Mémoire présenté le :

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

Par : Gretta CHAHWANE

Titre : *Provisionnement des sinistres climatiques en Multirisques Habitation : Techniques d'apprentissage statistique par évènement de grande ampleur.*

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

Membres présents du jury de l'Institut des Actuaires — Signature

Alexandre YOU

Sylvain JARRIER

Membres présents du jury de l'ISFA

Jean BRUNET

Entreprise :
Nom : AXA France

Signature :

Directeur de mémoire en entreprise :
Nom : David DELRIO

Signature : 

Invité :
Nom :

Signature :

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

Signature du responsable entreprise

Signature du candidat

GRANDE ÉCOLE D'ACTUARIAT ET DE GESTION DES RISQUES
INSTITUT DE SCIENCE FINANCIÈRE ET D'ASSURANCES

Université Claude Bernard Lyon 1

Institut de Science Financière et d'Assurances

Mémoire d'Actuariat

AXA France
Provisionnement des sinistres climatiques en Multirisques Habitation : Techniques d'apprentissage statistique par évènement de grande ampleur

---

Gretta Chahwane

Tuteur en entreprise : David DELRIO
Tuteur pédagogique : Aurélien COULOUMY

Résumé

L'activité d'assurance se caractérise par un cycle de production inversé : le paiement des primes a lieu avant la survenance des sinistres. Ainsi, le montant de ces derniers n'est pas nécessairement connu à l'avance, ce qui crée une dette pour l'assureur envers les assurés. Pour couvrir cet éventuel événement, cette dette figure au passif par la constitution d'une provision et doit être estimée. Le provisionnement est donc la solution à l'incertitude à laquelle toute entreprise d'assurance est confrontée.

En effet, il existe plusieurs méthodes de provisionnement en assurance non-vie et on en distingue les méthodes déterministes et stochastiques. Les méthodes déterministes sont des méthodes agrégées et ne tiennent pas compte des caractéristiques du contrat, sinistre par sinistre. De plus, ces méthodes nécessitent la vérification des hypothèses et ne sont pas toujours vérifiées dans un cadre opérationnel, surtout pour des branches volatiles, telles que les sinistres liés à des événements climatiques en assurance multirisques habitation.

De plus, les défis posés par les changements climatiques et les événements météorologiques inhabituels sont de plus en plus préoccupants de nos jours. Selon le rapport de France Assureurs sur l'impact du changement climatique sur l'industrie de l'assurance à l'horizon 2050, les paiements cumulés des assureurs pourraient atteindre environ 143 milliards d'euros d'ici 2050, comparé à 74,1 milliards d'euros de 1988 en 2019. Ainsi, au cours des 25 prochaines années, les dommages liés au climat pourraient presque doubler.

Étant donné l'impact financier que ces événements peuvent engendrer, il est crucial pour les compagnies d'assurance d'estimer avec précision la charge ultime de ces événements dès leur survenance. Ainsi, ce mémoire propose des méthodes de provisionnement basées sur des méthodes d'apprentissage statistique afin d'estimer, dès les premiers jours de survenance, la charge ultime d'un évènement climatique de grande ampleur.

Mots-clés : ***Assurance Multirisques Habitation, évènements climatiques, provisionnement, charge ultime, apprentissage statistique, prédiction, K-voisins les plus proches (KNN), arbre de décision, forêt aléatoire, Boosting, Bagging.***

Abstract

Insurance is a process of financial intermediation characterized by an inverted production cycle whereby policy holders pay premiums before the occurrence of claims. Due to the uncertain aspect of future claims, a debt is created between the insurer and the insured. The creation of this debt is recorded in the liabilities of the insurance company. This debt is what we call reserve and should be calculated using estimation and projection techniques. In the face of timing and amount uncertainty faced by insurance companies, reserving provides a solution to this problem.

Indeed, there are several reserving techniques in non-life insurance of which a distinction is made between deterministic and stochastic methods. The deterministic methods are aggregated methods and do not consider the characteristics of the contract, claim by claim. Additionally, these methods require the verification of assumptions and are not always verified in real life, especially for volatile sectors, such as claims related to climatic events in property insurance.

Specifically, the challenges posed by climate change and unusual weather events are increasingly concerning today. According to the « France Assureurs » report on the impact of climate change on the insurance industry by 2050, cumulative payments by insurers could reach around 143 billion euros by 2050, compared to 74,1 billion euros from 1988 to 2019. Thus, over the next 25 years, climate-related damages could almost double.

Given the financial impact that these events can generate, it is crucial for insurance companies to accurately estimate the ultimate cost of these events from their occurrence. Therefore, this thesis proposes reserving methods based on machine learning methods to estimate, from the very first days of occurrence, the ultimate cost of climatic events.

Keywords : ***Property insurance, climatic events, individual reserving, ultimate cost, Machine learning, prediction, K-Nearest Neighbors (KNN), decision tree, random forest, Boosting, Bagging.***

Remerciements

Tout d'abord, je tiens à remercier chaleureusement mon manager, M. Renaud MOUYRIN, pour son accueil au sein de l'équipe Data Analytics et ses précieux conseils qui ont enrichi mon expérience tout au long de mes deux dernières années d'alternance.

Il est essentiel pour moi de souligner l'aide précieuse et le soutien inestimable que j'ai reçus de mon tuteur en entreprise, M. David DELRIO, tout au long de l'élaboration de ce mémoire. Je tiens à le remercier pour sa disponibilité constante, son encadrement et le temps qu'il m'a accordé afin de me transmettre sa forte connaissance et compétence dans le domaine des sciences actuarielles.

Je tiens à exprimer ma gratitude à toutes les personnes qui m'ont accueilli au sein du service comptes PP d'AXA France IARD à Marseille et qui m'ont permis de mener à bien ces deux années d'alternance dans les meilleures conditions.

Mes remerciements s'adressent également à mon tuteur académique, M. Aurélien COULOUMY, pour son aide et sa disponibilité tout au long de mon mémoire.

Je n'oublie pas de remercier ma famille et mes amis qui, par leur soutien émotionnel et leurs encouragements constants pendant mes années d'études, ont joué un rôle crucial dans la réalisation de ce travail.

Note de synthèse

AXA France IARD (Incendie, Accidents et Risques Divers) PP (Professionnels et Particuliers) est une division d'AXA France qui propose des produits d'assurance pour les particuliers et les professionnels, tels que l'assurance automobile, l'assurance multirisques habitation (MRH), l'assurance responsabilité civile et l'assurance des professionnels et des petites entreprises.

Au sein de cette division se trouve l'équipe *Data And Business Analytics* dans laquelle ce mémoire a été développé et rédigé.

Initialement, l'équipe a été créée pour soutenir l'équipe d'inventaire non-vie en leur assurant l'exhaustivité et la fiabilité des données ainsi que des triangles de liquidation. Actuellement, elle dispose d'un large panel de processus mensuels et ses différents objectifs s'inscrivent dans :

- Le calcul des provisions techniques.
- L'estimation, l'extraction et le traitement des données liées aux contrats et sinistres des assurés.
- Le développement et la création d'outils optimisés pour l'analyse d'indicateurs de suivi de sinistralité et de rentabilité (par exemple : outils de suivi d'évolution de la fréquence des sinistres et du coût moyen des sinistres, et cela à des mailles assez fines).

A l'inverse d'une activité commercial classique, l'activité de l'assurance est caractérisée par une inversion de son cycle de production. Ainsi, la charge des prestations à régler n'est pas connu avant la survenance d'un sinistre. Cela étant dit, l'assureur se trouve dans l'obligation de provisionner et donc de constituer une provision : montant qu'il lui faut mettre de côté pour respecter ses engagements futurs. Le provisionnement est un mécanisme clé en Actuariat.

Bien que l'assureur ait l'obligation de provisionner afin de respecter ses engagements, ce processus est également soumis à un cadre réglementaire. Les provisions, constituées au passif du bilan de l'assurance, sont encadrées par la directive européenne Solvabilité II (S2). Cette dernière impose le concept de provisionnement en *Best Estimate* (BE) ou meilleure estimation et exige que les provisions soient évaluées de manière optimale en évitant tout sur-provisionnement ou sous-provisionnement.

Il existe deux principales familles de méthodes de provisionnement : les méthodes déterministes et les méthodes stochastiques. D'une part, les méthodes déterministes reposent sur des règles et des tendances historiques. D'autre part les méthodes stochastiques intègrent l'incertitude et la variabilité des sinistres futurs. Ainsi, la principale différence entre ces deux approches se trouve dans le fait que les méthodes stochastiques fournissent

des estimations de variance et d'intervalle de confiance des charges prédites. En pratique, les méthodes les plus couramment utilisées restent les méthodes déterministes, notamment la méthode de Chain Ladder, qui est une méthode agrégée.

Plusieurs types de provisions existent et ce mémoire se focalise sur la PSAP (Provision pour Sinistre A Payer). Cette provision représente le montant restant à payer pour les sinistres déjà déclarés ainsi que pour ceux qui ne le sont pas encore. Elle est composée de :

- Réserves observées : provisions établies par les gestionnaires de sinistres.
- Provisions *IBNR* (*Incurred But Not Reported*) qui incluent :
  - Provisions *IBNYR* (*Incurred But Not yet Reported*) : provisions pour les sinistres déjà survenus mais non déclarés.
  - Provisions *IBNER* (*Incurred But Not Enough Reported*) : provisions pour les sinistres déjà déclarés mais sous-provisionnés.

Enfin, pour obtenir la CFP (charge finale prévisible) ou en d'autres termes la charge des sinistres à l'ultime, il suffit d'additionner la PSAP aux règlements déjà effectués.

Dans un contexte où les risques climatiques émergents exercent une pression croissante sur l'industrie de l'assurance, la question du provisionnement des sinistres climatiques de grande ampleur devient un enjeu majeur. La branche MRH est particulièrement concernée, notamment face à l'augmentation de la fréquence des événements climatiques tels que la grêle et les tempêtes, observée ces dernières années en France et à l'échelle mondiale.

Ce mémoire s'inscrit dans cette problématique en explorant des méthodes de provisionnement adaptées à ces sinistres. Plus précisément, il propose des approches améliorées et alternatives aux méthodes agrégées traditionnelles (comme Chain Ladder ou la méthode budgétaire), en s'appuyant sur des méthodes individuelles basées sur l'apprentissage statistique ou *Machine Learning*. L'objectif est d'intégrer ces nouvelles approches tout en respectant les exigences réglementaires, afin d'améliorer la précision et la robustesse du provisionnement des sinistres climatiques.

Pour cela, le cadre général est présenté dans le premier chapitre. Ensuite, le deuxième chapitre décrit la méthodologie adoptée, en détaillant les méthodes de provisionnement existantes ainsi que celles candidates pour répondre à la problématique. Enfin, le dernier chapitre est consacré à l'application directe des méthodes de provisionnement présentées dans le chapitre 2 sur une base de données d'AXA. Les résultats de cette application sont analysés et comparés afin d'identifier le modèle de provisionnement le plus performant.

Initialement, l'équipe d'AXA applique une méthode budgétaire pour provisionner les sinistres climatiques. Cette approche repose sur l'établissement d'un budget initial pour

les sinistres climatiques, basé principalement sur la moyenne des coûts des années précédentes et des hypothèses sur le coût moyen, la fréquence et l'exposition. Ce budget est ensuite ajusté en fonction de l'évolution des événements climatiques au cours de l'année : il peut être maintenu, réduit en l'absence de sinistres pour financer d'autres dommages, ou augmenté si nécessaire. L'objectif est donc d'allouer une enveloppe de charge spécifique aux sinistres climatiques et de l'adapter selon la réalité des événements survenus.

Toutefois, cette approche reste globale et peut manquer de précision, d'où l'intérêt d'approches plus fines.

Une méthode alternative à l'approche budgétaire est une méthode agrégée par évènement climatique. Connaissant les sept premiers jours de développement du sinistre, cette approche repose sur :

- L'estimation du nombre finale prévisible (NFP) de sinistres : en projetant le nombre de sinistres connus selon le développement de tous les sinistres antérieurs à l'évènement.
- L'estimation du nombre de sinistres clos sans suite (NBCSS) : en estimant le taux de sinistre de sinistres clos sans suite pour l'année de survenance du sinistre, selon les taux historiques, tous évènements confondus.
- L'estimation du coût moyen (CM) des sinistres : en calculant le CM du sinistre sur la première semaine.

Cette méthode prédit la charge à l'ultime (différence entre le NFP et le NBCSS estimé et qui correspond au nombre de sinistres avec suite (NBAS), multiplié par le CM estimé) d'un événement survenu en tenant compte de son développement au cours de la première semaine. Ce choix est basé sur une étude concluante indiquant que le nombre/charge de sinistres tend à se stabiliser à partir du huitième jour et commence à se rapprocher de sa valeur finale. L'intérêt d'une telle approche est de déterminer rapidement et dès les premiers jours de survenance d'un sinistre climatique, la charge finale associée.

Ce mémoire propose, en plus de la méthode classique abordée par l'équipe d'AXA, une méthode agrégée par événement « améliorée ».

D'une part, une amélioration est apportée à l'estimation du nombre de sinistres. La détermination du NFP de sinistres est effectuée en procédant par regroupement d'événements selon le développement du nombre de sinistres, grâce à l'algorithme de *K-Means*. La deuxième amélioration concerne l'estimation du NBCSS de sinistres : des taux de sinistres clos sans suite sont déterminés en fonction du type d'événement.

D'autre part, une amélioration porte sur l'estimation du coût moyen des sinistres d'un événement, appliqué au NBAS de sinistres. En effet, la méthode classique se contente d'un coût moyen observé une semaine après la survenance. L'étude réalisée sur des sinistres

climatiques historiques montre que le coût moyen a tendance à diminuer avec le temps. Ainsi, il est pertinent de projeter le coût moyen observé au septième jour à l'ultime, en tenant compte de son évolution.

Le tableau ci-dessous décrit synthétiquement les deux méthodes agrégées :

Étapes	Méthode classique AXA	Méthode améliorée	Détails de l'amélioration
Estimation du NFP de sinistres	→ Le nombre de sinistres d'un nouvel événement est supposé se développer de la même façon que les événements antérieurs. → Projection du nombre de sinistres à l'ultime à partir du septième jour de survenance.	→ Regroupement des événements climatiques en fonction de l'évolution du nombre de sinistres au cours des sept premiers jours suivant leur survenance. → Projection du nombre de sinistres à l'ultime à partir du septième jour, en s'appuyant sur le développement des événements antérieurs appartenant au même groupe.	La méthode améliorée permet de **capturer les ressemblances de développement du nombre** de sinistres selon l'événement, plutôt que d'appliquer un développement unique à l'ensemble des événements antérieurs.
Estimation du NBCSS de sinistres	→ Détermination du taux de clos sans suite de l'année de survenance de l'événement, tous événements confondus.	→ Détermination du taux de clos sans suite de l'année de survenance de l'événement, par type d'événement.	Le taux de sinistres clos sans suite **par événement** permet d'obtenir un NBCSS plus proche de la réalité.
Estimation du coût moyen	→ Coût moyen des sinistres selon les sept premiers jours.	→ À partir du septième jour, projection du coût moyen à l'ultime.	La méthode améliorée permet **de capturer l'évolution (à la baisse) du coût moyen** à l'ultime et d'éviter un sur-provisionnement.

TABLE 1 – Description et comparaison des étapes des méthodes agrégées (classique et améliorée)

Pratiquement, ces deux méthodes sont implémentées sur une base d'évènements

climatiques des années 2021 et 2022. Chaque méthode est appliquée sur la base des évènements de l'année 2021 afin de prédire les charges ultimes des sinistres climatiques de l'année 2022. Pour une charge observée de 217 007 999€ de charges en 2022, la méthode agrégée classique estime 265 817 051€ (soit un écart de 48 809 052€), alors que la méthode améliorée estime 244 670 137€ (soit un écart de 27 662 138€). Les deux modèles sur-estiment la charge des sinistres climatique de 2022.

Le tableau ci-dessous récapitule les résultats obtenus grâce à l'application de ces deux méthodes à notre base de données.

Méthode	NFP de sinistres estimé	NFP de sinistres réel	NBCSS de sinistres estimé	NBCSS de sinistres réel	Charge estimée (en €)	Charge réelle (en €)	Écart (en €)
Agrégée classique	41 954	36 719	11 633	6 441	265 817 051	217 007 999	48 809 052
Agrégée améliorée	39 940		7 398		244 670 137		27 662 138

TABLE 2 – Résultats des prédictions des modèles agrégés pour les évènements climatiques de 2022

Bien que la méthode agrégée améliorée réduit l'écart entre la charge réellement observée et celle prédite, ce mémoire cherche à proposer une autre méthode plus rapide et plus robuste, qui réduit davantage les écarts entre les prédictions du modèle et la charge observée : la méthode de provisionnement individuelle. Cette méthode s'appuie sur des modèles d'apprentissage statistique ou *Machine Learning*.

Il existe plusieurs types d'apprentissage statistique mais ce mémoire s'intéresse aux modèles supervisés. Ces algorithmes établissent la relation entre une variable cible, ici la charge des sinistres, et plusieurs variables explicatives pouvant représenter les caractéristiques de l'assuré, de son contrat, de son sinistre et même des informations exogènes comme des variables de météorologie. Ainsi, l'objectif est de prédire la variable cible (quantitative ici) à partir des variables explicatives. La variable cible étant numérique, il s'agit d'un problème de régression.

Au sein de la famille des modèles supervisés, il existe des modèles paramétriques et non paramétriques mais cette étude se concentre uniquement sur les modèles non paramétriques, qui ne reposent sur aucune hypothèse concernant la loi de probabilité de la variable cible. Parmi ces algorithmes, se trouve les modèles les plus connus basés sur des techniques de *Bagging* et de *Boosting*.

Le modèle retenu pour la modélisation de la charge des sinistres est l'*eXtreme Gradient Boosting* (*XGBoost*). Ce choix a été orienté par les avantages qu'il présente : rapidité, précision et capacité de traitement de données volumineuses. Le *XGBoost* repose sur le modèle des arbres de décisions *Classification And Regression Trees* (*CART*) et permet

de réduire les biais et la variance des estimateurs par une application particulière du *Boosting*. Le principe de l'algorithme consiste à créer une série d'arbres *CART*, chacun se concentrant sur les erreurs commises par les précédents, ce qui améliore progressivement les prédictions. Le processus commence par la construction d'un arbre simple, puis les erreurs de cet arbre servent à créer un suivant, et ainsi de suite. Cette méthode produit une série d'arbres combinés qui offrent des prévisions plus précises pour les individus ayant des caractéristiques particulières. Ainsi, l'arbre final réduit considérablement la variance par rapport à celui de départ.

Les méthodes d'apprentissage statistique nécessitent un historique de données suffisamment profond. Ainsi, les modèles individuels sont entraînés sur les huit années précédant 2022, soit de 2014 à 2021. Cette approche repose sur une base de données détaillée, organisée ligne à ligne, intégrant la variable cible ainsi qu'un maximum d'informations afin d'optimiser la prédiction de la charge ultime. La base, qui recense les sinistres survenus entre 2014 et 2022, est extraite des bases de données d'AXA et enrichie à la fois par des variables explicatives fournies par les équipes d'AXA (comme le type de bien ou la qualité de l'occupant) et par des variables exogènes, notamment liées aux conditions météorologiques. La base de données finale a été nettoyée de toute anomalie et incohérence afin d'optimiser la performance des modèles. 171 083 sinistres et une quarantaine de variables sont enregistrés.

Une analyse des coûts et des nombres de sinistres sur les années de la base de données montre que le nombre de sinistres a régulièrement augmenté au cours des années, avec une forte croissance en 2022. La charge des sinistres en 2022 est plus élevée que celle des années précédentes, bien que la distribution de la charge des sinistres des années antérieures soit similaire. Cela suggère que l'année 2022 présente des caractéristiques atypiques qui influencent la charge des sinistres.

L'analyse des sinistres liés à la grêle montre que leur proportion a considérablement augmenté en 2022 par rapport aux années précédentes et les coûts des sinistres grêle sont significativement plus élevés que ceux des autres événements.

En raison de ces différences marquées dans les dynamiques de coûts et de fréquence des sinistres grêle, il est suggéré de développer deux modèles distincts :

- Un modèle pour les événements de grêle.
- Un modèle pour les autres types d'événements.

Sachant que les modèles sont entrainés sur les années antérieurs à 2022, cela permet d'éviter de biaiser le modèle unique et de sous-estimer les coûts des sinistres grêle. Pour pouvoir comparer les résultats du modèle individuelle à ceux des modèles agrégés, les prédictions des deux modèles seront sommées pour donner la charge ultime estimée de 2022.

Les charges de sinistres évoluent au fil du temps, influencées par des facteurs comme l'inflation et des événements exceptionnels, comme les épisodes de grêle en 2022. Ainsi, pour rendre les montants des charges de sinistres des années précédentes comparables à ceux de 2022, il est essentiel de les normaliser. Pour cela, les charges « *as if* » sont introduites et elles sont ajustées en appliquant un coefficient basé sur un indicateur spécifique, ce qui permet de neutraliser les effets liés aux variations temporelles, comme l'inflation et les événements extrêmes. Deux indicateurs ont été testé pour chaque modèle :

- L'indicateur de la Fédération Française du Bâtiment (FFB) et l'indicateur de Bâtiments-travaux (BT) :

  Cet indicateur permet de prendre en compte l'inflation en intégrant les évolutions et les variations du coût de la construction et des réparations, de la main-d'œuvre et d'autre charge pour les bâtiments. Ainsi, il reflète bien les coûts associés aux sinistres en MRH et justifie le choix d'un tel indicateur. La méthodologie consiste à ajuster la charge en fonction de l'évolution du cout de la construction et de la réparation des dommages à la suite des évènements climatiques.

  Cependant pour chaque modèle (grêle et hors grêle) l'approche sera légèrement différente. Pour le modèle hors grêle, on se base sur les indicateurs globaux de la FFB. Alors que pour le modèle grêle on se base sur un maillage de cet indicateur qui correspond aux indicateurs BT.

  Dans ces deux approches, c'est l'impact du coût de réparation et de reconstruction sur les charges des sinistres qui est pris en compte.

- L'indicateur de coût moyen J+7 :

  La méthode du CM J+7 repose sur une approche davantage orientée métier que macroéconomique. D'un point de vue actuariel, cette approche permet d'intégrer la dynamique d'évolution réelle du coût des indemnisations des sinistres au fil du temps. Elle vise à ajuster la charge des sinistres (grêle et hors grêle) en fonction du coût moyen observé après une semaine.

  Cette méthodologie permet de prendre en compte l'évolution réelle du coût des sinistres réglés au bout de sept jours.

Quatre modèles sont testés afin de modéliser les quatre charges « *as if* ». Avant toute modélisation, la base de données est divisée aléatoirement en trois sous-bases :

Base	Description	Intérêt
Apprentissage	80% de la base des sinistres de 2014 à 2021.	→ Entraînement du modèle. → Calibrage du modèle.
Test	20% de la base des sinistres de 2014 à 2021.	→ Évaluation des erreurs du modèle.
Validation	La base de données des sinistres de 2022	→ Choix des charges « *as if* ». → Comparaison de la charge estimée avec celle obtenue grâce aux modèles agrégés.

TABLE 3 – Division de la base de données

Sachant que la base de données contient plus de 40 variables, une première étape consiste à identifier les variables les plus significatives pour la prédiction de la charge des sinistres. Cette sélection permet d'améliorer la pertinence du modèle tout en réduisant le temps et la complexité de la modélisation. Bien que la sélection des variables ne soit pas une étape indispensable, elle permet d'alléger significativement la modélisation sous *XGBoost* et d'optimiser le temps de calcul.

Pour éviter le surapprentissage, le modèle est entraîné sur 80% de la base de données des sinistres antérieurs à 2021, tandis que les 20% restants, non utilisés pour l'entraînement, servent à évaluer ses performances. L'objectif est de sélectionner le modèle offrant les meilleures prédictions.

L'algorithme offre une large possibilité de paramétrage. Cependant, pour des raisons de contraintes opérationnelles liées aux temps de calcul, seuls les paramètres suivants ont été calibrés :

Hyperparamètre	Valeur
nrounds	Le nombre total d'arbres à agréger
max_depth	Le nombre de feuilles maximales des arbres
eta	Le taux d'apprentissage qui ajuste l'impact de chaque arbre sur la prédiction finale
subsample	La proportion des observations utilisées pour l'entraînement de chaque arbre
min_child_weight	Le poids minimum requis pour générer un nœud fils
colsample_bytree	La proportion des variables explicatives sélectionnées pour entraîner chaque arbre

TABLE 4 – Hyperparamètres à optimiser pour le modèle *XGBoost*

Le calibrage du modèle ou le tuning des hyperparamètres est une étape cruciale qui permet de sélectionner les paramètres de l'algorithme afin qu'il soit le plus ajusté possible aux données, tout en évitant le sur-apprentissage. Ce phénomène survient lorsque le modèle devient trop spécifique aux données historiques, rendant ainsi ses prédictions peu fiables sur de nouvelles données.

La méthode consiste à définir une grille de paramètres et à tester toutes les combinaisons possibles en entraînant un modèle sur la base d'apprentissage, puis en mesurant son erreur sur la base de test. Le modèle qui minimise le critère *Root Mean Square Error* (*RMSE*) ou racine de l'erreur quadratique moyenne est retenu.

Le tableau ci-dessous présente, pour chaque modèle calibré, la prédiction de la charge des sinistres climatiques en 2022 :

Modèle	« As if »	Valeurs prédites en (K€)	Valeurs observées (en K€)	RMSE
Grêle	**CM J+7**	175 127,00	173 902,49	16,74
	Indice BT	154 341,90		16,95
Hors grêle	**CM J+7**	43 130,14	43 105,50	9,17
	Indice FFB	48 403,30		9,21

TABLE 5 – Prédiction des quatre modèles selon la méthode « *as if* » *XGBoost* sur la base de validation

Les prédictions sur l'année 2022 de la base de validation des modèles sur les charges « as if » de CM semblent être les plus proches de la charge réellement observée. Ainsi, la charge

« *as if* » retenue pour les deux modèles est celle du CM à J+7.

La dernière étape de la modélisation individuelle consiste à estimer la charge des sinistres clos sans suite ainsi que celle des sinistres non encore déclarés. Ainsi, la charge des sinistres climatiques estimée correspond à la charge prédite par le modèle, diminuée de la charge des sinistres clos sans suite et augmentée de la charge des sinistres tardifs.

	Charge 2022 prédite (en K€)	**Charge 2022 observée (en K€)**
Méthode agrégée classique	265 817,05	217 007,99
Méthode agrégée améliorée	244 670,13	
Méthode individuelle	228 331,90	

TABLE 6 – Tableau comparatif des prédictions des différentes méthodes

En conclusion, l'approche individuelle se révèle plus performante pour prédire la charge des sinistres climatiques. En prenant en compte un ensemble de variables explicatives détaillées, cette méthode permet d'obtenir des estimations plus précises et mieux ajustées à la réalité. Contrairement aux approches agrégées, elle capture davantage les dynamiques spécifiques aux sinistres, réduisant ainsi les biais et améliorant la fiabilité des prédictions.

Summary

AXA France P&C (Property and Casualty) PP (Professional and Personal) is a branch of AXA France that provides insurance products for individuals and businesses, including car insurance, multi-risk home insurance (MRH), liability insurance, and coverage for professionals and small enterprises.

Within this branch, the Data and Business Analytics team is responsible for developing and writing this report.

Originally, the team was created to support the non-life inventory team by ensuring the completeness and reliability of data as well as claims settlement triangles. Currently, it oversees various monthly processes, with key objectives such as :

- Calculating technical reserving.
- Estimating, extracting, and processing data related to policyholders' contracts and claims.
- Developing and optimizing tools for analyzing claim trends and profitability indicators (for example, tools to monitor changes in claim frequency and average claim costs at a detailed level).

Unlike a typical commercial activity, the insurance sector operates with an inverted production cycle. This means that the cost of compensations to be paid is unknown until a claim actually occurs. As a result, insurers must allocate reserves, funds set aside to meet future obligations. Reserving is a fundamental aspect of actuarial science.

While insurers are required to establish reserves to meet their commitments, this process is also governed by regulations. These reserves, recorded as liabilities on the insurance company's balance sheet, are regulated by the European Solvency II (S2) directive. This directive enforces the concept of Best Estimate (BE) reserving and mandates that provisions be optimally evaluated to prevent over- or under-reserving.

There are two primary categories of reserving methods : deterministic and stochastic. Deterministic methods rely on historical trends and predefined rules, whereas stochastic methods incorporate uncertainty and variability in future claims. The main distinction between these approaches is that stochastic methods provide variance estimates and confidence intervals for predicted costs. In practice, deterministic methods remain the most widely used, particularly the Chain Ladder method, which is an aggregate approach.

Various types of reserves exist, and this report focuses on Claims Reserves (PSAP for Provisions pour Sinistres A Payer in french). This reserve represents the remaining amount to be paid for both reported and unreported claims. It consists of :

- Observed reserves : provisions set aside by claims adjusters.

- IBNR (Incurred But Not Reported) reserves, which include :
  - IBNYR (Incurred But Not Yet Reported) : reserves for claims that have occurred but have not yet been reported.
  - IBNER (Incurred But Not Enough Reported) : reserves for reported claims that have been underestimated.

To calculate the final expected cost (FEC), also known as the ultimate claims cost, the Claims Reserves is added to the amounts already paid.

As climate-related risks continue to rise, they are putting increasing pressure on the insurance industry. Large-scale weather-related claims are becoming a major concern, particularly in the P&C sector, due to the growing frequency of climate events such as hailstorms and windstorms observed in France and worldwide.

This thesis examines this issue by exploring reserving methods suited for such claims. Specifically, it evaluates improved and alternative approaches to traditional aggregate methods (such as Chain Ladder and budget-based approaches), incorporating individual-based methods that leverage Machine Learning. The aim is to integrate these innovative techniques while complying with regulatory requirements, enhancing the accuracy and robustness of climate claims reserving.

The first chapter introduces the general framework. The second chapter outlines the methodology, detailing both existing reserving methods and potential alternatives to address the issue. The final chapter focuses on applying these reserving methods to an AXA database, analyzing and comparing the results to identify the most effective approach.

Initially, the team applied a budget-based method to reserve climate claims. This approach involves setting an initial budget for climate-related claims, mainly based on the average costs of previous years and assumptions regarding average cost, frequency, and exposure. This budget is then adjusted based on climate events occurring throughout the year and it may be maintained, reduced if no claims arise (allowing funds to be redirected), or increased if necessary. The objective is to allocate a dedicated budget for climate claims and adjust it according to real-world occurrences.

However, this approach remains broad and may lack precision, highlighting the need for more refined methodologies.

An alternative to the budget-based approach is an aggregate method based on climate events. By analyzing the first seven days of claim development, this approach involves :

- Estimating the final expected number (FEN) of claims : by projecting the known claims count based on the development of past claims from similar events.

- Estimating the number of claims closed without follow-up (NBCC NFU) : by determining the closure rate of claims based on historical data for all events.
- Estimating the average claim cost (AC) : by calculating the average claim cost observed after the first week.

This method predicts the ultimate claim cost (the difference between the estimated FEN and NBCC NFU, representing the number of claims with follow-up (NBC FU), multiplied by the estimated AC) of an event while considering its development within the first week. This choice is supported by a study indicating that claim numbers and costs tend to stabilize from the 8th day onwards, approaching their final value. The advantage of this approach is that it enables a quick estimation of the final cost of a climate-related claim within just a few days.

This thesis proposes, in addition to the classical method used by the AXA team, an « enhanced » event-based aggregated method.

First, an improvement is made to the estimation of the number of claims. The determination of the FEC is carried out by grouping events based on the development of the number of claims, using the K-Means algorithm. The second improvement concerns the estimation of the NBCC NFU : closure rates are determined according to the type of event.

Furthermore, an improvement is introduced in estimating the AC for a given event, applied to the NBC FU. The basic method relies on the observed average cost one week after the occurrence. However, the study conducted on historical weather-related claims shows that the average cost tends to decrease over time. Therefore, it is relevant to project the observed average cost at day seven to its ultimate value while considering its evolution.

The table below provides a summary of both aggregate methods :

Steps	Aggregate method AXA	Enhanced method	Details of improvement
Estimation of FEC claims	→ The number of claims from a new event is assumed to develop in the same way as previous events. → Projection of the ultimate number of claims from the seventh day after occurrence.	→ Grouping of climate events based on the evolution of the number of claims during the first seven days following their occurrence. → Projection of the ultimate number of claims from the seventh day onwards, relying on the development of past events belonging to the same group.	The improved method **captures similarities in the development of the number of claims** based on the event rather than applying a single development pattern to all past events.
Estimation of NBCC NFU of claims	→ Determination of the claim closure rate without follow-up for the event occurrence year, across all events.	→ Determination of the claim closure rate without follow-up for the event occurrence year, by event type.	The claim closure rate without follow-up **per event** allows for a more accurate NBCSS.
Estimation of the AC	→ Average claim cost based on the first seven days.	→ From the seventh day onwards, projection of the ultimate average cost.	The improved method **captures the downward evolution of the** ultimate **average cost** and helps avoid over-reserving.

TABLE 7 – Description and comparison of the steps of the aggregated methods (classical and improved)

In practice, these two methods are implemented on a database of climate events from 2021 and 2022. Each method is applied to the 2021 events to predict the ultimate costs of climate claims for 2022. For an observed 2022 claims cost of 217,007,99K €, the classical aggregate method estimates 265,817,05K € (a difference of 48,809,05K €), while the imporved method estimates 244,670,13K € (a difference of 27,662,13K €). Both models overestimate the 2022 climate claims costs.

The table below summarizes the results obtained from applying these two methods to our database.

Method	FEC of estimated claims	FEC of real claims	NBCC NFU of estimated claims	NBCC NFU of real claims	Estimated cost (in €)	Real cost (in €)	Difference (in €)
Aggregate classical	41 954	36 719	11 633	6 441	265 817 051	217 007 999	48 809 052
Aggregate enhanced	39 940		7 398		244 670 137		27 662 138

TABLE 8 – Results of the aggregated model predictions for the 2022 climate events

While the improved aggregate method reduces the gap between the actual observed costs and predicted costs, this study aims to propose a faster, more robust method that further reduces the differences between model predictions and observed costs : the individual reserving method. This method uses Machine Learning models.

There are several types of statistical learning, but this study focuses on supervised models. These algorithms establish the relationship between a target variable (here, claim costs) and several explanatory variables that can represent policyholder characteristics, contract details, claim information, and even external data like weather variables. The goal is to predict the target variable (which is numerical here) from the explanatory variables. Since the target variable is numerical, this is a regression problem.

Within supervised models, there are parametric and non-parametric models, but this study focuses only on non-parametric models, which don't make assumptions about the probability distribution of the target variable. Among these algorithms are well-known models based on Bagging and Boosting techniques.

The chosen model for claim cost prediction is eXtreme Gradient Boosting (XGBoost). This choice was guided by its advantages : speed, accuracy, and ability to handle large datasets. XGBoost is based on Classification And Regression Trees (CART) and reduces bias and variance through a specific Boosting application. The algorithm creates a series of CART trees, each focusing on errors made by previous ones, gradually improving predictions. The process starts with a simple tree, then uses its errors to create the next tree, and so on. This produces a series of combined trees that offer more accurate predictions for cases with specific characteristics, significantly reducing variance compared to the initial tree.

Machine Learning techniques require sufficiently deep historical data. Thus, individual models are trained on the eight years before 2022 (2014-2021). This approach uses a detailed, row-by-row database including the target variable and as much information as possible to optimize ultimate cost prediction. The database, containing claims from 2014 to 2022, comes from team databases and is enriched with explanatory variables from AXA

teams (like property type or occupant quality) and external variables, particularly weather data. The final database was cleaned of all anomalies and inconsistencies to optimize model performance, containing 171,083 claims and around forty variables are reconded.

An analysis of costs and claim numbers shows that claim counts increased steadily each year, with strong growth in 2022. 2022 claim costs were higher than previous years, though cost distributions for earlier years were similar. This suggests 2022 had atypical characteristics affecting claim costs.

Analysis of hail-related claims shows their proportion increased significantly in 2022 compared to previous years, with hail claim costs being much higher than other events.

Due to these marked differences in hail claim cost and frequency patterns, it's suggested to develop two separate models :

- A model for hail events.
- A model for other event types.

Since models are trained on pre-2022 data, this avoids biasing a single model and underestimating hail claim costs. To compare individual model results with aggregate models, predictions from both models will be summed to give the estimated ultimate cost for 2022.

Claim costs change over time due to factors like inflation and exceptional events (for example, 2022 hail episodes). To make past years' claim costs comparable to 2022, normalization is essential. That's why « as if » costs are introduced, adjusted using coefficients based on specific indicators to account for time-related variations like inflation and extreme events. Two indicators were tested for each model :

- The indicator of the « Fédération Française du Bâtiment » (FFB) and the « Bâtiments-travaux » (BT) indicator :

  This indicator accounts for inflation by tracking changes in construction/repair costs, labor, and other building-related expenses. It well reflects P&C claim costs, justifying its use. The methodology adjusts costs based on construction/repair cost changes after climate events.

  However, approaches differ slightly between hail and non-hail models. The non-hail model uses global FFB indicators, while the hail model uses a more detailed BT indicator version. Both approaches account for repair/reconstruction cost impacts on claim costs.

  In both approaches, the impact of repair and reconstruction costs on claim expenses is taken into account.

- The D+7 AC indicator : :

  The D+7 AC method takes a more business-oriented than macroeconomic approach. From an actuarial perspective, it captures real compensation cost trends over time. It adjusts claim costs (hail and non-hail) based on average costs observed after one week.

  This methodology accounts for actual cost changes in claims settled after seven days.

Four models are tested to predict the four « as if » costs. Before modelling, the database is randomly split into three subsets :

Database	Description	Aim
Training	80% of the claims database from 2014 to 2021.	→ Training of the model. → Tuning of the model.
Test	20% of the claims database from 2014 to 2021.	→ Evaluation of model errors.
Validation	The 2022 claims database.	→ Choix des charges « *as if* ». → Comparaison of the estimated cost with the too aggregate method.

TABLE 9 – Division of the database

Given that the database contains more than 40 variables, a first step involves identifying the most significant variables for predicting claim costs. This selection improves the model's relevance while reducing modeling time and complexity. Although variable selection isn't absolutely necessary, it significantly streamlines XGBoost modeling and optimizes computation time.

To prevent overfitting, the model is trained on 80% of the pre-2021 claim database, while the remaining 20% (not used for training) serves to evaluate its performance. The goal is to select the model providing the best predictions.

The algorithm offers extensive parameter options. However, due to operational constraints related to computation time, only the following parameters were calibrated :

Hyperparameter	Value
nrounds	The total number of trees to be aggregated
max_depth	The maximum number of leaves per tree
eta	The learning rate, which adjusts the impact of each tree on the final prediction
subsample	The proportion of observations used for training each tree
min_child_weight	The minimum required weight to generate a child node
colsample_bytree	The proportion of explanatory variables selected to train each tree

TABLE 10 – Hyperparameters to optimize for the XGBoost model

Model calibration or hyperparameter tuning is a crucial step that involves selecting the algorithm's parameters to make it as well-adjusted as possible to the data, while avoiding overfitting. This phenomenon occurs when the model becomes too specific to historical data, making its predictions unreliable for new data.

The method consists of defining a parameter grid and testing all possible combinations by training a model on the training dataset, then measuring its error on the test dataset. The model that minimizes the Root Mean Square Error (RMSE) criterion is selected.

The table below shows, for each calibrated model, the predicted cost of climate-related claims in 2022 :

Model	« As if »	Estimated values in (K€)	Real values (in K€)	RMSE
Hail	**AC D+7**	175 127,00	173 902,49	16,74
	BT Indicator	154 341,90		16,95
Non Hail	**AC D+7**	43 130,14	43 105,50	9,17
	FFB Indicator	48 403,30		9,21

TABLE 11 – Prediction of the four models using the XGBoost method on the validation dataset

The 2022 predictions from the model validation set using « as if » AC appear closest to the actually observed costs. Thus, the retained « as if » cost for both models is the D+7 AC.

The final step of individual modelling involves estimating costs for closed-without-payment claims and not-yet-reported claims. Therefore, the estimated climate claims cost corresponds to the model's predicted cost, minus closed-without-payment claims costs, plus late-reported claims costs.

	Estimated 2022 cost (in K€)	Real 2022 cost (in K€)
Classical aggregate method	265 817,05	217 007,99
Enhanced aggregate method	244 670,13	
Individual method	228 331,90	

TABLE 12 – Comparative table of the results of the different methods

In conclusion, the individual approach proves more effective for predicting climate-related claims costs. By incorporating a set of detailed explanatory variables, this method yields more accurate estimates that better align with reality. Unlike aggregated approaches, it better captures claims-specific dynamics, thereby reducing biases and enhancing prediction reliability.

Table des matières

Introduction

Une compagnie d'assurance est une société qui propose des services d'assurance à ses clients, appelés assurés, en échange de la perception d'un montant appelé cotisation ou prime. Ces services sont conçus pour offrir des prestations aux assurés en cas de réalisation d'un événement incertain et aléatoire, communément appelé risque.

Dans ce secteur, on distingue l'assurance vie et l'assurance non-vie. Cette distinction est liée au principe d'indemnisation des sinistres :

- Le principe indemnitaire qui indique que l'assuré est remboursé en fonction du préjudice subi et sans possibilité d'enrichissement.
- Le principe forfaitaire qui indique que le remboursement ne dépend pas du préjudice subi et les montants d'indemnisation sont contractuels et préalablement déterminés.

L'assurance vie, également connue sous le nom d'assurance de personnes, garantit le versement d'un capital ou d'une rente à l'assuré en cas de survie, ou au bénéficiaire en cas de décès. Cette catégorie englobe divers types de contrats tels que l'Épargne Retraite, la Prévoyance (décès, incapacité et invalidité), la complémentaire santé, la dépendance, et d'autres encore.

Dans le cadre de l'assurance non-vie, ou assurance de dommages, souvent désignée sous l'appellation IARD pour « Incendie, Accidents et Risques Divers », sont regroupées les assurances de biens, visant à protéger les assurés contre les sinistres liés aux accidents, incendies, vols, etc. (par exemple l'assurance automobile, l'assurance Multirisques Habitation (MRH), Multirisques professionnel (MRP) ainsi que les assurances de responsabilité civile matérielle/corporelle, destinées à protéger les assurés contre les dommages corporels ou matériels causés à des tiers. Au sein de l'assurance IARD, une distinction est faite entre les particuliers et les professionnels.

Ce mémoire se place dans le cadre de l'assurance non-vie et plus précisément de la branche MRH. Cette branche permet à l'assuré de couvrir les dommages subis par les bâtiments et leurs aménagements (par exemple les caves et les garages), ainsi que le mobilier personnel. Elle s'applique uniquement aux particuliers et promet une couverture contre les risques qui touchent leurs biens. La branche MRH comprend de nombreuses garanties, comme le bris de glace, les dégâts des eaux, le vol, l'incendie, les catastrophes naturelles, etc. Certaines garanties sont obligatoires (par exemple la garantie dégâts des eaux et d'autres sont optionnelles (par exemple la garantie vol). Dans le cas où un sinistre MRH survient, ce dernier doit être déclaré, généralement, dans un délai de cinq jours ou de deux jours en cas de vol, et fournir divers justificatifs pour évaluer précisément le montant des pertes.

Contrairement à l'activité commerciale traditionnelle, l'activité d'assurance se caractérise par l'inversion de son cycle de production : le montant des prestations futures

est inconnu. Par conséquent, l'assureur détient des engagements envers ses assurés pour des montants indéterminés. Ainsi, l'assureur se trouve dans l'obligation de constituer une réserve, appelée provision, afin d'évaluer ses risques, de garantir sa solvabilité, de respecter les exigences réglementaires et d'assurer le règlement intégral de ses engagements à tout moment. C'est dans ce cadre que s'inscrivent un des objectifs de la direction actuariat IARD d'AXA France : assurer la suffisance des provisions techniques.

L'attention dans ce mémoire se porte sur les évènements climatiques de la branche MRH.

Ces dernières années et jusqu'à ce jour, la fréquence des événements climatiques a considérablement augmenté, tant au niveau mondial qu'en France. Cette tendance pèse lourdement sur les résultats financiers des assureurs, et il est crucial de comprendre les enjeux économiques liés à l'estimation rapide du coût de ces événements. Avoir une estimation précise de la charge finale des sinistres est essentiel pour les assureurs, d'autant plus que les événements climatiques représentent un coût croissant pour eux. En effet, le coût annuel moyen des sinistres climatiques en France est passé de 1,5 milliard d'euros entre 1982 et 1989 à 6 milliards d'euros entre 2020 et 2023, soit une augmentation de 300%.

Cependant, en raison de l'atypisme et de la variabilité de ces événements, prédire avec précision cette charge à l'ultime s'avère difficile pour les assureurs. Les méthodes traditionnelles de provisionnement telles que la méthode de Chain Ladder, habituellement utilisées pour le calcul des provisions, ne sont pas adaptées car les hypothèses fondamentales ne sont pas respectées. Ainsi, il est crucial de trouver d'autres moyens pour estimer au mieux cette charge.

Face à l'émergence de l'apprentissage statistique dans la gestion, la classification et la prédiction des données, les assureurs investissent de plus en plus dans l'utilisation de ces méthodes et cherchent à les mettre en œuvre pour estimer la charge des sinistres à l'ultime.

Ce mémoire se concentre sur l'estimation de la charge à l'ultime pour les évènements climatiques de la branche MRH. Pour cela, il remet en question les méthodes agrégées actuelles qui se basent uniquement sur l'observation du nombre de sinistres ouverts et de la typologie de l'événement. En enrichissant la base de données d'événements avec de nouvelles variables liées aux caractéristiques du sinistre ainsi qu'à des données exogènes de météorologie, il propose une approche d'estimation de la charge ultime des sinistres basée sur des méthodes d'apprentissage statistique.

Dans un premier temps, les caractéristiques et le marché de l'assurance MRH en France, les généralités sur le provisionnement en assurance non-vie et les risques climatiques en France sont présentés pour décrire le contexte général de ce mémoire. Dans un second temps, la méthodologie est introduite en expliquant les concepts mathématiques liés aux modèles agrégés d'une part et aux modèles d'apprentissage statistique d'autre

part. Enfin, dans une troisième partie et dernière partie, la création de la base de données et les résultats des modèles appliqués à cette base sont présentés et interprétés pour déterminer l'approche estimant au mieux la charge ultime.

I Cadre général

I.1 Généralités sur l'assurance Multirisques Habitation

Il est primordial de poser le contexte général de ce mémoire. Ce premier chapitre présente la branche de l'assurance qui nous intéresse et sur laquelle ce mémoire se base.

Dans un premier temps le marché français de l'assurance MRH est introduit pour évoquer dans un second temps les produits MRH au sein d'AXA France. Enfin, les problématiques de cette branche sont présentées dans un troisième temps.

I.1.A Le marché français de l'assurance Multirisques Habitation

Chaque année, la Fédération Française de l'Assurance « France Assureurs » publie des études statistiques sur les différentes branches de l'assurance, notamment sur le marché de l'assurance MRH.

Selon son rapport de 2024, portant sur l'année 2023, les cotisations des contrats MRH ont atteint 12,82 milliards d'euro, enregistrant une hausse de +5,3% d'évolution par rapport à 2022. Cette évolution confirme la dynamique de ce marché déjà observée avec une progression de 4,1% en 2022 par rapport à 2021. Cette performance place la branche des dommages aux biens des particuliers au deuxième rang du marché des assurances IARD, juste derrière l'automobile.

Cette augmentation est davantage marquée pour les contrats non-occupants (+8,9%) par rapport aux contrats occupants (+5,0%). Voici la répartition des cotisations des contrats occupants selon le type de résidence le type de bien :

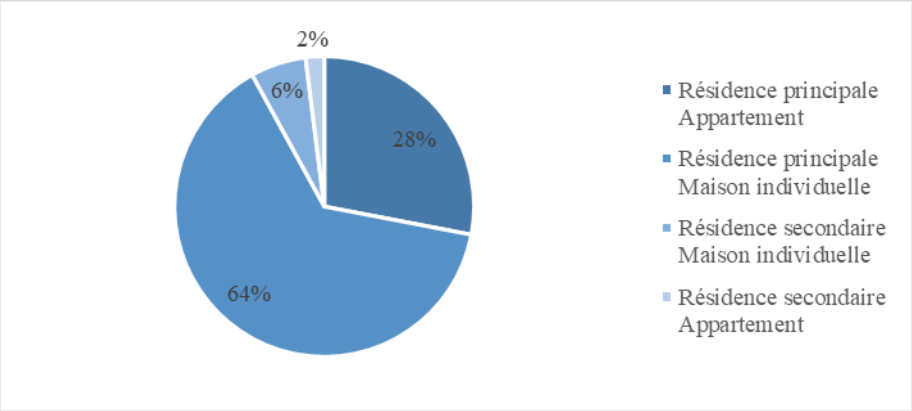


FIGURE 1 – La distribution des cotisations MRH des contrats occupants en 2023

Les résidences principales génèrent 92% des cotisations issues des contrats occupants, avec une majorité composée de maisons individuelles (64% de l'ensemble des contrats occupants). Les appartements déclarés comme résidences principales contribuent à un peu plus

du quart des cotisations totales (28%). Les résidences secondaires restent marginales, bien que la part des maisons individuelles dans cette catégorie ait légèrement augmenté (+1 point en un an).

Quant au nombre de contrats en MRH, l'année 2023 enregistre 45 899 milliers contrats et donc une hausse de 1,1% par rapport à 2022. Cette évolution est légèrement inférieure à celle de l'année 2022 (1,8%). Voici la répartition des contrats occupants selon le type de résidence :

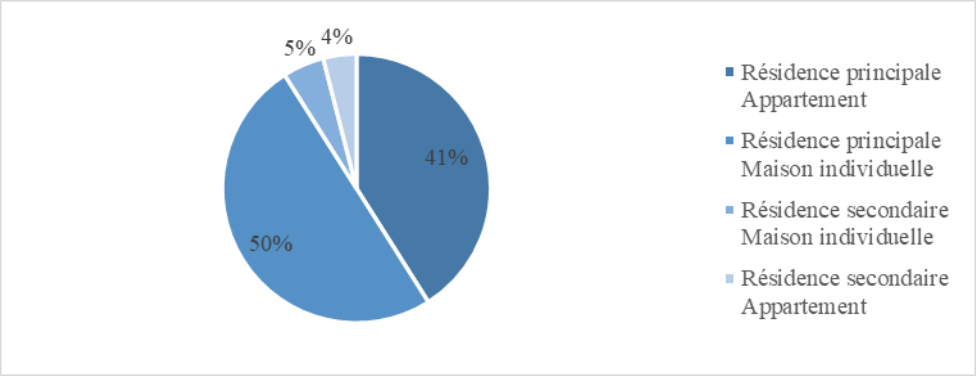


FIGURE 2 – La répartition des contrats occupants selon le type de résidence

Les résidences principales constituent 91% de l'ensemble des contrats occupants. La répartition entre maisons individuelles et appartements est relativement équilibrée, avec une légère prédominance des maisons individuelles, qui représentent 50% de l'ensemble des résidences principales et secondaires. En revanche, la part des résidences secondaires diminue, passant de 10% en 2022 à 9% en 2023 (-1 point).

La prime moyenne d'un contrat occupant s'élève à 303 € HT et celle d'un contrat non-occupant est estimée à 163 € hors taxe.

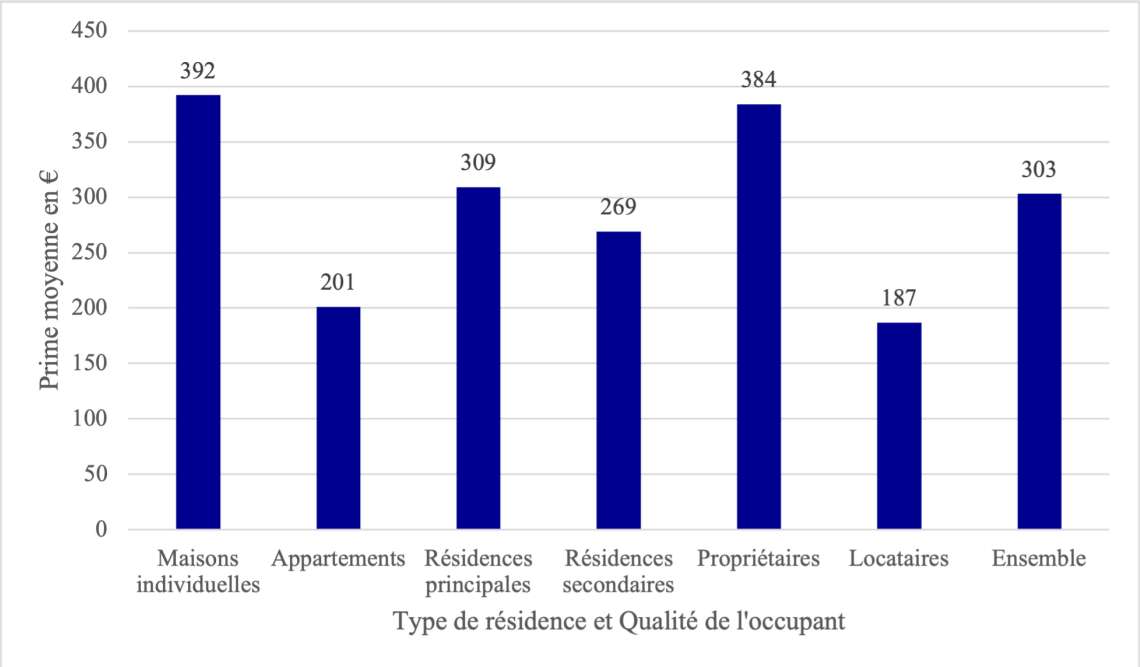


FIGURE 3 – La prime moyenne par type de résidence, type de bien et qualité de l'occupant

En fonction du type de résidence et du statut d'occupation, le montant de la prime moyenne diffère. Il est généralement plus élevé pour les résidences principales que pour les résidences secondaires. Les deux histogrammes ci-dessous illustre la distribution de la prime selon le type de résidence et la qualité de l'occupant pour les maisons et les appartements respectivement.

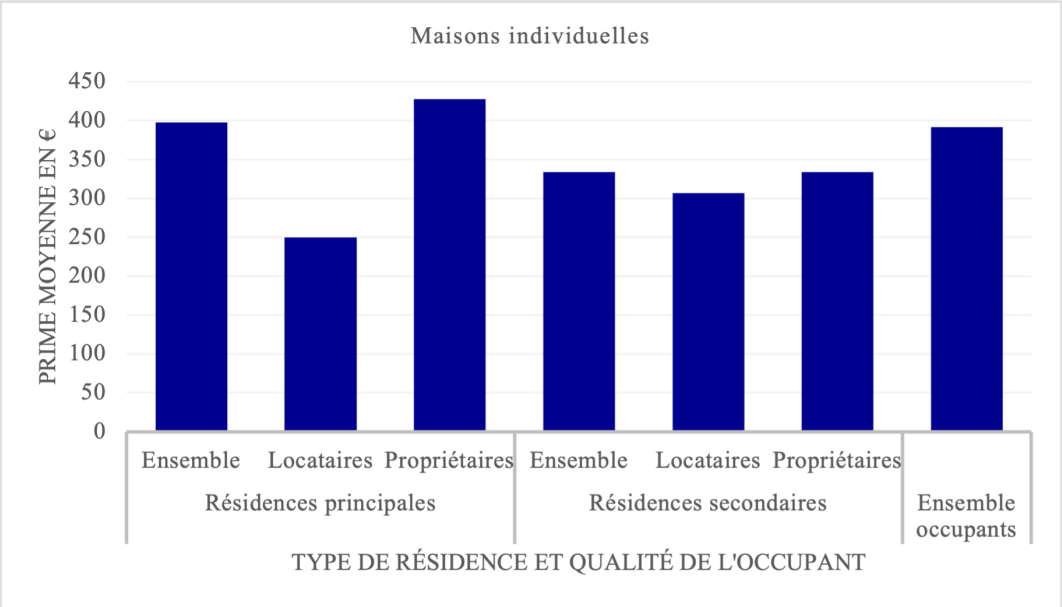


FIGURE 4 – La prime moyenne par type de résidence et qualité de l'occupant pour les maisons

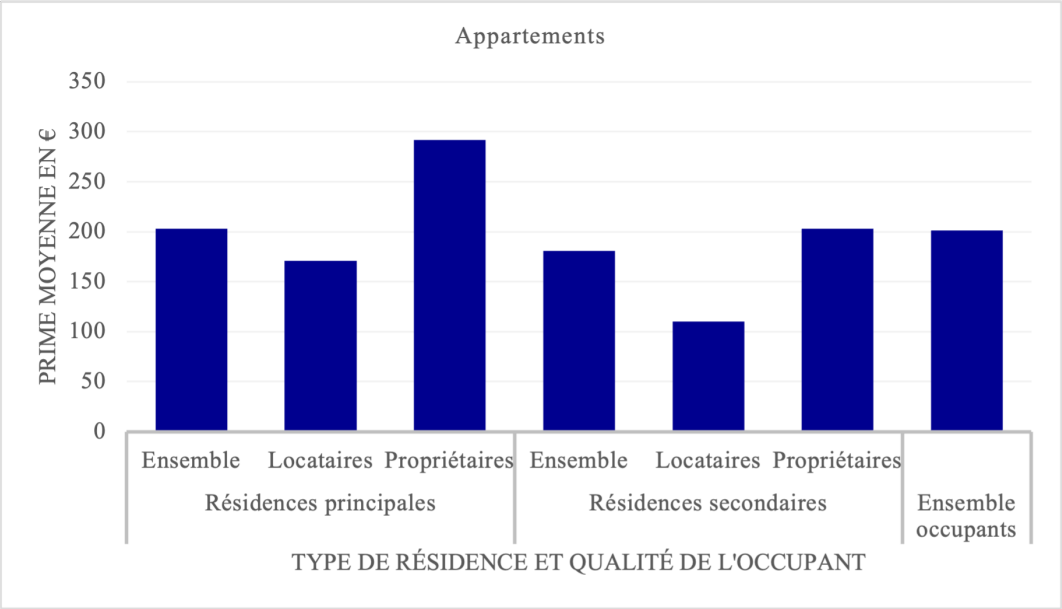


FIGURE 5 – La prime moyenne par type de résidence et qualité de l'occupant pour les appartements

En 2023, la prime moyenne pour les appartements atteint 203 € HT, tandis qu'elle s'élève à 398 € HT pour les maisons individuelles assurées comme résidences principales, engendrant ainsi des augmentations respectives de +3,9% et +5,3%. Par ailleurs, pour un même type de bien, comme un appartement, la prime moyenne est plus élevée pour les propriétaires (292 €) que pour les locataires (171 €).

Avec l'intensification des évènements climatiques ces dernières années, les catastrophes naturelles ont causé en 2023 une charge assurée de 6,5 milliards d'euros, dont 4,1 milliards pour l'assurance MRH. Depuis le début du suivi de ces événements en 1982, 2023 se classe comme la troisième année la plus coûteuse pour le secteur, après 1999 et 2022. La fin d'année a été particulièrement marquée par une forte sinistralité, avec les tempêtes Ciaran et Domingos en novembre, suivies des inondations dans le Nord-Pas-de-Calais. Ces événements ont représenté un coût total estimé à un peu plus de 1,4 milliard d'euros pour l'assurance habitation.

Le tableau ci-dessous décrit la fréquence, le cout moyen et le *Loss Ratio* (ratio des sinistres/primes) des sinistres des contrats MRH depuis 2019 jusqu'à 2023.

	Fréquence		Sinistre moyen		Loss Ratio	
Année	**Niveau**	**Évolution**	**Montants (en €)**	**Évolution**	**Niveaux**	**Évolution**
2019	9,33%	-4,70%	1 659	1,70%	56,80%	-3,1 pts
2020	8,84%	-5,30%	1 574	-5,10%	49,90%	-6,9 pts
2021	9,19%	4,00%	1 672	6,30%	54,60%	+4,7 pts
2022	9,59%	4,30%	2 137	27,80%	71,20%	+16,7 pts
2023	10,15%	5,80%	1 919	-10,20%	64,80%	-6,5 pts

TABLE 13 – La sinistralité des contrats MRH (y compris les catastrophes naturelles)

La fréquence des sinistres augmente de 5,8%, atteignant plus de 10 sinistres pour 100 contrats. En revanche, les coûts moyens, toutes garanties confondues, diminuent de manière significative de 10,2% par rapport à 2022. Cette baisse est principalement due à la diminution des sinistres liés à la garantie Tempête-grêle-neige (TGN).

Catégorie	Nombre	Répartition	Évolution répartition	Montant (en M€)	Répartition	Évolution répartition (montant)
Incendie	153 377	4%	-0,3 pt	2 019	26%	+2,0 pts
TGN	646 764	16%	+1,3 pt	1 893	24%	-9,9 pts
Vol	277 400	7%	-0,5 pt	566	7%	+0,7 pt
Dégâts des eaux	1 358 474	34%	+0,2 pt	1 846	24%	+3,6 pts
Responsabilité Civile	321 595	8%	-0,6 pt	466	6%	+0,4 pt
Bris de glaces	236 725	6%	-0,6 pt	145	2%	+0,1 pt
Dégâts électriques	265 231	7%	+0,4 pt	227	3%	+0,5 pt
Catastrophes naturelles	37 017	1%	+0,3 pt	337	4%	+3,4 pts
Protection juridique	162 955	4%	-0,1 pt	80	1%	+0,0 pt
Autres	592 623	15%	-0,1 pt	224	3%	-1,0 pt
Ensemble	**4 052 160**	**100%**		**7 802**	**100%**	

TABLE 14 – Le nombre et charge des sinistres par garantie MRH en 2023

Il est clair que la baisse du coût moyen a été le plus impactée par les épisodes de grêle intenses de 2022. La part de ces sinistres a chuté de 9,9 points de pourcentage de 2022 à 2023. Cependant, comparé à 2021, l'évolution annuelle moyenne du coût moyen reste dynamique, avec une hausse de 14%.

Ainsi, en 2023, la sinistralité des contrats MRH montre une détérioration en nombre, mais une amélioration en montant. Cela implique que le *Loss Ratio* s'améliore de 6,5 points par rapport à 2022.

En 2023, l'évolution de la charge des sinistres des contrats MRH a été marquée par des variations importantes selon les garanties :

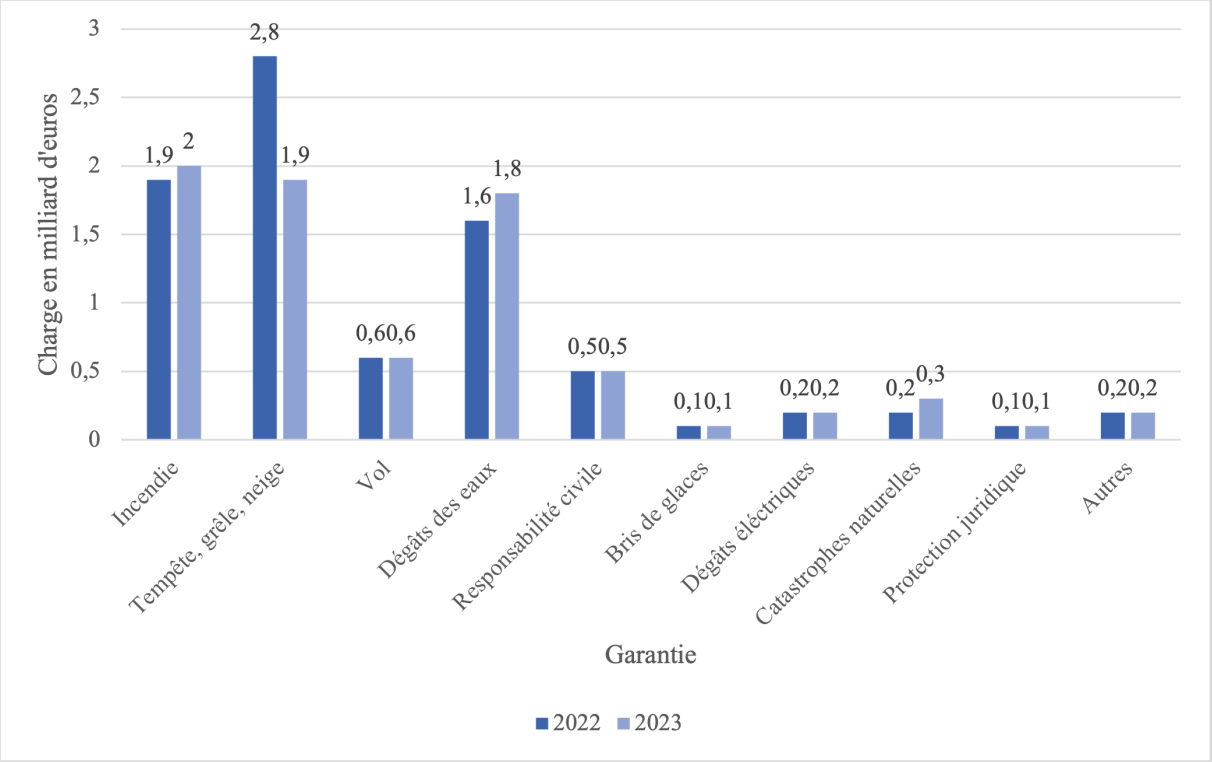


FIGURE 6 – Evolution de la charge des sinistres MRH par garantie

La charge liée à la garantie TGN a légèrement diminué, passant de 1,9 milliard d'euros en 2022 à 2,8 milliards d'euros en 2023, tandis que celle des autres garanties a connu une hausse.

Ainsi, bien que la charge liée à presque toutes les garanties des contrats MRH ait augmenté, le montant total des sinistres a diminué de 3,7% en 2023. Cette baisse s'explique par une réduction significative du coût des événements climatiques majeurs.

I.1.B La branche Multirisques Habitation chez AXA France

D'après le rapport de la Fédération Française de l'Assurance, voici la répartition du marché entre les 5 principaux groupes d'assurance en 2023 :

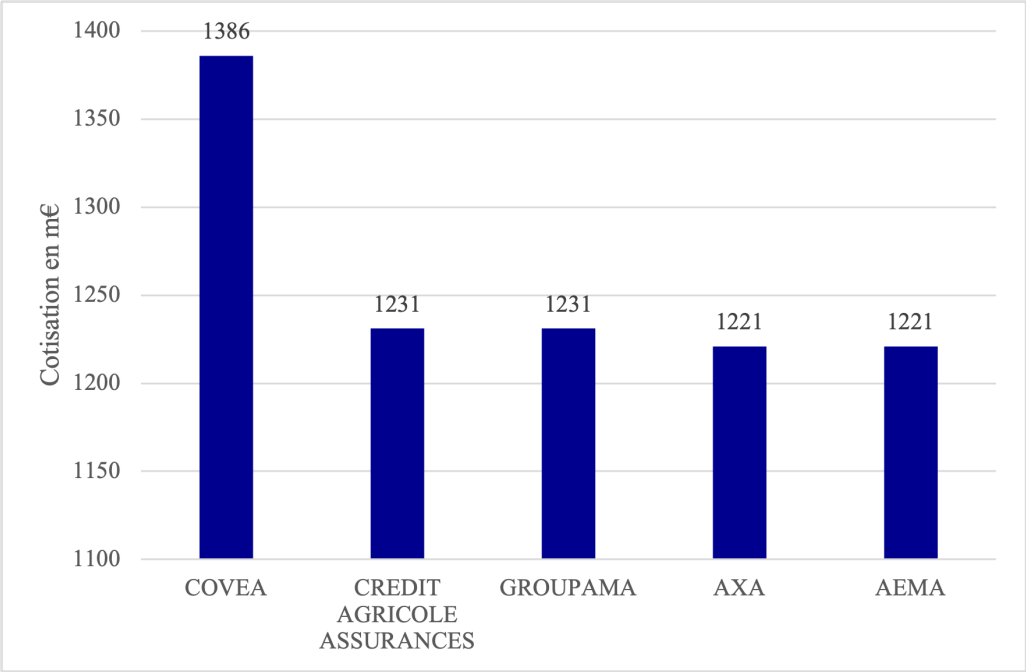


FIGURE 7 – Répartition des cotisations selon les principaux groupes d'assurances en 2023

Le groupe AXA se classe au quatrième rang des groupes d'assurances avec 1231 millions d'euros de cotisations.

Dans cette section, le produit d'assurance MRH chez AXA France est présenté. Le produit « Ma Maison » demeure le principal produit d'assurance MRH commercialisé chez AXA France. D'après les conditions générales d'AXA France (2024)[2], ce produit offre des garanties visant à protéger les biens des assurés. Les principales garanties proposées dans le produit « Ma Maison » sont les suivantes :

- La couverture incendie, qui inclut la protection contre les dommages causés par l'incendie, l'explosion, l'implosion, la fumée, les dommages matériels causés par les secours, ainsi que les dommages matériels dus à la chute de la foudre, le choc d'un véhicule terrestre à moteur et le choc d'un appareil aérien ou spatial ou des objets en provenance de ceux-ci.
- La couverture contre les dommages causés par l'eau et le gel, qui inclut la protection des bâtiments assurés contre les fuites, les ruptures/débordements de canalisations, ainsi que les infiltrations et le gel des conduites, des dispositifs de chauffage et des appareils à effet d'eau situés à l'intérieur des bâtiments d'habitation assurés, ainsi que les dommages matériels causés par les interventions de secours.
- La responsabilité civile, qui couvre les dommages causés à un tiers.
- La protection contre les catastrophes technologiques, qui couvre les dommages subis par les matériels à usage d'habitation et qui résulte d'une catastrophe technologique conformément aux articles L128-1 et suivants du Code des assurances.
- La garantie contre les attentats et les actes de terrorisme, qui couvre les dommages matériels et immatériels subis sur le territoire national à cause d'un éventuel attentat

ou acte de terrorisme, conformément aux articles 421-1 et 421-2 du Code pénal.

- La couverture contre les catastrophes naturelles, qui protège contre les dommages matériels non assurables à l'ensemble des biens garantis par le contrat et qui sont causés par l'intensité anormale d'un agent naturel. Cette garantie est déclenchée au cas où les mesures habituelles pour prévenir ces dommages n'ont pu empêcher leur survenance ou n'ont pu être prises.
- La couverture des événements climatiques, qui inclut la protection contre la tempête, l'ouragan, le cyclone, la grêle, la neige et l'inondation.

Le produit offre également diverses garanties facultatives, telles que la couverture contre le vol ou le bris de vitres.

Concernant le mode d'indemnisation de ce produit, il dépend des garanties souscrites et donc des conditions particulières d'AXA appliquées à chaque garantie. En revanche, il se peut que les conditions d'indemnisation ne soient pas indiquées pour une certaine garantie.

A l'exception des appareils électriques, les bâtiments et les aménagements immobiliers, qui ont subi des dommages, sont indemnisées selon la valeur de reconstruction à neuf, vétusté déduite au moment du sinistre, et dans la limite de la valeur vénale des bâtiments et des aménagements immobiliers à ce même moment.

Concernant les dommages aux biens mobiliers et aux appareils électriques et selon l'article L 121-5 du Code des assurances, une sanction est appliquée si les capitaux déclarés aux assureurs sont inférieurs à la valeur constatée à la survenance du sinistre : c'est la règle proportionnelle des capitaux. L'assuré a le droit de prouver l'existence des biens endommagé ainsi que leur valeur. Voici les modalités d'indemnisation :

- Rééquipement à neuf dans la limite des 10 ans : si l'ancienneté du bien n'excède pas les 10 ans, l'indemnisation est faite sur la base de la valeur de rééquipement à neuf le jour du sinistre. Sinon, indemnisation sur la base de la valeur de rééquipement à neuf avec déduction faite de la vétusté (prendre en compte de la dépréciation du bien en raison de son ancienneté).
- Rééquipement à neuf sans limite d'ancienneté : l'indemnisation est faite sur la base de la valeur de rééquipement à neuf le jour du sinistre quel que soit l'ancienneté du bien.
- Sans option : l'indemnisation est faite sur la base de la valeur de rééquipement à neuf le jour du sinistre avec déduction faite de la vétusté quel que soit l'ancienneté du bien.

I.1.C Les problématiques de la branche Multirisques Habitation

En raison du changement climatique dans le monde, les évènements climatiques sont en nette augmentation.

Dans le monde entier et notamment en France, la fréquence des évènements tels que les tempêtes, les inondations, les incendies et les sécheresses est en forte croissance. Cette augmentation engendre des défis considérables pour le secteur de l'assurance, plus particulièrement l'assurance MRH qui couvre ces évènements.

Cette augmentation des évènements climatiques engendre proportionnellement une augmentation des indemnisations des sinistres causés par de tels évènements et met une pression croissante sur le secteur assurantiel, qui doit ajuster ses primes, ses provisions et ses garanties pour faire face à ces événements coûteux.

Le graphique ci-dessous présente l'évolution des coûts des événements climatiques en France depuis 1984, exprimés en milliards d'euros.

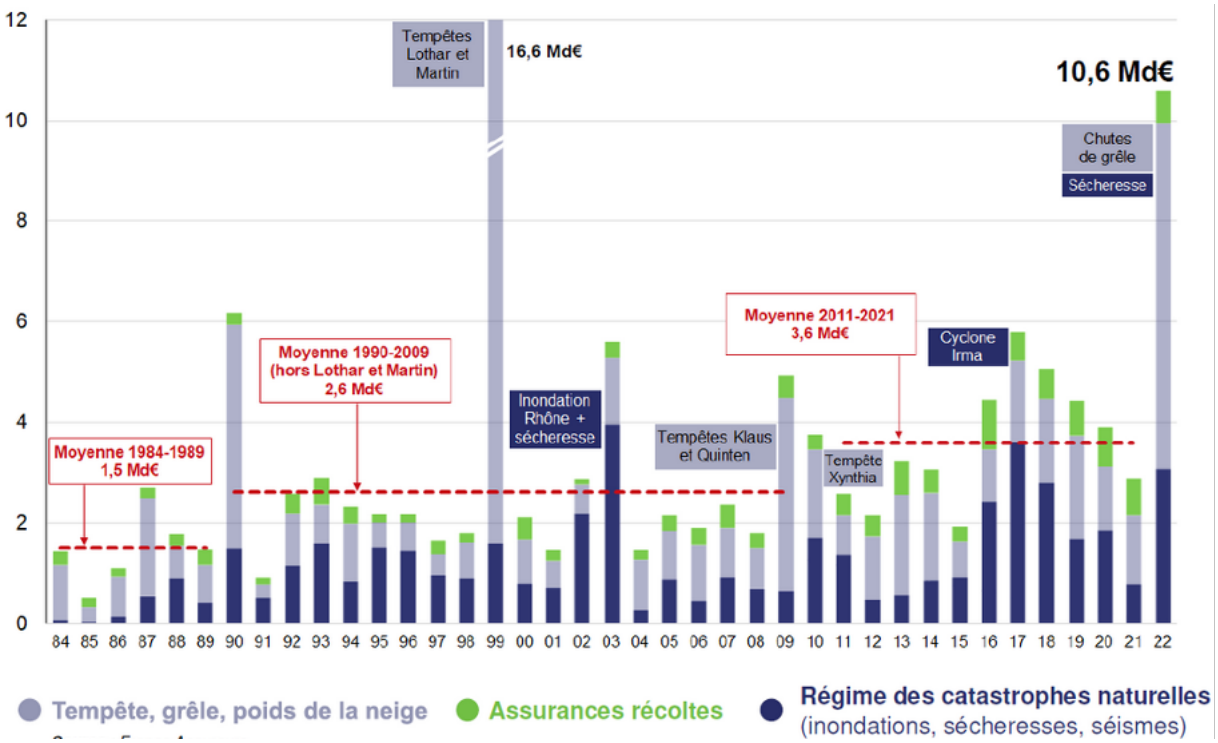


FIGURE 8 – La distribution de la charge des évènements climatiques depuis 1984

Une nette augmentation des coûts des événements climatiques est observée au fil du temps. La moyenne des coûts annuels a progressé de manière marquante : de 1,5 milliard d'euros en moyenne entre 1984 et 1989 à 2,7 milliards en moyenne entre 1990 et 2009, excluant les tempêtes Lothar et Martin, puis à 3,7 milliards en moyenne entre 2010 et 2019, et enfin à 6 milliards en moyenne entre 2020 et 2023. Ainsi, les évènements climatiques de ces 30 années ont marqué [1] 74 milliards d'euros d'indemnisations pour les assureurs.

1. Problèmes assurantiels des collectivités territoriales - Sénat

Cette évolution témoigne de l'intensification des phénomènes climatiques et de l'augmentation de leur coût pour les assurances. Plusieurs événements marquants ressortent : en 1999, les tempêtes Lothar et Martin ont généré un coût exceptionnel aux alentours des 17 milliards d'euros, un pic historique. En 2003, une combinaison d'inondations sur le Rhône et de sécheresse a également entraîné des coûts importants, tandis qu'en 2010, la tempête Xynthia a fait de cette année une année coûteuse.

Les années 2022 et 2023 sont particulièrement marquantes, avec un coût moyen atteignant 6,5 milliards d'euros en 2023, dépassant la moyenne récente. Les sinistres TGN représentent une part substantielle des dépenses, mais les catastrophes naturelles telles que les inondations, les sécheresses et les séismes, ainsi que les assurances récoltes, jouent également un rôle important, en particulier lors des années frappées par des événements comme des sécheresses ou des inondations.

Cette hausse constante des coûts moyens des évènements climatiques montre que ces derniers deviennent non seulement plus fréquents mais aussi plus destructeurs. Cette évolution est liée au changement climatique, qui intensifie les phénomènes extrêmes comme les tempêtes, les sécheresses ou les inondations.

Les années 2022 et 2023 sont l'une des années les plus coûteuses depuis l'année 1999. Cela confirme que les événements climatiques récents, bien que parfois moins spectaculaires que ceux de 1999, entraînent un coût financier important pour l'économie et les assurances. Les projections avenirs préviennent que les impacts du changement climatique vont continuer à s'aggraver d'ici les prochaines années, d'où la nécessité d'actions pour mieux gérer ces risques pour l'assurance MRH dans les années à venir.

Les assurés se prémunissent contre les risques climatiques en souscrivant des assurances, mais comment les assureurs, de leur côté, se protègent-ils face à ces risques croissants ? Quel que soit le type de sinistre, son ampleur ou le moment de sa survenance, un provisionnement adéquat est essentiel. Il permet aux assureurs de garantir leur capacité à honorer leurs engagements futurs envers les assurés en cas de sinistre.

Il existe plusieurs méthodes de provisionnement, classées en deux grandes catégories : les méthodes déterministes et les méthodes stochastiques. Ces méthodes nécessitent une vérification préalable de certaines hypothèses. Par exemple, la méthode du Chain Ladder exige la validation de l'indépendance et de la constance des cadences de règlements, avec un portefeuille considéré comme homogène et suffisamment vaste. Toutefois, pour les sinistres climatiques, le développement d'un sinistre peut varier en fonction du type d'événement. Ainsi, au lieu d'appliquer la méthode du Chain Ladder à l'ensemble des garanties climatiques, une alternative serait de regrouper les événements climatiques par type d'évènement (tempête, grêle ou orage), afin de créer une homogénéité spécifique à chaque catégorie.

Ces hypothèses n'étant pas toujours pleinement vérifiées pour toutes les branches, certaines entreprises d'assurance optent pour des approches plus simples telles que l'approche budgétaire. Bien que cette méthode soit plus facile à mettre en œuvre, elle reste éloignée de la réalité et constitue davantage une solution de simplification qu'une représentation fidèle des risques.

Dans ce contexte, la branche MRH se heurte à une problématique majeure : comment optimiser le provisionnement des sinistres climatiques face au dérèglement climatique et à l'augmentation de leur fréquence et de leur gravité ?

Ce mémoire s'attache à explorer cette question en mettant en lumière l'importance cruciale du provisionnement pour évaluer et anticiper efficacement les besoins financiers associés à ces sinistres.

I.2 Le provisionnement en assurance non-vie

I.2.A Les étapes de gestion d'un sinistre

Il est primordial de présenter le déroulement d'un sinistre. En effet, une fois le sinistre survenu (pendant la durée du contrat), l'assuré déclare son sinistre auprès de l'assureur. Les sinistres déclarés ne vont pas être réglés directement par l'assureur. La gestion du sinistre s'étale donc sur plusieurs délais appelés cadences de règlement des sinistres. Cependant les sinistres n'ont pas généralement la même cadence selon la garantie. C'est ainsi que nous distinguons les branches dites courtes et les branches dites longues. Par exemple, un sinistre dégât des eaux se clôture généralement plus rapidement qu'un sinistre responsabilité civile vie privée. De plus, les sinistres ne sont pas directement déclarés par l'assuré, c'est ce qu'on appelle les sinistres tardifs. Finalement, à la suite de la clôture de certains sinistres, ces derniers peuvent être rouverts en cas d'ajout d'informations supplémentaires au dossier du sinistre. Le schéma ci-dessous explique la vie du sinistre de la déclaration à la clôture.

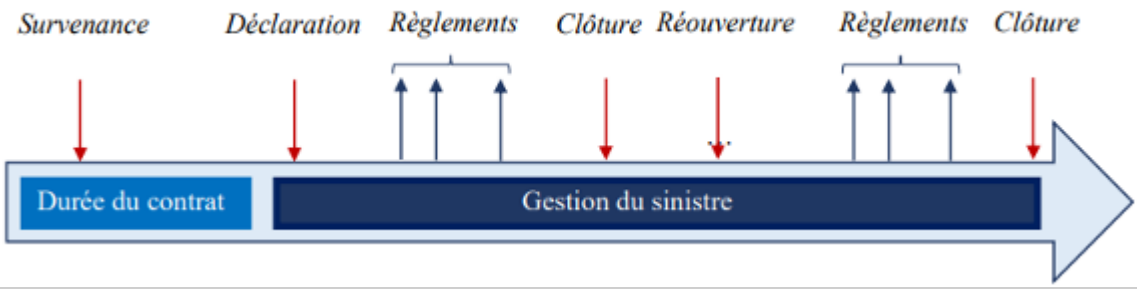


FIGURE 9 – Le déroulement d'un sinistre de sa survenance à sa clôture

Il est donc important de segmenter les sinistres selon différents critères. Généralement les sinistres sont segmentés selon la garantie et le produit (par exemple : les sinistres climatiques de la branche MRH). Cela permet de créer des groupes homogènes de sinistre

afin que les règlements soient stables et aboutir à un bon provisionnement en limitant au plus l'impact d'une volatilité dans le développement des sinistres.

Cette segmentation par garantie/branche/produit chez AXA, nous permet de récupérer facilement les sinistres de la branche MRH et de la garantie climatique et ainsi de se focaliser dessus.

I.2.B Les types de provision

Le Bilan d'un assureur se décompose en deux parties :

- L'actif : il correspond à tous les produits dans lesquels l'assureur investit les primes qui lui sont versées, valorisés en valeur de marché.
- Le passif : il correspond à l'ensemble des engagements de l'assureur. Il se décompose en deux parties :
  - Les fonds propres : c'est le capital apporté par les actionnaires. Ce sont des ressources disponibles permettant d'absorber les pertes liées à des événements rares et non anticipés.
  - Les dettes et provisions : ce sont les engagements de l'assureur liés à son activité d'assurance (paiement des sinistres, remboursements, etc.) et les dettes (emprunts, etc.).

Schématiquement, le Bilan économique se présente sous la forme suivante :

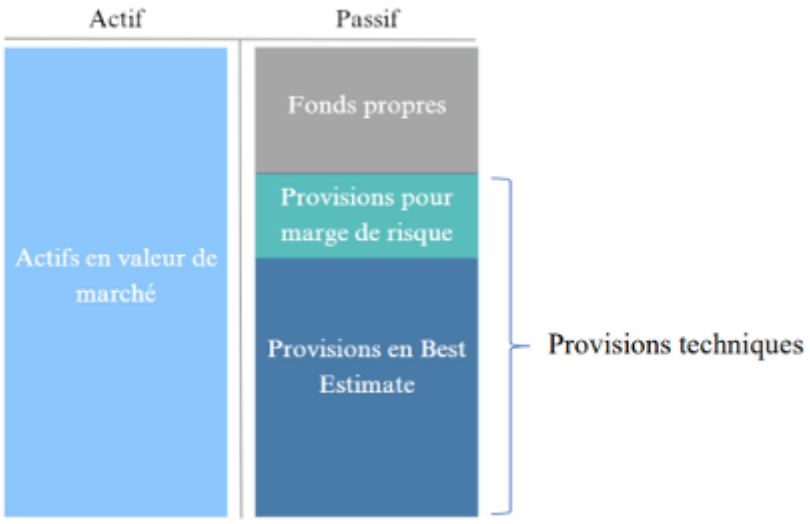


FIGURE 10 – Bilan économique sous Solvabilité II

Comme présenté en introduction, l'assureur détient une dette envers l'assuré en raison de l'inversion du cycle de production. De plus, on retrouve un délai entre l'ouverture d'un sinistre et la clôture de ce dernier. C'est pourquoi l'assureur doit estimer ces montants à la réalisation du bilan. Ces dettes figurent au Passif du bilan et constituent la

provision technique : montant de réserve couvrant les éventuels sinistres.

Il existe plusieurs types de provisions techniques qui englobent les provisions pour sinistres et les provisions pour prime. On ne traitera que la PSAP, provision principale, qui signifie Provision pour Sinistre A Payer. Cette provision présente le reste à payer pour les sinistres (même pour les sinistres qui ne sont pas encore déclarés).

Pour simplification « d/d » fait référence à « dossier/dossier », c'est à dire observé.

La PSAP est calculée en additionnant les réserves d/d (réserves émises par le gestionnaire de sinistre) et les *IBNR* qui désignent « *Incurred But Not Reported* ».

La provision *IBNR* est obtenue en sommant :

- Les *IBNYR* qui désigne « *Incurred But Not Yet Reported* » : ce sont les provisions pour les sinistres survenus mais non encore déclarés.
- Les *IBNER* qui désigne « *Incurred But Not Enough Reported* » : ce sont les provisions pour les sinistres survenus, déclarés mais sous évalués.

Les règlements constituent les paiements déjà effectués.

La charge d/d est la somme des charges des sinistres déclarés et par conséquent connus. Elle est obtenue en sommant les règlements et les réserves d/d.

Élément clé de l'étude de rentabilité, la charge finale prévisible (CFP) est la somme des PSAP et des règlements. Cette charge correspondant à la charge ultime.

I.2.C Le provisionnement sous un contexte réglementaire

L'assurance est étroitement encadrée par la réglementation, notamment le provisionnement. Entrée en application en 2016, la directive européenne Solvabilité II est une norme européenne qui se présente en trois piliers, selon l'Autorité de Contrôle Prudentiel et de Résolution « ACPR » :

- Pilier 1 : Exigences quantitatives
  Ces exigences introduisent un nouveau principe de valorisation du bilan : passage du bilan comptable à un bilan en vision économique.
  De nouvelles normes quantitatives sont introduites quant à l'évaluation des actifs et des passifs (dont les provisions techniques), des fonds propres et des exigences de capital.
  Cette nouvelle norme introduit deux seuils de capital réglementaire : le *SCR* (*Solvency Capital Requirement*) qui correspond au niveau de capital de solvabilité requis et le *MCR* (*Minimum Capital Requirement*) qui correspond au niveau de capital minimum réglementaire.

- Pilier 2 : Exigences qualitatives
  Ces exigences regroupent les règles de gouvernance et l'*ORSA* (*Own Risk and Solvency Assessment*)
- Pilier 3 : Transparence et communication
  Ce pilier consiste en la production de l'ensemble des rapports des piliers 1 et 2 à destination du régulateur et du public.

L'intérêt ici se porte au pilier 1 et plus précisément aux principes du calcul des provisions techniques en *Best Estimate* (*BE*).

Le *BE* représente la somme de la meilleure estimation des sinistres et de la marge de risque. Il est défini par la directive Solvabilité II (alinéa 2 de l'article 77 de la Directive 2009/138/CE [2]) comme suit : « *La meilleure estimation correspond à la moyenne pondérée par leur probabilité des flux de trésorerie futurs, compte tenu de la valeur temporelle de l'argent (valeur actuelle attendue des flux de trésorerie futurs), estimée sur la base de la courbe des taux sans risque pertinente.* »

La marge pour risque permet d'introduire une marge de prudence dans le calcul des provisions techniques. D'après l'alinéa 3 de l'article 77, elle est déterminée « *de manière à garantir une valeur des provisions techniques équivalente au montant que les entreprises d'assurance et de réassurance demanderaient pour reprendre et honorer les engagements d'assurance et de réassurance.* »

Comme expliqué précédemment, l'assureur doit construire des groupes de risque homogènes. C'est ce qu'impose également la directive Solvabilité II (article 80 de la directive) : « *Lorsqu'elles calculent leurs provisions techniques, les entreprises d'assurance et de réassurance segmentent leurs engagements d'assurance et de réassurance en groupes de risques homogènes et, au minimum, par ligne d'activité.* »

L'assureur peut être confronter à deux cas : une sous-estimation ou une sur estimation des provisions. Le premier cas va placer l'assureur dans l'impossibilité de respecter ses engagements futurs car il n'a pas les moyens suffisants. Dans le deuxième cas l'assureur constitue un manque de rentabilité et peut inévitablement amener à une immobilisation de capitaux.

C'est dans ce cadre que s'inscrit un des objectifs principaux de la norme Solvabilité II qui consiste à contrôler le calcul des provisions des compagnies d'assurance.

Ce mémoire vise à évaluer une approche alternative pour estimer au mieux les sinistres à payer des sinistres climatiques au sein de la branche MRH. L'objectif est d'obtenir

2. DIRECTIVE 2009/138/CE DU PARLEMENT EUROPÉEN ET DU CONSEIL du 25 novembre 2009 sur l'accès aux activités de l'assurance et de la réassurance et leur exercice (solvabilité II)

un modèle plus performant et aboutir à des résultats plus précis. La provision calculée suivra les principes fondamentaux de Solvabilité II à l'exception de l'actualisation des flux.

I.2.D La méthode classique de provisionnement : Chain Ladder

Il existe plusieurs méthodes de provisionnement. Ces méthodes se divisent en deux grandes catégories : les méthodes dites déterministes et celles dites stochastiques. La distinction des méthodes stochastiques est qu'elles fournissent des estimations de variance et d'intervalle de confiance des PSAP estimées.

Afin d'optimiser la capacité de l'assureur à respecter ses engagements vis-à-vis de l'assuré, le choix de la méthode doit être le plus adapté pour estimer au mieux les provisions.

La figure ci-dessous présente les principales méthodes déterministes :

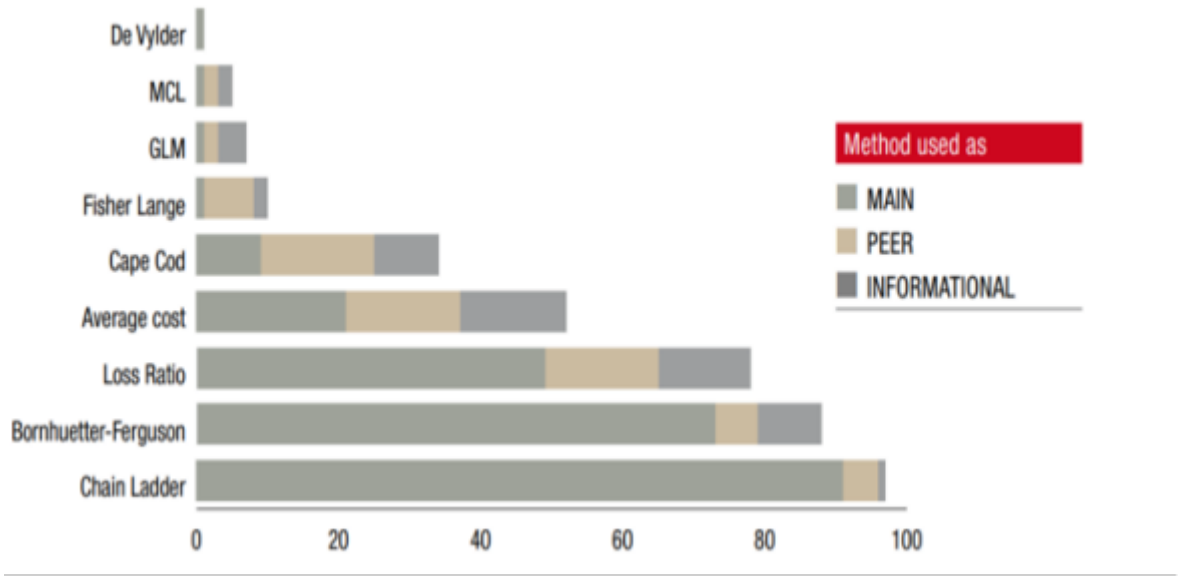


FIGURE 11 – Principales méthodes déterministes de provisionnement en non-vie

Ce mémoire se concentre principalement sur la méthode Chain Ladder, qui appartient aux méthodes déterministes. Comme illustré dans la figure ci-dessus, cette méthode est la plus couramment utilisée.

Avant de détailler la méthode *Chain Ladder*, il est essentiel d'introduire les triangles de liquidation. En effet, la méthode *Chain Ladder* repose sur une approche agrégée, nécessitant au préalable la construction de triangles de liquidation.

Les triangles de liquidation représentent le déroulement des sinistres et leur dynamique. Une distinction est faite entre le triangle incrémental et celui cumulé. Le triangle de liquidation de charges incrémental est présenté comme suit :

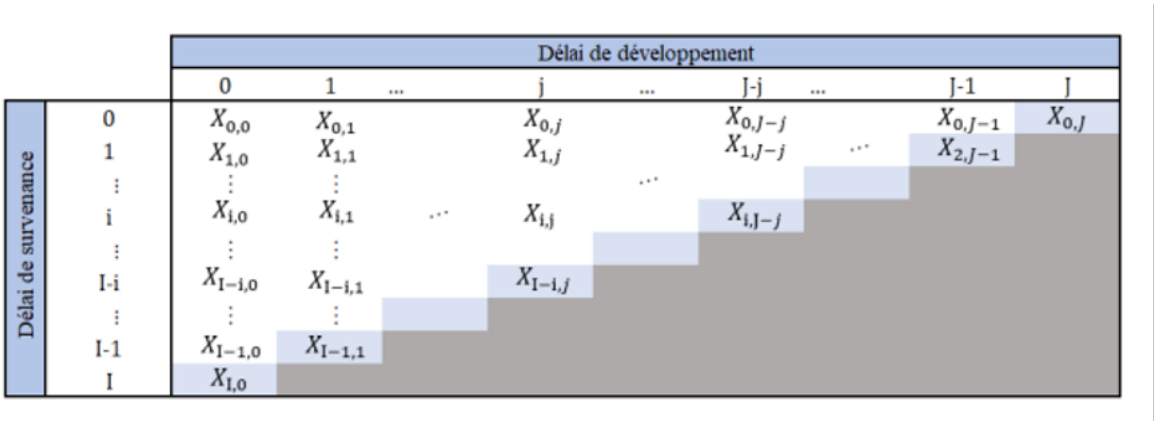


FIGURE 12 – Triangle des incréments

Les délais de survenance sont notés par i ; $(0 \leq i \leq I)$ et les délais de règlement/développement sont notés par j ; $(0 \leq j \leq J)$. Le montant $X_{i,j}$ représente l'incrément c'est-à-dire le paiement des sinistres survenus au $i^{\text{ème}}$ délai et réglé au $j^{\text{ème}}$ délai.

Le triangle de liquidation de charges cumulées est présenté comme suit :

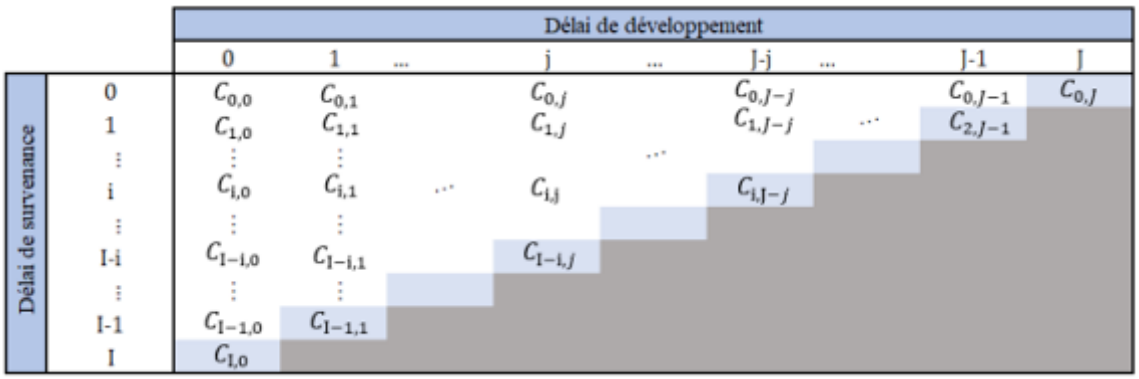


FIGURE 13 – Triangle des cumuls

Le montant $C_{i,j}$ représente le montant cumulé de tous les paiements effectués jusqu'au $j^{\text{ème}}$ délai pour un sinistre survenu au $i^{\text{ème}}$ délai et il est obtenu grâce à la formule suivante :

$$C_{i,j} = \sum_{k=0}^{j} X_{i,k}$$

Réciproquement, le passage de la valeur cumulée à la valeur incrémentale est possible :

$$X_{i,j} = C_{i,j+1} - C_{i,j}$$

Le triangle de liquidation est divisé en trois parties :

- La partie inférieure grisée du triangle (c'est-à-dire les $X_{i,j}$ tel que $i + j > n$) qui représente les montants à estimer.
- La partie supérieure blanche du triangle (c'est-à-dire les $X_{i,j}$ tel que $i + j \leq n$) qui représente les montants observés.
- La diagonale bleue du triangle (c'est-à-dire les $X_{i,j}$ tel que $i + j = n$) qui représente le dernier règlement effectué jusqu'à ce jour c'est-à-dire le montant des sinistres d'un exercice comptable.

Le but est d'estimer la partie inférieure du triangle et donc de projeter les sinistres survenus en délai i sur l'ensemble des délais de règlement j, c'est pour cela que les que les facteurs de développement sont introduits.

Les facteurs de développement individuels sont calculés comme suit :

$$f_{i,j} = \frac{C_{i,j+1}}{C_{i,j}}; j \in 0, \ \ldots, \ J-1$$

Le triangle des facteurs de développement appelé aussi d-triangle peut être représenté selon la figure ci-dessous :

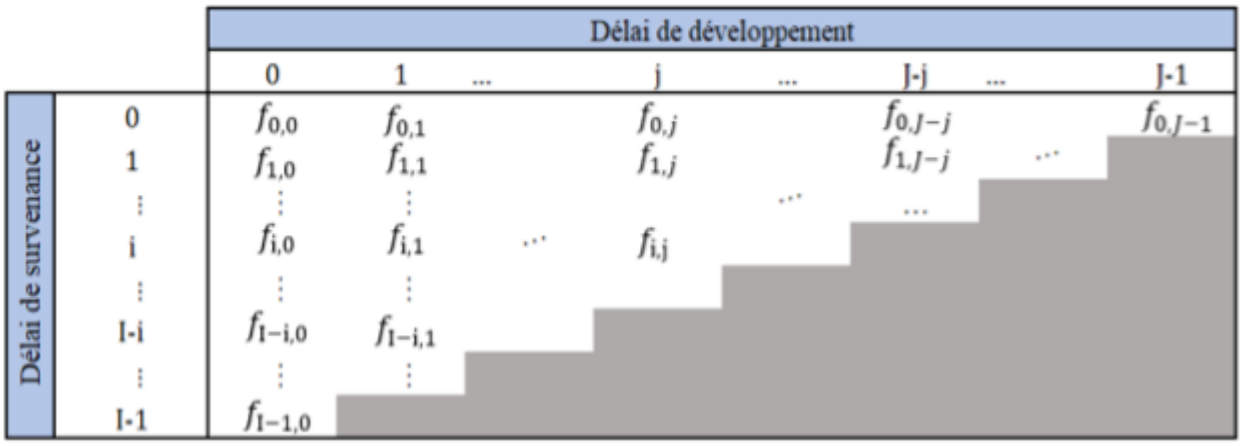


FIGURE 14 – Le d-triangle

Remarques :

- Les triangles de liquidation peuvent être également utilisés pour illustrer le nombre des sinistres (clos, clos sans suite, clos avec suite), les règlements des réserves, les recours, etc.
- Les délais de règlement et de survenance peuvent être exprimées en année ou en mois.

Une fois les triangles de liquidation expliqués, il est possible de présenter la méthode de Chain Ladder. Pour des raisons de simplification, le nombre de délais de survenance et de délais de développement sont considérés égaux ($I = J$).

En raison de sa simplicité, la méthode de Chain Ladder est celle la plus utilisée en provisionnement non-vie. Elle est la plus ancienne et a pour but d'estimer le coût final mais peut être aussi appliquée sur les règlements cumulés ou les nombres.

Les conditions d'application de cette méthode reposent sur deux hypothèses principales. Une première hypothèse est l'homogénéité de l'échantillon. Quant à la deuxième hypothèse, elle repose sur le fait que, pour un délai de survenance fixé, l'évolution des montants doit être identique aux cours des délais de développement. Cette évolution est traduite par le coefficient de passage, facteur de développement, expliqué plus haut. Plus précisément les facteurs de développement individuel $f_{i,j}$ doivent être indépendants du délai de survenance. Cela peut se traduire mathématiquement par :

$$f_{0,j} = f_{1,j} \ = \ldots = f_{I,j}; \forall \ j \ \in 0, \ldots, \ I$$

Autrement,

$$\frac{C_{0,j+1}}{C_{0,j}} = \frac{C_{1,j+1}}{C_{1,j}} = \ldots = \ \frac{C_{I,j+1}}{C_{I,j}}; \forall \ j \ \in 0, \ldots, \ I$$

En moyennant les facteurs de développement individuels, le facteur de développement cumulé ou le coefficient de passage d'un délai de développement à un autre est obtenue grâce à la formule suivante :

$$f_j = \frac{\sum_{i=0}^{I-j-1} C_{i,j+1}}{\sum_{i=0}^{I-j-1} C_{i,j}}; \forall \ j \ \in 0, \ldots, \ I-1$$

Pratiquement, pour chaque délai de survenance i, le passage d'un délai de développement j au $(j+1)^{\text{ème}}$ est réalisé en multipliant la charge du $j^{\text{ème}}$ délai de développement par le coefficient de passage f_j. La figure ci-dessous illustre ce passage :

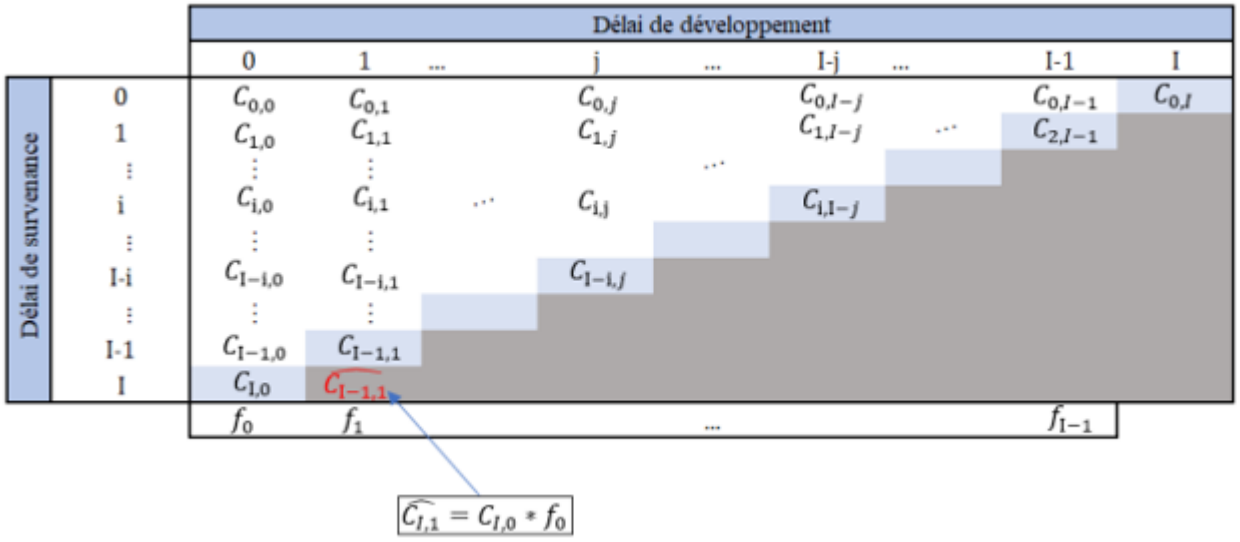


FIGURE 15 – Exemple illustratif du passage d'un délai de développement à un autre

L'importance d'estimer le triangle inférieur est de pouvoir estimer la charge ultime pour chaque délai de survenance. Cette charge réside dans la dernière colonne du triangle de

liquidation. Ainsi, $\widehat{S_i}$ représente le montant de la charge ultime/finale prévisible pour le délai de survenance i, c'est le montant estimé à payer d'ici le délai ultime I pour un sinistre survenu en i. Ainsi :

$$\widehat{S_i} = \widehat{C_{i,I}} \ \ \forall i = 0, \dots \ , \ I$$

La charge ultime peut également être directement estimer sans devoir passer par chaque délai de développement. Pour se faire, les facteurs de développement cumulés sont calculés :

$$\widehat{S_i} = \widehat{C_{i,I}} * \prod_{k=0}^{I-1} f_k \ \forall i = 0, \dots, I$$

Sachant que la diagonale du triangle correspond au dernier montant cumulé payé pour la survenance i, la provision pour chaque survenance $i = 0, \dots, I$, notée $\widehat{R_i}$, n'est autre que la différence entre la charge ultime (dernière colonne) estimée et le dernier paiement réglé (diagonale). Elle est calculée par la formule suivante :

$$\widehat{R_i} = \ \widehat{S_i} - C_{i,I-i} \, \forall i = 0, \ \dots \ , \ I$$

Finalement la réserve totale qui est la PSAP correspond à la somme des $\widehat{R_i}$:

$$\hat{R} = \sum_{i=0}^{I} \widehat{R_i}$$

Le tableau ci-dessous récapitule tous les éléments clés évoqués :

Délai de survenance	Règlements	Charge ultime	PSAP
0	$C_{0,I}$	$S_0 = C_{0,I}$	$R_0 = 0$
1	$C_{1,I-1}$	$\widehat{S_1} = \widehat{C_{1,I}}$	$\widehat{R_1} = \widehat{S_1} - C_{1,I-1}$
⋮	⋮	⋮	⋮
i	$C_{i,I-j}$	$\widehat{S_i} = \widehat{C_{i,I}}$	$\widehat{R_i} = \widehat{S_i} - C_{i,I-j}$
⋮	⋮	⋮	⋮
I-i	$C_{I-1,1}$	$\widehat{S_{I-i}} = \widehat{C_{I-1,1}}$	$\widehat{R_{I-i}} = \widehat{S_{I-i}} - C_{I-i,j}$
⋮	⋮	⋮	⋮
I	$C_{I,0}$	$\widehat{S_I} = \widehat{C_{I,I}}$	$\widehat{R_I} = \widehat{S_I} - C_{I,0}$

FIGURE 16 – Synthèse de Chain Ladder

Comme expliqué précédemment, l'application de Chain Ladder repose sur la validation d'hypothèses. Cette validation repose sur le d-triangle. La méthode est validée si $\forall j = 0, \dots, I-1$ les couples $(C_{i,j}, C_{i,j+1})$ sont alignés sur une droite passante par l'origine (grahiquement cela correspond au CC plot). En d'autres termes, $\forall j = 0, \ \dots \ , \ I-1$:

$$C_{i,j+1} = C_{i,j} * f_j$$

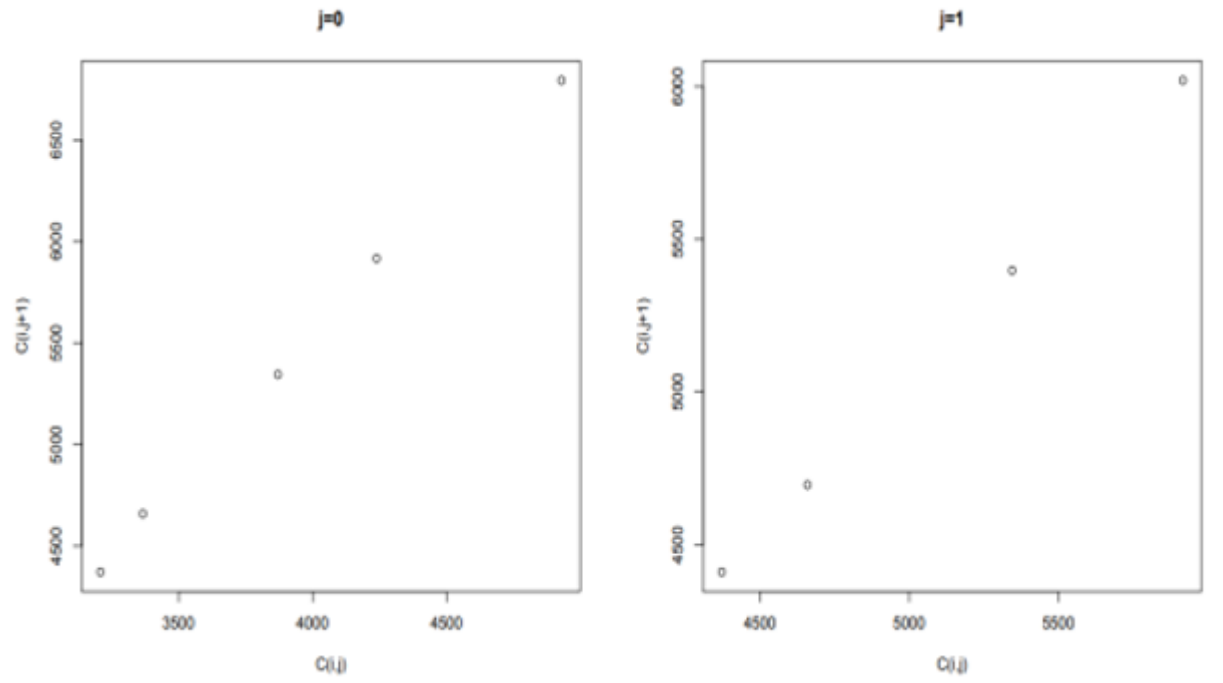


FIGURE 17 – CC plot pour le développement en 0 et en 1

La méthode Chain Ladder est facile à mettre en œuvre et simple à comprendre. Elle nécessite moins d'hypothèses que les méthodes stochastiques.

Cependant, le grand inconvénient est la potentielle présence de valeurs extrêmes dans les triangles. Cela peut aboutir à des résultats aberrants qui doivent donc être retirés pour le calcul des PSAP. C'est principalement le cas de la garantie climatique en MRH.

I.2.E Le provisionnement par approche budgétaire

Cette sous-section aborde la méthode de provisionnement des sinistres climatiques au sein de l'équipe inventaires d'AXA France IARD. En effet, pour les sinistres climatiques de la branche MRH, la méthode de provisionnement ne repose pas sur la méthode Chain Ladder mais sur une approche budgétaire. Cette approche consiste à établir un budget initial pour les sinistres climatiques. L'équipe inventaire examinent la moyenne des coûts financiers des sinistres climatiques au cours des dernières années pour établir le budget de l'année en cours. Ce budget est déterminé sur la base d'hypothèses concernant le coût moyen, la fréquence et l'exposition) et il est ensuite révisé au cours de l'exercice comptable, en fonction de l'évolution des événements climatiques. Il peut être maintenu, réduit en cas d'absence de sinistres climatiques pour financer d'autres dommages, ou augmenté si nécessaire. Ainsi, cette approche consiste à allouer une enveloppe de charge pour des événements de type climatique et à ajuster cette enveloppe en fonction des événements réels au cours de l'année.

À quel moment ce budget est-il généralement déterminé et révisé ? L'équipe Inventaire a pour objectif de produire des exercices prévisionnels tout au long de l'année

(budgets, prévisions, plans). Lors de ces exercices, l'équipe projette les résultats sur un horizon temporel, ce qui permet d'envisager une première estimation de la sinistralité de 2025 dès 2024. Ces simulations reposent sur des hypothèses concernant l'évolution du chiffre d'affaires, du *Loss Ratio*, ainsi que sur les effets prix et volume. Ce dernier se réfère à une augmentation des volumes de contrats vendus et des charges associées.

Une première prévision budgétaire est alors établie. Lors de la détermination du budget, l'équipe intègre également une projection de l'atterrissage de fin d'année. Le budget de l'année N est généralement finalisé en décembre de l'année N-1, en se basant sur le chiffre d'affaires de la dernière prévision de septembre. Au cours de l'année, cet atterrissage est ensuite mis à jour avec des données plus récentes : c'est ce qu'on appelle le budget révisé. Le schéma ci-dessous illustre les étapes de la détermination et de la revue du budget tout au long de l'année :

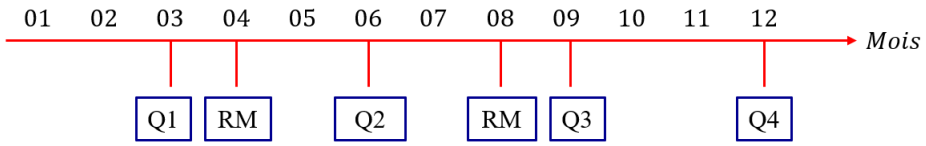


FIGURE 18 – Déroulement d'une année comptable de l'équipe inventaire

C'est selon cette approche que les sinistres climatiques en MRH sont provisionnés.

Bien que cette méthode soit facile à implémenter, l'approche budgétaire ne manque pas d'inconvénient. En effet, même si le budget est révisé plusieurs fois dans l'année, l'approche manque énormément de précision. Le but de ce mémoire est de proposer des méthodes améliorées par rapport à cette approche. Par exemple, selon l'historique des années passées, si l'équipe met X€ de provisions en décembre de l'année $N-1$ et entre janvier et février (visée 1) et que le modèle amélioré détecte Y€ de CFP, ce dernier va être capable de détecter l'écart entre le montant de CFP estimé et le budget des comptes. Le modèle amélioré proposé dans ce présent mémoire va donc aider à regarder la consommation sur l'année en cours. Dès l'observation d'un évènement, la consommation du budget alloué aux sinistres climatiques est évaluée.

I.3 Le risque climatique

Cette partie aborde le risque climatique. Elle commence par définir le risque climatique et les événements de grande ampleur puis présente leur impact sur les assureurs en France et examine les risques individuels de manière plus spécifique.

I.3.A Définition du risque climatique et son impact sur les assureurs en France

Les phénomènes climatiques extrêmes, tels que les vagues de chaleur, les tempêtes violentes, les crues et la sécheresse, sont de plus en plus fréquents et intenses en raison du changement climatique. Ce phénomène est principalement causé par l'accumulation des gaz à effet de serre dans l'atmosphère, qui modifie les régimes météorologiques, causant une augmentation des températures, des modifications des précipitations et l'intensification des événements climatiques extrêmes.

Le risque climatique désigne l'ensemble des dangers liés au dérèglement climatique, qui peuvent affecter les biens, les individus et les économies. Les risques climatiques sont en forte augmentation, notamment en raison du changement climatique, et représentent aujourd'hui des risques émergents pour les compagnies d'assurance. Ces risques se manifestent principalement par des événements extrêmes causant des dommages matériels, des perturbations économiques ou une perte de valeur des biens assurés.

Les risques climatiques peuvent être divisés en trois catégories :

- Les risques physiques : Ils sont directement liés aux phénomènes météorologiques, et ce sont ceux dont il est généralement question lorsqu'on parle de risques climatiques.
- Les risques de transition : Ils concernent les entreprises qui ne parviennent pas à s'adapter à un environnement en mutation pendant la transition climatique.
- Les risques de responsabilité : Ils surviennent lorsqu'une entreprise est poursuivie en justice pour non-respect d'engagements environnementaux.

Ce mémoire se concentre uniquement sur les risques physiques, appelés simplement « risques climatiques ».

Dans le secteur de l'assurance, ces risques se matérialisent par des événements extrêmes, tels que la grêle, les tempêtes, les inondations ou les sécheresses et peuvent entraîner des sinistres nécessitant des indemnisations importantes pour les assureurs surtout pour la branche MRH. L'identification, et la gestion de ces risques sont essentielles pour garantir une couverture efficace des assurés tout en préservant la stabilité financière des compagnies d'assurance.

En raison de l'incertitude croissante associée à ces phénomènes, il est crucial pour les assureurs d'adopter des approches avancées de modélisation. Ces méthodes doivent permettre de prendre en compte à la fois les impacts immédiats des événements climatiques et leurs tendances à long terme et donc d'ajuster les modèles de provisionnement en conséquence.

La gestion du risque climatique en assurance MRH représente un défi majeur, nécessitant une compréhension approfondie des phénomènes météorologiques et des stratégies

d'adaptation face à des risques de plus en plus incertains.

I.3.B Définition d'un évènement de grande ampleur

Ce mémoire s'intéresse aux évènements de grande ampleur, les « EGA ». Ils se distinguent par l'occurrence d'un ou plusieurs sinistres d'intensité exceptionnelle. Ce type d'évènement peut être défini selon différentes perspectives.

- Perspective statistique :
  C'est un évènement climatique extrême dont les valeurs mesurées (comme la vitesse du vent, les précipitations, la température ...) dépassent les seuils habituellement observés. Le caractère extrême est donc défini sur la base de valeurs numériques.
- Perspective physique :
  Contrairement à l'approche statistique, qui se base uniquement sur des valeurs numériques inhabituelles, les physiciens considèrent qu'un événement est extrême selon sa nature et son comportement dans un contexte géographique donné, comme les canicules.
- Perspective des sciences sociales :
  Un évènement devient de grande ampleur selon les conséquences (dommages humains, matériels ou économiques) sur la société qu'il engendre.

I.3.C Panorama des risques climatiques

Dans la perspective des projections du Groupe d'experts intergouvernemental sur l'évolution du climat, appelé GIEC, et d'après l'étude « Impact du changement climatique sur l'assurance à l'horizon 2050 » publiée en 2021 par France Assureurs[11], les montants indemnisés par les assureurs pour les sinistres climatiques entre 1989 et 2019 ont atteint 74 milliards d'euros.

Avec l'intensification et la fréquence croissante des événements climatiques, les pertes cumulées dues aux risques climatiques pourront évoluer voire doubler en atteignant 143 milliards d'euros sur la période 2020-2050.

D'après cette étude, les montants indemnisés par les assureurs entre les années 1989 et 2019 des sinistres climatiques (dont les tempêtes, la sécheresse et les inondations) se répartit principalement selon la figure ci-dessous :

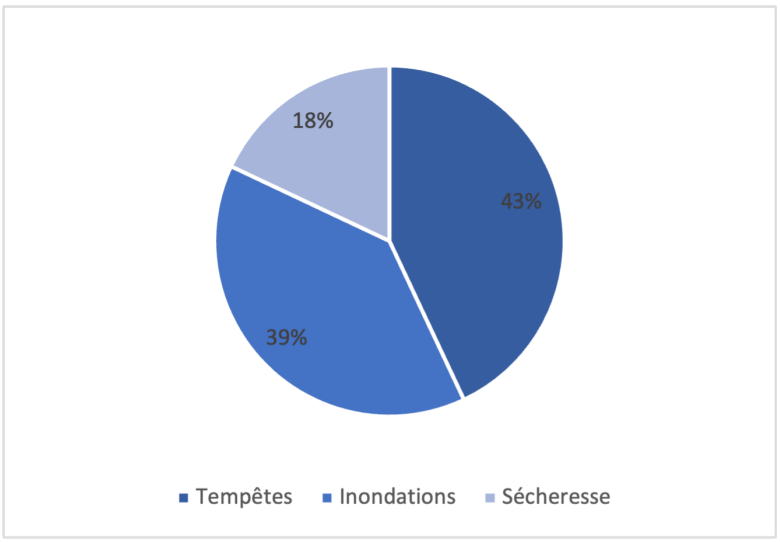


FIGURE 19 – Répartition des montants indemnisés des sinistres climatiques

Cette répartition montre que 43% des indemnisations représentent les indemnisations des sinistres causés par des tempêtes. En deuxième place, 39% des montants indemnisés représentent les sinistres inondations. Enfin, l'évènement avec le moins d'impact est la sécheresse avec 18%.

Ces pourcentages peuvent être expliqués par le nombre de sinistres associé à chaque événement. Plus précisément, le tableau ci-dessous représente le nombre de sinistres indemnisées ainsi que la charge associée pour chaque évènement :

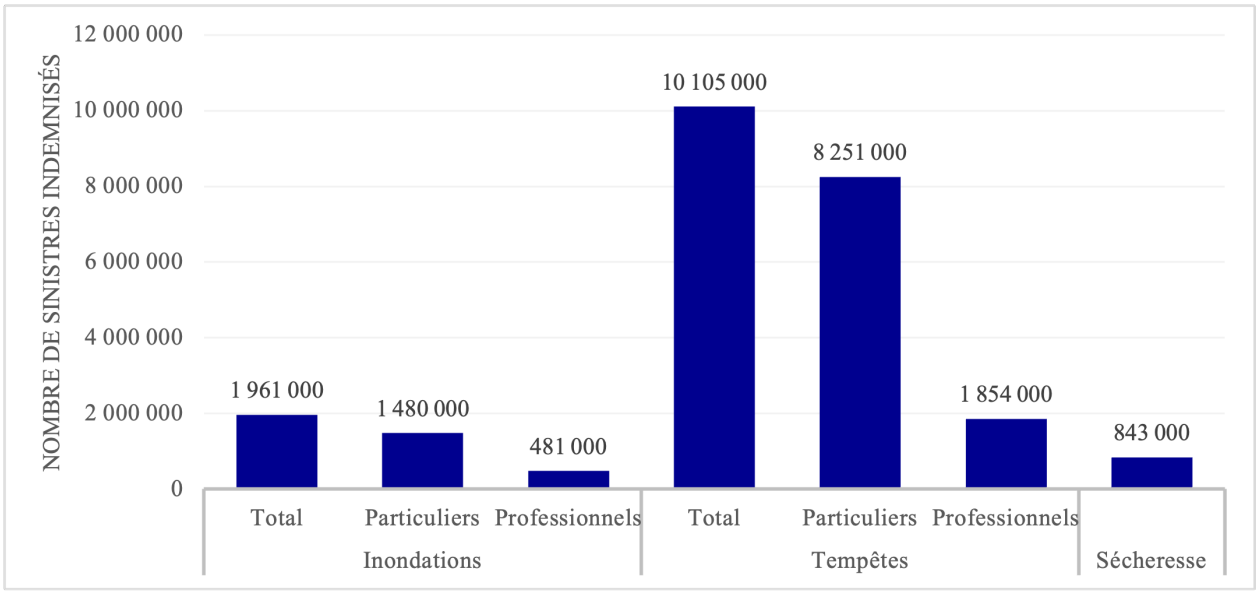


FIGURE 20 – Répartition du nombre de sinistres indemnisés de 1989 à 2019

Effectivement, la charge importante des tempêtes est expliquée par le nombre élevé de sinistres (10 105 000 sinistres contre 1 961 000 sinistres pour les inondations et 843 000

sinistres pour la sécheresse).

Dans la suite, les principaux évènements climatiques sont examinés et présentés plus en détail :

Les tempêtes :

La tempête représente un vent assez fort et crée généralement avec cela des précipitations. Pour que le risque soit qualifié de tempête, il faut que le vent moyen dépasse ou atteigne les 89 km/h pendant au moins 10 minutes. Ces évènements météorologiques se composent de masses nuageuses et s'étendent sur de grandes distances à l'ordre de plusieurs milliers de kilomètres. Lorsque les nuages sont plus denses, les précipitations accompagnant la tempête sont plus intenses aussi. De plus, les vents les plus violents se situent généralement aux endroits où les différences de température sont les plus marquées, comme sur le front froid.

En France, les perturbations météorologiques se forment généralement au-dessus de l'océan Atlantique ou de la mer Méditerranée, lorsque de l'air chaud en surface rencontre un fort vent en altitude (de 10 à 12 km). Au cours de sa phase de développement, la perturbation voit le renforcement du vent en surface, la multiplication des nuages, leur extension verticale, et l'arrivée de pluies. Si ces conditions persistent, la perturbation peut évoluer en tempête, avec des vents dépassant les 89 km/h.

Les tempêtes engendrent des dégâts matériels impactant l'assurance MRH, comme des toitures et des cheminées endommagées. De plus, sachant que les tempêtes sont généralement accompagnées par des précipitations assez violentes, des inondations peuvent aussi être attendu causées par des ruissellements ou des crues ou aussi des orages : le risque tempête est ainsi triple.

Les dégâts engendrés par la tempête sont aussi humains : le vent engendré par la tempête est même capable de faire tomber un individu qui ne se tient pas à quelque chose de stable.

Les inondations :

Le risque d'inondation est considéré comme le premier risque naturel en France menaçant les habitations et les territoires français. En effet, plus de 17,1 millions d'habitants sont exposés aux conséquences des inondations causé par un débordement de cours d'eau et 1,4 millions d'habitants exposés au submersion marine.

Il existe deux types principaux d'inondation en France :

- Les inondations par débordement.
- Les inondations par ruissellement pluvial.

Le tableau ci-dessous explique davantage ces deux types :

	Inondation par débordement	Inondation par ruissellement pluvial
COMMENT	La submersion de zones sèches rapidement ou lentement.	L'accumulation rapide d'eau de pluie sur des surfaces imperméables entraîne le ruissellement et la submersion de zones environnantes.
POURQUOI	Pluies d'intensités et de durées variables accompagné parfois de la fonte des neiges.	Des pluies intenses sur des surfaces imperméables ou saturées en eau peuvent causer le ruissellement.
CONSEQUENCES	Endommagement des biens, perturbation des activités humaines, risque pour la vie humaine, contamination des eaux, ...	Risques d'inondations soudaines, dommages matériels, perturbation des transports, ...

TABLE 15 – Caractéristique de chaque type d'inondation

La grêle :

La grêle se forme à partir de particules de glace de plus de 5 mm de diamètre, appelées grêlons, qui se forment généralement lors d'orages violents dans les cumulonimbus. Moins de 10% des cumulonimbus produisent de la grêle touchant le sol. Lorsque des particules de glace de moins de 5 mm de diamètre tombent d'un cumulonimbus et rebondissent au sol sans se briser, on parle de grésil. Les grêlons, quant à eux, ont un diamètre supérieur à 5 mm et une forme irrégulière, souvent circulaire. Ils se forment à l'intérieur du nuage par dépôts successifs de glace sur des particules appelées « noyaux glaçogènes » avant de tomber au sol sous forme d'averses de grêle.

Les orages les plus violents génèrent les plus gros grêlons en raison de courants forts, permettant aux particules de rester longtemps en suspension dans le nuage et d'absorber plus d'eau.

Les impacts de la grêle sont nombreux, surtout sur l'assurance MRH. Ce risque cause des dégâts sur les toitures, fenêtres, volets et autres parties exposées des biens assurés.

L'étude de France Assureurs et la projection des risques climatiques à 2050 soulignent l'intérêt pour les assureurs, notamment pour AXA France, de considérer davantage ces risques et d'adapter leurs méthodes de provisionnement.

L'objectif de ce mémoire est ainsi d'explorer les méthodes de provisionnement les plus adaptées pour mieux appréhender les risques associés aux EGA en MRH.

II Les nouvelles méthodes de provisionnement

Cette deuxième partie se concentre sur la présentation théorique des méthodes qui seront ensuite appliquées. Elle commence par une exposition des méthodes agrégées, comprenant à la fois la méthode classique d'AXA et une amélioration de celle-ci. Ensuite, les méthodes basées sur l'apprentissage statistique sont introduites.

II.1 Les méthodes agrégées

II.1.A La méthode de provisionnement AXA

Outre la méthode par approche budgétaire, la méthode actuelle d'estimation de la charge à l'ultime des sinistres climatiques utilisée par l'équipe repose sur une approche agrégée par événement.

L'équipe surveille quotidiennement l'évolution de ces sinistres. Cette méthode vise à estimer la charge à l'ultime d'un événement survenu en tenant compte de son développement au cours des sept premiers jours. Ce choix est basé sur une étude concluante indiquant que le nombre/charge de sinistres tend à se stabiliser à partir du 8ème jour et commence à se rapprocher de son ultime valeur. Cette tendance est illustrée par le graphe ci-dessous représentant certains événements climatiques.

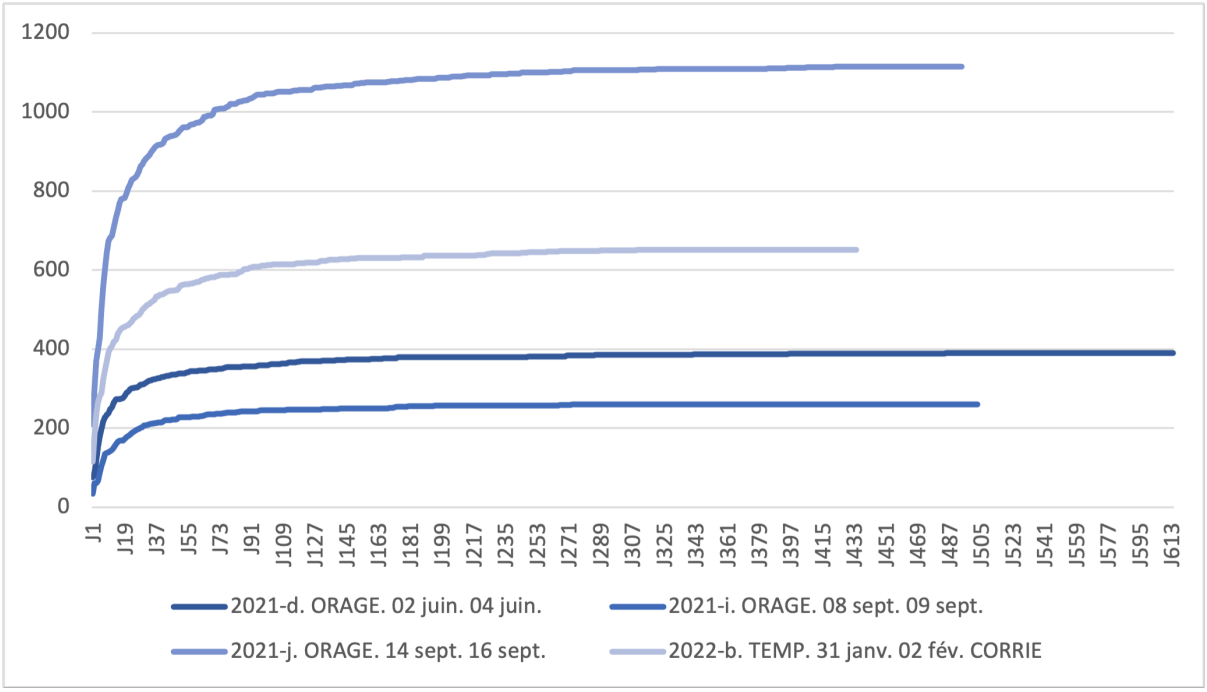


FIGURE 21 – Évolution du nombre cumulé de sinistres par jour selon l'évènement

Ainsi, en se plaçant au 8ème jour, il est possible de connaitre une estimation de la charge ultime de cet évènement et donc une idée de ce chiffrage à la suite d'un évènement climatique spécifique.

Connaissant le nombre de sinistres et les couts associés des sept premiers jours à la suite de la survenance d'un évènement, le calcul de la charge finale prévisible consiste à estimer d'une part le nombre final prévisible (NFP) et le nombre de sinistres clos sans suite (NBCSS) et d'autre part le coût moyen (CM). Le produit des deux donne la charge finale prévisible (CFP).

L'estimation de la charge finale prévisible d'un évènement repose sur quatre étapes.

Etape 1 : Détermination du NFP à l'ultime des sinistres d'un évènement

La méthode d'estimation des NFP des sinistres d'un évènement consiste à se placer au $8^{\text{ème}}$ jour et à projeter le nombre de sinistre à partir du nombre cumulé au $7^{\text{ème}}$ jour de l'évènement en question, selon le développement de tous les événements qui le précèdent. Voici la formule qui calcule le nombre projeté du $i^{\text{ème}}$ jour avec $j \geq 1$ pour un évènement noté E :

$$\text{NombreCumulé}_{j+7}(E) = \text{NombreCumulé}_{(j+7)-1}(E) \times \frac{\sum_{\text{Evenement, date(Evenement)}<\text{date}(E)} \text{NombreCumulé}_{7+j}(\text{Evenement})}{\sum_{\text{Evenement, date(Evenement)}<\text{date}(E)} \text{NombreCumulé}_{7+j-1}(\text{Evenement})}$$

Par exemple, pour j égal à 1, il est possible de calculer le nombre de sinistre sur $8^{\text{ème}}$ jour. En incrémentant j, le nombre de sinistre au $9^{\text{ème}}$ jour est obtenu, etc.

L'ultime est atteint quand le nombre se stabilise. L'estimation du NFP des sinistres d'un évènement est déterminé en ne connaissant que les nombres de sinistres des sept premiers jours à la suite de sa survenance.

Cette méthode peut aussi être rapprocher de la méthode de *Chain Ladder*, qui consiste à calculer des coefficients de passage et à projeter les montants/nombres connus selon un coefficient.

Etape 2 : Estimation du NBCSS FP des sinistres d'un évènement

Comme son nom l'indique, les sinistres clos sans suite correspondent aux sinistres qui ont été déclarés et qui n'ont pas abouti à une indemnisation.

La méthode actuelle de l'estimation du NBCSS FP consiste à multiplier le NFP obtenu à la première étape par le taux du nombre de sinistres clos sans suite. Ce taux est déterminé grâce à une régression linéaire. En effet, cette approche est motivée par l'effet croissant du taux de clos sans suite au cours des dernières années. La graphe ci-dessous illustre l'évolution du taux de clos sans suite au cours des dernières années.

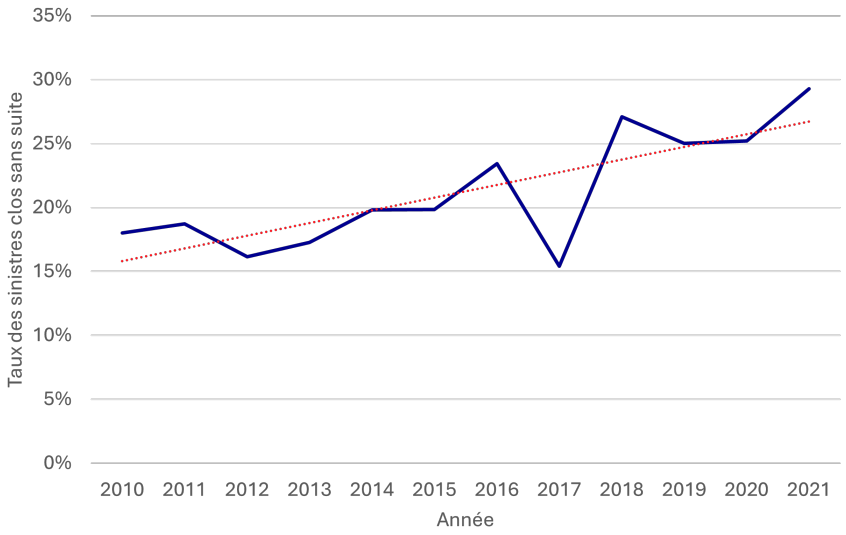


FIGURE 22 – Évolution des taux du nombre de sinistres clos sans suite au cours des dernières années

Il est clair que le taux augmente chaque année (sauf en 2017 avec une légère baisse). Le taux estimé suivant la régression linéaire sera ainsi appliqué à tous les évènements de l'année en cours.

Les NBCSS FP des sinistres d'un évènement E survenu à l'année A est donné par :

$$\text{NBCSS FP}\,(E) = \text{NFP}(E) * \text{tx}_{\text{NBCSS}}(A)$$

Tel que :

- NFP(E) : correspond au nombre final prévisible des sinistres de l'évènement E obtenu à l'étape 1.
- $\text{tx}_{\text{NBCSS}}(A)$: correspond au taux du nombre de sinistres clos sans suite de l'année A obtenu grâce à la régression linéaire des taux de clos sans suite sur les années précédentes en fonctions des années.

Etape 3 : Détermination du CM des sinistres d'un évènement

Cette étape consiste à déterminer le CM de l'évènement E en se positionnant au 7ème jour. Il est obtenu grâce à la formule suivante :

$$CM_7(E) = \frac{\sum_{j=1}^{7} Charge(E)_j}{\sum_{j=1}^{7} Nombre_j(E) - \sum_{j=1}^{7} NBCSS_i(E)}$$

Etape 4 : Détermination de la CFP des sinistres d'un évènement

Connaissant les informations des sinistres sur les sept jours, la charge ultime de l'évènement E est obtenue par la formule suivante :

$$CFP\,(E) = (NFP\,(E) - NBCSSFP(E)) * CM_7(E)$$

Comme toute méthode de provisionnement et dans la même logique que la méthode de Chain Ladder, des hypothèses d'indépendance du nombre de sinistres par évènement

et du caractère homogène des développements devront être vérifiées.

Pour récapituler, la méthode par événement repose sur trois éléments clés :

- Le NFP des sinistres.
- Le NBCSS des sinistres.
- Le coût moyen des sinistres.

En estimant séparément ces trois paramètres, une meilleure compréhension de l'évolution des sinistres et des coûts est faite impliquant un meilleur provisionnement.

Cette méthode représente une amélioration de celle d'AXA, basée sur l'approche budgétaire. En effet, elle consiste à provisionner événement par événement au lieu d'appliquer un seul budget à tous les événements climatiques, en tenant compte des développements du nombre et de la charge des sinistres journaliers. Chaque événement climatique étant unique, l'approche budgétaire, qui fonctionne avec des enveloppes globales, peut biaiser le provisionnement et ne pas refléter l'unicité de chaque événement. Ainsi, l'approche agrégée par événement permet d'adapter les prévisions en fonction des tendances réelles observées sur les dernières années plutôt que de s'appuyer uniquement sur des projections historiques générales. Elle assure une meilleure réactivité, car les premières données de sinistres sont extraites rapidement.

Grâce au suivi quotidien et à l'analyse des sept premiers jours, la méthode affine l'estimation en fonction des premières observations réelles, évitant ainsi le sur-provisionnement ou le sous-provisionnement. Contrairement à une approche budgétaire qui fixe des montants à l'avance, elle réduit les écarts entre les charges prédites et celles réellement observées.

De plus, cette méthode améliore la gestion des sinistres clos sans suite. La prise en compte de ces derniers par régression linéaire permet d'affiner le provisionnement et de s'approcher davantage de la réalité. Cette dynamique évolutive du taux de clos sans suite n'est pas forcément intégrée dans l'approche budgétaire.

Ainsi, cette méthode offre une estimation plus fiable et adaptée aux spécificités de chaque événement, contrairement à une approche budgétaire classique, fondée sur des prévisions globales.

II.1.B La méthode de provisionnement améliorée

II.1.C Motivation de l'amélioration

La méthode de provisionnement actuellement utilisée par AXA présente plusieurs limites qui justifient son amélioration.

Tout d'abord, elle repose sur l'estimation du NFP des sinistres d'un événement en prenant en compte l'ensemble des événements antérieurs, sans distinction ni regroupement. Cette

approche peut biaiser l'estimation du NFP des sinistres car elle suppose que le nombre de sinistres se développe de la même manière quelques soit l'évènement. Or, au fil des années, il est apparu que certains événements partagent des dynamiques similaires en termes de développement quotidien du nombre de sinistres. Plutôt que d'appliquer une approche globale regroupant tous les événements passés, il serait plus pertinent de regrouper ces événements selon des critères de similarité. Ainsi, le NFP des sinistres d'un nouvel événement pourrait être estimé en se basant sur les événements antérieurs les plus comparables en termes d'évolution quotidien du nombre de sinistres.

Ensuite, la méthode actuelle applique un taux de clos sans suite des sinistres unique à tous les événements, sans aucune distinction. Cependant, en pratique, ce taux varie considérablement en fonction de la nature des événements climatiques. Certains sinistres présentent un taux de clos sans suite plus élevé et d'autres des taux bien plus faibles. L'utilisation d'un taux unique conduit donc à des estimations biaisées qui ne reflètent pas toujours la spécificité de chaque événement. Une approche plus fine consisterait à différencier ces taux en fonction de typologies d'événements identifiées à partir des données historiques.

Enfin, la méthode actuelle se base sur le coût moyen des sinistres calculé au $7^{\text{ème}}$ jour après la survenance de l'évènement. Or, cette hypothèse peut induire un biais, car certains sinistres évoluent encore au-delà de cette période avant d'atteindre leur coût ultime. Une meilleure approche consisterait à estimer le coût moyen à son niveau ultime, en tenant compte des tendances observées sur des événements similaires. Cela permettrait d'ajuster plus finement les prévisions et d'éviter des sous-évaluations ou des sur-évaluation des charges à l'ultime.

Ainsi, l'amélioration de cette méthode passe par une approche plus segmentée et dynamique, prenant en compte la typologie des événements, l'évolution réelle des sinistres, une meilleure estimation des taux de sinistres clos sans suite et des coûts des sinistres afin d'améliorer le provisionnement.

II.1.D Méthode d'estimation du nombre final prévisible des sinistres et du coût moyen par évènement

Cette méthode repose sur la même méthodologie que celle d'AXA mais en améliorant chaque étape de celle-ci. Ainsi, l'estimation de la CFP d'un évènement repose sur quatre étapes.

Etape 1 : Détermination du NFP des sinistres par rapprochement d'évènements

L'estimation du NFP des sinistres pour un nouvel événement repose sur l'identification des événements ayant une évolution quotidien similaire en termes de développement du nombre de sinistres. L'objectif est de regrouper les événements présentant des tendances comparables afin d'affiner l'estimation du NFP.

Pour cela, il est essentiel de disposer d'une base de données structurée, où les coefficients de passage du nombre de sinistres sont organisés en lignes, tandis que les événements et leurs dates figurent en colonnes.

Les coefficients de passages quotidiens sont :

$$\text{Coefficient\_De\_Passage}(\text{jour}_i, \text{jour}_j) = \frac{\text{Nombre\_De\_Sinistres}_{\text{jour}_j}}{\text{Nombre\_De\_Sinistres}_{\text{jour}_i}}$$

L'objectif de cette étape est d'identifier des groupes d'évènements ayant des caractéristiques similaires, afin de regrouper les évènements présentant des comportements de développement comparables. L'enjeu est double : former des groupes homogènes tout en garantissant une forte différenciation entre eux. Une fois ces clusters définis, il sera possible d'affecter de nouveaux évènements à ces groupes, dans le but d'estimer les NFP de sinistres de ce nouvel évènement.

Des algorithmes de classification doivent être utiliser, qui permettent de segmenter un ensemble de données en fonction de critères spécifiques. Ces méthodes se divisent en deux grandes catégories :

- Les méthodes supervisées : elles nécessitent un jeu de données labellisé, c'est-à-dire que chaque observation est associée à une catégorie connue au préalable.
- Les méthodes non supervisées : elles ne nécessitent aucune connaissance préalable des classes. L'algorithme détecte des structures naturelles dans les données en fonction des similarités entre observations.

Dans le cadre de l'estimation des NFP de sinistres, il n'existe pas de classification prédéfinie des évènements. Ainsi, les méthodes non supervisées sont privilégiées. Parmi les différentes approches disponibles, les plus utilisées pour ce type de problème sont :

- *K-Means* : algorithme pour les données numériques
- *K-Modes* : basé sur le principe du *K-Means* mais utilise des variables qualitatives pour affecter les observations aux classes.
- *K-Prototype* : algorithme pour les données mixtes.

L'algorithme utilisée est le *K-Means* dans le but de partitionner des données numériques (ici les évènements) en groupes homogènes.

Le clustering *K-Means* est une technique utilisée pour regrouper des observations similaires en plusieurs ensembles distincts. Il repose sur l'identification de K groupes, représentés par des points centraux appelés centroïdes. L'objectif de l'algorithme est d'attribuer chaque observation au groupe le plus approprié tout en minimisant la variance intra-cluster.

Avant de commencer l'application du *K-Means* il est essentiel de déterminer le nombre K de clusters à créer. Une des approches les plus courantes pour se faire est la méthode de coude ou « *Elbow* » en anglais. Cette méthode consiste à effectuer plusieurs exécutions de *K-Means* pour différents nombres de clusters K et à mesurer la somme des carrés des distances intra-cluster. L'inertie diminue généralement à mesure que le nombre de clusters augmente, car plus de clusters permettent de mieux ajuster les données. Le principe de cette méthode est de chercher le nombre de clusters où l'inertie cesse de diminuer rapidement. En d'autres termes, le « coude » est recherché dans le graphique qui montre l'évolution de l'inertie en fonction du nombre de clusters. Le point où l'inertie commence à diminuer moins rapidement, ou quand l'inertie se stabilise, indique le nombre optimal de clusters, car au-delà de ce point, l'ajout de clusters supplémentaires n'apporte plus une réduction significative de l'inertie.

Une fois le nombre de cluster choisi, l'algorithme de *K-Means* (*KNN - K-Nearest Neighbors* en anglais) est appliqué. Son fonctionnement repose sur un processus itératif comprenant deux phases principales :

- Assignation des observations : Chaque point de données est rattaché au centroïde le plus proche en utilisant une mesure de distance, généralement la distance euclidienne.
- Mise à jour des centroïdes : Pour chaque groupe nouvellement formé, le centroïde est recalculé en prenant la moyenne des observations qui y sont associées.

Ces étapes se répètent jusqu'à ce que la position des centroïdes se stabilise, indiquant la convergence de l'algorithme. Voici l'algorithme détaillé :

Algorithm 1 Algorithme K-Means

Initialisation :
— Choisir des variables quantitatives
— Sélectionner un nombre K d'observations tel que K est le nombre de clusters

Tant que Individus réaffectés à des nouveaux groupes **faire**
Étape 1 : Calculer la distance euclidienne de chaque couple (observation, centroïdes)
Étape 2 : Affecter chaque observation au centroïde le plus proche
Étape 3 : Calculer le centre de gravité des clusters créés (devenant les nouveaux centroïdes)
Fin

Résultat : K classes d'observations

Les centroïdes initiaux sont souvent sélectionnés de manière aléatoire, bien que des techniques comme *K-Means++* permettent d'optimiser ce choix pour améliorer la qualité du regroupement. Son fonctionnement repose sur un processus itératif comprenant deux phases principales :

- Choix du premier centroïde : Le premier centroïde est sélectionné de manière aléatoire parmi toutes les observations, exactement comme dans l'algorithme *K-Means*

classique.

- Sélection des centroïdes suivants : Pour chaque observation x, la distance minimale $D(x)$ est calculée entre x et le centroïde déjà choisi le plus proche. L'idée est de favoriser les points qui sont loin des centroïdes précédemment choisis.

  La probabilité de choisir un point x comme prochain centroïde est proportionnelle à $D\,(x)^2$, c'est-à-dire que plus un point est éloigné des centroïdes existants, plus il a de chances d'être sélectionné.

Ces étapes sont répétées jusqu'à ce que K centroïdes aient été choisis. Ensuite, l'algorithme continue avec les étapes classiques de l'algorithme *K-Means* : l'assignation des observations aux centroïdes et la mise à jour des centroïdes.

Dans le cadre du mémoire, l'algorithme *K-Means* est appliqué à une base de données contenant les différents facteurs correspond au développement quotidien du nombre de sinistres d'un événement pour les sept premiers jours après sa survenance. Il permet ensuite d'assigner ce nouvel évènement au cluster formé par des évènements dont le profil de développement est le plus similaire.

Une fois le cluster identifié, l'estimation du NFP des sinistres d'un nouvel évènement E sera déterminé selon la méthode d'AXA présentée dans la partie précédente :

$$\text{NombreCumulé}_{j+7}(E) = \text{NombreCumulé}_{(j+7)-1}(E) \times \frac{\sum_{\substack{\text{Evenement}\in\text{Cluster} \\ \text{dateEvenement}<\text{date}(E)}} \text{NombreCumulé}_{7+j}(\text{Evenement})}{\sum_{\substack{\text{Evenement}\in\text{Cluster} \\ \text{dateEvenement}<\text{date}(E)}} \text{NombreCumulé}_{7+j-1}(\text{Evenement})}$$

Tel que Cluster : l'ensemble des évènements du cluster auquel l'évènement E appartient.

L'amélioration apportée à la méthode d'AXA réside dans le fait que, au lieu de considérer tous les événements antérieurs à l'événement E pour calculer le NFP des sinistres, seuls les événements, dont le développement quotidien du nombre de sinistres est le plus similaire à celui de l'événement E, sont sélectionnés. Cette approche permet de mieux cibler les événements pertinents pour le calcul, diminuant ainsi le biais dans la formule de calcul du NFP.

Cette version améliorée de la méthode d'AXA rapproche davantage cette approche de la méthode de provisionnement Chain Ladder, car elle permet de valider implicitement l'une de ses hypothèses fondamentales, à savoir l'homogénéité de l'échantillon. En créant des groupes de risque homogènes, cette approche renforce la pertinence et la précision des calculs, en s'assurant que les événements permettant le calcul du NFP de sinistres d'un nouvel évènement partagent des caractéristiques similaires avec ce dernier en termes de

développement des sinistres.

Etape 2 : Estimation du NBCSS FP des sinistres d'un évènement

Comme mentionné au début de cette sous-partie, la première méthode applique un taux unique de sinistres clos sans suite, indépendamment de la nature de l'événement. Toutefois, il est probable que ce taux varie en fonction de la nature de l'événement (tempête, grêle, inondation). Il est donc plus pertinent d'appliquer un taux de sinistres clos sans suite spécifique à chaque type d'événement, ce qui permettrait d'obtenir des estimations plus précises et conformes à la réalité.

Ainsi, le taux de sinistres clos sans suite d'un événement E est obtenu en moyennant les taux de sinistres clos sans suite de tous les événements antérieurs à E et de même type que E. On note $T = \{\textit{tempête, grêle, orage, inondation}\}$:

$$\text{tx}_{\text{NBCSS}}(t, E) = \text{moyenne}_{e \in \text{EvenAnt}(E,t)} \left(\text{tx}_{\text{NBCSS}}(e)\right)$$

Tel que t le type de l'événement E appartenant à T et $\text{EvenAnt}(E, t)$ l'ensemble des événements de type t antérieur à E.

Les NBCSS FP des sinistres d'un évènement E appartenant à un type d'évènement T est :

$$\text{NBCSS FP}\,(E) = \text{NFP}(E) * \text{tx}_{\text{NBCSS}}(t, E)$$

Le $\text{NFP}(E)$ correspond au NFP des sinistres de l'évènement E obtenu à l'étape 1, le $\text{tx}_{\text{NBCSS}}(t, E)$ correspond au taux du nombre de sinistres clos sans suite du type d'évènement t auquel l'évènement E appartient et t peut faire référence à tempête, orage/inondation ou grêle.

Pour un nouvel événement, le NBCSS est estimé de manière similaire à la première méthode. Cependant, le taux de sinistres clos sans suite appliqué est désormais spécifique au type d'événement auquel il appartient, offrant ainsi une estimation plus fine du comportement des sinistres pour chaque catégorie d'événements.

Etape 3 : Détermination du CM des sinistres d'un évènement

Dans la méthode classique d'AXA, le coût moyen est calculé à partir des sept premiers jours suivant la survenance d'un événement. Toutefois, cette approche présente une limite : elle ne prend pas en compte l'évolution du coût moyen au-delà de cette période, ce qui peut entraîner une sous-estimation ou une surestimation de la charge ultime des sinistres.

L'amélioration apportée par la nouvelle méthode consiste à adopter une vision à long terme du coût moyen à l'ultime, afin de mieux refléter la réalité et d'affiner l'estimation

de la charge finale des sinistres d'un nouvel événement. Le graphique ci-dessous montre l'évolution du coût moyen des évènements climatiques des années 2021 et 2022 en fonctions du délai de déclaration.

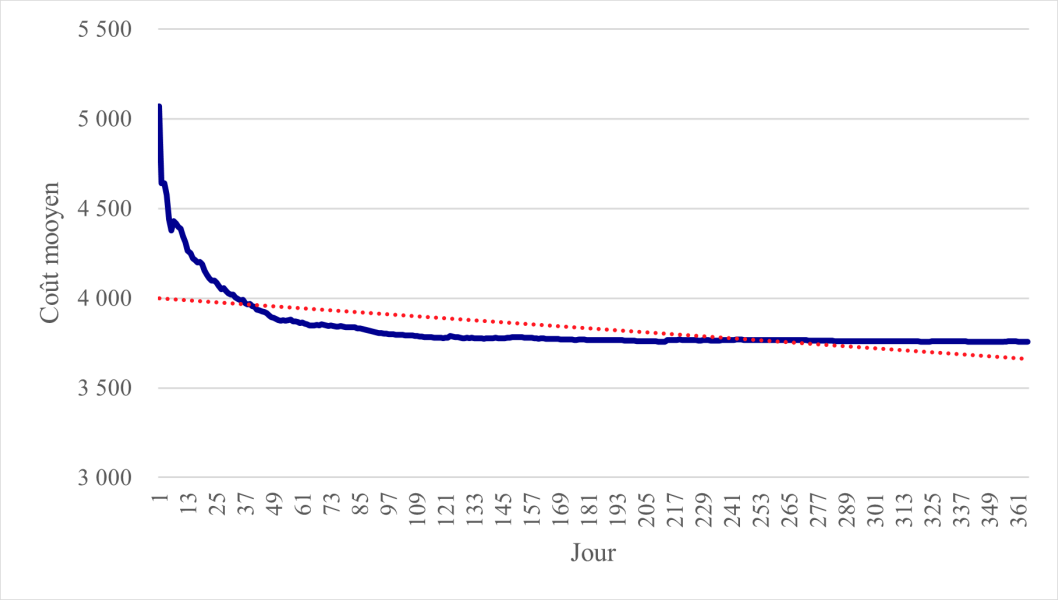


FIGURE 23 – Évolution du coût moyen des évènements climatiques 2021 et 2022 en fonction des jours

Cette analyse montre que le coût moyen diminue progressivement au fil des jours avant de se stabiliser à partir d'un certain seuil. Ainsi, le coût moyen utilisé dans la première méthode et qui est basé uniquement sur les sept premiers jours, entraîne une surestimation de la charge ultime, puisque l'évolution naturelle du coût moyen au-delà de cette période n'est pas prise en compte. Afin de corriger ce biais et de mieux refléter la réalité, la méthode améliorée propose de projeter le coût moyen à l'ultime et de l'intégrer dans la formule de calcul de la charge ultime.

A l'aide des données historiques des événements climatiques, l'évolution du coût moyen de tous les évènements passés jour après jour est obtenue grâce à la formule suivante :

$$\text{Evol}_i = \frac{CM_i}{CM_{i-1}} - 1 \; ; i > 7$$

Le CM_i correspond au coût moyen de tous les évènements passés pour le jour i :

$$CM_i = \frac{CHG_i}{NB_i - NBCSS_i} \; ; i > 7$$

Les $Evol_i$ calculés représentent des coefficients de passage qui sont ensuite appliqués au CM des évènements à prédire.

Ainsi, pour un nouvel évènement, le coût moyen hebdomadaire est calculé comme suit :

$$CM_i = CM_{i-1} * (1 + Evol_i) \; ; i > 7$$

Le coût ultime CM_{ultime} correspond à la dernière valeur obtenue lorsque le coût moyen cesse d'évoluer de manière significative voire devient constant.

En intégrant cette méthodologie, la nouvelle approche permet une estimation plus précise et plus réaliste du coût moyen à l'ultime, se rapprochant ainsi davantage des valeurs finales effectivement observées tout en prenant en compte l'évolution réelle du coût moyen observé.

Etape 4 : Détermination de la CFP des sinistres d'un évènement

Enfin, de manière similaire à la méthode d'AXA, la charge ultime d'un nouvel événement E, sept jours après sa survenance, est calculée selon la formule suivante :

$$\text{CFP}(E) = (\text{NFP}(E) - \text{NBCSSFP}(E)) * CM_{\text{ultime}}(E)$$

Cette approche, fondée sur le rapprochement des événements, présente un net avantage par rapport à la méthode agrégée classique d'AXA. En effet, elle permet d'estimer de manière plus précise la CFP d'un événement en fonction du groupe spécifique auquel il appartient.

Par nature, les événements climatiques sont des phénomènes hautement volatiles et imprévisibles avec des variations marquées selon leur type (tempête, grêle, inondation). Regrouper tous ces événements en une seule catégorie, comme le fait la méthode agrégée classique, ne reflète pas correctement la diversité de leurs comportements, tant en termes de développement du nombre de sinistres que de clôture des sinistres. Chaque type d'événement climatique suit une dynamique particulière qui nécessite une approche spécifique pour estimer avec précision son comportement futur.

De plus, il est bien plus pertinent de prendre en compte le coût moyen ultime plutôt que le CM à $J+7$ après la survenance de l'évènement. En effet, estimer CFP à l'ultime permet de se rapprocher davantage de la réalité, car cette approche prend en compte l'intégralité du développement du sinistre, au lieu de se limiter à une projection à sept jours. Ainsi, l'estimation est plus précise.

Même si les méthodes agrégées sont des méthodes faciles à implémenter et à interpréter, elles demeurent toujours moins précises. En effet, ces méthodes ne prennent pas en compte des caractéristiques spécifiques à chaque sinistre comme le font les méthodes de provisionnement individuelles qui sont présentées dans la partie suivante.

II.2 Les méthodes individuelles par apprentissage statistique

II.2.A Motivation et définitions

Avec la forte croissance du domaine de l'apprentissage statistique ces dernières années et dans la volonté de mesurer au mieux les provisions techniques des évènements

climatiques en forte augmentation, les entreprises d'assurances doivent investir de plus en plus dans des méthodes de provisionnement ligne à ligne basées sur des méthodes d'apprentissage statistique ou *Machine Learning* en anglais. Ce mémoire propose une approche basée sur les méthodes d'apprentissage statistique pour estimer la charge ultime des EGA climatiques.

Dans les années 1950 et avec les premiers travaux sur l'intelligence artificielle, Alan Turing pose les bases du raisonnement automatique et de la capacité des machines à apprendre à partir de données. L'apprentissage statistique s'est développé davantage à la suite des avancées en statistiques, en algorithmique et en puissance de calcul. Grace au développement de l'internet et du Big Data dans les années 1990 et 2000, l'apprentissage automatique est devenu un élément clé de l'intelligence artificielle permettant ainsi d'exploiter des données volumineuses pour automatiser la prise de décision.

Il existe deux types d'apprentissage statistique :

- **L'apprentissage supervisé** : ensemble d'algorithmes qui apprennent à partir d'un ensemble de données étiquetées (avec des entrées et des sorties connues) et établissent un modèle pour prédire les sorties de nouvelles données.
- **L'apprentissage non supervisé** : ensemble d'algorithmes dont aucune sortie n'est fournie, qui identifient des structures sous-jacentes et qui regroupent les données en fonction de leurs similarités.
- **L'apprentissage par renforcement** : ensemble d'algorithmes inspirés du comportement humain et qui permettent d'apprendre par essais et erreurs pour atteindre un objectif optimal.

Ce mémoire se concentre uniquement sur l'apprentissage supervisé. Mathématiquement, les modèles d'apprentissage supervisé visent à analyser la relation entre une variable cible Y et un ensemble de variables explicatives $X = (X_1, \dots X_n)$. Dans le cadre de ce mémoire, Y correspond à la charge ultime des évènements climatiques et X aux variables exogènes représentant les caractéristiques liées au contrat, à l'assuré et à sa sinistralité, et peuvent être de nature qualitative ou quantitative. L'objectif est alors de déterminer la fonction f qui permet de prédire au mieux Y à partir de X, selon la relation :

$$f(X) = Y$$

L'apprentissage supervisé se divise en deux grandes familles de modèles :

- **Les modèles paramétriques** :

  Ces modèles supposent une structure sous-jacente des données à disposition et sont définis par un ensemble fixe de paramètres. Les régressions linéaires et les régressions logistiques font partie de ces modèles.

Toutefois, ces modèles sont difficiles à mettre en place parce qu'il n'est pas souvent évident de trouver une distribution qui s'adapte bien aux données à disposition. Ainsi, ces modèles sont peu présents dans le cadre du provisionnement non-vie.

- **Les modèles non paramétriques** :

  Ces modèles s'avèrent plus adapter au provisionnement non-vie. En effet, ces méthodes ne font pas d'hypothèses sur la distribution des données et ne nécessitent pas d'estimations de paramètres assurant ainsi une meilleure flexibilité. Les modèles non paramétriques sont surtout adaptés aux branches de l'assurance qui sont plus volatiles que d'autres. Ce mémoire se concentre uniquement sur les modèles non paramétriques. Parmi ces modèles on peut citer par exemple les forets aléatoires, les méthodes basées sur les k plus proches voisins (*K-NN*) et d'autres encore...Seuls les modèles utilisés dans le cadre de ce mémoire seront présentés dans la suite.

Dans la suite seuls les modèles de régression seront présentés parce que la variable (charge ultime des sinistres) à modéliser est une variable continue.

II.2.B Arbre de décision

Les arbres de décision *CART* (*Classification And Regression Trees*) ont été développés par Léo Breiman et ses collaborateurs en 1984. Ils constituent une méthode d'apprentissage supervisé permettant de modéliser des relations entre une variable qu'on appelle variable cible et un ensemble de variables explicatives sous la forme d'un arbre de décision, facilement interprétable. En d'autres termes, pour prédire une variable cible Y, le modèle va classifier un ensemble de données selon des règles binaires successives afin de former des sous-groupes homogènes et va représenter graphiquement ces décisions sous forme d'arbre. Ces divisions sont choisies de manière à minimiser l'hétérogénéité des sous-groupes formés selon un critère prédéfini.

Les deux types d'arbres *CART* sont :

- Les arbres de classification, lorsque la variable cible est qualitative.
- Les arbres de régression, lorsque la variable cible est quantitative.

Sachant que la variable à modéliser est une variable numérique, ce sont les arbres de régression qui seront expliqués et mis en œuvre dans le cadre de ce mémoire.

Un arbre de décision est formé par plusieurs éléments :

Sens de la lecture de l'arbre

- La racine qui représente l'ensemble des données initiales.
- Les nœuds intermédiaires qui résultent de la séparation des observations à la suite de l'application des règles de division.
- Les branches qui matérialisent les chemins empruntés par les observations en fonction des règles de segmentation.
- Les feuilles ou nœuds terminaux qui regroupent les sous-populations homogènes obtenues après plusieurs divisions successives. Les dernières feuilles représentent les décisions finales et résultent de l'application successive des règles de division binaires à chaque nœud.

Pratiquement, un nœud correspond à la sélection d'une variable explicative et d'un seuil de division. Ce seuil détermine une séparation de la variable choisie en deux groupes distincts. L'objectif est d'identifier, à chaque nœud, la variable et la règle de division qui réduisent au maximum l'hétérogénéité entre les deux sous-groupes formés. Ce processus est répété jusqu'à ce qu'une règle d'arrêt s'active. Enfin, la valeur moyenne des observations dans les nœuds finaux constitue alors la prédiction de la variable cible pour un individu qui satisfait les conditions qui ont mené à ce nœud. La figure ci-dessous représente un exemple d'arbre *CART*.

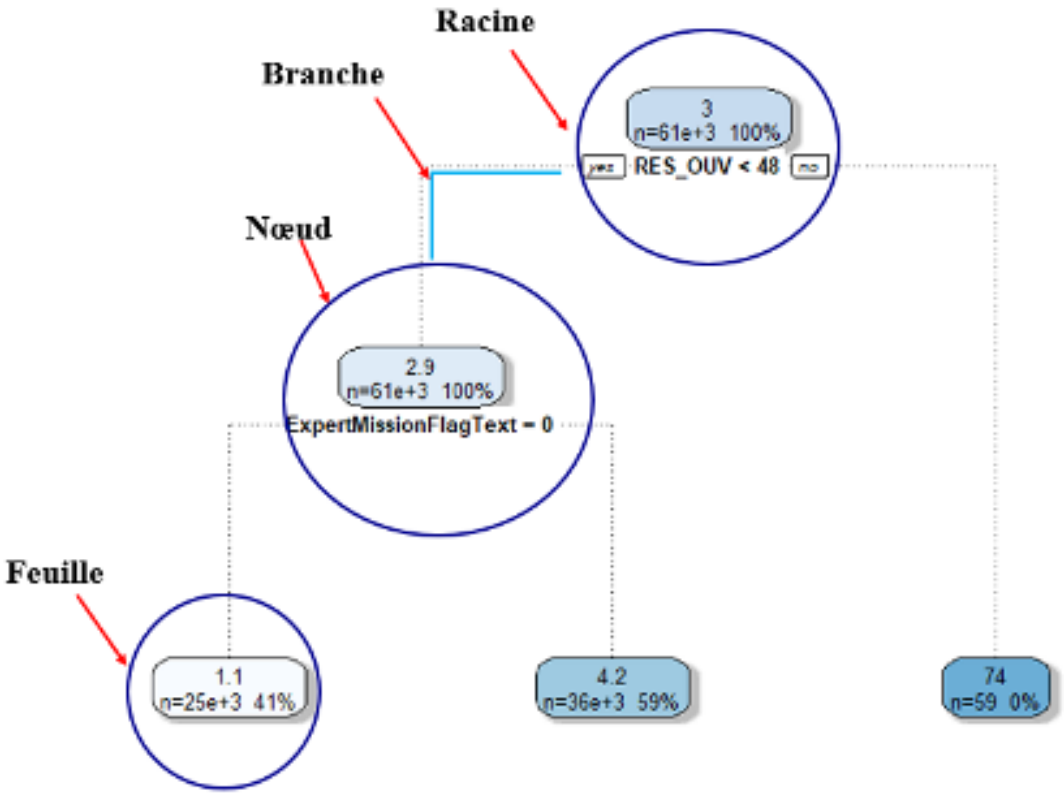


FIGURE 24 – Illustration d'un arbre CART

A partir d'un nœud, deux cas sont possibles :

- Le nœud aboutit à deux branches, dans ce cas on dit que le nœud est coupé ;

- Aucune branche n'est créée.

La division d'un nœud est effectuée selon des :

- Règles d'arrêt :

  Le critère couramment utilisé est d'exiger un nombre minimal d'observations dans un nœud (*min_ samples_ split*) avant d'envisager une nouvelle division. Cela évite de créer des nœuds trop petits et donc d'avoir un modèle trop complexe qui pourrait sur-apprendre aux données (la notion de sur-apprentissage est présentée dans la partie application).

  Il existe aussi d'autres règles d'arrêt défini par une profondeur maximale de l'arbre (*max_ depth*) ou d'une valeur fixée du paramètre de complexité (*cp*).
- Règles de division :

  Les règles utilisées pour diviser un nœud en deux sont basées sur la réduction de l'hétérogénéité qui est mesurée par l'erreur quadratique moyenne. Il faut noter qu'une règle de coupure dépend d'une seule variable.

Soit Y la variable cible à prédire et $X = (X_i),\ i \in 1, \ldots,\ p$ le vecteur des variables explicatives. Si la variable explicative X_i est quantitative alors, pour $c \in \mathbb{R}$:

- Si $X_i < c$, $\bar{y}_{i,c,gauche}$ représente la moyenne de Y et $EG_{i,c}$ représente la somme des carrés des écarts entre les valeurs de Y et $\bar{y}_{i,c,gauche}$
- Si $X_i \geq c$, $\bar{y}_{i,c,droite}$ représente la moyenne de Y et $ED_{i,c}$ représente la somme des carrés des écarts entre les valeurs de Y et $\bar{y}_{i,c,droite}$

Ainsi, pour un nœud donné, l'erreur commise lors de la séparation des individus en fonction des règles $X_i < c$ et $X_i \geq c$ est exprimée par :

$$E_{i,c} = \ EG_{i,c} \ + ED_{i,c}$$

Lors de chaque division, il est souhaitable de minimiser cette erreur.

Donc, la règle de division consiste à trouver le couple optimal (i^*, c^*) tel que $X_i^* \geq c^*$ (qui donne « OUI » ou « NON » en réponse) et $E\,(i^*, c^*) \leq \ E\,(i,\ c)$. X_i^* correspond ainsi au caractère de division et c^* au seuil de division.

En pratique et dans le but d'optimisation du temps de la modélisation, une grille de valeur pour les valeurs de la condition c est mise en place.

L'indice d'amélioration d'un nœud est donné par :

$$I \ = \ 1 \ - \frac{E\,(i^*,\ c^*)}{E}$$

Pour les individus du nœud en question, E est la somme des carrés des écarts entre les valeurs de Y et la moyenne de Y. L'interprétation de cet indice est que plus la valeur de ce dernier du nœud est proche de 1, plus l'amélioration apportée par la règle de division est importante.
Il faut noter que si la variable explicative est qualitative la condition $X_i \geq c$ est remplacée par $X_i = c$ et $X_i < c$ par $X_i \neq c$.

La figure ci-dessous récapitule les étapes de la construction d'un arbre *CART*.

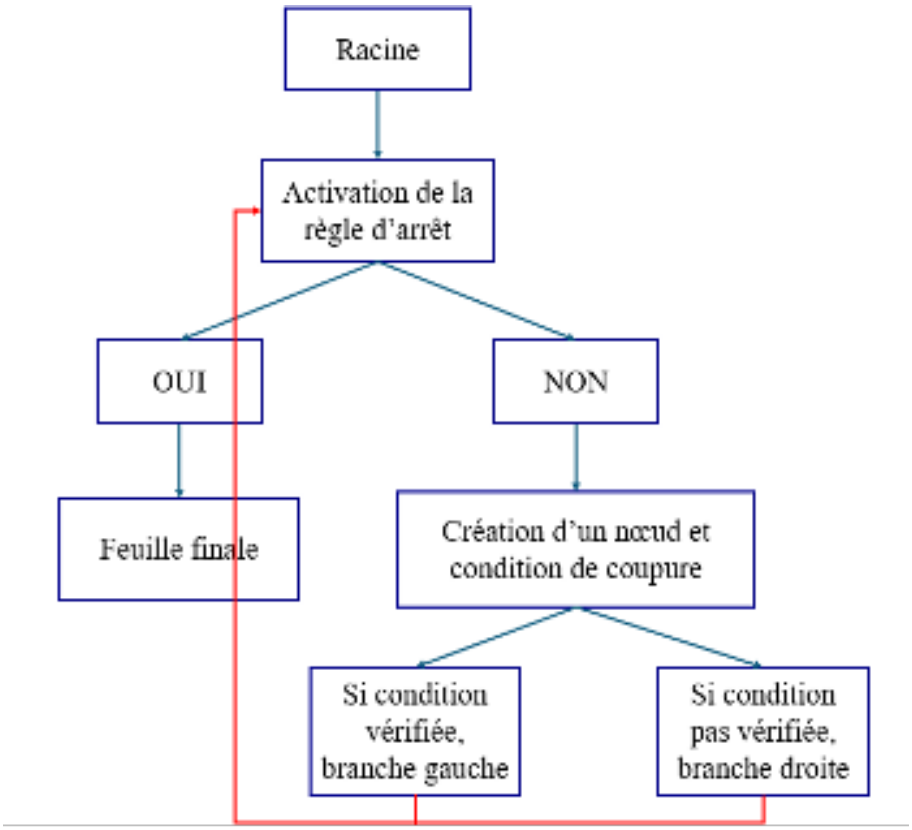


FIGURE 25 – Illustration des étapes de la construction d'un arbre *CART*

Toujours dans le but d'optimiser les prédictions de la variable cible, il est important que l'arbre construit soit le moins complexe possible avec les meilleures prédictions possibles. Ainsi, si l'arbre comporte un nombre excessif de feuilles, il est possible de le simplifier en élaguant ses branches de bas en haut.

Un élagage (ou *pruning* en anglais) optimal vise à trouver un équilibre entre la complexité de l'arbre et la précision des prédictions et ainsi consiste à extraire un sous-arbre optimal à partir de l'arbre maximal, permettant ainsi d'améliorer la généralisation du modèle. Ce compromis peut être atteint grâce à une méthode de validation croisée qui teste différentes versions élaguées de l'arbre. Un élagage pertinent correspond à une valeur du paramètre de complexité *cp* minimisant une erreur spécifique, appelée erreur de validation croisée ou *xerror*. Comme évoqué précédemment, la valeur de *cp* peut être utilisée comme critère

d'arrêt dans la construction d'un nouvel arbre, conduisant ainsi à une version élaguée de l'arbre initial.

L'élagage peut également être appliqué en amont, c'est ce qu'on appelle pré-élagage. Il consiste à appliquer les règles d'arrêt pour arrête la croissance de l'arbre dès que la condition d'arrêt est remplie. Cependant, le risque est d'arrêter trop tôt et d'empêcher l'arbre de capturer des structures pertinentes dans les données.

Une autre méthode est de diviser l'échantillon en données d'apprentissage et données test, d'entrainer le modèle sur les données d'apprentissage puis de considérer l'erreur quadratique moyenne *MSE* , comme fonction de perte pour l'arbre, en cherchant à la minimiser :

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

Où y_i représente les valeurs observées de la variable cible, $\widehat{y_i}$ les valeurs prédites, et n le nombre total d'observations.

Les arbres CART sont connus pour être des modèles simples à implémenter et à interpréter. Contrairement à d'autres modèles, les arbres de décision n'imposent pas de restrictions sur les types de variables utilisées et peuvent donc traiter à la fois les variables quantitatives et qualitatives sans nécessiter de conversion préalable. Ce sont aussi des modèles connus pour leurs capacité à identifier des relations complexes entre les variables explicatives sans nécessiter d'hypothèses linéaires.

Malgré leurs nombreux atouts, elles présentent certaines limites. En effet, lorsque l'arbre est trop profond et complexe, il s'adapte excessivement aux données d'apprentissage, ce qui réduit sa capacité de généralisation créant ainsi du sur-apprentissage. Cet inconvénient est limité grâce à l'élagage présenté plus haut.
De plus, les arbres de décision sont sensibles à tout léger changement dans les données rendant le modèle difficile à généraliser. Comme expliqué déjà, la construction d'un arbre repose sur une recherche du meilleur seuil de division à chaque nœud, ce qui peut être coûteux en temps de calcul, en particulier sur de grandes bases de données.

Les limitations des arbres de décision, notamment leur tendance au sur-ajustement et leur instabilité face aux variations des données, ont conduit au développement d'algorithmes plus robustes, comme les forêts aléatoires (expliqué dans la partie II.2.C).

II.2.C Le *Bagging* et son application aux forêts d'arbres de décisions

Les forêts aléatoires, ou *Random Forest* en anglais, sont des algorithmes d'apprentissage statistique qui reposent sur la méthode du *Bagging* (*Bootstrap Agrégation*). Il est donc essentiel d'expliquer les principes de cette technique.

Le *Bagging* est une méthode d'ensemble, une approche qui combine plusieurs modèles entrainés indépendamment sur des sous-ensembles de données afin d'améliorer la performance globale du modèle. Le but principal de ces méthodes est de réduire la variance, de limiter le sur-ajustement et d'augmenter la robustesse du modèle.

Les méthodes d'ensemble reposent sur deux étapes principales :

- Etablir plusieurs modèles entrainés indépendamment sur chaque sous-ensemble de données.
- Agréger les prédictions de chaque modèle, par exemple en prenant la moyenne des prédictions de chaque modèle dans le cas d'une régression.

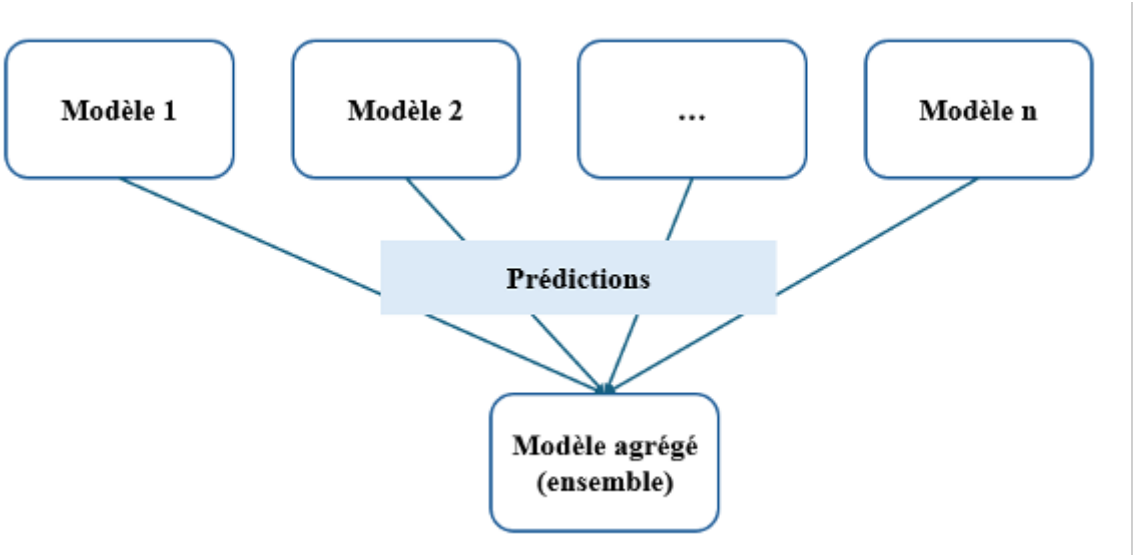


FIGURE 26 – Méthode d'ensemble

Formalisée par Léo Breiman en 1994[4], le *Bagging* est une technique spécifique de méthode d'ensemble qui repose sur la technique de *Bootstrap*. En effet, il consiste à générer plusieurs échantillons *Bootstrap* en effectuant des tirages aléatoires avec remise sur l'échantillon initial selon une distribution empirique. Ensuite, un modèle est entraîné sur chaque sous-échantillon, et les prédictions finales sont obtenues en agrégeant les résultats des différents modèles. Dans le cas d'une régression, les prédictions agrégées sont représentées par la moyenne des prédictions de chaque modèle. Ainsi, le *Bagging* produit un meilleur résultat en réduisant la variance et le sur-ajustement et permet ainsi d'améliorer la précision des modèles.
Le schéma ci-dessous illustre les étapes de la technique de *Bagging* :

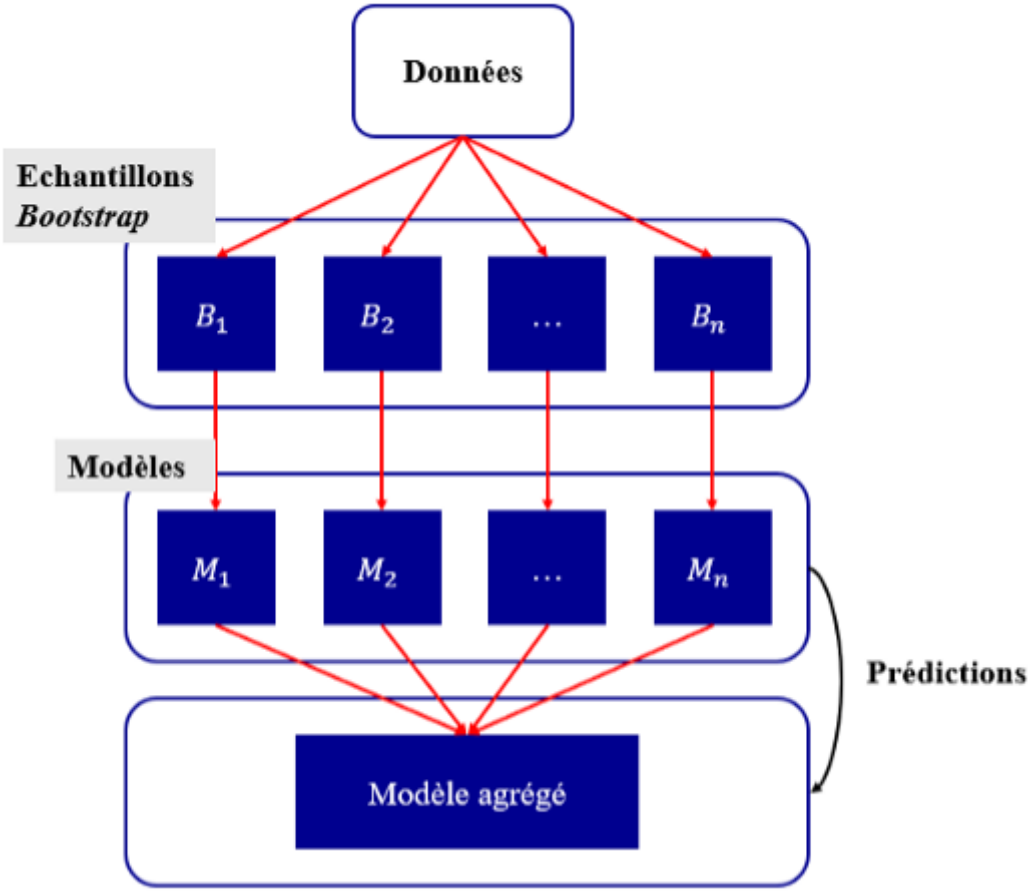


FIGURE 27 – Les étapes du *Bagging*

En 1996[4], Breiman applique le *Bagging* aux arbres de décision, une méthode consistant à entraîner des modèles d'arbres CART sur des échantillons *Bootstrap* obtenus par tirage aléatoire avec remise à partir de l'échantillon de données initial : c'est la technique de *tree bagging* en anglais. Cette application améliore la performance des modèles basés sur les arbres de décision en réduisant la variance de l'estimateur final et, par conséquent, en diminuant l'erreur de prévision.

Voici les quatre étapes principales de ce processus :

- Construire plusieurs arbres de décisions sur des échantillons tirés aléatoirement avec remise de l'échantillon initial.
- Entraîner chaque arbre obtenu.
- Prédire à l'aide de chaque arbre obtenu.
- Moyenner les prédictions pour avoir la prédiction finale.

Voici un exemple de 3 arbres de régression de type *CART* :

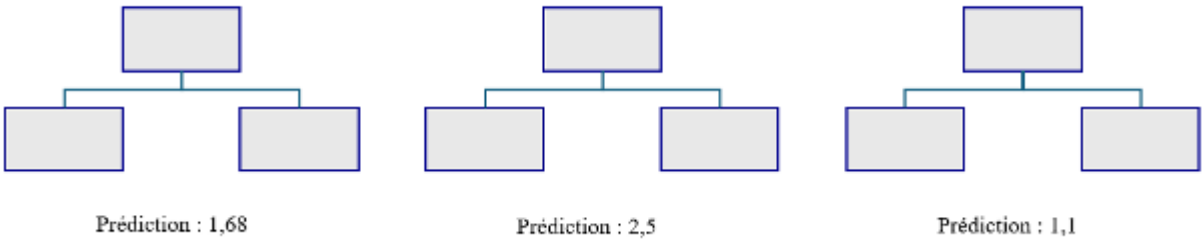


FIGURE 28 – Exemple illustratif d'une forêt d'arbre de décision avec leurs prédictions

Il est évident que chaque arbre fournit une prédiction différente. Si l'on s'était basé uniquement sur un seul arbre, la prédiction aurait été différente. Cependant, en prenant le moyenne des prédictions de l'ensemble des arbres, la valeur prédite est de 1,76. D'où l'intérêt du *Bagging*.

C'est en 2001 que Breiman introduit l'algorithme *Random Forest*[5]. C'est une extension du *Bagging* visant à améliorer encore davantage la performance des arbres *CART* en introduisant une source supplémentaire d'aléa.

En effet, pour que le *Bagging* soit efficace, il est essentiel que les arbres générés soient performants mais aussi suffisamment diversifiés. Pour cela, le *Random Forest* applique la même technique d'échantillonnage aléatoire avec remise que le *Bagging*, mais ajoute une contrainte lors de la construction des arbres : à chaque division d'un nœud, la meilleure séparation est choisie non pas parmi toutes les variables explicatives, mais parmi un sous-ensemble aléatoire de celles-ci. Ce processus est ce qu'on appelle *feature sampling* en anglais. Cette sélection aléatoire des variables permet de réduire la corrélation entre les arbres, limitant ainsi la variance du modèle global. Comme pour le *Bagging*, les prédictions des arbres sont moyennées pour produire une estimation finale plus robuste.

Le schéma de l'algorithme de *Random Forest* est décrit ci-dessous :

Algorithm 2 Algorithme de Forêt Aléatoire (régression)

Entrée :

— Données D avec n observations et p variables explicatives
— Nombre d'arbres N
— Nombre de variables à sélectionner par nœud k parmi les p variables

Pour $i = 1$ à N **faire** :

— Échantillonnage bootstrap : tirer un sous-échantillon E_i avec remplacement depuis D
— Construire l'arbre T_i sur E_i :
Pour chaque nœud n dans T_i :
• Tirer aléatoirement k variables parmi les p
• Choisir la meilleure variable X_j selon un critère MSE
• Partitionner E_i en S_1 et S_2 sur la base de X_j
• Répéter jusqu'à atteindre un critère d'arrêt (profondeur max, nombre minimal d'observations, etc.)
— Ajouter T_i à la forêt

Prédiction pour une observation x **:**

— **Pour chaque** arbre T_i dans la forêt :
• Effectuer une prédiction : $\hat{y}_i = f_{T_i}(x)$
— **Sortie (régression)** : $\hat{y} = \frac{1}{N} \sum_{i=1}^{N} f_{T_i}(x)$

Un élément clé du *Random Forest* en régression est l'utilisation des échantillons *out-of-bag* (*OOB*). Lors de la construction de l'algorithme, chaque observation $z_i = (x_i, y_i)$ peut être absente de certains échantillons *Bootstrap* utilisés pour entraîner les arbres. Ainsi, la réponse de z_i est estimée en agrégeant uniquement les prédictions des arbres pour lesquels z_i n'a pas été utilisé lors de l'apprentissage.
L'erreur *OOB* obtenue est alors une estimation non biaisée de l'erreur de généralisation et peut être assimilée à une forme de validation croisée intégrée. Contrairement à d'autres estimateurs non linéaires, le Random Forest bénéficie de cette approche pour ajuster son hyperparamétrage en continu, grâce à la stabilisation de l'erreur *OOB*.

Un avantage majeur du *Random Forest* est sa capacité à mesurer l'importance des variables utilisées dans les prédictions. À chaque nœud d'un arbre, la variable sélectionnée impacte la qualité de la partition, ce qui permet de quantifier son influence dans la construction du modèle. Cette contribution est cumulée sur l'ensemble des arbres pour obtenir une mesure globale de l'importance des variables.

Les échantillons *OOB* sont également utilisés pour calculer un indicateur supplémentaire d'importance. La méthode consiste à :

- Évaluer l'erreur *OOB* : on utilise les échantillons *OOB* pour mesurer la précision des prédictions.
- Perturber chaque variable : on réarrange aléatoirement les valeurs d'une variable dans les échantillons *OOB*.

- Mesurer la nouvelle erreur *OOB* : si cette modification entraîne une augmentation significative de l'erreur, cela signifie que la variable joue un rôle crucial dans la prédiction.

En agrégeant cette perte de précision sur l'ensemble des arbres, un score d'importance pour chaque variable est obtenu, permettant d'identifier celles qui contribuent le plus à la performance du modèle.

Bien que ces méthodes offrent un avantage significatif en réduisant la variance et en améliorant la précision des prédictions, elles présentent aussi des inconvénients. D'une part, elles diminuent l'interprétabilité des modèles, donnant naissance à des systèmes souvent qualifiés de « boîtes noires ». D'autre part, elles nécessitent une puissance de calcul et une capacité mémoire plus élevée, ce qui peut être un frein pour des ensembles de données volumineux.

II.2.D Le *Boosting* et son application à l'XGBoost

Le *Boosting* est une autre méthode d'ensemble. Toujours dans le même objectif d'amélioration de performance et de la robustesse des modèles, c'est une méthode qui est souvent confondue avec le *Bagging*. En effet, le *Boosting* apporte une amélioration supplémentaire en réduisant non seulement la variance mais le biais des prédictions également.

Cette méthode d'apprentissage statistique vise à améliorer la performance des modèles en combinant plusieurs modèles à faible pouvoir prédictif, appelé *weak learners* en anglais, de manière séquentielle. Les algorithmes *weak learners* sont transformés en algorithmes à haut pouvoir prédictif appelé *strong learners* en anglais.
Contrairement au *Bagging*, où les modèles sont entraînés indépendamment les uns des autres, le *Boosting* entraîne chaque modèle en mettant davantage l'accent sur les erreurs commises par les modèles précédents. Donc, la distinction principale entre le *Bagging* et le *Boosting* réside dans leur méthode d'entraînement : le *Bagging* améliore la précision des modèles faibles en les entraînant simultanément sur plusieurs sous-ensembles de données et le *Boosting*, quant à lui, entraîne les modèles faibles successivement, l'un après l'autre.
Le principe des algorithmes de *Boosting* repose sur l'amélioration progressive des prédictions en accordant une attention particulière aux observations mal classées à l'étape précédente. Pour cela, la notion de résidus est introduite. Elle correspondant à la différence entre les valeurs réelles et les valeurs prédites. Un nouveau modèle est alors entraîné pour prédire ces résidus et ses prédictions sont combinées avec celles du modèle précédent afin d'affiner l'estimation. Ce processus est répété itérativement, chaque nouveau modèle cherchant à corriger les erreurs des précédents en accordant davantage de poids aux observations mal prédites.
La prédiction finale est obtenue en agrégeant les modèles successifs, généralement par une moyenne ou une médiane pondérée par les poids attribués aux classes mal classées en fonction de la qualité des prédictions.
Les types d'algorithmes de *Boosting* les plus connus sont :

- *AdaBoost* (*Adaptative Boosting*) :

  Cet algorithme a été introduit par Freund et Schapire[13] en 1997 et correspond au premier algorithme de *Boosting*. Son principe est d'ajuster les poids des observations mal classées à chaque itération afin de concentrer l'apprentissage sur ces dernières. Il combine plusieurs modèles faibles en un modèle robuste. Cet algorithme n'est adapté qu'au variable binaire, qui n'est pas le case de ce mémoire.

  À partir d'un vecteur de variables explicatives X, le classifieur $G(X)$ prédit une des deux valeurs possibles de Y. L'erreur résultante est exprimée par :

$$E = \frac{1}{N}\sum_{i=1}^{N} I(y_i \neq G(x_i))$$

  L'espérance du taux d'erreur pour les prédictions futures se note $E_{X,Y} I(Y \neq G(X))$.

  Un classifieur faible est un classifieur dont l'erreur d'estimation est seulement légèrement inférieure à celle d'une prédiction aléatoire. Le principe du *Boosting* repose sur l'application répétée de classifieurs faibles sur des versions modifiées des données, générant ainsi une séquence de ces classifieurs faibles : $G_m(x)$, $m = 1, \ldots, M$. La prédiction finale est obtenue par un vote majoritaire pondéré :

$$G(x) = sgn(\sum_{m=1}^{M} \alpha_m G_m(x))$$

  Les coefficients $\alpha_1, \ldots, \alpha_M$ sont calculés par l'algorithme *Boosting* et représentent les poids respectifs de $G_m(x)$, l'objectif étant d'attribuer plus d'importance aux classifieurs les plus performants. La modification des données à chaque étape du *Boosting* consiste à appliquer des poids $w_1, \ldots, w_N$ à chaque observation (x_i, y_i), $i = 1, \ldots, N$. Initialement, les poids sont fixés à $w_i = \frac{1}{N}$ afin d'entraîner le premier classifieur de manière classique. Ensuite, pour m=1,...,M, les poids des observations sont ajustés, et l'algorithme est réappliqué. À chaque itération, les poids des observations mal prédites augmentent, tandis que ceux des observations correctement prédites diminuent. Ainsi, au fur et à mesure des itérations, les observations difficiles à classifier prennent un poids plus important, et chaque modèle se concentre sur ces observations mal classées.

  Voici le schéma de l'algorithme « AdaBoost[27] » :

Algorithm 3 *AdaBoost.M1*

1. Initialisation des poids des observations $w_i = 1/N$, $i = 1, 2, \ldots, N$

2. Pour $m = 1$ à M **faire** :

(a) Entraîner un premier modèle $G_m(x)$ sur l'échantillon d'apprentissage en utilisant les poids w_i.

(b) Calculer :

$$\text{err}_m = \frac{\sum_{i=1}^{N} w_i \cdot \mathbb{I}(y_i \neq G_m(x_i))}{\sum_{i=1}^{N} w_i}$$

(c) Calculer :

$$\alpha_m = \ln\left(\frac{1 - \text{err}_m}{\text{err}_m}\right)$$

(d) Réattribuer les poids :

$$w_i \leftarrow w_i \cdot \exp\left[\alpha_m \cdot \mathbb{I}(y_i \neq G_m(x_i))\right], \quad i = 1, \ldots, N$$

Résultat : $G(x) = \text{signe}\left(\sum_{m=1}^{M} \alpha_m G_m(x)\right)$

À chaque itération, il est essentiel de s'assurer que le modèle obtient de meilleures performances qu'une prédiction aléatoire, c'est-à-dire que son taux d'erreur reste inférieur à 0,5.

Par la suite, de nombreuses variantes de cet algorithme ont été développées pour traiter des problèmes de classification et de régression. Ces variantes se distinguent par plusieurs aspects : la méthode utilisée pour renforcer l'importance des observations mal estimées, la manière dont les modèles sont pondérés lors de l'agrégation, leur finalité (classification ou régression) ainsi que la fonction de perte adoptée pour évaluer l'erreur d'ajustement. Par exemple, Drucker (1997) a introduit l'une des premières extensions de cet algorithme dans le domaine de la régression.

- *Gradient Boosting Machine* (*GBM*) :

Cet algorithme a été proposé par Friedman[14] en 2001. Il repose sur l'optimisation du résidu à chaque étape en minimisant une fonction de perte via la descente de gradient. Il est plus flexible qu'*AdaBoost* et permet d'adapter le modèle aux erreurs précédentes de manière plus progressive.

Ce modèle combine les prédictions de plusieurs arbres de décision pour produire une prédiction finale plus précise. Il est important de noter que tous les *weak learners* dans un GBM sont des arbres de décision.

L'intérêt d'utiliser plusieurs arbres de décisions réside dans la manière dont chaque arbre est construit :

- Sélection de sous-ensembles de variables : À chaque nœud, les arbres utilisent un sous-ensemble différent de variables explicatives pour déterminer le meilleur critère de séparation. Cela permet à chaque arbre d'apprendre des informations légèrement différentes.
- Correction des erreurs précédentes : Contrairement à des arbres indépendants (comme en *Random Forest*), chaque nouvel arbre dans le *Gradient Boosting* est entraîné sur les erreurs des arbres précédents. Ainsi, chaque arbre successif apprend à mieux corriger les prédictions incorrectes du modèle précédent.

Ce processus itératif, où chaque arbre est construit séquentiellement en tenant compte des erreurs des arbres antérieurs, permet d'améliorer progressivement la précision des prédictions finales.

Le *Gradient Boosting*, introduit par Friedman[14] (2001), est aujourd'hui l'un des algorithmes les plus couramment utilisés avec les arbres de décision. Il repose sur l'agrégation séquentielle de multiples modèles de régression à faible pouvoir prédictif afin de construire un modèle global plus performant. Pour accélérer la convergence vers un modèle de meilleure qualité, chaque itération ajuste le modèle en exploitant le gradient de la fonction de perte.

Pour rappel, la descente de gradient est un algorithme d'optimisation visant à minimiser une fonction réelle et différentiable f. Son principe repose sur des mises à jour successives à partir d'un point initial x_0 et d'un seuil de tolérance $\varepsilon > 0$. L'algorithme suit les étapes suivantes :

→ Calcul du gradient de f : $\nabla f(x_k)$

→ Si $\nabla f(x_k) \geq \varepsilon$ alors l'algorithme s'arrête.

→ Si $\nabla f(x_k) < \varepsilon$ alors mise à jour du point $x_{k+1} = x_k - \alpha_k \nabla f(x_k)$ où $\alpha_k > 0$ et représente le pas d'apprentissage.

C'est cette approche qui est reprise dans l'algorithme de *Gradient Boosting*. Pour la suite, voici les notations :

- x_i : variable explicative ;
- y_i : variable cible ;
- n : nombre d'observations ;
- $D = \{(x_i, y_i)\}_{i=1,\dots,n}$: échantillon d'apprentissage ;
- L : fonction de perte différentiable.

A noter que, dans le cas d'une régression, la fonction de perte la plus utilisé est l'erreur quadratique moyenne *MSE* :

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

Où $\hat{y}_i$ est la valeur prédite par le modèle de la valeur réel y.

De même le *RMSE* correspond à la racine carrée du *MSE*.

À l'itération s, le modèle f_s, obtenu à l'étape précédente, est amélioré en ajoutant un terme basé sur le produit d'un pas de descente ρ_s et d'une approximation du gradient opposé de la fonction de perte, estimée par un arbre de régression $\delta_s(x)$. Les pseudo-résidus sont définis comme suit :

$$\widetilde{y}_{i,m} = -\nabla L(y_i, f_{s-1}(x_i))$$

Le modèle mis à jour s'écrit alors :

$$f_s(x) = f_{s-1}(x) - \rho_s \sum_{i=1}^{n} \nabla L(y_i, f_{s-1}(x))$$

Le meilleur pas de descente ρ_s est déterminé en minimisant la fonction de perte :

$$\rho_s = argmin_\rho \sum_{i=1}^{n} L(y_i, f_{s-1}(x_i) + \rho\delta_s(x_i)$$

Ainsi, l'algorithme ajuste progressivement le modèle en se concentrant sur les erreurs passées, améliorant ainsi ses performances à chaque itération.

Ci-dessous est représenté le schéma de l'algorithme « *Gradient Boosting Machine* » :

Algorithm 4 *Gradient Boosting Machine*

Initialisation :
$f_0(x) = \arg\min_\rho \left(\sum_{i=1}^{n} L(y_i, \rho)\right)$

Pour $\mathbf{s} = 1$ à S **faire** :
- Calculer les pseudo-résidus : $\widetilde{y}_{i,s} = -\nabla L(y_i, f_{s-1}(x_i))$
- Estimer le weak learner $\delta_s(x)$ sur les pseudo-résidus $\{(x_i, \widetilde{y}_{i,s})\}_{i=1,\dots,n}$
- Estimer le meilleur pas de descente :

$$\rho_s = \arg\min_\rho \sum_{i=1}^{n} L\left(y_i, f_{s-1}(x_i) + \rho \cdot \delta_s(x_i)\right)$$

- Mettre à jour la fonction :

$$f_s(x) = f_{s-1}(x) + \rho_s \cdot \delta_s(x)$$

Fin

Résultat : $f_S(x)$

Dans le but de limiter le sur-apprentissage, il est courant d'introduire un taux d'apprentissage $\eta < 1$, qui permet de contrôler la vitesse d'ajustement du modèle. Ainsi,

l'actualisation du modèle suit la règle :

$$f_s(x) = f_{s-1}(x) + \eta\, \rho_s \delta_s(x)$$

Ainsi, cette approche ralentit la convergence et évite que le modèle ne s'adapte trop précisément aux données d'entraînement.

Pour optimiser les performances prédictives du modèle, plusieurs paramètres clés doivent être fixés/optimisés :

- Le nombre de *weak learners* à agréger.
- La profondeur des arbres qui correspond à la taille des arbres, définie par le nombre de feuilles.
- Le taux d'apprentissage qui contrôle l'impact de chaque itération sur le modèle global.

L'optimisation de ces paramètres, connue sous le nom de *tuning* des hyperparamètres, est généralement réalisée à l'aide d'un échantillon de validation pour trouver la meilleure combinaison.

- *XGBoost* (*eXtreme Gradient Boosting*) :

Cet algorithme a été introduit par Chen & Guestrin[6] en 2016 et représente une application plus optimisée du gradient boosting. Ainsi il se base sur la descente de gradient dans le but de minimiser la fonction de perte, la plus utilisée étant le *MSE* en régression. L'algorithme vise à optimiser la vitesse d'apprentissage grâce à une gestion efficace de la mémoire et une régularisation avancée pour éviter le sur-apprentissage. Il est connu pour sa haute performance et sa rapidité et est utilisé à la fois pour la classification et la régression, reposant sur des arbres de décision de type *CART*.

En pratique, l'algorithme *XGBoost* sera employé dans la partie application car il offre plusieurs avantages. Il permet notamment d'obtenir des prédictions précises et robustes, tout en assurant un temps de calcul optimisé, un critère essentiel dans un contexte opérationnel. De plus, *XGBoost* est particulièrement adapté au traitement de grands volumes de données, ce qui en fait l'un des algorithmes de *Machine Learning* les plus utilisés aujourd'hui. La fonction de perte de *XGBoost* est définie comme la somme d'une fonction de coût L et d'un terme de régularisation Ω :

$$\Omega(f) = \sum_{s=1}^{S} \left(\gamma L_s + \frac{1}{2}\lambda \left|\left|\omega_s\right|\right|_2^2 + \alpha \left|\left|\omega\right|\right|_1\right)$$

Où :

- γ : paramètre qui applique une pénalisation le nombre de feuilles des arbres et correspond au gain minimal requis pour créer un nouveau nœud.
- λ : paramètre qui applique une pénalisation L2 sur les valeurs des prédictions.
- α : paramètre qui applique une pénalisation L1 sur les valeurs des prédictions.

Tout comme les forêts aléatoires et dans le but de limiter le biais et la variance, *XGBoost* sélectionne aléatoirement des observations et des variables explicatives grâce à des tirages aléatoires. L'algorithme offre plusieurs hyperparamètres à ajuster, comme :

→ $Nrounds$: nombre total d'arbres à agréger.

→ Eta : taux d'apprentissage qui ajuste l'impact de chaque arbre sur la prédiction finale.

→ Min_child_weight : poids minimal requis pour générer un nœud fils.

→ $Subsample$: proportion des observations utilisées pour l'entraînement de chaque arbre.

→ $Colsample\_bytree$: proportion des variables explicatives sélectionnées pour entraîner chaque arbre.

→ Max_depth : profondeur ou nombre de feuilles maximale des arbres.

Ainsi, le *Boosting* améliore la précision du modèle en réduisant les erreurs de prédiction, tout en étant flexible, capable de s'adapter à n'importe quelle fonction de perte et vise surtout à diminuer le sur-apprentissage.

III Mise en application

Dans ce chapitre, des données d'AXA sont appliquées directement aux méthodes présentées dans le chapitre précédent. Tout d'abord, les résultats de la méthode agrégée sont présentés et ensuite ceux de la méthode individuelle.

III.1 Méthodes agrégées

III.1.A Présentation de la base de données

Des bases mensuelles de contrats sont à disposition (regroupant les informations des assurés) et des bases mensuelles de sinistres (regroupant les sinistres des assurés par unité de prestation (UP) sinistrée). Ces données contiennent des sinistres AXA depuis 1989.

L'objectif principal est de récupérer les données journalières relatives à la charge, au nombre de sinistres et au nombre de sinistres clos sans suite pour les événements climatiques survenus en 2021 et 2022. Pour se faire, sur SAS, les données de chaque base mensuelle des années concernées sont extraites et agrégées afin d'obtenir une base unique contenant les informations par mois, de chaque année en question.
Pour s'assurer que les sinistres aient atteint leur charge et leur nombre ultime, les bases mensuelles des années 2021 et 2022 sont observées à décembre 2023, ce qui donne suffisamment de recul pour évaluer la vision ultime des sinistres.

Dans le cadre de la modélisation agrégée, il est essentiel d'avoir une base journalière. Ainsi, il faut transformer les bases mensuelles en bases journalières. Une nouvelle variable JO a été créée pour mesurer le délai écoulé entre l'occurrence d'un événement et son enregistrement officiel. Elle correspond à la différence entre ces deux dates, permettant ainsi d'analyser le temps de réaction ou de déclaration.
De plus, il faut créer une variable décrivant l'événement pour identifier les sinistres liés aux événements climatiques. Cette information n'étant pas directement disponible dans la base, il a fallu utiliser une liste d'événements et leurs dates d'occurrence pour effectuer ce rapprochement. Cette liste a été récupérée à partir d'un historique de suivi des climatiques AXA France.

Afin d'exploiter ces informations de manière structurée, la base est organisée en fonction des événements et des jours ouvrés. Ainsi, un traitement d'ordonnance de la base est effectué pour associer à chaque événement une colonne de jours ouvrés $(J_0,\ J_1,\ \dots J_{max})$ avec les variables numériques associées. Pour chaque évènement, le maximum de jour ouvrés est récupéré à l'aide d'un macro-programme SAS. Ensuite, la liste $(J_0,\ J_1,\ \dots J_{max})$ est créée et est mergé à la base. Ainsi, cette étape est répétée pour chaque évènement pour avoir le bon ordre.

Une clé unique est ensuite définie, basée sur l'événement et le jour et les variables suivantes sont créées :

- La somme cumulée de la variable charge des sinistres est calculée :
  - $CHG\_\text{cumulé}_0 = CHG_0$
  - $CHG\_\text{cumulé}_i = CHG\_\text{cumulé}_{i-1} + CHG_i$ pour $i \geq 1$
- La variable nombre des sinistres :
  - $NB\_\text{cumulé}_0 = NB_0$
  - $NB\_\text{cumulé}_i = NB\_\text{cumulé}_{i-1} + NB_i$ pour $i \geq 1$
- La variable nombre des sinistres clos sans suite :
  - $NBCSS\_\text{cumulé}_0 = NBCSS_0$
  - $NBCSSC\_\text{cumulé}_i = NBCSS\_\text{cumulé}_{i-1} + NBCSS_i$ pour $i \geq 1$

Pour rappel, cette approche repose sur un modèle agrégé et par conséquent, la base doit être agrégée par événement et par jour ouvré (JO).

La finalité est une base agrégée formée par les variables suivantes :

Nom de la variable	Description
Évènement	La concaténation de l'évènement en question et de l'année de survenance
Année	L'année de survenance
JO	Les jours ouvrés de développement
Clé	Concaténation de la variable évènement et JO
NB	Le nombre des sinistres
NB_cumulé	Le nombre des sinistres cumulé
NBCSS	Le nombre des sinistres clos sans suite
NBCSS_cumulé	Le nombre des sinistres clos sans suite cumulé
CHG	La charge des sinistres
CHG_cumulé	La charge des sinistres cumulée

TABLE 16 – Liste des variables de la base de données agrégées

III.1.B Modèle mis en œuvre par AXA

La méthode agrégée est appliquée sur Excel. Cette section présente la construction de la maquette ainsi que les résultats obtenus en appliquant cette méthode (présentée dans la section II.1.A) à la base de données créée.

La première étape consiste à estimer le NFP de sinistres climatiques pour l'année 2022 à partir des données des sinistres climatiques de l'année 2021.

Dans le but de projeter le nombre de sinistres, la feuille « Cadence développement » a été structurée dans un format inspiré des triangles de liquidation : les jours $(J1,\ J2,\ \ldots)$ sont disposés en ligne et les événements climatiques de 2021 à 2022 en colonne. Ainsi, les nombres cumulés de sinistres pour les événements de 2021 sur l'ensemble des jours sont récupérés, et uniquement les nombres des 7 premiers jours pour les événements de 2022.

Pour estimer le NFP des sinistres de 2022, une hypothèse est établie : tous les événements climatiques suivent une évolution similaire en termes de développement du nombre de sinistres. Cette hypothèse peut être vérifier graphiquement selon le cc plot des développement aléatoirement choisis :

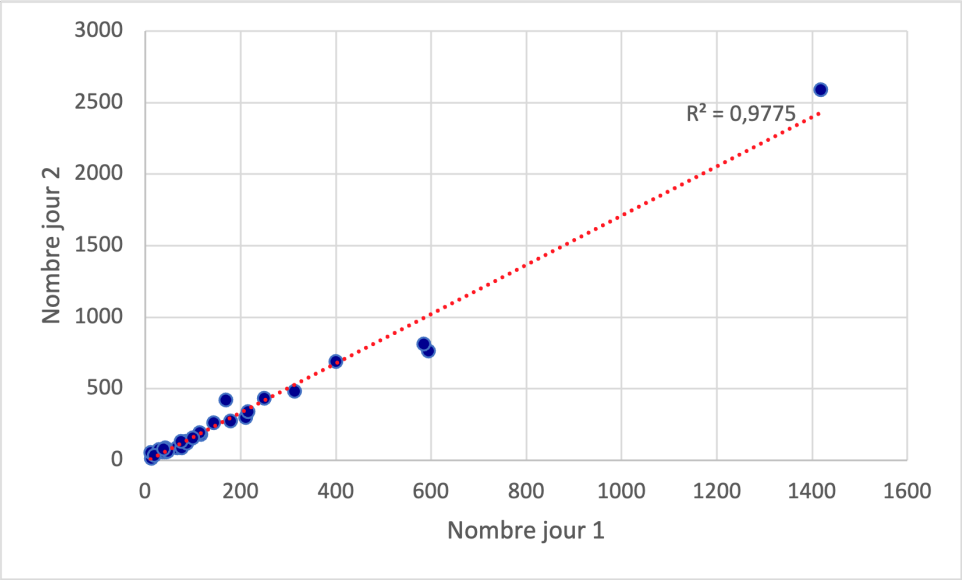


FIGURE 29 – CC plot du développement J2/J1

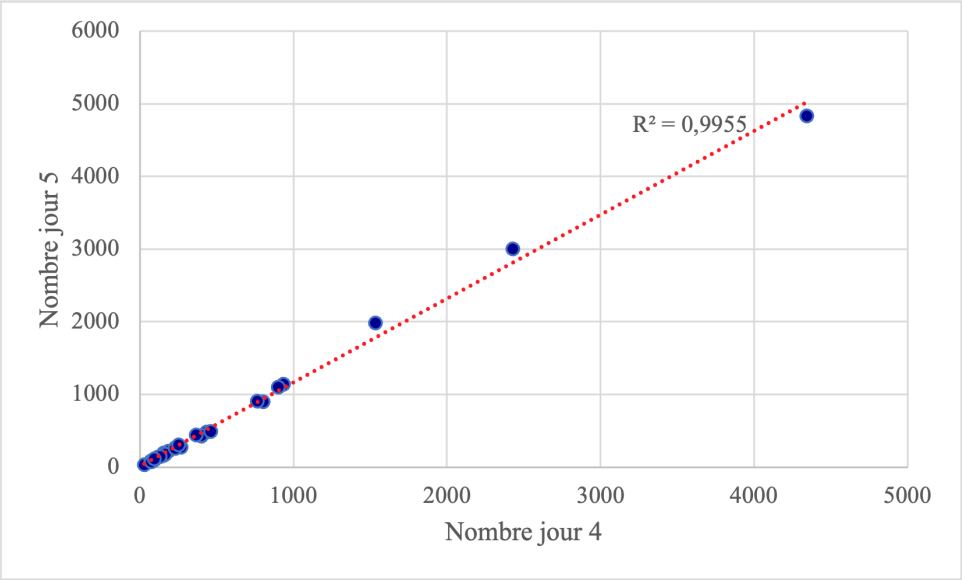


FIGURE 30 – CC plot du développement J5/J4

La représentation graphique permet de valider l'hypothèse émise sur le développement du nombre de sinistre.

En s'inspirant de la méthode de projection Chain Ladder, pour chaque événement de 2022, les nombres de sinistres sont projeter en partant du septième jour. Dès lors que la valeur projetée se stabilise, le nombre de sinistres atteint son ultime.

Ensuite, pour chaque événement de 2022, le coût moyen des sept premiers jours de survenance (CM J+7) est déterminé :

EVENEMENT	JOUR_OUV	NB	NB_CUMUL	NBCSS	NBCSS_CUMUL	CHG (en €)	CHG_CUMUL (en €)	CM à J+7 (en €)
2022-a. TEMP. 08-11 janv.	J7	11	176	4	66	9	464	4 194
2022-b. TEMP. 31 janv. 02 fév. CORRIE	J7	34	324	14	86	47	694	2 909
2022-c. TEMP. 06-07 fév.	J7	7	110	1	24	6	384	4 469
2022-d. TEMP. 17-21 fév. EUNICE FRANKLIN	J7	545	3 988	95	836	1 477	10 129	3 213
2022-e. INOND. 12-15 mars .	J7	8	206	3	62	11	418	2 915
2022-f. TEMP. 07-09 avril DIEGO	J7	63	575	17	147	150	1 248	2 916
2022-g. ORAGE. 15 mai	J7	-	80	-	24	-	281	5 054
2022-h. GRELE. 19-23 mai	J7	25	974	8	183	94	5 909	7 468
2022-i. GRELE - INOND 02-06 juin	J7	149	2 490	27	411	795	27 026	13 001
2022-j. GRELE. 19-23 juin	J7	562	5 628	83	804	6 092	66 831	13 854
2022-k. GRELE. 24-25 juin	J7	14	177	7	50	12	635	5 014
2022-l. GRELE. 26-30 juin	J7	32	1 028	5	176	109	4 225	4 961
2022-m. GRELE. 03-04 juillet	J7	6	272	-	33	39	1 574	6 577
2022-n. GRELE. 20 juillet	J7	32	392	3	55	199	5 356	15 887
2022-o. ORAGE. 14-18 aout	J7	191	1 523	36	319	908	7 301	6 066
2022-p. GRELE. 05-08 septembre	J7	43	571	13	153	107	2 046	4 898
2022-q. GRELE. 14 septembre	J7	17	152	2	24	63	657	5 119
2022-r. TEMP. 23-24 oct. 2022	J7	3	268	-	58	6	3 044	14 493

FIGURE 31 – Résultats des CM des 7 premiers jours des évènements climatiques 2022

Le nombre de clos sans suite est estimé via une régression linéaire basée sur les 10 dernières années.
À partir d'un code SAS, on extrait les taux de sinistres clos sans suite des années 2010 à 2022 et effectuons une régression linéaire pour estimer le taux applicable en 2022.

Année	Taux clos sans suite
2010	18%
2011	19%
2012	16%
2013	17%
2014	20%
2015	20%
2016	23%
2017	15%
2018	27%
2019	25%
2020	25%
2021	29%

FIGURE 32 – Taux de sinistres clos sans suite en fonction de l'année (2014 à 2021

Les résultats du modèle sont synthétisés sous forme de tableau avec les valeurs suivantes :

EVENEMENT	NB_estimé	Nb_observé	CM à J+7	NBCSS_estimé	NBCSS_observé	CHG_estimée (en €)	CHG_observée (en €)	MSE
2022-a. TEMP. 08-11 janv.	391	420	4 194	108	129	1 184 294	987 143	38 868 739 153
2022-b. TEMP. 31 janv. 02 fév. CORRIE	718	654	2 909	199	149	1 510 741	1 459 104	2 666 340 040
2022-c. TEMP. 06-07 fév.	243	198	4 469	68	42	786 425	560 083	51 230 544 655
2022-d. TEMP. 17-21 fév. EUNICE FRANKI	8 838	8 011	3 213	2 450	1 426	20 519 190	18 653 778	3 479 763 758 394
2022-e. INOND. 12-15 mars .	456	437	2 915	126	126	960 021	844 226	13 408 460 297
2022-f. TEMP. 07-09 avril DIEGO	1 274	1 208	2 916	353	255	2 684 257	2 613 958	4 941 892 798
2022-g. ORAGE. 15 mai	176	179	5 054	49	43	643 982	500 674	20 537 102 644
2022-h. GRELE. 19-23 mai	2 158	1 821	7 468	598	293	11 646 336	9 258 051	5 703 903 559 290
2022-i. GRELE - INOND 02-06 juin	5 517	5 758	13 001	1 530	875	51 840 527	54 177 113	5 459 635 700 014
2022-j. GRELE. 19-23 juin	12 471	9 407	13 854	3 458	1 424	124 871 933	89 016 487	1 285 613 021 025 430
2022-k. GRELE. 24-25 juin	392	308	5 014	109	84	1 422 308	1 091 280	109 579 700 874
2022-l. GRELE. 26-30 juin	2 277	1 798	4 961	631	314	8 164 840	6 610 470	2 416 067 207 464
2022-m. GRELE. 03-04 juillet	603	517	6 577	167	63	2 865 166	2 585 674	78 115 776 530
2022-n. GRELE. 20 juillet	869	676	15 887	241	100	9 976 419	6 541 342	11 799 752 147 299
2022-o. ORAGE. 14-18 aout	3 374	3 394	6 066	936	709	14 793 952	13 697 484	1 202 241 810 387
2022-p. GRELE. 05-08 septembre	1 266	1 139	4 898	351	266	4 480 030	3 661 230	670 432 989 359
2022-q. GRELE. 14 septembre	338	328	5 119	94	49	1 248 931	1 251 429	6 238 576
2022-r. TEMP. 23-24 oct. 2022	594	468	14 493	165	92	6 217 699	3 498 472	7 394 196 342 260
TOTAL	41 954	36 719		11 633	6 441	265 817 051	217 007 999	8 576 643

FIGURE 33 – Résultats de la méthode agrégée classique d'AXA

Pour chaque évènement 2022, les valeurs observées sont :

- NB_observé : le nombre de sinistres ultime observé.
- NBCSS_observé : le nombre de sinistres clos sans suite observé.
- CHG_observé : la charge totale observée

Pour chaque évènement 2022, les valeurs estimées sont :

- NB_estimé : le nombre de sinistres ultime estimé après projection.
- CM à J+7 : Le coût moyen des sinistres calculé à J+7.
- NBCSS_estimé : le nombre de sinistres clos sans suite ultime estimé

Pour chaque évènements 2022, on obtient ainsi :

- CHG_estimée : La charge ultime estimée, calculée comme suit :

$$\text{CHG estimée} = (\text{NB\_estimé} - \text{NBCSS\_estimé}) \times \text{CM à J+7}$$

- *RMSE* : Pour mesurer la qualité de l'estimation, le *RMSE* global est calculé

$$RMSE = \sqrt{\frac{1}{\text{nombre d'évènements 2022}} \sum_{i=1}^{\text{nombre d'évènements 2022}} (\text{CHG\_observé}_i - \text{CHG\_estimée}_i)}$$

Cette évaluation permet d'analyser la fiabilité et la qualité de prédiction du modèle.

Cette méthode estime 265 817 051€ de charge en 2022, pour 217 007 999€ réellement observé. Cela montre que le modèle sur-estime la charge.
Le *RMSE* est de l'ordre de 8 576 643.

III.1.C Modèle AXA amélioré

Cette méthode constitue une amélioration du modèle agrégé présenté précédemment, visant à affiner l'estimation des sinistres en intégrant une approche plus détaillée et plus adaptée à l'évolution réelle des événements climatiques.

L'objectif principal de cette méthode est de regrouper les événements selon des comportements similaires et d'adapter les calculs en fonction de ces regroupements, notamment en estimant le nombre final prévisible et le taux de sinistres clos sans suite par évènement et en projetant le CM à l'ultime.

L'un des éléments clés de cette approche repose sur le calcul des coefficients de passage entre deux jours consécutifs, ce qui permet de mieux comprendre la dynamique du développement des sinistres au fil du temps. Pour chaque jour i, le rapport entre le nombre de sinistres cumulés de ce jour et celui observé le jour suivant est défini par la formule suivante :

$$\frac{J_{i+1}}{J_i} = \frac{\text{NombreCumulé}_{i+1}}{\text{NombreCumulé}_i}$$

Ces coefficients servent de base pour la suite du modèle. En effet, une nouvelle feuille de calcul intitulée « Coefficient_développement » a été créée. Elle suit le même format que la feuille « Cadence développement », mais au lieu d'y retrouver directement les nombres de sinistres cumulés, elle contient les coefficients de passage entre chaque jour successif. Ainsi, en colonne se trouvent les évènements de 2021 et de 2022 et en ligne se trouvent les coefficients ($\frac{J_2}{J_1}$, $\frac{J_3}{J_2}$, ...). Ces coefficients sont calculés directement à partir des données issues de la feuille « Cadence_développement », ce qui permet une automatisation du calcul. Grâce à la similarité des structures, il suffit alors d'appliquer la formule de calcul et de l'étendre pour obtenir l'ensemble des coefficients nécessaires.

Une fois ces coefficients extraits, l'étape suivante consiste à déterminer le NFP des évènements 2022.
Pour se faire un rapprochement est fait pour regrouper les évènements qui se ressemblent en termes de développement de nombres de sinistres par jour. Le regroupement est réalisé à l'aide d'un modèle non supervisé de type *K-Means*. Afin de construire les groupes d'évènements similaires, le modèle utilise les coefficients de passage des sept premiers jours ($\frac{J_2}{J_1}$, $\frac{J_3}{J_2}$, ..., $\frac{J_7}{J_6}$) des événements survenus en 2021, calculés préalablement.
Ainsi la base formée par ces coefficients en colonne et des évènements en ligne sert de base d'apprentissage pour définir des groupes d'évolution similaires.

Tout d'abord la base est divisée en deux : une base pour les événements de 2021 et l'autre pour ceux de 2022. Voici un exemple de la base formée par les évènements de 2021 :

	J2/J1	J3/J2	J4/J3	J5/J4	J6/J5	J7/J6
0	1.240833	1.604433	1.083717	1.347625	1.171969	1.220103
1	1.349226	1.419157	1.463491	1.212758	1.152382	1.129179
2	1.552717	1.269588	1.232279	1.119924	1.131846	1.118311
3	1.172803	1.346367	1.283441	1.172767	1.111857	1.094467
4	1.580435	1.348107	1.267469	1.196870	1.135101	1.112549
5	5.232727	1.797255	1.482262	1.299270	1.175124	1.051256
6	1.425130	1.622473	1.172523	1.235212	1.176317	1.107920
7	2.370673	1.637112	1.218015	1.079145	1.037010	1.000000
8	1.746755	1.000000	1.115588	1.296033	1.204317	1.123281
9	1.433933	1.230006	1.078207	1.071709	1.172727	1.114768
10	1.842006	1.261083	1.206699	1.101611	1.022430	1.059648
11	1.295052	1.050484	1.154782	1.228258	1.144180	1.081966

FIGURE 34 – Base de données des coefficients de passage du nombre de sinistres des 7 premiers jours par évènement

La première étape d'un algorithme *K-Means* consiste à trouver le nombre de *cluster*, c'est à dire le nombre de groupes. Il est déterminé grâce à la méthode de coude ou « *elbow* » en anglais. Le graphe ci-dessous représente la somme des distances *intra-cluster* (l'inertie des *clusters*) en fonction du nombre de *clusters*.

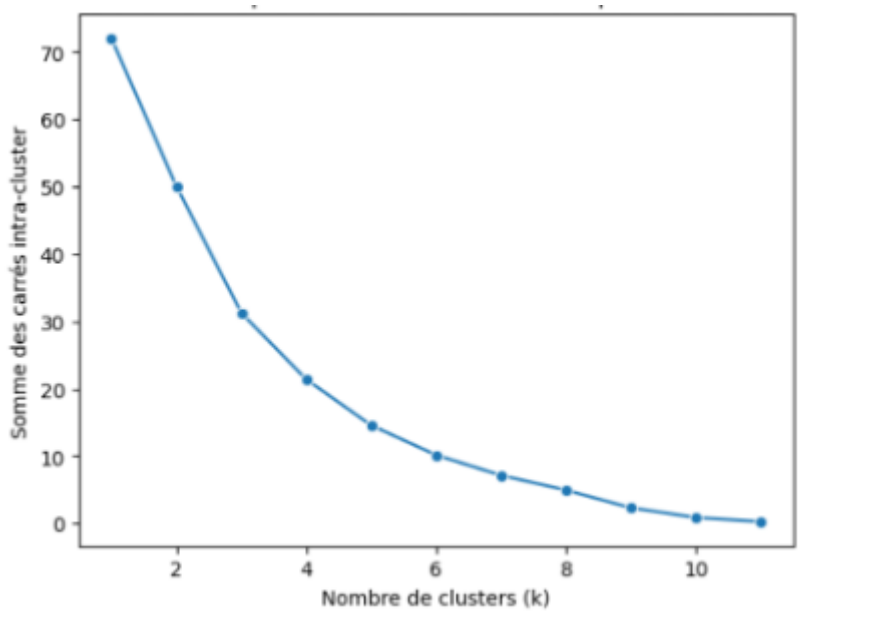


FIGURE 35 – Résultat de la méthode du coude pour déterminer le nombre optimal de *cluster*

Normalement, l'inertie diminue en fonction du nombre de *cluster*. Cependant, cette diminution n'est pas linéaire : la diminution de l'inertie devient plus ou moins marginale à partir d'un certain nombre de *cluster*. Ainsi, le choix du nombre de *cluster* correspond

au point du graphe ou la pente change brutalement en formant une forme de « coude », d'où le nom de la méthode. En effet, avant ce point, l'inertie diminue beaucoup quand on rajoute un autre *cluster*, en revanche, après ce point l'inertie diminue de moins en moins en rajoutant un *cluster* supplémentaire. Cela montre que rajouter plus de *cluster* n'apporte pas de gain mais complexifie juste le modèle.
On observe que l'inertie chute fortement entre 1 et 4 *clusters* et se stabilise après 4 *clusters*. Pour cela, 3 et 4 semblent pertinents pour le nombre de *clusters*. Vu le faible nombre de données et pour ne pas complexifier le modèle, le choix s'est orienté vers 3 *clusters*.

L'inconvénient d'une telle méthode est qu'elle est basée que sur une interprétation graphique, ce qui peut biaiser le choix et le résultat. Pour cela, une mesure est calculée pour évaluer la qualité d'un *clustering* : c'est la mesure de silhouette. En effet, cet indicateur mesure le degré de cohésion et de séparation des *clusters* et les comparent. Pratiquement, la mesure de silhouette calcule :

- La compacité *intra-cluster* : au sein du même *cluster*, la moyenne des distances entre un point et les autres points.
- La séparation *inter-cluster* : entre plusieurs *clusters*, la moyenne des distances entre le point et les autres points des *clusters* les plus proches.

La formule mathématique de cette mesure pour un point x est donnée par :

$$s\left(x\right) = \frac{h\left(x\right) - c(x)}{\max(h\left(x\right) - c\left(x\right))}$$

Avec $h\left(x\right)$ la séparation inter-*cluster* et $c(x)$ la compacité intra-*cluster*.
Ainsi, l'indicateur global S pour tous les points est obtenue en moyennant toutes les mesures de silhouette des points déjà calculées. L'indicateur est analysé comme suit :

- Si le score S est proche de 1 : les points sont associés au bon *cluster* ce qui implique un très bon *clustering*.
- Si le score S est proche de 0 : les points sont à la frontière entre plusieurs clusters.
- Si le score S est proche de -1 : les points sont dans le mauvais cluster ce qui implique un mauvais *clustering*.

Dans le cas de cette application, pour un nombre de *clusters* de l'ordre de 3, l'indice de silhouette est de 0,4, ce qui implique un bon choix du nombre de *cluster* et donc un bon *clustering*.

Par la suite, une Analyse en Composantes Principales ACP (*PCA - Principal Component Analysis* en anglais) est effectuée pour réduire la dimensionnalité des données. En effet, vu que les données sont formées de sept variables, cela peut compliquer l'interprétation et le *clustering*. L'ACP permet de représenter les données avec moins de dimensions tout en conservant le maximum d'information. Ainsi, en réduisant les données à 2 dimensions principales, il est plus facile de visualiser les *clusters* obtenus grâce au *K-Means*. De plus, pour améliorer davantage la qualité du *clustering*, l'ACP transforme les variables

initiales en nouvelles variables qu'on appelle composantes principales et qui ne sont pas corrélées.

La première étape d'une ACP consiste à créer un objet ACP en normalisant les données. L'extraction des variances expliquées par chaque composante permet de voir combien d'information est conservée par chaque nouvelle dimension.
Ensuite la variance cumulée expliquée par les composantes est déterminée. La courbe associée permet d'identifier le nombre de composante nécessaire pour expliquer un seuil donné de variance (généralement le seuil est à 95%).
Ainsi, grâce à l'analyse graphique de l'évolution de la variance expliquée en fonction du nombre de *clusters*, il est possible de déterminer le nombre minimal de dimensions à retenir tout en conservant l'essentiel de l'information des données.
L'analyse de ce graphique se fait dans la même logique que pour la méthode du coude utilisée pour déterminer le nombre de clusters : détermination du point où le coude se forme.
En général, la courbe suit une évolution croissante : plus des composantes principales sont conservées, plus la variance expliquée est importante, ce qui est attendu. Cependant, au-delà d'un certain nombre de composantes, le gain en variance expliquée devient marginal et stagne. C'est pourquoi le nombre optimal de composantes principales correspond au point où la courbe ralentit significativement son augmentation.

Le graphe ci-dessous montre l'évolution de la variance expliquée cumulée en fonction du nombre de composantes principales.

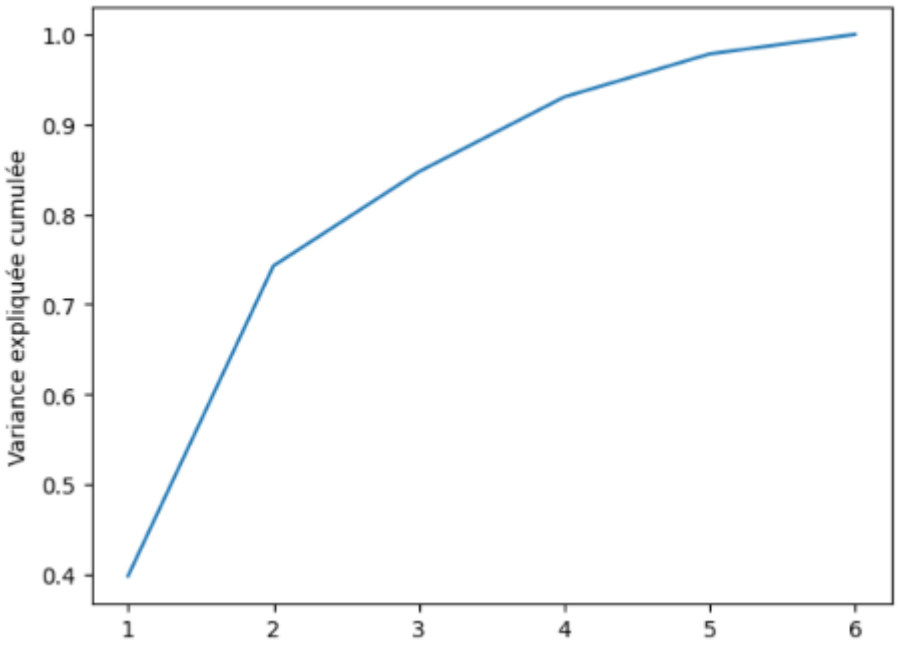


FIGURE 36 – Variance expliquée cumulée en fonction du nombre de composante principale

D'après le graphique, le coude apparaît à partir de la deuxième composante principale,

avec une variance expliquée cumulée de 75%. Ainsi, deux composantes principales sont retenues.

Afin de déterminer les composantes principales, il faut calculer leurs coefficients appelés *loadings* en anglais, qui permettent d'indiquer l'importance de chaque variable initiale dans chaque composante principale.
Sur les données standardiser $X_{\text{standardisé}} = \frac{X-\bar{X}}{\sqrt{v_x}}$ (tel que, X le vecteur des données, $\bar{X}$ étant le vecteur des moyennes et v_x le vecteur des variances), la matrice de covariance Σ est déterminée :

$$\Sigma = \frac{1}{n} X^T_{\text{standardisé}} X_{\text{standardisé}}$$

Ensuite, il faut identifier les vecteurs propres ν_i et les valeurs propres λ_i pour chaque composante principale i :

$$\sum \nu_i = \lambda_i \nu_i$$

Ainsi les valeurs des *loadings* sont obtenues $L_{ij} = \nu_{ij}\sqrt{\lambda_j}$, avec ν_{ij} la contribution de la $i^{\text{ème}}$ variable au $j^{\text{ème}}$ vecteur propre et λ_i la valeur propre associée à la $j^{\text{ème}}$ composante principale.
Plus la valeur du *loading* est élevé, plus la variable contribue à la composante principale.
Ces valeurs sont stockées dans une matrice :

- En ligne : la variable initiale.
- En colonne : les composantes principales, notée $PC1 = \frac{J_2}{J_1}, \ \ldots, \ PC6 = \frac{J_7}{J_6}$.

Le graphe ci-dessous montre les coefficients en fonctions des sept composantes principales $(PC1, \ldots, PC6)$et permet d'identifier quelles variables contribuent le plus aux nouvelles dimensions.

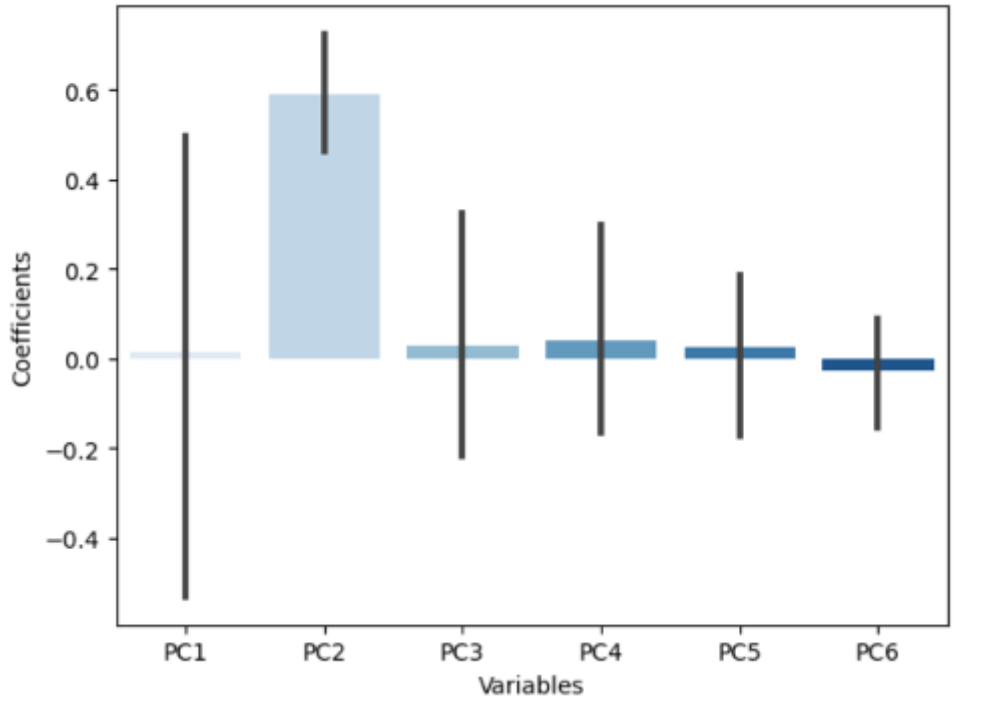


FIGURE 37 – Contribution des variables aux composantes principales

Les barres bleues montrent l'importance des variables dans la construction des composantes principales et les barres noires représentent la variabilité dans les coefficients estimés. Ainsi, les deux composantes principales retenue correspondent à celles ayant le coefficient le plus élevés :

- La composante principale la plus influente est PC2 et qui correspond pratiquement à $\frac{J_3}{J_2}$.
- La deuxième composante la plus influente peut être PC1 : son influence semble non négligeable mais incertain (variabilité expliquée par la barre d'erreur).
- PC4 peut être une bonne option sachant qu'elle est plus stable que PC1.

Une étude doit être faite pour choisir entre PC1 et PC4. En effet, il faut s'assurer si PC1 est vraiment incertaine. Si PC4 explique une part plus importante de variance après PC2 donc le choix doit s'orienter vers PC4 au lieu de PC1. Ainsi, il faut calculer les variances expliquées de PC1 et PC4 :

- Si PC1 explique beaucoup plus de variance que PC4, il vaut mieux garder PC1 malgré son incertitude.
- Si PC4 explique presque autant que PC1 et est plus stable, alors le choix est justifié.

La variance expliquée des composantes est obtenue grâce à la formule suivante :

$$V\left(PC_i\right) = \frac{\lambda_i}{\sum_{j=1}^{6} \lambda_j}$$

Les variances de PC1 et de PC4 valent respectivement 30% et 40%. Ainsi, le choix de PC4 est justifié.

L'objectif étant de réduire la dimensionnalité, les deux variables les plus influentes sur les deux premières composantes principales sont sélectionnées : $PC2 = J2/J1$ et $PC4 = J5/J4$.
Le graphique ci-dessous représente le regroupement des évènements de 2021 en trois clusters obtenus via le *K-Means* en 2 dimensions selon les deux composantes principales retenues.

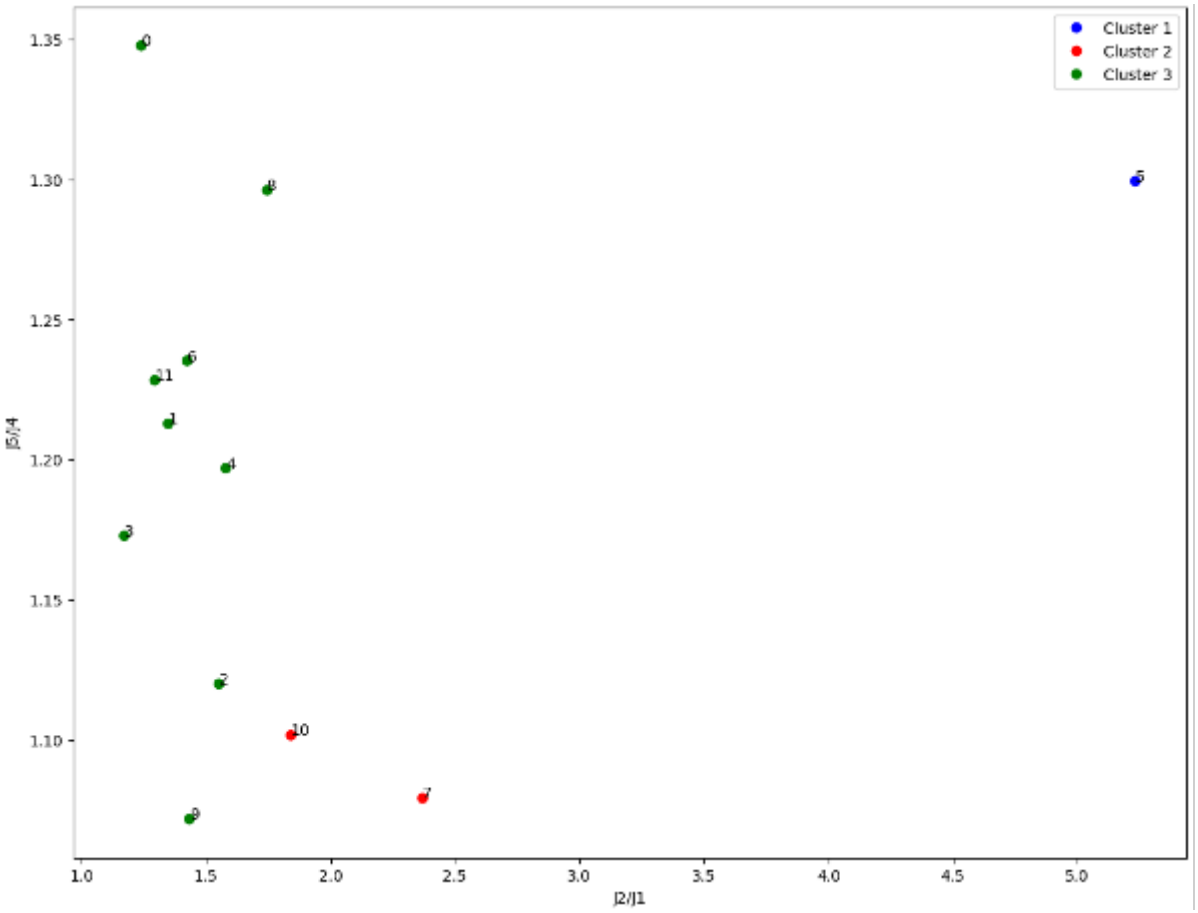


FIGURE 38 – Représentation des *clusters* en 2-dimensions

Les couleurs indiquent l'appartenance à un *cluster* et les évènements sont numérotés, pour un intérêt de lisibilité. Ce graphique montre bien trois *clusters* bien distincts :

- Le cluster 3 est le cluster principal : la plupart des évènements appartiennent à ce *cluster*. Ces évènements sont donc similaires en termes de développement quotidien du nombre de sinistre.
- Le *cluster* 1, quant à lui, est formé que d'un seul évènement « 2021-f. GRELE. 23 juil. 24 juil. ». Cela peut indiquer un point atypique et donc un groupe très différent des autres. En effet, c'est le seul évènement grêle dans la base de 2021.

- Le *cluster* 2 contient quelques points qui semblent être un peu isolé du cluster principal.

Le tableau ci-dessous récapitule les résultats du modèle *K-Means* sur la base des évènements de 2021 :

EVENEMENT	*CLUSTER*
2021-a. TEMP. 13 janv. 15 janv.	2
2021-b. TEMP. 20 janv. 23 janv.	2
2021-c. INOND. 28 janv. 19 fév.	2
2021-d. ORAGE. 02 juin. 04 juin.	2
2021-e. ORAGE. 12 juil. 17 juil.	2
2021-f. GRELE. 23 juil. 24 juil.	0
2021-g. TEMP. 12 août	2
2021-h. TEMP. 21 août	1
2021-i. ORAGE. 08 sept. 09 sept.	2
2021-j. ORAGE. 14 sept. 16 sept.	2
2021-k. ORAGE. 02 oct. 05 oct.	1
2021-l. TEMP. 20 oct. 22 oct. Tempête Aurore	2

TABLE 17 – Les évènements 2021 associés à leur *cluster*

La dernière étape consiste à effectuer les prédictions de cet algorithme sur la base des évènements de 2022. Chaque événement de l'année 2022 est attribué à l'un des *clusters* ainsi constitués, en fonction de la dynamique de développement observée au cours des sept premiers jours. Ainsi en appliquant ce modèle sur la base des évènements de 2022, chaque évènement est associé à un numéro de *cluster*. Le tableau ci-dessous montre les prédictions des *clusters* pour chaque évènement :

EVENEMENT	*CLUSTER*
2022-a. TEMP. 08–11 janv.	1
2022-b. TEMP. 31 janv. 02 fév. CORRIE	1
2022-c. TEMP. 06–07 fév.	1
2022-d. TEMP. 17–21 fév. EUNICE FRANKLIN	2
2022-e. INOND. 12–15 mars	1
2022-f. TEMP. 07–09 avril DIEGO	2
2022-g. ORAGE. 15 mai	1
2022-h. GRELE. 19–23 mai	1
2022-i. GRELE - INOND 02–06 juin	0
2022-j. GRELE. 19–23 juin	1
2022-k. GRELE. 24–25 juin	2
2022-l. GRELE. 26–30 juin	1
2022-m. GRELE. 03–04 juillet	1
2022-n. GRELE. 20 juillet	2
2022-o. ORAGE. 14–18 août	2
2022-p. GRELE. 05–08 septembre	1
2022-q. GRELE. 14 septembre	2
2022-r. TEMP. 23–24 oct. 2022	1

TABLE 18 – Tableau des prédictions du *K-means* pour les évènements de 2022

Pour chaque événement de 2022, le nombre final projeté (NFP) est estimé en appliquant la même méthodologie que celle de la première méthode, mais cette fois en considérant uniquement les événements appartenant au même *cluster*. Plutôt que de projeter le nombre de sinistres en se basant sur l'ensemble des données historiques, seuls les événements qui présentent une évolution similaire sont impliqués. À partir du septième jour, le nombre de sinistres cumulés est ainsi projeté en utilisant les tendances des événements antérieurs de son *cluster*. Progressivement, les projections des événements suivants intègrent celles des événements déjà estimés, à condition qu'ils appartiennent au même regroupement. Comme dans la première méthode agrégée, le NFP final est déterminé à partir du moment où le nombre projeté se stabilise.

Cette classification permet d'affiner l'estimation du nombre final de sinistres en tenant compte des similitudes entre les évènements et permet d'obtenir des estimations plus précises et mieux adaptées aux spécificités de chaque type d'événement.

En parallèle de l'amélioration apportée à l'estimation du NFP des évènements, une

seconde optimisation est introduite concernant l'estimation du coût moyen des sinistres. Comme expliqué préalablement, l'analyse des données a révélé que le CM tend à diminuer au fil du temps, ce qui signifie que considérer le CM à J+7, comme dans la première méthode, ne reflète pas nécessairement la réalité finale des coûts. Afin d'intégrer cette évolution, cette nouvelle approche propose de prendre en compte le CM ultime, qui correspond à la valeur stabilisée après une période suffisamment longue. L'objectif est d'obtenir une estimation plus proche de la réalité en tenant compte de la tendance à la baisse du coût moyen.

Pour cela, l'évolution journalière du CM est calculée et les CM à J+7 des évènements de 2022 calculé à la première méthode sont projetés selon cette évolution. Les CM ultime sont déterminés au moment où les valeurs se stabilisent.

Enfin, une dernière amélioration concerne l'estimation du NBCSS de sinistres. Contrairement à la première méthode, où un taux de sinistres clos sans suite unique était appliqué à tous les événements sans distinction, cette approche vise à affiner l'estimation en différenciant les taux en fonction du type d'événement. Ainsi, au lieu d'un taux global, trois taux distincts sont calculés : un pour les événements de grêle, un autre pour les tempêtes, et un dernier pour les inondations et orages.
Cette segmentation repose sur l'analyse des sinistres survenus en 2021. Pour chaque événement de 2021, le taux de sinistres clos sans suite est déterminé en calculant le rapport entre le nombre de sinistres clos sans suite ultime et le nombre de sinistres total à l'ultime (à partir de la base de données à disposition).

	NBCSS ultime	NB Ultime	Taux clos sans suite
2021-a. TEMP. 13 janv. 15 janv.	49,89	217,45	23%
2021-b. TEMP. 20 janv. 23 janv.	178,9	768,39	23%
2021-c. INOND. 28 janv. 19 fév.	524,03	1652,88	32%
2021-d. ORAGE. 02 juin. 04 juin.	140,58	394,6	36%
2021-e. ORAGE. 12 juil. 17 juil.	293,46	750,16	39%
2021-f. GRELE. 23 juil. 24 juil.	99,67	587,35	17%
2021-g. TEMP. 12 aout	79,23	413,58	19%
2021-h. TEMP. 21 aout	70,56	378,8	19%
2021-i. ORAGE. 08 sept. 09 sept.	97,45	260,79	37%
2021-j. ORAGE. 14 sept. 16 sept.	388,48	1119,32	35%
2021-k. ORAGE. 02 oct. 05 oct.	350,03	918,63	38%

FIGURE 39 – Taux de sinistres clos sans suite par évènement observé sur l'année 2021

Ensuite, ces valeurs sont moyennées pour chaque catégorie d'événement afin d'obtenir un taux de référence propre à chaque catégorie. Voici le résultat obtenu :

Evenement	Taux clos sans suite
TEMP	22%
INNOND/ORAGE	36%
GREL	17%

FIGURE 40 – Taux de sinistres clos sans suite par type d'évènement sur l'année 2021

Ces taux sont ensuite appliqués aux nombres de sinistres pour déterminer le nombre de sinistres clos sans suite pour chaque évènement de 2022.

Le tableau ci-dessous présente les différentes valeurs de taux de sinistres clos pour chaque évènement de 2022 selon la méthode utilisée. Les écarts entre le taux de sinistre clos sans suite observé et celui estimé selon les deux méthodes sont aussi présentés.

Evenement	Tx de css observé	Tx de css estimé (méthode classique)	Tx de css estimé (Méthode améliorée)	Ecart (Méthode classique)	Ecart (Méthode améliorée)
2022-a. TEMP. 08-11 janv.	31%	28%	22%	6%	-9%
2022-b. TEMP. 31 janv. 02 fév. CORRIE	23%	28%	22%	6%	-1%
2022-c. TEMP. 06-07 fév.	21%	28%	22%	6%	0%
2022-d. TEMP. 17-21 fév. EUNICE FRANKLIN	18%	28%	22%	6%	4%
2022-e. INOND. 12-15 mars .	29%	28%	36%	-8%	7%
2022-f. TEMP. 07-09 avril DIEGO	21%	28%	22%	6%	0%
2022-g. ORAGE. 15 mai	24%	28%	17%	11%	-7%
2022-h. GRELE. 19-23 mai	16%	28%	17%	11%	1%
2022-i. GRELE - INOND 02-06 juin	15%	28%	17%	11%	2%
2022-j. GRELE. 19-23 juin	15%	28%	17%	11%	2%
2022-k. GRELE. 24-25 juin	27%	28%	17%	11%	-10%
2022-l. GRELE. 26-30 juin	17%	28%	17%	11%	-1%
2022-m. GRELE. 03-04 juillet	12%	28%	17%	11%	5%
2022-n. GRELE. 20 juillet	15%	28%	17%	11%	2%
2022-o. ORAGE. 14-18 aout	21%	28%	17%	11%	-4%
2022-p. GRELE. 05-08 septembre	23%	28%	17%	11%	-6%
2022-q. GRELE. 14 septembre	15%	28%	17%	11%	2%
2022-r. TEMP. 23-24 oct. 2022	20%	28%	22%	6%	2%
TOTAL			19%	8%	-1%

FIGURE 41 – Comparaison des méthodes de détermination du taux de sinistre clos sans suite

L'application de ces nouveaux taux aux événements de 2022 permet d'obtenir une estimation plus précise du nombre de sinistres clos sans suite. En effet, en comparant les taux observés aux taux estimés, une erreur moyenne de seulement -1% est constatée, ce qui est largement acceptable. En revanche, lorsque la méthode classique est appliquée (basée sur la régression linéaire des taux de CSS des dix dernières années), l'écart moyen avec la réalité est de 8%. Cet écart significatif prouve l'apport de cette nouvelle approche qui permet d'améliorer la qualité des prévisions.

Comme pour la première méthode, l'ensemble des résultats est récapitulé dans un tableau comparatif, où sont recensées à la fois les valeurs observées et les estimations pour le nombre de sinistres, le nombre de sinistres clos sans suite, la charge totale et le coût

moyen. À partir de ces informations, la charge ultime est calculée, ainsi que le *RMSE* global.

EVENEMENT	NB_estimé	Nb_observé	CM projeté à l'ultime	NBCSS_estimé	NBCSS_observé	CHG_estimée	CHG_observée	MSE
2022-a. TEMP. 08-11 janv.	354	420	3 557	77	129	987 347	987 143	41 913
2022-b. TEMP. 31 janv. 02 fév. CORRIE	651	654	2 468	141	149	1 259 507	1 459 104	39 839 344 745
2022-c. TEMP. 06-07 fév.	221	198	3 790	48	42	655 643	560 083	9 131 729 924
2022-d. TEMP. 17-21 fév. EUNICE FRANKLIN	8 576	8 011	2 725	1 856	1 426	18 310 602	18 653 778	117 769 529 740
2022-e. INOND. 12-15 mars .	413	437	2 472	149	126	652 654	844 226	36 699 976 678
2022-f. TEMP. 07-09 avril DIEGO	1 236	1 208	2 473	268	255	2 395 336	2 613 958	47 795 608 862
2022-g. ORAGE. 15 mai	160	179	4 287	27	43	568 910	500 674	4 656 141 721
2022-h. GRELE. 19-23 mai	1 956	1 821	6 334	332	293	10 288 672	9 258 051	1 062 178 633 467
2022-i. GRELE - INOND 02-06 juin	5 942	5 758	11 027	1 008	875	54 402 271	54 177 113	50 695 913 660
2022-j. GRELE. 19-23 juin	11 307	9 407	11 750	1 919	1 424	110 315 068	89 016 487	453 629 526 850 237
2022-k. GRELE. 24-25 juin	381	308	4 253	65	84	1 344 918	1 091 280	64 332 162 404
2022-l. GRELE. 26-30 juin	2 065	1 798	4 208	350	314	7 213 029	6 610 470	363 077 744 189
2022-m. GRELE. 03-04 juillet	546	517	5 578	93	63	2 531 161	2 585 674	2 971 657 895
2022-n. GRELE. 20 juillet	843	676	13 475	143	100	9 433 584	6 541 342	8 365 060 976 075
2022-o. ORAGE. 14-18 aout	3 275	3 394	5 145	556	709	13 988 986	13 697 484	84 973 350 858
2022-p. GRELE. 05-08 septembre	1 147	1 139	4 154	195	266	3 957 773	3 661 230	87 937 826 695
2022-q. GRELE. 14 septembre	328	328	4 342	56	49	1 180 974	1 251 429	4 963 813 111
2022-r. TEMP. 23-24 oct. 2022	538	468	12 292	116	92	5 183 703	3 498 472	2 840 002 295 999
TOTAL	39 940	36 719		7 398	6 441	244 670 137	217 007 999	5 092 541

FIGURE 42 – Résultats de la méthode agrégée améliorée

Les résultats finaux montrent bien une amélioration significative des performances du modèle. En effet, cette méthode améliorée prédit 244 670 137€ de charge, comparé à 265 817 051€ estimé grâce à la méthode classique. Cependant, cette méthode sur estime encore la charge ultime réel (217 007 999€).
Le *RMSE* est de 5 092 541 (comparé à 8 576 643 pour la méthode classique), marquant une réduction de l'erreur par rapport au modèle agrégé classique.
Cette diminution du *RMSE* confirme que le rapprochement des évènements selon leur développement du nombre de sinistre quotidien, ainsi que l'intégration d'un coût moyen à l'ultime et d'un taux de clos sans suite personnalisé pour chaque type d'évènement, permettent d'améliorer la précision des estimations.

III.2 Méthodes individuelles

Le développement d'un modèle basé sur des méthodes d'apprentissage statistique nécessite une base de données structurée, contenant à la fois la variable cible et les variables explicatives. Une fois entraîné, le modèle peut prédire la variable cible pour de nouvelles observations, en se basant uniquement sur les variables explicatives. Ainsi, la préparation des données historiques est une étape essentielle dans la mise en place du modèle individuel.

III.2.A Présentation de la base de données

Cette sous-partie présente la base de données, depuis sa création jusqu'aux traitements appliqués pour corriger les anomalies et enrichir les variables exogènes.

La base utilisée pour le modèle agrégé est issue de la base individuelle agrégée. Contrairement à la modélisation individuelle, le modèle agrégé ne nécessite pas de variables exogènes telles que les caractéristiques des sinistres. Cette section détaille donc la structure et le

contenu de la base de données principale.

Pour la simplicité du modèle agrégé, seules les deux années 2021 et 2022 ont été impliqué dans le modèle. Cependant, pour la modélisation individuelle, les évènements de 2014 à 2022 sont considérés dans le modèle. En effet, pour les modèles d'apprentissage statistique, il est important d'avoir un recul dans le temps pour avoir de bonnes prédictions.

L'extraction des sinistres climatiques survenus entre 2014 et 2022 a été réalisée sous SAS, à une vision de décembre 2023, garantissant ainsi que la charge ultime soit déjà atteinte. Ce processus d'extraction est exactement pareil que celui de la base agrégée, à la différence qu'aucune agrégation n'a été effectuée. Ainsi, la fusion des bases de contrats et de sinistres donne une base tel que chaque ligne représente un sinistre et les informations associées : c'est une base ligne à ligne.

Lors du travail d'extraction de données, le but est de récupérer le maximum d'information concernant les sinistres et les contrats associés afin que les modèles d'apprentissage statistique puissent estimer au mieux la charge ultime. Les bases de données utilisées par l'équipe ne contiennent pas systématiquement les caractéristiques des contrats. En effet, ces informations sont plutôt présentes dans les bases de la direction de l'offre en charge de la tarification. Ainsi, elles ont été intégrées à la base de données. Le tableau ci-dessous recense les variables constituant la base de données individuelles :

Nom de la variable	Description	Type/valeurs/modalités de la variable
NMSIN	Numéro du sinistre	Variable numérique
UP	Unité de prestation	Variable catégorielle : → GEL : gel → GRELE : grêle → INOND : inondation → NATUR : naturelle → TEMP : tempête
DTOUVUP	Date d'ouverture de l'UP	Variable de type date
ETATUP	État de l'UP	Variable catégorielle : → 0 : sinistre en cours → 1 : sinistre clos sans suite → 3 : sinistre clos
DTSURV	Date de survenance du sinistre	Variable de type date
Surv	Année de survenance du sinistre	Variable numérique : → De 2014 à 2022
Chg	Charge d/d ou observé	Variable numérique : **En K euro**
Dep	Département	Variable catégorielle de tous les départements de la France

Nom de la variable	Description	Type/valeurs/modalités de la variable
Réseau	Réseau de distribution	Variable catégorielle : → AGTSA : Agents → SALSA : Salarié → Courtiers
EVENEMENT	Nom de l'évènement	Variable caractère
NBPIECS	Nombre de pièce du bien	Variable numérique : → De 0 à 46
CDRESID	Type de résidence	Variable catégorielle : → PLO : Principale / Secondaire → O : bien occupé → U : bien inoccupé (PNO)
AGE	Ancienneté du bien	Variable catégorielle : → 0 → 1 → 2 → 3
OPT_AMG_INST	Option couverture installations extérieures	Variables indicatrices : → 1 : oui → 0 : non
OPT_BDG_INT2	Option couverture casse intérieure	Variable catégorielle : → 0 → 1 → 2 → 3
NMSIN	Numéro du sinistre	Variable numérique
UP	Unité de prestation	Variable catégorielle : → GEL : gel → GRELE : grêle → INOND : inondation → NATUR : naturelle → TEMP : tempête
DTOUVUP	Date d'ouverture de l'UP	Variable de type date
ETATUP	État de l'UP	Variable catégorielle : → 0 : sinistre en cours → 1 : sinistre clos sans suite → 3 : sinistre clos
DTSURV	Date de survenance du sinistre	Variable de type date
Surv	Année de survenance du sinistre	Variable numérique : → De 2014 à 2022
Chg	Charge d/d ou observé	Variable numérique : **En K euro**
Dep	Département	Variable catégorielle de tous les départements de la France

Nom de la variable	Description	Type/valeurs/modalités de la variable
Reseau	Réseau de distribution	Variable catégorielle : → AGTSA : Agents → SALSA : Salarié → Courtiers
EVENEMENT	Nom de l'évènement	Variable caractère
NBPIECS	Nombre de pièce du bien	Variable numérique : → De 0 à 46
CDRESID	Type de résidence	Variable catégorielle : → PLO : Principale / Secondaire → O : bien occupé → U : bien inoccupé (PNO)
AGE	Ancienneté du bien	Variable catégorielle : → 0 → 1 → 2 → 3
OPT_AMG_INST	Option couverture installations extérieures	Variables indicatrices : → 1 : oui → 0 : non
OPT_BDG_INT2	Option couverture casse intérieure	Variable catégorielle : → 0 → 1 → 2 → 3
OPT_BDG_TL3	Option couverture casse des appareils nomades	Variables indicatrices : → 1 : oui → 0 : non
OPT_AMG_PISC	Option couverture des piscines, spa et jacuzzi	Variables indicatrices : → 1 : oui → 0 : non
OPT_BS_MAT_PRO_1	Option remboursement matériel professionnel 5 000€	Variable numérique en euro
OPT_BS_MAT_PRO_2	Option remboursement matériel professionnel 15 000€	Variable numérique en euro
OPT_VAN	Option valeur de remplacement	Variable indicatrice : → 1 : oui → 0 : non
NBM2DEP	Si elle existe, superficie de la dépendance du bien	Variable numérique en m^2
MTCAPASS	Montant du capital assuré	Variable numérique en euro
Prime_TTC	Prime HT	Variable numérique en euro

Nom de la variable	Description	Type/valeurs/modalités de la variable
ETAGE	Étage du bien	Variable catégorielle : → D : grande demeure → I : intermédiaire → M : maison → R : rez-de-chaussée
DEDUCTIBLE_TYPE	Type de franchise	Variable catégorielle : → FRANCHISE150 → MAJORATION → RACHAT
ton_etudiant	Si l'assuré est étudiant ou pas	Variable indicatrice : → 1 : oui → 0 : non
LossCauseDetail	Description du sinistre	Variable caractère
ThirdPartyInvolvedFlagText	Si une tierce personne est impactée par le sinistre	Variable catégorielle : → Oui → Non
HowReported	Moyen de déclaration du sinistre	Variable catégorielle : → Assignation → Bordereau courtier → Courrier → EDI → Courtier → Email → Internet → Mise en Cause → Téléphone → Visite
ExpertMissionFlagText	Expert missionné ou non	Variable catégorielle : → Oui → Non
Description	Description du sinistre	Variable caractère
RES_OUV	Réserve à l'ouverture	Variable numérique
TX_OBJ_VALEUR	Taux d'objet de valeur	Variable numérique
DTCLOT	Date de clôture du sinistre	Variable de type date

TABLE 19 – Liste des variables de la base de données individuelles

III.2.B Création de nouvelles variables

La préparation de la base de données implique la création de nouvelles variables à partir des informations existantes afin d'améliorer la modélisation. Cette transformation permet de simplifier certaines variables et d'enrichir la base en structurant davantage l'information.

« CDHABIT » et « CDQUALP » sont deux variables catégorielles qui indiquent respectivement le type du bien et la qualité de l'occupant. Cependant ces deux variables présentent plusieurs modalités qui doivent être regroupées pour une meilleure modélisation.
Ainsi, à partir de la variable « CDHABIT », la variable « Appart_Maison » a été créée, distinguant les appartements des maisons. De plus, la modalité « G.Demeure » a été rajoutée à la variable « Appart_Maison » , qui fait référence aux grandes demeures, afin de mieux caractériser certains types de logements.

La variable « CDQUALP » a été utilisée pour générer la variable « Propriétaire_Locataire » permettant d'identifier la qualité de l'occupant du bien.

De la même manière et afin de simplifier la modélisation, le regroupement des modalités a été réalisé pour d'autres variables. Par exemple, les différentes formes de déclarations ont été consolidées en grandes catégories :

- « Communication écrite/électronique » : regroupent « Courrier », « Email », « Fax » et « Internet ».
- « Communication orale/physique » regroupe « Téléphone » et « Visite ».
- « Assignation et mise en cause » regroupe « Assignation » et « Mise en Cause ».
- « Courtier » regroupe « EDI Courtier » et « Bordereau courtier ».

La variable « PostalCode » a permis d'extraire les deux premiers caractères et ainsi créer la variable « DEP », représentant le département du bien assuré.

La variable « TOP_REOUV » a été également créée, et indique si un sinistre a été rouvert ou non. Elle est définie par la condition suivante : si la date « DTREOUV » existe, alors « TOP_REOUV » prend la valeur 1, sinon 0.
Par ailleurs, le délai d'ouverture des sinistres a été calculé entre « DTSURV » et « DTOUVUP », permettant de mesurer le temps écoulé entre la survenance et l'ouverture du sinistre.

Une autre transformation importante concerne la création de la variable « TX_OBJ » à partir de « TX_OBJ_VALEUR », qui catégorise la variable continue en trois classes distinctes :

- Si $TX\_OBJ\_VALEUR \leq 0.10$ alors $TX\_OBJ = 0,1$ (peu onéreux)
- Si $0.1 < TX\_OBJ\_VALEUR \leq 0.20$ alors $TX\_OBJ = 0,2$ (coût moyen)
- Si $TX\_OBJ\_VALEUR > 0.20$ alors $TX\_OBJ = 0,3$ (très cher)

Sans tenir compte de la superficie de la dépendance si elle existe, la variable « top_dep » a été créée à partir de « NBM2DEP » pour indiquer simplement la présence ou non d'une dépendance.

III.2.C Enrichissement de la base

La base de données a été enrichie de variables exogènes liées à la météorologie, informations utiles pour l'analyse des sinistres liés aux évènements climatiques. Les deux variables sont :

- « RAFALES » : cette mesure indique la vitesse des rafales de vent permettant de déterminer les dommages causés par les tempêtes.
- « PRECIPITATIONS » : cette mesure représente la quantité de précipitations et influe directement sur les risques d'inondation.

L'intégration de ces deux variables permet donc d'améliorer la précision des modèles de prédiction. En exploitant les données de la base mensuelle Synop de Météo France, chaque station météorologique a été associée à un département, en récupérant deux nouvelles variables. Toutefois, certains départements de la base de données ne possèdent pas de station météorologique associée, ce qui a entraîné des valeurs manquantes lors de la jointure de la base Synop et de la base de données.

Afin d'imputer ces valeurs de manière cohérente, la méthode du *krigeage* a été implémentée. En effet, c'est une méthode d'interpolation spatiale qui permet d'estimer des valeurs manquantes en exploitant la corrélation spatiale des observations disponibles et la structure spatiale des données.

Avant d'appliquer le *krigeage*, un traitement préalable des données départementales a été réalisé pour identifier les observations manquantes et structurer l'information de manière homogène notamment en rajoutant les mesures de latitude et longitude de chaque département. La base de données utilisée pour effectuer le *krigeage* est formée des colonnes :

Variable	Description
DATE_METEO	La date de l'évènement
DEP	Le département touché par l'évènement
Latitude	Latitude du département
Longitude	Longitude du département
PRECIPITATIONS	Mesure de précipitations
RAFALES	Mesure de rafales
PRESENT	Variable indicatrice : → 1 : si les variables PRECIPITATIONS et RAFALES sont présentes et non nulles → 0 : sinon

TABLE 20 – Tableau récapitulatif des variables de la base pour le *krigeage*

Pour chacune des variables (« RAFALES » et « PRECIPITATIONS »), un modèle de krigeage est déterminé et appliqué aux données manquantes pour les prédire.

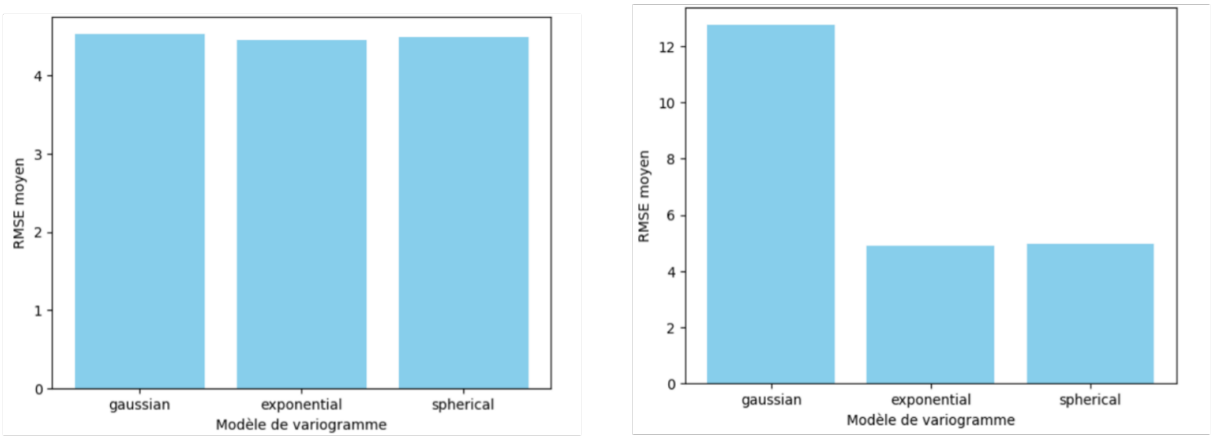
La première étape de la technique de *krigegage* et de déterminer un modèle de variogramme optimal. En effet, ce dernier décrit la variation de la variable en question en fonction de la distance entre les différentes observations. Il en existe plusieurs types et chacun est caractérisé selon la manière dont la variance évolue en fonction de la distance entre les points :

- Sphérique : La variance augmente rapidement avant d'atteindre un seuil, à partir duquel elle devient stable.
- Exponentiel : La variance augmente d'une façon progressive et asymptotique sans atteindre un seuil comme dans le cas sphérique.
- Gaussien : La variance augmente plus lentement que dans le modèle exponentiel, avec une continuité plus lisse entre les points.

Le choix du modèle optimal pour les variables de rafales et de précipitations est réalisé grâce à une validation croisée selon les 3 types de variogramme évoqués plus haut en évaluant la performance de chaque type et en calculant la racine de l'erreur quadratique moyenne. Voici les étapes :

- Division des données : grâce à la variable « PRESENT » créée, séparation des stations avec des valeurs connues de précipitations et rafales.
- Validation croisée : en appliquant la technique de validation croisée ou *K-Fold cross validation* en anglais, entrainement des modèles de krigeage selon les 3 types de variogramme.
- Modèle optimal : pour chaque variable, sélection du modèle optimal selon le critère de RMSE. Ainsi le meilleur modèle est celui qui présente le plus faible *RMSE*.

Pratiquement, voici les résultats des *RMSE* de chaque variogramme :



(a) RMSE moyen pour chaque modèle de variogramme pour la variable « PRECIPITATIONS »

(b) *RMSE* moyen pour chaque modèle de variogramme pour la variable « RAFALES »

FIGURE 43 – *RMSE* moyen pour chaque modèle de variogramme pour les variables « PRECIPITATIONS » et « RAFALES »

Il est clair que le modèle de variogramme qui minimise le *RMSE* est le modèle exponentiel pour la variable « RAFALES ». Cependant, pour la variable « PRECIPITATIONS » les 3 modèles semblent être bon pour la modélisation en termes de minimisation du critère de *RMSE*. Le modèle exponentiel reste celui qui minimise le plus (même si la différence n'est pas énorme) le *RMSE*.

Une fois les modèles optimaux déterminés, le *krigeage* est appliqué aux points où les valeurs de précipitations et rafales sont manquantes. Pour chaque département où la station météorologique n'est pas disponible, le modèle optimisé prédit les valeurs en utilisant les stations voisines.

Pour assurer la cohérence de ces deux variables, les valeurs négatives de précipitations et de rafales par 0.

Enfin, ces deux colonnes sont rajoutées à la base de données pour la suite de la modélisation.

III.2.D Traitement des données incohérentes et manquantes

Le dernier travail sur les données constitue le nettoyage de la donnée des incohérences *inter* et *intra* variable ainsi que l'imputation des données manquantes.

Il consiste tout d'abord à s'assurer que les variables à disposition sont au bon format afin de garantir leur compatibilité avec les analyses. Certaines variables numériques ont été converties en caractère :
« ETATUP », « OPT_BDG_INT2 », « OPT_BDG_TL3 », « OPT_AMG_PISC », « OPT_BS_MAT_PRO_1 », « OPT_BS_MAT_PRO_2 », « OPT_VAN », « OPT_AMG_INST », « top_etudiant », « TOP_REOUV » et « top_dep ».

Les modalités de certaines variables ont également été harmonisées. Par exemple, les modalités « Oui » et « Non » des variables « ThirdPartyInvolvedFlagText » et « ExpertMissionFlagText » ont été remplacées respectivement par 1 et 0, pour homogénéiser les variables indicatrices.

Le traitement des données incohérentes a impliqué plusieurs ajustements : les valeurs négatives des variables « PRECIPITATIONS », « RAFALES », « NBDEP », « TX_OBJ_VALEUR », « chg » (les charges négatives représentent les recours et sont très peu représentes lors de sinistres climatiques) et RES_OUV ont été remplacées par 0.

Un contrôle de cohérence *inter*-variables a été effectué. Par exemple, lorsque « NBPIECS » valait 0 pour des maisons et des grandes demeures, la valeur a été remplacé par la modalité la plus fréquente dans chaque catégorie. De même, la variable « ETAGE » a été ajustée : si le bien était une maison ou une grande demeure, l'étage a été forcé à « M » pour assurer la cohérence avec la typologie du logement.

Enfin, le traitement des données manquantes est réalisé en appliquant différentes méthodes selon le type de variable.

Pour les variables continues comme la prime ou le montant du bien assuré, les valeurs manquantes sont remplacées par la moyenne de l'ensemble de la base. Si la valeur de la variable nombre de mètre au carré de dépendance est vide alors elle vaut 0, c'est-à-dire qu'il n'existe pas de dépendance. L'imputation des valeurs manquantes pour la variable nombre de pièces du bien est faite en fonction du type de bien : en moyenne les maisons présentent 6 pièces et les appartements 3 pièces.

Concernant les variables catégorielles, une approche probabiliste d'imputation a été réalisé sur Python à l'aide de deux fonctions créées selon si la variable est binaire ou non. Cette méthode repose sur l'estimation des proportions des différentes modalités observées dans la base, puis sur l'imputation aléatoire des valeurs manquantes en respectant ces proportions. Ainsi, pour les variables binaires, les valeurs 0 ou 1 sont attribuées en fonction des fréquences observées. Pour les variables non binaires, les probabilités cumulées de chaque modalité sont calculées et une valeur est assignée en fonction d'un tirage aléatoire respectant ces distributions. Cette approche permet de préserver la structure statistique des données imputées et d'éviter des biais liés à une imputation systématique.

Ces différents traitements ont permis de garantir une base de données propre et cohérente afin d'optimiser et de fiabiliser la qualité des analyses et des modèles développés par la suite.

Pour récapituler, voici la liste des variables finalement retenue pour la modélisation. Les variables en gris ne seront pas prises en compte dans la suite de la modélisation, en jaune ce sont les variables modifiées et les variables en vert correspondent aux variables

ajoutées à la base de données.

Nom de la variable	Description	Type/valeurs/modalités de la variable
NMSIN	Numéro du sinistre	Variable numérique
UP	Unité de prestation	Variable catégorielle : → GEL : gel → GRELE : grêle → INOND : inondation → NATUR : naturelle → TEMP : tempête
DTOUVUP	Date d'ouverture de l'UP	Variable de type date
ETATUP	État de l'UP	Variable catégorielle : → 0 : sinistre en cours → 1 : sinistre clos sans suite → 3 : sinistre clos
DTSURV	Date de survenance du sinistre	Variable de type date
Surv	Année de survenance du sinistre	Variable numérique : → De 2014 à 2022
Chg	Charge d/d ou observé	Variable numérique : **En K euro**
Dep	Département	Variable catégorielle de tous les départements de la France
Reseau	Réseau de distribution	Variable catégorielle : → AGTSA : Agents → SALSA : Salarié → Courtiers
EVENEMENT	Nom de l'évènement	Variable caractère
NBPIECS	Nombre de pièce du bien	Variable numérique : → De 0 à 46
CDRESID	Type de résidence	Variable catégorielle : → PLO : Principale / Secondaire → O : bien occupé → U : bien inoccupé (PNO)
AGE	Ancienneté du bien	Variable catégorielle : → 0 → 1 → 2 → 3
OPT_AMG_INST	Option couverture installations extérieures	Variables indicatrices : → 1 : oui → 0 : non

Nom de la variable	Description	Type/valeurs/modalités de la variable
OPT_BDG_INT2	Option couverture casse intérieure	Variable catégorielle : → 0 → 1 → 2 → 3
OPT_BDG_TL3	Option couverture casse des appareils nomades	Variables indicatrices : → 1 : oui → 0 : non
OPT_AMG_PISC	Option couverture des piscines, spa et jacuzzi	Variables indicatrices : → 1 : oui → 0 : non
OPT_BS_MAT_PRO_1	Option remboursement matériel professionnel 5 000€	Variable numérique en euro
OPT_BS_MAT_PRO_2	Option remboursement matériel professionnel 15 000€	Variable numérique en euro
OPT_VAN	Option valeur de remplacement	Variable indicatrice : → 1 : oui → 0 : non
NBM2DEP	Superficie de la dépendance du bien	Variable numérique en m^2
MTCAPASS	Montant du capital assuré	Variable numérique en euro
Prime_TTC	Prime HT	Variable numérique en euro
ETAGE	Étage du bien	Variable catégorielle : → M : pour les maisons → I : étage intermédiaire → R : rez-de-chaussée → D : grandes demeures
DEDUCTIBLE_TYPE	Type de franchise	Variable catégorielle : → FRANCHISE150 → MAJORATION → RACHAT
top_etudiant	Si l'assuré est étudiant ou pas	Variable indicatrice : → 1 : oui → 0 : non
Appart_Maison	Type du bien	Variable catégorielle : → Appart → Maison → Grande demeure
Proprietaire_Locataire	Qualité de l'occupant	Variable catégorielle : → Locataire → Propriétaire
LossCauseDetail	Description du sinistre	Variable caractère

Nom de la variable	Description	Type/valeurs/modalités de la variable
ThirdPartyInvolvedFlagText	Tierce personne impactée	Variable catégorielle : → Oui → Non
HowReported	Moyen de déclaration du sinistre	Variable catégorielle : → Communication écrite/électronique → Communication orale/physique → Assignation et mise en cause → Courtier
ExpertMissionFlagText	Expert missionné ou non	Variable catégorielle : → Oui → Non
Description	Description du sinistre	Variable caractère
TOP_REOUV	Sinistre réouvert ou non	Variable indicatrice : → 1 : oui → 0 : non
RES_OUV	Réserve à l'ouverture	Variable numérique
TX_OBJ	Taux d'objet de valeur	Variable catégorielle : → 0,1 → 0,2 → 0,3
DTCLOT	Date de clôture du sinistre	Variable de type date
top_dep	Présence d'une dépendance	Variable indicatrice : → 1 : oui → 0 : non
TOP_REOUV	Sinistre rouvert ou pas	Variable indicatrice : → 1 : oui → 0 : non
delai_ouverture	Temps entre la survenance et la déclaration	Variable numérique
region	Région de survenance du sinistre	Variable catégorielle
PRECIPITATIONS	Somme des précipitations du lieu sinistré	Variable numérique
RAFALES	Rafales de vent enregistrées en km/h	Variable numérique

TABLE 21 – Liste des variables de la base de données finale

III.2.E Statistiques descriptives et études de corrélation

Une étape importante avant de commencer la modélisation est d'analyser le comportement d'une part de la variable cible et d'autre part les relations entre la variable cible et

les autres variables exogènes. Ainsi, cette sous-partie présente les statistiques descriptives univariées et bivariées.

La variable cible à modéliser est la charge des sinistres des évènements climatiques. La charge de la base de données est en moyenne d'environ 3K €. Selon le graphique de la boite à moustache, la médiane qui est d'environ 1K €. Il existe des sinistres à 0 € (minimum) et des sinistres à 842K € (maximum).

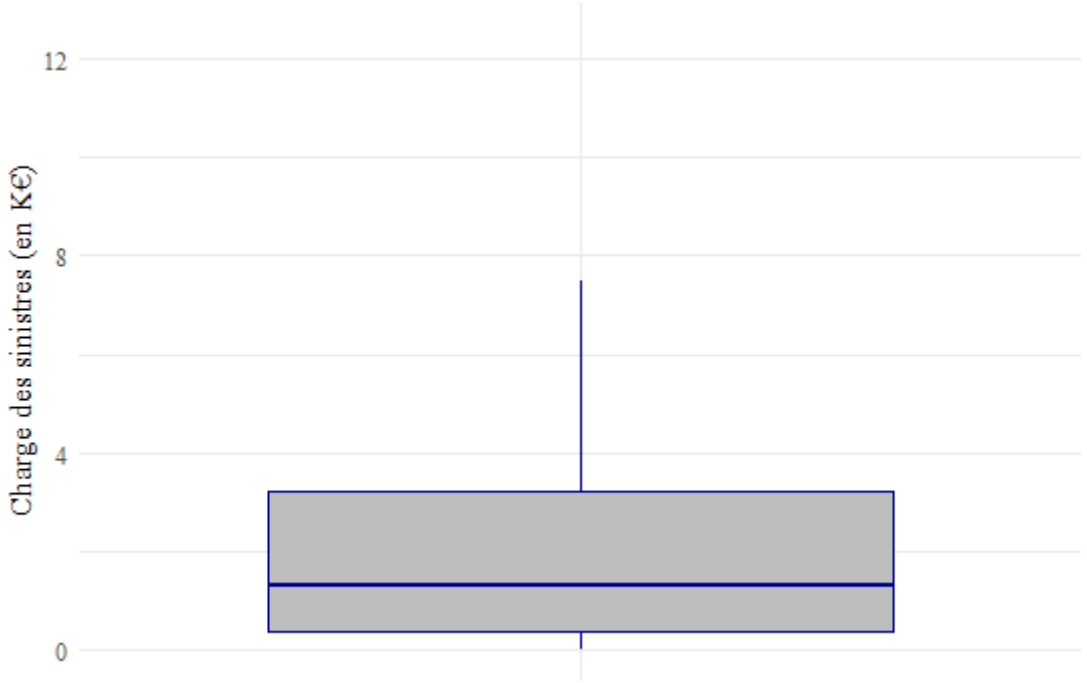


FIGURE 44 – Boîte à moustache de la variable cible charge des sinistres

Dans la suite, il est intéressant de s'attarder aux variables explicatives. Les graphiques ci-dessous représentent les principales conclusions sur ces différentes variables.

La plupart des sinistres correspondent à des tempêtes et en deuxième place se trouve les sinistres de grêles.

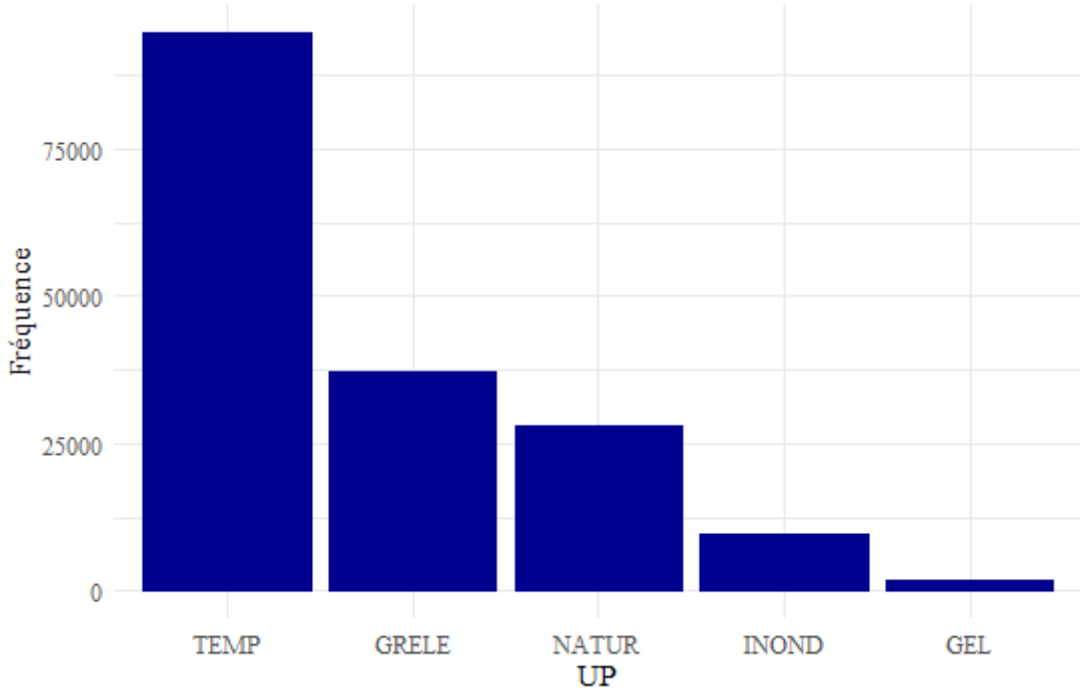


FIGURE 45 – Répartition des sinistres selon l'unité de prestation

Sur les 171 083 sinistres, 86 concernent des étudiants. Ainsi, la majorité du portefeuille n'est pas composé d'étudiants.

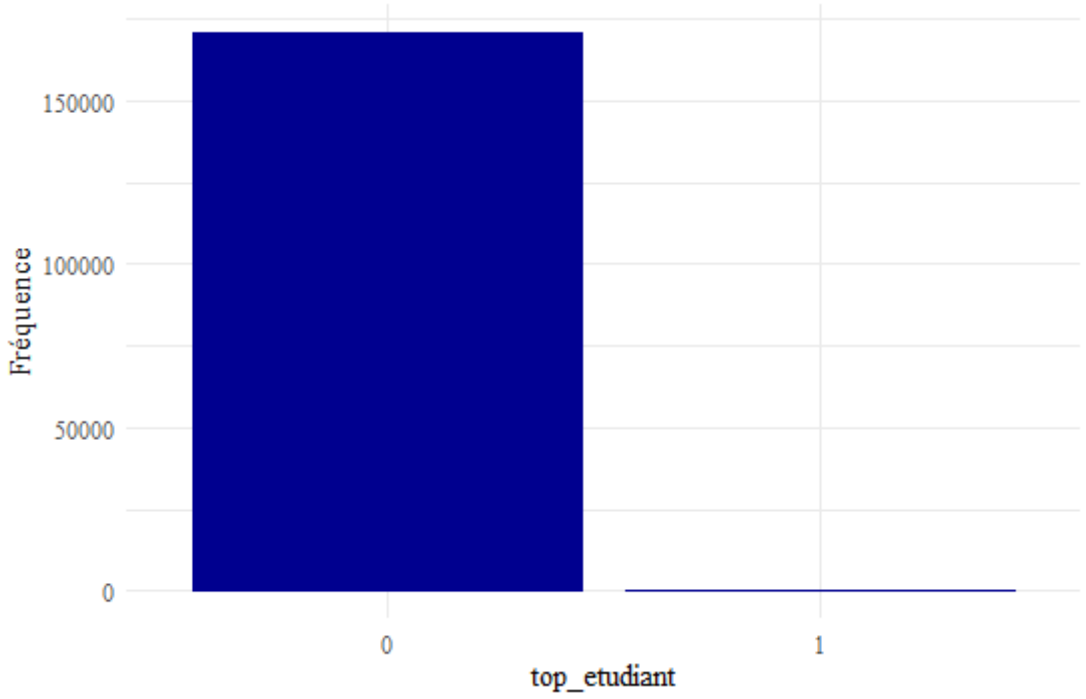


FIGURE 46 – Répartition des sinistres selon le statut d'étudiant de l'assuré

De même, le portefeuille est majoritairement constitué de résidences principales.

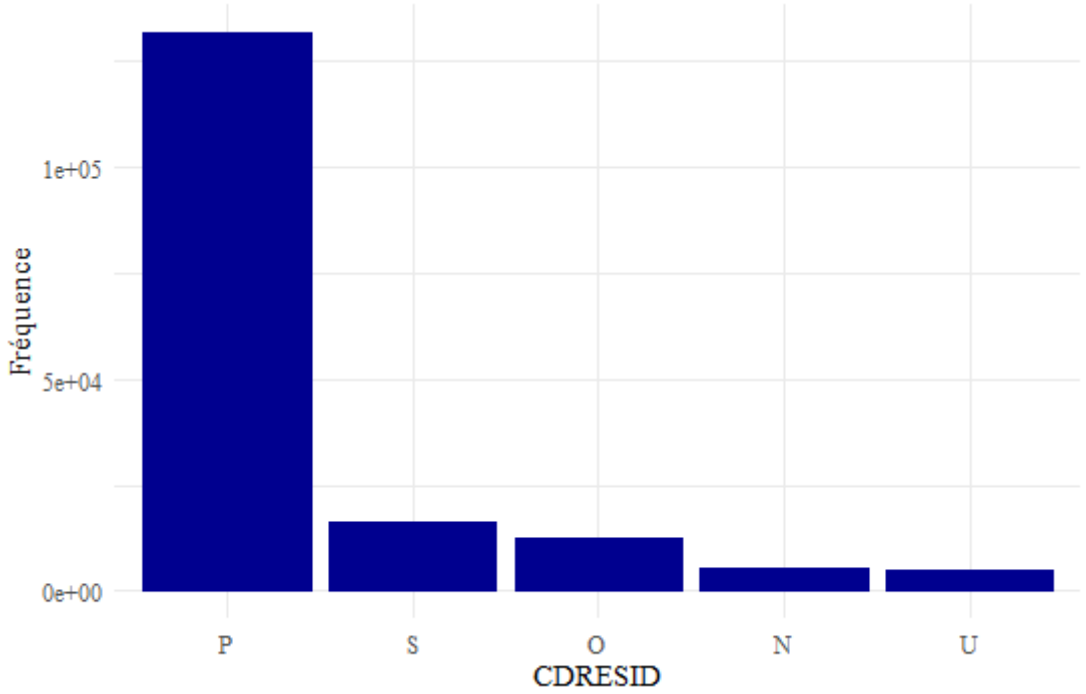


FIGURE 47 – Répartition des sinistres selon le type de résidence

L'histogramme ci-dessous montre la répartition des sinistres selon la qualité de l'occupant :

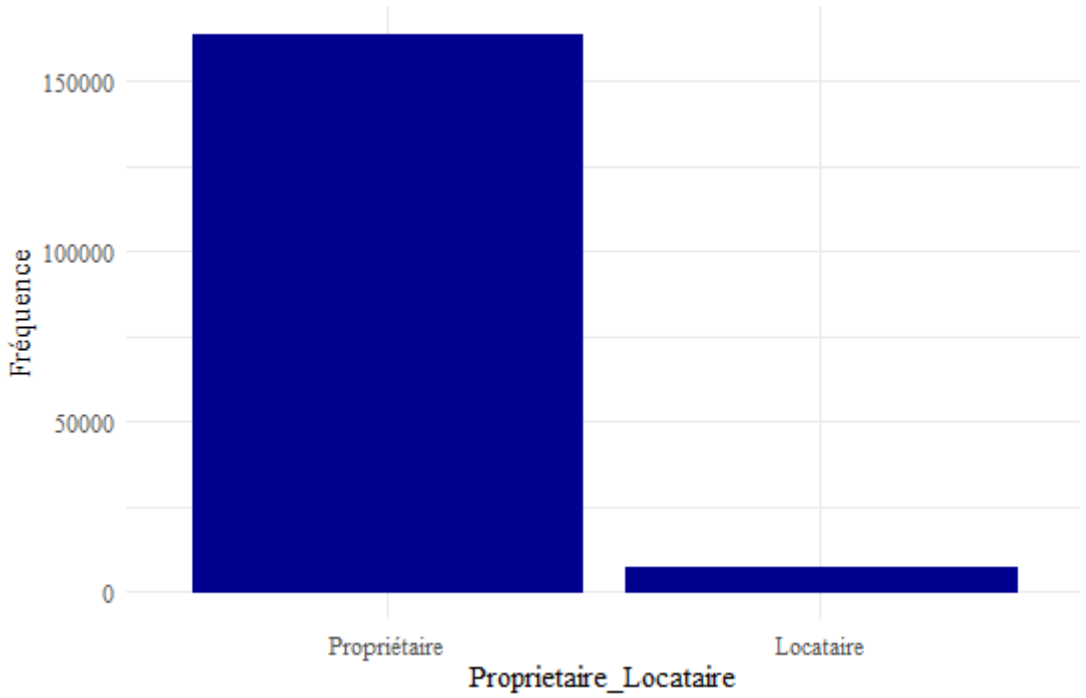


FIGURE 48 – Répartition des sinistres selon la qualité de l'occupant

Les propriétaires sont à l'origine de 95% des sinistres ce qui est expliqué par le fait qu'ils assurent directement leur bien et ont plus de responsabilités liées aux réparations et aux

entretiens.
Les locataires quant à eux sont moins impactés dans notre portefeuille parce que l'assurance du locataire n'est pas systématiquement activée au détriment de l'assurance du propriétaire.

Le graphique ci-dessous montre la répartition des sinistres selon le type du bien assuré :

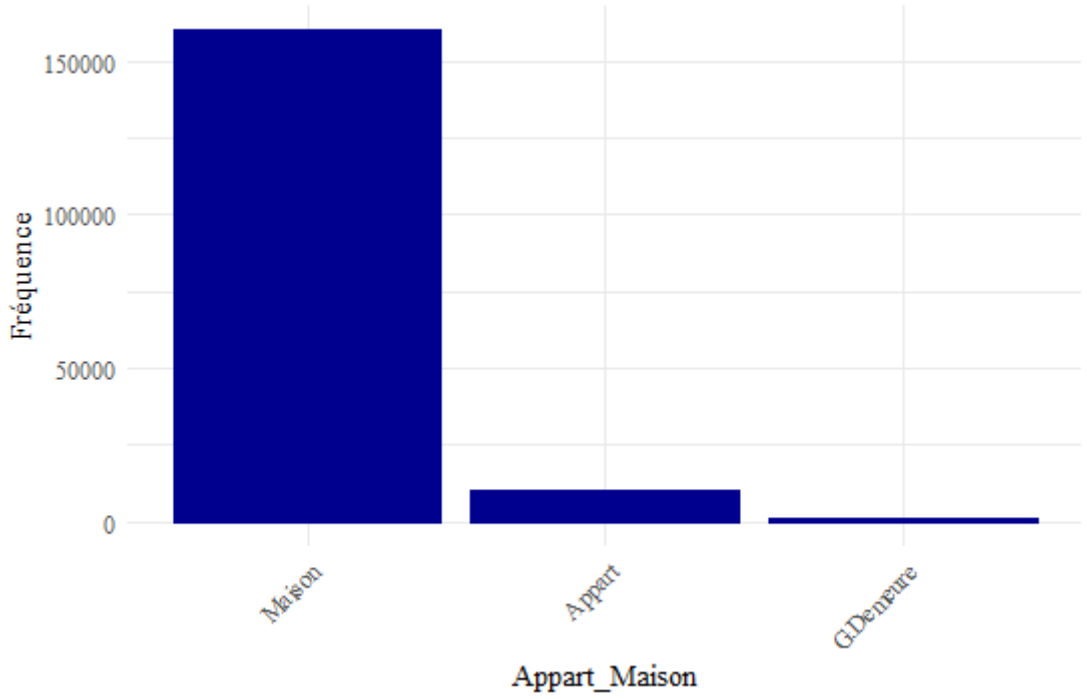


FIGURE 49 – Répartition des sinistres selon le type du bien assuré

Les maisons sont majoritaires et représentent 93,6% des sinistres. Cela peut s'expliquer par une plus grande exposition aux risques climatiques ou un effet de volume si la majorité des assurés possèdent une maison. Cependant, les appartements représentent une faible part des sinistres de 5,89%. Les grandes demeures sont très peu concernées avec 0,50% des sinistres.

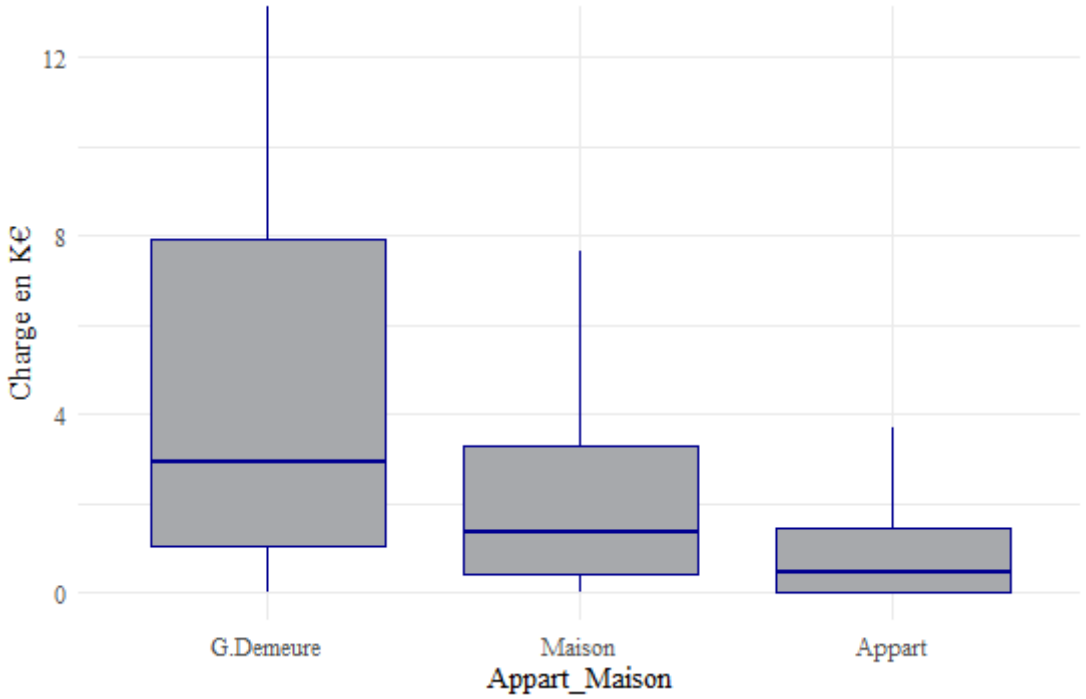


FIGURE 50 – La charge des sinistres en fonction du type du bien assuré

Bien qu'elles soient les moins présentes dans la base, leur cout reste élevé 10K € comparé à 1,5K € pour les appartements et 3,5K € pour les maisons.

Selon le graphique ci-dessous, il s'avère que 90% des sinistres de la base sont des sinistres de propriétaire de maisons.

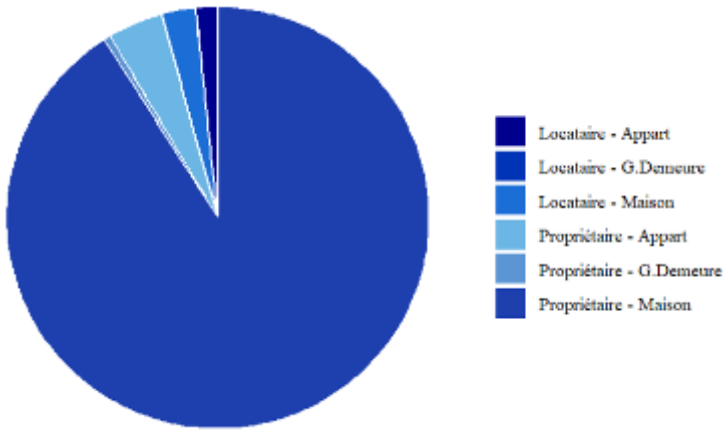


FIGURE 51 – Répartition des sinistres selon le type de logement et la qualité de l'occupant

Les locataires de grande demeure ne sont pas très représentés mais représentent cependant le charge moyenne la plus élevée.

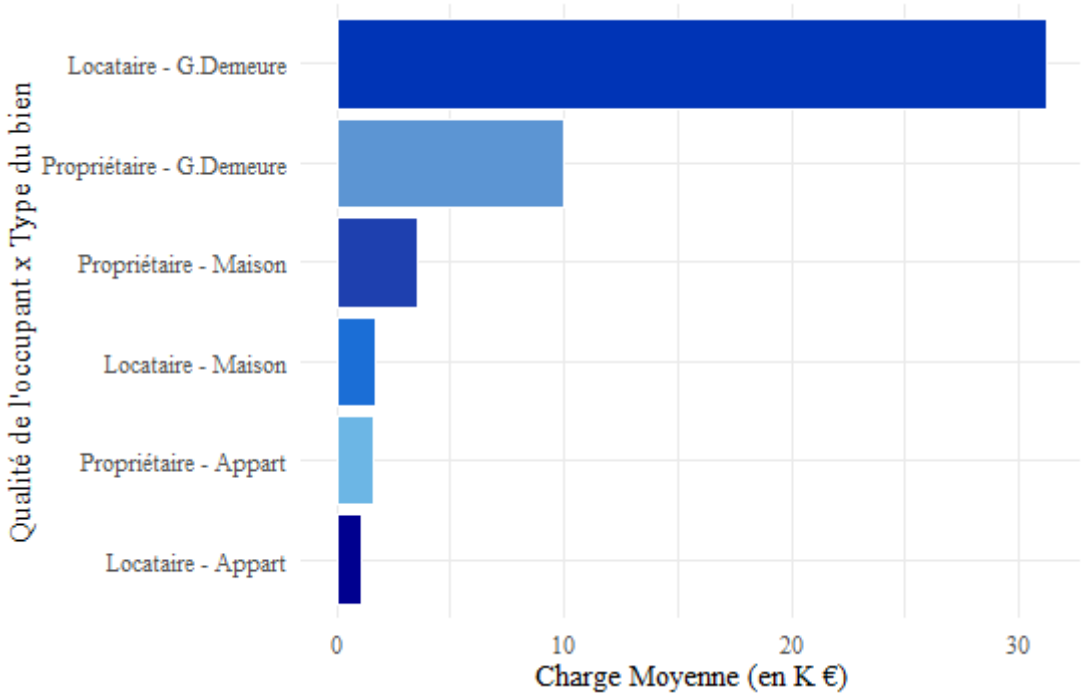


FIGURE 52 – Charge moyenne des sinistres par type de logement et qualité de l'occupant

La manière dont le sinistre est déclaré peut impacter la charge des sinistres. La majorité des sinistres de notre portefeuille sont déclarés via une communication écrite/électronique. En deuxième place, c'est plutôt la communication orale/physique. En dernier on retrouve les autres moyens de déclaration.

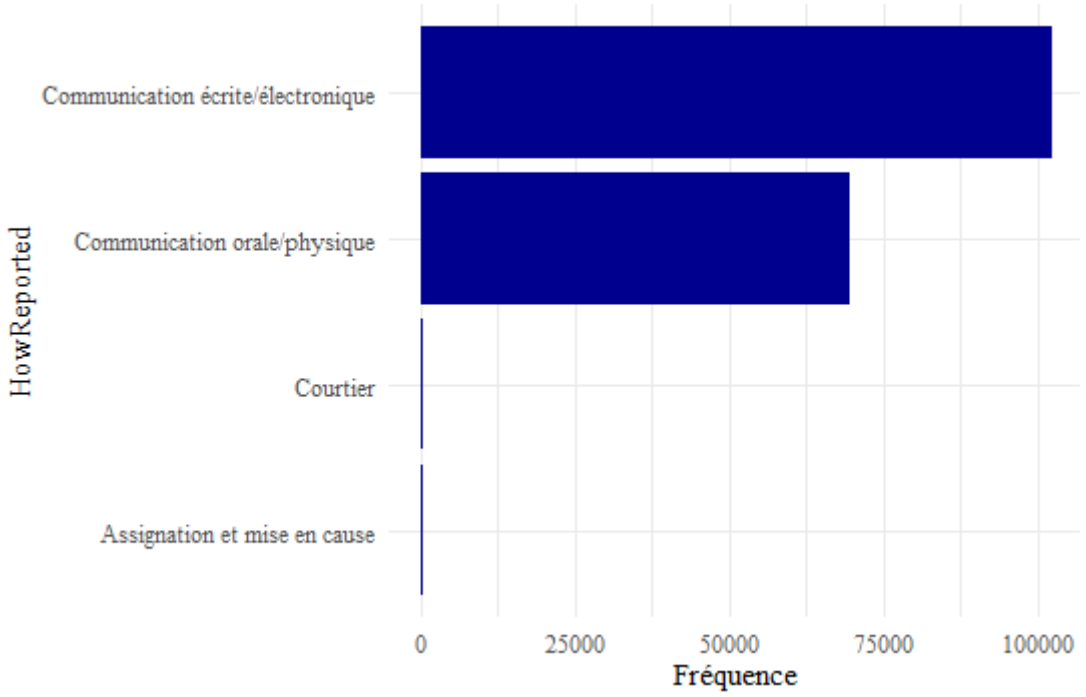


FIGURE 53 – Répartition des sinistres en fonction du moyen de déclaration

Bien que le nombre de sinistres déclarés via les courtiers ou une assignation/mise en cause soit très faible comparé aux autres moyens de déclaration, la charge moyenne de ces deux moyens de déclarations reste légèrement proche des autres.

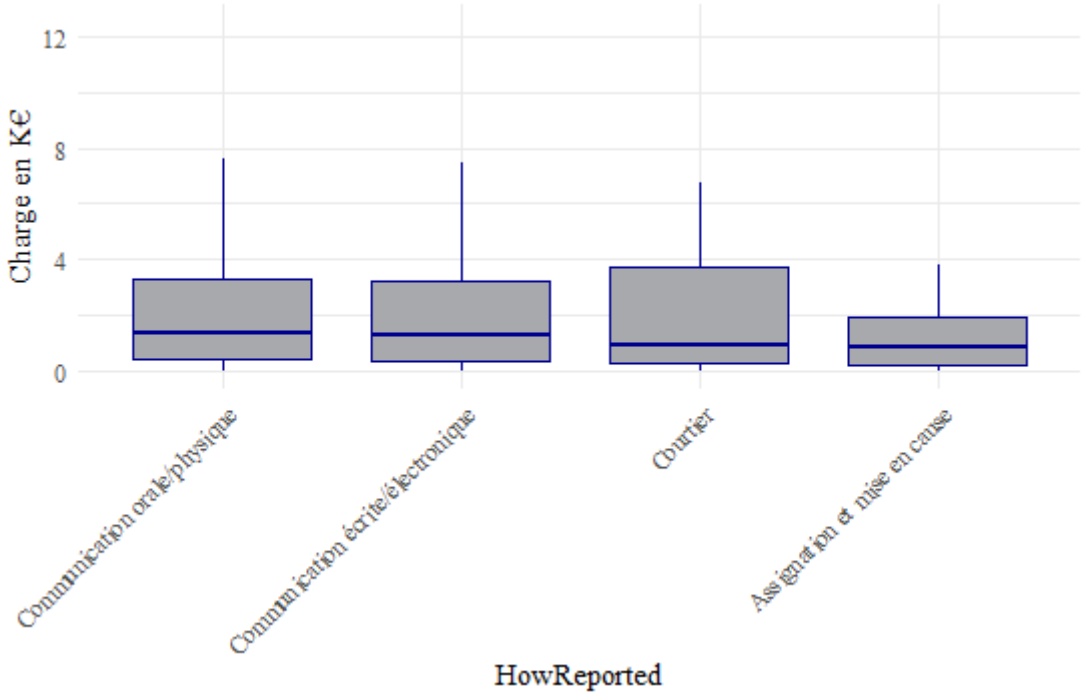


FIGURE 54 – La charge des sinistres en fonction du moyen de déclaration

Un peu plus de la moitié des sinistres ont nécessité l'intervention d'un expert.

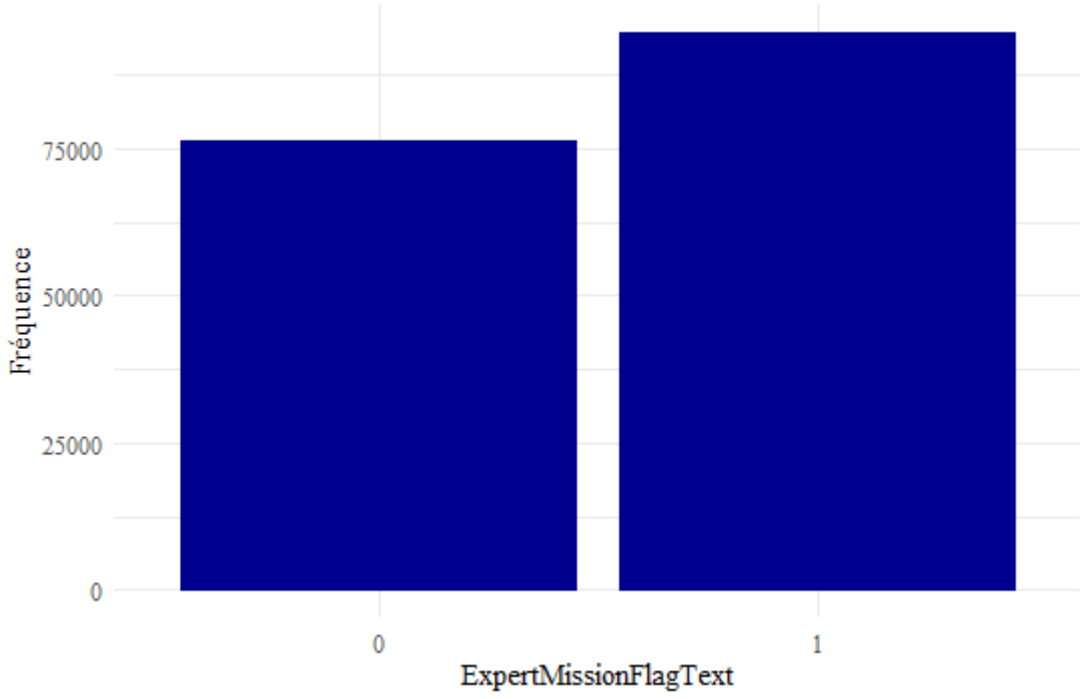


FIGURE 55 – La répartition des sinistres selon l'intervention ou pas d'un expert missionné

Le coût des sinistres nécessitant un expert missionné (5,28K €) est plus élevé que ceux qui ne l'ont pas été (1,14K €). Cette variable serait donc probablement corrélée à la charge des sinistres.

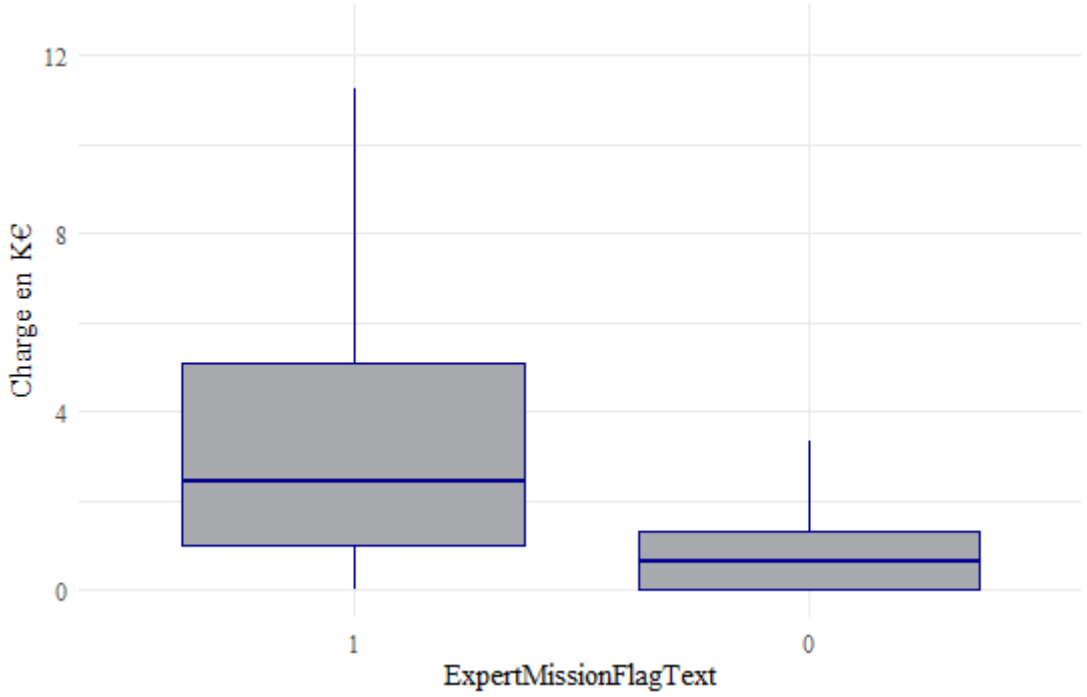


FIGURE 56 – Distribution de la charge de sinistres selon l'intervention ou pas d'un expert missionné

Même si la plupart du portefeuille est constitué de maisons, l'absence de dépendance reste dominante.

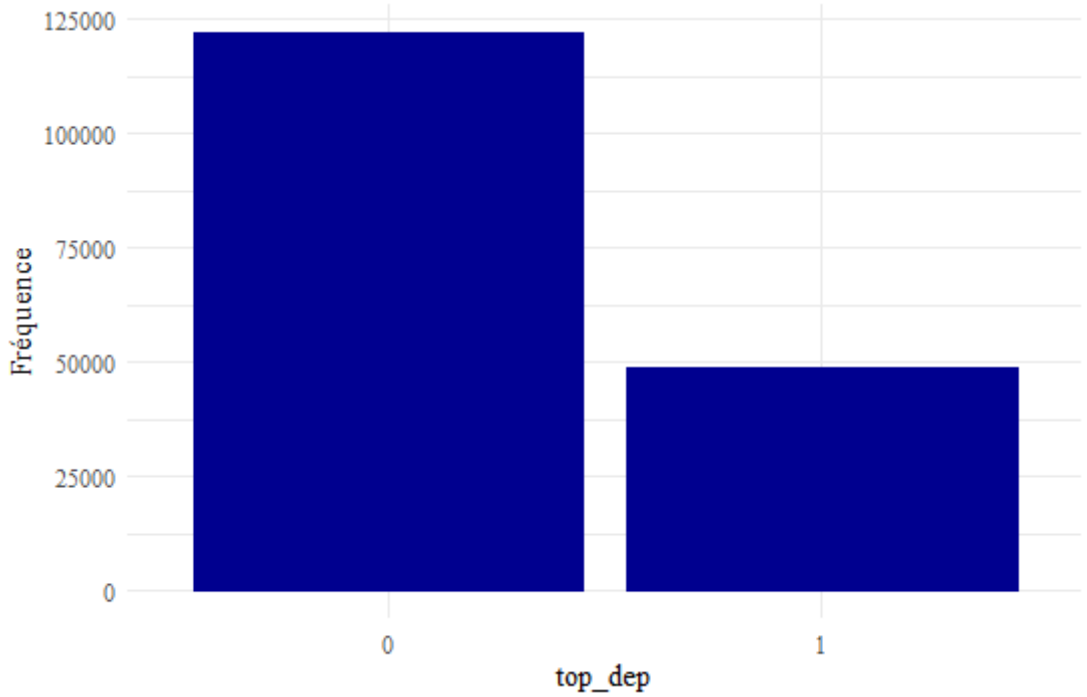


FIGURE 57 – Répartition des sinistres selon l'existence ou pas d'une dépendance au bien assuré

La présence de dépendance entraîne une charge assez élevée (4,61K €) par rapport aux bien sans dépendance (2,96K €). En effet, les biens ayant des dépendances sont moins présents dans la base mais engendrent une charge moyenne assez proche de ceux sans dépendance. Cette variable pourrait donc être candidate dans le modèle de la charge des sinistres.

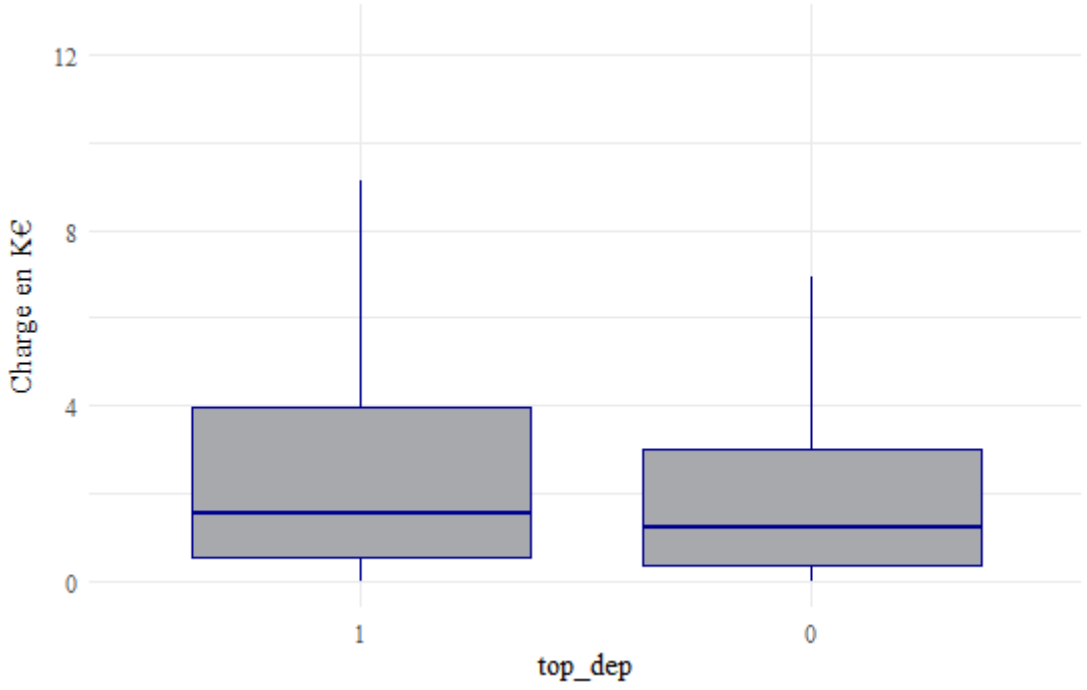


FIGURE 58 – Distribution de la charge de sinistres selon la présence ou pas d'une dépendance au bien

Les régions les plus touchées par les sinistres climatiques dans ce portefeuille sont : la Nouvelle-Aquitaine, les Hautes-de-France, l'Auvergne-Rhône-Alpes, l'Occitanie et l'Île-de-France.

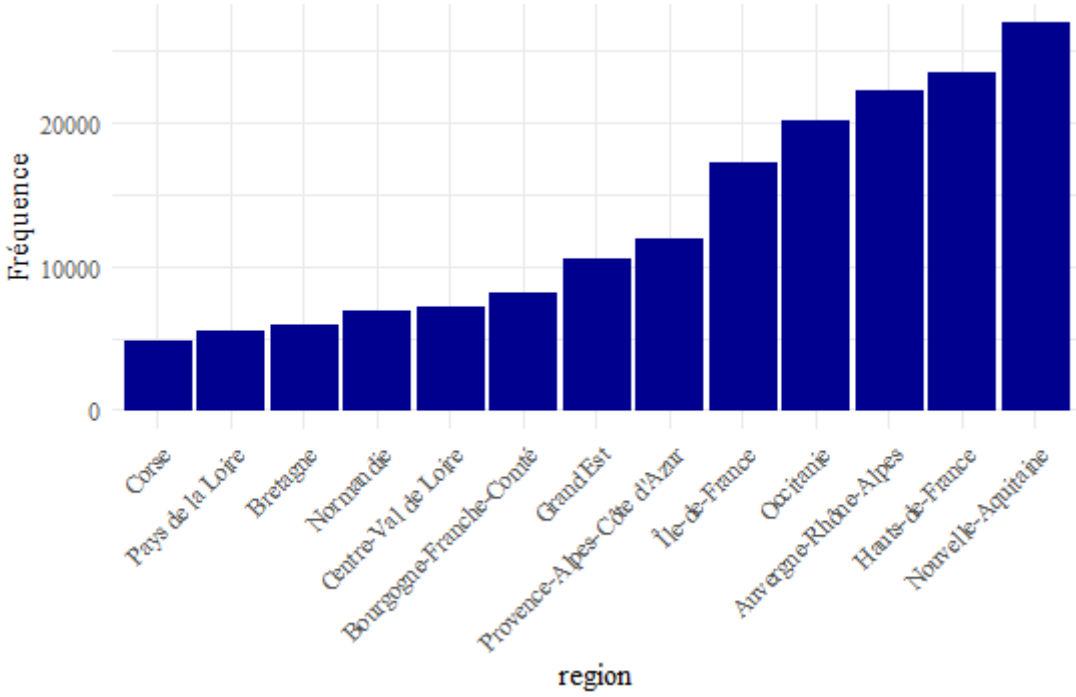


FIGURE 59 – Répartition des sinistres selon la région

L'analyse statistique bivariée permet de trouver une relation entre la charge des sinistres et le reste des variables à disposition. L'analyse de dépendance diffère selon le type des variables. Tout d'abord, dans le cas des variables quantitatives, il est intéressant d'observer graphiquement la dépendance via le nuage de point.

La variable réserve d'ouverture, correspond au montant que les gestionnaires de sinistres définissent selon leur expertise, c'est un montant forfaitaire et qui est réévalué chaque année. Cette variable est logiquement corrélée avec la charge des sinistres.

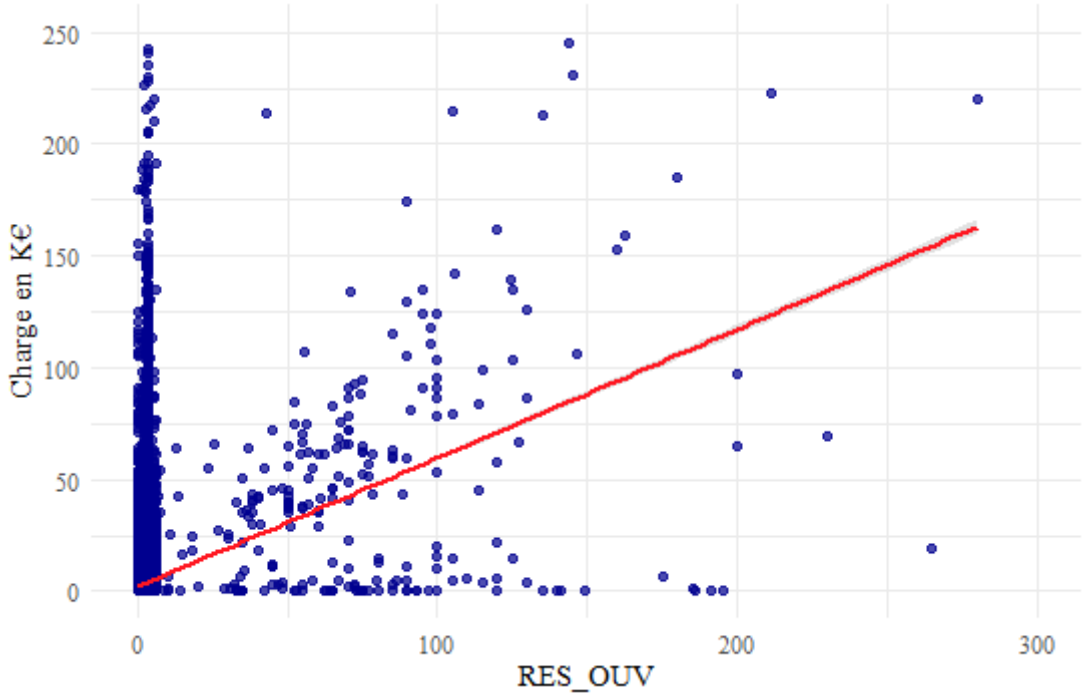


FIGURE 60 – La charge des sinistres en fonction de la réserve d'ouverture

Une forte concentration de points à zéro est observable. En effet, de nombreux sinistres ayant une réserve initialement faible, aboutissent à l'ultime à une charge non nulle pouvant être élevée. Ce phénomène se produit notamment lorsque la réserve est sous-estimée lors de l'ouverture du sinistre.

Le montant du capital assuré semble également être corrélé positivement avec la charge des sinistres. Plus la montant du bien assuré est élevé plus le sinistre est coûteux.

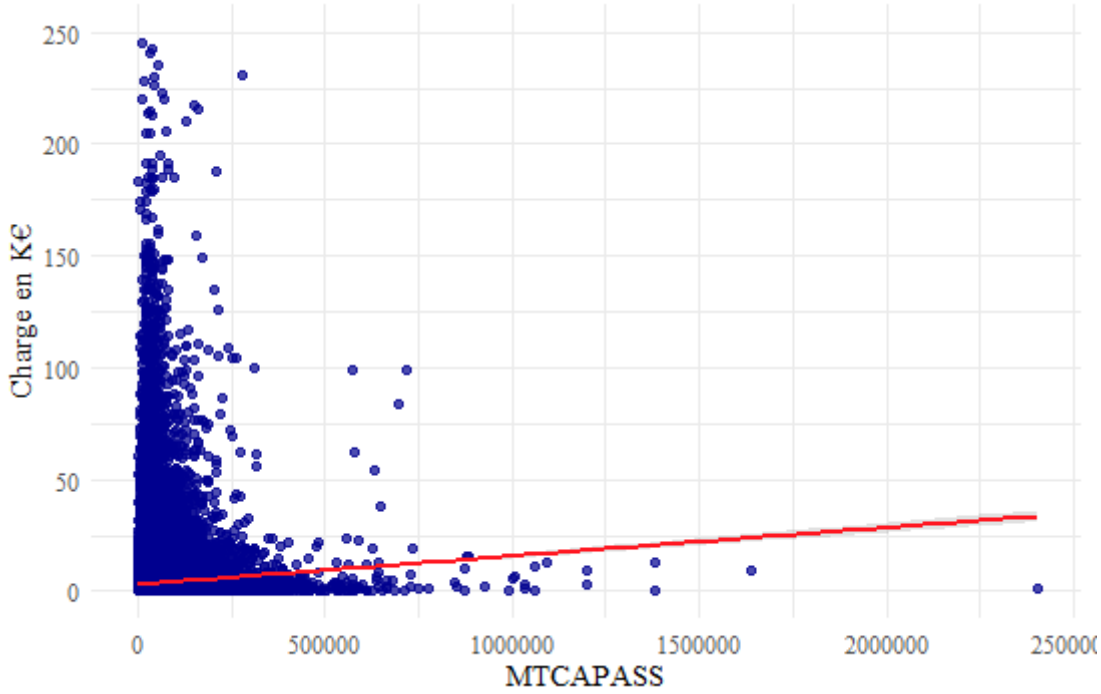


FIGURE 61 – La charge des sinistres en fonction du montant du capital assuré

Les mêmes conclusions graphiques peuvent être faites sur la prime et le nombre de pièces du bien assuré.

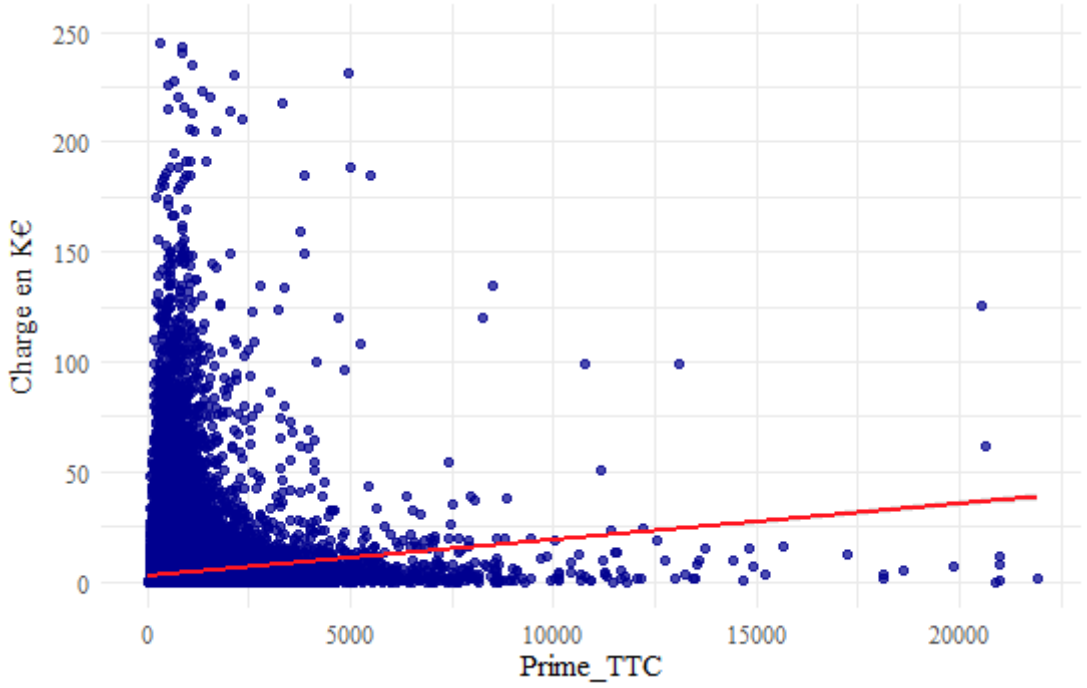


FIGURE 62 – La charge des sinistres en fonction du montant de la prime

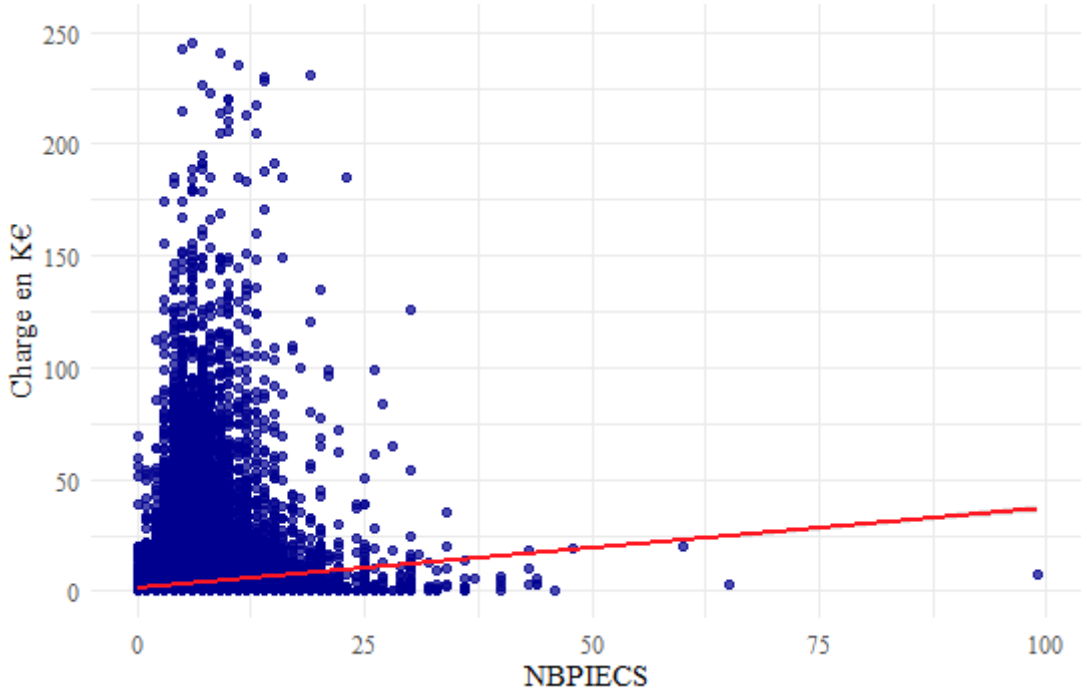


FIGURE 63 – La charge des sinistres en fonction du nombre de pièces du bien assuré

Cependant il est important d'accompagner les analyses graphiques par des analyses statistiques. Pour cela, l'indicateur le plus couramment utilisé est le coefficient de corrélation de Pearson. Le coefficient de Pearson, entre une variable X et une variable Y, est obtenu via la formule suivante :

$$\rho_{X,Y} = \frac{COV\,(X,Y)}{\sigma_X \sigma_Y} = \frac{\sum (X-\bar{X})(Y-\bar{Y})}{\sqrt{\sum (X-\bar{X}^2)\ \sum (Y-\bar{Y})^2}}$$

Les valeurs de ce coefficient varie entre -1 et 1 et sont interprétées comme suit :

- Si $\rho_{X,Y}$ est proche de 1 alors X est positivement corrélé à Y (Si X augmente, Y augmente)
- Si $\rho_{X,Y}$ est proche de -1 alors X est négativement corrélé à Y (Si X augmente, Y diminue)
- Si $\rho_{X,Y}$ est proche de 0 ou nul alors X et Y ne sont pas corrélées (X et Y évoluent d'une façon indépendante)

Il faut noter que cette mesure suppose que les variables suivent une distribution normale.

Le corrélogramme des coefficients de corrélation est représenté ci-dessous :

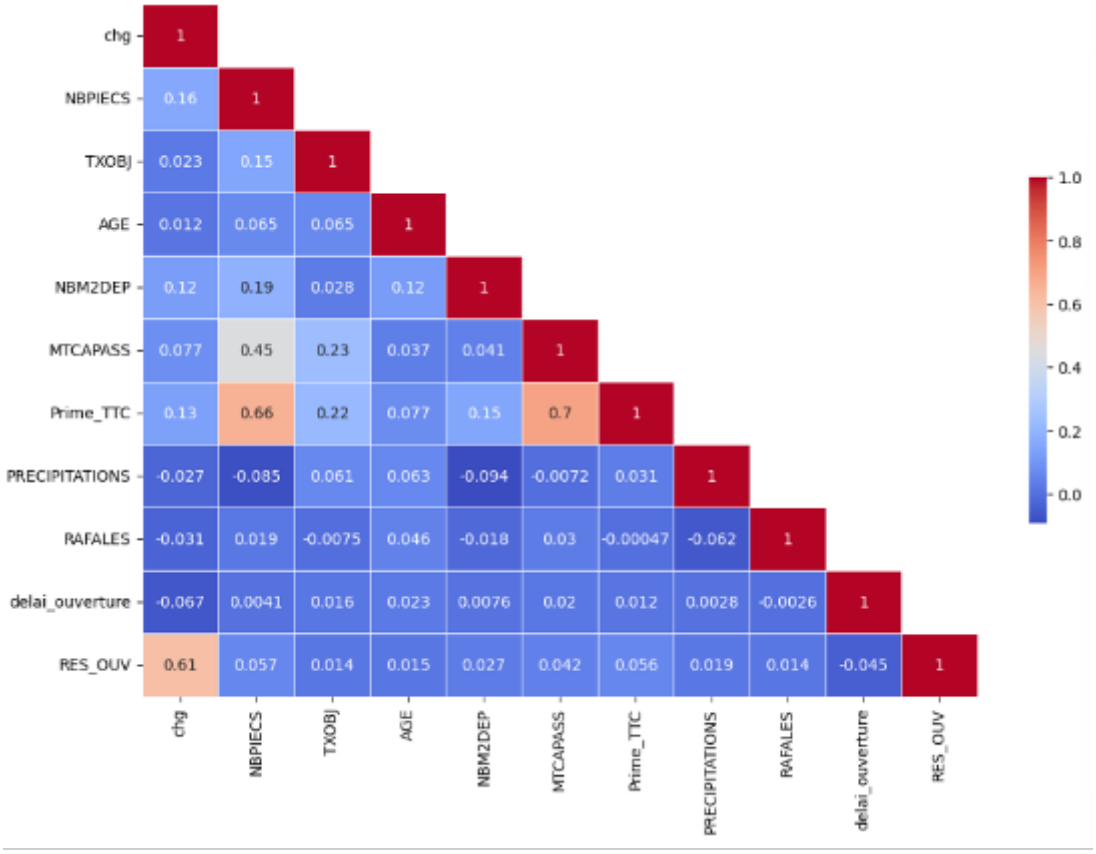


FIGURE 64 – Corrélogramme des variables quantitatives

La lecture de la matrice de corrélation est réalisée via un code couleur :

- Le rouge foncé indique une forte corrélation positive (valeur du coefficient proche de 1)
- Le bleu foncé indique une forte corrélation négative (valeur du coefficient proche de -1)
- Le blanc indique une absence de corrélation (valeur du coefficient proche de 0)

L'intérêt de ce corrélogramme est de valider ou non les conclusions graphiques et d'identifier les variables ayant une corrélation avec la variable cible qui est la charge des sinistres.

La seule corrélation positive est celle avec la réserve d'ouverture : le coefficient de corrélation est de 0,61. Cela signifie que plus la réserve est élevée, plus la charge des sinistres est élevée. Vu qu'un sinistre jugé coûteux à l'ouverture a de fortes chances d'évoluer selon ce jugement, ce résultat est donc attendu. Ce coefficient vérifie aussi la conclusion graphique effectué sur la corrélation entre la réserve d'ouverture et la charge des sinistres.
Les autres variables comme la prime (0,13), le nombre de mètre au carré de dépendance (0,12) ou même le nombre de pièce du bien (0,16) semble être très faiblement corrélé avec la charge des sinistres.

Pour valider ces conclusions, des tests statistiques sont effectués. Les tests les plus couramment utilisés sont :

- Le test de Pearson : utilisé lorsque la corrélation est linéaire, sous l'hypothèse que les données suivent une distribution normale.
- Le test de Spearman : utilisé lorsque la relation entre les variables est monotone mais non nécessairement linéaire. Ce test ne requiert pas l'hypothèse de normalité.
- Le test de Kendall : utilisé lorsque la corrélation est ordinale.

Dans le cadre de la vérification des corrélations entre la charge des sinistres et les autres variables, le test de Spearman est retenu dans cette étude. Le *rho* de Spearman est obtenu selon la formule suivante :

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

Tel que d_i correspond à la différence entre les rangs des valeurs des deux variables en question pour une observation i et n le nombre d'observations.

Ce test est basé sur les deux hypothèses suivantes :

- H_0 : Les deux variables ne sont pas corrélées.
- H_1 : Les deux variables sont corrélées.

L'indicateur de p-value permet de vérifier ou pas l'hypothèse nulle H_0. Cet indicateur est une mesure statistique.

Suivant le seuil d'acceptation choisi (5% dans ce cas), l'hypothèse nulle est rejetée ou acceptée :

- Si la p-value est inférieur à 0,05, l'hypothèse H_0 est rejettée et les variables sont corrélées.
- Sinon, H_0 n'est pas rejettée et les variables ne sont pas corrélées.

Ainsi, selon la p-value, les tests de Spearman ont permis d'identifier les variables ayant une corrélation statistiquement significative avec la charge des sinistres :

- Les variables corrélées avec la charge (p-value $< 5\%$) : Réserve à l'ouverture (« RES_OUV »), Rafales de vent (« RAFALES »), Nombre de départements touchés (« NBM2DEP »), Nombre de pièces (« NBPIECS ») et prime (« Prime TTC »). Graphiquement, une corrélation a été identifié graphiquement entre la charge et ces variables, à l'exception des variables « RAFALES » et « NBM2DEP ».
- Les variables non corrélées avec la charge ultime (p-value $> 5\%$) : Délai d'ouverture du sinistre (« delai_ouverture »), Précipitations (« PRECIPITATIONS »), Montant capital assuré (« MTCAPASS »). Graphiquement une corrélation a été identifié entre la charge et le montant du capital assuré (« MTCAPASS »). Cependant le test de Spearman rejette cette hypothèse.

Ainsi, ces résultats permettent de former un premier avis sur les variables à retenir dans la modélisation et permettent d'orienter la modélisation avec une attention sur les variables significatives et en écartant celles qui ne montrent pas d'influence sur la charge des sinistres.

Pour étudier la corrélation entre une variable quantitative (la charge des sinistres) et une variable qualitative, le test le plus couramment utilisé et qui ne nécessite pas d'hypothèses sur la distribution des variables est le test non paramétrique *Kruskal-Wallis*. Il va tester les hypothèses suivantes :

- H_0 : les médianes des sous-populations sont identiques
- H_1 : les médianes des sou-populations sont différentes

La statistique du test est calculée comme suit :

$$K = (N-1)(\sum_{i=1}^{m} n_i \left(r_i - r\right)^2) / \sum_{i=1}^{m}\sum_{j=1}^{n_i} \left(r_{ij} - r\right)^2$$

Tel que :

- N : le nombre total d'observations
- m : le nombre de groupes
- n_i : le nombre d'observations dans le groupe i
- r_{ij} : le rand de l'observation j dans le groupe i
- r_i : la moyenne des rangs du groupe i
- r : la moyenne globale des rangs et qui est obtenue par $r = \frac{N+1}{2}$

Sous l'hypothèse H_0, la statistique K suit une loi de χ^2 à $(m-1)$ degrés de liberté. L'application de ce test via la fonction kruskal.test du logiciel R montre que pour toutes variables qualitatives, l'hypothèse H0 est rejetée (p-value $< 5\%$). Cela indique que ces variables influencent la charge des sinistres.

Cependant, le rejet de H_0 ne renseigne pas sur la significativité de cette dépendance. Pour mesurer l'ampleur de la relation entre chaque variable qualitative et la charge des sinistres, il est pertinent de calculer un indicateur spécifique, le η^2, défini par la formule suivante :

$$\eta^2 = \frac{K}{N-1}$$

Cet indicateur représente à proportion de variance expliquée par la variable qualitative et est interprété suivant les règles suivantes :

- $\eta^2 \approx 1$: la variable qualitative explique presque toute la variance de la charge
- $\eta^2 \approx 0$: la variable qualitative n'explique presque rien.

En appliquant cette approche aux différentes variables qualitatives, seule la variable indiquant l'intervention ou pas d'un expert missionné « ExpertMissionFlagText » présente un η^2 de 0,2, ce qui indique une corrélation significative avec la charge des sinistres. Les

autres variables présentent des η^2 inférieurs à 0,1, suggérant un effet faible sur la charge.

Ainsi, bien que le test de *Kruskal-Wallis* montre une dépendance statistique entre ces variables et la charge des sinistres, l'analyse de l'η^2 permet de relativiser cette dépendance en mettant en évidence l'importance réelle de cette relation.

III.2.F Observations préalables et focus sur les évènements grêles

Dans cette sous-partie, une analyse exploratoire sur les coûts et les nombres des sinistres est présentée, notamment sur les différentes années de la base de données à disposition.

Comme déjà évoqué, la base de données à disposition contient les sinistres des années 2014 à 2022. Le graphique ci-dessous illustre le nombre de sinistres selon l'année :

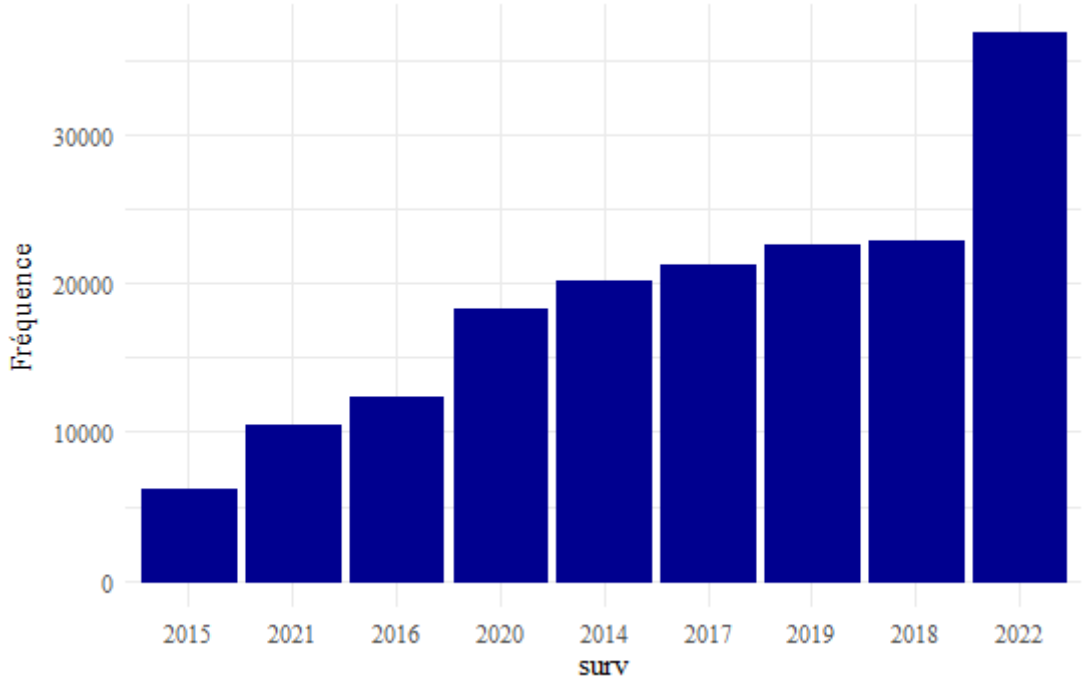


FIGURE 65 – Répartition des sinistres selon l'année de survenance

Il est clair que le nombre de sinistre augmente avec l'année et que l'année 2022 est marquée par le plus grand nombre de sinistres climatiques. Ainsi, il est intéressant de s'attarder sur la distribution de la charge selon l'année.

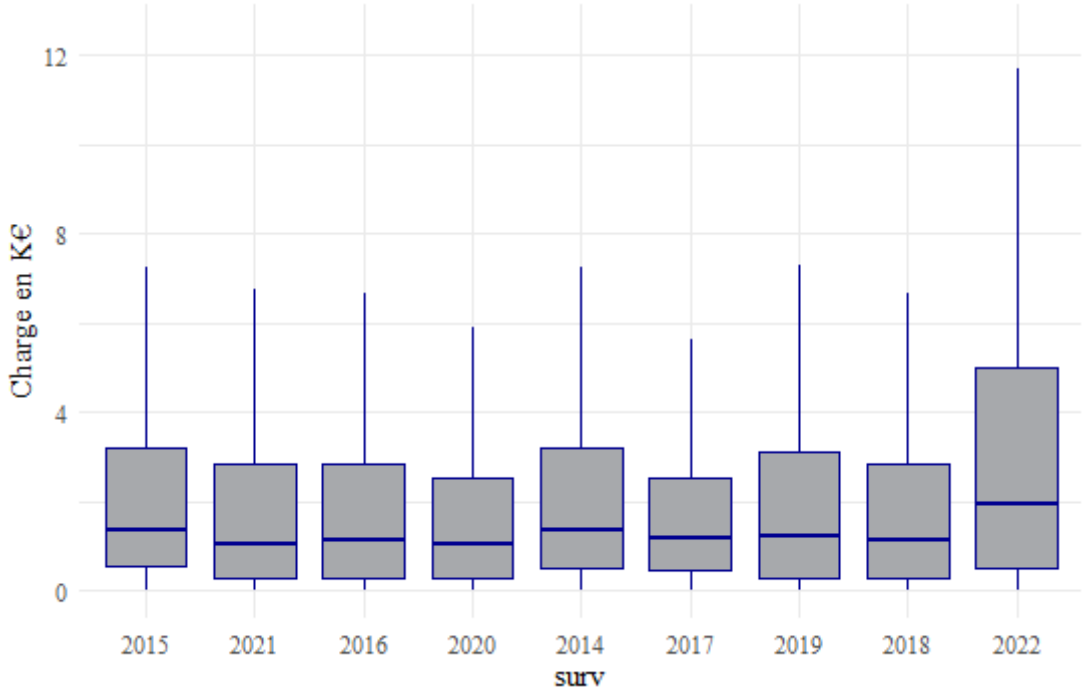


FIGURE 66 – Charge des sinistres selon l'année de survenance

La charge moyenne des sinistres survenu en 2022 est plus importance que la moyenne de ceux des années précédentes. Sur les années antérieurs à 2022, les boîtes à moustaches de la charge sont très similaires. Ainsi, l'année n'influent pas sur la charge des sinistres mais la forte sinistralité de l'année 2022 est impacté par une autre raison.
Sachant que le but de ce mémoire est de proposer un modèle pour la charge ultime et de prédire l'année 2022, il est important de comprendre d'où vient l'atypisme de cette année.

L'intérêt va se porter sur les sinistres grêles. Le graphique ci-dessous montre l'évolution de la proportion de sinistres grêles par année.

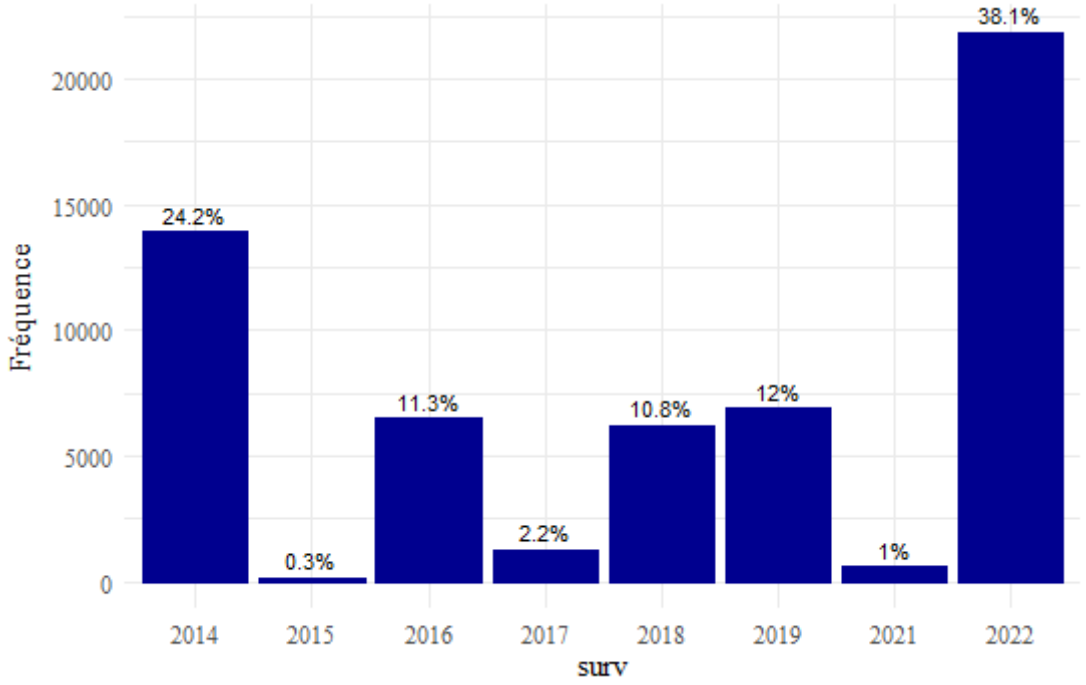


FIGURE 67 – Fréquence des sinistres grêle par année de survenance

Avant l'année 2022, la proportion de sinistres grêle varie entre 0,3% et 12% selon les années avec une année atypique de 24% en 2014. En 2022, cette proportion augmente à 38%. Cela montre que le nombre de sinistres grêle a été anormalement élevé cette année-là. Ainsi, les données historiques ne sont pas représentatives de 2022 et donc le modèle risque d'être biaisé.

Le graphique ci-dessous montre la comparaison des coûts de sinistres suivant le type d'évènement : évènements grêles ou hors grêle.

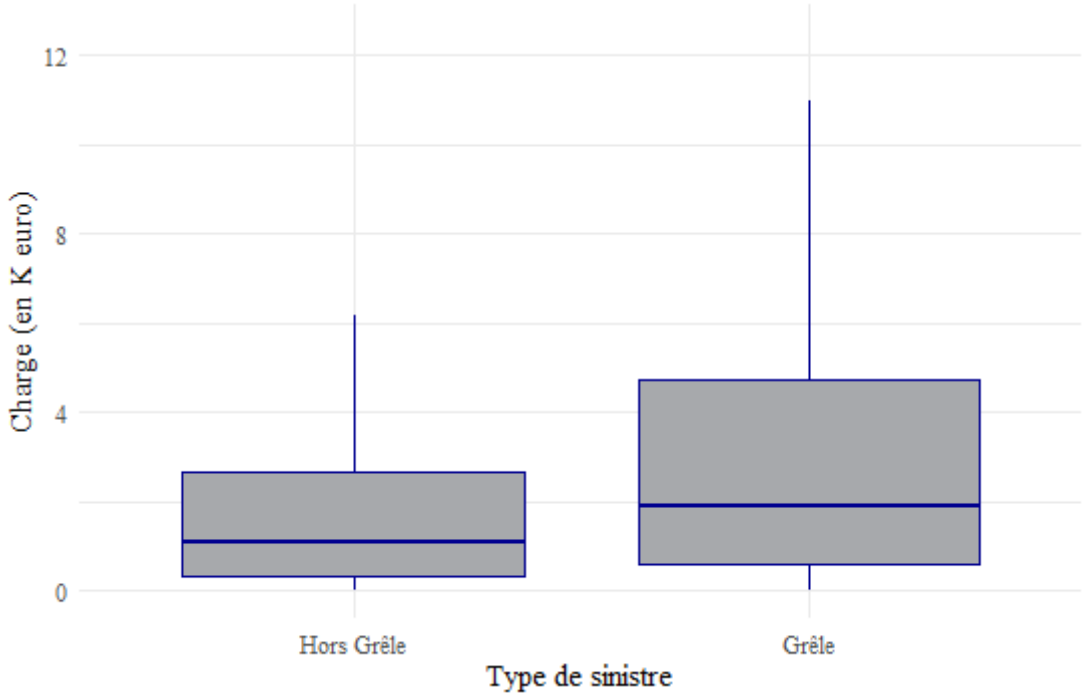


FIGURE 68 – Comparaison de la charge des sinistres Grêle et Hors-Grêle

Toutes années confondues, les sinistres hors grêle ont un coût moyen à 2,48K euro et une médiane à 1,1K euro comparé respectivement à 5,32K et 1,92K euro pour les évènements grêles. Donc le coût moyen global des sinistres grêles est plus important que celui des autres évènements confondus.

Le graphe ci-dessous montre l'évolution du coût moyen des sinistres grêle et des sinistres hors grêle sur les années de 2014 à 2022 :

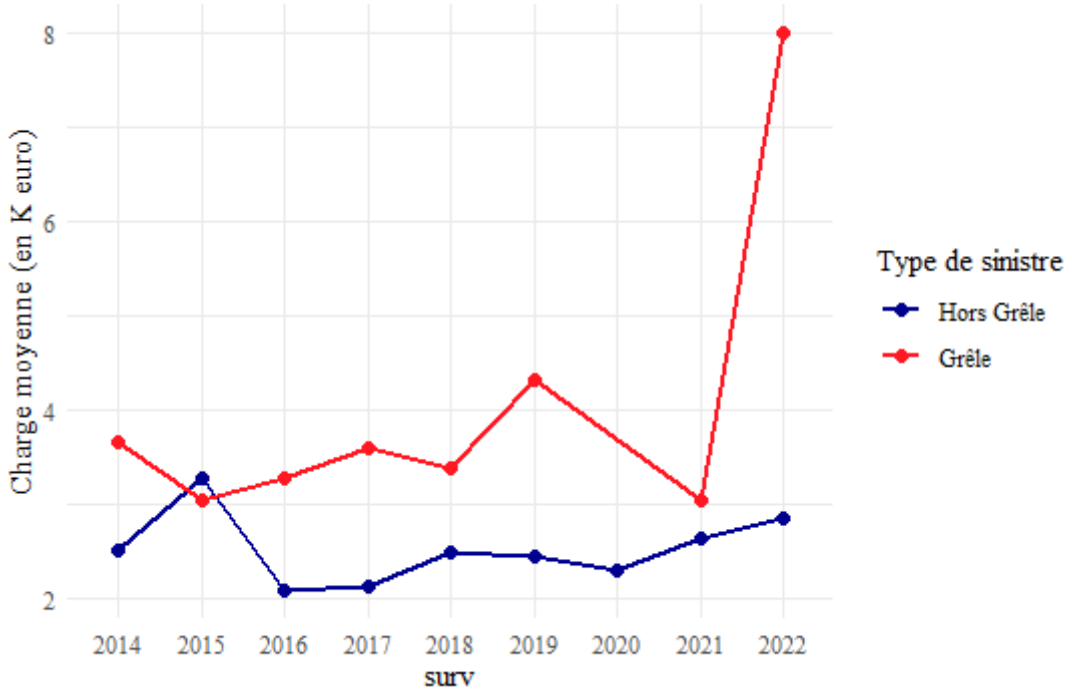


FIGURE 69 – Évolution de la charge moyenne des sinistres (Grêle VS Hors-Grêle)

Sur les années antérieurs à 2022, le coût moyen des sinistres grêle varie entre 3K et 4,3K euro, ce qui n'est pas très loin des montants des sinistres des évènements hors grêle. C'est en 2022 que le coût moyen des sinistres grêle bondit à 8,01K euro.

Ainsi, la différence énorme entre les sinistres grêle et les autres sinistres montre qu'ils n'ont pas la même dynamique de coût. En 2022, le coût des sinistres grêle a explosé, ce qui risque de fausser un modèle unique qui aurait appris sur des données où la grêle était moins fréquente et moins coûteuse.

Les années 2014 à 2021 ne reflètent pas la situation de 2022. Si le modèle s'entraîne sur tous les évènements, cela risque de biaiser le modèle et ainsi de sous-estimer la charge ultime. En conséquence, le modèle va sur-apprendre sur les dynamiques spécifiques des évènements.

La charge des sinistres grêle étant significativement plus élevée, les dynamiques de ces deux catégories ne sont pas les mêmes et incitent donc à proposer deux modèles distincts pour la charge : un modèle pour les évènements de grêles et un autre pour tous les autres événements. Un seul modèle risque d'être déséquilibré et sous-estimer les coûts des évènements grêle en particulier si le modèle prédisant 2022 est entraîné sur les années antérieures à 2022.
En procédant de cette manière, la somme des estimations des deux modèles donnera la prédiction de la charge ultime.

III.2.G Présentation des charges *as if*

Le but est d'entraîner un modèle sur les années 2014 à 2021 et d'évaluer sa qualité de prédiction sur l'année 2022. Il faut noter que les charges de sinistres annuelles évoluent temporellement. Cette évolution peut être due par exemple à l'effet de l'inflation. De plus, les charges de sinistres peuvent être très volatiles et impactées par des sinistres exceptionnels (comme de forts épisodes de grêle). Ainsi, pour que les montants des coûts de sinistres antérieurs à 2022 deviennent comparables, il est important de les normaliser par rapport à 2022.

C'est pour ces différentes raisons que les charges « *as if* » sont introduites. Concrètement, les charges « *as if* » sont des charges obtenues en leur appliquant un coefficient basé sur un indicateur spécifique et bien choisi et permettent d'ajuster les valeurs historiques à une année de référence. Le but est de pouvoir corriger les effets liés aux variations temporelles sur les montants de charge de sinistres en neutralisant les effets de l'inflation et les effets des évènements extrêmes tout au long des années.

L'année 2022 est supposée être l'année de référence. Pour un indicateur I et pour une année i tel que $2014 \leq i \leq 2021$, le coefficient C_i appliqué à la charge est obtenu par :

$$C_i = \frac{I_{2022}}{I_i}$$

La charge « *as if* » chg_i pour l'année i est obtenue en supposant que les conditions de l'année 2022 s'appliquent à cette année :

$$chg_i = chg_i * C_i$$

Pratiquement, il faut choisir des indicateurs qui ont un impact sur la charge des sinistres dans le cadre de l'assurance MRH. Deux options d'indicateurs peuvent être testés :

- L'indicateur de FFB (Fédération Française du Bâtiment) et les indicateurs BT (Bâtiments-Travaux)
- Le coût moyen observé à J+7

Pour chaque indicateur, le choix de l'indicateur est justifié par la suite et la méthodologie abordée est expliquée. Ainsi, le but sera de comparer ces deux approches à la fois pour la modélisation des évènements grêles et hors grêles.

1) Indicateur FFB :

Cette approche utilise d'une part l'indicateur de FFB et d'autre part les indicateurs de BT, pour créer les charges « *as if* ».

Cette approche est macroéconomique. L'indicateur FFB permet de prendre en compte l'inflation en intégrant les évolutions et les variations du coût de la construction et de

réparation, de la main-d'œuvre et d'autre charge pour les bâtiments. Par exemple, si un sinistre tempête survient la charge de ce sinistre repose principalement sur la reconstruction/réparation des biens touchés. Ainsi, il reflète bien les coûts associés aux sinistres en MRH et justifie le choix d'un tel indicateur.

La méthodologie consiste à ajuster la charge en fonction de l'évolution du coût de la construction et de la réparation des dommages à la suite des évènements climatiques. Cependant, pour chaque modèle (grêle et hors grêle) l'approche sera légèrement différente. Le modèle hors grêle se base sur les indicateurs globaux de la FFB, alors que le modèle grêle se base sur un maillage de cet indicateur qui correspond aux indicateurs BT. Dans ces deux approches, c'est l'impact du coût de réparation et de reconstruction sur les charges des sinistres qui est pris en compte.

Chaque année, la FFB publie mensuellement ces indicateurs. Les indicateurs du mois de décembre sont considérés dans la suite car ils reflètent plus la réalité de l'année.

Pour le modèle Hors Grêle :

Depuis le site de la Fédération Française du Bâtiment (FFB), les indices de FFB des années 2014 à 2022 sont récupérés et le coefficient d'évolution par rapport à 2022 est calculé.

Pour l'année i tel que $2014 \leq i \leq 2021$, le coefficient de l'indicateur de FFB par rapport à l'année de référence 2022 C_i^{FFB} est obtenu par :

$$C_i^{FFB} = \frac{I_{2022}^{FFB}}{I_i^{FFB}}$$

Le tableau ci-dessous recense les indicateurs FFB annuels ainsi que les coefficients qui seront appliqués à la charge pour obtenir la charge « *as if* ».

Année	Indice FFB	Évolution par rapport à 2022
2014	930,8	1,22
2015	929,5	1,22
2016	942	1,21
2017	974,8	1,17
2018	988,2	1,15
2019	994,3	1,14
2020	1 000,5	1,14
2021	1 066,4	1,07
2022	1 137	1,00

TABLE 22 – Les indicateurs FFB annuel et les coefficients associés

Ainsi, les conditions économiques de l'année 2022 seront bien prises en compte dans les charges « *as if* » FFB des sinistres hors grêle.

Pour le modèle Grêle :

Un maillage plus fin peut être apporter à la charge des sinistres grêle. Au lieu d'utiliser les indicateurs FFB, les indicateurs de BT seront plutôt utilisés.

En effet, l'indice FFB intègre directement ces indices car il reflète l'évolution global des coûts liés aux bâtiments. Les indices BT sont des composantes spécifiques pour refléter l'évolution d'un coût spécifique lié aux réparations des bâtiments. Ils en existent plusieurs comme « BT07 Ossatures et charpentes » qui représente le coût des structures en bois, métal ou béton.

La question qui se pose est : quel indicateur prendre en compte dans le calcul des charges « *as if* » ? En effet, grâce à la variable « DESCRIPTION », les éléments les plus touchés dans les bâtiments seront identifiés. Par exemple, si les descriptions évoquent les vitres, l'indice « BT45 Vitrerie » pourra être considérer. C'est ainsi qu'une analyse des descriptions des sinistres a été menée afin d'extraire les informations les plus pertinentes et de déterminer les indices BT les plus représentatifs des sinistres. En complément de l'indice FFB, cette approche permet d'affiner le choix des indices les plus représentatifs des conséquences spécifiques des sinistres de grêle.

Toutefois, la variable représentant la description du sinistre est très détaillée et il est donc impossible de l'exploiter ligne à ligne. Pour cela, du traitement automatique du langage NLP (*Natural Language Processing* en anglais), ou « *text mining* » en anglais, doit être réaliser sur cette variable afin d'identifier les éléments les plus fréquemment affectés.

Le *text mining* est une technique d'apprentissage statistique qui a pour but d'extraire des données à partir de données textuelles afin d'identifier des tendances, des relations ou des informations au sein d'un corpus de texte. Sans devoir lire l'intégralité du texte ou le contenu d'une variable ligne à ligne, cette approche permet de créer de l'information plus rapidement. Elle repose sur plusieurs concepts, par exemple l'extraction des termes clés ou l'analyse de leur fréquence.

Dans le cadre de l'identification des éléments les plus touchés par les sinistres, la méthodologie abordée a été inspiré de la méthode de *text mining*. Contrairement au *text mining* qui repose sur des algorithmes de NLP ou de classification, la méthodologie abordée dans ce cas repose plutôt sur une analyse exploratoire des descriptions des sinistres et une sélection manuelle des mots-clés les plus significatifs.

L'analyse des descriptions de sinistres permet de repérer les termes les plus fréquemment utilisés (par exemple, « toiture », « tuiles cassées », « vitres brisées », etc.) et guide dans le choix des indices BT les plus pertinents.

Pour ce faire, plusieurs étapes sont nécessaires à la réalisation de cette méthode. Tout d'abord, les données liées à la charge « chg », l'année « surv », la description du sinistre « DESCRIPTION » sont extraites en filtrant sur les données non manquantes dans la variable « DESCRIPTION ». Ensuite, la base est divisée selon l'année pour obtenir ainsi sept bases. Le processus suivant sera réalisé pour chaque année et donc sur chaque base.

La description du sinistre est saisie manuellement par les gestionnaires et continent des erreurs de mise en forme du texte ou d'orthographe pouvant affecter cette analyse. Pour résoudre cette problématique, un pré-traitement des données doit être effectué afin de les convertir dans un format exploitable et extraire les informations significatives. Aussi, il est essentiel de considérer l'élimination de certains éléments textuels, tels que la ponctuation, les majuscules, les caractères spéciaux et les « *stopwords* » (mots courants comme « je », « et », « que », etc.). Les termes répétitifs seront supprimés, comme « client », « suite » ou « type » » qui n'apportent pas d'information efficace.

Dans le but d'améliorer la standardisation du texte, le traitement de la donnée inclut aussi d'autres étapes importantes. La tokenisation consiste à découper le texte en plusieurs mots appelées tokens. Ensuite, l'étape de lemmatisation permet de ramener chaque mot à sa forme initiale, par exemple « impactent » devient « impacter ».

Le nuage de mots est parmi les différentes méthodes de *text mining*. on Il est basé sur la fréquence d'apparition des termes. Cette approche repose sur l'identification des mots les plus récurrents dans la variable « DESCRIPTION » et la création d'une représentation graphique où la taille des mots reflète leur fréquence d'occurrence, facilitant ainsi l'interprétation des résultats.

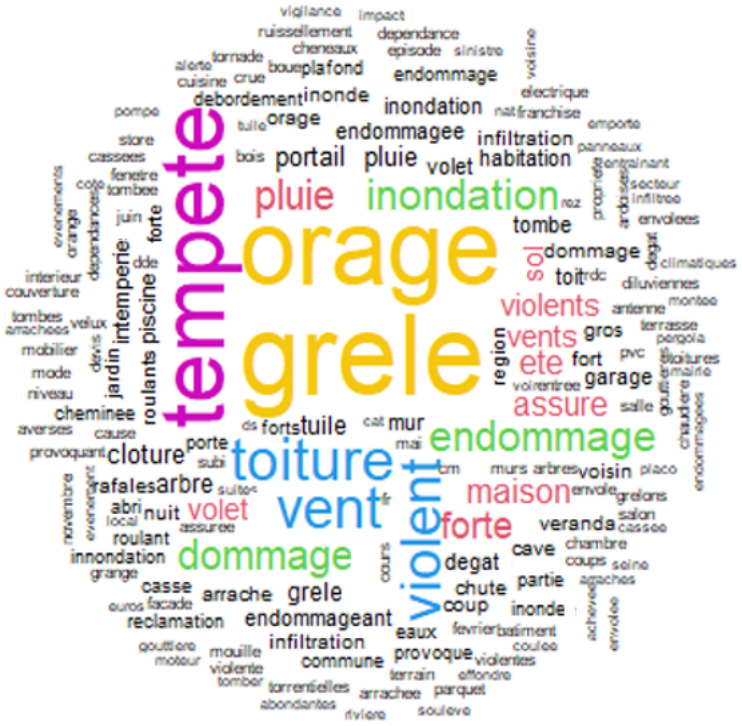


FIGURE 70 – Nuage de mots

L'analyse de ce graphique met en évidence la forte occurrence de certains termes : « toiture », « vitrerie », « tuiles » et « menuiseries ». Grâce à l'analyse des fréquences et à la classification des mots, quatre indices de construction se dégagent comme les plus représentatifs des dégâts causés par les évènements de grêle :

- BT07 Toitures et charpentes
- BT45 Vitrerie
- BT32 Tuiles en terre cuite
- BT51 Menuiserie PVC

Les données des indices retenus ont été récupéré du site de l'INSEE à la vision de décembre et sont présentés dans le tableau ci-dessous :

Indice	2014	2015	2016	2017	2018	2019	2020	2021	2022
BT07 ossatures et charpentes	100,5	98,8	100,8	108,7	113,6	110,6		150,1	160,2
BT45 Vitrerie	106,9	108,5	109,8	112,8	118,4	120,9		127,5	148,8
BT32 Tuiles terre cuite	105,9	105,5	109,1	110,5	111,8	113,3		121,1	133,8
BT51 Menuiserie PVC	104,8	104,5	105,8	106,1	108,3	110,1		114,5	126,6
Moyenne	36%	37%	34%	30%	26%	25%		11%	

TABLE 23 – Tableau des indices BT

Pour optimiser encore plus les résultats et au lieu de juste considérer la moyenne annuelle des indices comme coefficient, on va plutôt opter pour une moyenne des indices pondéré par la proportion des sinistres ayant des dégâts liés à l'indice en question. Pour calculer les différentes proportions, on va créer 4 variables indicatrices :

- Indicatrice Toiture : cette variable prend la valeur 1 si au moins l'un des termes toiture, toit, plafond, plafond ou taule apparaît dans le contenu de la variable « DESCRIPTION », sinon elle prend la valeur 0
- Indicatrice Vitrerie : cette variable prend la valeur 1 si au moins l'un des termes véranda, velux, vitrage, verrière, vitre ou verre apparaît dans le contenu de la variable « DESCRIPTION », sinon elle prend la valeur 0
- Indicatrice Tuile : cette variable prend la valeur 1 si au moins le terme tuile apparaît dans le contenu de la variable « DESCRIPTION », sinon elle prend la valeur 0
- Indicatrice Menuiserie : cette variable prend la valeur 1 si au moins l'un des termes volet ou roulant apparaît dans le contenu de la variable « DESCRIPTION », sinon elle prend la valeur 0

Ces indicatrices permettent ensuite d'évaluer la part de chaque type de dommage. 1_x représente la variable indicatrice pour chaque élément $x \in A = \{toiture, vitrerie, tuile, menuiserie\}$. Ainsi, la proportion $x \in A$ est calculée de la manière suivante :

$$proportion_x = \frac{\sum 1_x}{\sum_{y \in Ax} 1_y}$$

Les résultats des proportions sont recensés dans le tableau ci-dessous :

	2014	2015	2016	2017	2018	2019	2020	2021	2022
proportion_toiture	55,9%	37,5%	48,0%	52,6%	56,2%	39,5%		38,0%	46,1%
proportion_vitrerie	13,7%	16,7%	10,4%	17,9%	9,1%	11,6%		7,0%	9,7%
proportion_tuile	16,8%	16,7%	6,8%	8,4%	12,3%	13,2%		11,3%	16,2%
proportion_menuiserie	13,7%	29,2%	34,7%	21,1%	22,4%	35,6%		43,7%	28,0%

TABLE 24 – Tableau des proportions selon l'année

Les moyennes des indices d'évolutions par rapport à 2022 et pondéré par la proportion sont obtenues selon chaque année i tel que $2014 \leq i \leq 2021$:

$$\begin{aligned}\text{MoyennePondérée}_i = \frac{1}{4}(&BT07_i * \text{proportion}_{\text{toiture}} + BT45 * \text{proportion}_{\text{vitrerie}} \\ &+ BT32 * \text{proportion}_{\text{tuile}} + BT51 * \text{proportion}_{\text{menuiserie}})\end{aligned}$$

Le tableau ci-dessous présente les résultats des moyennes pondérées sur les différentes années.

	2014	2015	2016	2017	2018	2019	2020	2021
Moyenne pondérée	45,8%	40,1%	40,4%	36,5%	31,6%	28,1%		9,5%
Moyenne	36,4%	36,8%	34,2%	29,9%	25,8%	25,3%		11,1%

TABLE 25 – Moyenne des indices pondérée par la proportion

Ces moyennes correspondent aux coefficients C_i^{BT} qui seront appliqués pour le calcul de la charge « *as if* » pour chaque année i. La moyenne pondérée est plus élevée que la moyenne classique : ce qui va donc aider à faire face à la problématique de sous-estimations de la charge ultime des évènements grêle.

2) Indicateur Coût Moyen à J+7 (CM J+7)

La méthode du CM J+7 repose sur une approche davantage orientée métier que macroéconomique.
D'un point de vue actuariel, cette approche permet d'intégrer la dynamique d'évolution réelle du coût des indemnisations des sinistres au fil du temps. Elle vise à ajuster la charge des sinistres (grêle et hors grêle) en fonction du coût moyen observé après une semaine. Cette méthodologie permet de prendre en compte l'évolution réelle du coût des sinistres réglés au bout de sept jours, en comparant les valeurs des différentes années à l'aide du ratio $CM_{J+7}(2022)/CM_{J+7}(i)$.

Pratiquement, l'extraction des données est réalisée via SAS en agrégeant par année le nombre total de sinistres (NB), le nombre de sinistres clos sans suite (NBCSS) et la charge journalière des sinistres (chg). Le calcul du CM à J+7 pour chaque année i entre 2014 et 2021 s'effectue selon la formule suivante :

$$CM_{J+7}(i) = \frac{chg_i}{NB_i - NBCSS_i}$$

Le coefficient C_i^{CM} « *as if* » pour l'année i correspond à l'évolution du coût moyen et est déterminée pour les années 2014 à 2021 via la formule suivante :

$$C_i^{CM} = \frac{CM_{J+7}(2022)}{CM_{J+7}(i)}$$

Cette approche permet ainsi d'analyser l'évolution du coût des sinistres dans le temps et d'ajuster les charges des sinistres en conséquence.

Les tableaux ci-dessous recense les résultats des coefficients obtenus grâce à cette méthode :

→ Pour les évènements hors grêle :

HORS GRELE					
surv	**NB**	**NBCSS**	**chg**	**CM_7**	**Evol CM**
2014	2 070,0	738,0	7 343,5	**5,5**	0,0%
2015	1 946,0	393,0	9 538,5	**6,1**	0,0%
2016	1 920,0	392,0	5 104,7	**3,3**	27,4%
2017	5 819,0	1 017,0	16 471,1	**3,4**	24,1%
2018	5 471,0	1 741,0	17 619,9	**4,7**	0,0%
2019	4 828,0	1 379,0	15 477,9	**4,5**	0,0%
2020	6 665,0	1 796,0	17 748,7	**3,6**	16,8%
2021	4 453,0	1 485,0	13 240,5	**4,5**	0,0%
2022	7 296,0	1 631,0	24 110,5	**4,3**	

TABLE 26 – Résultats de la méthode de CM pour les évènements hors grêle

→ Pour les évènements grêle :

GRELE					
surv	**NB**	**NBCSS**	**chg**	**CM_7**	**Evol CM**
2014	5 187,0	736,0	26 326,7	5,9	117%
2015	36,0	7,0	110,5	3,8	236%
2016	2 852,0	1 126,0	10 517,3	6,1	110%
2017	529,0	85,0	2 337,0	5,3	144%
2018	1 913,0	471,0	9 474,9	6,6	95%
2019	2 544,0	511,0	14 459,9	7,1	80%
2021	247,0	43,0	819,8	4,0	219%
2022	11 751,0	1 899,0	114 832,1	11,7	

TABLE 27 – Résultats de la méthode de CM pour les évènements grêle

Pour ces deux approches, des bases de données sont formées par les colonnes suivantes :

- Année
- Coefficient FFB hors grêle (HG)
- Coefficient BT grêle (G)
- Coefficient CM hors grêle (HG)
- Coefficient CM grêle (G)

Ensuite, cette base est fusionnée à notre base de données selon la clé « Année » pour pouvoir ainsi calculer les charges « *as if* ». Pour l'année i tel que $2014 \leq i \leq 2021$, les évènements e tel que $e \in \{G,\ HG\}$ et M la méthode tel que $M \in \{FFB,\ BT,\ CM\}$, la charge « *as if* » $chg_{asif}\,(i, e, m)$ est obtenue selon l'équation suivante :

$$chg_{asif}\,(i, e, m) = chg\,(i) * C_i^m(e)$$

III.2.H Préparation du jeu de données

L'objectif est de proposer une méthode d'estimation, sinistre par sinistre, des charges ultimes des évènements climatiques survenus de 2014 à 2022. Pour ce faire, la base de données est structurée de manière à définir une variable cible – la charge finale une fois le sinistre clos – ainsi que les variables explicatives présentées dans les parties précédentes.

Comme expliqué, un modèle pour les sinistres des évènements grêle et un modèle pour tous les autres évènements sont développés. La base est donc scindée en deux : une base avec les évènements de grêle et l'autre avec le reste des évènements.

Pour chaque modèle deux sous modèles seront proposés : un modèle pour la charge « *as if* » avec l'indice FFB/BT et un autre avec l'indice de CM.

La finalité est donc 4 modèles :

- Modèle Grêle charge « *as if* » BT qu'on note modèle G_BT
- Modèle Grêle charge « *as if* » CM J+7 qu'on note modèle G_CM
- Modèle Hors Grêle charge « *as if* » FFB qu'on note modèle HG_FFB
- Modèle Hors Grêle charge « *as if* » CM J+7 qu'on note modèle HG_CM

La modélisation est faite via l'algorithme *XGBoost*. L'objectif est d'aboutir à un modèle qui permet d'estimer la charge ultime pour chaque observation grâce à l'ensemble des données (variables explicatives) à disposition.

Pour se faire, la base de données est scindée en trois bases :

- Une base d'apprentissage : cette base est formée de 80% des données des évènements choisis aléatoirement et elle servira à entraîner le modèle et donc à ajuster ses paramètres.
- Une base de test : cette base est formée de 20% des données choisi aléatoirement et elle servira à évaluer les performances du modèle et affiner ses hyperparamètres.
- Une base de validation : cette base est formée des évènements de l'année 2022 et servira à mesurer la capacité du modèle à généraliser sur de nouvelles données.

La base de validation permet d'éviter le sur-apprentissage. En effet, ce phénomène se produit lorsque le modèle mémorise trop précisément les données d'entraînement, au détriment de sa capacité à faire des prédictions pertinentes sur des cas inconnus.

III.2.I Phénomène de censure

L'analyse exploratoire des données met en évidence une disparité marquée entre les sinistres clos et ceux encore en cours.
Le graphique ci-dessous montre la répartition des sinistres selon son état (ouvert ou clos).

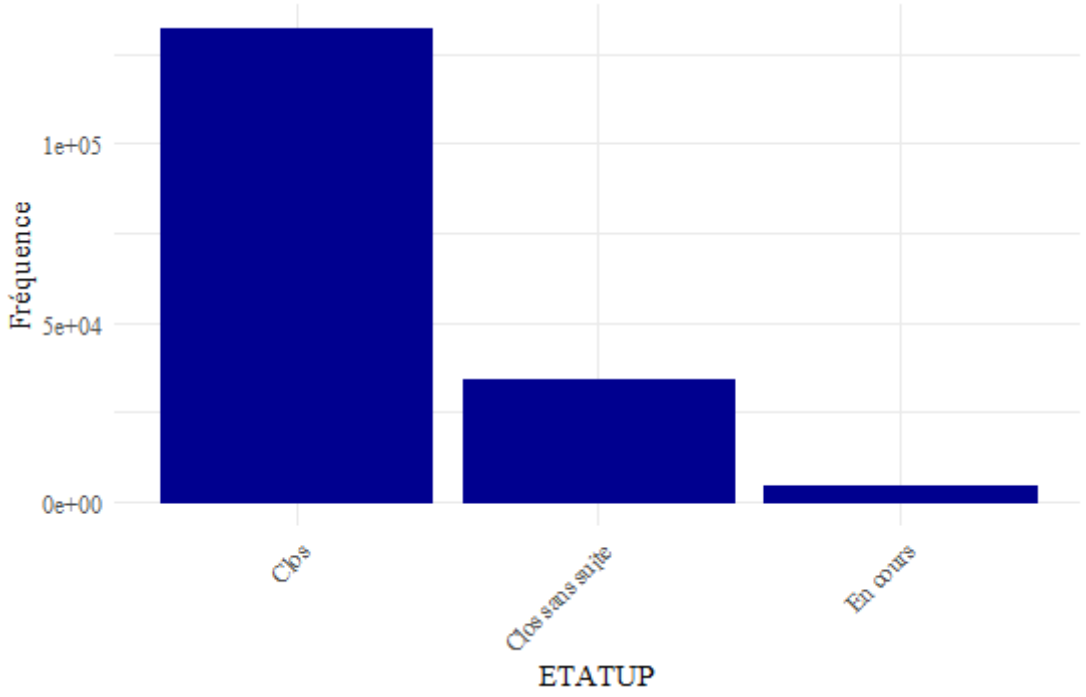


FIGURE 71 – Répartition des sinistres selon leur état

La majorité des sinistres dans la base sont clos ou clos sans suite. Très peu de sinistres sont toujours en cours (2,7%).
Le graphique ci-dessous représente la distribution de la charge selon l'état du sinistre :

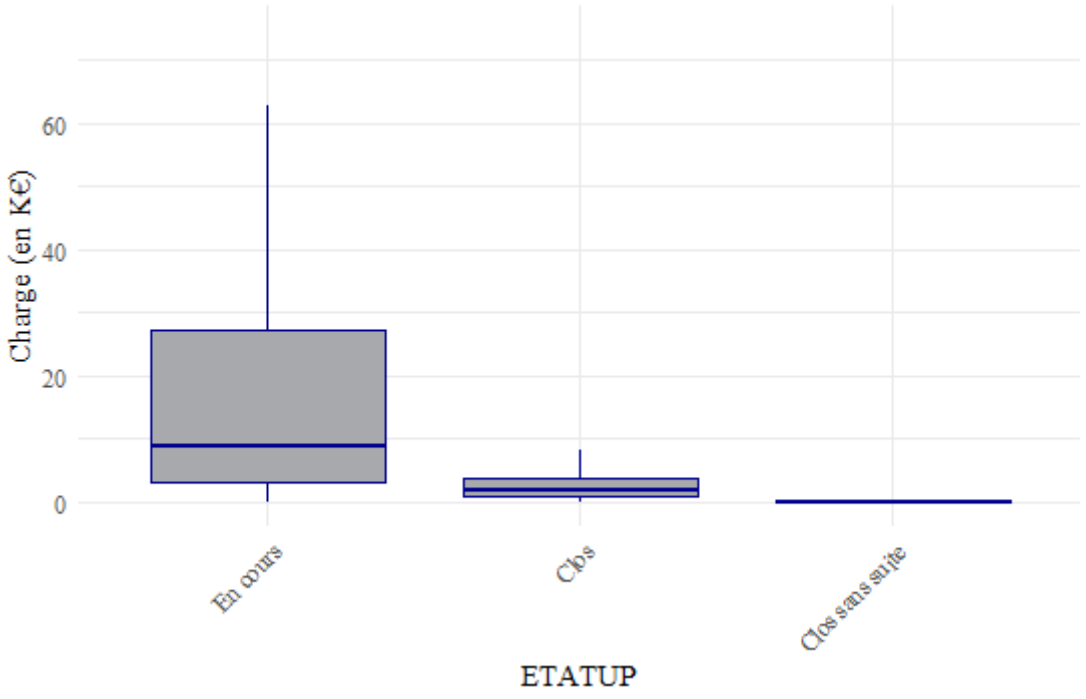


FIGURE 72 – Charge des sinistres selon leur état

Les sinistres en cours présentent une particularité majeure. En effet, leur charge moyenne

est nettement plus élevée que celle des sinistres clos. Les sinistres en cours ont une charge moyenne de 22 646 K€, les sinistres clos 3 611 K€ et les sinistres clos sans suite 133 K€ pour. Bien que les sinistres en cours ne soient pas très nombreux dans la base, ils représentent une charge élevée. Ainsi, s'ils ne sont pas pris en compte dans la modélisation, un biais pourrait être introduit. Un modèle basé uniquement sur les sinistres clos sans ajustement préalable, peut entraîner une sous-estimation des charges ultimes, notamment pour les sinistres complexes et de longue durée, qui sont généralement plus coûteux.

Ainsi, une méthodologie adaptée doit être mise en place pour intégrer ces sinistres en cours dans l'analyse. Une approche pertinente consiste à utiliser des poids de correction, comme les poids de *Kaplan-Meier*, afin d'ajuster l'impact des sinistres en cours et de garantir une estimation plus robuste des charges ultimes. Ces poids sont proportionnels à la durée d'observation du sinistre, permettant ainsi d'atténuer l'effet de la censure sur l'estimation des charges ultimes. Soit le vecteur aléatoire (M, T, X) :

- $X \in \chi \subset \mathbb{R}^d$: ensemble de variables explicatives.
- $T \in \mathbb{R}^+$: Durée de vie du sinistre qui correspond au délai entre l'ouverture du sinistre et sa clôture.
- $M \in \mathbb{R}^+$: Montant de charge ultime du sinistre, c'est-à-dire la charge totale une fois le sinistre clôturé.

Dans un contexte de censure, la variable C représente la durée entre l'ouverture du sinistre et la fin de la période d'observation (date d'inventaire ou annulation du sinistre). Ainsi, il n'est pas toujours possible d'observer directement (T, M), mais plutôt les variables suivantes :

- $Y = \min(T, C)$
- $\delta = 1_{T \leq C}$
- $N = \delta M$

Les données disponibles se présentent donc sous la forme de n observations indépendantes et identiquement distribuées :

$$(N_i, Y_i, \delta_i, X_i)_{1 \leq i \leq n}$$

C est est supposé indépendant de (M, T, X) et que :

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

Pour corriger le biais introduit par la censure, les poids de *Kaplan-Meier* sont appliqués et définis par :

$$\mu_i = \frac{\delta_i}{n(1 - \hat{G}\left(Y_i^-\right))}$$

Tel que $\hat{G}$ l'estimateur de $Kaplan-Meier$ de la fonction de répartition $G(t) = P(C \leq t)$, de la variable de censure C :

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

Les poids μ_i sont nuls pour les observations censurées, tandis que pour les sinistres non censurés, plus la durée d'observation est longue, plus le poids attribué est élevé. Cela permet ainsi de compenser l'absence des sinistres en cours lors du calibrage du modèle prédictif.

L'application de cette méthode sur le logiciel R permet d'obtenir la courbe de survie des sinistres.

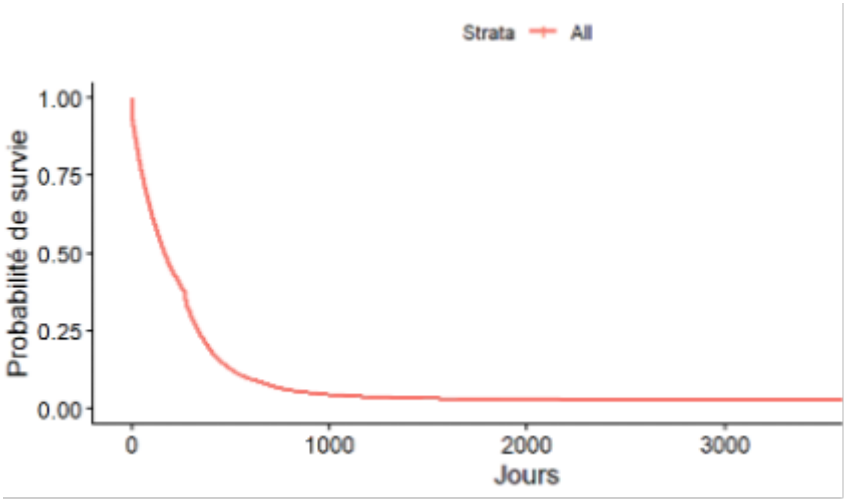


FIGURE 73 – Courbe de survie de Kaplan-Meier

L'estimation de *Kaplan-Meier* indique que :

→ 25% des sinistres ont une durée de vie inférieure à 52 jours.

→ 50% des sinistres ont une durée de vie inférieure à 164 jours.

→ 75% des sinistres ont une durée de vie inférieure à 335 jours.

Les poids de *Kaplan-Meier* ainsi calculés seront ensuite intégrés dans le modèle, afin d'améliorer l'estimation des charges ultimes en prenant en compte l'impact de la censure sur les sinistres en cours.

III.2.J Choix des variables

Avant de procéder à la modélisation *XGBoost* et afin de réduire la complexité de la modélisation, il est essentiel d'effectuer une sélection des variables, tout en conservant les informations les plus pertinentes pour la prédiction de la charge des sinistres. En effet, la base de données présente un grand nombre de variables (une quarantaine à peu près), notamment après son enrichissement avec de nouvelles informations. Cependant, l'intégration de toutes les variables dans le modèle est susceptible d'entraîner plusieurs risques :

- Le modèle devient plus complexe et plus difficile à interpréter.
- Le modèle aura tendance à sur-apprendre.

- Le temps de calcul sera plus long et n'apportera pas un gain significatif en performance.

Ainsi, une première sélection des variables les plus influentes est nécessaire afin de simplifier le problème et d'améliorer la robustesse du modèle.

Tout d'abord, il faut noter que certaines variables présentes dans la base de données ne contribuent pas directement à la prédiction et seront exclues de la modélisation : le numéro de sinistre, l'année de survenance, la description du sinistre, la date de clôture. Ces variables sont donc retirées de la base avant la sélection des variables explicatives.

Pour identifier les variables les plus influentes sur la charge des sinistres, une analyse exploratoire basée sur un modèle de *Random Forest* est réalisée. Deux modèles distincts sont construits :

- Un modèle *Random Forest* pour les sinistres liés à la grêle.
- Un modèle *Random Forest* pour les sinistres hors grêle.

Ces modèles sont utilisés avec leurs paramètres par défaut, car leur objectif n'est pas l'optimisation des prédictions, mais uniquement la sélection des variables les plus pertinentes. Ainsi, cette démarche permet d'identifier les variables ayant le plus grand impact sur la prédiction de la charge des sinistres.

Pour chaque modèle, les résultats de la sélection des variables sont visualisés sous forme de graphiques d'importance des variables.

Le graphe ci-dessous représente l'importance des variables dans le modèle des évènements de grêle.

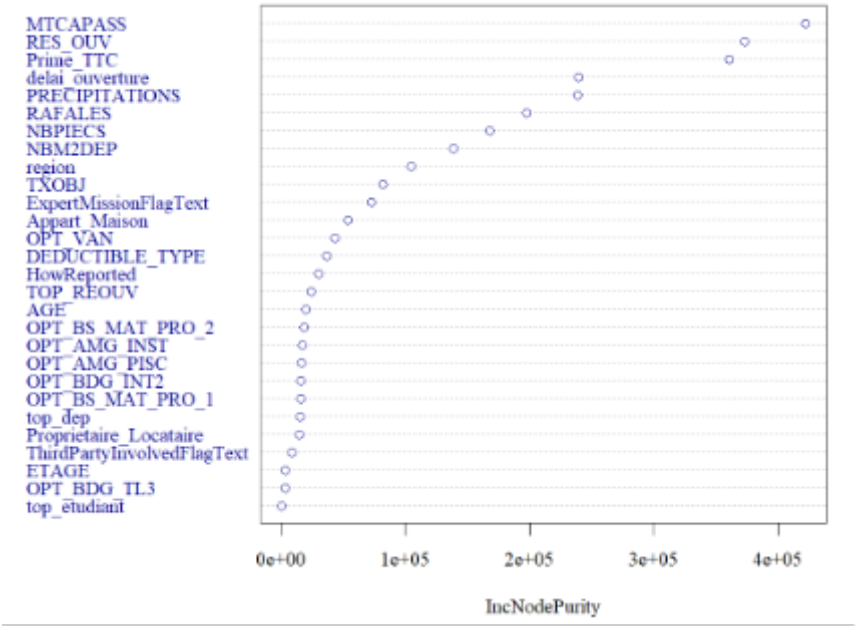


FIGURE 74 – Importance des variables dans le modèle *Random Forest* Grêle

Ainsi, d'après ce graphique, 15 premières variables les plus influentes sont :

Variables	Description
MTCAPASS	Montant assuré
RES_OUV	Réserve d'ouverture fixée par le gestionnaire
Prime_TTC	La prime payée
PRECIPITATIONS	La somme des précipitations du lieu sinistré
delai_ouverture	Temps entre la survenance et la déclaration
NBPIECS	Le nombre de pièces du bien
RAFALES	Les rafales de vent en km/h enregistrées
NBM2DEP	La superficie de la dépendance
region	La région où se situe le contrat sinistré
TXOBJ	Objets de valeurs présent dans l'habitation
ExpertMissionFlagText	Expert missionné ou non
Appart_Maison	Appartement vs Maison
OPT_VAN	Option remplacement valeur à neuf
DEDUCTIBLE_TYPE	Type de franchise
HowReported	Type de déclaration

TABLE 28 – Liste des 15 variables retenues pour la modélisation de la charge des sinistres des évènements grêle

De même, le graphe ci-dessous représente l'importance des variables dans le modèle des évènements hors grêle.

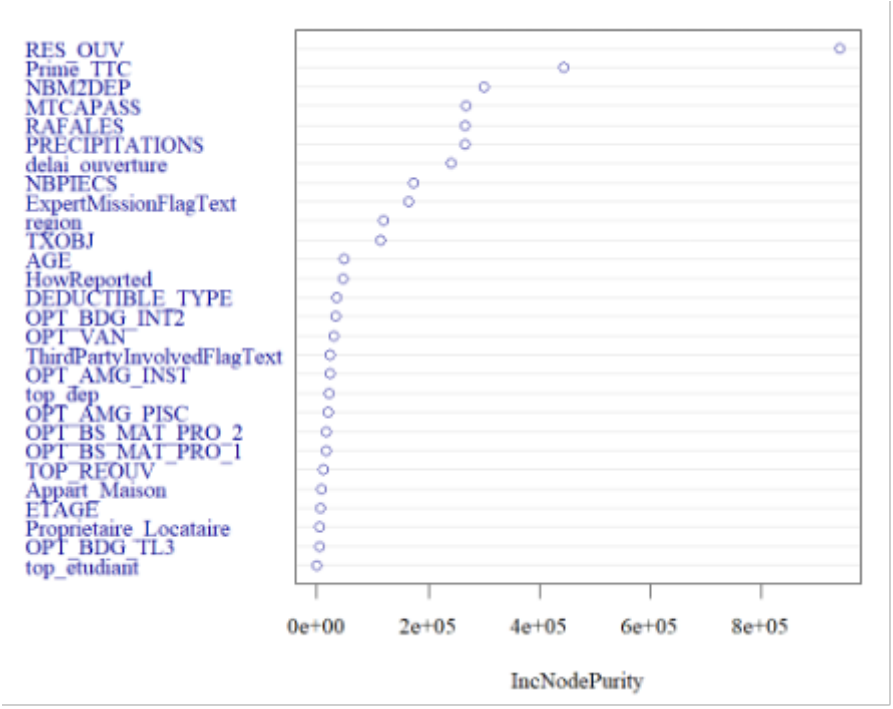


FIGURE 75 – Importance des variables dans le modèle *Random Forest* Hors Grêle

Selon le même raisonnement, les 15 premières variables les plus influentes sont :

Variables	Description
RES_OUV	Réserve d'ouverture fixée par le gestionnaire
Prime_TTC	La prime payée
NBM2DEP	La superficie de la dépendance
PRECIPITATIONS	La somme des précipitations du lieu sinistré
RAFALES	Les rafales de vent en km/h enregistrées
MTCAPASS	Montant assuré
delai_ouverture	Temps entre la survenance et la déclaration
ExpertMissionFlagText	Expert missionné ou non
NBPIECS	Le nombre de pièces du bien
region	La région où se situe le contrat sinistré
TXOBJ	Objets de valeurs présent dans l'habitation
AGE	Ancienneté du bien
HowReported	Type de déclaration
OPT_BDG_INT2	Option bris de glace
DEDUCTIBLE_TYPE	Type de franchise

TABLE 29 – Liste des 15 variables retenues pour la modélisation de la charge des sinistres des évènements hors grêle

III.2.K Modèle *CART*

Dans le but d'avoir une idée générale du modèle et de valider les variables retenues, un premier modèle *CART* est proposé. Le graphe ci-dessous montre la représentation graphique de l'arbre :

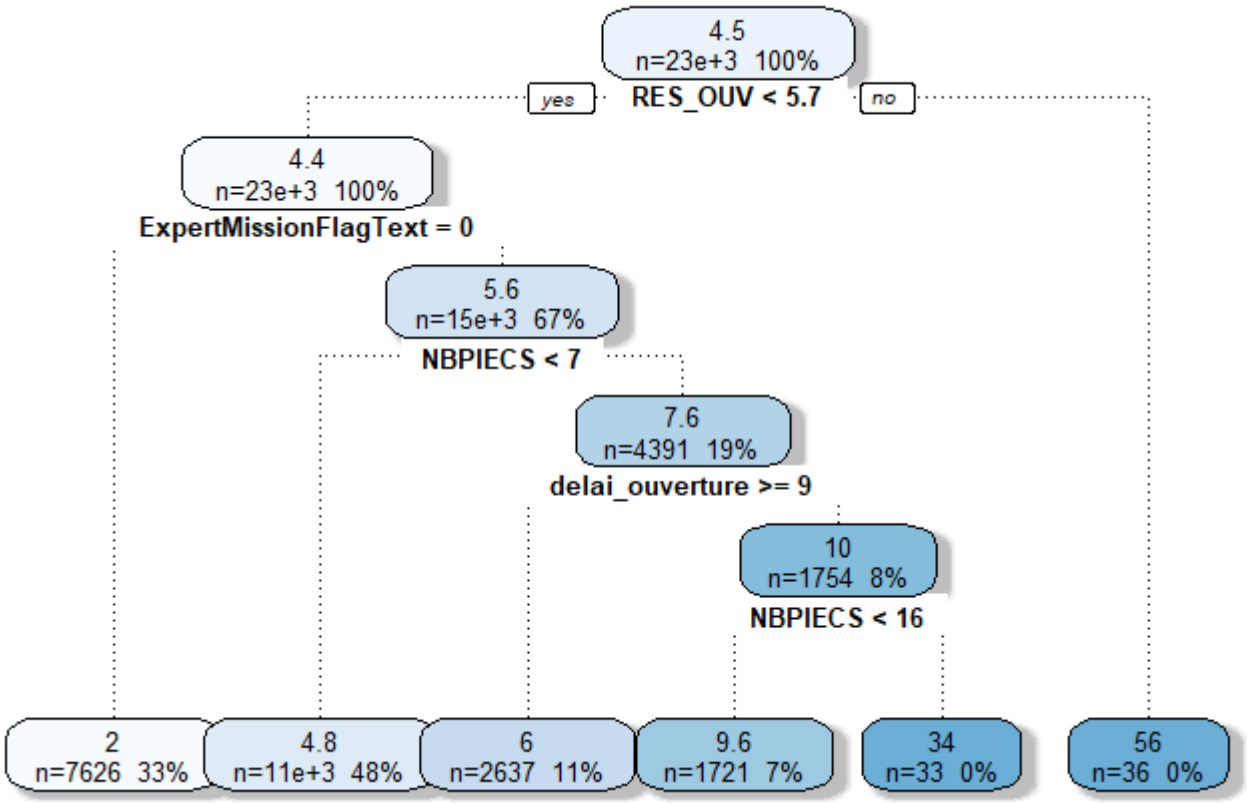


FIGURE 76 – Arbre *CART* pour le modèle Grêle

Il est clair que la variable réserve d'ouverture « RES_OUV » semble être la plus importante. En effet, il a été démontré dans la partie statistique descriptive que cette variable était la variable la plus corrélée avec la charge des sinistres. En deuxième et troisième place on retrouve respectivement les variables « ExpertMissionFlagText » et « NB_PIECS ». De même, il a été démontré que la variable catégorielle la plus corrélée avec la charge des sinistres est la variable présence d'un expert missionné « ExpertMissionFlagText ».

Cependant, la représentation graphique de l'arbre montre que certaines branches sont plus homogènes que d'autres. Cela peut être lié à des cas extrêmes.

L'arbre *CART* concernant les évènements hors grêle est représenté ci-dessous :

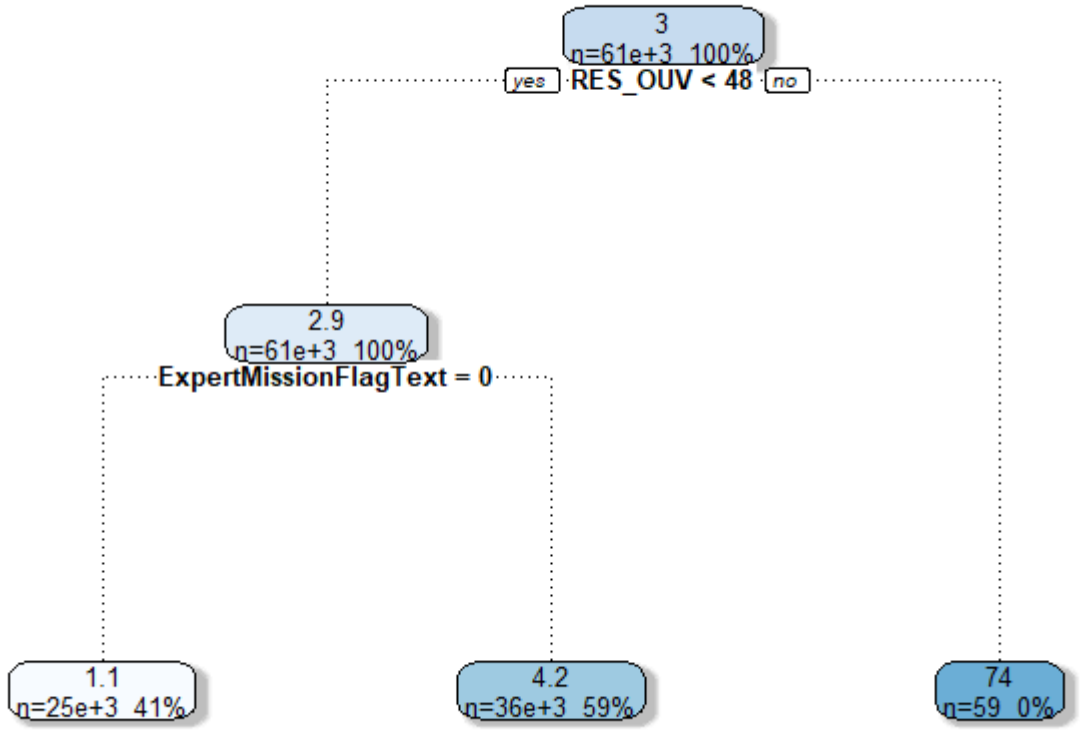


FIGURE 77 – Arbre *CART* pour le modèle Hors Grêle

La variable réserve d'ouverture « RES_OUV » reste la plus influente sur la charge des sinistres.

Ainsi, ces deux modèles *CART* confirment en partie la pertinence des variables sélectionnées pour la modélisation.

Une première estimation sur l'échantillon test peut être réalisée à l'aide de ces modèles :

Modèle	Charge estimée (en K€)	Charge observée (en K€)
Grêle	25 440,05	24 629,88
Hors grêle	45 475,44	45 727,87

TABLE 30 – Résultat des prédictions des modèles *CART* (grêle et hors grêle)

Le modèle dédié aux sinistres liés à la grêle prédit une charge de 25 440,05 K€ contre une valeur réelle de 24 629,88 K€ et le modèle hors grêle estime 45 475,44 K€ pour une valeur observée de 45 727,87 K€.
Ainsi tout sinistre confondu de l'échantillon test, 70 915,49€ sont prédit contre une valeur réelle de 70 357,75€.
Cela montre que le modèle parvient à estimer une charge globale relativement proche des

valeurs réelles.

III.2.L Modélisation des évènements grêle

Comme expliqué précédemment, pour chacun des modèles grêle et hors grêle, deux sous modèles seront entrainés. Le premier, en considérant la charge des sinistres « *as if* » CM J+7 et le deuxième en considérant la charge « *as if* » FFB/BT.

La base des sinistres grêles est constituée de :

- 30 206 sinistres dans la base d'apprentissage avec une charge associée de :
  - 142 416,80K € pour la charge chg_asif_BT
  - 193 891,20K € pour la charge chg_asif_CM
- 12 839 sinistres dans la base test avec une charge associée de :
  - 34 032,39K € pour la charge chg_asif_BT
  - 46 484,41K € pour la charge chg_asif_CM
- 21 425 sinistres dans la base de validation contenant les sinistres de 2022 avec une charge associée de 173 902,49K €.

Que ce soit pour l'échantillon d'entraînement ou l'échantillon test, les charges « *as if* » sont différentes selon la méthode utilisée. En effet, cela est dû à la différence entre les indices obtenus dans la section III.2.G.

La première modélisation pour les évènements grêle est la charge « *as if* » CM J+7.

Par défaut et pour avoir un modèle de base, il est toujours recommandé de débuter par un premier modèle « par défaut », c'est-à-dire avec les paramètres du modèle *XGBoost* par défaut :

Hyperparamètre	Description	Valeur
nrounds	Le nombre total d'arbres à agréger	1 000
gamma	Le seuil de gain minimal pour diviser un nœud	0
alpha	La régularisation L1 (Lasso)	0
Max_depth	Le nombre de feuilles maximales des arbres	6
Eta	Le taux d'apprentissage qui ajuste l'impact de chaque arbre sur la prédiction finale	0,3
subsample	La proportion des observations utilisées pour l'entraînement de chaque arbre	1
min_child_weight	Le nombre minimal d'observations dans un nœud	1
colsample_bytree	La proportion des variables explicatives sélectionnées pour entraîner chaque arbre	1
eval_metric	La métrique d'évaluation	*RMSE*

TABLE 31 – Liste des hyperparamètres du modèle grêle « *as if* » CM J+7 par défaut

Ainsi, un premier modèle est entraîné avec les hyperparamètres ci-dessus.

Le choix du nombre d'arbres dans un modèle *XGBoost* est essentiel pour éviter le phénomène de sur-apprentissage. En général, ce paramètre est ajusté en analysant l'évolution du *RMSE*, qui mesure l'écart entre les valeurs observées et celles prédites par le modèle.

Lors de l'entraînement, le *RMSE* est évalué à la fois sur l'échantillon d'apprentissage et sur l'échantillon de test. Tant que le *RMSE* du jeu de validation se stabilise ou diminue, le modèle reste performant. En revanche, si elle commence à augmenter alors que le *RMSE* du jeu d'apprentissage continue de baisser, cela indique un sur-ajustement : le modèle devient trop spécifique aux données d'entraînement et perd en capacité de généralisation.

Pour limiter ce phénomène, plusieurs stratégies peuvent être mises en place. Une approche consiste à ajuster dynamiquement le pourcentage d'observations et de variables utilisées à chaque itération, ce qui permet d'introduire plus de diversité dans les données d'entraînement. Une autre méthode repose sur une recherche systématique des hyperparamètres, en testant différentes combinaisons et en sélectionnant celle qui offre la meilleure performance selon un critère comme le *RMSE*.

Dans le cadre de cette étude, une analyse empirique a permis d'identifier le seuil optimal du nombre d'arbres à partir duquel le sur-apprentissage devient notable. Cette identification repose sur l'observation du point où le *RMSE* du jeu de test cesse de s'améliorer et commence à croître, tandis que celle du jeu d'entraînement poursuit sa diminution. L'ajustement du nombre d'itérations est donc un élément clé pour obtenir un modèle performant et robuste face à de nouvelles données.

Le schéma ci-dessous illustre cette explication :

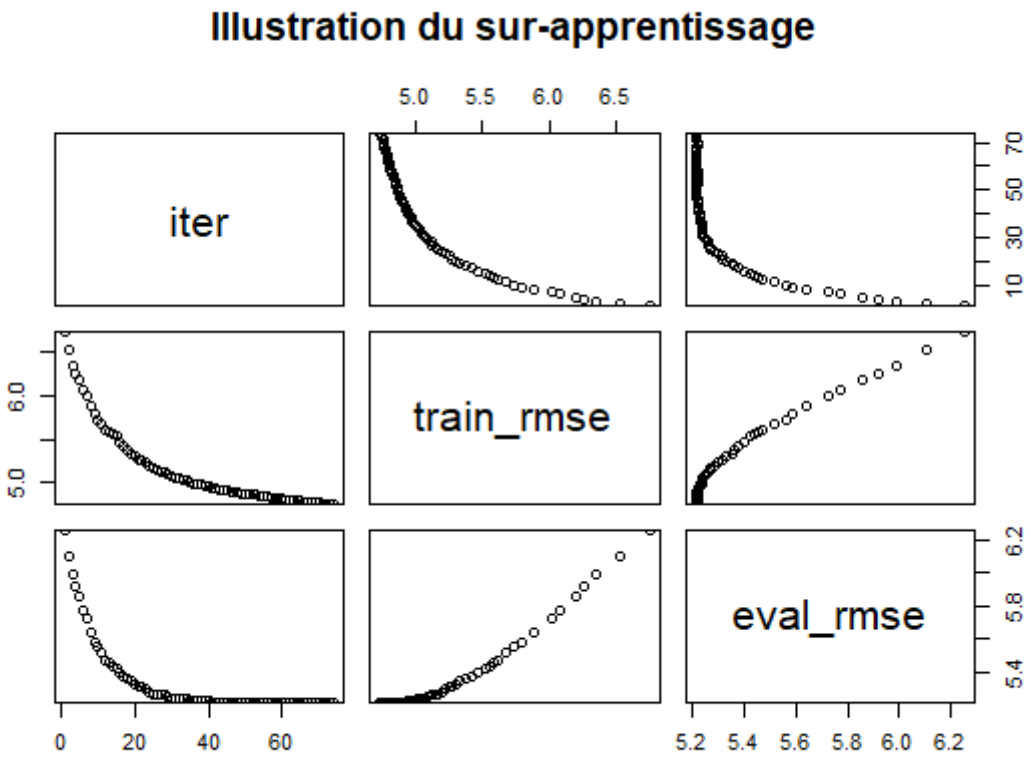


FIGURE 78 – Illustration du sur-apprentissage pour le modèle grêle de la charge « *as if* » CM J+7

Au-delà de 10 arbres, le *RMSE* sur l'échantillon d'entraînement poursuit sa diminution, tandis que celui sur l'échantillon de test tend à stagner, voire à légèrement augmenter. Par conséquent, augmenter le nombre d'arbres au-delà de ce seuil dans le modèle *XGBoost* risque soit de conduire à un sur-apprentissage, soit de n'apporter qu'un gain négligeable en termes de performance.

Ainsi, en appliquant le nombre d'arbres retenu au modèle par défaut, le modèle entrainé sur la base d'entrainement prédit 54 783,65K € de charges sur la base test comparé à 46 484,41K € en réalité aboutissant ainsi à un écart de 8 299,24K €.

Le *RMSE* est de 18,90 et l'erreur quadratique moyenne *MSE* de 356,10. Cependant, ce modèle est « naïf » est doit être optimisé via le *tuning* des hyperparamètres afin de diminuer d'avantage cet écart.

Le calibrage d'un modèle *XGBoost* consiste à optimiser simultanément plusieurs hyperparamètres en fonction d'un critère de performance, ici le *RMSE*. Ces hyperparamètres étant interconnectés, il est essentiel de ne pas les ajuster séparément, mais plutôt de les optimiser ensemble. Par exemple, le nombre d'arbres (*nrounds*) est fortement lié au taux d'apprentissage (*eta*), et un mauvais réglage de l'un peut impacter négativement l'autre.

Pour effectuer cette optimisation, le package *caret* du logiciel R permet de tester automati-

quement plusieurs combinaisons d'hyperparamètres en définissant une grille de recherche. De plus, pour éviter le surajustement, la validation croisée ou *cross-validation* en anglais, est utilisée afin d'évaluer la robustesse du modèle sur plusieurs sous-échantillons du jeu de données. Cela permet de sélectionner la meilleure combinaison d'hyperparamètres en garantissant une bonne généralisation du modèle sur de nouvelles données.

La grille suivante est proposée :

Hyperparamètre	Valeur
nrounds	→ 10
gamma	→ 0
Max_depth	→ 3 → 5 → 8 → 10
Eta	→ 0,01 → 0,1 → 0,15
subsample	→ 0,5 → 1
min_child_weight	→ 1 → 5 → 15
colsample_bytree	→ 0,7 → 1

TABLE 32 – Grille d'hyperparamètres pour le calibrage du modèle grêle avec la charge « *as if* » CM J+7

Pour chaque combinaison des 144 possibles, le modèle s'entraîne sur la base d'entraînement et calcule un *RMSE* sur la base de test. Le résultat des *RMSE* en fonction de chaque combinaison possible d'hyperparamètres est illustré dans le graphique ci-dessous :

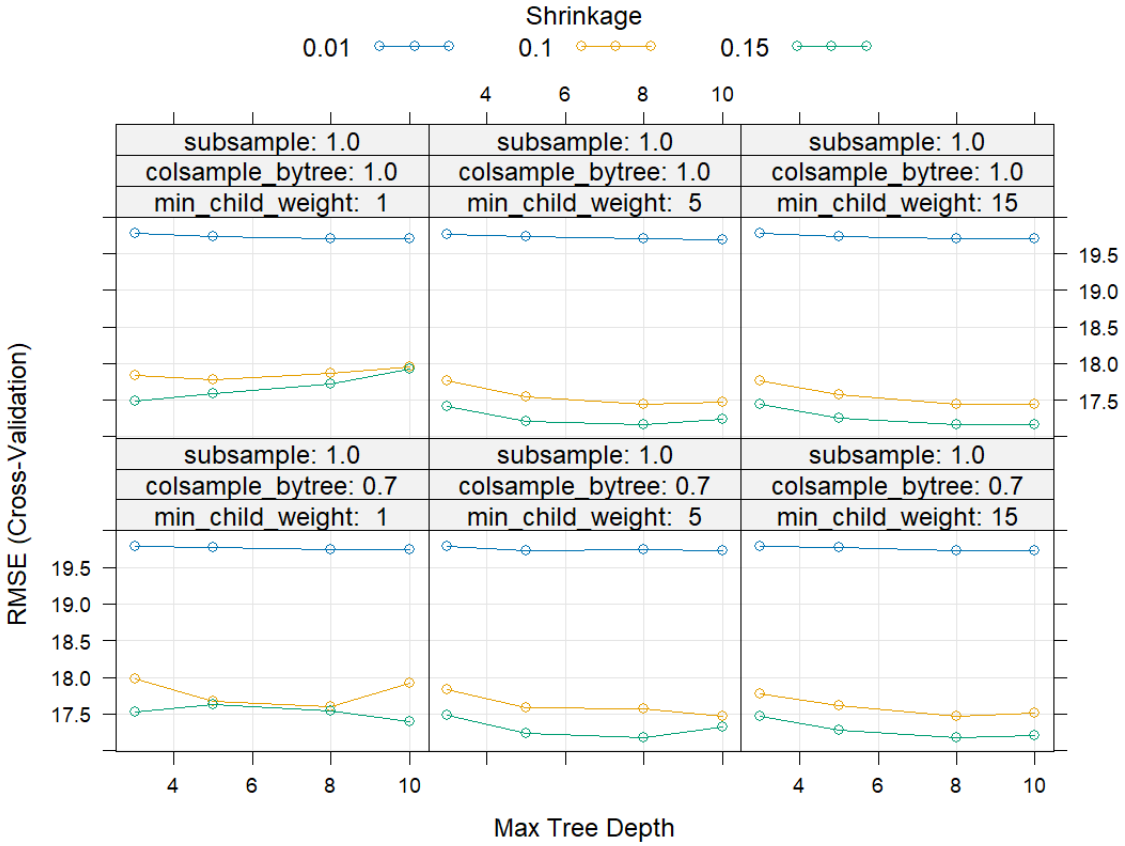


FIGURE 79 – Résultats du *RMSE* en fonction des combinaison d'hyperparamètres pour le modèle grêle « *as if* » CM J+7

Chaque courbe en couleur représente une valeur d'*eta* et chaque carré représente une combinaison de (*subsample*, *colsample_bytree*, *min_child_weight*) en fonction de *max_tree_depth*.

La combinaison d'hyperparamètres minimisant le *RMSE* correspond à :

Hyperparamètre	Valeur
nrounds	10
gamma	0
Max_depth	8
Eta	0,15
subsample	1
min_child_weight	5
colsample_bytree	1

TABLE 33 – Hyperparamètres optimaux selon le critère de minimisation du *RMSE* pour le modèle grêle « *as if* » CM J+7

La mesure de l'erreur quadratique moyenne obtenue avec ces hyperparamètres est de 280. Elle est inférieure à celle obtenue avec les paramètres par défaut (356,10), ce qui souligne l'importance du calibrage des hyperparamètres.

Pour le reste des trois modèles, la même méthodologie est abordée.

Le deuxième modèle des événements grêle correspond à la modélisation de la charge « *as if* » BT.

Le graphique ci-dessous illustre la variation du *RMSE* en fonction du nombre d'arbres. La courbe bleue représente le *RMSE* sur l'échantillon d'entraînement et la courbe rouge représente celui sur l'échantillon de test.

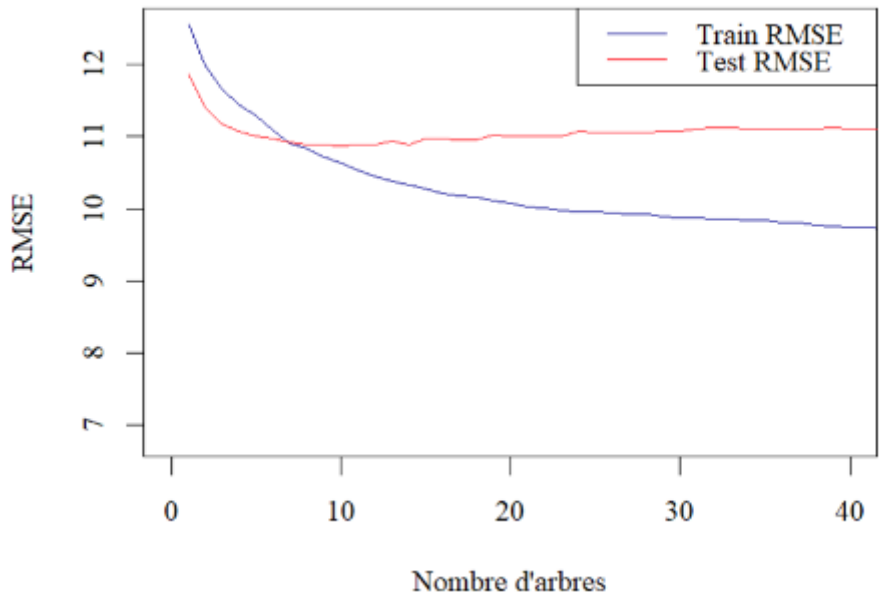


FIGURE 80 – Illustration du sur-apprentissage pour le modèle grêle de la charge « *as if* » BT

A partir de 10 arbres, le *RMSE* sur l'échantillon d'entraînement continue à diminuer alors que sur l'échantillon de test, le *RMSE* continue légèrement d'augmenter voire de stagner. Ainsi, augmenter le nombre d'arbres au-delà de 10 dans le modèle *XGBoost* risque soit de provoquer du sur-apprentissage, soit de n'apporter aucune amélioration significative aux performances.
Le modèle *XGBoost* avec les hyperparamètres par défaut et le nombre d'arbres à 10 prédit 35 629,52K € de charges (chg_asif_BT) sur la base de test comparé à 34 032,39K € de charge en réalité donnant ainsi un écart de 1 597,13K €. Le *RMSE* est de 18,31 et le *MSE* de 335,59.

Les différentes combinaisons possibles d'hyperparamètres sont testées, selon la grille d'hyperparamètre présentée dans le tableau de la table 32. Les résultats du *RMSE* en fonction des combinaisons d'hyperparamètres sont illustrées dans le graphique ci-dessous :

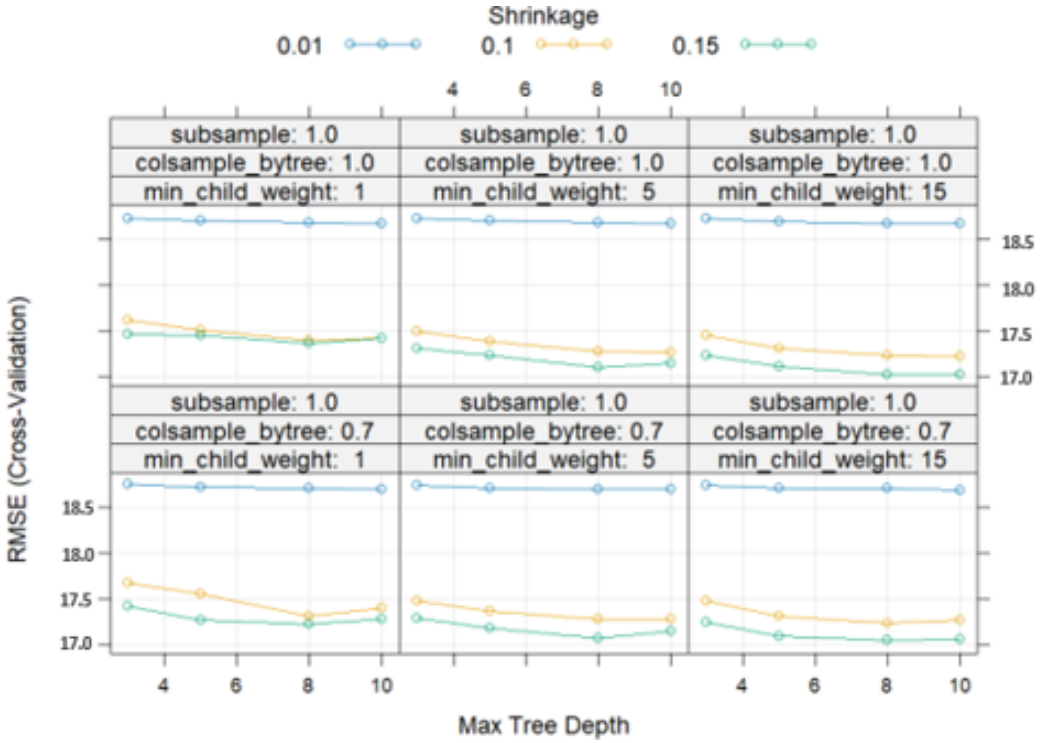


FIGURE 81 – Résultats du *RMSE* en fonction des combinaison d'hyperparamètres pour le modèle grêle « *as if* » BT

Le modèle optimal selon la minimisation du critère de *RMSE* est le modèle *XGBoost* avec les hyperparamètres suivants :

Hyperparamètre	Valeur
nrounds	10
gamma	0
Max_depth	10
Eta	0,15
subsample	1
min_child_weight	15
colsample_bytree	1

TABLE 34 – Hyperparamètres optimaux selon le critère de minimisation du *RMSE* pour le modèle grêle

De même, la mesure de l'erreur quadratique moyenne obtenue avec ces hyperparamètres est de 286. Cette valeur est inférieure à celle obtenue avec les paramètres par défaut 335,59, ce qui souligne l'importance du calibrage des hyperparamètres.

III.2.M Modélisation des évènements hors grêle

Cette partie est dédiée à la modélisation des évènements hors grêle. Suivant la même méthodologie, d'une part, la charge des sinistres « *as if* » CM est modélisée, et d'autre part la charge « *as if* » FFB.

La composition de la base des évènements hors grêle est la suivante :

- 68 461 sinistres dans la base d'apprentissage avec une charge associée de
  - 212 212,20K € pour la charge chg_asif_FFB
  - 200 006,70K € pour la charge chg_asif_CM
- 22 404 sinistres dans la base test avec une charge associée de
  - 52 784,14K € pour la charge chg_asif_FFB
  - 49 687,19K € pour la charge chg_asif_CM
- 15 294 sinistres dans la base de validation contenant les sinistres de 2022 avec une charge associée de 43 105,50K €.

La première charge « *as if* » testée est la charge « *as if* » CM J+7.

Comme déjà présenté, il est important de se baser tout d'abord sur un modèle dit modèle « par défaut » avec des hyperparamètres fixés. La même liste d'hyperparamètres que celle des modèles grêle est fixée pour les deux modèles hors grêle qui suivent.

Le nombre d'arbres retenus pour éviter le sur-apprentissage est de 24.

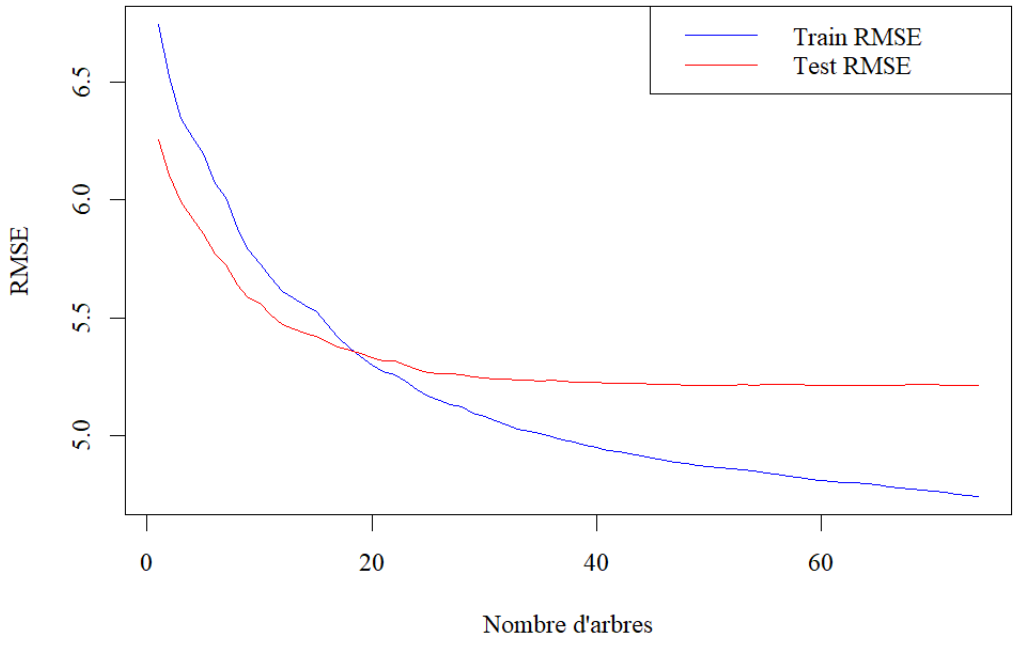


FIGURE 82 – Illustration du sur-apprentissage pour le modèle hors grêle de la charge « *as if* » CM J+7

Le modèle *XGBoost* par défaut prédit 49 618,03K € de charges sur la base test comparé à 49 687,19K € en réalité donnant ainsi un écart de 69,16. Le *RMSE* est de

11,31 et le *MSE* de 128,04.

Pour le calibrage du modèle, voici la grille d'hyperparamètres proposée :

Hyperparamètre	Valeur
nrounds	→ 24
gamma	→ 0 → 0,1 → 0,9
Max_depth	→ 3 → 6 → 10
Eta	→ 0,01 → 0,1 → 0,3
subsample	→ 0,7
min_child_weight	→ 1
colsample_bytree	→ 0,5 → 0,8 → 0,9

TABLE 35 – Grille d'hyperparamètres pour le calibrage du modèle hors grêle avec la charge « *as if* » CM J+7

81 combinaisons d'hyperparamètres est possible. Le résultat des *RMSE* en fonction de chaque combinaison possible d'hyperparamètre est illustré dans le graphique ci-dessous :

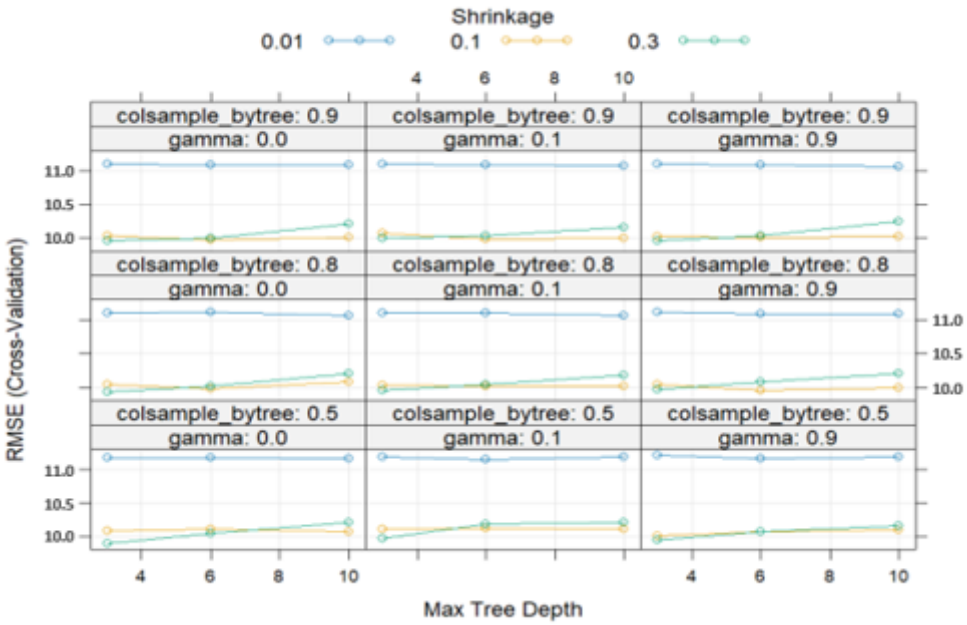


FIGURE 83 – Résultats du *RMSE* en fonction des combinaison d'hyperparamètres pour le modèle hors grêle « *as if* » CM J+7

Les combinaisons sont légèrement différentes que celles des deux modèles grêle. Chaque

courbe en couleur représente une valeur d'*eta* et dans ce cas, chaque carré représente une combinaison de (*colsample_ bytree*, *gamma*) en fonction de *max_ tree_ depth*.
Selon le critère de minimisation du *RMSE*, le modèle retenu est celui avec les hyperparamètres suivants :

Hyperparamètre	Valeur
`nrounds`	24
`gamma`	0,1
`Max_depth`	6
`Eta`	0,1
`subsample`	0,7
`min_child_weight`	1
`colsample_bytree`	0,5

TABLE 36 – Hyperparamètres optimaux selon le critère de minimisation du *RMSE* pour le modèle hors grêle « *as if* » CM J+7

La mesure de la racine de l'erreur quadratique moyenne obtenue avec ces hyperparamètres est de 9,165 et est inférieure à celle obtenue avec les paramètres par défaut 11,31, soulignant encore une fois l'importance du calibrage des hyperparamètres.

Enfin, le dernier modèle *XGBoost* est celui appliqué à la charge des sinistres « *as if* » FFB.

Le graphique ci-dessous montre l'évolution du *RMSE* de la base d'entraînement et de la base de test en fonction du nombre d'arbre :

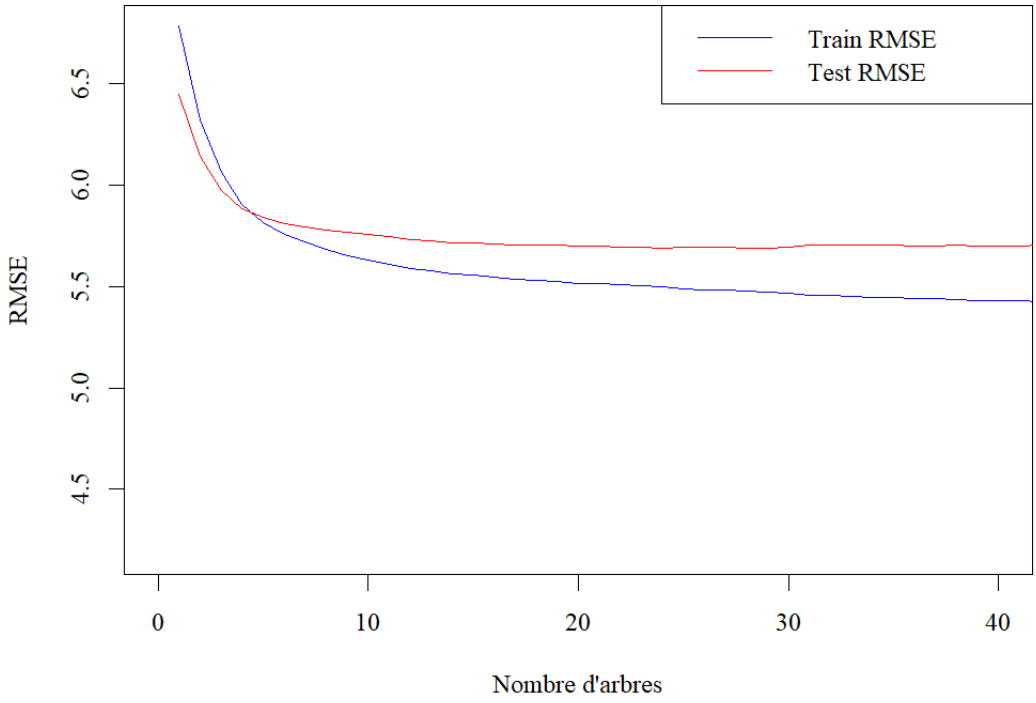


FIGURE 84 – Illustration du sur-apprentissage pour le modèle hors grêle de la charge « as if » FFB

Afin d'éviter le sur-apprentissage, il s'avère que 24 soit le nombre d'arbres optimal.
En partant toujours d'un modèle par défaut en appliquant le nombre d'arbre optimal retenu, le modèle prédit 52 391,46K € contre 52 784,14K € en réalité. L'écart ainsi observé est de 39,68K € avec un RMSE de 39,68 et un *MSE* de 135,59.

La dernière étape consiste à calibrer le modèle suivant la même grille de d'hyperparamètres que précédemment. Le résultat de la validation croisée suivant les différentes combinaisons possible d'hyperparamètres est présentée dans la figure ci-dessous :

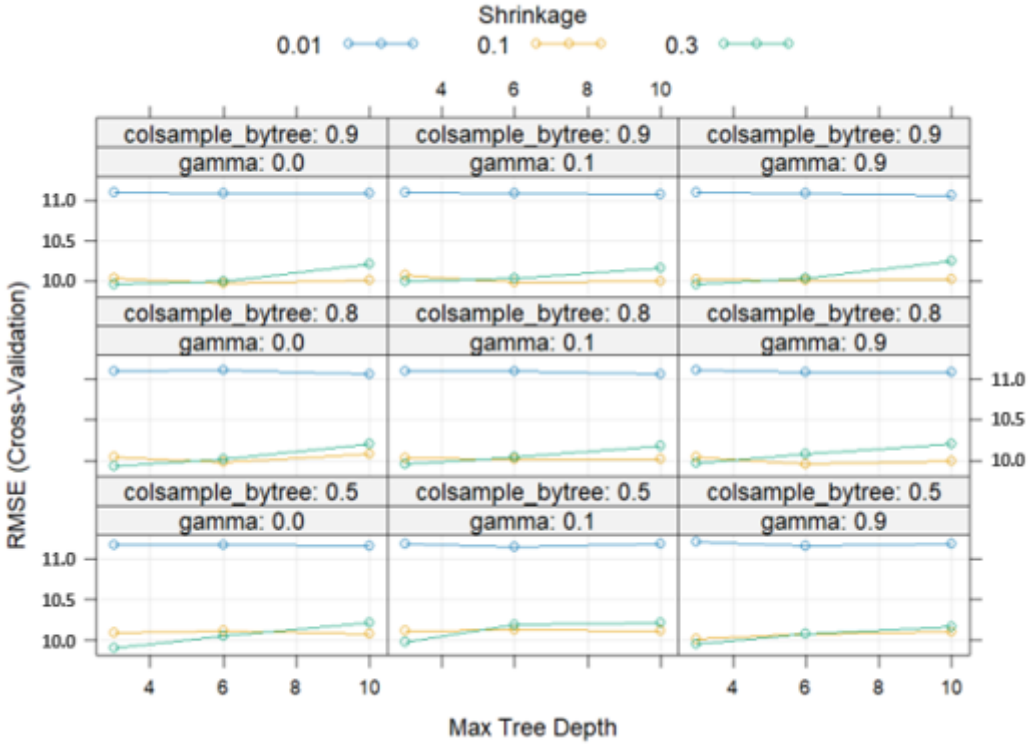


FIGURE 85 – Résultats du *RMSE* en fonction des combinaison d'hyperparamètres pour le modèle hors grêle « *as if* » FFB

Ainsi, les hyperparamètres retenus sont :

Hyperparamètre	Valeur
nrounds	24
gamma	0
Max_depth	3
Eta	0,3
subsample	0,7
min_child_weight	1
colsample_bytree	0,5

TABLE 37 – Hyperparamètres optimaux selon le critère de minimisation du *RMSE* pour le modèle hors grêle « *as if* » FFB

La mesure de l'erreur quadratique moyenne obtenue avec ces hyperparamètres est de 84,8 et est inférieure à celle obtenue avec les paramètres par défaut 135,59.

III.2.N Choix de l'*as if*

À ce stade, pour chacun des deux modèles, événements de grêle et événements hors grêle, deux sous-modèles sont proposés :

- L'un avec une charge à laquelle on applique un « *as if* » de CM J+7.
- L'autre avec une charge à laquelle on applique un « *as if* » d'indice FFB ou d'indice BT.

Pour déterminer lequel retenir, chaque modèle sera utilisé pour prédire la charge sur la base de validation. En effet, l'objectif du « *as if* » appliqué à la charge est d'aligner au mieux la charge passée sur celle de 2022. Par conséquent, valider les modèles sur un échantillon test basé sur des années antérieures à 2022 ne serait pas optimal, d'où l'intérêt de prédire la charge sur la base des évènements de 2022.

Le tableau ci-dessous présente, pour chaque modèle avec les paramètres optimaux choisis à l'étape précédente, la prédiction de la charge en 2022 ainsi que le *RMSE* associé :

Modèle	« *As if* »	Valeurs prédites en (K€)	Valeurs observées (en K€)	*RMSE*
Grêle	**CM J+7**	175 127,00	173 902,49	16,74
	Indice BT	154 341,90		16,95
Hors grêle	**CM J+7**	43 130,14	43 105,50	9,17
	Indice FFB	48 403,30		9,21

TABLE 38 – Prédiction de la charge ultime des quatre modèles *XGBoost* sur la base de validation 2022

Les prédictions sur l'année 2022 de la base de validation des modèles sur les charges « *as if* » de CM J+7 semblent être les plus proches de la charge réellement observée. Les quatre valeurs de *RMSE* sont du même ordre deux à deux mais ceux des modèles « *as if* » de CM restent légèrement inférieurs à ceux de l'indice FFB/BT.

Même en utilisant un indice spécifique en MRH, et plus particulièrement adapté à la construction, des éléments susceptibles d'être impactés par des événements climatiques, qu'il s'agisse de grêle ou d'autres évènements (hors grêle), il reste impossible d'atteindre les coûts moyens de 2022, ce qui souligne le caractère atypique de cette année.

Ainsi, la méthode du coût moyen parvient à capturer l'inflation imprévisible liée aux épisodes de grêle en 2022, là où même les indices FFB et BT, pourtant les plus spécialisés, ne parviennent pas à l'expliquer.

III.2.O Validation des modèles

La dernière étape de la modélisation consiste à valider le modèle notamment en testant la qualité des prédictions.

Pour déterminer l'erreur commise par le modèle, une analyse des résidus est établie (différence entre la charge réelle et la charge prédite). Les prédictions sont faites sur la base de test et sont ensuite comparées aux valeurs réellement observées de cette base.

Le modèle retenu pour les événements de grêle estime une charge de 175 127,00K €, contre une valeur observée de 173 902,49K €, indiquant ainsi une légère surestimation. Le schéma ci-dessous illustre les prédictions (courbe rouge) et les valeurs observées (courbe bleue) sur des observations choisies aléatoirement :

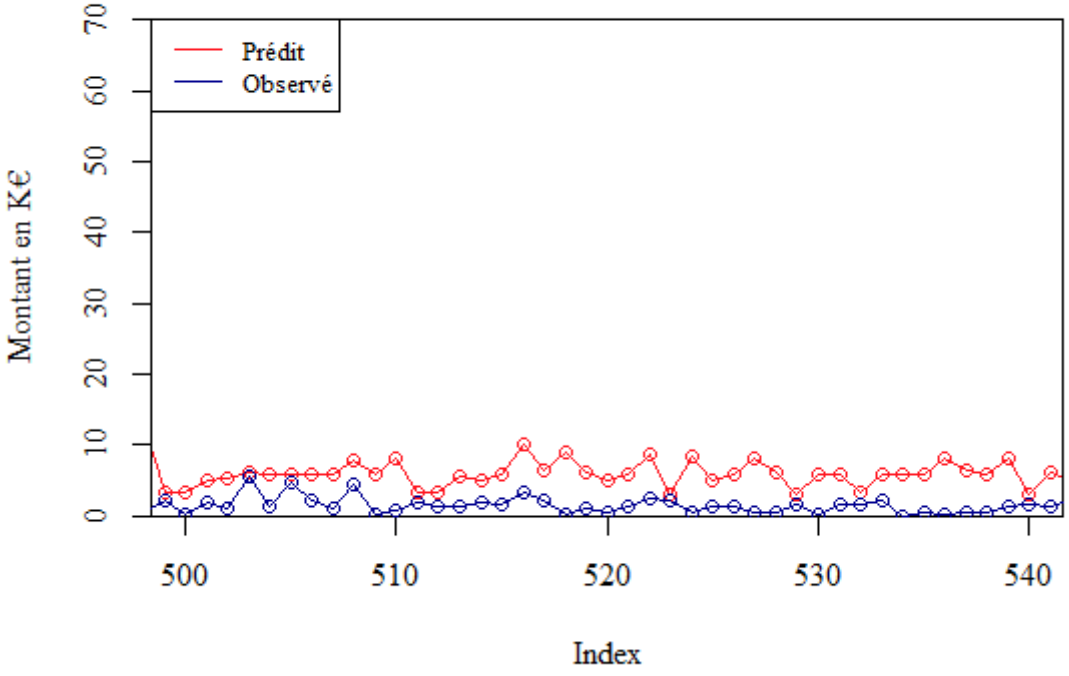


FIGURE 86 – Valeurs prédites et observées du modèle grêle retenu

Il est évident d'après le graphique ci-dessus que la prédiction du modèle ne correspond pas toujours exactement à la valeur observée. Sachant que l'écart entre la charge prédite et celle observée est de 1 224,51K €, la prédiction reste globalement proche du montant réel. C'est pourquoi, dans le cadre de l'analyse des résidus, il est pertinent d'examiner la boîte à moustaches des résidus.

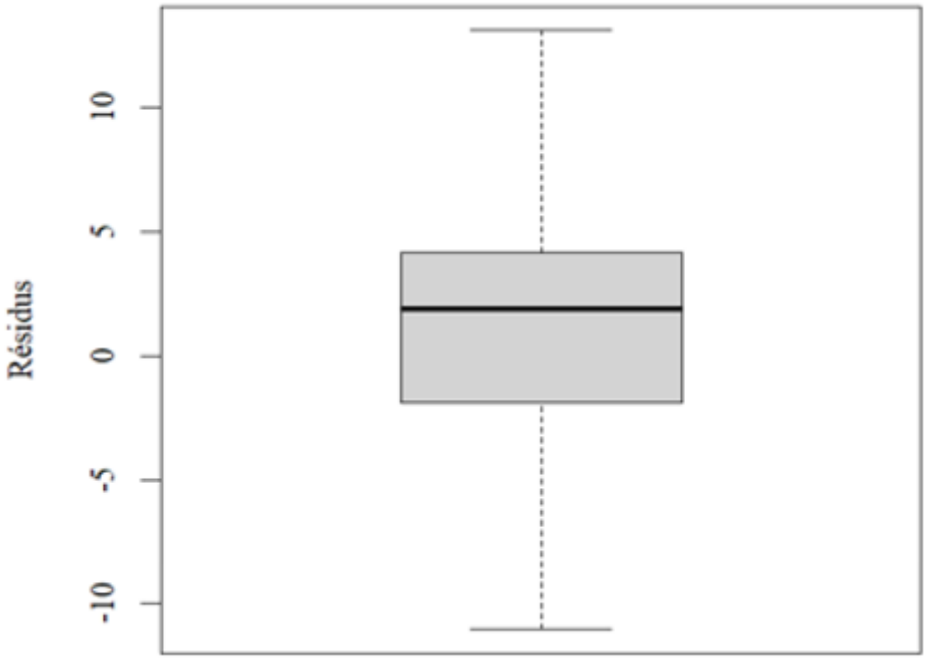


FIGURE 87 – Boite à moustache des résidus du modèle grêle retenu

La boîte à moustaches est concentrée entre -10 et +10, ce qui indique que la majorité des prédictions individuelles présentent un écart compris entre -10 000 € et +10 000 € par rapport aux valeurs observées, ce qui confirme les écarts observés sinistres par sinistres. De plus, la boîte à moustaches étant centrée autour de 0, cela signifie que, bien que le modèle puisse présenter des erreurs à l'échelle individuelle, la somme des prédictions agrégées reste globalement proche de la somme des observations réelles.

Dans la même optique d'analyse des résidus, les prédictions sont faites sur la base des évènements hors grêle via le modèle hors grêle retenu. Le modèle optimal prédit 43 130,14K € pour les évènements hors grêle contre une valeur observée de 43 105, 50K €.

Pour des observations aléatoirement choisies, le graphe ci-dessous illustre les prédictions en rouge et les valeurs observées en bleue :

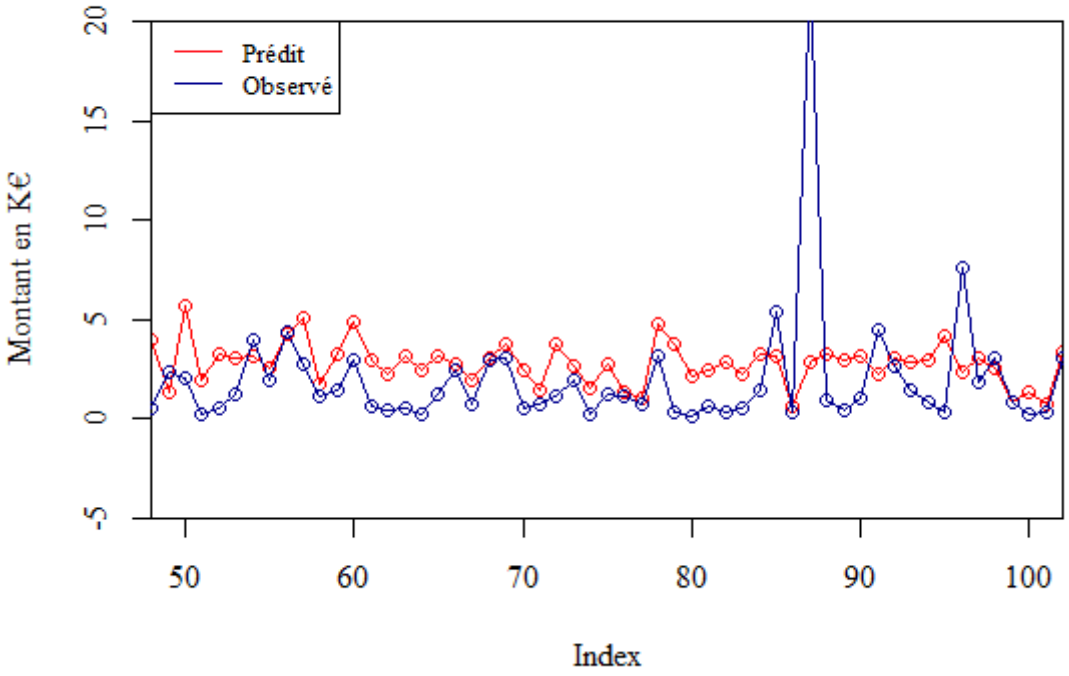


FIGURE 88 – Valeurs prédites et observées du modèle hors grêle retenu

La même conclusion sur les prédictions du modèle grêle peut être faite sur le modèle hors grêle : la prédiction ne coïncide pas parfaitement avec la valeur observée mais l'écart de 24,64K € entre la charge estimée et celle réellement constatée montre que le modèle prédit globalement très bien la charge.

Similairement, la boite à moustache des résidus est présentée ci-après :

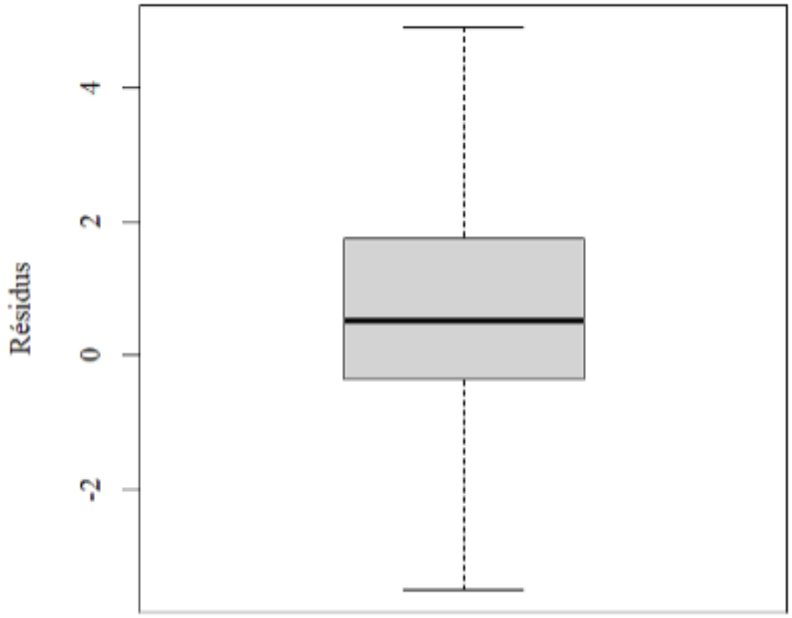


FIGURE 89 – Boite à moustache des résidus du modèle hors grêle retenu

La boite à moustaches est concentrée entre -2 et +4 : les écarts entre les valeurs observées et les valeurs prédites varient entre -2 000 € et 4 000 €. Ce montant reste élevé mais beaucoup plus faible que celui du modèle grêle. Comme pour le modèle grêle, la concentration en 0 permet de conclure aussi que la somme des prédictions agrégées est globalement proche des valeurs réelles.

Ainsi, la charge prédite ($\text{chg}_{\text{prédite}}$) pour l'ensemble des sinistres de 2022 est obtenue en additionnant les charges prédites des événements de grêle et hors grêle, estimées par leurs modèles respectifs.

La charge ultime estimée s'élève à 218 257,10K€.

III.2.P Estimation des sinistres clos sans suite et des sinistres tardifs

Les modèles individuels ont été entraînés sur des sinistres hors sinistres clos sans suite. Ainsi, leurs prédictions sur l'échantillon de validation 2022 ne tiennent pas compte de la probabilité qu'un sinistre soit clôturé sans suite au cours de l'année 2022. En d'autres termes, chaque prédiction suppose qu'un sinistre a donné lieu à une indemnisation.

Ainsi, il est essentiel d'évaluer la part des sinistres qui seront finalement clos sans suite dans la base de données. Une approche simplifiée consiste à utiliser le taux de sinistres clos sans suite estimé d'un évènement E, grâce à la méthode agrégée améliorée, noté $tx_{css}(E)$.

Connaissant le nombre total à date de sinistres en 2022 ($Nb_{total}\,(2022)$), le nombre de sinistres clos sans suite pour l'année 2022 ($Nb_{css}(2022)$) est obtenu comme suit :

$$Nb_{css(2022)} = tx_{css}(E) * Nb_{total}\,(2022)$$

Avec cette approche, on estime que 10 344 sinistres seront clos sans suite en 2022.
Pour évaluer la charge des sinistres clos sans suite, il faut estimer leur coût moyen, et cela à partir des prédictions du modèle sur la base de validation (CM_{2022}) .
Le montant total des sinistres clos sans suite est alors donné par :

$$cout_{css} = Nb_{css}\,(2022) * CM_{2022}$$

Ce montant est de 5 904 K€ et doit être retirer de la charge totale prédite sur la base 2022, pour éviter une surestimation liée à la non-prise en compte de la probabilité de clôture sans suite dans le modèle individuel.

De plus, à ce stade, la CFP de 2022 est estimée sans prendre en compte les sinistres tardifs, qui correspondent aux *IBNYR*. Ainsi, l'estimation actuelle ne tient compte que des sinistres déjà déclarés en 2022. Pour chaque sinistre connu dans la base de 2022, on associe une charge ultime définie comme :

$$charge_{ultime} = CFP - charge_{dossier/dossier}$$

Ainsi, il reste à déterminer les sinistres qui ne sont pas encore déclarés. La charge des sinistres qui ne sont pas déclarés est obtenue comme suit :

$$IBNYR = (NBAS\ ultime_{\text{modèle agrégé}} - NBAS_{2022}) * CM_{2022}$$

Tel que :

- $NBAS\ ultime_{\text{modèle agrégé}}$: correspond au NBAS ultime, obtenu grâce à la méthode agrégée améliorée.
- $NBAS_{2022}$: correspond au NBAS obtenu grâce à la formule suivante : $NBAS_{2022} = Nb_{total}\,(2022) - Nb_{css}(2022)$
- CM_{2022} : correspond au coût moyen des prédictions de la base de validation.

Le montant des sinistres tardifs estimé s'élève à 15 974,82K €. Ce montant est rajouté à la charge préalablement estimée.

Même si la base a été extraite en 2023 et que la plupart des sinistres de 2022 ont déjà été déclarés, il reste possible que de nouvelles ouvertures interviennent. En d'autres termes, même si la base est extraite en 2022, le nombre de sinistre clos sans suite ($NBAS\ ultime_{\text{modèle agrégé}} - NBAS_{2022}$) aurait été plus grand. C'est pourquoi l'ajout des IBNYR permet une estimation plus précise de la charge totale des sinistres.

À l'issue de ces ajustements, la charge globale des sinistres est estimée en intégrant :

- Les charges dossier/dossier
- Les réserves
- Les *IBNR* (*IBNER* + *IBNYR*)
- L'exclusion des sinistres clos sans suite

Donc, la charge ultime prédite y compris les tardifs et hors sinistres clos sans suite est donnée par :

$$charge_{ultime}\,(2022) = chg_{\text{prédite}}(2022) + IBNYR - cout_{css}$$

Sachant que la méthode individuelle prédit 218 257,10 K€., la charge ultime prédite y compris les tardifs et hors sinistres clos sans suite est de 228 331,90K € (=218 257,10 + 15 974,82 - 5 900K)K€.

Cette démarche souligne la complémentarité entre les méthodes individuelle et agrégée. En effet, la méthode agrégée joue un rôle clé dans l'estimation du nombre de sinistres clos sans suite et des sinistres tardifs, éléments qui ne sont pas directement pris en compte par les modèles individuels. Ainsi, l'approche combinée permet d'obtenir une estimation plus robuste et plus fiable de la charge des sinistres climatiques en assurance MRH.

III.2.Q Importance des variables

À la suite de l'identification des 15 variables les plus importantes grâce à l'algorithme *Random Forest*, le modèle *XGBoost* permet également de détecter les variables les plus influentes dans la modélisation.

Les deux graphiques ci-dessous illustrent, d'un côté l'importance des variables pour le modèle grêle et de l'autre côté l'importance des variables pour le modèle hors grêle.

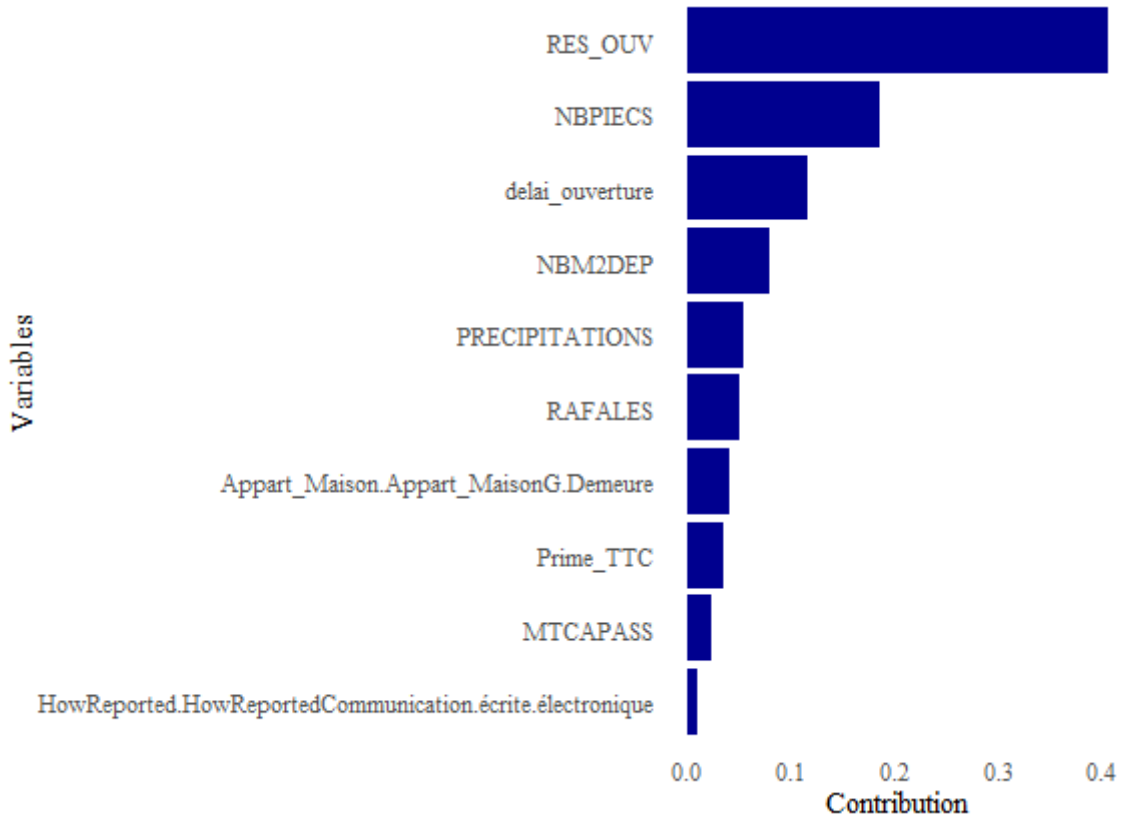


FIGURE 90 – Importance des variables - modèle grêle

On remarque que les trois variables les plus importantes dans la prédiction de la charge des sinistres grêle sont : la réserves d'ouverture, le nombre de pièce dans le bien et le délai d'ouverture.

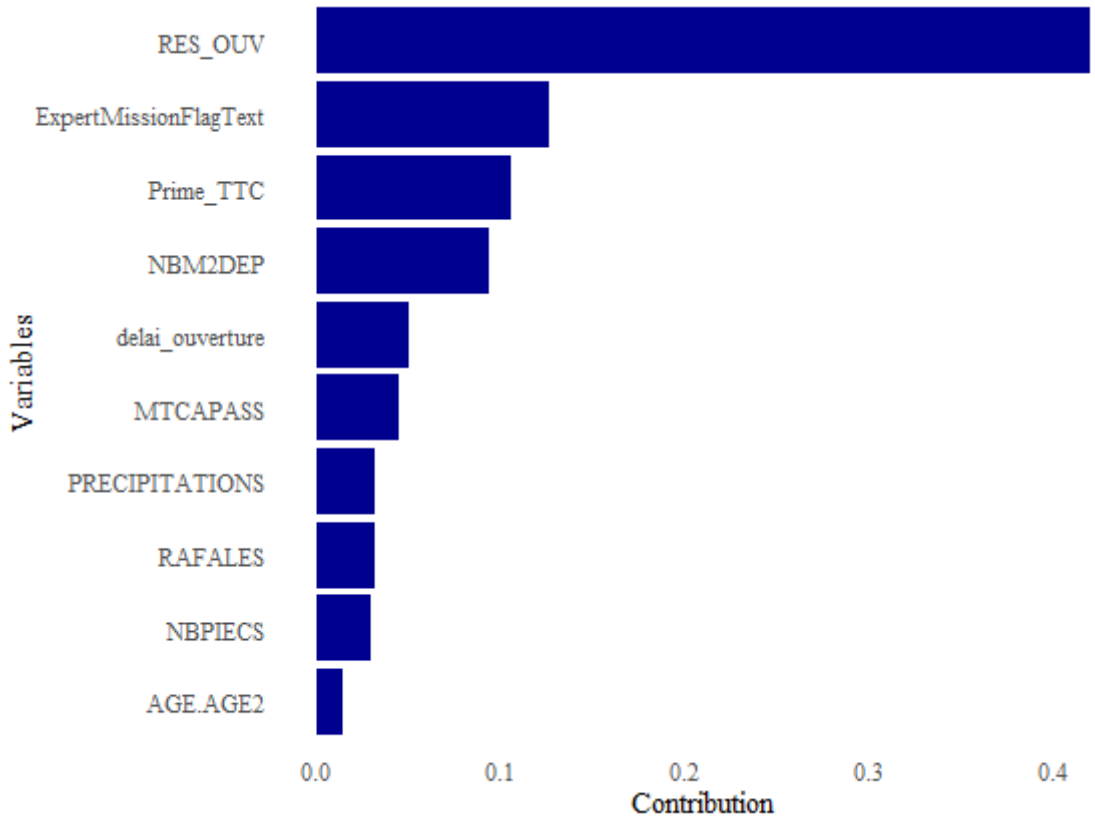


FIGURE 91 – Importance des variables - modèle hors grêle

Quant au modèle hors grêle, il est clair que les trois variables les plus importantes dans la prédiction de la charge des sinistres hors grêle sont : la réserves d'ouverture, la présence d'un expert et la prime.

Dans les deux modèles, la réserve d'ouverture semble être la variable la plus influente sur la prédiction de la charge des sinistres. Ainsi, la réserve d'ouverture joue un rôle central dans l'estimation de la charge ultime des sinistres. Son importance traduit la pertinence des premières évaluations réalisées lors de l'ouverture du sinistre, qui conditionnent l'évolution de la charge du sinistre. Cela incite à améliorer l'analyse des facteurs comptables, notamment la gestion des réserves, pour améliorer la précision des estimations et optimiser le provisionnement des sinistres.

Les variables exogènes sur le caractère du bien tel que le nombre de pièces et l'intervention d'un expert lors d'un sinistre semblent aussi avoir une forte importance sur la prédiction de la charge.

III.2.R Comparaison des méthodes

Une fois les modèles construits, la dernière étape consiste à identifier la méthode la plus pertinente en évaluant la minimisation de l'erreur de prédiction et la proximité avec la charge réellement observée. Pour se faire, une comparaison des prédictions sur l'année

2022 entre les modèles agrégés et individuels est effectuée.

Le tableau ci-dessous présente les résultats obtenus, en confrontant la charge prédite à la charge effectivement observée pour chaque modèle :

	Charge 2022 prédite (en K€)	**Charge 2022 observée (en K€)**
Méthode agrégée classique	265 817,05	217 007,99
Méthode agrégée améliorée	244 670,13	
Méthode individuelle	228 331,90	

TABLE 39 – Tableau comparatif des prédictions des différentes méthodes

On peut confirmer que la charge prédite par la méthode individuelle se rapproche le plus de la charge ultime observée de 2022. Même si cette méthode sur-estime la charge ultime, elle reste quand même la méthode la plus performante en terme de minimisation de l'écart entre la charge observée et la charge prédite.

Conclusion

L'objectif de ce mémoire est de provisionner, rapidement, rigoureusement et par évènement, les sinistres climatiques de la branche multirisques habitation. Ce mémoire propose des méthodes pour estimer rapidement la charge des événements climatiques de grande ampleur en proposant deux approches : une méthode agrégée par événement et une méthode individuelle.

Au sein d'AXA France, les méthodes de provisionnement utilisées pour ce type de sinistres reposent sur une approche budgétaire, consistant à établir un budget en fonction de l'historique des coûts et de la fréquence des sinistres climatiques. Cependant, en raison de la volatilité de cette branche, notamment en ce qui concerne les sinistres climatiques, ainsi que de l'augmentation exponentielle du nombre et du coût de ces événements, cette méthode est de plus en plus remise en question.

Ainsi, ce mémoire propose deux nouvelles approches : une approche agrégée par évènement et une approche individuelle.

L'approche agrégée par évènement est déjà mise en place chez AXA mais ce mémoire propose une amélioration de cette méthode. Sur une base de données d'évènements climatiques de 2021 et 2022, les deux méthodes agrégées estiment une charge assez proche de la charge observée. Bien que les deux modèles surestiment cette charge, la méthode basée sur les évènements améliore la prédiction de la charge par le modèle classique d'AXA.

Cela étant dit, la méthode améliorée par évènement proposée par ce mémoire est validée.

Outre la méthode agrégée améliorée, ce mémoire propose également une approche de provisionnement individuelle, basée sur des méthodes d'apprentissage statistique.

L'intérêt d'une telle approche est de prédire une variable cible à partir d'un ensemble de variables explicatives. Ici, la variable cible est la charge des sinistres et les variables explicatives regroupent toutes variables apportant des informations susceptibles d'améliorer la prédiction, qu'il s'agisse de caractéristiques liées aux sinistres, aux contrats ou encore de facteurs exogènes.

Les méthodes d'apprentissage statistique nécessitent un historique de données suffisamment profond. Un travail conséquent a été mené pour la construction et le nettoyage de la base de données, car les informations disponibles ne sont jamais parfaitement structurées. Les sinistres climatiques survenus entre 2014 et 2022 ont ainsi été recensés et consolidés dans une base et sur laquelle a été construit le modèle de prédiction de la charge des sinistres climatiques.

L'année 2022 choisie pour la prédiction correspond à une année particulièrement marquée par des événements de grêle, ce qui la distingue des années précédentes. Pour mieux capter cette spécificité, deux modèles distincts ont été proposé :

- Un modèle pour aux événements de grêle
- Un modèle pour les événements hors grêle

En effet, les événements climatiques n'ont pas tous le même coût et certains sont plus graves que d'autres. Aussi, certaines années peuvent présenter des scénarios extrêmes, rendant nécessaire une approche différenciée. Ainsi, la segmentation par type d'événement permet d'améliorer la pertinence des prévisions et de mieux anticiper l'évolution des charges.

Le but étant de prédire la charge des événements survenu en 2022, il a été indispensable d'introduire les charges de sinistres « *as if* » afin de rendre les montants des années antérieurs à 2022 comparables. Deux indicateurs ont été testé :

- L'indicateur de FFB et BT : L'indicateur FFB intégrant les évolutions du coût de la construction des bâtiments et d'autres charges (pour le modèle hors grêle) et l'indicateur BT basé sur des indicateurs plus spécifiques à la construction des bâtiments (pour le modèle grêle).
- L'indicateur de CM à J+7 : l'indicateur intégrant l'évolution du coût réellement observé les sept premiers jours après la survenance d'un sinistre.

Ainsi, quatre modèles ont été entraînés pour prédire

- La charge « *as if* » BT des évènements de grêle
- La charge « *as if* » CM J+7 des évènements de grêle
- La charge « *as if* » FFB des évènements hors grêle
- La charge « *as if* » CM J+7 des évènements hors grêle

Parmi les différents modèles, *XGBoost* est le modèle choisi pour la modélisation de la charge car il est très performant en termes de précision et de rapidité d'exécution.

Pour chaque modélisation, un modèle *Random Forest* a été entraîné afin d'identifier les variables les plus significatives qui serviront ensuite à l'entraînement du modèle *XGBoost*.

Le choix de l'indicateur « *as if* » s'est basé sur la capacité du modèle à prédire au mieux la charge de 2022. À l'issue de cette analyse, l'indicateur du CM J+7 a été retenu pour les deux modèles, aussi bien pour les événements de grêle que pour ceux hors grêle. Cela montre que, c'est la méthode du coût moyen qui capture mieux l'inflation imprévisible liée aux épisodes de grêle en 2022 que celle avec les indices de construction FFB et

BT.

À la suite de la prédiction de la charge des sinistres suivant les méthodes retenues, des traitements sont poursuivis. D'une part, la méthode individuelle développée modélise l'ensemble des sinistres comprenant les sinistres clos sans suite. Sans information sur l'année 2022, le cas de clôture d'un sinistre sans suite doit être pris en compte. Ainsi, ces coûts ont été retirés de la charge estimée pour éviter toute surestimation. D'autre part, les sinistres tardifs ne sont pas directement intégrés dans la modélisation *XGBoost*. Un traitement complémentaire a donc été effectué afin de les inclure dans la charge prédite.

L'évaluation des trois différentes approches (méthode individuelle, méthode agrégée classique et méthode agrégée améliorée) confirme que la meilleure méthode pour la modélisation de la charge des sinistres climatiques est la méthode individuelle, au sens de la minimisation de l'écart entre la valeur réelle de 2022 et la valeur prédite.

Toutefois, une limite majeure a été identifiée : bien que le modèle parvienne à prédire correctement la charge globale, il commet des erreurs sur chaque sinistre individuellement. Des pistes d'amélioration pourraient être explorées pour affiner ces prédictions, notamment par l'amélioration de la qualité de certaines variables telles que la description des dommages par l'assuré, ou encore par ajout d'une photographie des dommages, pouvant être analysée par des processus d'Intelligence Artificielle.
Une autre limite d'une modélisation ligne à ligne repose aussi dans la complexité d'interprétation. Un tel modèle saurait avoir sa place dans le cadre d'une seconde opinion *Best Estimate* de la charge afin d'anticiper une prise en charge ou non de la réassurance.

Bibliographie

[1] Autorité de Contrôle Prudentiel et de Résolution. *Solvabilité II*. 2019.

[2] AXA France. *Conditions générales ma maison*. 2024.

[3] Blent.ai. *XGBoost : Tout savoir sur le Boosting*. 2023.

[4] Breiman Leo. Bagging Predictors. 1994.

[5] Breiman Leo. Random Forests. 2001.

[6] Chen Tianqi et Guestrin Carlos. XGBoost: A Scalable Tree Boosting System. 2016.

[7] Cours Krigeage par D. Marcotte.

[8] CyberInstitut. *Boosting : techniques d'amélioration des modèles en machine learning*. 2023.

[9] DataScientest. *Machine Learning : Définition, fonctionnement, utilisations*. 2023.

[10] Delrio David. *Méthodes d'apprentissage statistique pour la segmentation des sinistres et l'évaluation des provisions non-vie*. Master's thesis, ISFA, Université Claude Bernard Lyon 1, Lyon. 2021.

[11] France Assureurs. *Impact du changement climatique à l'horizon 2050*. 2022.

[12] France Assureurs. *L'assurance des événements naturels en 2022*. 2022.

[13] Freund Yoav et Schapire Robert E. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. 1997.

[14] Friedman Jerome H. Greedy function approximation: A gradient boosting machine. 2001.

[15] Fédération Française de l'Assurance. *L'assurance habitation en 2023*. 2024.

[16] Fédération Française du Bâtiment. *Indice FFB classique : Indice de la construction avec la FFB*. 2023.

[17] Géorisques. *Tempête | Géorisques*. 2023.

[18] Ichi.pro. *Guide du débutant sur le réglage des hyperparamètres de forêt aléatoire*. 2023.

[19] Ichi.pro. *XGBoost: un guide complet pour affiner et optimiser votre modèle*. 2023.

[20] INSEE. *Indices FFB : Séries Index bâtiment, travaux publics et divers de la construction*. 2023.

[21] Milhaud Xavier. *Cours de modélisation actuarielle*. Supports de cours, IMSA, Aix Marseille.

[22] Ministère de la Transition écologique. *Généralités sur le risque inondation en France*. 2023.

[23] Ministère de la Transition écologique. *Publication du 6e rapport de synthèse du GIEC*. 2023.

[24] Mon Coach Data. *Tous les modèles de Machine Learning expliqués en 8 minutes*. 2023.

[25] Météo-France. *Tempêtes et changement climatique*. 2023.

[26] SDEA. *Les risques d'inondations*.

[27] Zhou Zhi-Hua et Yu Yang. AdaBoost. 2009.

Annexes

Région	Département
Région Auvergne-Rhône-Alpes	01 à 07, 26, 38, 42, 69, 73, 74
Région Bourgogne-Franche-Comté	21, 25, 39, 58, 70, 71, 89, 90
Région Bretagne	35, 22, 56, 29
Région Centre-Val de Loire	18, 28, 36, 37, 41, 45
Région Corse	2A, 2B
Région Grand Est	08 à 10, 51, 52, 54, 55, 57, 67, 68, 88
Région Hauts-de-France	59, 62, 60, 80, 02
Région Île-de-France	75, 77, 78, 91, 92, 93, 94, 95, 28
Région Normandie	27, 76, 14, 50
Région Nouvelle-Aquitaine	16, 17, 19, 23, 24, 33, 40, 47, 64, 79, 86, 87
Région Occitanie	09, 11, 12, 30, 31, 32, 34, 46, 48, 65, 66, 81, 82
Région Pays de la Loire	44, 49, 53, 72, 85
Région Provence-Alpes-Côte d'Azur	04, 05, 06, 13, 83, 84

TABLE 40 – Les régions et départements de la France

Table des acronymes

ACP : Analyse en Composantes Principales
BE : Best Estimate
BT : Bâtiments-travaux
CART : Classification And Regression Trees
CFP : Charge finale prévisible
CL : Chain Ladder
CM : Coût moyen
CSS : Clos sans suite
DD : dossier/dossier
EGA : Evènements de grande ampleur
FFB : Fédération Française du Bâtiment
FP : Finale prévisible
GEIC : Groupe d'experts intergouvernemental sur l'évolution du climat
HT : hors taxe
IARD : Incendie Accident et Risques Divers
IBNER : Incurred But Not Enough Reported
IBNR : Incurred But Not Reported
IBNYR : Incurred But Not yet Reported
KNN : K-Nearest Neighbor
MCR : Minimum Capital Requirement
MRH : Multirisque habitation
MSE : Mean Squared Error
NBAS : Nombre avec suite
NBCSS : Nombre de sinistres clos sans suite
NFP : Nombre finale prévisible
ORSA : Own Risk and Solvency Assessment
PP : Professionnels et Particuliers
PSAP : Provision pour Sinistre A Payer
RMSE : Root Mean Square Error
S2 : Solvabilité II
SCR : Solvency Capital Requirement
TGN : Tempête-grêle-neige
UP : Unité de prestation
XGBoost : eXtreme Gradient Boosting

Table des figures

Liste des tableaux

Fin du Mémoire