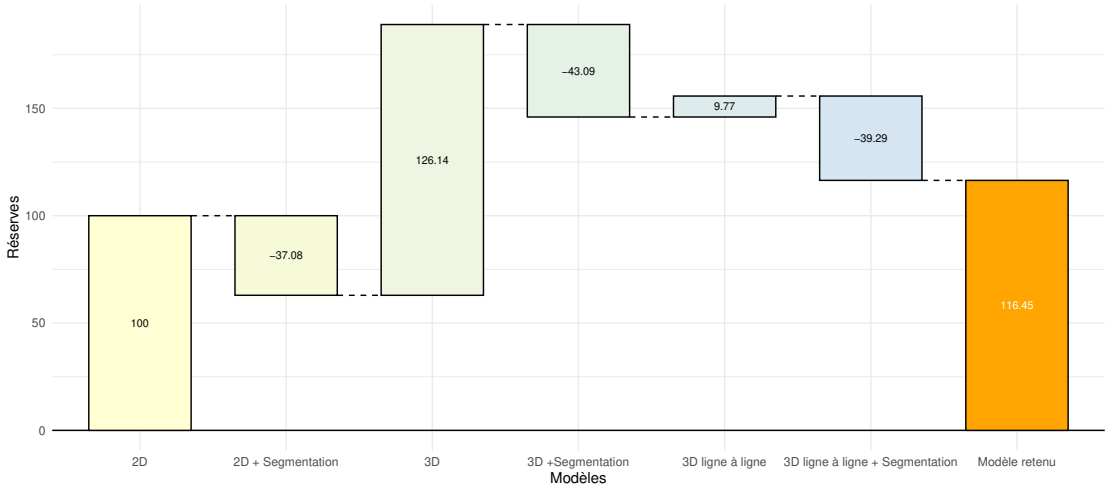

FIGURE 8 – Net reserves of recoveries expressed relative to the 2D model

The models produce a very varied level of reserve compared to the two-dimensional modelling.

Conclusion

The modeling proposed in this dissertation made it possible to exploit the typologies of claims. Several limitations and areas for improvement should be considered.
The backtest carried out is limited to the year 2021 and does not allow us to generalize our conclusions on the relevance of the models. Modeling assumptions were also made, in particular with respect to the restatement of **inflation**, the choice of marginal distributions and the dependence model despite the non-validation of the statistical tests. The data quality step, relating to the pre-processing of the **textual features** of the destination, also shows a major limit given that mixed destinations are grouped together in the class of unclassified destinations, creating additional heterogeneity.
Finally, our work can see a direct use in the different normative approaches to reserve calculations, in particular according to the new IFRS17 framework which encourages segmentation selection according to profitability, which can also have a different assessment depending on the destination of the structures. All this in the hope of clarifying the risks in relation to a branch of insurance that is difficult to control.

Remerciements

Ce mémoire a été réalisé au sein de l'équipe d'études et d'analyses actuarielles IARD d'Allianz France. Que M. Mohamed Zaimi, qui m'a accueilli dans son équipe, trouve ici l'expression de ma respectueuse gratitude.

Je tiens à remercier M. Sébastien Farkas pour avoir accepté d'encadrer ce mémoire, pour la qualité de son suivi continu. Merci pour l'ardeur et la passion mises dans cette aventure. Qu'il soit assuré de ma plus profonde gratitude.
Un grand merci à M. Nicolas Trouilh pour m'avoir accueilli au sein d'Allianz France, pour l'intérêt constant qu'il a porté à ce travail, pour son suivi de ce mémoire et pour les conseils éclairés qu'il m'a prodigués tout au long de cette aventure.
Je remercie amplement M. Bernard Bailleul, qui m'a orienté avec soin dans le choix du sujet de ce mémoire, pour la confiance accordée. Il a suivi avec un intérêt constant la réalisation de ce travail.

Je tiens également à remercier tous les professeurs de l'EURIA, pour la qualité de leurs cours et pour leur suivi continu. Je remercie tout particulièrement M. Franck Vermet, directeur de l'EURIA et M. Rainer Buckdahn, mon tuteur académique.
Il m'est aussi agréable de remercier le corps professoral de l'Institut national de statistique et d'économie appliquée (INSEA), qu'ils trouvent ici ma respectueuse reconnaissance.

Je remercie finalement Mme. Julia Simaku pour son intérêt et ses remarques pertinentes, Mme. Khadija Chehabi et M. Aymane Rbaati pour leur relecture.

À mon père, ma mère et à mes deux frères. Pour tous leurs sacrifices, leur amour et leur soutien tout au long de mes études,

Table des matières

Introduction

La garantie Dommages-ouvrage en assurance construction connait des spécificités particulières. Définie selon un régime décennal, cette garantie est donc à développement long, et nécessite une attention particulière lors de l'estimation des réserves.

On propose donc dans ce mémoire d'illustrer le potentiel de la prise en compte des typologies des sinistres à travers différents modèles de provisionnement afin d'affiner ce dernier.

Afin de poser les bases nécessaires pour effectuer la modélisation souhaitée, on commencera par une première partie dont le but est d'expliquer comment les données permettant de prendre en compte les typologies des sinistres ont été retraitées, et comment les mailles de provisionnement ont été déterminées. Par ailleurs, des rappels sur le provisionnement en trois dimensions, spécifique à l'assurance construction, sont abordés.

Dans un second temps, on procédera à la quantification des IBNeR à travers une modélisation individuelle des sinistres. Pour cela, on proposera une modélisation de l'état des sinistres à la sortie (sans ou avec suite en principal). On étudiera le modèle proposé par [Lopez 2018] qui se base sur une étude de la durée et du coût des sinistres ainsi que sur leur structure de dépendance. Ce modèle est adapté aux données censurées, et part du constat que les coûts des sinistres censurés se trouvent être supérieurs aux coûts des sinistres clos. L'introduction des poids dits IPCW permet de corriger le biais d'une estimation ne se basant que sur les sinistres clos.

Finalement, on proposera une méthodologie combinant les estimations individuelles des IBNeR avec la méthode 3D (agrégée) par année d'enregistrement. On espère capter avec cette idée les bénéfices des deux types de modélisation. Par la suite, une comparaison entre les différentes méthodes est effectuée ainsi qu'une estimation des taux de recours.

Chapitre 1

Assurance construction et provisionnement

1.1 Généralité sur l'assurance construction :

Ce chapitre a pour but d'introduire les généralités concernant l'assurance construction, de bien éclaircir les différentes garanties proposées et leurs spécificités juridiques et enfin de détailler le rôle des différents intervenants.
En toute généralité, l'assurance construction vise à protéger toute personne physique ou morale qui s'engage dans la réalisation des travaux de construction. Les dommages peuvent survenir avant la réception de chantier tel que la dégradation et sont assurés par des garanties spécifiques :

- Garantie de bon fonctionnement : Valide jusqu'à deux ans après la réception des travaux, elle impose à l'entreprise qui a réalisé les travaux de remplacer ou réparer les éléments d'équipements séparables à l'ouvrage (*dissociable*).
- Garantie de parfait achèvement : Valide jusqu'à un an après la réception des travaux, elle impose à l'entreprise qui a réalisé les travaux de réparer le désordre et les malfaçons signalées.

D'autres, par ailleurs, sont relatives aux dommages qui peuvent survenir après la date de réception de chantiers :

- Responsabilité civile décennale (RCD) : Valable 10 ans après la réception des travaux. Elle oblige les constructeurs à réparer les dommages qui apparaissent après la réception des travaux.
- Dommages-ouvrage (DO) : Valable aussi sur une période décennale, son objectif est de préfinancer les travaux de réparation des dommages relevant de la garantie responsabilité civile décennale sans recherche de responsabilité. L'assureur DO, par la suite, fait son recours contre l'assureur RCD.

Nous nous focaliserons dans tout ce qui suit sur les garanties décennales et spécialement sur la dommages-ouvrage qui fait l'objet de ce mémoire.

1.1.1 Contexte juridique :

L'assurance construction est définie par la loi Spinetta introduite en janvier 1978. Bien que les constructeurs des ouvrages soient responsables des dommages afférant à l'ouvrage dans le cadre de la construction :
« Tout constructeur d'un ouvrage est responsable de plein droit, envers le maître ou l'acquéreur de l'ouvrage, des dommages, même résultant d'un vice du sol, qui compromettent la solidité de l'ouvrage ou qui, l'affectant dans l'un de ses éléments constitutifs ou l'un de ses éléments d'équipement, le rendent impropre à sa destination. » Article 1792 du code civil.

L'article spécifie à la fin que :
« Une telle responsabilité n'a point lieu si le constructeur prouve que les dommages proviennent d'une cause étrangère. » Article 1792 du code civil.

Ceci rendait compliqué le préfinancement des travaux de réparation pour le maitre d'ouvrage vu que les démarches d'identification des responsabilités après la survenance étaient longues et fastidieuses avant la loi Spinetta.
L'assurance dommages-ouvrage introduite permet à l'assuré d'obtenir une indemnisation rapide des sinistres qu'il déclare, indépendamment de toute recherche de responsabilité décennale des constructeurs concernés. Elle institue un principe dit à *« double détente »*. Ainsi, l'article L. 242-1 du code des assurances prévoit une procédure rapide d'instruction de la demande de l'assuré, faisant peser sur l'assureur de strictes obligations en termes de délais : *« Lorsqu'il accepte la mise en jeu des garanties prévues au contrat, l'assureur présente, dans un délai maximal de quatre-vingt-dix jours, courant à compter de la réception de la déclaration du sinistre, une offre d'indemnité, revêtant le cas échéant un caractère provisionnel et destinée au paiement des travaux de réparation des dommages. En cas d'acceptation, par l'assuré, de l'offre qui lui a été faite, le règlement de l'indemnité par l'assureur intervient dans un délai de quinze jours. »* Article L242-1 du code des assurances.

Le Processus de construction et les principaux acteurs :

Une opération de construction se définit comme étant l'ensemble des travaux de réalisation d'un ou plusieurs ouvrages, exécutés entre les dates d'ouverture du chantier, communément abrégée en « DROC » ou « DOC », et de réception de cette opération. [1]
Le schéma suivant trace les étapes de déroulements des 4 garanties obligatoires de l'assurance construction :

1. *Source : la jurisprudence, Cour d'appel de Paris, 15 mai 2013, n° 09/16662*

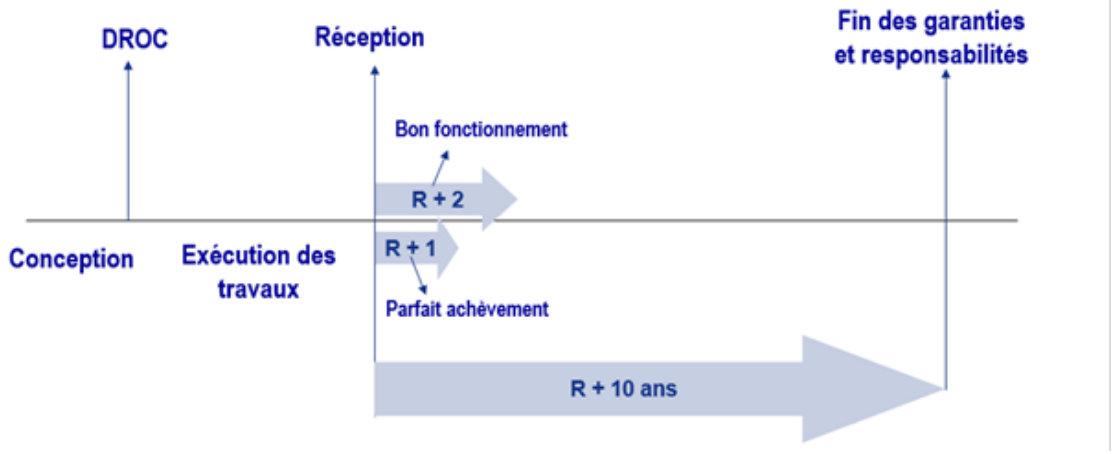

FIGURE 1.1 – Étapes de déroulement des garanties constructions

Différents intervenants sont à distinguer :

— **Maitre d'ouvrage** : Il s'agit de la personne physique ou morale pour le compte de laquelle les travaux sont exécutés [2]. C'est le donneur d'ordre, au moins pour ce qui concerne l'engagement initial. Il s'agit par exemple de particuliers, de vendeurs ou de promoteurs qui rénovent ou construisent des ouvrages afin de les revendre, des collectivités territoriales, des entreprises et autres.
— **Maitre d'œuvres** : Chargé par le maitre d'ouvrage de l'exécution des travaux, il s'agit de la personne, physique ou morale qui vielle à la bonne exécution des travaux et est chargée de concevoir l'ouvrage en respectant le cahier des charges. Selon le type de travaux, le maître d'œuvre peut être un architecte, un ingénieur, un bureau d'études et parfois le maître d'ouvrage peut lui-même être maitre d'œuvre.
— **Les entrepreneurs** : Les entrepreneurs représentent quant à eux l'ensemble des entreprises participant à la construction et devant exécuter les travaux prévus au marché conformément aux règles de l'art.

1.1.2 Spécificité de la Dommages-Ouvrage :

Gestion de temps :

La DO comme on l'a introduite a pour objectif de préfinancer le paiement des travaux, sans recherche de responsabilité et d'accélérer le paiement de cette indemnisation. Le délai de la réponse de l'assureur est de 90 jours après la réception de la déclaration du sinistre où il doit proposer une offre d'indemnisation. L'assureur n'a après qu'un délai de 15 jours à compter de la réception de l'indemnité pour la payer si l'assuré accepte. [3]

2. Norme Afnor NF-P03001
3. Article A. 243-1, Annexe II - B – OBLIGATIONS DE L'ASSUREUR EN CAS DE SINISTRE – 3

Les sanctions du non-respect de ce délai sont définies dans l'article L.242-1 du code des assurances qui stipule que *« l'assuré peut, après l'avoir notifié à l'assureur, engager les dépenses nécessaires à la réparation des dommages. L'indemnité versée par l'assureur est alors majorée de plein droit d'un intérêt égal au double du taux de l'intérêt légal. ».*

Destination des ouvrages :

La DO est applicable après l'apparition de dommages après la réception des travaux, résultants d'un vice caché au moment de la réception et surtout sous un degré de gravité requis. Une possibilité d'exonération est parfois possible dans le cas d'une cause étrangère, c'est-à-dire à un fait d'un tiers ou d'une force majeure par exemple.

Le degré de gravité est précisé dans l'article 1792 du code civil sur deux termes : les dommages doivent compromettre la solidité de l'ouvrage ou le rendent impropre à sa destination en affectant l'un de ses éléments constitutifs ou d'équipements. On parle également d'**impropriété à la destination.**

Ce critère, bien qu'il soit encadré, reste subjectif et entraîne différentes interprétations de la jurisprudence. Un ouvrage à destination habitation se voit donc selon ses critères différent d'un ouvrage à destination industriel ou commercial et un désordre pourrait être considéré éligible à la DO pour une destination et non éligible pour une autre.

La destination de l'ouvrage diffère parfois entre ce qui est contractuel et ce qui est subjectif. La notion d'impropriété reste à l'appréciation des juges.
L'exemple ci-dessous est très claire sur cette subjectivité : [4]

« On peut ainsi lire, dans un arrêt de la troisième chambre civile de la Haute juridiction, qu'une cour d'appel « a retenu, à bon droit, que la destination de l'immeuble à prendre en compte était celle prévue initialement entre le maître de l'ouvrage et les concepteurs. En outre, dans une hypothèse où un immeuble bénéficiait à la fois d'un chauffage traditionnel et d'un chauffage solaire, la Cour de cassation a approuvé les juges du fond d'avoir décidé que l'incapacité de l'installation solaire à fonctionner rendait le bâtiment impropre à sa destination telle qu'elle avait été prévue et commercialisée. »

4. THIOYE Moussa, « La responsabilité décennale à raison de l'impropriété de l'immeuble à sa destination », Droit et Ville, 2015/2 (N° 80), p. 53-66. DOI : 10.3917/dv.080.0053. URL : https ://www.cairn.info/revue-droit-et-ville-2015-2-page-53.htm

1.2 Introduction au provisionnement en assurance construction

1.2.1 Les principes du provisionnement en assurance non-vie

L'assurance est définie par un engagement de la part d'une personne morale (assureur) à réaliser une prestation au profit d'une autre personne, physique ou morale, en cas de réalisation d'un risque défini *a priori*, et ce, en échange d'un encaissement d'une prime. Le coût réel que l'assureur devrait payer est donc inconnu au moment de la souscription du contrat d'assurance. On parle de cycle de production inversé. Il est donc primordial que les assureurs identifient et valorisent les risques sous-jacents aux contrats en amont pour faire face à leurs engagements. Pour les garanties décennales de l'assurance construction, ce problème s'aggrave puisque ces derniers prennent effet à la date de réception de l'ouvrage pour une durée de dix ans et doivent être souscrites avant le début du chantier. Pour ces raisons, les assureurs comptabilisent dans le passif de leur bilan différentes provisions.

Exemple de provisions en IARD

Certaines provisions sont réglementaires. En effet, Le Code des Assurances dans l'article R331-6 définit les « Provisions Techniques » en assurance non-vie. Trois de ces provisions, liées à la suite de nos travaux, sont intéressantes à retracer :

- PSAP (Provision pour sinistres à payer) : « Valeur estimative des dépenses en principal et en frais, tant internes qu'externes, nécessaires au règlement de tous les sinistres survenus et non payés, y compris les capitaux constitutifs des rentes non encore mises à la charge de l'entreprise » [5].
- PPNA (Provision pour primes non acquises) : « Provision, calculée selon les méthodes fixées par arrêté du Ministre de l'Économie, destinée à constater, pour l'ensemble des contrats en cours, la part des primes émises et des primes restant à émettre se rapportant à la période comprise entre la date de l'inventaire et la date de la prochaine échéance de prime ou, à défaut, du terme du contrat » [6].
- PSNEM (Provision pour sinistres non encore manifestés) : Elle correspond à une provision pour sinistres non encore manifestés pour les garanties Dommages-ouvrage et Responsabilité civile décennale et qui devraient se manifester avant l'expiration de la période de couverture décennale.

En une date de calcul donnée, des sinistres sur lequels on est engagé sont survenus et la PSAP est la provision relative aux règlements futurs de ces sinistres. La PPNA est par ailleurs constatée pour les contrats avec une période de couverture restante, à la date de calcul, puisqu'ils pourront encore être sinistrés.

5. Selon l'article R331-6 du code des Assurance
6. Selon l'article R331-6 du code des Assurance

Remarque : La distinction entre la PPNA et la PSNEM est importante. En effet, bien que des garanties en assurance construction s'étalent sur une période décennale, on considère que l'événement d'origine d'un sinistre se manifestant après la date de réception des travaux a eu lieu durant le chantier. Le risque qu'on couvre est donc rattaché à la durée du chantier et non à la période décennale. On estime que la période de couverture de ces contrats est d'environ deux ans, ce qui correspond à la durée moyenne de chantier.

La PPNA est donc équivalente aux réserves constatées pour les deux DOC récentes. Ceci impose à tracer des correspondances entre les différentes normes assurantielles, à savoir les normes IFRS17, Solvabilité II et la norme française. Plus de détails seront apportés dans la fin de ce chapitre.

Cycle de vie d'un sinistre en DO :

Après la date de réception du chantier, la durée décennale de la garantie commence. Le maître d'œuvre pourra déclarer à l'assureur DO les dégâts remarqués qui, comme spécifié dans le chapitre précédent, doivent porter atteinte à la solidité de l'ouvrage ou le rende impropre à sa destination. L'assureur DO, après réception de la déclaration, désigne un expert et doit proposer à l'assuré une indemnité avant un délai de soixante jours. Des recours sont généralement attendu de la part de l'assureur RCD après que le dernier paiement de l'assureur DO soit effectué

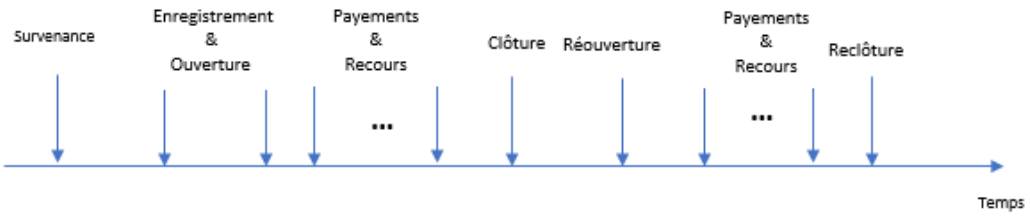

FIGURE 1.2 – Cycle de vie d'un sinistre

Au moment de la déclaration d'un sinistre, les gestionnaires de sinistres lui évaluent une charge sur la base de rapport d'experts, des grilles forfaitaires ou encore sur la base des factures fournies par l'assuré. Ils calculent une charge dossier-dossier, qui est le montant directement lié aux données connues au moment de l'évaluation. Cette estimation ne prend pas en compte le niveau de l'aggravation future du sinistre. Périodiquement, les gestionnaires réadaptent leurs estimations aux nouvelles informations à leurs dispositions. L'assureur en parallèle effectue des paiements en fonction de cette évaluation. Le sinistre est considéré clôturé après la connaissance de la charge réelle du sinistre et après son paiement intégral. Des réouvertures peuvent avoir lieu dans des cas spéciaux.

Cela étant dit, l'assureur, devant constituer des provisions pour faire face à ses engagements futurs, a besoin d'estimer, en plus des charges dossier-dossier, le montant qui

permet d'écouler l'ensemble des sinistres à l'ultime, c.-à-d. à leurs clôtures. En plus des sinistres connus, l'assureur a besoin d'estimer le coût des sinistres survenus, mais non encore déclarés, et spécialement en Dommages-ouvrage, les sinistres survenus non encore manifestés.

En Dommages-ouvrage, on s'attend naturellement à des recours. Les gestionnaires estiment avec le même principe des prévisions de recours à encaisser (PRAE) et leur projection à l'ultime est nécessaire.

L'un des rôles clés des actuaires est d'estimer la quantité de charge qu'il faut ajouter à la charge dossier-dossier pour obtenir la charge ultime. Ce montant est plus communément appelé IBNR, pour *Incurred but not reported.* Il se décompose en un montant d'IBNeR (*Incurred but not enough reported*), qui représente l'aggravation future de la charge des sinistres connus, non prise en compte par le gestionnaire de sinistres, et en un montant d'IBNyR (*Incurred but not yet reported*), qui correspond à la charge finale prévisible des sinistres survenus, mais non encore connus par l'assureur.

$$\text{Charge Ultime} = \text{Charge Dossier-Dossier} + \text{IBNR}$$

$$\text{IBNR} = \text{IBNeR} + \text{IBNyR}$$

On expliquera en détail le calcul réglementaire de la PSNEM dans la fin de ce chapitre.

1.2.2 Provisionnement agrégé

On notera dans cette partie $C^k_{i,j}$: la charge totale du sinistre k survenu à l'année i et au développement j, le développement étant égal à la différence entre l'année d'observation et l'année de survenance. Comme il a été introduit précédemment, les gestionnaires évaluent une charge au moment de la connaissance du sinistre. Cette information évolue au fur et à mesure selon l'état du sinistre, le retour de l'expertise, les règlements effectués par l'assureur et l'état judiciaire si le sinistre est contentieux.
À une date d'arrêté t, l'assureur doit constituer des provisions pour les sinistres survenus et son estimation se fait naturellement à travers la dynamique des flux historiques observés, qu'il s'agit de charges, de règlements ou encore de recours.
En assurance non-vie, une façon d'analyser l'évolution temporelle des charges de sinistres survenus est de les agréger par année de survenance, *i.e* $C_{i,j} = \sum_k C^k_{i,j}$.

La partie des charges connues est représentée par une forme triangulaire puisqu'on connaît que les charges telles que $(C_{i,j})_{i+j\leq 2022}$. En prenant l'année 2010 comme année de référence :

Survenance\Développement	0	1	...	11	12
2010	$C_{2010,0}$	$C_{2010,1}$	...	$C_{2010,11}$	$C_{2010,12}$
2011	$C_{2011,0}$	$C_{2011,1}$	...	$C_{2011,11}$	
⋮	⋮	⋮	⋰		
2021	$C_{2021,0}$	$C_{2021,1}$			
2022	$C_{2022,0}$				

Pour estimer le montant d'IBNR, les actuaires utilisent classiquement des triangles regroupant des typologies de risques homogènes en développement, que ce soit selon une séparation par garanties ou selon des mailles plus fines, et essayent de les projeter jusqu'à maturité en utilisant des modèles statistiques.
Pour une estimation des IBNeR, on pourrait penser à tracer un triangle par année d'enregistrement [7]. La projection de ce dernier fournira une estimation des flux de l'aggravation des sinistres, sans prendre en compte les sinistres tardifs.

Triangles en construction

La nature décennale de l'assurance construction nous oblige à rattacher les sinistres à leurs dates d'ouverture de chantier. Ceci permet aussi de rapporter les sinistres aux primes qui les couvrent par DOC.

Avec le triplet année d'ouverture de chantiers, année de survenance, année d'observation, différent types de triangles pourront être envisagé selon ce que l'on cherche à estimer. On parle de *triangle 3D*.

Comme vu précédemment, estimer la partie inférieure d'un triangle selon la vision Survenance-Développement permet d'estimer les IBNR. Une distinction entre les DOC permettant de tracer un triangle Survenance-Développement par DOC est également possible (1.1).

Un triangle selon la vision DOC-Développement dessinera l'évolution des flux de charges sans considérer la survenance. Une estimation de la partie inférieure inconnue nous permettra de calculer les réserves relatives à la fois aux sinistres survenus et aux sinistres non encore manifestés rattachés aux DOC.
Finalement, un triangle croisant la vision DOC-Survenance permettra de projeter à l'ultime le montant de sinistres non encore manifestés par DOC. Ce triangle change de vision en vision. En effet, Une aggravation d'un sinistre en vision $n+1$ changera la même cellule du triangle vu en vision n. Il faudra donc vieillir la partie supérieure de ce triangle à l'ultime avant de le projeter.

Notons également qu'en séparant par DOC, le nombre de développements par année d'observation ou par rapport aux sinistres survenus décroit. Une séparation selon les DOC ne nous permettra donc pas d'estimer les flux futurs sans considérer la dynamique de sinistralité des DOC antérieurs :

7. On ne fait pas de différence dans la suite entre la date d'enregistrement, la date de déclaration et la date d'évalutation

DOC	Survenance\Développement	0	1	...	J-1	J
0	0	$C^0_{0,0}$	$C^0_{0,1}$		$C^0_{0,J-1}$	$C^0_{0,J}$
	1	$C^0_{1,0}$	$C^0_{1,1}$	$\ddots$	$C^0_{1,J-1}$	
	$\vdots$	$\vdots$	$\vdots$	$\ddots$		
	J-1	$C^0_{J-1,0}$	$C^0_{J-1,1}$			
	J	$C^0_{J,0}$				
1	1	$C^1_{1,0}$	$C^1_{1,1}$		$C^1_{1,J-1}$	
	$\vdots$	$\vdots$	$\ddots$			
	J	$C^1_{J,0}$				
$\vdots$	$\vdots$		$\ddots$			
J-1	J-1	$C^{J-1}_{J-1,0}$	$C^{J-1}_{J-1,1}$			
	J	$C^{J-1}_{J,0}$				
J	J	$C^J_{J,0}$				

TABLE 1.1 – Triangle Survenance-Développement par DOC

La durée de développement des sinistres sera naturellement supérieure à 10 ans vu la nature décennale de la garantie Dommages-ouvrage. Également, le nombre de survenances par rapport à la DOC en première ou en deuxième année est souvent très faible. Ceci s'explique par le fait que la durée de construction des chantiers est en moyenne de deux ans et que les sinistres pris en charge avant la réception des travaux sont marginaux.

1.2.3 Méthode d'estimation des réserves

Comme présenté auparavant, les actuaires estiment à travers des modèles statistiques les parties non connues relatives à l'évolution future des écoulements de charges, de nombres de sinistres, de règlements ou encore de recours.
On notera dans la suite de ce chapitre $c_{i,j}$, la variable décrivant ces agrégats financiers pour l'année de survenance i au développement j. On notera également $C_{i,j}$ le cumul des incréments $c_{i,j}$ par année de survenance i. c.-à-d. : $C_{i,j} = \sum_{k=0}^{j} c_{i,k}$

L'algorithme Chain-Ladder

La méthode de Chain-ladder est une méthode simple, robuste et directement interprétable, servant à projeter les triangles d'écoulement. Il s'agit certainement de l'algorithme le plus connu en provisionnement non-vie.

La méthode Chain-Ladder repose sur l'hypothèse d'un écoulement proportionnel du

triangle de développement. Elle suppose donc l'existence de facteur $(f_j)_{2,\dots,n}$ tel que :

$$\frac{C_{1,j+1}}{C_{1,j}} \approx \frac{C_{2,j+1}}{C_{2,j}} \approx \dots \approx \frac{C_{n,j+1}}{C_{n,j}} \approx f_j$$

Ces facteurs sont estimés par :

$$\hat{f}_j = \frac{\sum_{i=1}^{n-j+1} C_{i,j}}{\sum_{i=1}^{n-j+1} C_{i,j-1}} \quad 2 \leq j \leq n-1$$

Le développement des charges futures (triangle inférieur) est donc estimé d'une manière récursive. Le montant de réserves pour l'année i se calcule par la suite comme différence entre le montant ultime et le montant connu pour de l'année i (diagonale du triangle) :

$$\hat{C}_{i,j} = C_{i,n-i} \prod_{k=n-i}^{j-1} \hat{f}_k, \qquad \hat{R}_0 = 0, \qquad \hat{R}_i = \hat{C}_{i,n} - C_{i,n-i}, \quad i \in \{0,\dots,n\}$$

Cadre mathématique de la méthode Chain-Ladder

La méthode décrite ci-dessus a été historiquement vue comme un simple algorithme qui complète la partie inconnue d'un triangle sans fondement mathématique. Thomas Mack en 1993 introduit un cadre stochastique en se basant sur deux hypothèses dans son article [Mack 1993]

1. Les montants cumulés des sinistres encourus $C_{i,j}$ d'années de survenance différentes sont indépendants.
2. $\exists \lambda_2,\dots,\lambda_n > 0$, $\sigma_2^2,\dots,\sigma_n^2 > 0$ tel que

$$\mathbb{E}(C_{i,j}|C_{i,1},\dots,C_{i,j-1}) = \mathbb{E}(C_{i,j}|C_{i,j-1}) = \lambda_{j-1} C_{i,j-1}$$
$$\mathbb{V}\mathrm{ar}(C_{i,j}|C_{i,1},\dots,C_{i,j-1}) = \mathbb{V}\mathrm{ar}(C_{i,j}|C_{i,j-1}) = \sigma_{j-1}^2 C_{i,j-1}$$

   pour tout $i = 1,\dots,n$ et $j = 2,\dots,n$.

Mack montre avec ces hypothèses que l'estimateur f_j définit auparavant est un estimateur sans biais et non corrélé.

Un autre cadre stochastique a été introduit par *Renshaw et Verrall* en 1998 à travers le modèle de Poisson surdispersé avec une fonction de lien logarithmique, aboutissant au même niveau de réserves que la méthode Chain-Ladder.

Le modèle se définit comme :

1. Les incréments $c_{i,j}$ sont indépendants.
2. $c_{i,j}$ suit une loi de poisson surdispersé avec :
   — $\mathbb{E}(c_{i,j}) = \mu_{i,j} = \exp(\mu + \alpha_i + \beta_j)$
   — $\mathbb{V}\mathrm{ar}(c_{i,j}) = \phi \mu_{i,j}$
3. $\sum_{j=1}^{j=n} \exp \beta_j = 1$

pour $\alpha_1 \dots \alpha_n$, $\beta_1 \dots \beta_n$ et ϕ des constantes positives. Plus de détails sont présents dans l'article [Renshaw and Verrall 1998]

1.2.4 Un provisionnement selon différentes normes

L'objectif *in fine* des travaux de provisionnement est d'estimer le montant de réserves qui permettent de faire face au niveau de sinistralité ultime. Les diverses normes actuarielles ou comptables sont différentes dans leurs appréciations du niveau du risque et dans la forme de présentation de leur reporting. La norme comptable française (*French GAAP)* est centrée sur l'avoir et le patrimoine de l'entreprise et est conçue en intégrant les contraintes fiscales dans son esprit. La norme provisoire IFRS4 ou encore la norme IFRS17 sont orientées plutôt par rapport à la réalité économique, l'objectif est de pouvoir comparer l'exposition au risque et la situation financière des acteurs du marché, et ce, d'une façon transparente. Certaines entreprises d'assurance européennes sont également soumises à la norme prudentielle Solvabilité II qui exige un calcul économique des provisions techniques, un niveau de fond propres dépendant de la nature des risques encourus, un suivi et une surveillance des risques et exige finalement un reporting particulier et une gouvernance spéciale.
Ces différentes normes, de par leurs multiples approches, diffèrent dans l'estimation des provisions, notamment dans le cas de la PSNEM qui va nous servir comme exemple :
On rappelle que la PSNEM est la provision pour sinistres non encore manifestés.

Exemple de la PSNEM en assurance construction :
La norme locale présente le calcul de la PSNEM selon des coefficients de manifestation définies dans le code des assurances :

" Pour effectuer l'estimation mentionnée au 2° de l'article R. 331-17, les entreprises calculent, pour chaque exercice d'ouverture de chantier, séparément pour les garanties décennales de responsabilité civile et pour les garanties décennales de dommage aux ouvrages, l'ancienneté n des chantiers ainsi que les montants A_n et B_n, définis comme suit :
n = différence de millésime entre l'exercice sous inventaire et l'exercice d'ouverture des chantiers ;
A_n = coût total, estimé dossier par dossier, des sinistres afférents aux garanties décennales d'assurance construction délivrées pour des chantiers d'ancienneté n et qui se sont manifestés jusqu'à la date de l'inventaire, diminué des recours encaissés ou à encaisser ;
B_n = montant des primes émises et des primes restant à émettre, nettes des primes à annuler et des frais d'acquisition, afférent à ces mêmes garanties.
L'estimation des sinistres non encore manifestés, effectuée séparément pour les garanties décennales de responsabilité civile et pour les garanties décennales de dommage aux ouvrages, est égale au plus élevé des deux montants M_{S_n} et M_{P_n} suivants :
$M_{S_n} = a_n \times A_n$;
$M_{P_n} = b_n \times B_n$,
a_n et b_n prenant les valeurs suivantes :

n	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**
a_n	0	0	3,4	2	1,4	1	0,7	0,5	0,35	0,25	0,20	0,15	0,10	0,05
b_n	1	1	0,95	0,85	0,75	0,65	0,55	0,45	0,35	0,25	0,20	0,15	0,10	0,05

"

La vision économique de l'entreprise et son niveau de risque sont peu présents dans cette estimation.
Dans le cadre de la norme prudentielle Solvabilité II, les provisions techniques non-vie sont séparées en trois parties :

— *Best Estimate* de sinistres : une évaluation économique des coûts de sinistres survenus.
— *Best Estimate* de primes : une évaluation économique des coûts de sinistres non survenus pour lesquels on est engagés à la date de l'évaluation.
— Marge pour risques : mesure l'incertitude liée au calcul de la meilleure estimation des provisions techniques. Différentes approches sont utilisées pour la mesurer.

Pour le cas de la PSNEM selon Solvabilité II, l'ACPR a indiqué qu'il s'agit d'une provision de sinistres [8]. Ainsi, la différence entre la composante sinistre et prime du *Best Estimate* dans le cas de l'assurance construction est que la première devient relative aux coûts des sinistres futurs, survenus ou non, des polices déjà souscrites. Le *Best Estimate* de prime se définit dans ce cas comme une évaluation économique primes relatifs aux contrats souscrits avant la date de l'évaluation.

Pour la nouvelle norme comptable IFRS17, et dans le cas d'une valorisation du passif selon le modèle simplifié dénommé PAA [9], les provisions techniques diffèrent également entre une partie relative aux sinistres survenus appelée LIC (*Liability for Incurred Claims*), une partie relative aux sinistres non survenus des contrats pour lesquels on est engagés à la date d'évaluation appelée LRC (*Liability for Remaning Coverage*) et une composante nommée RA (*Risk Adjustement*) relative à une mesure de l'incertitude d'une évaluation économique sans marge de prudence.

La correspondance avec l'évaluation du passif selon la norme prudentielle Solvabilité II est immédiate. Cependant, quelques différences de fond sont existantes. Par exemple, elles n'ont pas la même définition de la frontière de contrats, servant à tracer la différence entre les composantes LIC et LRC.

En toute rigueur, la valorisation de la PSNEM selon la norme IFRS17 demande de distinguer une partie LIC et une partie LRC. Une méthode de les séparer après une estimation des réserves pour chaque DOC, faite par exemple après la projection d'un triangle DOC-Développement, serait de considérer qu'un pourcentage des réserves des deux dernières DOC à la date de l'évaluation est relatives aux durées de chantiers restant à couvrir.

Ceci vient du fait que le risque générant le sinistre, est rattaché à la période de la

8. Source : `https://www.institutdesactuaires.com/global/gene/link.php?news_link=2016110706_2016133810-npa3-1.pdf&fg=1`
9. PAA : *Premium Allocation Approach*

construction des travaux et que c'est la manifestation qui est observée par la suite. Aussi, la durée moyenne des travaux de construction ne dépasse pas généralement deux ans [10]. Elle est en moyenne de deux ans chez Allianz.

La vision de ce mémoire, que ce soit par rapport à nos travaux sur la maille d'analyse ou par rapport à la modélisation individuelle des sinistres, s'inscrit dans le cadre d'une estimation reflétant une vision économique et non pas dans le cadre d'une approche normative ou comptable.

10. Différentes selon le type de construction : Par exemple, elle est entre 11 et 13 mois pour les maisons individuelles et entre 16 et 23 mois pour les logements collectifs, selon les statistiques de l'année 2010 du Commissariat général au développement durable. Source : `https://www.statistiques.developpement-durable.gouv.fr/sites/default/files/2018-10/LPS%20122%20Dur%C3%A9e%20de%20construction%20des%20logements.pdf`

Chapitre 2

Analyse des données

2.1 Présentation des données

La préparation des données constitue une étape principale pour la compréhension des liaisons entre les variables et la détection des sources d'aberrances. Cette étape est nécessaire avant toute tentative de modélisation pour une meilleure orientation vers des modèles répondant aux enjeux identifiés et pour une vérification ultérieure avec les constats de cette analyse.
Nous présenterons dans ce chapitre une analyse des variables utilisées dans cette étude, les retraitements et les hypothèses prises pour la construction des données finales et par la suite une analyse textuelle de destination des ouvrages.

2.1.1 Périmètre des données

Les données à notre disposition sont relatives au produit DO obligatoire d'Allianz France, commercialisées en France Métropolitaine. Différentes bases de données sont utilisées :

- Bases sinistres annuelles : commençant de la vision 31/12/2003 jusqu'à la vision 31/12/2021. Un sinistre ouvert et clos en 2006 ne sera pas présent dans les bases de vision supérieure à la vision 31/12/2007 et sera présent à la vision 31/12/2006.
- Base portefeuille : représentant la vision au 31/12/2021 des polices DO sinistrées. Il s'agit d'une base historisée, contrairement aux bases sinistres. Elle nous servira à lier les sinistres avec leurs caractéristiques statiques tels que la destination des ouvrages. *On ne s'intéressera pas dans ce qui suit aux mouvements de portefeuille ou à l'analyse de l'exposition des polices.*

Pour les données sinistres, on récupère également les sinistres enregistrés avant le 31/12/2002 sans information sur la première estimation de la charge ou sur le premier règlement effectué. Ces données représentent donc une troncature à gauche. Différentes techniques peuvent être utilisées pour les inclure dans les modèles de provisions, que ce soit agrégés [1] ou individuels. Nous ferons le choix de se limiter aux sinistres de DOC

1. Des techniques d'imputations de données pourront être utilisées en introduisant des variables

supérieure à 2003 pour contourner cette problématique.

Hypothèses Prises :

— On considère que la durée de vie d'un sinistre clos est la différence entre sa date de première ouverture et sa date de dernière clôture : Un sinistre peut être réouvert après l'avoir déclaré comme étant clos. Nos bases présentent ces deux informations. Pour la modélisation ligne à ligne des IBNeR, nous jugeons que cet impact n'est pas significatif puisqu'on entraine nos modèles sur les sinistres ouverts avant fin 2019. Un sinistre réouvert et en vie entre temps ne sera pas pris donc en considération. Le chapitre cinq détaille cette analyse. Pour le choix de la maille de projection, ceci n'est pas impactant puisque les données sont agrégées sans aucune modélisation de la durée de vie des sinistres.

— Données manquantes :

  — On manquait d'information par rapport à la nature de la garantie DO sinistrée sur 248 sinistres. On a fait l'hypothèse qu'elle représente la garantie obligatoire.

  — La destination des ouvrages sera classée en catégorie *Non renseigné* lorsque le numéro de police, considéré comme clé primaire lors de la jointure avec la base portefeuille, n'est pas identifiable.

2.1.2 Mise en *As-If*

La modélisation des flux financiers, sur un périmètre de données s'étalant sur une longue durée, représentée également dans le cadre des garanties décennales, devrait se faire après une considération de la dépendance temporelle causée par l'inflation. La méthode Chain-Ladder, servant à la projection des triangles, est sensible à l'inflation vu qu'elle perturbe l'hypothèse d'indépendance par année de DOC (ou Survenance, selon le triangle considéré) en créant une dépendance temporelle par année calendaire représentée par les diagonales du triangle.

Nous choisirons donc de retraiter l'inflation en déflatant les flux de paiements, $i.e$ en considérant que tous les flux de paiements ont été réglés en euro 2021. Un calcul des réserves devrait se faire donc après cette étape en projetant l'inflation au futur.
Dans la suite de nos travaux, La projection des triangles permettant ce calcul se fera à l'euro vu à fin 2021.
Notre choix vient du fait que nos travaux ont principalement pour but d'affiner la qualité de l'estimation des réserves en intégrant des informations sur les sinistres tel que la destination des ouvrages ou de les modéliser selon une dynamique de provisionnement individuelle.

En plus, l'environnement actuel ne permet pas une modélisation statistique pertinente du niveau d'inflation sans avis d'expert et sans considération macroéconomique forte.

Les indices BT

explicatives de l'évolution des provisions pour estimer la partie tronquée des triangles *(mémoire de Mohammed Amine EL AIDOUNI)*

L'inflation dans le secteur de la construction est représentée par exemple par l'évolution des coûts des matières premières, du coût du travail ou aussi des coûts de transport. L'INSEE publie mensuellement des indices (BT) qui représentent l'inflation sur différents secteurs de sous-activité. La figure suivante représente l'évolution de l'indice du bâtiment - tout corps d'état (BT01), l'indice du bâtiment - terrassement (BT02), l'indice du bâtiment - Peinture, tenture, revêtements muraux (BT46), l'indice de bâtiment Électricité (BT47) et de l'indice de bâtiment - Ascenseurs (BT48), en base 2010.

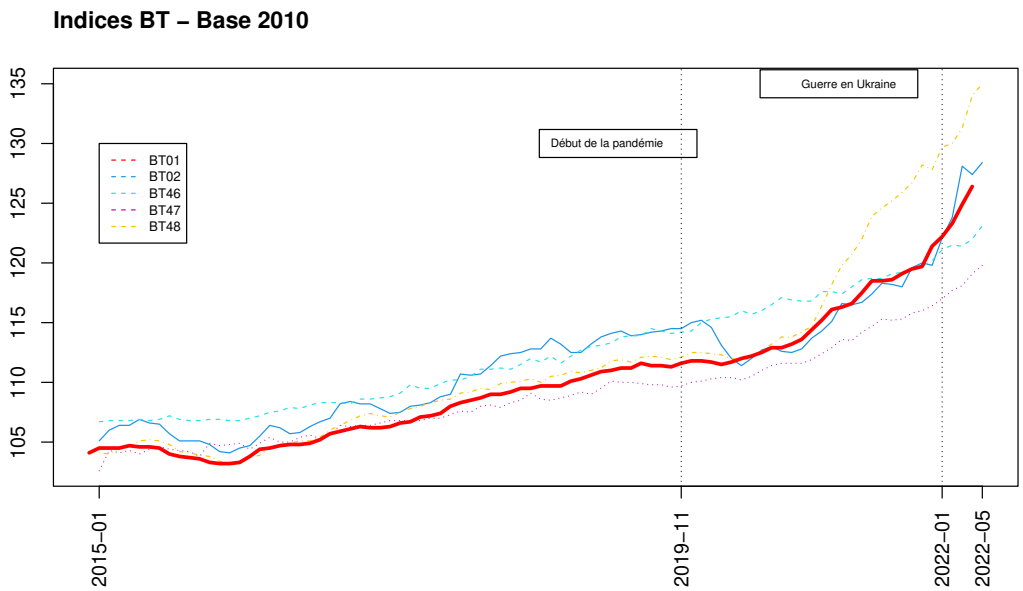

FIGURE 2.1 – L'évolution de quelques indices BT de janvier 2015 à mai 2022

L'indice général BT01 représente l'évolution globale des coûts de construction. Il est calculé selon la pondération suivante :

— Salaires et charges : 43 %
— Matériaux : 32 %
— Équipements : 4 %
— Frais de transports : 3 %
— Frais d'énergie : 3 %
— Frais divers : 15 %

La figure montre une augmentation considérable de la valeur du BT01 à partir de l'année 2020. Cette surinflation continue après la guerre en Ukraine au début de l'année 2022. Les orientations géopolitiques pourront expliquer une partie de cette évolution vu la montée du prix du gaz et de l'électricité, corrélés positivement avec l'évolution générale des prix et des frais de transport et d'énergie, composantes du BT01.
L'inflation du secteur de construction se comporte différemment selon les sous-activités : le BT48 représentant l'indice de bâtiment - ascenseurs voit une évolution plus forte après

la crise pandémique.

Pour retraiter l'ensemble des flux de paiements de nos données, nous choisirons l'indice BT01 comme déflateur :
Un règlement effectué à l'année i est égal, vu en 2021, à :

$$\text{RéglésASIF}_{annee_i} = \text{Réglé}_{annee_i} \times \frac{BT01_{2021}}{BT01annee_i}$$

Une modélisation complexe

Que ce soit par rapport à une inflation sectorielle ou à l'inflation générale des prix des biens et des services, la modélisation de l'inflation reste une étape complexe nécessitant une maîtrise globale de tous les effets de l'environnement économique. L'explication de l'inflation devrait en toute rigueur se faire sur une base d'analyse macroéconomique pertinente.
La projection de l'inflation future nous pousse à étudier si la surinflation qui commence à se dessiner est un phénomène structurel ou conjoncturel.

L'approche de la modélisation dépend de cet avis : les modèles stochastiques utilisés dans le cadre des sujets de gestion actif-passif supposent un retour à la moyenne du type *Vasicek* (1977) :

$$di_t = k(\mu - i_t)dt + \sigma dW_t$$

où i_t représente le taux d'inflation, μ le niveau vers lequel le taux d'inflation converge au long terme, σ et $(W_t)_t$ représentent respectivement la volatilité et le mouvement brownien de la dynamique stochastique. Les modèles macroéconomiques, comme les modèles d'équilibre, intègrent des variables économiques pouvant influer sur le niveau d'inflation générale et décrivent une convergence du niveau de prix en fonction de la croissance économique ou de la demande.

D'autres théories, telle que la courbe de Philips, mise en évidence en 1958 et qui présente une relation inverse entre le taux d'inflation et le taux des salaires nominaux, sont plus simples à analyser, mais ne permettent pas une interprétation sur le long terme.

La projection de l'inflation future étant un sujet complexe à part entière, nous ne la retraiterons pas dans le cadre de ce mémoire. Les différents calculs et projections présentés dans ce mémoire sont donc vus à fin 2021

2.1.3 Exploration de données

Analyse des covariables

Une exploration des variables sera proposée dans cette partie. Le retraitement du champ textuel de la destination des ouvrages, supposé ici effectué, sera analysé dans la section suivante. On a proposé de la représenter selon deux variables : une segmentation selon trois classes et une autre selon sept classes.

Destination :

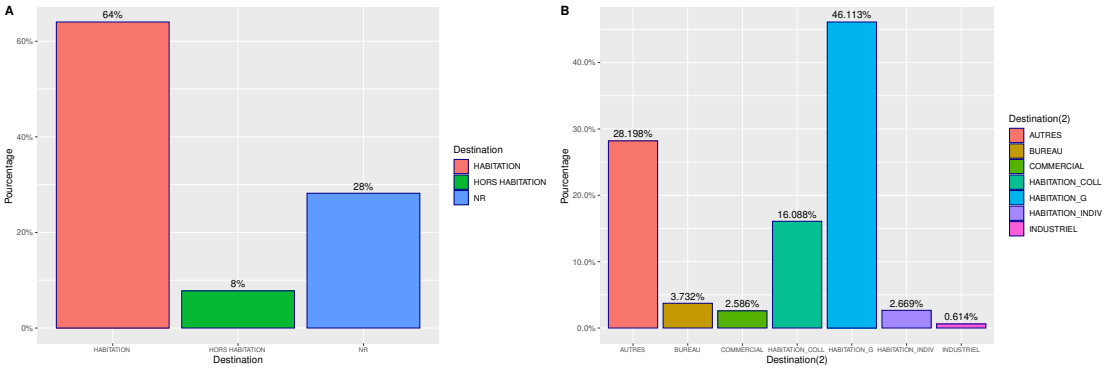

FIGURE 2.2 – A : Pourcentage de la destination selon la segmentation à trois classes. B : Pourcentage de la destination selon la segmentation à sept classes

La classe NR (ou AUTRES), contenant à la fois les champs de destination vides et les champs de destination non classifiées, diminue pour les DOC de plus en plus récentes :

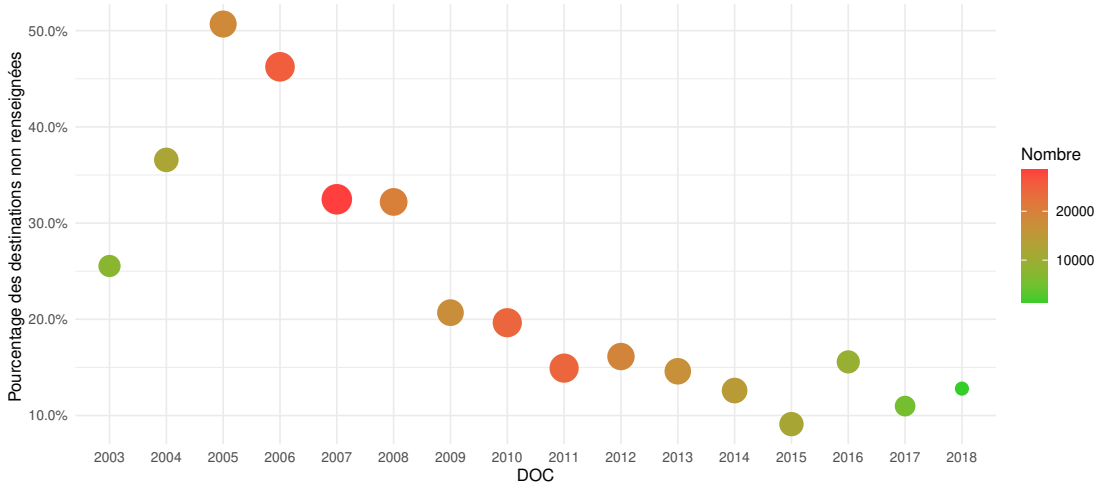

FIGURE 2.3 – Pourcentage des destinations non renseignées par DOC

Réseau et Statut :

Nos données sinistres montrent également différentes typologies. Certains sinistres sont sans suite en principal, ayant engendrés des frais d'honoraires, d'expertise ou des frais d'étude de dossiers, mais n'ayant pas connus un paiement à l'assuré. D'autres sinistres sont clos avec suite en principal et d'autres sont en cours d'évaluation.
Également, la souscription se fait selon deux canaux de distribution, à savoir le réseau des agents et le réseau des courtiers :

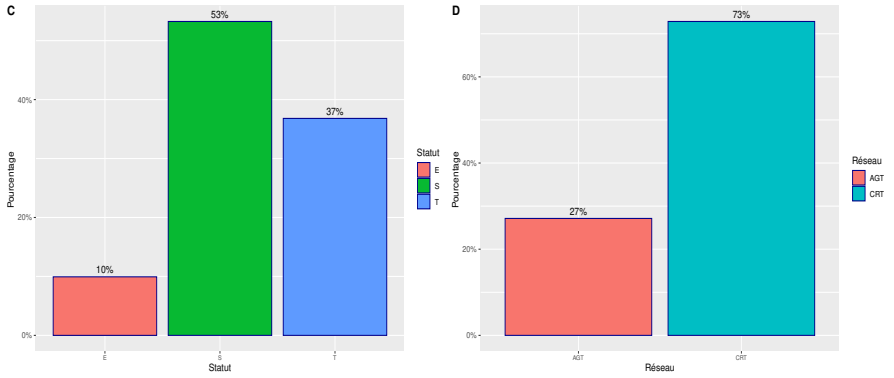

FIGURE 2.4 – C : Pourcentage des sinistres de statut ouvert, clos et en cours. D : Pourcentage des sinistres selon le réseau de distribution

Analyse des coûts [2]

Distribution des sinistres : sans suite vs clos

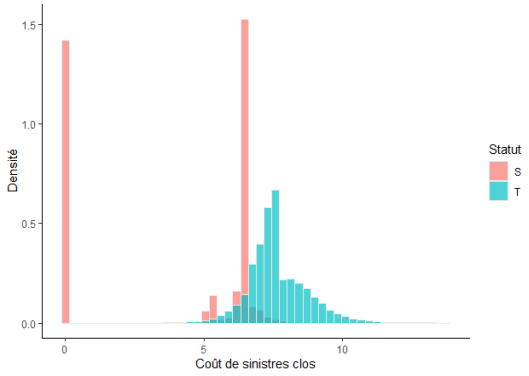

FIGURE 2.5 – Distribution des coûts des sinistres clos [a], par statut des sinistres après transformation en $x \to \log(x+1)$
Statut - S : Sans suite en principal, T : clos avec suite

a. L'axe des ordonnées a été limité à 12, pour des raisons de confidentialité.

Les sinistres sans suite en principal ont une distribution atypique, avec une médiane à 750 euros *As-If*, et avec plus de 40% de sinistres clos à zéro.
La distribution des coûts de sinistres avec suite en principal représente une asymétrie à droite avec une forme de distribution mélange vu la concentration autour de 1200 euros *As-If*.
Ceci montre une hétérogénéité de la distribution des coûts des sinistres, que ce soit par le fait que le coût étudié ici est la somme des frais d'honoraire et du règlement en principal ou par les différentes déformations du portefeuille. La mise en *As-If* ajoute également une couche de perturbation vu que c'est une estimation générale de l'inflation des coûts.

Coût moyen (CM) - Destination :

2. Rappellons que les données présentées ici sont mis en *As-If*

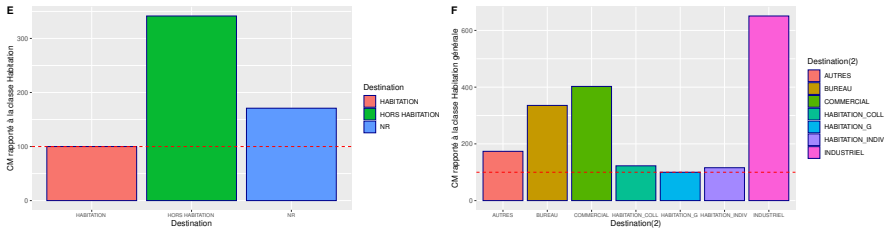

FIGURE 2.6 – E : Coût moyen des sinistres par destination (première classification), F : Coût moyen des sinistres par destination (deuxième classification)

Le coût moyen des sinistres à destination hors habitation représente plus de trois fois celui des sinistres à destination habitation. La classe NR, qui est un mixte de ces deux destinations à un coût moyen supérieur à la destination habitation.
La destination à sept classes montre des différences significatives entre les classes à destination hors habitation. Les sinistres d'ouvrages à destination industrielle ont un coût ultime important dépassant de plus de six fois le coût moyen des sinistres à destination habitation générale.

Coût moyen (CM) - Réseau et Statut :

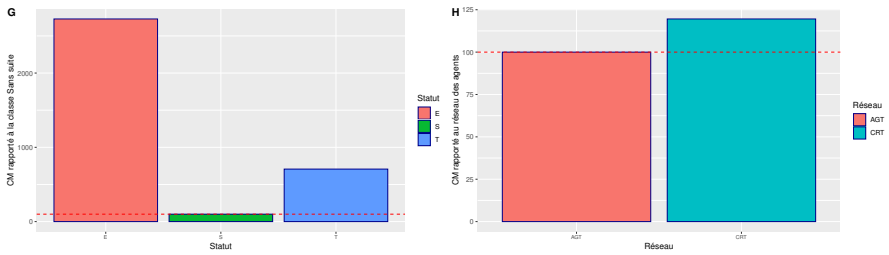

FIGURE 2.7 – E : Coût moyen des sinistres par statut, F : Coût moyen des sinistres par réseau de distribution

Le coût des sinistres ouverts est défini ici comme la somme des paiements effectués pour ces sinistres au 31/12/2021.
On remarque que les sinistres ouverts ont un coût moyen s'élevant à plus de 25 fois du coût moyen des sinistres sans suite et plus de trois fois du coût moyen des sinistres avec suite en principal. Ces sinistres ne représentent en nombre que 10% de l'ensemble des sinistres.

Une modélisation individuelle servant à estimer les montants de réserves à constituer pour les sinistres venant de se déclarer, risque donc de sous-évaluer l'estimation des réserves si elle se base uniquement sur les sinistres clos.

2.2 Traitement textuel de la destination des ouvrages

Cette section a pour but d'analyser et de retraiter la variable décrivant la destination des ouvrages.
On commencera par présenter la motivation qui nous a poussé à retraiter ce champ. Une description des traitements proposés et de la segmentation finale sera discutée dans un second temps.

2.2.1 Valeurs manquantes et anomalies

La base portefeuille permet de retrouver les variables statiques qui peuvent dégager des tendances de développement de sinistralités particulières. Nous étudierons la variable qui définie la destination des ouvrages relative aux polices.
La variable relative à la nature des travaux permet de compléter l'information sur la destination. Par simplification, elle ne sera pas utilisée dans le cadre de ce mémoire. Elle présente aussi des anomalies qui pourront se corriger selon la même démarche qui sera suivie pour la destination.

Destination	**Nombre de sinistres**
HABITATION	57 753
LOGEMENTS	9 799
COLLECTIF D'HABITATION	9 295
LOGEMENT ACCESSION	8 720
HABITAT COLLECTIF ACCESSION	8 102

TABLE 2.1 – Les 5 labels du champ des destinations les plus représentés parmi les polices sinistrées

Pour la variable relative à la nature des travaux :

Nature des travaux	**Nombre de sinistres**
NEUF	48 441
NEUFS	31 825
	12 364
CONSTRUCTION NEUVE	2 768
CONSTRUCTION	1 618
MAISON INDIVIDUELLE	1 036
TRAVAUX NEUFS	1 030

TABLE 2.2 – Les 7 labels du champ de nature des travaux les plus représentés parmi les polices sinistrées

Deux variables décrivant la destination des ouvrages sont présentes dans nos bases :

Le champ de destination brut renseigné par les gestionnaires est une variable ("DESTINAT") qui sépare la destination des ouvrages en **Habitation** et **Autres**. Cette variable a été ajoutée et retraitée à partir du champ textuel brut. Cependant, différentes anomalies ont été détectées :

Table 2.3 – Statistiques du champs "DESTINAT"

Habitation/Autres	**Nombre**
Habitation	191 543
Autres	27 026

Table 2.4 – Le nombre de sinistres selon la variable "DESTINAT"

Destination	**Habitation/Autres**	**Nombre**
HABITATION	Habitation	56 925
HABITATION	Autres	828

Table 2.5 – Exemple d'anomalies détectées pour la variable "DESTINAT"

DSTCSC	DESTINAT	Nombre de sinistres
HABITATION		2 396
HABITATION	Autres	828
HABITATION	Habitation	56 925
HABITATION COLLECTIF EN ACCESSION	Autres	70
HABITATION COLLECTIF EN ACCESSION	Habitation	715
HABITATIONS		549
HABITATIONS	Autres	89
HABITATIONS	Habitation	2 152

Table 2.6 – Exemple d'anomalies détectées

On a également constaté que la variable "DESTINAT", qui pourra potentiellement être utilisée pour étudier une segmentation par nature de destination, est peu renseignée pour les DOC inférieures à 2010.

Ce dernier point, en addition du fait que cette variable présente des anomalies, nous motive à proposer ***une classification*** *des champs de destination.*

Retraitement de la destination

Après qu'on ait défini la nature de la variable qui décrit la destination des ouvrages, on évoquera dans cette partie les méthodes de retraitement, de classification et de validation de ce champ.

Le plan ci-dessous a été suivi :

1. **Retraitement manuel des 100 destinations (brutes) les plus présentes :**

On a observé qu'elles représentent 75% des destinations de polices sinistrées non vides.

2. **Détection de classification potentielle :**

   L'analyse des 100 destinations les plus fréquentes nous a donné une idée plus claire sur la nature des destinations d'ouvrage présentes dans notre portefeuille et a permis de proposer une classification plus fine à sept classes :
   {Habitation individuelle, Habitation générale, Habitation collective, Bureau, Usage commercial, Usage industriel}.

3. **Retraitement automatique des champs textuels restants à partir des outils de traitement des données textuelles :**

   (a) Nettoyage et standardisation des champs textuels :
   Bien que la classification manuelle des 100 phrases permet de bien regrouper 75% des sinistres, la non prise en compte du pourcentage restant pourra biaiser notre analyse, surtout en présence de fautes d'orthographe, de codes parfois non unifiés. En effet, on remarque une dizaine de variantes du même champ de destination {Maison individuelle, MI, Maison Indiv, M Indiv, ... } en plus de fautes d'orthographe : {Mai0sons Individuelle, Maiisons, ... }.
   Un sinistre de coût élevé ou un ensemble de sinistres de développement de sinistralité atypique non retraité risque de changer nos analyses. Aussi, une haute fréquence de phrases non retraitées pourra être significative en poids de polices sinistrées vu que les 100 phrases retraitées ne présentent que 0.15% de l'ensemble des destinations prises d'une manière unique.
   Finalement, la classe de destinations non renseignée est toujours majoritaire en nombre, d'où la nécessité d'affiner la classification.

   Les techniques de traitement automatique du langage présentent un paradigme bien défini dans la littérature. Leur utilisation dans nos travaux sera à but opérationnel. On suivra les procédures de traitement textuel décrites dans [Feinerer 2007]

   Pour classifier des phrases textuelles, un prétraitement de standardisation est généralement fait avec des techniques de *Racinisation*.
   Cette étape permet la transformation des mots en leur radical ou racine. La racine d'un mot correspond à la partie du mot restante une fois que l'on a supprimé son (ses) préfixe(s) et suffixe(s), à savoir son radical [3]. Il existe différents algorithmes pour réaliser cette racinisation. Martin Porter développa un algorithme en 1980 connu sous le nom de *Poter Stemmer*, qui comporte des phases de traitements successives pour détecter la racine.

3. Source : `https://fr.wikipedia.org/wiki/Traitement_automatique_des_langues`

Exemple [4] **:**
"cheval, chevaux, chevalier, chevalerie, chevaucher → cheva."
La suppression des mots vides (*stopwords*) et des caractères spéciaux est également nécessaire avant cette étape. Ce sont des mots qui n'apportent pas d'informations contextuelles :
Exemple de stopwords : et, ou, a, avec, aucun, aussi
À la fin de cette étape, les séquences de textes brutes sont transformées en mots (*terms*) sous une forme de racine.
Exemple :

$$\text{"Habitations individ"} \underset{racinisation}{\rightarrow} \text{"hab ind"}$$

(b) Détection des mots (*terms*) les plus fréquents : L'étape précédente permet obtenir un dictionnaire de tous les *terms* détectés. Pour détecter les mots les plus fréquents dans l'ensemble des champs des destinations, on pourrait calculer la fréquence de leurs apparitions. Un mapping entre les structures les plus fréquentes et les phrases sous forme matricielle de colonnes de mots binaires permettra d'effectuer un tel tri.
Le choix de réduction de nombre de mots se fait à l'aide de pourcentage de *sparsity* qui est défini, pour un mot, comme le seuil de fréquence (en pourcentage) des phrases qui au-dessus duquel il sera omis.
Par exemple, un terme qui apparaît cinq fois dans l'ensemble des champs de phrases de taille 1000, aura une fréquence d'apparition de 0.005 = 0.5%. Sa *sparsity* serait de $\frac{1000-5}{1000}$ = 0.995 = 99.5%. Dans ce cas, choisir une *sparsity* de 99.5% permet de sélectionner tous les termes qui apparaissent au moins cinq fois dans l'ensemble des phrases.
Pour une meilleure détection des mots les plus fréquents, on a choisi une *sparsity* de 99.9%. Une augmentation de ce seuil a été tester et à confirmer ce choix, en concluant que les mots ajoutés ne sont pas pertinents pour notre analyse.
Ceci nous a permis de conserver 16 mots selon des fréquences d'apparition différentes (figure 2.8).
Les radicales des mots sélectionnés sont relatifs aux significations suivantes :
coll pour collectif, ind pour individuel, hab pour habitation, indu pour industriel, lgmt pour logment, comm pour commercial et bur pour bureau.

4. Source : https ://fr.wikipedia.org/wiki/Racinisation

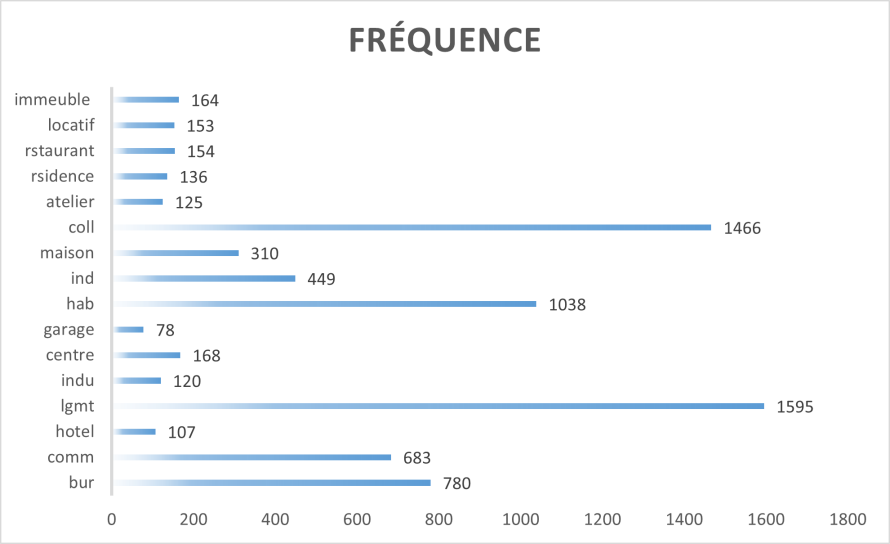

FIGURE 2.8 – La fréquence des 16 mots sélectionnés.

(c) Classification : Après cette étape, un choix des règles de classification selon les mots détectés est nécessaire. On a distingué, parmi les 16 termes sélectionnés, ce qui est directement relatif à une destination, à savoir les termes suivant : immeuble, rstaurant, rsidence, atelier, maison, hab, garage, centre, indu, lgmt, hotel, comm et bur. Les termes restants représentent une description de l'usage : locatif, coll, et ind.

Une phrase représentée par un unique terme, détectée après la racinisation dans les mots choisis, se verra s'attribuer directement une classification.
Cette dernière, pour la classe habitation, est selon le choix suivant : rsidence, maison, hab, lgmt, hotel, ind, coll $\underset{choix}{\rightarrow}$ Habitation.
Pour créer une segmentation plus fine, on propose de classifier l'habitation en habitation individuelle et collective. Les phrases comportant une référence à l'habitation sans détails sur son type seront classées en habitation générale. On propose également de séparer la classe des destinations hors habitation en : industriel, bureau et commercial.

Pour cela, les phrases avec plusieurs termes seront classifiées soit par rapport à leur état "individuel" ou "collectif" si un mot référant à une destination d'habitation est aussi présent.
Exemple récapitulatif :
"Habitations individ" $\underset{racinisation}{\rightarrow}$ "hab ind" $\underset{classification}{\rightarrow}$ "Habitation individuelle"

DSTCSC	Classification 1	Classification 2	N
HABITATION	HABITATION	HABITATION_G	9818
BUREAUX	HORS HABITATION	BUREAU	4423
COMMERCE	HORS HABITATION	COMMERCIAL	4035
IMM	HABITATION	HABITATION_G	3761
BAT COMMERCIAL	HORS HABITATION	COMMERCIAL	3754
IMME	HABITATION	HABITATION_G	2274
INDU	HORS HABITATION	INDUSTRIEL	1522
MI	HABITATION	HABITATION_INDIV	920

TABLE 2.7 – Résultat de la classification : huit phrases non vides des destinations les plus fréquentes

Les structures textuelles restantes sont classées en *Autres*.

(d) *On présentera en annexeA une description d'une méthode utilisant des algorithmes d'**apprentissage profond** dites Transformers. Elle a été testée et a servi pour un traitement manuel de certains champs. Elle n'a cependant pas été utilisée pour la sélection de mots clés ni pour une détection directe des classes.*

Chapitre 3

Prise en compte de la destination et segmentation optimale

Nous proposerons dans cette partie une analyse de la maille à considérer pour l'estimation des réserves sur des triangles DOC - Développement. La projection de ce triangle fournit donc un niveau de charge à l'ultime regroupant les aggravations de sinistres, les sinistres tardifs et les sinistres non encore manifestés.

Nous avons choisi de travailler sur des triangles de règlements. Un triangle de charges brutes pourrait être mieux adapté pour plus de robustesse au niveau de l'estimation des facteurs de Chain-Ladder pour les premiers développements des DOC récentes. En effet, nos travaux ont pour but de capter une homogénéité au niveau de la vitesse du cadencement des différents segments. De plus, étant donné que les charges dossier-dossier sont exclues, les cadences ne sont pas biaisées par des changements dans la philosophie de provisionnement. Un triangle de règlements permet donc de telles considérations.

Rappelons que les flux de paiements sont mis en *As If*, vus en 2021. Ceci permettra d'isoler l'effet de l'inflation et donc de stabiliser la projection par Chain-Ladder. Le choix final de la maille se fera par *backtesting* sur l'année 2021, ce qui revient à se limiter à la vision à fin 2020 et comparer les paiements observés en 2021 avec les différentes projections.

3.1 La segmentation et le choix de la maille

Rappelons que la vision triangulaire en assurance construction est construite sur trois axes : la DOC, la date de survenance et la date d'observation.

On analysera dans la suite la vision DOC-Développement, regroupant comme réserves à la fois la PNSEM et les IBNR. On expliquera à la fin de ce chapitre une méthode combinant les deux visions triangulaires et analysera ses résultats dans le cinquième chapitre de ce mémoire.

L'analyse de la segmentation se fera à partir de trois variables :

— La destination : *Habitation, Hors habitation, Autres.*
Un split plus fin de la destination aurait été possible vu qu'on a construit à travers notre analyse textuelle une classification à sept catégories : *Habitation Individuelle, Habitation collective, Habitation générale, Industriel, Commercial, Bureau et Autres.*

Notre choix était basé sur des contraintes opérationnelles, relatives au nombre de sinistres limités sur certaines classes. Nous tenterons aussi l'idée de regrouper la classe *Autres*[1] et la classe *Hors habitation* sur un seul segment pour bien capter les performances de la classe *Habitation*. Une modélisation ligne à ligne pourrait réussir à mieux détecter les structures de l'hétérogénéité des différentes destinations.

— Le réseau : *Agents, Courtage.*

Le comportement des sinistres pourrait être différent en termes de cadencement par réseau de distribution. En effet, le portefeuille courtage est composé majoritairement des accords cadre et des prometteurs. Le réseau des agents est quant à lui composé majoritairement de polices individuelles.[2] Ces distinctions peuvent impacter la dynamique du provisionnement.

— La première estimation de la charge : *Inférieur ou supérieur au ticket modérateur.*

Selon les règles de la convention CRAC, on s'attend à un changement de dynamique de provisionnement sur deux seuils : le ticket modérateur (TM) et l'avenant 1. Ces derniers changent chaque année et s'appliquent par rapport à l'année de déclaration des sinistres. Nous faisons la règle de classifier par rapport à la première évaluation de la charge pour une simplification du traitement opérationnel et pour rester en adéquation avec les règles de segmentations des sinistres attritionnels/graves faites par rapport à d'autres périmètres, où on considère qu'un sinistre classifié en classe grave selon sa première évaluation ne changera pas de classe même après une réévaluation venant diminuer sa charge. De la même manière, un sinistre classé en $"< TM$[3] $"$ selon la première estimation de la charge restera dans la même classe même après un changement de cette estimation dans les arrêtés ultérieurs à son enregistrement.
Nous nous limiterons à une segmentation par rapport au ticket modérateur. En effet, on a remarqué que le nombre de sinistres avec une première évaluation de la charge dépassant l'avenant 1 n'est pas suffisamment élevé pour pouvoir construire des classes et les croiser sur des mailles plus fines.

1. La classe *Autres* est parfois nommée *NR* pour Non renseigné
2. Il s'agit ici d'informations descriptives du portefeuille DO d'Allianz obtenues en interne
3. TM pour ticket modérateur

Il est important de préciser que le seuil du ticket modérateur est introduit dans le cadre de la convention CRAC pour que l'assureur DO ne conteste de recours à l'assureur de responsabilité que si l'indemnisation du sinistre est supérieur aux règlements en principal augmenté de 50% des frais d'honoraires. Une segmentation par rapport à la comparaison de la première évaluation de la charge par rapport au ticket modérateur ne permet donc pas *a priori* de sélectionner la maille de projection sur un critère d'homogénéité en développement. Cependant, cette première estimation nous permettra d'affiner le calcul des taux de recours, et pourrait également donner une idée sur la typologie des sinistres, puisque leur évaluation est basée sur un retour d'expert ou sur des informations communiquées par l'assuré, qui ne sont pas à notre disposition.

3.1.1 Méthodologie pour la sélection de la maille :

Pour la sélection de la maille la plus optimale, nous proposerons une comparaison des segmentations (croisements de classes) suivantes :

(a) Sans segmentation
(b) Destination avec 3 classes (Habitation - Hors habitation - Autres)
(c) Destination avec 2 classes (Habitation - Autres)
(d) Réseau
(e) Destination (2 classes) x Réseau
(f) Destination (3 classes) x Réseau
(g) TM
(h) Destination (3 classes) x TM
(i) Destination (3 classes) x TM x Réseau

Pour identifier la maille optimale, nous essaierons de prédire, pour les DOC de 2002 à 2020, le développement sur l'année 2021. Nous pourrons donc comparer notre estimation avec le niveau réel des flux payés en 2021.

La comparaison des différentes segmentations se fera selon les indicateurs RMSE et MAE après la projection de chaque triangle d'une segmentation donnée. Nous analyserons les performances sur chaque DOC.

Pour valider le choix de la classification, nous appliquerons le *back-testing* $\mathbf{N}$[4] fois sur un échantillon représentant 75% des sinistres pris aléatoirement en chaque itération. En particulier, nous testerons systématiquement les neuf segmentations définies ci-dessus, chacune comprenant plusieurs segments (*i.e triangles*).

4. En pratique, on prendra $N = 100$

Nous pourrons ainsi comparer également la variabilité des MSE et des MAE des échantillons et conclure sur la segmentation optimale.
Le diagramme suivant récapitule la méthodologie suivie :

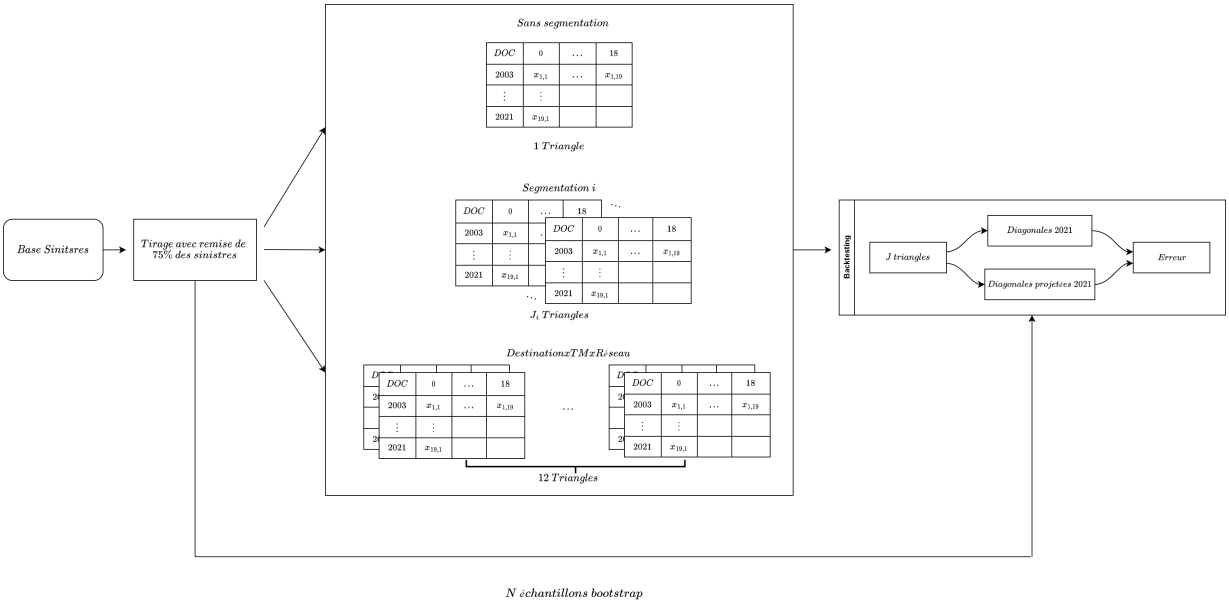

FIGURE 3.1 – Méthodologie suivie pour la sélection de la segmentation

3.1.2 Résultats

En première étape, on commence par analyser le choix de segmenter par la destination en deux ou en trois modalités : *Habitation, Autres*, en regroupant les destinations non renseignées avec les destinations classées en *Hors habitation* d'une part et une segmentation avec les trois modalités séparées d'autre part. En concluant sur l'utilisation de la destination à trois modalités, on fait le choix de ne tester que les mailles ayant cette séparation afin de réduire le temps de calcul, qui augmente considérablement selon le nombre de triangles considérés [5].
Pour chaque segmentation κ, on calcule la différence entre les règlements prédits et les valeurs connues en 2021, pour chaque DOC de 2003 à 2020 et pour chaque triangle $p \in \{1, \dots, n_\kappa\}$ d'élément $(C_{i,j}^{\kappa,p})_{i+j\leq 2021}$, n_κ étant le nombre de triangles de segmentation la κ :

$$loss_{i,j}^{\kappa,p} = \hat{C}_{i,j}^{\kappa,p} - C_{i,j}^{\kappa,p} \text{ tel que } i+j = 2021\, p = 1, \dots, n_\kappa$$
$$RMSE_\kappa = \sqrt{\sum_{i+j=2021} (\sum_{p=1}^{n_\kappa} loss_{i,j}^{\kappa,p})^2}$$
$$MAE_\kappa = \sum_{i+j=2021} |(\sum_{p=1}^{n_\kappa} loss_{i,j}^{\kappa,p})|$$

Le tableau ci-dessous montre la moyenne de ces deux indicateurs sur 100 échantillons bootstrap :

5. Le temps de calcul sur 100 échantillons bootstrap des neuf mailles proposées est de 3h50min

	RMSE	MAE
Destination(2)	2858.980	2177.750
Destination(2)xRéseau	2793.772	2079.744
Destination(3)	2761.916	2120.247
Destination(3)xRéseau	2720.015	2078.379
Réseau	2900.688	2188.198
Sans segmentation	3015.953	2324.213
TM	2960.020	2308.415
TMxDestination	2687.550	2072.351
TMxDestinationxRéseau	3037.050	2242.311
TMxRéseau	2868.397	2198.093

TABLE 3.1 – La moyenne de la RMSE et de la MAE sur 100 échantillons bootstrap

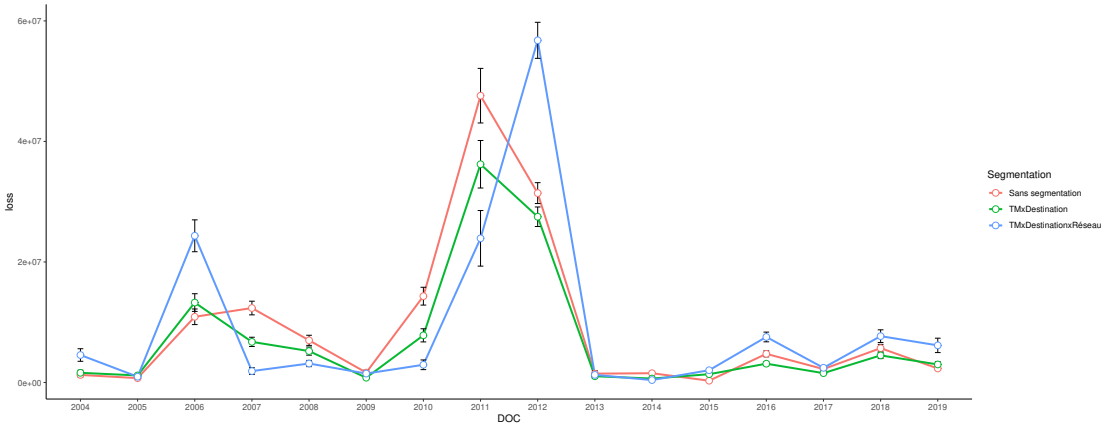

FIGURE 3.2 – MSE par DOC

Commentaire et choix de la segmentation :

— Les deux segmentations testées pour la segmentation de la variable destination en trois modalités minimisent la RMSE et la MAE plus que la segmentation à deux modalités. Suite à ce constat, la segmentation de la destinations qui est testée dans la suite des croisements est la segmentation à trois destinations.

— Le croisement des trois variables semble créer de la volatilité puisqu'on construit douze modèles à projeter.

— La segmentation qui minimise les indicateurs et qui sera retenue par la suite est le croisement **TMxDestination**

L'évaluation des écarts par DOC est nécessaire pour s'assurer de l'adéquation de la segmentation sur toutes les DOC. La figure 3.2 illustre les écarts de MSE par DOC des différentes segmentations, notamment en montrant l'intervalle représentant la variabilité de la MSE, défini par $[MSE \pm q_{95\%}(MSE)]$

La segmentation retenue minimise la MSE sur la majorité des DOC. La segmentation sur la maille TMxDestinationxRéseau est volatile, ce qui est certainement dû au manque de robustesse de la projection en 12 triangles. Notons que les DOC 2011 et 2012 connaissent des règlements élevés par rapport aux autres lignes du triangle global. La MSE pour ces DOC connait de grandes variations, pouvant être dû à une volatilité des calculs des facteurs de développement Chain-Ladder selon l'échantillonnage effectué ou encore à la présence de sinistralité atypique.

Sur toute la base des sinistres, une comparaison (*backtest* 2021) en relative entre la segmentation retenue et la projection sans segmentation est montrée sur la figure ci-dessous :

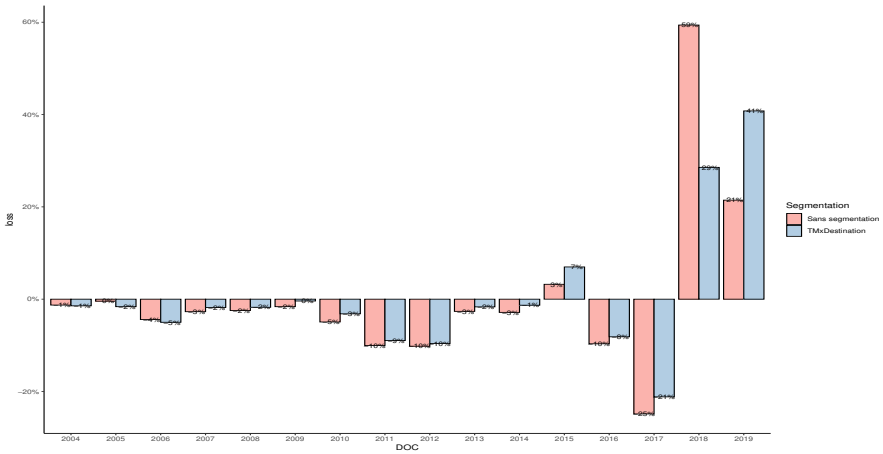

FIGURE 3.3 – Erreur relative globale. Rouge : Sans segmentation, Bleu : TMxDestination

On constate qu'on sous-provisionne sur les deux modèles pour les DOC inférieures à 2014, le modèle avec segmentation montre des écarts inférieurs pour les DOC de 2007 à 2014. Pour les DOC récentes, il sera plus pertinent d'utiliser des méthodes qui intègrent le niveau d'expositions (Primes), tel que la méthode Bornhuetter–Ferguson apparue en 1972 dans un papier de *Ron Bornhuetter* et *Ron Ferguson*. La méthode Chain-Ladder estime un niveau d'ultime non adapté, vu le manque de données en première année.

Aussi, cette erreur relative est considérée importante pour les DOC 2011 et 2012, retraçant ainsi les mêmes phénomènes de MSE obtenu par la procédure d'échantillonnage et montrant une volatilité de la projection pour ces DOC, pouvant être liée à une sur-sinistralité non modélisé par Chain-Ladder.

Finalement, les cadences de règlements par rapport à la vision DOC sont illustrées sur la figure suivante :

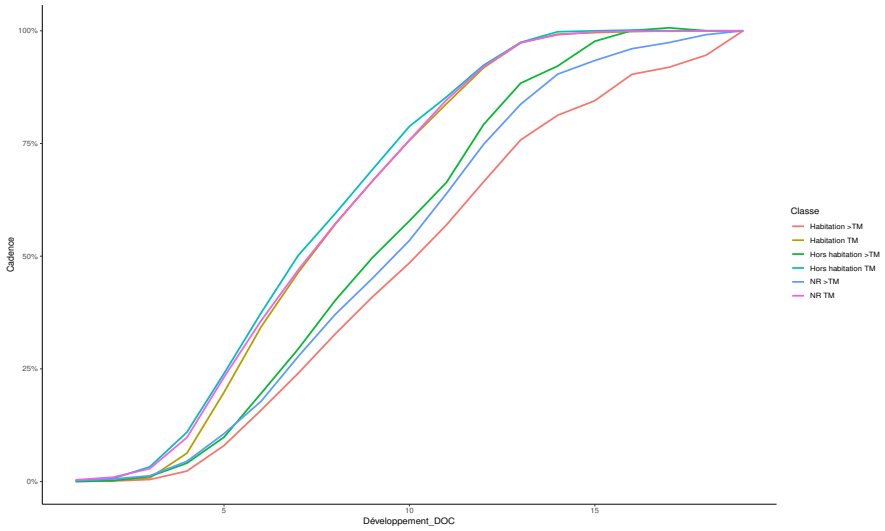

FIGURE 3.4 – Cadences sur la maille sélectionnée

Les cadences sur chaque sous-segment montrent clairement que les triangles des sinistres de charge avec une évaluation supérieure au ticket modérateur ne sont pas entièrement développés. Par manque de données, une extrapolation est nécessaire pour une projection plus adaptée à l'ultime.

3.1.3 Estimation d'un tail factor :

L'évaluation d'un tail factor se fera selon un modèle de régression linéaire du type suivant :

$$f_j^k = exp(\alpha^k j + \beta^k) + 1 \qquad \forall j \in \{0, \dots, n-1\}$$

où n représente le nombre d'années de développement des triangles et $f^k = (f_j^k)_{0,\dots,n-1}$ les facteurs de développements relatifs à la classe k considérée.
Nous devons, à partir de l'estimation du facteur de queue, projeter au-delà du dernier développement connu et déterminer, selon un seuil qu'on définit, à quel horizon l'ultime est atteint.
On fait le choix de considérer qu'un triangle est développé entièrement lorsque :

$$\hat{f_j}^k - 1 < 10^{-5}$$

Le nombre d'années nécessaires pour atteindre l'ultime selon cette définition est donc :

$$n_u^k = argmin_j\{j \geq n, \hat{f_j^k} - 1 < 10^{-5}\}$$

Nous calculons ensuite le facteur f_u qui permet de développer la dernière colonne de nos triangles à l'ultime :

$$f_u^k = \prod_{j=n}^{n_u^k} \hat{f_j^k}$$

	Habitation	Hors habitation	NR
f_u^k	1.05649	1.025424	1.033852
R_{adj}^k	0.9162	0.9400	0.9451
n_u^k	35	29	33

TABLE 3.2 – Estimation d'un *tail factor* sur la segmentation retenue pour les classes > TM

3.1.4 Impact sur l'estimation des réserves

L'algorithme de Chain-Ladder nous fournit une projection de la partie inférieure du triangle. On peut donc estimer le montant des réserves à provisionner pour la DOC i :

$$R_i = \hat{C}_{i,n} - C_{i,n-i}$$

Les réserves sur p segments sont la somme des réserves calculées sur chaque triangle. L'estimation $\hat{C}_{i,n}$ en regroupant tous les segments étant différente de $\sum_p \hat{C}_{i,n}^p$, nous souhaitons mesurer l'impact de la segmentation sélectionnée par rapport à un modèle sans segmentation :

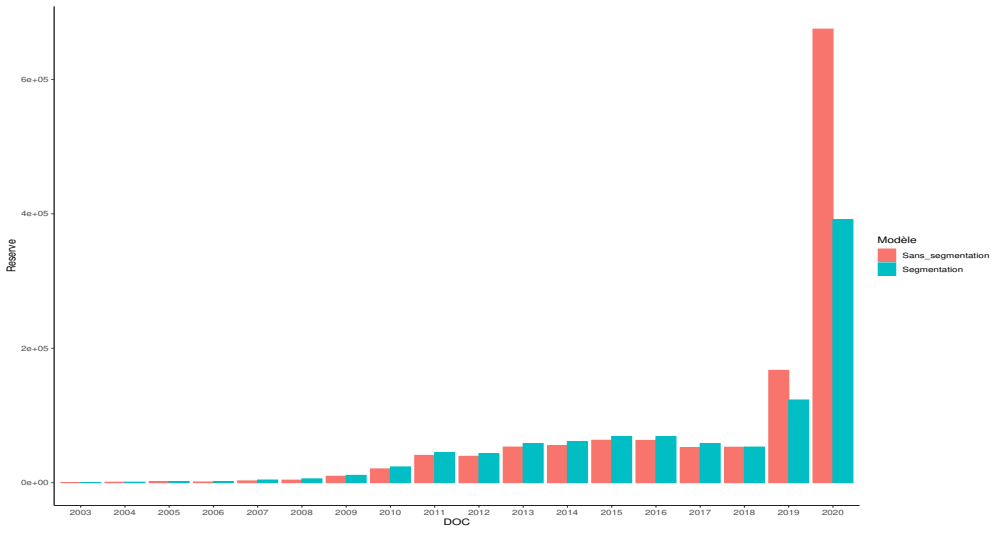

FIGURE 3.5 – Réserves sans et avec segmentation pour les DOC de 2003-2020 en K€

On remarque qu'une projection sans segmentation sous-évalue le niveau des réserves par rapport à la projection relative à la segmentation retenue pour les anciennes DOC. La projection sur les DOC récentes n'est pas trop adapté sans prendre en considération une mesure d'exposition en utilisant la méthode Bornhunter-Ferguson comme exemple.

3.1.5 Vers un provisionnement en trois dimensions

La méthode de projection sur un triangle DOC-Développement estime un montant de réserves global pour chaque DOC sans séparation entre la PSNEM et les PSAP. Projeter directement un triangle DOC - Survenance n'est pas une solution envisageable pour estimer la PSNEM car la partie supérieure de ce triangle change de vision en vision.

Exemple : le règlement relatif à la DOC 2005 et à la survenance 2008 vu en 2020 peut augmenter en 2021 suite à l'aggravation d'un sinistre ou à l'apparation d'un sinistre tardif lié à cette DOC et à cette survenance.
L'idée du provisionnement en 3D revient à vieillir à l'ultime le triangle DOC-Survenance avec une ventilation des IBNR.
La segmentation, bien qu'elle n'ait pas été sélectionnée sur un critère relatif à la liquidation des sinistres survenus, devrait impacter également l'estimation des IBNR. Nous proposons donc d'évaluer cet impact en appliquant la méthode 3D à la fois par rapport à la vision sans segmentation et pour chaque classe de la segmentation retenue.

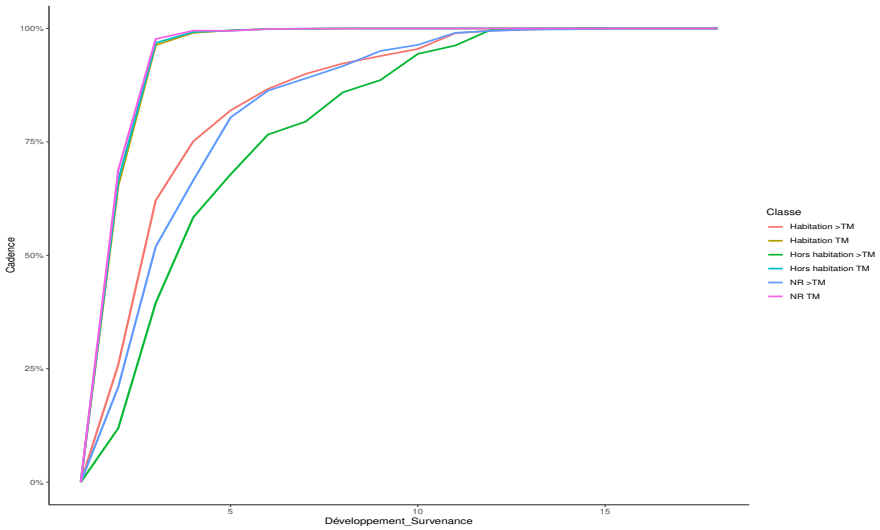

FIGURE 3.6 – Cadences de liquidation (survenance) par classe de segmentation retenue

On remarque également une liquidation différente selon la séparation *inférieur ou supérieur au TM*. Les sinistres de première charge inférieure au ticket modérateur sont à développement rapide sans différences significative par rapport à la destination. Les sinistres de classe > TM sont assez hétérogène, avec un cadencement plus long pour la classe hors habitation comparé à la classe habitation. La classe de destination NR paraît s'approcher de la classe habitation, laissant conjecturer sur la proportion de sinistres habitation à destination non identifiées.

Remarque : Dans tout ce qui suit, on fournira une explication de la méthode 3D par rapport à la vision Enregistrement. Les réserves estimées sur un triangle DOC-Enregistrement vieilli représentent à la fois la PSNEM et les IBNyR.
La dimension Enregistrement nous servira à tracer le lien avec la partie suivante, relative à la modélisation ligne à ligne des IBNeR.
Les développements suivants peuvent également s'appliquer à la dimension Survenance, en remplaçant les IBNeR par les IBNR.

Ventilation des IBNeR sur les triangles DOC-Enregistrement

La ventilation des IBNeR effectuée ici est au prorata des règlements cumulés. Soit $(P^n_{i,k})_{i+k\leq n}$ les triangles DOC-Enregistrement des paiements en vision n, k étant la DOC et i l'année d'enregistrement et soit $IBNeR_k$ notre estimation IBNeR par année d'enregistrement k. La projection à l'ultime de ce triangle selon cette méthode est présentée par l'équation suivante :

$$\forall i \in \{0,\dots,n\},\ \forall k \in \{0,\dots,n\},\ P^{ult}_{i,k-i} = P^n_{i,k-i} + \frac{P^n_{i,k-i}}{\sum_{l=0}^{k} P^n_{i,k-i}} \times \hat{IBNeR}_i$$

L'hypothèse sous-jacente à cette ventilation est qu'on considère que pour une DOC donnée, plus les règlements sont élevés, plus les paiements relatifs aux aggravations futurs seront élevés.

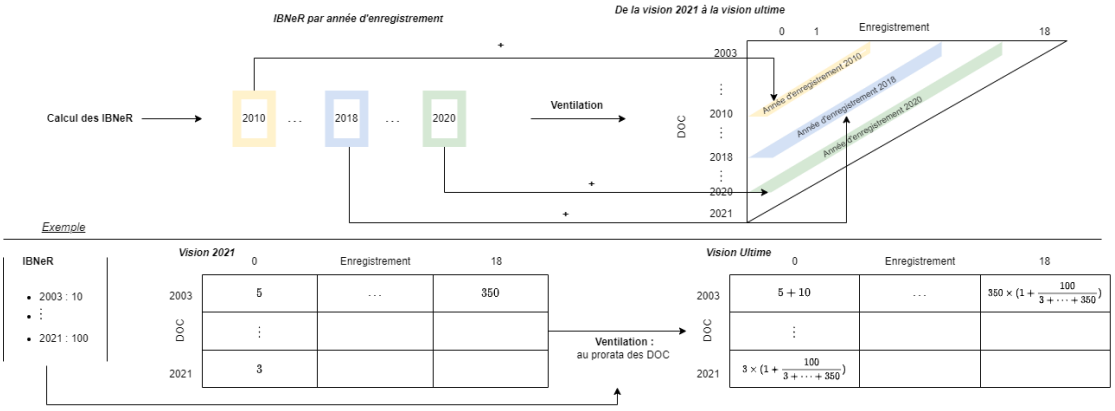

FIGURE 3.7 – Explication de la ventilation en 3D

La méthode 3D sera appliquée d'une part par rapport à la vision sans segmentation et d'autre part à la segmentation retenue, c'est-à-dire aux 6 classes du croisement de la variable DestinationxTM

Backtesting de la méthode 3D

La méthode 3D nous permettra à avoir une vision ultime de la partie supérieure d'un triangle DOC-Enregistrement, après l'intégration de l'estimation des aggravations. Après une projection par Chain-Ladder, on pourra en déduire les réserves pour les sinistres non encore manifestés augmentés des IBNyR pour chaque année de DOC. Les diagonales de ce triangle en vision décumulée représentent les flux à payer pour chaque DOC sur une année d'enregistrement donnée.

itLa vision de l'année d'observation étant absente, on ne peut pas directement estimer le montant qui sera payé pour une DOC j et pour un enregistrement en $j+5$, réglé en $j+6$ par exemple. Une méthode simple pour estimer ces flux sera d'estimer des cadences de paiements sur un triangle Enregistrement - Développement (*i.e : Année d'observation - Année d'enregistrement*). On pourra par la suite cadencer la totalité de la partie inférieure du triangle projeté qui sert à estimer la PSNEM + IBNyR.

Avec les mêmes notations définies supra, soit $(\hat{P}^{ult}_{i,j})_{i,j\leq n}$ le triangle DOC-Enregistrement vieilli à l'ultime et développé par Chain-Ladder, et soit, $(\hat{P}^{s}_{i,j})_{i,j\leq n}$ le triangle Survenance - Développement. Les cadences de paiements sur ces derniers triangles sont définies comme :

$$\begin{cases} \tilde{K}_i = \frac{1}{\prod_{j=i}^{j=n-1} \hat{f}_j}, & i = 1, \dots, n-1 \\ \tilde{K}_n = 1 \end{cases}$$

Où les $\hat{f}_j$ sont estimés par Chain-Ladder :

$$\hat{f}_j = \frac{\sum_{i=1}^{i=n-j} \hat{P}^{s}_{i,j+1}}{\sum_{i=1}^{i=n-j} \hat{P}^{s}_{i,j}}$$

L'estimation des montants à payer à l'année $r = k+j+i$ future pour la DOC k et l'enregistrement $l = k+j$ selon le triangle $(\hat{P}^{ult}_{k,j})_{k,j\leq n}$ est donc donnée par :

$$\hat{F}^{i}_{k,l} = \hat{P}^{ult}_{k,l} \times \tilde{K}_i$$

Backtester sur l'année 2021 revient donc à reproduire l'estimation en se limitant à l'horizon 2020 et à comparer les flux réellement perçu en 2021 avec les flux estimés :

$$Flux_{2021} = \sum_{k=2003}^{i=2020} \sum_{l=k+1}^{l=2020} \hat{F}^{2021-l}_{k,l}$$

Le Backtest de la méthode 3D, à la fois sans et avec segmentation, sera analysé dans le cinquième chapitre. Une comparaison des réserves globale des différentes méthodes est donnée en A.1.

3.1.6 Conclusion et limites

Ce chapitre a permis d'analyser la projection par Chain-Ladder selon une vision DOC-Développement agrégée. On a pu sélectionner la maille optimale qui minimise la MSE selon un *backtesting* par rapport à l'année 2021.
Également, la maille sélectionnée montre des cadences de liquidation différentes selon la vision d'année de survenance. La méthode 3D, permettant une estimation de la PSNEM, devrait également produire des résultats différents avec ou sans la segmentation.
Cette méthode et son *backtest* ont été introduits par la suite selon l'année d'enregistrement, ce qui permet d'estimer une PSNEM augmentée des IBNyR. Les résultats de *backtest* de ces méthodes, sans et avec segmentation, sont illustrés sur la figure 5.1. Le niveau des réserves globales est présenté en A.1.

Nous présenterons dans la suite les limites d'une telle modélisation et les perspectives d'améliorations possibles.

Limite de la projection Chain-Ladder : La méthode a été utilisée selon sa vision algorithmique et non selon son cadre stochastique qui n'est valable que selon ses hypothèses sous-jacentes. Également, des avis d'expert permettant de modifier les facteurs de développements lié à des cas de sinistres sérielles ou à des développements atypiques, de nature conjoncturelle, n'ont pas été pris en compte. Une méthode alternative serait de sélectionner la maille optimale à travers la minimisation de la MSEP, obtenu selon un cadre stochastique :

$$\text{MSEP} = \mathbb{E}((R - \hat{R})^2)$$

et avec un choix adéquat de facteurs de développement, en supprimant les facteurs individuels extrêmes par exemple.

Limites de la segmentation : La méthode de projection Chain-Ladder à l'ultime est une méthode non additive dans le sens où une projection sur deux triangles n'aura pas les mêmes résultats qu'une projection d'un triangle globale. La segmentation choisie pourra créer une dépendance non captée par la modélisation en Chain-Ladder classique, qui suppose que le développement d'un triangle ne dépend que de ces propres anciennes valeurs. En voulant capter des structures de développements homogènes, on crée de la volatilité additionnelle avec une projection sur six triangles en ignorant la dépendance crée.
La méthode de Chain-Ladder multivarié introduite dans [Pröhl and Schmidt 2005] et [Zhang 2010] suppose, selon un cadre stochastique, une dépendance de différents triangles de *run-off* et estime des facteurs de développements incorporant les multiples corrélations.

Limite de méthode 3D : la méthode 3D utilisée se base sur une ventilation au prorata des règlements cumulés par DOC. Ce critère de ventilation n'a pas été remis en cause, bien qu'il constitue une hypothèse forte pouvant influencer les

résultats d'une manière significative. Plus d'analyses sont présentées dans le mémoire [Bourry 2016] qui analyse différentes techniques de ventilation : au prorata des nombre de sinistres, de charges et en distinguant les IBNyR des IBNR. La projection par des triangles de règlements semble aussi être plus volatile, comparés aux triangles de charges, surtout au niveau des premiers développements. Ceci rend également la ventilation selon les règlements moins adaptés.

Backtest : Le *backtesting* servant à la sélection de la maille optimale est fait l'année 2021. Ceci est considéré comme une limite de la méthode.

Nous présenterons dans le chapitre suivant une modélisation ligne à ligne des IBNeR, prenant en compte les données censurées. Cette modélisation permettra d'estimer également la PNSEM augmentée des IBNyR en la combinant avec la méthode 3D.

Chapitre 4

Vers une modélisation ligne à ligne

La garantie Dommages-ouvrage présente des spécificités particulières. Les méthodes de provisionnement classiques risquent un manque de robustesse sur un risque atypique et à développement long.

L'intégration des caractéristiques des sinistres dans la modélisation est importante. D'une part, on a montré à partir des méthodes agrégées triangulaires que le développement des sinistres connaît une hétérogénéité selon la destination des ouvrages. Le réseau de distribution ou encore le délai entre la survenance et l'année de DOC peuvent également impacter la dynamique de développement. D'autre part, plus de 50% des sinistres enregistrés sont sans suite en principal, c'est-à-dire qu'on s'engage sur des frais d'honoraires pour les étudier et on conclut *in fine* que les conditions pour une indemnisation en principal ne sont pas respectées. Ceci présente donc une autre source d'hétérogénéité.

Sur une garantie à développement long, une modélisation ligne à ligne de l'évolution des coûts des sinistres pourrait capter des structures que ce soit par rapport à la durée ou par rapport au niveau de l'aggravation de la charge des sinistres. Nous testerons dans cette partie une modélisation ligne à ligne inspirée de l'article [Lopez 2018] : L'article propose de modéliser la dépendance existante entre la durée et le montant des sinistres sur des données censurées. Nous adapterons cette idée en intégrant une modélisation intermédiaire de l'état des sinistres, l'objectif étant de pouvoir prédire en quel état les sinistres ouverts seront clôturés et pouvoir donc prédire leur coût final.

Que ce soit pour la modélisation de l'état à la sortie des sinistres ou pour la projection à l'ultime, on fait face à des données censurées. Nous présenterons dans un premier lieu un formalisme mathématique servant à la modélisation des données censurées. Nous traiterons ensuite le modèle [Lopez 2018] et nous finirons par une analyse et une comparaison des résultats obtenus avec un développement

en vision agrégée.
La figure ci-dessus présente le schéma de la modélisation proposée :

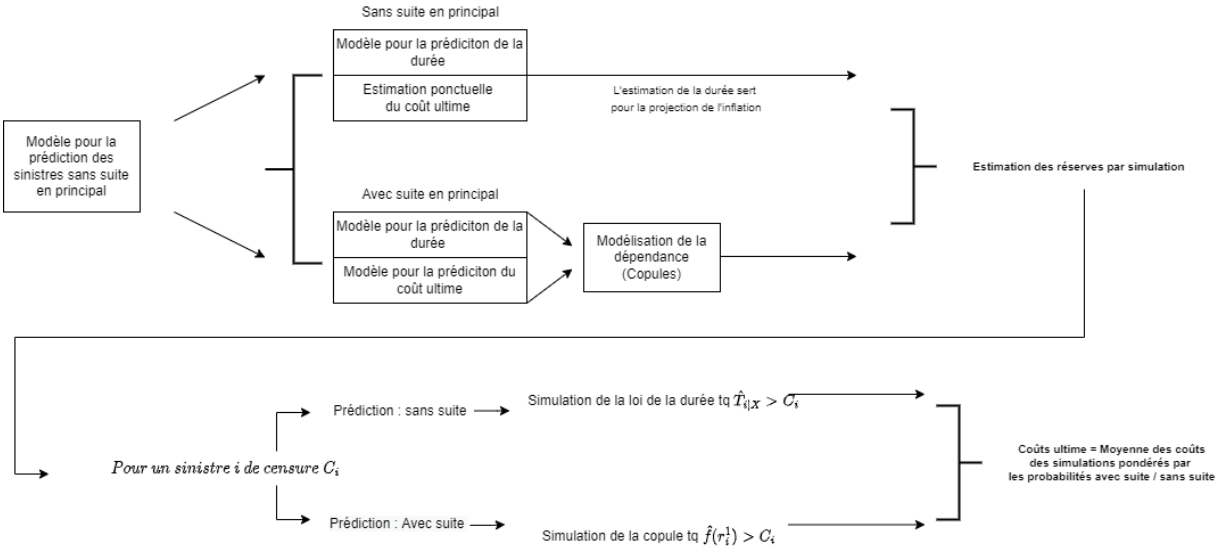

FIGURE 4.1 – Modélisation ligne à ligne proposée

Par simplification, l'estimation de la durée de vie pour les sinistres sans suite, servant à projeter l'inflation, ne sera pas étudier dans ce mémoire.
Le coût des sinistres sans suites sera fixé à ***850*** €. sans modélisation particulière de sa distribution.

Variables explicatives utilisées : On a fait le choix d'utiliser cinq variables explicatives, que ce soit pour la modélisation de la typologie des sinistres ou pour la prédiction du coût ultime.

— Destination à sept modalités : *Habitation individuelle, Habitation collective, Habitation générale, Industrielle, Bureau, Commerciale, Autres.* Elle est utilisée uniquement dans le cadre de la modélisation de l'état des sinistres.
— Destination à trois modalités : *Habitation, Hors Habitation, Autres.*
— Réseau : *Courtage et Agent.*
— Délai entre l'année de DOC et l'année de survenance : Cette variable prend des valeurs entre 0 et 10 et sera notée "delsur". Le délai entre l'année de DOC et l'année d'enregistrement a été aussi testé et a donné des performances similaires.
On fait le choix de standardiser cette variable selon la méthode min-max :

$$delsur = \frac{delsur - \min(delsur)}{\max(delsur) - \min(delsur)} = \frac{delsur}{10}$$

— Charge à l'ouverture supérieure ou inférieur à 850 € : Elle sera notée dans la suite "ch_nulle" en faisant référence à un paiement en principal nul, dans l'espérance de le détecter à ce niveau pour les sinistres sans suite.
Nous avons vu que la distribution des coûts des sinistres sans suite présente un pic à ce niveau. Cette variable sera utiliser uniquement pour la prédiction de l'état à la clôture des sinistres. La charge comme étant une variable continue ne sera pas utilisée, pour une simplification de l'entraînement des modèles, et aussi pour ne pas intégrer les corrélations temporelles non prises en compte par la mise *As-If*.
L'intégration de cette variable risque aussi de biaiser la modélisation de l'état des sinistres, puisqu'on ne récupère que la dernière estimation sur l'année au 31 décembre de la même année. On n'observera par exemple que la dernière évaluation de la charge pour un sinistre déclaré en janvier et ayant eu deux évaluations avant le 31 décembre. On pourra supposer, malgré cela, que cette estimation vient a priori de la clôture des sinistres.

Les variables utilisées ici sont supposées statiques et connues au moment de l'ouverture du sinistre.

Découpage de la base de données : On a fait le choix d'entraîner nos modèles uniquement sur les sinistres enregistrés avant le 31/12/2019. Cette base sera divisée en base d'apprentissage et de test dans le cas de la modélisation des typologies des sinistres. Les sinistres encore ouverts à fin 2019 et clôturés avant le 31/12/2020 servent pour la validation du modèle de dépendance. Les sinistres encore ouverts ou ceux ouverts en 2021 ne seront pris en compte que pour le *backtest* qui sert pour la mesure de performance du modèle combinant la projection ligne à ligne et la méthode 3D.

Ci-dessus le schéma du découpage choisi :

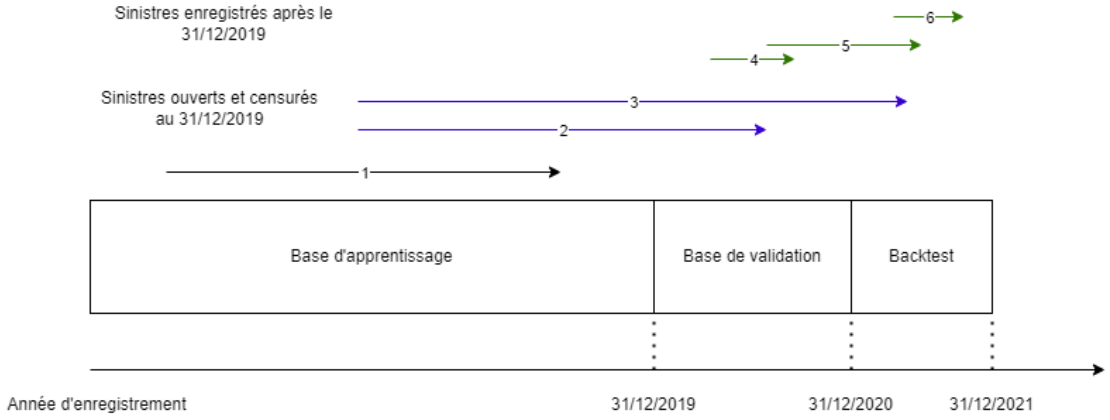

FIGURE 4.2 – Découpage de la base de données. Les flèches numérotées de un à six représentent les différents cas de figures des sinistres traités

Les modèles lignes à lignes, sur une garantie à risque juridique et à développement

long telle que la Dommages-ouvrage, ne peuvent être estimés directement sur des données à historique réduit vu la forte dépendance existant entre le coût moyen des sinistres et leurs durées.
On introduira dans la section suivante une méthode permettant de réduire ce biais à travers l'estimation des poids qui prennent en compte la censure des sinistres et qui permettent une estimation directe des modèles sur la base des sinistres clos.

4.1 Une modélisation des données censurées : les poids IPCW

Soit T la variable aléatoire représentant la durée des sinistres. On définit un objectif de modélisation de la durée des sinistres. Différents estimateurs, paramétriques comme non paramétriques, pourront servir pour tracer cette dynamique.

La spécificité de la modélisation de la durée est que l'information est manquante sur certains individus statistiques. Pour estimer la durée des sinistres, les données à disposition pourront contenir des sinistres ouverts, mais non clos, des sinistres ouverts avec un état inconnu sur leur suite ou encore des sinistres qui seront ouverts dans le futur. On parle de troncature si l'information n'est pas observable ou si l'observation n'a lieu que conditionnellement à un évènement. On parle de censure, ou que T est censurée par une variable aléatoire C, si on observe parfois C à la place de T.

Pour un arrêté au $31/12/N$, la figure suivante représente l'état des sinistres possible :

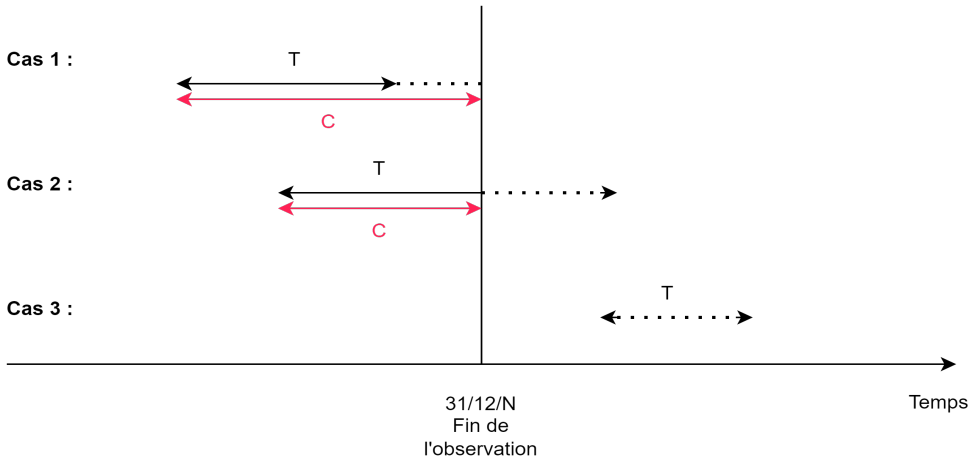

FIGURE 4.3 – L'évolution de la durée des sinistres et la censure

Nous confrontons ici des cas de censures à droite. Nous ne nous intéresserons pas dans cette partie au troisième cas de la figure.

On note Y la durée observée, δ l'indicatrice représentant si le sinistre est censuré :

$$\begin{cases} Y = \inf\{T, C\} \\ \delta = \mathbb{1}_{T \leq C} \end{cases}$$

Soit $(Y_i, T_i, C_i)_{i=1...n}$ des réalisations de ces variables aléatoires, n étant le nombre de sinistres observés.
Pour une estimation paramétrique de la loi de la durée T, l'expression de la vraisemblance dans le cas où C est indépendante de T est la suivante :

$$L(\theta) = \prod_{i=1}^{n} (f_\theta(T_i) S_C(T_i))^{\delta_i} (f_C(T_i) S_\theta(T_i))^{1-\delta_i}$$

Avec f_θ (*resp.* S_θ) la densité (*resp.* la fonction de survie) de T, f_C (*resp.* S_C) la densité (*resp.* la fonction de survie) de C.

Pour une estimation non paramétrique, l'estimateur le plus classique est celui de Kaplan-Meier (1958). Il représente une adaptation de la fonction de répartition empirique en présence de données censurées à droite. Il s'exprime comme :

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

Avec l'hypothèse de l'indépendance de T et de C, [Stute 1993] montre un équivalent de la loi forte des grands nombres pour les données censurés à droite à travers l'estimateur de Kaplan-Meier :
Avec l'hypothèse d'indépendance $C \perp T$ et en notant $\tau = \inf\{t \geq 0, \mathbb{P}(C > t) = 0\}$, on a :

$$\sup_{t<\tau} |\hat{S}_T^{(n)}(t) - S_T(t)| \underset{n \to +\infty}{\to} 0$$

4.1.1 Kaplan-Meier sous une représentation additive

Yohann Le Faou dans ces travaux de thèse [Le Faou 2019] fournit une explication claire d'une forme additive de l'estimateur Kaplan-Meier. Cette forme, analysée par [Satten et al. 2001], nous permettra de définir des poids sur les observations non censurées pour une réadaptation des méthodes d'estimation en ne se basant que sur les sinistres non censurés.
En premier lieu, remarquons qu'en absence de censure, l'estimateur de Kaplan Meier coïncide avec la fonction de survie empirique définie pour une variable Z de réalisations $Z_1, \ldots, Z_n$:

$$\hat{S}_Z(t) = \frac{1}{n} \sum_{i=1}^{i=n} \mathbb{1}_{Z_i > t}$$

Où on attribue des sauts de poids uniforme pour tout les observations. L'idée qui sera développée dans la suite est d'essayer de définir un poids de sauts pour les observations censurés.
La situation de censure est symétrique. Quitte à définir $\delta' = 1 - \delta = \mathbb{1}_{C \leq T}$ [1], l'estimateur de Kaplan-Meier pour C s'écrit comme :

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

On a donc en notant $\hat{S}_Y^{(n)}$ l'estimateur de survie empirique : $\hat{S^{(n)}}_Y(t) = \frac{1}{n} \sum_{i=1}^{i=n} \mathbb{1}_{y_i > t}$:

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

En retraçant le raisonnement fourni dans la thèse de *Yohann Le Faou (2017)* :
La différenciation de la relation ci-dessus implique :

$$d\hat{S}_Y^{(n)}(t) = d\hat{S}_C^{(n)}(t) \hat{S}_T^{(n)}(t) + \hat{S}_C^{(n)}(t) d\hat{S}_T^{(n)}(t)$$

Pour une variation infinitésimale [2] en un point non censuré $t = Y_i$: $d\hat{S}_C^{(n)}(t) = 0$ et $d\hat{S}_Y^{(n)}(t) = \frac{1}{n}$
On déduit donc :

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

L'estimateur de Kaplan-Meier pourra s'exprimer donc sous la forme additive suivante :

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

On remarque qu'on attribue à chaque observation un poids différent sur les observations à la place d'une pondération uniforme de $\frac{1}{n}$ en absence de censure. Intuitivement, les sinistres de longues durées seront sous représentés à cause de la censure. Estimer la fonction de survie que sur les sinistres clos biaisera l'estimation. Les poids définis ci-dessus réajustent ceci en pondérant les sinistres clos selon leur durée.
On notera par la suite ces poids par les poids IPCW signifiant : *Inverse probability of censoring weighting*

$$W_{i,n} = \frac{\delta_i}{n \hat{S}_C^{(n)}(Y_i)}$$

La partie suivante traite la correction du biais d'une estimation se basant sur des observations non censurées avec une pondération par les poids IPCW.

1. On suppose que $\mathbb{P}(C = T) = 0$
2. S'agissant de processus à nombre fini de sauts sur un intervalle de temps fini, la différentiation est justifiée

4.1.2 La méthode IPCW : une estimation non biaisée

Soit T la durée censurée par C et $Y = \inf\{T, C\}$. Soit $(T_i, Y_i, C_i)_{i=1,\dots,n}$ des réalisations *i.i.d.* de ces variables aléatoires non toujours observées (pour T et C).
Estimer T par Y risque de biaiser l'estimation. En effet, pour tout $t > 0$

$$\frac{\sum_{i=1}^{n} \mathbb{1}_{Y_i>t,\delta_i=1}}{\sum_{i=1}^{n} \mathbb{1}_{\delta_i=1}} = \frac{\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{T_i>t,T_i\leq C_i}}{\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{T_i\leq C_i}} \underset{n\to+\infty}{\to} = \frac{\mathbb{P}(t \leq T \leq C)}{\mathbb{P}(T \leq C)} \neq \mathbb{P}(t \leq T)$$

Le théorème suivant justifie l'introduction des poids IPCW : Soit ϕ une fonction réelle définit sur $[0, \tau[$ avec $\tau = \inf\{t \geq 0, \mathbb{P}(C > t) = 0\}$ Sous l'hypothèse d'indépendance entre C et T, on a :

$$\mathbb{E}[\frac{\delta}{S_C(Y)}\phi(Y)] = \mathbb{E}[\phi(T)]$$

On a :

$$\mathbb{E}[\frac{\delta}{S_C(Y)}\phi(Y)] = \mathbb{E}[\frac{\mathbb{1}_{T\leq C}}{S_C(T)}\phi(T)] = \mathbb{E}[\mathbb{E}[\frac{\mathbb{1}_{T\leq C}}{S_C(T)}\phi(T)|T]] = \mathbb{E}[\frac{\phi(T)}{S_C(T)}\mathbb{E}[\mathbb{1}_{T\leq C}|T]]$$

Sous indépendance entre T et C, on a :

$$\mathbb{E}[\mathbb{1}_{T\leq C}|T] = \mathbb{P}(C \geq T) = S_C(T)$$

D'où le résultat.

4.1.3 Utilisation des poids IPCW dans nos travaux

L'utilisation des poids IPCW sera fréquente dans le cadre de ce mémoire.
On introduit ici les cas étudiés :

— **Modélisation ligne à ligne des sinistres ouverts** : On testera dans ce mémoire une modélisation proposée par [Lopez 2018] qui propose une prédiction du coût final des sinistres enregistrés. Le modèle est basé sur un constat de dépendance entre la durée des sinistres et leur montant à l'ultime et également un conditionnement de ces derniers avec des variables explicatives statiques. On proposera une estimation par les modèles linéaires généralisés, effectuée par la méthode de maximum de vraisemblance, mais pondérée par les poids IPCW. Nous capterons la structure de dépendance à travers une estimation de copule par la méthode de maximum de vraisemblance, également pondérée par les poids IPCW.
Des explications détaillées seront apportées dans la partie de la modélisation ligne à ligne.

— **Modélisation et prédiction des sinistres sans suite** : Nos bases sinistres connaissent plus de 50% de sinistres sans suite en principal où on ne s'engage en tant qu'assureur que sur des frais d'expertises et de gestion. On propose d'étudier un modèle de régression logistique pour prédire l'état d'un sinistre enregistré. Le taux de sans suite connaissant des fluctuations selon la durée des sinistres, on propose également de pondérer la vraisemblance par les poids IPCW. Un modèle d'arbre de décision sera également testé en l'adaptant au cas des données censurées avec les poids IPCW.

— **Modélisation du taux de recours** : La Dommages-ouvrages est une garantie qui propose une indemnisation sans recherche de responsabilité *a priori*. On s'attend donc à un recours de l'assureur de responsabilité par la suite. Les recours ne sont pas toujours à la hauteur de l'indemnisation proposée et sont nuls sur certains sinistres. Les assureurs estiment des provisions brutes de recours et calculent des taux de recours historiques pour estimer la charge nette de recours à provisionner.
Le taux de recours change significativement selon la stratégie de souscription. En effet, les ouvrages à destination d'habitation connaissent des taux de recours inférieurs aux ouvrages à destination industrielle par exemple (voir 5.3). Un changement de politique de souscription ne sera pas directement observé et il faut attendre une dizaine d'années pour qu'une DOC ne connaisse plus de manifestations et donc de recours par la suite. Cela montre que son estimation ne pourra pas se faire directement sur une dynamique historique des recours. On propose dans ce mémoire de calculer des taux de recours avec une pondération par les poids IPCW et en se limitant à un historique relativement assez récent pour incorporer les derniers changements de portefeuille.

4.2 Modélisation de la typologie des sinistres

Nos données sinistres sont présentées selon deux typologies : sans suite et avec suite en principal. Après l'enregistrement du sinistre, on engage des frais pour l'étude du sinistre et les dédommagements y affairant à travers un expert. Il pourra conclure par exemple que le sinistre ne correspond pas aux critères de la garantie DO. La clôture de ce sinistre engage donc des frais, sans règlement en principal. C'est ce qu'on entend par sinistre *sans suite*.
Naturellement, on s'attend à un comportement différent entre ces deux typologies. Également, la typologie est connue qu'après la clôture du sinistre. Cette donnée est donc censurée.

L'objectif de la modélisation ligne à ligne étant de prédire le coût ultime des sinistres enregistrés, nous proposons dans cette partie de construire un modèle de prédiction de l'état du sinistre connu *a posteriori*.

4.2.1 Modèles et adaptation

La régression logistique

Soit $y = (y_i)_{i=0,1}$ l'état du sinistre (*i.e. sans ou avec suite*), $X = (X_j)_{j=1,...,J}$ des variables explicatives.
Notons par $p(y = 1|X)$ (*resp.* $p(y = 0|X)$) la probabilité pour qu'un sinistre de caractéristique X soit de classe $\{y = 1\}$ (*resp.* $\{y = 0\}$).
La régression logistique se définit par l'équation suivante :

$$\log \frac{p(y = 1|X)}{p(y = 0|X)} = \theta_0 + \theta_1 X_1 + ... + \theta_J X_J$$

L'estimation des paramètres se fait par la méthode de maximum de vraisemblance. La présence des données censurées ici rend l'estimation se basant que sur les sinistres clos biaisés. Un autre problème se présente dans notre cas qui est que la sortie des individus statistiques (sinistres) est à causes multiples (sans ou avec suite).

Formalisme mathématique : Soit T_1 (*resp.* T_2) la variable aléatoire décrivant la durée de vie d'un sinistre avant sa clôture en état "sans suite" (*resp. en état "avec suite"*). La variable T_1 (*resp.* T_2) n'est pas observable si le sinistre se clôture en état "sans suite" (*resp. "avec suite"*). La durée observée pour le sinistre est égale donc :

$$T = \inf\{T_1, T_2\}$$

De plus, la durée de vie est une variable censurée. La méthode la plus naturelle pour estimer la distribution de T, pourrait être d'estimer par Kaplan-Meier T_1 et T_2 et de supposer l'indépendance entre ces deux variables aléatoires. On peut également estimer $T|X$ à partir de $T_1|X$ et $T_2|X$ en supposant l'indépendance.

Par simplification, on tentera d'ajuster l'estimation en introduisant les poids IPCW, en supposant que les poids ne dépendent pas de l'état de sortie des sinistres.

L'estimation des paramètres du modèle de la régression logistique se fera donc par pseudo-vraisemblance, après une pondération par les poids IPCW :

$$\hat{\theta} = \text{argmax}_\theta \sum_i W_i * \left\{ y_i * \log\left(\frac{\exp\left({}^t\theta X_i\right)}{1 + \exp\left({}^t\theta X_i\right)} \right) + (1 - y_i) * \log\left(1 - \frac{\exp\left({}^t\theta X_i\right)}{1 + \exp\left({}^t\theta X_2\right)} \right) \right\}$$

Les arbres de décision

Cette partie est inspirée de [Franck EURIA 2021].

L'arbre de décision est un modèle non paramétrique supervisé représentant l'ensemble des choix sous forme d'arbre. L'objectif est de partitionner l'ensemble des données en groupes homogènes sur lesquels une prédiction ponctuelle sera estimée. La prédiction est donnée en identifiant la partition ou le nœud auquel il appartient.
Les arbres de décisions pour être utilisé pour expliquer une variable quantitative ou qualitative. On fournira ici des explications introductives pour problème de classification.

On appelle nœud un sous échantillon de données qui sera défini au fur et à mesure dans l'arbre. Une feuille est un segment d'un nœud, considéré homogène, et représente un estimateur de la variable endogène sur le nœud en question. La racine est la population des données d'entrainement avant segmentation.

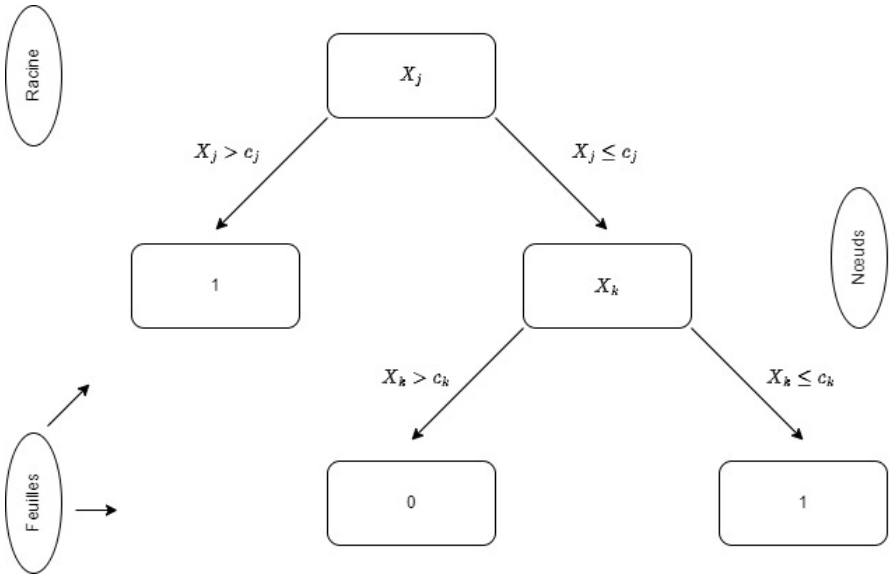

FIGURE 4.4 – Exemple d'arbre de décision (classification binaire)

Sélectionner l'arbre optimal revient à choisir les variables qui partitionnent en chaque nœud et par rapport à quels seuils.

De nombreuses techniques ont été proposées pour développer des arbres de décision, différentes principalement dans les critères utilisés pour décider comment diviser un nœud et pour éviter les situations de sur-apprentissage. Une technique populaire [Leo Breiman and Stone. 1987] segmente à chaque étape en deux sous partitions. Elle procède en deux étapes :
— Développement de l'arbre maximal : En se plaçant au niveau de la racine, on choisit la variable qui permet de mieux segmenter les données en deux

populations selon le programme d'optimisation suivant :

$$\min_{X_j \leq c_j} P_1 i(t_1) + P_2 i(t_2)$$

Avec : P_1 (*resp.* P_2) le pourcentage d'individus dans le nœud de gauche (*resp. de droite*) et $i(t)$ une fonction d'impureté servante à mesurer l'homogénéité des individus d'une sous partition. L'indice de GINI est la fonction d'impureté la plus répandue pour l'algorithme *CART*, qui se définie comme :
Pour le nœud t et pour c classes possibles, soit $p(i|t)$ la proportion d'individus de classe i dans le nœud t :

$$i(t) = 1 - \sum_{i=1}^{c} p(i|t)(1 - p(i|t))$$

On effectue cette procédure jusqu'à ce qu'on ne puisse plus segmenter.
La segmentation à la fin de cette étape est sans erreurs : on reproduit exactement les structures des données d'apprentissage, ce qui conduit à un sur-apprentissage.

— Élagage et sélection de l'arbre optimal : L'élagage se définit comme une procédure permettant d'extraire un sous arbre de l'arbre maximal, en minimisant le taux des feuilles mal classées. L'objectif est de réduire le sur-apprentissage présent dans l'arbre maximal. Le lecteur pourra se référer à [Leo Breiman and Stone. 1987] pour des explications exhaustives.

IPCW :

L'ajout des poids IPCW à l'indice de GINI, comme proposé par [Vock et al. 2016] permet de prendre en compte la censure :

$$i_C(t) = 1 - \sum_{i=1}^{c} W_i p(i|t)(1 - W_i p(i|t))$$

Le choix de P_1 et de P_2 doit également prendre en compte cette pondération.
Cette technique a été analysée également dans [Lopez et al. 2015] et appliquée dans un mémoire d'actuariat dans [Barbaste 2017].

Remarque :

L'implémentation de la pondération par les poids IPCW a été faite dans le cas des arbres de décisions en dupliquant les lignes représentant les sinistres $[\frac{W_i}{\min(W_i>0)}]$[3] fois. Les sinistres ouverts sont absents de la base finale puisqu'ils reçoivent un poids

3. [.] est la fonction partie entière

W_i = 0 et les lignes relatives aux sinistres de poids importants sont dupliquées plus. Les scores de validation se voient également modifier avec cette pondération, ce qui correspond au fonds de la méthode discutée dans [Vock et al. 2016]. Cependant, la base de test n'est pas modifiée.

Les méthodes se basant sur la maximisation de la vraisemblance, que ce soit pour la régression logistique ou le modèle de provisionnement qui sera discuter dans la suite de ce mémoire, ont été implémentés par optimisation directe sans duplication de lignes.

4.2.2 Résultats

La modélisation a été effectuée en splittant les données en base d'apprentissage - base test avec une proportion de 75% - 25%.

La comparaison des deux modèles se fera donc sur les deux bases de modélisation : on calculera le score des prédictions bien classé défini par : $\frac{TP+FN}{N}$ où :

— TP (*True Positive*) : Le nombre de sinistres classés comme avec suite et le sont.

— FN (*False Négative*) : Le nombre de sinistres classés comme sans suite et le sont.

— N : Le nombre total de sinistres sur la base d'étude (apprentissage ou test)

Pour bien mesure le pouvoir de généralisation, on tracera la courbe de ROC sur la base de test. Cette courbe compare les performances des classificateurs en variant le seuil de discrimination, définit comme le seuil de probabilité qui permet de prédire la classe d'un individu pour un modèle donné (pris par défaut à 0.5).

On calculera également l'indice AUC qui représente la surface au-dessus de la courbe ROC :

	B.A	B.T	AUC
Régression logistique	72%	68%	0.71
Arbres de décision	80%	77%	0.84

TABLE 4.1 – Taux d'erreur calculé sur : B.A pour la base d'apprentissage et B.T pour la base de test

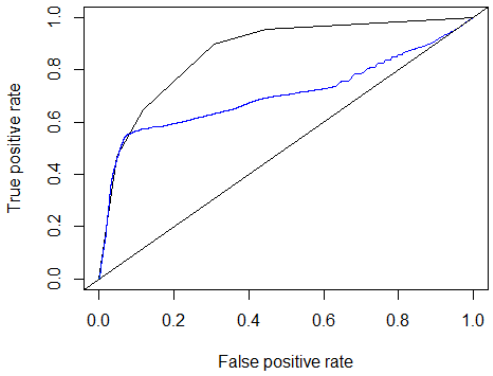

FIGURE 4.5 – Courbe ROC : En noir, le modèle d'arbres de décisions. En bleu, le modèle de régression logistique

Le modèle choisi est le modèle d'arbre de décision. La représentation de l'arbre est la suivante.

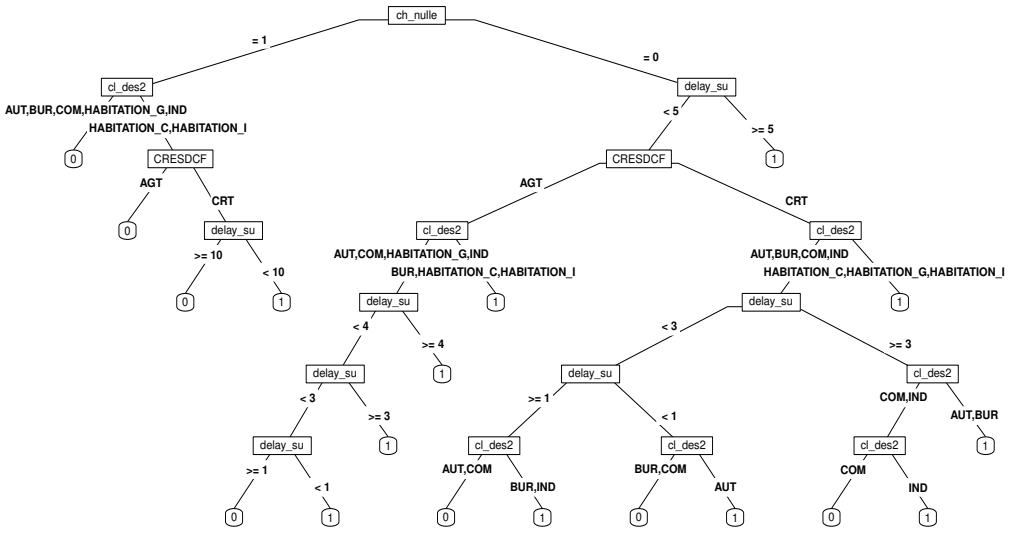

FIGURE 4.6 – Arbre CART après élagage. 0 : Sortie sans suite, 1 : Sortie avec suite.

L'arbre prédit, dans la majorité des branches, un état avec suite pour les sinistres à première évaluation de la charge dépassant 850 €. La destination à un impact direct sur la prédiction de l'état à la sortie, s'agissant de la variable la plus sélectionnée par l'arbre.

La modélisation testée dans cette section a montré une performance assez élevée pour ce type de problématique à challenges variés, que ce soit au niveau de la présence de la censure ou par rapport aux différentes sources pouvant changer le comportement de l'état à la sortie. Les arbres CART présentent cependant différents biais : ils sont généralement non robustes et il suffit d'un petit changement dans les données d'entraînement pour rendre la structure arborescente instable, ce qui peut entraîner une volatilité.
Également, la séparation hiérarchique crée un sous-apprentissage si des classes sont déséquilibrées. On remarque par exemple que l'arbre détecte une rupture selon si le délai entre la DOC et la survenance est supérieur ou égale à dix ans et prédit un état sans suite, l'inverse étant possible que sur une minorité de sinistres bénéficiant de l'application de la prescription biennale.
En effet, l'article L. 114-1 du code des assurances précise que : *"Toutes actions dérivant d'un contrat d'assurance sont prescrites par deux ans à compter de l'événement qui y donne naissance".*

Des techniques de *bagging*, introduits en 1996 par Léo Breiman, permettent de réduire la variance de l'estimation des arbres CART en construisant une agrégation de différents estimateurs construits sur des strates de données pris par échantillonnage. Le lecteur intéressé pourra se référer à l'algorithme de forêt aléatoire ou aux *gradient boosting*. [Lopez et al. 2015] utilisent les techniques de *bagging* sur des données censurées et l'adapte par la pondération IPCW. [Le Faou 2019] adapte l'algorithme des forêts aléatoires aux données censurées par pondération IPCW également et l'utilise pour la prédiction des durées des contrats en assurances santé.
Après cette étape, on s'intéressera à la modélisation des composantes marginales du modèle, à savoir la durée et le coût des sinistres.

4.3 Modélisation des marginaux et résultats

La durée de vie des sinistres, censurée, peut être estimée selon différents modèles. Nous commencerons par tracer l'estimateur non paramétrique de Kaplan-Meier pour illustrer les différences significatives de la durée selon la destination. Que ce soit pour les deux typologies de sinistres définies dans la section précédente, nous testerons une adaptation des modèles linéaires généralisés pour les données censurées.

4.3.1 La destination des ouvrages et la durée

La destination des ouvrages agit différemment sur la durée de vie des sinistres. L'estimateur non paramétrique de Kaplan-Meier, sous sa forme discrète, peut

s'écrire comme :

$$\hat{S}(t) = \prod_{t_i<t} \frac{n_i - d_i}{n_i}.$$

Sur une durée d'observation $t_1 \leq t_2 \leq \cdots \leq t_N$, n_i étant le nombre de sinistres de durée supérieur à t_{i-1}, d_i le nombre de sinistres clos au temps t_i.
En traçant la fonction de survie de Kaplan-Meier pour chaque classe de destination, on s'aperçoit que les sinistres à destination habitation sont de durée plus courtes que les sinistres à destination industrielle ou commerciale.

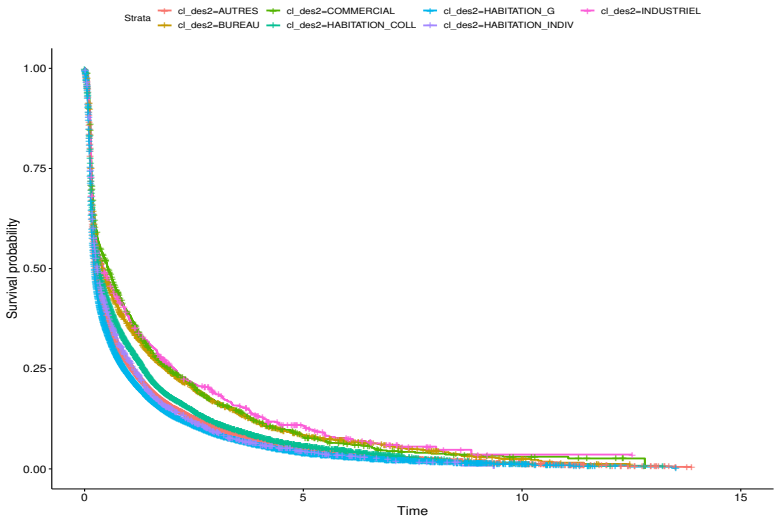

FIGURE 4.7 – La durée selon la destination

Ceci motive la prise en compte des variables endogènes comme la destination dans la modélisation de la durée des sinistres.

4.3.2 Modélisation de la marginale de la durée

L'objectif *in fine* de ce chapitre est de pouvoir estimer des lois marginales pour simuler une distribution de coûts ultimes pour chaque sinistre non clos.
Différentes modélisations de durée ont été étudiés dans la littérature servante à proposer une estimation de la loi de T, censurée :
En définissant la fonction de hasard, en suivant les notations de [Lopez 2018] :

$$\mu(t|x) = lim_{dt-0^+} \frac{\mathbb{P}(t \leq T \leq t + dt | X = x)}{dt} = \frac{f_{T|X=x}(t)}{S_{T|X=x}(t))}$$

Sous réserve de l'existance de cette limite.

Le modèle de Cox

Le modèle de Cox suppose l'hypothèse du hasard proportionnel.
En notant $S(t)$ une fonction de survie, $S_\theta(t)$ la fonction de survie de la variable aléatoire tel que X, $S_\theta(t) = S(t)^\theta$ avec θ un paramètre inconnu. On a donc :

$$\mu_\theta(t) = \frac{f_\theta(t)}{S_\theta(t)} = \theta \frac{f(t)}{S(t)}$$

Pour estimer θ, le modèle de Cox considère : $\theta = exp(\alpha_1 + \cdots + \alpha_J X_J)$
Ce qui revient à estimer la fonction du hasard comme :

$$\mu_\theta(t|x) = \mu_0(t) exp(\alpha^T x)$$

On parle d'hypothèse des risques proportionnels puisque pour deux individus x_1 et x_2 ayant les mêmes caractéristiques sauf pour une covariable X_j, $j \in \{0, \ldots, J\}$

$$\frac{\mu_\theta(t|x_2)}{\mu_\theta(t|x_1)} = \mu_0(t) exp(\beta_j(x_j^2 - x_j^1))$$

Le rapport des fonctions de hasard est donc indépendant du temps.

Le modèle AFT

Le modèle AFT (*Accelerate Failure Time model*) suppose que :

$$\mu_\theta(t|x) = \mu_0(exp(\alpha^T x)t) exp(\alpha^T x)$$

Martinussen and Scheike dans [Elangovan et al. 2020] montrent que ce modèle est équivalent à un modèle linéaire de la variable $log(T)$ sous certaines hypothèses. En pratique, une sélection de loi prenant en compte la censure devrait se faire *a priori*. La complexité de l'estimation de la vraisemblance des données censurées nous motive à reconsidérer l'introduction des poids IPCW. l'article [Lopez 2018], sur lequel cette partie du mémoire est basée, utilise une loi log normal.

Dans ce qui suit, on propose de tester deux ajustements : une loi gaussienne pour la variable $log(T)$ et une loi Gamma pour T. On retiendra après une comparaison de la qualité de l'ajustement le modèle gaussien :

$$log(T) = \alpha_0 + \alpha_1 \mathbb{1}_{CRT} + \alpha_2 \mathbb{1}_{Habitation} + \alpha_3 \mathbb{1}_{HorsHabitation} + \alpha_4 delsur + \epsilon$$

[4] Avec $\epsilon \sim \mathcal{N}(0, \sigma^2)$.
On rappelle que l'objectif de cette partie est de proposer une estimation des lois marginales. La finalité est d'avoir des pseudo-observations estimées qui suivent une loi uniforme. La sélection du modèle se fera sur ce critère.

4. CRT référant au réseau coutage, delsur pour le délai entre la survenance et la DOC

Pour ajuster une marginale des coûts ultimes des sinistres, on suivra la même démarche appliquée pour la loi de la durée. On testera, comme pour le modèle de durée, une régression gaussienne pour $\log(1+M)$ et une régression Gamma.
Le modèle sélectionné est le modèle Gamma qui s'écrit :

$$\log(1+M) \sim \Gamma(r(X), \lambda)$$

Avec :

$$r(X) = \beta_0 + \beta_1 \mathbb{1}_{CRT} + \beta_2 \mathbb{1}_{Habitation} + \beta_3 \mathbb{1}_{HorsHabitation} + \beta_4 delsur$$

L'ajustement et la sélection se fera par pseudo-vraisemblance, pondérée par les poids IPCW :
— Pour la loi de la durée :

$$\log(\hat{L}(\theta)) = \sum_{i=1}^{n} W_{i,n} \log(f_\theta^1(\log(Y_i)|\alpha^T X_i)), \text{ avec } \theta = (\alpha^T, \sigma)$$

— pour la loi des coûts :

$$\log(\hat{L}(\theta)) = \sum_{i=1}^{n} W_{i,n} \log(f_\theta^2(\log(1+N_i)|\beta^T X_i), \text{ avec } \theta = (\beta^T, \lambda)$$

Où :

$$f^1_{\sigma,\mu}(x) = \frac{1}{\sigma\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

$$f^2_{\lambda,r}(x) = \frac{m^{r-1}}{\Gamma(r)\lambda^r} \exp(-m/\lambda)\mathbf{1}_{m\geq 0}.$$

Les formules sont identiques pour les deux autres modèles testés.

4.3.3 Résultats

Nous analyserons la qualité de l'ajustement des modélisations en calculant les pseudo-observations définies comme : pour une variable aléatoire X de fonction de répartition F avec $(X_i)_{i=1,\dots,n}$ des copies de X :

$$U_i = F_X(X_i)$$

En effet, si G est la fonction de répartition inverse définie par :

$$G(y) = \inf\{x \in \mathbb{R}, F_X(x) \geq y\}$$

On a :

$$\mathbb{P}(U \leq u) = \mathbb{P}(F_X(X) \leq u) = \mathbb{P}(X \leq G(u)) = F_X(G(u)) = u$$

Sous condition d'inversibilité de la fonction de répartition F.
On déduit que $(U_i)_{i=1,\dots,n}$ suit une loi uniforme.

On comparera dans ce qui suit les pseudo observations ajustées à la loi uniforme. Les tests statistiques d'adéquation permettent de conclure sur la qualité de l'ajustement. On utilisera trois tests, à savoir le test de Kolmogorov-Smirnov, le test de Cramer-Von Mises et le test de d'Anderson-Darling.

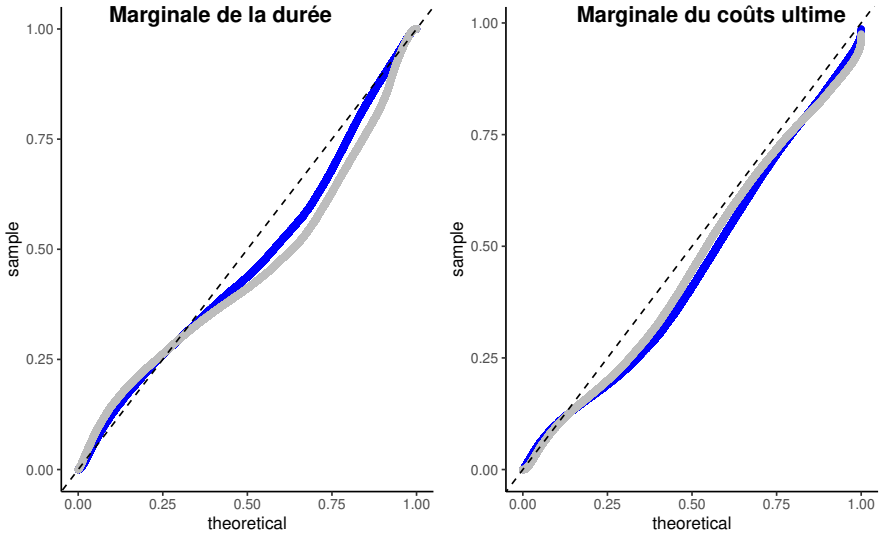

FIGURE 4.8 – QQplot des lois marginales comparées à la loi uniforme : En bleu GLM log-normal, En gris : GLM Gamma

	KS Test	CvM Test	Ad Test
Lognormal	0.1067	0.08017	0.0141
Gamma	0.1639	0.1087	0.0185

TABLE 4.2 – Statistiques des tests de la marginale de la durée

	KS Test	CvM Test	Ad Test
Lognormal	0.1521	0.0038	0.0137
Gamma	0.1349	0.0028	0.0090

TABLE 4.3 – Statistiques des tests de la marginale des coûts ultime

En se référant aux valeurs théoriques de statistique, les tests ne sont pas validés au seuil de 5%. On choisira, malgré la non-adéquation théorique, la loi qui s'ajuste le mieux, à savoir le modèle log gaussien pour la durée et le modèle GLM Gamma pour le coût.

Estimation des paramètres des lois marginales

Les tableaux ci-dessous présente les valeurs estimées des modèles sélectionnés :

	Valeurs estimées
α_0	-0.0145
α_1	-0.0081
α_2	-0.6291
α_3	0.3153
α_4	-0.0144
σ	1.3502

TABLE 4.4 – Modèle GLM lognormal sur la durée

	Valeurs estimées
β_0	2.1469
β_1	-0.0846
β_2	-0.0767
β_3	0.0839
β_4	-0.0719
λ	3.8886

TABLE 4.5 – Modèle GLM Gamma sur le coût ultime

Les résultats confirment l'analyse univariée des données. Les paramètres relatifs à la destination Hors habitation (α_3, β_3) montrent des différences relativement significatives par rapport aux autres classes de la destination. Cette significativité pourrait être mesurée par le test de Student pour le modèle de durée ou le test de Wald pour le cas du modèle GLM Gamma.
L'introduction des poids IPCW au niveau de l'optimisation rend leur implémentation compliquée. L'optimisation de cette pseudo vraisemblance risque également de fausser les propriétés statistiques de convergences sur lesquelles ces tests sont basés.

4.4 Modélisation de la dépendance entre la durée et le coût ultime

La section suivante se base sur [Brice EURIA 2021]. Elle est dédiée à quelques rappels de la théorie des copules.

4.4.1 Quelques rappels sur les copules

Nous souhaiterons rappeler dans cette partie les principales propriétés des copules bivariées utilisés dans le cadre de cette mémoire.
Un couple de variable aléatoire (X, Y) ne peut être complètement défini avec seulement les fonctions de répartition des marges, *i.e* F_X *et* F_Y. L'intuition derrière l'objet mathématique *"copule"* est de modéliser la dépendance entre X et Y, permettant ainsi une caractérisation complète de la loi de (X, Y).

On appelle copule la fonction qui permet de représenter la fonction de répartition du couple (X, Y) avec les fonctions de répartition des lois marginales respectives.
[Copule] Une copule bivariée est une fonction $\mathbb{C} : [0,1]^2 \longrightarrow [0,1]$ telle que :

(a) $\forall\ u \in [0,1], \mathbb{C}(u,0) = \mathbb{C}(0,u) = 0$

(b) $\forall\ v \in [0,1], \mathbb{C}(v,1) = \mathbb{C}(1,v) = v$

(c) $\forall\ 0 \leq u_1 \leq u_2 \leq 1,\ 0 \leq v_1 \leq v_2 \leq 1,\ \mathbb{C}(u_2,v_2) - \mathbb{C}(u_2,v_1) - \mathbb{C}(u_1,v_2) + \mathbb{C}(u_1,v_1) \geq 0$

[Sklar, 1959] Soit (X, Y), un couple de variable aléatoire de fonction de répartition F_X et F_Y. Il existe une fonction C tel que :

$$\forall (x,y) \in \mathbb{R}^2,\ C(F_X(x), F_Y(y)) = F_{(X,Y)}(x,y)$$

Réciproquement, si $\mathbb{C}$ est une copule et si F_X et F_Y sont des fonctions de répartitions alors $F(x,y) = \mathbb{C}(F_X(x), F_Y(y))$ définit une fonction de répartition dont les marges sont exactement F_X et F_Y
Remarque : Si les lois marginaux sont continues, la copule est définit d'une manière unique à partir de la relation :

$$\mathbb{C}(u,v) = F(F_X^{-1}(u), F_Y^{-1}(v))$$

On associe donc $\mathbb{C}$ au vecteur (X, Y).
La fonction de répartition du couple $(F_X(X), F_Y(Y))$ est la copule $\mathbb{C}$ associé au vecteur (X, Y). La loi de $F_X(X)$ étant une loi uniforme, la copule permet donc d'avoir des simulations du vecteur aléatoire (X, Y) :
Si $(\hat{u}, \hat{v})$ *sont générés à partir de la copule C, l'échantillon* $(F_X^{-1}(\hat{u}), F_Y^{-1}(\hat{v}))$ *est une réalisation du couple* (X, Y).

Des copules classiques :

1 - La copule indépendante :

$$\forall (u,v) \in [0,1]^2, \mathbb{C}(u,v) = uv$$

2 - Les copules elliptiques : La copule gaussienne et la copule de Student sont parmis les deux exemples les plus utilisés. Elles sont définis par leurs fonctions de répartition multivariées.

3 - Les copule archimédiennes définis par :

$$\forall (u,v) \in [0,1]^2, \mathbb{C}(u,v) = \phi(\phi^{-1}(u) + \phi^{-1}(v))$$

où ϕ est une fonction nommée générateur de la copule archimédienne, vérifiant $\phi(1) = 0$, continue et décroissante convexe.
Nous citerons trois copules archimédiennes qui seront testés dans le cadre de ce travail.

a - La copule de Clayton : $\phi(t) = \frac{t^{-\theta}-1}{t}$ avec $\theta \geq -1$ et $\theta \neq 0$

b - La copule de Gumbel : $\phi(t) = (-\log(t))^{\theta}$ avec $\theta \geq 1$

c - La copule de Frank : $\phi(t) = -\log \frac{\exp(\theta t)-1}{\exp(\theta)-1}$ avec $\theta \neq 0$

Nous nous limerons dans ce travail à l'étude des trois copules archimédiennes cités supra. Nous noterons dans ce qui suit la copule bivarié : $\mathbb{C}_\theta$, θ étant le paramètre du générateur de la copule.

Mesure de dépendance

— Le coefficient de corrélation de Pearson : $r = \frac{<X,Y>}{\sqrt{Var(X)Var(Y)}}$, ce coefficient permet de mesure la corrélation linéaire exitante entre X et Y. Il n'est pas adaptés pour mesurer la corrélation des lois non gaussiennes. En particuler, il n'est pas invariant par transformation croissante.

— le tau de Kendall :

$$\tau = \mathbb{P}[(X^1 - X^2)(Y^1 - Y^2) > 0] - \mathbb{P}[(X^1 - X^2)(Y^1 - Y^2) < 0]$$

où (X^1, Y^1) et (X^2, Y^2) deux copies indépendantes de mêmes loi (X, Y).
On peut montrer que le tau de Kendall s'écrit comme :

$$\tau = 4 \int_{[0,1]} \int_{[0,1]} C(u,v) dC(u,v) - 1$$

Ceci montre que ce coefficient ne dépends que de la copule. Il est également invariant par transformation croissante.

— Le rho de Spearman :

$$\rho = 3\{\mathbb{P}[(X^1 - X^2)(Y^1 - Y^3) > 0] - \mathbb{P}[(X^1 - X^2)(Y^1 - Y^3) < 0]\}$$

où : (X^1, Y^1), (X^2, Y^2) et (X^3, Y^3) trois copies indépendantes de mêmes loi (X, Y).

Comme pour le tau de Kendall, on peut montrer que le rho de Spearman s'écrit comme :

$$\rho = 12 \int_{[0,1]} \int_{[0,1]} C(u,v)dudv - 3$$

Aussi, il ne dépend que de la structure de dépendance.

Nous utiliserons pour estimer les paramètre de la copule le tau de Kendall. Le tableau suivant permet de tracer les correspondances avec le coefficient générateur de la copule archimédienne :

Copule	Tau de Kendall
Clayton	$\frac{\theta}{\theta+2}$
Gumbel	$\frac{\theta-1}{\theta}$
Franck	$1 + 4\frac{\theta^{-1}\int_{[0,\theta]} \frac{t^k dt}{\exp(t)-1} - 1}{\theta}$

TABLE 4.6 – Correspondance entre le paramètre de la copule et le tau de Kendall

4.4.2 Modélisation de la structure de dépendance

La modélisation proposée dans ce chapitre se base sur l'idée de détecter la dépendance entre la durée des sinistres et leur coûts ultimes. une estimation de la loi marginale conditionnelle a été effectué dans la section précédente.

L'introduction des copules semble assez naturel à cette étape vu qu'elle permettent une modélisation de la dépendance à travers les lois marginaux. Suivant les notations définit auparavant, la loi bivarié du couple s'écrit comme :

$$F_{(T,M)_{|X=x}}(t,m|X=x) = \mathbb{C}_{\theta|X=x}(F_{T|X=x}(t), F_{M|X=x}(m))$$

Pour des raisons de simplification, on suppose que la copule ne dépends pas des covariables : $\mathbb{C}_{\theta|X} = \mathbb{C}_\theta$. On suppose également une structure paramétrique sur la copule. L'estimation pourra se faire donc par la méthode de maximum de vraisemblance.

En notant : $U_i = F_{T|X}(T_i|X_i)$, $V_i = F_{M|X}(M_i|X_i)$, l'estimateur de maximum de vraisemblance se définit par : $\hat{\theta}_{ML} = \text{argmax}_\theta \sum_{i=1}^{i=n} \log(\frac{\delta^2}{\delta_u \delta_v}\mathbb{C}_\theta(U_i, V_i))$

Dans notre cas, les pseudo observation ne sont pas observables pour les sinistres censurés. On procédera donc à la maximisation d'une pseudo-vraisemblance en pondérant par les poids IPCW.

Avec $\hat{U}_i = F_{T|X}(Y_i|X_i)$, $\hat{V}_i = F_{M|X}(N_i|X_i)$:

$$\hat{\theta}_{pseudoL} = \text{argmax}_\theta \sum_{i=1}^{i=n} W_{i,n} \log(\frac{\delta^2}{\delta_u \delta_v}\mathbb{C}_\theta(\hat{U}_i, \hat{V}_i))$$

Nous testerons trois structures de dépendance : Gumbel, Clayton et Frank.

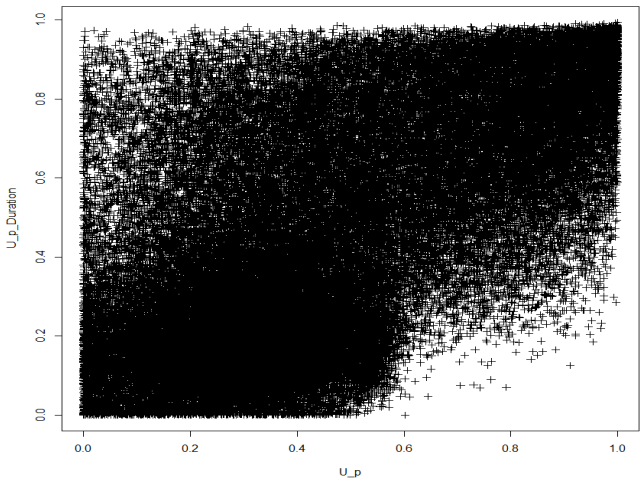

FIGURE 4.9 – Les pseudo observations de la durée et du coûts ultime

Une forte dépendance des marginaux est remarquée : Sur le modèle ajusté, les sinistres de longues durée ont un coût élévé, expliquant ainsi la forte concentration au voisinage du point de coordonné $(1, 1)$ du nuages de points.

La partie gauche montre cependant une concentration importante au voisinage du point de coordonné $(0, 0)$ mais moins condensées que la partie de droite, venant de la présence de sinistres à durée courte mais à coûts importants.

Résultats

Les résultas de l'estimation des copules sont représentées dans le tableau ci-dessous.

	Log-vraisemblance	Taux de Kendall	Paramètre associé	$RMSE_{validation}$
Gumbel	0.2179	0.4072	4.2614	48666.22
Clayton	0.1263	0.2808	1.5232	50102.47
Frank	0.2375	0.3435	0.7809	49811.42

TABLE 4.7 – Résultats de l'estimation des modèles de dépendances

La $RMSE_{validation}$ est relative aux coûts des sinistres sur la base de validation, c'est à dire sur les sinistres présents dans la base d'apprentissage mais censurés (encore ouverts après le 31/12/2019) et clos avant le 31/12/2020. Cette base sert pour contrôler les différentes modélisation effectués.

Les copules simulées selon les paramètres ci-dessous sont représentés dans la figure suivante :

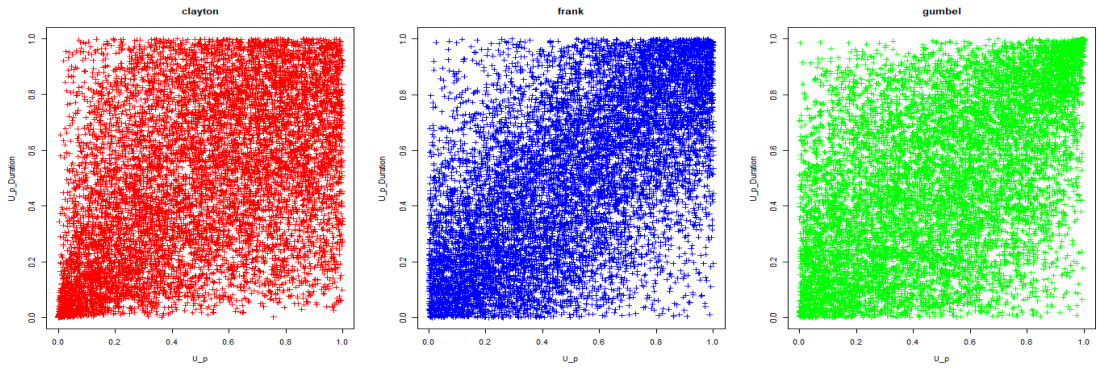

FIGURE 4.10 – Copule de Gumbel, Clayton et Frank simulés selon les paramètres estimés

Choix de la copule :

Nous comparons ici la qualité de l'ajustement des copules. Pour le cas univarié, le test de Kolmogrov-Smirnov se base sur une distance entre la loi ajustée et la fonction de répartition empirique et permet une validation l'estimation. En présence de données bivariés, la fonction de fonction empirique se définit comme :

$$F_{(X,Y)}(t,m) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(X_i \leq t, Y_i \leq m)$$

avec $(X_i, Y_i)_{i=1,\dots,n}$ des réalisations de (X, Y)

En présence des données censurées, cette estimateur non empérique est biaisée. [Gribkova and Lopez 2015] étudient les propriétés statistiques d'une pondération par les poids IPCW :

$$\hat{F}_{(X,Y)}(t,m) = \sum_{i=1}^{n} W_{i,n}\mathbb{1}(X_i \leq t, Y_i \leq m)$$

Une première étape pour visualiser la qualité de l'ajustement serait de tracer les niveau de différence entre la fonction de répartition empirique et la copule ajustée :

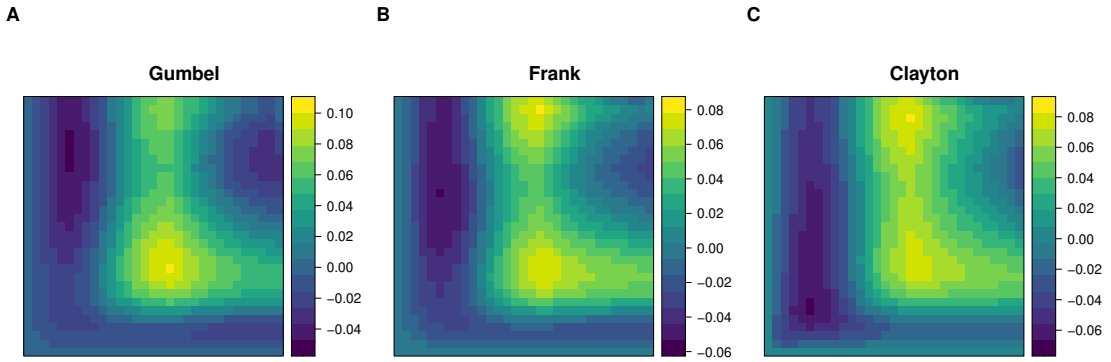

FIGURE 4.11 – *Heatmap* des modèles de copules : A - Gumbel, B - Frank, C - Clayton

Les graphiques, comportant des structures, montre une qualité d'ajustement médiocre par rapport à la fonction de répartition empirique estimée. Les copules sont choisies dans l'espérance de représenter la dépendance des deux marginaux, mais ne sont en aucun cas représentative de la structure des nuages de point des valeurs ajustées (4.9).

On calculera ensuite, pour chaque copule, la distance entre l'estimation théorique et l'estimateur de la fonction de répartition empirique. Les deux distances suivantes, similaires à la statistique du test de Kolmogrov-Smirnov et Cramer-Von Mises, seront considérées :

$$d_1\left(\hat{\mathbb{C}}, \mathbb{C}_{\hat{\theta}}\right) = \left\|\hat{\mathbb{C}} - \mathbb{C}_{\hat{\theta}}\right\|_\infty$$

$$d_2\left(\hat{\mathbb{C}}, \mathbb{C}_{\hat{\theta}}\right) = \left\{\int \left(\hat{\mathbb{C}}(a,b) - \mathbb{C}_{\hat{\theta}}(a,b)\right)^2 d\hat{\mathbb{C}}(a,b)\right\}^{1/2}.$$

Copule	$d_1\left(\hat{\mathbb{C}}, \mathbb{C}_{\hat{\theta}}\right)$	$d_2\left(\hat{\mathbb{C}}, \mathbb{C}_{\hat{\theta}}\right)$
Clayton	0.10733	0.03908
Frank	0.07549	0.03083
Gumbel	0.09781	0.05293

TABLE 4.8 – Les distances d_1 et d_2 par copule

La copule de Frank sera sélectionner dans la suite. Remarquons que la $RMSE_{validation}$ est minimale pour la copule de Gumbel. Cette dernier représentant une distance de Cramer-von Mises très dégradée et ne sera pas sélectionné. La $RMSE_{validation}$ du modèle de Frank s'approche du niveau du modèle de Gumbel et minimise les distances d_1 et d_2.

4.5 Prédiction des coûts des sinistres ouverts

Pour prédir le coûts des sinistres ouverts, on suit le schéma de la modélisation proposée en 4.1.

Pour un sinistre ouvert i, soit p_i la probabilité de clôture avec état avec suite en principal prédit par l'arbre CART. Soit M_i^b son coût ultime prédit par la modélisation de dépendance pour une simulation b.

Le montant de réserves serait donc égale à :

$$R_{\text{IBNeR}} = \sum_{i=1}^{n}(1-\delta_i)\{p_i \times (\frac{1}{B}\sum_{j=1}^{B} M_i^j) + (1-p_i) \times SS\}$$

Avec, le coût des sinistres sans suite, fixé à 850 €. En pratique, on prendra $B = 100$.

Résultats et comparaison

Les IBNeR par année d'enregistrement sont présentés dans la figure 4.12 . On comparera les résultats avec la méthode Chain-Ladder sur un triangle Enregistrement-Développement.

Notons que la méthode Chain-Ladder utilise l'information du développement des sinistres et projette par developpment selon les facteur f_j. Notre modélisation représente un vision à l'ultime des coûts.

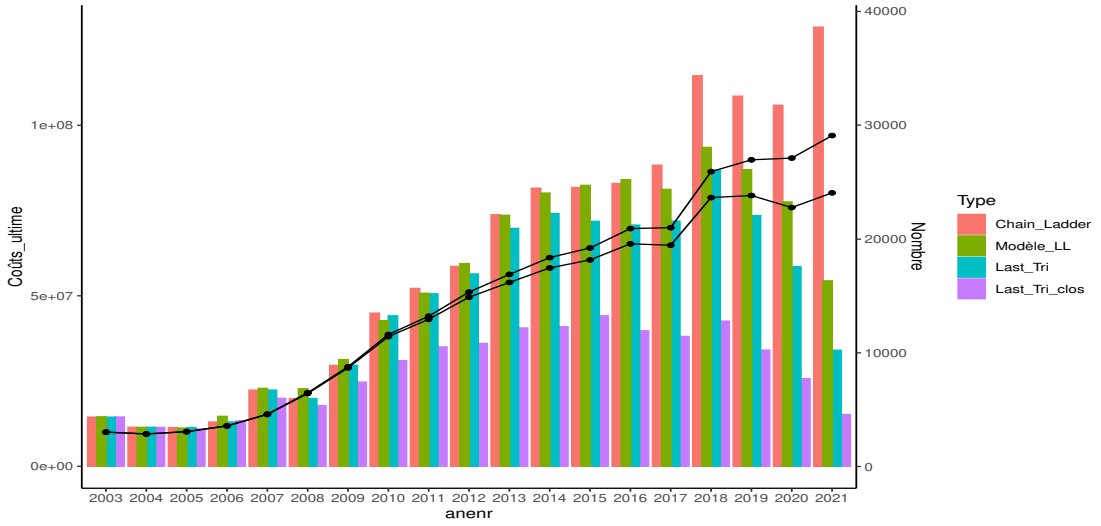

FIGURE 4.12 – Comparaison des modèles

On comparera entre :

— Chain_Ladder : ultimes par la méthode CL (rouge)
— Modèle_LL : ultime par la modélisation proposée (en vert)
— Last_Tri : Somme des derniers paiements (en turquoise)
— Last_tri_clos : coûts des sinistres clos (en violet)

— Les deux lignes présentées sur l'axe de droite représente le nombre de sinistres totale et clos.

Les résultats montrent que la non prise en compte de la dernière vision des sinistres impacte amplement les résultats, ce qui conduit à un niveau d'ultime parfois inférieur à ce qui est payé (pour l'année d'enregistrement 2010 par exemple). Notre modélisation, incorporant la censure, sous estime les ultimes des années d'enregistrement récentes vu que la simulation de la copule est conditionnée par la censure des sinistres, inférieure à une année pour ceux enregistrés en 2021. Cette sous-estimation pourrait également être due au prédictions des sans suite et à des défauts de modélisation de l'arbre CART sélectionnée.
Une prise en compte de l'exposition est nécessaire pour une comparaison plus adaptée.

Nous proposerons dans le chapitre suivant une idée de combinaison des prévisions effectués par le modèle ligne à ligne avec la méthode 3D.
En effet, le vieillissement du triangle DOC - Enregistrement et sa projection à l'ultime permet une estimation d'un montant de réserves à la fois pour les sinistres non encore manifestés et pour les sinistres tardifs. Ce vieillissement, effectué selon la méthode 3D sur une vision de survenance, par l'allocation des IBNR (estimées sur un triangle Survenance - Développement) au prorata des charges ou des règlements, pourra par ailleurs se faire en allouant les IBNeR estimées par le modèle ligne à ligne si on travaille avec des triangles d'enregistrement à la place des triangles de survenance.
Pour mesurer la performance de cette méthode, on procédera par *backtesting* des flux payés par rapport à l'année 2021.

Chapitre 5

Résultats, analyse et ouverture

La modélisation individuelle, consistant à prédire un coût aux sinistres ouverts, nous permettra d'estimer un montant d'IBNeR. La méthode, bien qu'elle permet d'inclure les informations sinistres et d'affiner la modélisation, ne propose pas une estimation des IBNeR ou encore de la PSNEM qui représente une grande partie des provisions en Dommages-ouvrage.

La méthode agrégée de provisionnement en trois dimensions permet après le vieillissement d'un triangle DOC-Survenance, à travers la ventilation des IBNR, d'avoir une estimation de la PSNEM. Les PSAP ajoutées à la PSNEM constituent donc l'ensemble des provisions DO.

Une idée développée dans ce mémoire serait de vieillir un triangle DOC-Enregistrement par la ventilation des IBNeR estimées avec la modélisation ligne à ligne. Ceci sera testé par rapport aux visions sans segmentation et avec segmentation.

Il est important de préciser ici que les réserves estimées sur un triangle DOC - Enregistrement ne représentent pas uniquement la PSNEM mais plutôt une estimation regroupant à la fois les sinistres tardifs et les sinistres non encore manifestés, contrairement à la vision par DOC-Survenance.

L'ensemble des résultats analysés dans ce chapitre se basent sur un backtesting des flux payés en année 2021

5.1 Résultats

5.1.1 Méthodologie

Comme pour le *backtesting* effectué dans le troisième chapitre relatif à la sélection de la maille optimale, l'estimation des flux prévus en 2021 avec les méthodes 2D (triangle DOC-Développement), sans et avec segmentation, est directement présente au niveau de la projection de la première diagonale.

Deux étapes sont nécessaires pour estimer les cash-flows qui seront payés en 2021 avec la vision 3D :

La première est de projeter la première diagonale du triangle DOC-Enregistrement en vision décumulée, représentant ici les montants des sinistres enregistrés en 2021 pour les DOC inférieures ou égales à 2020. Les flux payés la première année pour ces sinistres peuvent être estimés, dans une deuxième étape, en calculant une cadence de paiements sur un triangle Enregistrement - Développement. Ceci nous permettra d'estimer le montant qui serait payé en 2021 pour les sinistres non encore enregistrés. La partie relative au paiement des aggravations des sinistres déjà survenus est obtenue en projetant la première diagonale d'un triangle Enregistrement - Développement. Notons que pour les méthodes avec segmentation, les cadences de paiements seront calculées sur chaque classe.

Finalement, ce *backtest* tracera les différences de flux de paiements effectués en 2021 avec ce qu'on prévoit par DOC et par segmentation pour les méthodes avec segmentation, permettant ainsi de détecter les sources d'aberration et de volatilité des différents modèles.

La ventilation des IBNeR par DOC se fera au prorata des règlements, en supposant que plus les règlements sont élevés pour une DOC donnée, plus les aggravations seront élevées.

5.1.2 Comparaison et analyses

Au global :

Payés	2D sans seg	2D avec seg	3D sans seg	3D avec seg	3D làl sans seg	3D làl avec seg
77360	102646	98653	99849	100365	73947	73973
	32,69%	27,52%	29,07%	29,74%	-4,41%	-4,38%

TABLE 5.1 – Flux 2021 estimés par les différentes méthodes, écart relatifs

Ces premiers résultats confirment les conclusions de la comparaison entre la projection en 2D par segmentation. Avec la segmentation, on surestime moins avec la méthode 2D. La segmentation vient diminuer les performances, au global, de la modélisation 3D.

La modélisation ligne à ligne améliore clairement les performances au global. Une analyse générale des résultats ne permet pas de conclure sur la meilleure modélisation. Des écarts de projections pourront se compenser par DOC ou par segmentation, fournissant ainsi une meilleure estimation au global.

Par DOC :

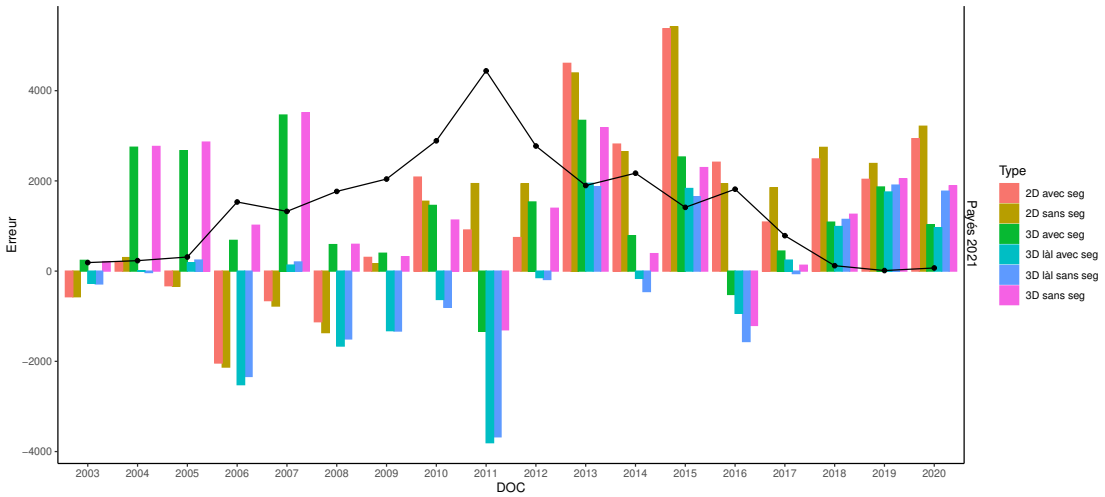

FIGURE 5.1 – L'erreur, représentée par la différence entre les flux prédits et les flux payés, par modèle et par DOC. La ligne noire, représentée dans l'axe des ordonnées à droite, montre les flux payés en 2021 par DOC.

Généralement, les différentes modélisations n'ont pas les mêmes tendances de projection :

— Pour les DOC inférieurs à 2007, la modélisation en 3D surestime significativement les paiements, contrairement à la modélisation en 2D qui sous-estime et paraît globalement adéquate. Ceci s'explique par le fait qu'en 3D, la ventilation des IBNeR vient ajouter des variations de développement qui pour ces DOC anciennes, retardent l'avènement à l'ultime plus qu'une vision en 2D. La modélisation 3D combinée avec le modèle ligne à ligne s'ajuste globalement bien, vu qu'on ne projette pas les sinistres clos, ce qui se traduit par une ventilation d'une quantité d'IBNeR bien adaptée à la réalité.
La DOC 2006 connait une hausse de règlements, bien prévue par les modélisations 3D vu qu'elles surestiment les règlements au global.

— Pour les DOC de 2008 à 2011, les modèles ligne à ligne combinés sous-estiment la projection. La modélisation des IBNeR n'incorpore pas un suivi par développement des sinistres dans le sens où les aggravations d'un sinistre enregistré ne seront pas prises en compte. A titre d'exemple, plus de 75% des sinistres de la DOC 2011, encore ouverts en 2021, ont été enregistrés avant 2020. Leurs aggravations en 2020 n'est donc pas prise en compte, ce qui conduit systématiquement à une évaluation biaisée à l'ultime.

— Pour les DOC supérieures à 2012, on remarque que la modélisation de sinistres individuels contribue à affiner la méthode 3D. La projection en 2D, surestimant toujours les paiements prévus, n'est pas adéquate. La segmentation permet de réduire les divergences des différentes méthodes de projection, notamment sur

la DOC 2012 et la DOC 2017.
La RMSE et la MAE, calculés par DOC, montrent que le modèle 3D ligne à ligne est le meilleur selon ces deux mesures :

	2D sans seg	2D avec seg	3D sans seg	3D avec seg	3D làl sans seg	3D làl avec seg
MSE	2661	2050	1840	1799	1505	1471
MAE	2134	1706	1528	1483	1167	1083

TABLE 5.2 – RMSE par modèle, calculées par DOC

Pour les modèles avec segmentation, il est intéressant de comparer les performances par classe :

Pour les modèles avec segmentation : Par classe

Par classe TMxDes	**Payés**	**2D avec seg**	**3D avec seg**	**3D làl avec seg**
NR <TM	324	1325	1765	1440
NR >TM	29059	27653	26035	18869
HABITATION <TM	1731	6594	8139	6698
HABITATION >TM	33009	46342	48980	35387
HORS HABITATION <TM	109	563	730	610
HORS HABITATION >TM	13126	19177	14717	10969
RMSE		6341	7191	4836

TABLE 5.3 – Flux 2021 estimés pour les modèles avec segmentation, par classe. En vert, les classes minimisant l'erreur quadratique.

— Pour les classes de destination "NR", la modélisation agrégée, que ce soit globale (2D) ou en distinguant les IBNeR (3D), paraît mieux adaptée. La classe "NR" représente les destinations atypiques et celles inconnues. Le modèle individuel combiné à la méthode 3D n'a pas réussi à capter les structures de mixte destination de cette classe.

— Pour les classes "<TM", le modèle 2D est le plus performant. Ces classes caractérisées par les sinistres à cadence de manifestation rapide se projettent bien par cette méthode puisque les IBNeR relatifs à cette classe ne sont généralement pas significatifs, vus les typologies des sinistres qui les caractérisent modulo les biais de segmentation.

— La modélisation ligne à ligne, combinant une projection des sinistres sans suite, qui représentent la grande partie des sinistres à première évaluation inférieure au TM, se trouve non adaptée. Ce modèle combine une modélisation *a priori* par les arbres CART de l'état final des sinistres sans inclure le montant de la première évaluation de la charge, ce qui pourra créer des divergences illustrées

par une sous-évaluation des sinistres classés ">TM" ou une surestimation pour la classe "<TM".

— Finalement, les modèles 2D et 3D surestiment la classe "Habitation >TM", cette dernière représentant le volume de règlement le plus important parmi les classes de la segmentation. La modélisation individuelle combinée à la méthode 3D réussit à capter les caractéristiques de cette classe avec une projection se rapprochant du niveau réel des paiements.

Conclusion

La modélisation combinée (3D ligne à ligne avec segmentation) permet de minimiser la MSE par DOC.

Elle permet de prendre en compte les spécificités individuelles des sinistres en la combinant avec la robustesse de la projection d'un triangle agrégée :

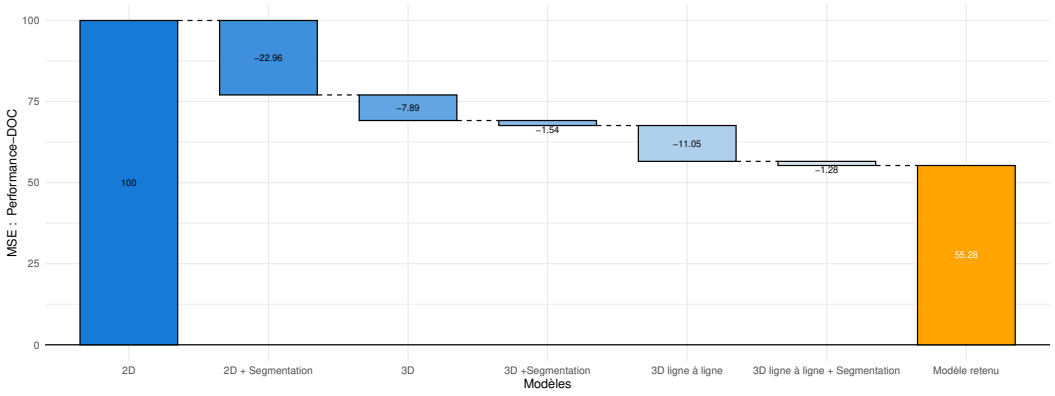

FIGURE 5.2 – Résultalt du *backtesting* : Performance à base de la MSE par DOC

L'impact des différentes modélisations sur le calcul des réserves est présenté en figure 5.4

5.2 Taux de recours

La modélisation en Dommages-ouvrage, garantie décennale servante à proposer une indemnisation à l'assuré sans recherche de responsabilité, devrait naturellement prendre en considération les recours et leur dynamique. De plus, la convention CRAC permet une simplification de la garantie et définit le ticket modérateur, seuil au-dessous duquel l'assureur DO ne reçoit pas de recours de la part de l'assureur de responsabilité.
La segmentation qu'on a retenue dans ce mémoire sépare entre les sinistres de première estimation inférieure ou supérieure au ticket modérateur. On s'attend donc à des dynamiques de recours très différentes sur les deux classes, autant plus différentes après la considération de la destination des ouvrages.
La section suivante présente une méthode pour estimer les recours.

5.2.1 Méthodologie

L'estimation des recours peut se faire selon différentes méthodes. Comme pour les charges dossier-dossier, les gestionnaires estiment des prévisions de recours sur chaque sinistre. Un modèle de projection de recours observés ou de prévisions de recours nous permettra d'estimer le montant de recours ultime comme pour les modèles de charges brutes de recours. Aussi, à travers le calcul d'un taux de recours, les recours pourront être estimés à partir des PSAP et de la PSNEM.
Pour un sinistre i après j année de développement à l'enregistrement, on définit un taux de recours individuel comme :

$$tr_j^i = \begin{cases} \frac{recours_j^i}{reglement_j^i} & \text{si } reglement_j^i \neq 0 \\ 0 & \text{si } reglement_j^i = 0 \end{cases}$$

Les recours et les réglements étant en cumulés sur les j année de développement du sinistre.
Ce taux change de vision en vision. De plus, ce taux n'est pas toujours croissant si le paiement des règlements se chevauche avec la réception des recours sur une même période. Cette situation pourra correspondre à une aggravation d'un sinistre venant après la clôture du dossier du sinistre. La réévaluation du sinistre peut créer une interaction des deux dynamiques.
Ceci étant dit, évaluer un taux de recours à l'ultime risque d'introduire un biais :

— D'une part, on introduit un biais en calculant ce taux sur tous les sinistres observés (ouverts et clos) puisqu'on n'est pas à l'ultime sur les sinistres ouverts. Ce taux se voit sous-évalué parce qu'on prend en compte en dénominateur les règlements des sinistres encore ouvert et qui donneront vraisemblablement lieu à des recours.
— D'autre part, le calculer que sur les sinistres clos biaise l'estimation puisque les sinistres non clos ont un coût moyen plus élèves, ce qui pourrait se traduire par une dynamique de recours différente.

Les déformations de portefeuille impactent également le taux de recours. La DO est une garantie à risque juridique et sa maîtrise nécessite une adaptation continue de la politique de souscription au changement de jurisprudence.

Ce taux est différent selon la typologie des sinistres. On s'attend à un taux de recours de 0% pour les sinistres de charge inférieur au ticket modérateur (Convention CRAC).

Nous proposons dans la section suivante de modéliser le taux de recours à l'ultime. Afin de l'estimer, on se basera sur les sinistres clos. On corrigera le biais introduit en élevant les sinistres encore ouverts par la méthode d'estimation des données censurés *IPCW*.

5.2.2 Formalisme mathématique

Soit T la variable aléatoire modélisant la durée des sinistres, C la censure (*i.e* la durée entre la date d'ouverture du sinistre et la date d'observation) et X la variable aléatoire modélisant le vecteur des caractéristiques de ce sinistre supposé statique (*i.e connu a priori*). Le règlement cumulé après l'enregistrement est noté Reg, les recours Rec.

On définit :

$$\begin{cases} Y = inf(T, C) \\ \delta = \mathbb{1}_{T \leq C} \\ Reg^{\star} = \delta Reg \\ Rec^{\star} = \delta Rec \end{cases}$$

$(T_i, C_i, Y_i, \delta_i, Reg_i, Rec_i, Reg_i^{\star}, Rec_i^{\star})_{1 \leq i \leq n}$ les réalisations de ces variables aléatoires, n étant le nombre de sinistres.

Le taux de recours en incluant tous les sinistres observés est défini comme :
$t_r = \frac{\sum_i^n Rec_i}{\sum_i^n Reg_i}$ qui s'écrit également en fonction des taux de recours individuels :
$t_r = \frac{\sum_i^n t_r^i Reg_i}{\sum_i^n Reg_i}$ Comme expliqué auparavant, ce taux correspond à une estimation biaisée puisqu'il prend en compte des sinistres encore ouvert.

On attribue pour chaque sinistre i un poids IPCW définit comme :

$$W_{i,n} = \frac{\delta_i}{n\hat{S}_C(Y_i)}$$

On attribue aux sinistres de statut non encore clos (censurés à droite) un poids nul, ce qui les exclus de l'analyse. On attribue au sinistre i de statut clos un poids inversement proportionnels à la probabilité estimée d'être censuré après son temps de suivi observé.

En procédant comme les sections précédente de ce mémoire, l'estimation de la fonction de survie de la censure se fera sur la base de la fonction de survie empirique.

On suppose, en suivant les hypothèses du modèle d'Olivier Lopez, que (T, X, Reg, Rec) sont indépendantes de C.
On définit le taux de recours IPCW comme :

$$t_r^{IPCW} = \frac{\sum_i^n W_{i,n} t_r^i Reg_i^*}{\sum_i^n W_{i,n} Reg_i^*}$$

Période de calcul

Comme introduit précédemment, les changements de portefeuille pourront significativement influencer le calcul du taux de recours. En effet, on constate que le taux de recours observé relatif aux sinistres d'ouvrages à destination habitation est inférieur à ceux à destination hors habitation. Un changement récent de politique de souscription, ne sera pas constaté directement en calculant un taux de recours observé.

Pour bien analyser cela, on calculera un taux de recours observé par rapport aux DOC de 2003 à 2007 et des DOC de 2008 à 2012, ceci par destination pour la classe des sinistres "$>$TM" et au global. Un calcul sur les DOC récentes, non matures, ne permettra pas de tirer de grandes conclusions. Aussi, le taux observé est inférieur à 10^{-3} pour les classes "$<$TM". On l'omettra des analyses et on le fixera, pour toute comparaison, à ce niveau.

5.2.3 Résultats

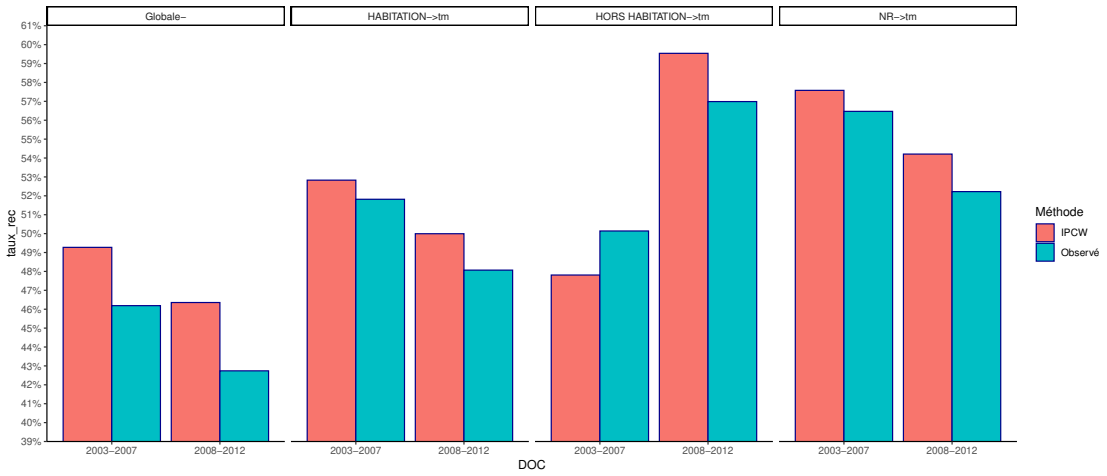

FIGURE 5.3 – Taux de recours par segment. En bleu, les taux observés calculés sur tous les sinistres. En rouge, les taux après pondération par la méthode IPCW.

Commentaires :

— Pour le calcul du taux de recours global, On remarque que le taux observé sur la strate 2008-2012 est inférieur au taux observé sur la strate 2003-2007. La méthode IPCW estime un taux 2008-2012 qui représente le même niveau du taux observé sur 2003-2008.
— Pour les diverses classes de destination, le taux de recours connait des différences significatives. La classe habitation voit une diminution du taux observé sur les DOC 2008-2012 contrairement à la classe hors habitation. La méthode IPCW permet d'estimer un taux plus élèves que le taux observé pour la classe habitation, mais évalue un taux plus bas pour la classe hors habitation de la strate 2003-2008. Ceci s'explique par la présence de sinistres de durée dépassant dix ans avec des taux de recours individuels inférieurs à 30% ce qui est peut être dû à des carences de l'assureur Dommages-ouvrage sur des sinistres déclarés après la période décennale.
— La classe NR, contenant une mixte destination, voit un taux de recours important qui diminue sur la strate 2008-2012.

On fera le choix de considérer le taux de recours estimé par IPCW sur la deuxième période.

Taux de recours au global et avec segmentation :

Après ces considérations, un taux de recours global en considérant la segmentation pourra être calculé comme :

$$t_s = \frac{\sum_i^6 t_r^{classe_i} Reg_{classe_i}}{\sum_i^6 Reg_{classe_i}}$$

	Globale : observé	Globale : retenu	Segmentation : retenu
Taux de recours	46.35%	49.21%	50.32%

TABLE 5.4 – Taux de recours global et avec segmentation

On constate que la segmentation permet d'améliorer le taux de recours d'environ 4%.

Application : Calcul des réserves
En choisissant le taux de recours global retenu, les réserves calculées par type de modélisation en net de recours sont représentées dans la figure suivante :

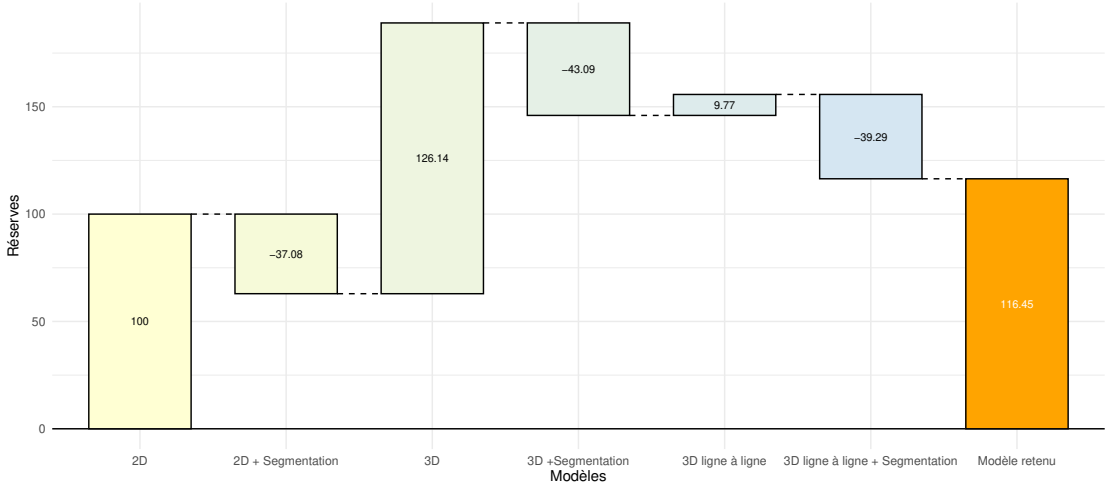

FIGURE 5.4 – Réserves nettes de recours exprimé relativement au modèle 2D

On remarque que les modèles produisent un niveau de réserves très varié par rapport à la modélisation en deux dimensions avec des écarts de [-37.08% ;+89.06%] en relatif.
Les modélisations sans segmentation sur-provisionnent par rapport aux modélisations avec segmentation. Les réserves en 3D sont surestimées par rapport à la 2D alors que la combinaison de la méthode 3D et du modèle ligne à ligne retenue, estime un niveau de réserves supérieur à la 2D de 16.45%.

Les différentes variations s'expliquent en partie par la non-adéquation de la méthode Chain-Ladder, non adéquate pour les DOC récentes. La ventilation au prorata des règlements peut également être source de volatilité en la comparant par une ventilation au prorata des charges dossier-dossier.

En conclusion, ce chapitre nous a permis de comparer les différentes modélisations testées dans ce mémoire. On a retenu selon un critère de *backtesting* la méthode qui combine la modélisation individuelle des IBNeR et la méthode 3D.
L'étude des recours nous a permis d'estimer des taux de recours par classe de destination, ce qui peut servir à une meilleure prise en compte des recours en cas de changements de stratégies de souscription selon la destination des ouvrages. Finalement, les réserves nettes de recours nous ont permis de comparer les différentes méthodes et d'apprécier notre niveau de provisionnement en interne.

Conclusion

On a proposé à travers ce mémoire d'actuariat une intégration des typologies des sinistres à des fins d'amélioration de provisionnement, dans le but d'éclairer les risques par rapport à une branche d'assurance difficilement maîtrisable. Ce travail est le résultat d'une combinaison de diverses considérations actuarielles, notamment sur les aspects de qualité des données, de considérations normatives, juridiques et surtout de modélisations statistiques.

Nos travaux illustrent le potentiel de la prise en compte des caractéristiques des sinistres sur l'estimation des provisions liées à la garantie DO. Nous avons également mis en évidence l'impact des choix de modélisation sur les performances et les résultats des méthodes, permettant de suggérer une modélisation optimale et à notre connaissance novatrice, reposant sur la combinaison de triangles agrégés avec quantification des engagements liées aux sinistres IBNeR selon une modélisation individuelle.

Néanmoins, ces considérations souffrent de plusieurs faiblesses et limites :
— Nous avons basé toutes nos comparaisons sur un *backtest* effectué par rapport à l'année 2021, ce qui ne permet pas de généraliser nos conclusions sur la pertinence des modèles.
— La méthode Chain-Ladder a été également utilisée sans validation de ses hypothèses. Elle présente à part cela des limites alarmantes, surtout pour les projections des DOC récentes.
— Des hypothèses de modélisation ont été également faites, notamment par rapport au retraitement de l'inflation, aux choix des lois marginales et du modèle de dépendance, malgré la validation médiocre de ces dernières.
— Le modèle [Lopez 2018] utilisé prédit un coût final pour les sinistres ouverts sans prendre en compte leurs aggravations, ce qui peut causer une sous-estimation des IBNeR. Le retraitement du champ textuel de la destination montre aussi une grande limite vu qu'on regroupe des mélanges de destination dans la classe des destinations non classifiées, créant une hétérogénéité additionnelle.

Il conviendrait par ailleurs, afin d'améliorer les conclusions de nos travaux et de leurs utilité, d'intégrer un traitement de l'inflation plus fin, de tester d'autres hy-

pothèses de modélisation ou encore d'améliorer la qualité de nos données.
Le retraitement de l'inflation pourrait être effectué selon des indices de coûts de construction adaptés à la nature de destination. L'arbre de décision permettant d'estimer une probabilité de clôture en état sans suite des sinistres pourrait être modifié, par exemple en intégrant des techniques de *bagging* [Le Faou 2019] ou en supposant une structure de dépendance dépendante de variables exogènes. Les algorithmes d'intelligence artificielle, introduits en ouverture à la fin du deuxième chapitre, pourront mieux affiner la classification textuelle de la destination.

Finalement, nos travaux peuvent permettre une utilisation directe dans les différentes approches normatives de calcul des réserves, notamment selon la nouvelle norme IFRS17 qui incite à une sélection de maille selon la profitabilité, pouvant par ailleurs avoir une appréciation différente selon la destination des ouvrages.

Annexe A

Classification textuelle non supervisée

Classification non supervisée et détection des structures ne présentant aucun champs de destination (bruits) :

La partie suivante a été testée et a servi pour un traitement manuel de certains champs. Elle n'a cependant pas été utilisée pour la sélection de mots clés ni pour une détection directe des classes. On fournira dans la suite qu'une explication succincte. Le lecteur intéressé pourra se référer au livre [Rothman 2021] ou au livre [Tunstall et al. 2022] pour plus de détails sur l'implémentation des modèles.

En 2017, le monde du traitement textuel du langage (NLP [1]) a été bouleversé par l'introduction des modèles d'apprentissage profond, dites *Transformers*.
Il s'agit de modèles qui essaient de détecter des structures de langages, selon un entraînement d'architectures neuronales particulières et avec des millions de paramètres.
En effet, selon une architecture donnée qui comporte plusieurs paramètres, ce type de modèles se base sur un dictionnaire de mots et peuvent effectuer une correspondance avec les documents textuels. Après une initialisation des paramètres, ils essaient par la suite de prédire pour chaque *item* le mot ou la structure des mots qui suit selon une logique séquentielle. À travers une mesure d'erreur adéquate, une comparaison avec les vrais *items* permet d'entrainer au fur et à mesure le modèle avec des techniques d'optimisation.
En supposant que le modèle soit bien entrainé, on pourrait obtenir avec les paramètres optimisés, *in fine,* une représentation vectorielle pour chaque *item* textuel.
En réalité, les *Transformers* sont plus sophistiqués et traitent les données selon un mécanisme séquentiel introduisant des couches neuronales qui encodent les séquences d'entrées et des couches neuronales qui la décodent à la sortie.

1. Natural Language Processing

Nous avons testé un modèle *Transformer* appelé *CamemBERT*, développé par les chercheurs de l'INRIA [2] et apparu en 2020 [Martin et al. 2020]. Il s'agit d'un modèle entraîné sur 138 Gigabyte de données textuelles en français avec 110 millions paramètres.

Ce modèle pré-entrainé nous permettra d'avoir une représentation vectorielle (*Embedding*) de chaque phrase des destinations après l'étape de racinisation. En l'utilisant, on peut représenter chaque phrase selon un vecteur de taille 749, selon la définition de l'architecture du modèle.

Nous pourrons ensuite essayer de regrouper les vecteurs qui représentent des similarités selon une mesure de distance adéquate, à partir de méthodes de classification non supervisée telles que *Kmeans (1957)* ou encore d'algorithmes comme *DBSCAN (1996) ou HDBSCAN (2013)* qui permettent de détecter des points aberrants (bruits), s'agissant de vecteurs qui ne s'approchent pas localement d'un regroupement de points pouvant constituer une classe.

Ces étapes ont été suivies et ont permis de corriger manuellement quelques champs de destinations bruts qu'on n'a pas détecté après la sélection des mots les plus fréquents.
Par exemple, on a pu détecter une classe de destinations de type industrielles qu'on a retraité manuellement, cette classe contient des destinations brutes de type : *stockage*, *entrepôt* ou encore *Bâtiment logistique.*

champs_bruts	clus	clus_hdbscan	score_outlier_hdbscan
PARC LOGISTIQUE 2	2	10	0.030483
VENTE ET STOCKAGE	2	10	0.000000
VENTE ET SOCKAGE	2	10	0.000000
BATIMENT LOGISTIQUE	2	10	0.019710
ENTREPOT - STOCKAGE	2	10	0.002553
VENTE ET LOCATION	2	10	0.000000
ENTREPOT LOGISTIQUE	2	10	0.015696

FIGURE A.1 – Une classe détectée par classification par l'algorithme HDBSCAN

Remarquons que la racinisation des séquences "PARC LOGISTIQUE" et "VENTE ET STOCKAGE" ne représente pas de similarité de racine. L'algorithme a réussi à détecter leurs sens à travers leur représentation vectorielle (*Embedding*)
Le *score outlier* présent ici permet de détecter des structures d'aberrance localement. Plus le score est élevé, plus le point est susceptible d'être une valeur

2. Institut national de recherche en sciences et technologies du numérique

aberrante. Le lecteur pourra se référer à l'article introduisant l'algorithme HDBSCAN fournit plus de détails [McInnes et al. 2017].
Ce score nous a permis de détecter d'autres structures de textes aberrantes.

Bibliographie

Barbaste, M. (2017), 'Une méthode de provisionnement individuel par apprentissage automatique', Link. site consulté le 20 août 2022.
URL: *https ://www.institutdesactuaires.com/docs/mem/a6ab5fa0f52f8a741ba6f85c5f17bfaf.pdf*

Bornhuetter, R. and Ferguson, R. (1972), 'The actuary and ibnr', *Casualty Actuarial Society, .*

Bourry, C. (2016), 'Evaluation des provisions techniques et du capital économique assoccié au risque de réserve en assurance construction', Link. site consulté le 20 août 2022.
URL: *https ://www.institutdesactuaires.com/docs/mem/334249b8ec2b0fc32e22230bb083fcd5.pdf*

Brice, F. (EURIA 2021), Modèles de dépendance.

Elangovan, R., Mohanasundari, R. and Susiganeshkumar, E. (2020), 'Accelerated failure time model for survival analysis', *Journal of Information and Computational Science* **9**, 10–33.

Feinerer, I. (2007), Introduction to the tm package text mining in r.

Franck, V. (EURIA 2021), Cours d'apprentissage statistique.

François-Xavier Ajaccio, Albert Caston, R. P. (2022), L'assurance construction.

Gerber, G., Le Faou, Y., Lopez, O. and Trupin, M. (2018), The impact of churn on client value in health insurance, evaluation using a random forest under random censoring. working paper or preprint.
URL: *https ://hal.archives-ouvertes.fr/hal-01807623*

Gribkova, S. and Lopez, O. (2015), 'Non-parametric copula estimation under bivariate censoring', *Scandinavian Journal of Statistics* **42**(4), 925–946.
URL: *https ://EconPapers.repec.org/RePEc :bla :scjsta :v :42 :y :2015 :i :4 :p :925-946*

Le Faou, Y. (2019), Contributions à la modélisation des données de durée en présence de censure : application à l'étude des résiliations de contrats d'assurance santé, Theses, Sorbonne Université.
URL: *https ://tel.archives-ouvertes.fr/tel-03017164*

Leo Breiman, Jerome H. Friedman, R. A. O. and Stone., C. J. (1987), 'Classification and regression trees', *Cytometry* **8**(5), 534–535.
URL: *https ://onlinelibrary.wiley.com/doi/abs/10.1002/cyto.990080516*

Lopez, O. (2007), Réduction de dimension en présence de données censurées, Theses, ENSAE ParisTech.
URL: *https ://pastel.archives-ouvertes.fr/tel-00195261*

Lopez, O. (2018), A censored copula model for micro-level claim reserving. working paper or preprint.
URL: *https ://hal.archives-ouvertes.fr/hal-01706935*

Lopez, O., Milhaud, X. and Thérond, P.-E. (2015), 'Tree-based censored regression with applications to insurance', *Electronic Journal of Statistics* **10**.

Mack, T. (1993), 'Distribution-free calculation of the standard error of chain ladder reserve estimates', *ASTIN Bulletin* **23**(2), 213–225.

Martin, L., Muller, B., Ortiz Suárez, P. J., Dupont, Y., Romary, L., de la Clergerie, É., Seddah, D. and Sagot, B. (2020), CamemBERT : a tasty French language model, *in* 'Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics', Association for Computational Linguistics, Online, pp. 7203–7219.
URL: *https ://aclanthology.org/2020.acl-main.645*

McInnes, L., Healy, J. and Astels, S. (2017), 'hdbscan : Hierarchical density based clustering', *Journal of Open Source Software* **2**(11), 205.
URL: *https ://doi.org/10.21105/joss.00205*

Pröhl, C. and Schmidt, K. (2005), 'Multivariate chain-ladder'.

Renshaw, A. and Verrall, R. (1998), 'A stochastic model underlying the chain-ladder technique', *British Actuarial Journal* **4**(4), 903–923.

Rothman, D. (2021), *Transformers for Natural Language Processing : Build Innovative Deep Neural Network Architectures for NLP with Python, PyTorch, TensorFlow, BERT, RoBERTa, and More*, Packt Publishing.
URL: *https ://books.google.fr/books ?id=Ua03zgEACAAJ*

Satten, G. A., Datta, S. and Robins, J. (2001), 'Estimating the marginal survival function in the presence of time dependent covariates', *Statistics & Probability Letters* **54**(4), 397–403.
URL: *https ://ideas.repec.org/a/eee/stapro/v54y2001i4p397-403.html*

Stute, W. (1993), 'Consistent estimation under random censorship when covariables are present', *Journal of Multivariate Analysis* **45**(1), 89–103.
URL: *https ://www.sciencedirect.com/science/article/pii/S0047259X83710286*

Tunstall, L., von Werra, L. and Wolf, T. (2022), *Natural Language Processing with Transformers : Building Language Applications with Hugging Face*, O'Reilly Media.
URL: *https ://books.google.fr/books ?id=7hhyzgEACAAJ*

Vock, D., Wolfson, J., Bandyopadhyay, S., Adomavicius, G., Johnson, P., Vazquez-Benitez, G. and O'Connor, P. (2016), 'Adapting machine learning techniques to censored time-to-event health record data : A general-purpose approach using inverse probability of censoring weighting', *Journal of Biomedical Informatics* **61**, 119–131.

Zhang, Y. (2010), 'A general multivariate chain ladder model', *Insurance : Mathematics and Economics* **46**, 588–599.