IRM | institut du risk management BY Institut.

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

Par : Anne-Sophie Dyevre

Titre : Estimation de la prime pure d'une multirisque professionnelle dans un contexte de Covid

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

Membres présents du jury de l'Institut des actuaires :

_____

_____

_____

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

_____

_____

_____

_____

_____

_____

_____

_____

Secrétariat :

Bibliothèque :

Entreprise : Allianz

Nom : ___

Signature et Cachet :

Directeur de mémoire en entreprise :

Nom : PASCAL FORCHIONI

Signature :

Invité :

Nom : _____

Signature :

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

Signature du responsable entreprise

Signature(s) du candidat(s)

Résumé

Dans un contexte de Covid, les différentes mesures de restrictions modifient les habitudes des français. Le marché des professionnels est particulièrement touché par ces changements de comportement. En 2020, la sinistralité de la multirisque des professionnels s'est écartée des tendances observées les années précédentes. L'objectif de ce mémoire est de décrire les étapes de modélisation de la prime pure d'une multirisque des professionnels dans un contexte de Covid, avec un focus sur la segmentation des activités et du risque géographique.

La démarche consiste à introduire des données externes liées à la Covid dans les modèles. Pour chacune des étapes de la modélisation, après avoir évalué l'impact de la pandémie sur les données à modéliser, une méthode est proposée, le cas échéant, pour adapter les modèles de prime pure.

Une approche traditionnelle est retenue pour la modélisation des attritionnels : modèle fréquence, coût moyen dont l'effet des variables explicatives sur le risque à expliquer est modélisé par un algorithme de régression de type GLM. En revanche, pour l'analyse des sinistres atypiques plus volatils et dont l'effet est plus difficile à capter, l'approche GLM est comparée à un modèle de type *Gradient boosting*. Le but étant de mieux segmenter la réallocation de sinistres atypiques par profils de risque.

Les travaux menés sont illustrés par une application sur les deux garanties principales de la multirisque des commerçants et des artisans d'Allianz : l'incendie et le dégât des eaux. Ainsi une garantie d'intensité et une garantie de fréquence sont traitées.

Mots clés : multirisque professionnelle, fréquence, coût moyen, prime pure, Covid, sinistres attritionnels, sinistres atypiques, classification, zonier, Analyse en Composante Principale : ACP, Classification Ascendante Hiérarchique : CAH, modèle linéaire généralisé : GLM, G*radient boosting.*

Abstract

In the context of Covid, the various restrictive measures are changing the habits of the French. The small and medium-sized enterprises (SME) market is particularly affected by these changes. In 2020, the claims experience of multi-risk for professionals deviated from the trends observed in previous years. The objective of this paper is to describe the steps involved in modelling the pure premium of SME multi-risk in a Covid context, with a focus on the segmentation of activities and geographical risk.

The approach consists in introducing external data related to the Covid into the models. For each of the modelling stages, after assessing the impact of the pandemic on the data to be modelled, a method is proposed, if necessary, to adapt the pure premium models.

A traditional approach is adopted for the modelling of attritional claims: frequency model, average cost, where the effect of the rating factors on the risk to be explained is modelled by a GLM. On the other hand, for the analysis of more volatile large claims, whose effect is more difficult to identify, the GLM approach is compared to a gradient boosting model, with the aim of better segmenting the reallocation of severe claims by risk profile.

The work carried out is illustrated by an application on the two main guarantees: fire and water damage of Allianz's multi-risk insurance for retail businesses and craftsmen, thus making it possible to deal with an intensity guarantee and a frequency guarantee.

Keywords: SME multi-risk, frequency, severity, pure premium, Covid, attritional claims, large claims, classification, microzoning, Principal Component Analysis: PCA, Agglomerative Hierarchical Clustering: AHC, Generalized Linear Model: GLM, G*radient boosting*.

Synthèse

L'année 2020 a été mouvementée par la crise sanitaire de la Covid-19. Le marché des professionnels a été très impacté au niveau de la garantie perte financière. D'après les chiffres de France Assureurs, le S/P de la multirisque des artisans, commerçants et prestataires de services s'est dégradé de 42 points en passant de 61 % en 2019 à 103 % en 2020. À la suite du premier confinement en mars 2020, beaucoup d'assureurs ont été amenés à revoir leur couverture, la rédaction de leurs dispositions générales ou même à proposer de nouvelles couvertures en cas de pandémie. Cette crise a eu des conséquences plus larges sur le marché des professionnels, car elle a significativement accéléré l'évolution des comportements des français avec des répercussions sur la sinistralité de l'ensemble des garanties de la multirisque des professionnels (MRP).

Les garanties ont été différemment impactées par les mesures de restrictions avec des conséquences distinctes sur la fréquence et sur le coût moyen. Les graphiques ci-dessous montrent les distorsions observées en 2020 sur les deux garanties principales (incendie et dégât des eaux) du produit MRP d'Allianz :

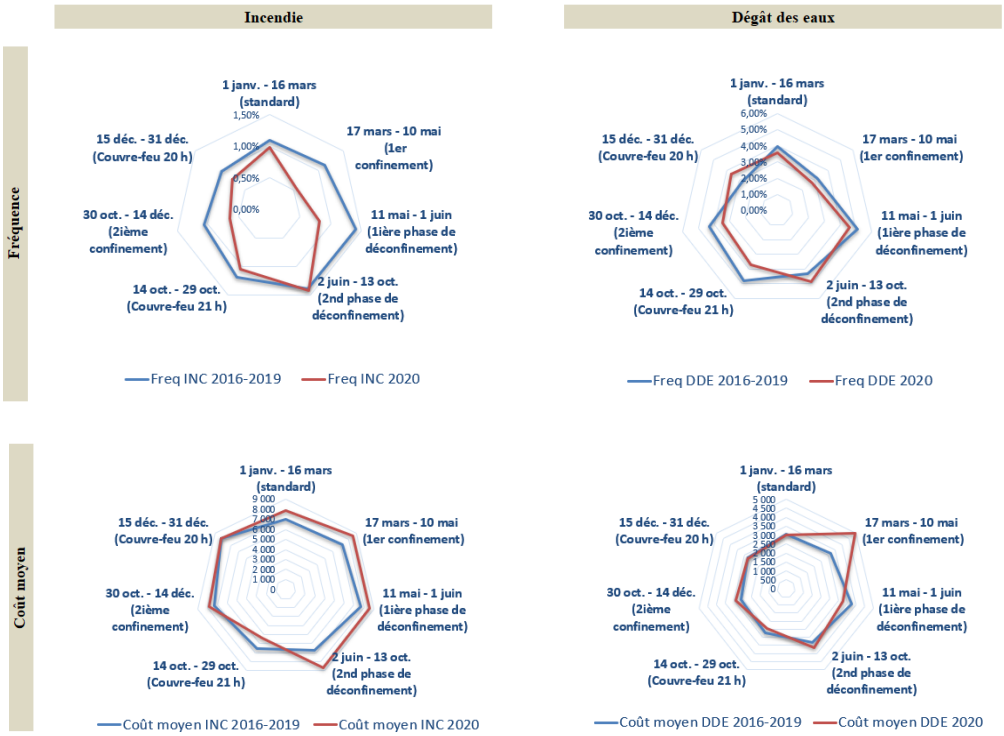


Figure 1 - Fréquence et coût moyen écrêté par périodes de restrictions

Problématique

En 2020, la crise sanitaire liée à la Covid-19 a sensiblement écarté les tendances observées des années précédentes. A quel point les modèles de tarifications actuels sont-ils à remettre en cause ?

Un tel contexte requiert l'excellence technique en termes de modélisations actuarielles et d'utilisation de données, en particulier sur un marché aux multiples avantages qui reste très prisé des assureurs malgré la situation actuelle. Dans un environnement concurrentiel, une segmentation adéquate est indispensable pour identifier les segments cibles de développement et éviter l'antisélection.

L'objectif de ce mémoire est de construire, pour une MRP, des modèles de prime pure par garanties permettant de prédire les effets sur la sinistralité des différentes mesures de restriction sanitaire : confinement, couvre-feu … Une attention particulière est portée à la valeur ajoutée de l'utilisation des données internes et externes, ainsi qu'à la segmentation des activités et du risque géographique.

L'étude porte sur le produit MRP des artisans et des commerçants d'Allianz France, distribué par des agents et des courtiers sur la Métropole. Les analyses sont illustrées sur les deux garanties principales : l'incendie et le dégât des eaux. Pour des raisons de confidentialité, tous les chiffres spécifiques au produit ont été transformés.

L'ensemble des travaux a été réalisé sous les logiciels suivants : SAS Entreprise Guide, Python et les logiciels de tarification : Emblem, Classifier et Radar de chez Willis Towers Watson.

La base de données

La base de données, socle de la modélisation, est construite sur un historique de cinq ans, de 2016 à 2020. Plusieurs sources de données **internes** et **externes** sont exploitées.

Les données internes caractéristiques du **risque** assuré sont présentes dans la base portefeuille. Toutes les réponses aux questions posées à la souscription sont stockées dans cette base. Ces données, mises à jour tous les mois, tiennent compte de toute modification du risque au cours de la vie du contrat. Des traitements de mises en forme des données sont nécessaires pour pouvoir les exploiter. Les données sont formatées pour une meilleure interprétation et les variables continues sont discrétisées pour les besoins du logiciel de tarification Emblem.

Les données **sinistres** distinguées par garanties sont actualisées tous les mois. Les sinistres sont vus à fin mars 2021, vision la plus récente au moment de l'analyse. Les sinistres sans suite ou de montant aberrant, inférieur à 10 €, sont retraités ainsi que les coûts des sinistres au forfait d'ouverture pour ne pas perturber la distribution des coûts moyens. Un seuil d'écrêtement permet de distinguer les sinistre attritionnels (de fréquence) des sinistres atypiques de forte intensité et de les modéliser séparément. Les **seuils d'écrêtement** des sinistres sont déterminés par garantie à l'aide de plusieurs méthodes. Sur la garantie dégât des eaux, un seuil de 50 k€ est retenu en analysant la **distribution** des nombres de sinistres et de la charge. Sur la garantie incendie, la distribution du coût des sinistres à queue lourde est modélisable par une loi de **Pareto généralisée**. Trois méthodes visuelles exploitant les propriétés d'une distribution de Pareto généralisée (GPD) sont analysées pour définir le seuil adéquat : le **paramètre de forme** de la GPD, la fonction **moyenne des excès** et le test de **Kolmogorov-Smirnov** (KS). Ces trois méthodes ont conduit à fixer le seuil d'écrêtement des sinistres incendie à 50 k€. Sur chacune des garanties, le coût des sinistres est écrêté par rapport au seuil retenu.

Les données des **clients** Allianz sont stockées dans la base client mise à jour tous les mois. Cette base recense des informations propres aux clients Allianz comme la multi-détention avec le nombre de contrats d'un client sur chacune des branches assurantielles, mais aussi des données externes sur l'activité du professionnel : la note de risque de faillite transmise par **Euler Hermes** (filiale d'Allianz) et des informations issues de la base externe **Sirene** comme le nombre d'employés.

Les données externes **géographiques** proviennent de différentes bases disponibles en Open Data sur les sites www.data.gouv.fr et www.insee.fr. La **distance** aux pompiers, à la gendarmerie et à la police a également pu être calculée à partir de coordonnées géographiques récupérées sur le site www.openstreetmap.fr. Au total, près d'une quarantaine d'indicateurs sur les caractéristiques du lieu géographique ont pu être calculés à la maille Insee et département. Cette maille a été retenue, car le portefeuille n'est pas géocodé et l'information géographique disponible dans les bases au niveau du contrat est le code de la commune : code Insee (près de 35 000 codes Insee en France). Le code Insee enregistré est normalisé et contrôlé à la souscription. Le produit étudié couvre uniquement des contrats monosites : un contrat ne peut être rattaché qu'à un seul site. Les indicateurs géographiques sont utilisés pour identifier le risque géographique.

Les données **Covid** s'appuient sur la source officielle du site de l'ECDC (European Centre for Disease Prevention and Control : Centre européen de prévention et de contrôle des maladies) : www.ecdc.europa.eu/sites/default/files/document, et sur les variations de fréquentations transmises par Google sur le site www.google.com/covid19/mobility/. L'analyse de ces données a permis de construire plusieurs variables. Une variable restituant les variations de mobilité, une variable regroupant les rubriques d'activités en fonction de l'impact des mesures de fermetures (par exemple, l'alimentation et la santé, commerces de première nécessité pouvant rester ouverts en période de confinement ont été regroupées). Et surtout d'identifier six phases liées

aux restrictions en 2020 : 2 mesures de confinement, 2 phases de déconfinement lors du 1er confinement et 2 mesures de couvre-feu comme le montre le graphique ci-dessous.

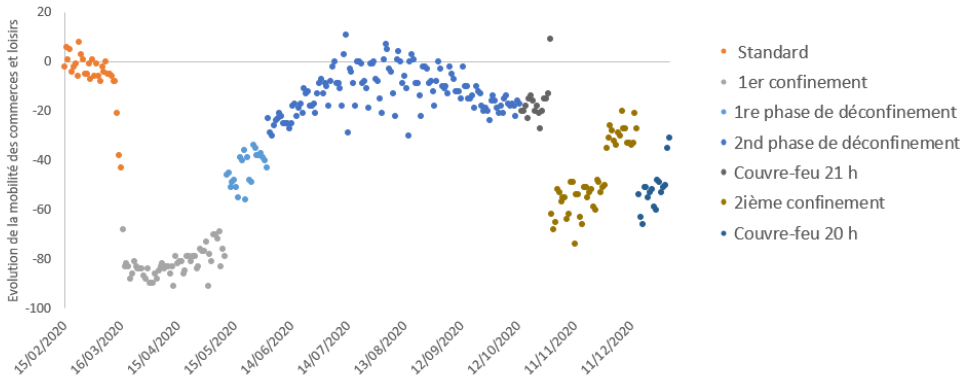


Figure 2 - Tendances de mobilité pour les commerces et loisirs (source Google)

La vision par année d'accident habituellement retenue dans les modèles ne permet pas de capter les changements de comportement qui ont fait suite à la mise en place des différentes mesures de restriction. Les données Covid ont été ajoutées aux données descriptives du risque, et la variable année d'accident a été croisée avec les phases de restrictions déterminées ci-dessus pour créer le critère « Année x Période_Covid » : 2016-Standard, 2017-Standard, 2018-Standard, 2019-Standard, 2020-Standard, 2020-Premier_confinement, 2020-Première_ phase_déconfinement, etc. Les expositions ont été calculées à cette maille pour identifier chaque changement de situation du risque par année.

Les données caractéristiques du risque segmentées par durées d'**expositions** ont ensuite été fusionnées aux données sinistres, puis enrichies des données clients et des données géographiques. Une dernière étape : le **regroupement des activités** a été nécessaire pour pouvoir exploiter les données dans l'outil de modélisation. Le nombre de modalités par variable est limité à 255 dans le logiciel de modélisation utilisé. La nomenclature du produit étudié recense près de 400 activités très variées, définies selon un découpage et une codification propre à Allianz (un code activité Allianz peut correspondre à plusieurs codes NAF : nomenclature d'activité française). Afin de réduire le nombre de modalités, les activités ont été rassemblées en classes de risques homogènes. Une base des activités a été construite avec en ligne les activités et en colonnes les données les caractérisant : surface, chiffre d'affaires, effectif (nombre de salariés), etc. Le regroupement des activités s'est fait en deux temps : une **ACP** pour analyser les activités en fonction des critères retenus, puis une **CAH** pour regrouper les activités à partir des axes de l'ACP. L'analyse visuelle du dendrogramme, complétée des trois métriques : le coefficient de **silhouette**, l'indice de **Calinski-Harabasz** et l'indice de **Davies-Bouldin** ont conduit à retenir un nombre optimal de 5 classes sur chacune des deux garanties.

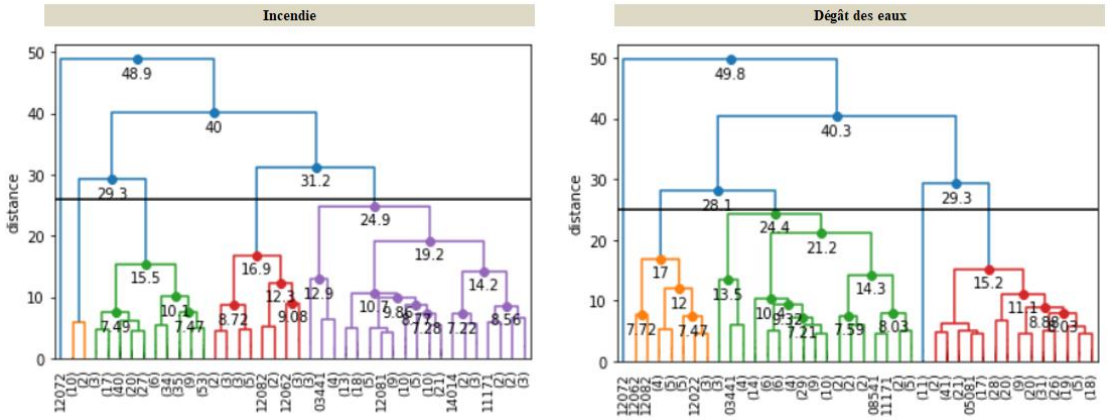


Figure 3 – Dendrogramme classification des activités

La classification a permis de distinguer les activités de petite taille, les activités avec des moyens de préventions importants, les commerces de rues, les activités présentes dans les centres commerciaux, mais surtout de faire ressortir les palaces avec restaurant qui se différencient nettement des autres activités. La classification a été réalisée sur 80 % des données sur les années 2016 à 2019. Afin de valider la stabilité de la classification des activités, l'indice de **Rand** a été calculé sur les 20 % de l'échantillon de test et sur l'année 2020. Des valeurs de l'indice comprises autour de 0,7 valident la classification.

Pour s'assurer que la nouvelle variable classe d'activité, apportait de l'information complémentaire et n'était pas redondante avec les données dont elle était issue, une matrice de corrélation a été construite. Aucune corrélation forte (supérieure ou inférieure à 0,5) n'est ressortie.

L'ajout des classes d'activités est venu finaliser l'étape de création de la base de données. Cette base a été fractionnée en trois échantillons : un échantillon **d'apprentissage** représentant **60 %** de la base pour construire les modèles, un échantillon de **validation** de **20 %** pour estimer les paramètres des modèles qui en requièrent et un échantillon de **test 20 %** pour évaluer les performances du modèle. Les échantillons ont été construits par une méthode de bootstrap en maximisant la somme des écarts au carré (SSE) de sorte que la fréquence et le coût moyen de chaque échantillon soit représentatif de la base globale.

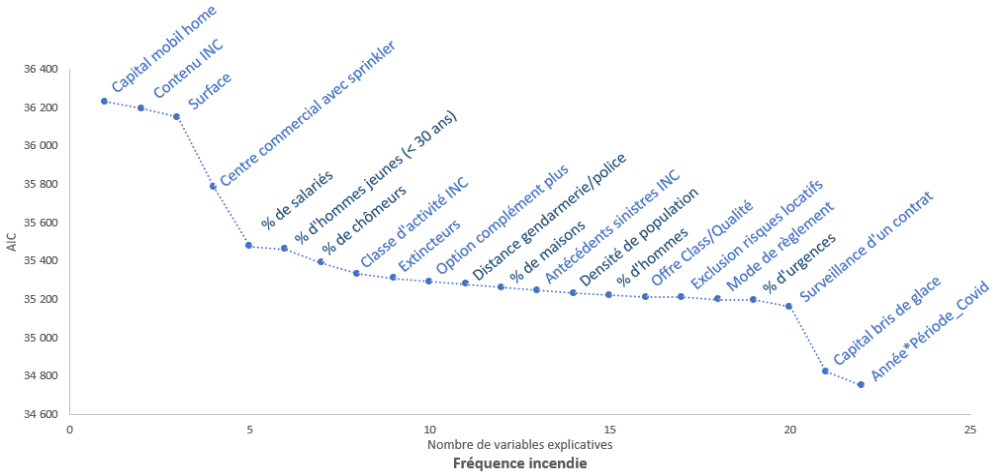
La modélisation de la prime pure est séparée selon deux types de sinistres : les sinistres attritionnels (coût des sinistres écrêté), et les sinistres atypiques (sur-crête).

La modélisation des sinistres attritionnels

La sélection des variables explicatives

Des modèles de **fréquence** (loi de Poisson) et de **coût moyen** (loi Gamma) de type GLM ont été utilisés pour modéliser les sinistres attritionnels, permettant ainsi de distinguer les facteurs explicatifs de la fréquence et du coût moyen.

La première étape a consisté à identifier les **variables explicatives** des modèles de fréquence et de coût moyen sur les garanties incendie et dégât des eaux. Le produit étudié couvre des activités appartenant à des secteurs variés : restauration, hébergement, artisanat du bâtiment, santé, alimentation, habillement, loisirs, équipement de la maison, tourisme ... Cette diversité conduit à un nombre important d'informations descriptives du risque recueillies lors de la souscription. En tout, près de 100 variables sont recensées dans les bases de données. Pour faciliter l'identification des variables descriptives, pour chaque modèle, une présélection de 50 variables classées par ordre d'importance a été effectuée par la méthode « mRMR » (minimum Redundancy and Maximum Relevance : redondance minimale et pertinence maximale). Un regard critique en fonction de la connaissance du produit et des garanties a permis de modifier et/ou de compléter cette sélection. Les variables ont ensuite été ajoutées une à une dans les modèles en optimisant les critères d'ajustement et de prédiction : la déviance et l'AIC. Pour exemple, les variables retenues dans les modèles incendie sont présentées ci-dessous :



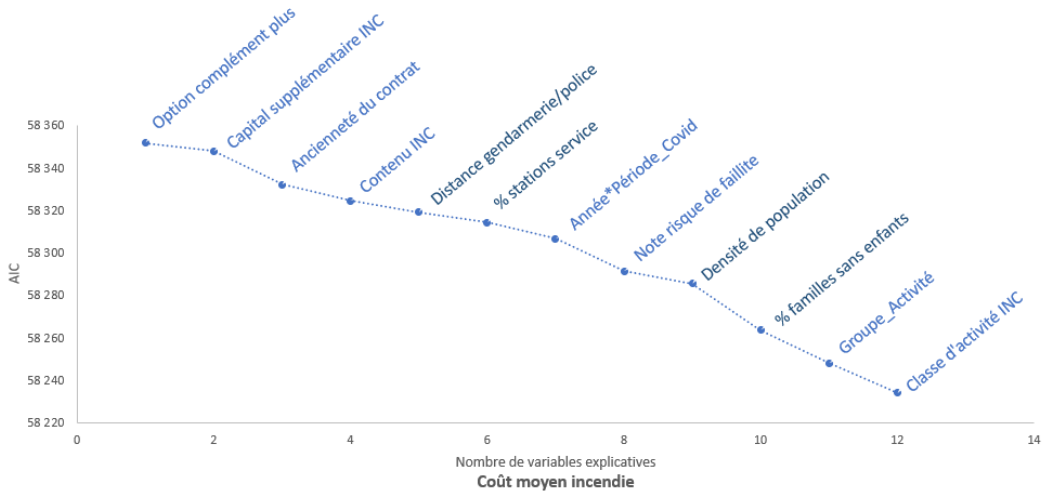


Figure 4 – Variables explicatives des modèles de fréquence et coût moyen incendie

Parmi les variables introduites dans les modèles, la variable **Année x Période_Covid** a toujours été prise en compte, elle a permis de capter les tendances dans l'évolution de la sinistralité, non liées à un phénomène structurel, et de distinguer les différents effets des périodes de restrictions. Les **variables de taille** de risque : capital, superficie se retrouvent dans les modèles. D'autres variables sont plus caractéristiques du risque couvert par la garantie comme les moyens de prévention : extincteurs, sprinkler. Les variables **géographiques** (en bleu plus foncé sur les graphiques ci-dessus) apportent une information sur l'environnement géographique, sociodémographique du risque.

Afin de capter les effets des différentes périodes de restrictions sanitaires, l'interaction de chacune de ces variables avec la variable Année x Période_Covid a été testée. Pour les variables dont les tendances étaient différentes par périodes de restriction, l'interaction avec la variable Année x Période_Covid a été rajoutée dans le modèle. Par exemple, l'interaction Surface x Année x Période_Covid a été retenue dans le modèle de fréquence incendie.

Les variables géographiques retenues dans les modèles de fréquence et de coût moyen ont seulement capté une partie de l'effet géographique. L'analyse et le lissage des résidus de ces modèles ont permis d'identifier l'effet géographique global : le zonier.

La construction des zoniers

Le schéma ci-dessous résume les étapes de construction des modèles :

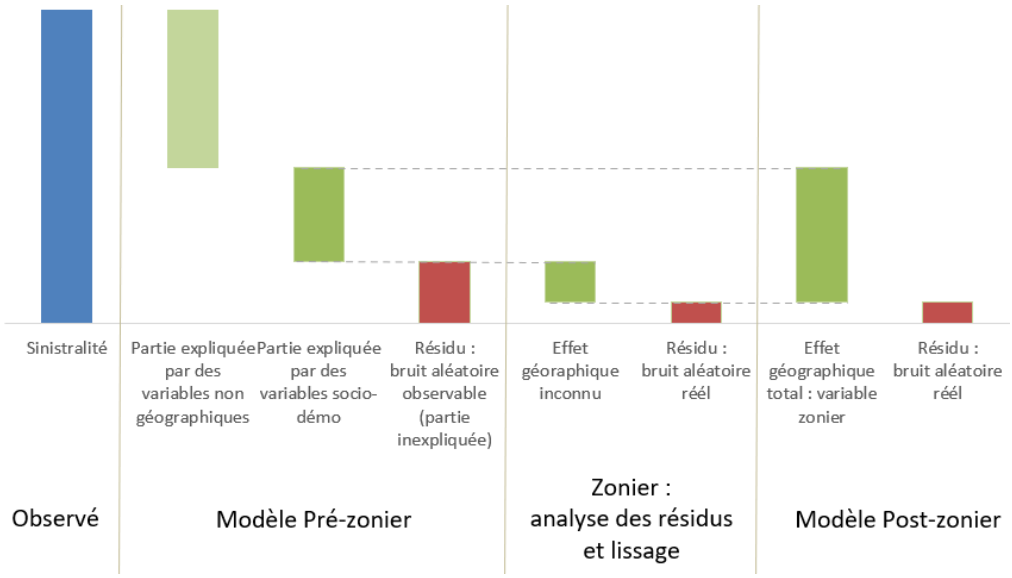


Figure 5 – Etapes de construction des zoniers

Pour chacun des modèles, la construction des zoniers a été réalisée par un lissage des résidus pour identifier l'effet géographique inconnu. Après neutralisation de l'effet non-géographique (standardisation), les résidus projetés sur la carte des codes postaux de la France ont été lissés par la méthode ***Adjacency*** : méthode de lissage local basée sur l'approche bayésienne qui prend en compte le risque des codes postaux immédiatement voisins, c'est-à-dire **adjacents**. Pour éviter tout risque de surajustement, le niveau de lissage a été estimé sur l'échantillon de validation en minimisant la moyenne des écarts au carré (MSE) entre les résidus lissés et les résidus originaux. L'effet résiduel capté par le lissage est venu compléter l'effet des variables géographiques. La combinaison de ces deux effets a constitué l'effet géographique total recherché.

Afin d'analyser l'impact du lissage dans les modèles, l'effet seul des variables géographiques a été comparé à l'effet géographique global incluant l'effet géographique résiduel (les résidus lissés). Ces deux statistiques calculées par codes postaux ont été regroupées en zones, et ont été dénommées « zonier non lissé » et « zonier lissé ». Afin d'avoir un bon compromis entre granularité et robustesse compte tenu des volumes de données, un regroupement en 50 zones a été implémenté avec la méthode de Ward.

La validation des modèles

Les zoniers non lissés et lissés créés par garantie pour la fréquence et le coût moyen ont été testés dans les modèles, en remplacement des variables géographiques. Les graphiques de comparaison de l'observé et des prédictions sur l'échantillon de test ont amené à retenir un zonier lissé pour les modèles de fréquence et de coût moyen dégât des eaux, un zonier non lissé pour le modèle de fréquence incendie et aucun zonier pour le modèle de coût moyen incendie. L'apport du zonier pour les modèles concernés a été mesuré par une augmentation du coefficient de Gini entre les modèles pré et post zonier. Comme pour toute variable intégrée dans les modèles, la stabilité dans le temps du zonier a été contrôlée. La performance et la capacité des modèles à généraliser les tendances apprises, ont été évaluées par la différence de déviance sur l'échantillon de modélisation et l'échantillon de test entre le modèle sans variables et le modèle retenu. Pour chacun des modèles, la baisse de déviance constatée sur l'échantillon de modélisation est restée suffisamment proche de celle constatée sur l'échantillon de test pour conclure à l'absence de surapprentissage (*overfitting*).

La modélisation des sinistres atypiques

Un seuil d'écrêtement déterminé à 50 k€ lors de la création de la base de données a permis de distinguer les sinistres attritionnels des sinistres atypiques. Sur la garantie dégât des eaux, la sur-crête concerne 0,7 % des sinistres et représente 7,6 % de la charge. La crise sanitaire a eu un impact défavorable sur le nombre et coût moyen des sinistres atypiques sur cette garantie. Toutefois, le faible nombre des sinistres atypiques en dégât des eaux n'a pas permis de les modéliser, et la sur-crête a été mutualisée sur l'ensemble des contrats. La garantie incendie quant à elle, a la spécificité de couvrir des risques pouvant atteindre plusieurs millions d'euros. La charge est portée par un petit nombre de sinistres : les sinistres supérieurs à 50 k€ représentent 7,3 % des sinistres incendie et portent 73,4 % de la charge incendie. Le choix de seuil d'écrêtement de 50 k€ a permis d'obtenir un volume suffisant de sinistres atypiques incendie pour pouvoir les modéliser.

La propension de sinistres atypiques en incendie (pourcentage de sinistres de forte intensité parmi les sinistres incendie) a été modélisée par un modèle logit. Dans le but d'identifier au mieux les profils de risque, deux méthodes de régression logistique ont été testées : une méthode classique de type GLM et une méthode de *Machine Learning* de type *Gradient boosting : LightGBM*. L'amélioration de 6 points du coefficient de Gini et les variables sélectionnées avec le *LightGBM* ont conclu à privilégier ce modèle pour estimer la propension.

Pour pouvoir modéliser par profils de risques le coût moyen des sinistres incendie atypiques, un second seuil d'écrêtement a été déterminé à 750 k€. La partie de la sur-crête comprise entre 50 k€ et 750 k€ a pu être modélisée par une loi Gamma, tandis que la partie supérieure à 750 k€ a été mutualisée sur l'ensemble des sinistres incendie atypiques. Afin de rester cohérent avec le modèle de propension, un modèle *LightGBM* a été retenu pour estimer le coût moyen des sinistres incendie atypiques.

Les variables des modèles de propension et de coût moyen des sinistres incendie atypiques sont représentées ci-dessous par ordre d'importance :

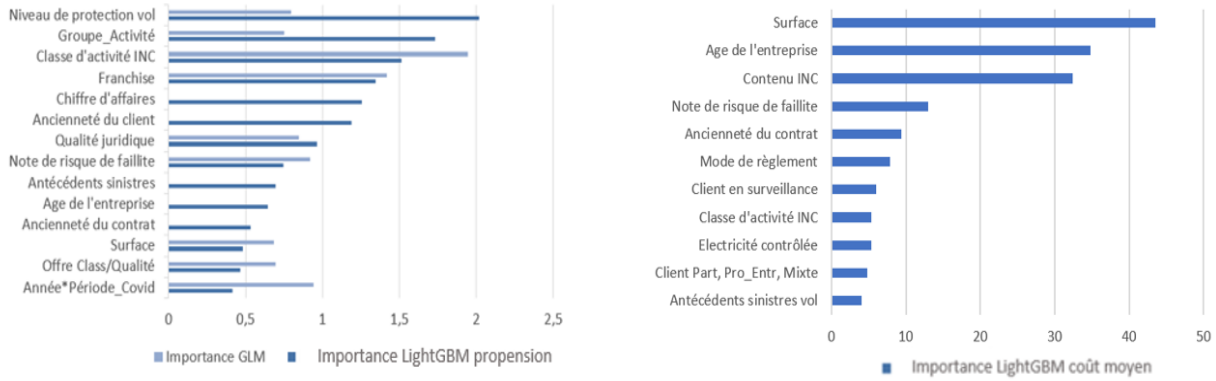


Figure 6 – Importance des variables des modèles de sinistres incendie atypiques

L'importance des variables a été évaluée en fonction du pouvoir discriminant (intervalle des beta) pour le GLM et selon le gain pour le *LightGBM* (ces deux métriques ont été rapportées à leur moyenne pour la comparaison des modèles de propension). La propension et le coût moyen des sinistres incendie atypiques sont expliqués par des critères différents, d'où l'importance de les modéliser séparément.

Prime pure et interprétations

Sur chacune des garanties étudiées, les modèles de fréquence, coût moyen et de sinistres atypiques ont été consolidés pour avoir une estimation de la prime pure. La prime pure estimée a été comparée à l'observée par critères. Sur les périodes de restrictions, les tendances ont bien été captées par les modèles. En incendie, la sinistralité a tendance à baisser avec les restrictions et à augmenter en période de déconfinement avec la reprise de l'activité. En dégât des eaux, les périodes de confinement et de déconfinement ont eu un impact négatif sur la sinistralité tandis que les périodes de couvre-feu ne semblent pas avoir eu de répercussions.

Sur certains critères comme la surface, pour les profils avec très peu d'exposition, la prime pure estimée est ressortie avec un niveau très inférieur à l'observée du fait de la mutualisation d'une partie des sinistres atypiques. L'interprétation d'un arbre de décision a permis de mettre en place une règle de sélection des profils de risque les plus exposés aux sinistres incendie atypiques.

Par ailleurs, les travaux menés sur la classification des activités et les zones géographiques ont conduit à proposer des évolutions de classe tarifaire pour les activités et de zone tarifaire pour les communes, par rapport à la classification et aux zoniers actuellement en place dans le tarif commercial.

Conclusion

Toutes les étapes de modélisation de la prime pure d'une MRP dans un contexte de Covid ont été présentées dans ce mémoire. Les modèles traditionnels GLMs de fréquence et de coût moyen ont été adaptés pour pouvoir capter les effets des mesures de restriction dans la modélisation des attritionnels. Tandis que des méthodes plus innovantes de *Machine Learning* ont été appliquées pour une meilleure segmentation des sinistres incendie atypiques.

Les limites de ce mémoire ont été le manque de recul sur l'évolution de la crise sanitaire au moment de l'étude. En 2022, la situation sanitaire est loin d'être terminée et continue à évoluer. Une mise à jour des modèles avec un historique plus grand des données impactées par la crise est nécessaire pour généraliser les effets à venir du Covid sur la sinistralité.

Synthesis

Year 2020 has been turbulent time due to the Covid-19 health crisis. The small and medium-sized enterprises (SME) market has been heavily impacted in terms of financial loss cover. According to France Assureurs' figures, the loss ratio of multi-risk for craftsmen, traders and service providers has increased by 42 points from 61% in 2019 to 103% in 2020. Following the first national lockdown in March 2020, many insurers have had to review the wording of their cover or even offer new pandemic cover solution. This crisis initiated consequences on the SME market, as it significantly boosted the behavior changes, with impact on the claims experience of all SME multi-risk cover.

The effects of mobility restriction periods vary according to the cover and impact differently the frequency and the severity differently, as shown in the charts below for the two main covers (fire and water damage) of Allianz's SME product:

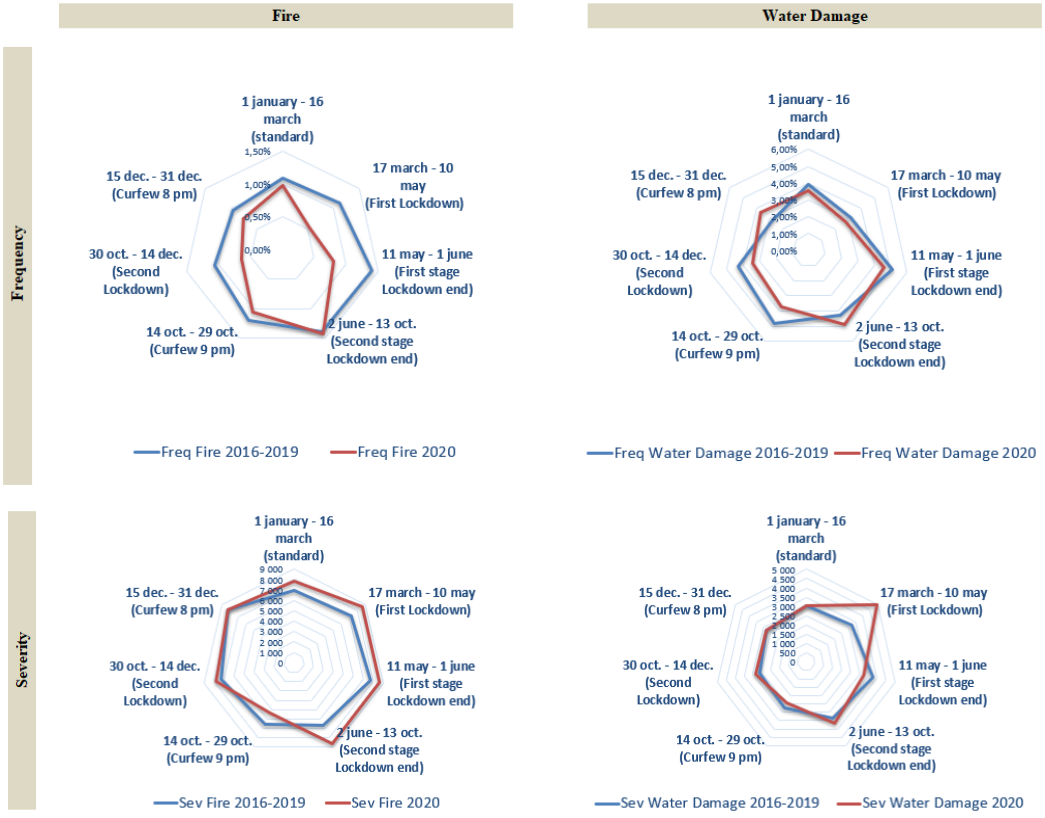


Figure 1 - Frequency and capped severity by restrictions

Problematics

The Covid-19 health crisis has significantly reversed the 2020 trends of previous years. To what extent should current pricing models be challenged?

Such a context requires technical excellence in terms of actuarial modelling and use of data, especially in a market with multiple advantages that remains very popular with insurers despite the current situation. In a competitive environment, adequate segmentation is essential to identify target segments for development and avoid anti-selection.

The objective of this thesis is to build, for a SME product, pure premium models by coverages allowing the prediction of the effects on the loss experience of different health restriction measures: lockdown, curfew... Particular attention is paid to the added value of the use of internal and external data, as well as to the segmentation of activities and geographical risk.

The study focuses on Allianz France's SME product for retail businesses and craftsmen, distributed by agents and brokers in mainland France. The analyses are illustrated on the two main guarantees: fire and water damage. For reasons of confidentiality, all product-specific figures have been transformed.

All the work was carried out using the following software: SAS Entreprise Guide, Python and the underwriting software: Emblem, Classifier and Radar from Willis Towers Watson.

The database

The database, the basis of the modelling, is built on a five-year history, from 2016 to 2020. Several internal and external data sources are used.

Internal data characteristic of the insured risk is present in the portfolio database. All the answers to the questions asked at underwriting are stored in this database. This data, updated monthly, considers any changes in risk during the life of the contract. Data formatting processes are necessary to be able to use them. The data is formatted for better interpretation and the continuous variables are discretized for the needs of the Emblem pricing software.

The claims data, by cover, are taken from the claims database which is updated every month. This database contains only claims that have occurred at Allianz. Claims made prior to subscription and declared by the insured at the time of subscription are recorded in the portfolio database. The claims are seen at the end of March 2021, the most recent vision at the time of the analysis. Claims with no follow-up or outliers (less than €10) are restated, as well as the costs of claims with an opening flat rate, so as not to disrupt the distribution of average costs. A capping threshold is used to distinguish between attritional (frequency) and large (intensity) claims and to model them separately. The capping thresholds for claims are determined by cover using several methods. For water damage cover, a threshold of €50 k is selected by analysing the distribution of the number of claims and the cost. For fire insurance, the distribution of the cost of heavy tail claims can be modelled by a generalised Pareto distribution. Three visual methods exploiting the properties of a generalized Pareto distribution (GPD) are analysed to define the appropriate threshold: the shape parameter of the GPD, the mean excess function and the Kolmogorov-Smirnov (KS) test. These three methods led to a capping threshold for fire claims of €50 k. For each of the coverages, the cost of claims is capped against the threshold.

Allianz customer data is stored in the customer database, which is updated every month. This database contains information specific to Allianz clients, such as the number of contracts held by a client in each insurance branch, as well as external data on the professional's activity: the bankruptcy risk rating provided by Euler Hermes (an Allianz subsidiary) and information from the external Sirene database, such as the number of employees.

The external socio-demographic data were taken from various databases available in Open Data at www.data.gouv.fr and www.insee.fr. The distance to the fire brigade, the gendarmerie and the police could also be calculated from geographical coordinates retrieved from the website www.openstreetmap.fr. In total, nearly forty indicators on the characteristics of the geographical location could be calculated at the Insee and department level. This grid was chosen because the portfolio is not geocoded and the geographic information available in the databases at the contract level is the city code: Insee code (nearly 35 000 Insee codes in France). The Insee code recorded is standardized and checked at the time of subscription. The product studied covers only single-site contracts: a contract can be attached to only one site. Socio-demographic indicators are used to identify the geographical risk.

The Covid data is based on the official source of the ECDC (European Centre for Disease Prevention and Control) website: www.ecdc.europa.eu/sites/default/files/document, and on the variations in mobility transmitted by Google on the www.google.com/covid19/mobility website. The analysis of these data made it possible to construct a variable showing the variations in mobility, a variable grouping the activity headings according to the impact of the closure measures (for example, food and health, which are essential shops that can remain open during the period of confinement, were grouped together), but above all to identify six phases linked to the restrictions in 2020: 2 lockdown measures, 2 lockdown end phases during the first lockdown and 2 curfew measures, as shown in the graph below.

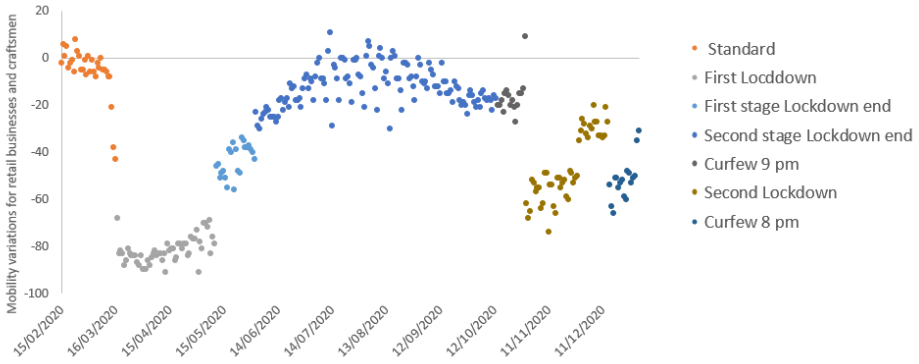


Figure 2 – Mobility variations for retail businesses and craftsmen (source Google)

The accident year view usually used in the models does not capture the changes that have occurred because of the different restriction measures. The Covid data was added to the risk descriptive data, and the accident year variable was cross-referenced with the restriction phases determined above to create the "Year x Covid_period" criterion: 2016-Standard, 2017-Standard, 2018-Standard, 2019-Standard, 2020-Standard, 2020-First_lockdown, 2020-First_lockdownEnd, etc. Exposures were calculated at this grid for each change in risk status.

The risk characteristic data segmented by exposure duration was then merged with the claims data and enriched with customer and socio-demographic data. A final step was to group the activities to use the data in the modelling tool. The number of modalities per variable is limited to 255 in the modelling software used. The nomenclature of the product studied lists nearly 400 very varied activities, defined according to a breakdown and coding specific to Allianz (an Allianz activity code can correspond to several NAF codes: French activity nomenclature). To reduce the number of modalities for the activity variable, the activities were grouped into homogeneous risk classes. A database of activities was constructed with the activities in rows and the data characterizing in columns: surface area, turnover, workforce (number of employees), etc. The grouping of activities was done in two stages: a PCA to analyze the activities according to the selected criteria, and then an AHC to group the activities according to the PCA axes. The visual analysis of the dendrogram, completed by three metrics: the silhouette coefficient, the Calinski-Harabasz index and the Davies-Bouldin index, led to the selection of an optimal number of 5 classes on each of the two guarantees.

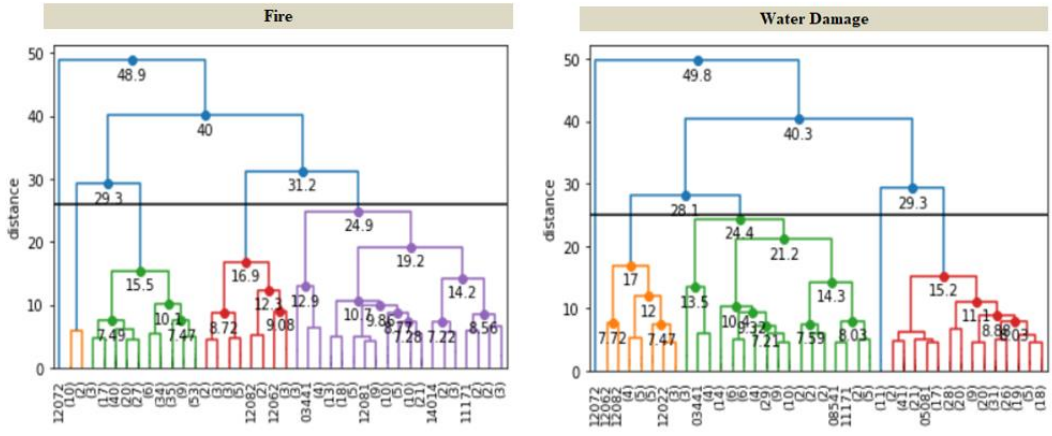


Figure 3 – Dendrogram activities classification

The classification made it possible to distinguish between small activities, activities with significant prevention resources, street shops, activities in a mall, but above all to highlight palaces with restaurants, which are clearly different from other activities. Classification was performed on 80% of the data for the years 2016 to 2019. In order to validate the stability of the activity classification, the Rand index was calculated on the 20% of the test sample and on the year 2020. Index values around 0.7 validate the classification.

To ensure that the new activity class variable, created from other variables, provided additional information and was not redundant with the data from which it was derived, its correlation with the data from which it was created was tested through a correlation matrix. No strong correlation (greater than or less than 0.5) was found.

The addition of the activity classes finalized the database creation stage. This database was split into three samples: a learning sample representing 60% of the database to build the models, a validation sample of 20% to estimate the parameters of the models that require them and a test sample of 20% to evaluate the performance of the model. The samples were constructed using a bootstrap method by maximizing the sum of squared deviations (SSE) so that the frequency and average cost of each sample is representative of the overall base.

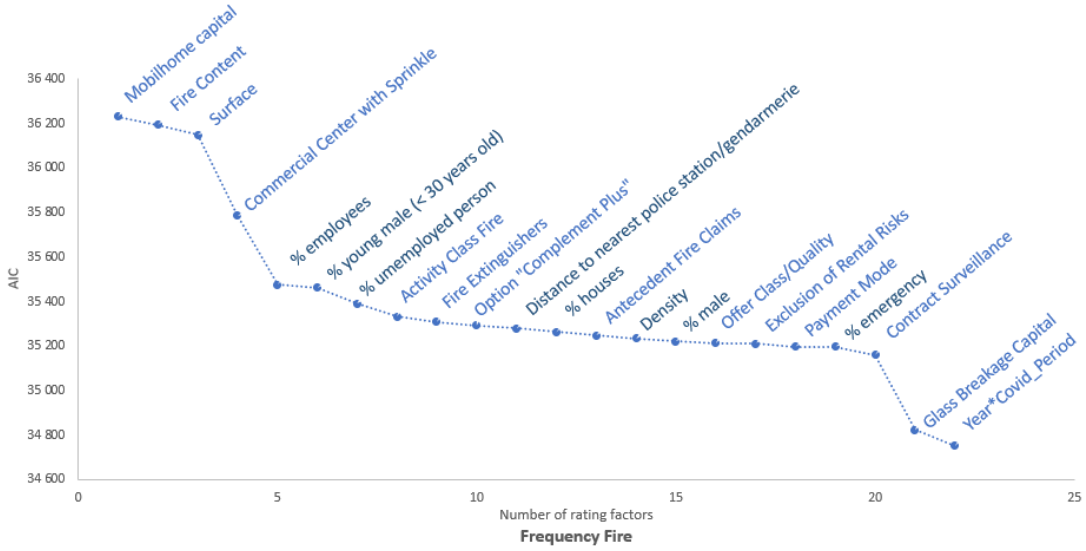
The pure premium modelling is separated into two types of claims: attritional claims (capped claims cost), and large claims (over-peak).

Modelling attritional claims

Selection of rating factors

Frequency (Poisson distribution) and average cost (Gamma distribution) models of the GLM type were used to model attritional claims, thus making it possible to distinguish the rating factors of frequency and severity.

The first step was to identify the rating factors for the frequency and severity models for fire and water damage cover. The product studied covers activities belonging to various sectors: restaurants, accommodation, building craftsmen, health, food, clothing, leisure, household equipment, tourism, etc. This diversity leads to a large amount of descriptive information on the risk collected at the time of subscription. In all, nearly 100 variables are listed in the databases. To facilitate the identification of descriptive variables, for each model, a pre-selection of 50 variables ranked in order of importance was carried out using the "mRMR" method (minimum redundancy and maximum relevance). A critical look at the product and guarantee knowledge allowed to modify or complete this selection. The variables were then added one by one to the models by optimizing the fit and prediction criteria: deviance and AIC. For example, the variables retained in the fire models are presented below:



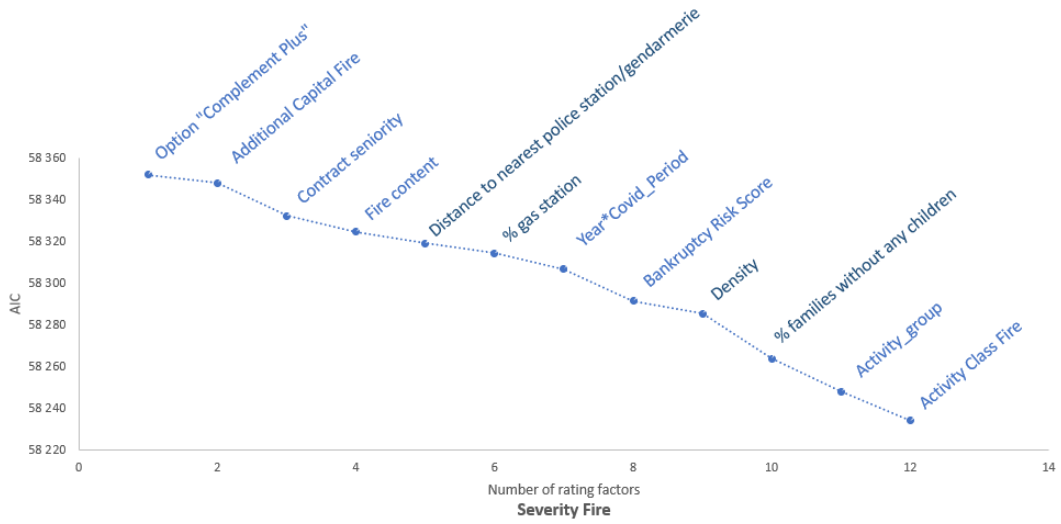


Figure 4 – Rating factors for fire frequency and severity

Among the variables introduced in the models, the variable Year x Period_Covid was always considered, it allowed to capture the trends in the evolution of the loss experience not linked to a structural phenomenon, and to distinguish the different effects of the restriction periods. The variables of risk size: capital, area are found in the models. Other variables are more characteristic of the risk covered by the guarantee, such as the means of prevention: extinguishers, sprinklers. The socio-demographic variables (in darker blue on the graphs above) provide information on the geographical risk.

To capture the effects of the different periods of health restrictions, the interaction of each of these variables with the variable Year x Covid_period was tested. For variables with different trends per restriction period, the interaction with the variable Year x Covid_period was added to the model. For example, the interaction Area x Year x Covid_Period was retained in the fire frequency model.

The socio-demographic variables retained in the frequency and average cost models only captured part of the geographical effect. The analysis and smoothing of the residuals of these models allowed the identification of the overall geographical effect: the microzoning.

The construction of the microzoning

The diagram below summarizes the steps involved in building the models:

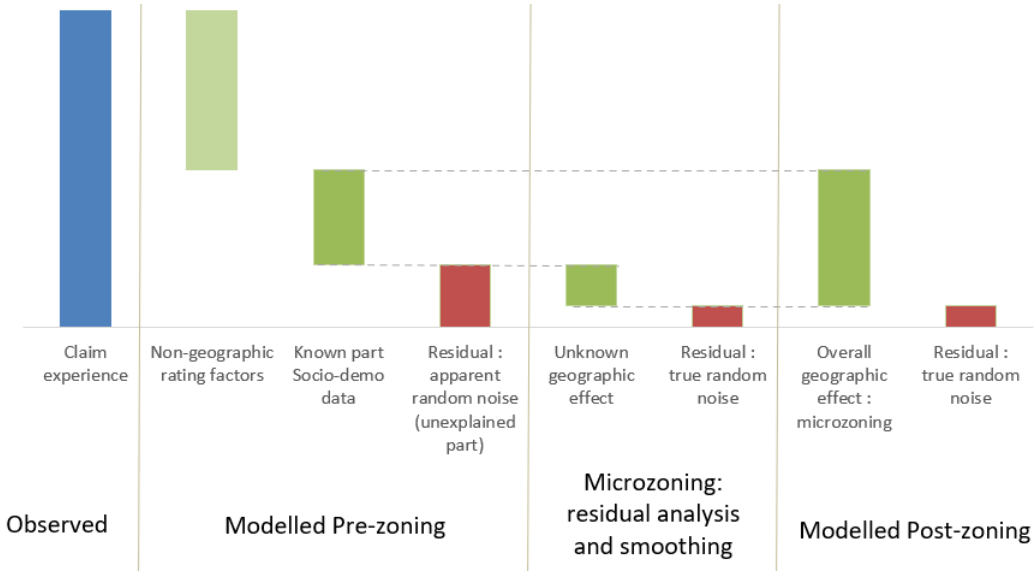


Figure 5 – Microzoning construction steps

For each model, the construction of the microzoning was carried out by smoothing the residuals to identify the unknown geographical effect. After neutralizing the non-geographical effect (standardization), the residuals projected onto the postcode map of France were smoothed using the Adjacency method: a local smoothing method based on the Bayesian approach which takes into account the risk of the immediately neighboring, i.e. adjacent, postcodes. To avoid any risk of overfitting, the level of smoothing was estimated on the validation sample by minimizing the mean squared deviation (MSE) between the smoothed and the original residuals. The residual effect captured by the smoothing complemented the effect of the socio-demographic variables. The combination of these two effects constituted the total geographical effect sought.

To analyze the impact of smoothing in the models, the effect of the socio-demographic variables alone was compared to the overall geographical effect including the residual geographical effect (the smoothed residuals). These two statistics calculated by postcode were grouped into zones and were named "unsmoothed microzoning" and "smoothed microzoning". To have a good compromise between granularity and robustness considering the volumes of data, a grouping in 50 zones was implemented with the Ward method.

Model validation

The unsmoothed and smoothed microzoning created by the guarantee for frequency and severity were tested in the models as a replacement for the socio-demographic variables. The comparison graphs of the observed and the predictions on the test sample led to retain a smoothed microzoning for the water damage frequency and severity models, an unsmoothed microzoning for the fire frequency model and no microzoning for the fire severity model. The contribution of the microzoning to the models was measured by an increase in the Gini coefficient between the pre and post microzoning models. As for any variable integrated in a model, the stability over time of the microzoning was checked. Finally, the performance and ability of the models to generalize the learned trends was assessed by the difference in deviance between the model without variables and the model retained on the modelling sample and the test sample. For each model, the decrease in deviance observed on the modelling sample remained sufficiently close to that observed on the test sample to conclude that there was no overfitting.

Modelling of large claims

A capping threshold set at €50 k when the database was created made it possible to distinguish attritional claims from serious claims. In the case of water damage cover, the excess peak concerns 0,7% of claims and represents 7,6% of the cost. The health crisis had an unfavorable impact on the number and average cost of serious claims. However, the low number of serious claims for this cover did not allow them to be modelled, and the excess peak was mutualized over all the insureds. Fire insurance has the specificity of covering risks that can reach several million euros. The burden is borne by a small number of claims: claims over €50k represent 7,3% of fire claims and carry 73,4% of the fire burden. For this cover, the year 2020 was not atypical in terms of serious claims. This choice of capping threshold has made it possible to model severe fire claims.

The propensity of large fire claims (percentage of large claims among fire claims) was modelled by a logit model. To identify the risk profiles as well as possible, two logistic regression methods were tested: a classical GLM method and a Machine Learning method of the Gradient boosting type: *LightGBM*. The 6 points improvement in the Gini coefficient and the variables retained with the *LightGBM* led to the decision to use this model to estimate propensity.

To model the average cost of large fire claims by risk profile, a second capping threshold was determined at €750 k. The part of the over-peak between €50 k and €750 k could be modelled by a Gamma law, while the part above €750 k was mutualized over all severe fire claims. To remain consistent with the propensity model, a *LightGBM* model was used to estimate the severity of large fire claims.

The rating factors in the propensity and the severity of large fire claims models are shown below in order of importance:

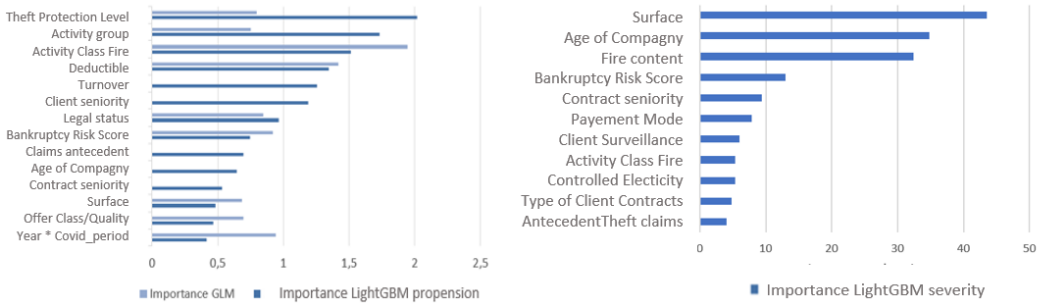


Figure 6 – Importance rating factors for fire large claims

The importance of variables was assessed according to the discriminating power (beta range) for the GLM and according to the gain for the *LightGBM* (these two metrics were related to their mean for the comparison of propensity models). Propensity and severity of large fire claims are explained by different criteria, hence the importance of modelling them separately.

Pure premium and interpretations

For each coverage studied, the frequency, severity and large models were consolidated to provide an estimate of the pure premium. The estimated pure premium was compared to the observed by criteria. Over the restriction periods, the trends were well captured by the models. In fire, the loss experience tends to decrease with the restrictions and to increase in the lockdown end period with the resumption of activity. In water damage, the lockdown and lockdown end periods had a negative impact on the loss experience while the curfew periods did not seem to have any impact.

On certain criteria such as surface area, for profiles with very low exposure, the estimated pure premium came out much lower than observed due to the mutualization of part of the large claims. However, the knowledge of the profiles most exposed to large fire claims has made it possible to reinforce the underwriting rules. A rule for selecting these risk profiles has been put in place thanks to the interpretation of the results of a decision tree.

In addition, the work carried out on the classification of activities and geographical zones led to the proposal of changes in the price class for activities and the price zone for postal codes, compared to the classification currently in place in the commercial price.

Conclusion

All the steps for modelling the pure premium of a SME product in a Covid context have been presented in this thesis. The traditional GLMs frequency and severity models have been adapted to capture the effects of restrictive measures in the modelling of attritionals claims. While more innovative Machine Learning methods have been applied for better segmentation of large fire claims.

The limitations of this thesis were the lack of hindsight on the evolution of the health crisis at the time of the study. In 2022, the health situation is far from over and continues to evolve. An update of the models with a larger history of data impacted by the crisis is necessary to generalize the future effects of Covid on claims.

Remerciements

Avant toutes choses, je tiens à remercier Magali VACHEROT, responsable du département *Pricing*, de m'avoir offert l'opportunité de travailler au sein de son équipe et permis de traiter un sujet d'actualité en MRP.

Je tenais également à remercier vivement Pascal FORCHIONI pour ses nombreuses qualités professionnelles et personnelles, et pour son expertise sur la tarification des produits du marché des professionnels.

J'adresse mes remerciements à tous mes collègues d'Allianz pour leurs encouragements, et plus particulièrement à Claire LAMON pour ses conseils avisés.

Je remercie également l'ensemble des enseignants du CEA pour la qualité de la formation et des cours dispensés.

J'adresse enfin mes remerciements à mes responsables, passés et actuels, qui m'ont soutenu dans ma démarche d'entreprendre une formation d'actuaire, et qui m'ont donné les moyens de la mener à son terme.

Table des matières

INTRODUCTION

« Le marché des pros est un segment prioritaire pour plusieurs raisons. Sur la partie strictement IARD, les artisans, TNS et TPE constituent une part très importante du tissu économique et donc de la matière assurable. Ensuite, le marché des pros nous intéresse car ce sont des assurés aux multiples besoins, tant dans la sphère professionnelle que dans la sphère privée, en dommages, en protection de la personne et en gestion de patrimoine ». Ce propos tenu par François Nédey, responsable produits et assurances de biens et de responsabilité au comex d'Allianz France, dans un article de L'Argus de l'assurance en novembre 2019, expose les avantages du marché des professionnels pour les assureurs.

Le marché des professionnels, très convoité par les assureurs, a subi dernièrement les effets de la crise de la Covid-19. Cette crise a certes eu un impact négatif sur la rentabilité de ce marché, cependant, elle n'a pas diminué le regain d'intérêt pour ce secteur d'activité marqué par une revisite des offres et l'arrivée de nouveaux acteurs.

Dans ce contexte particulier de crise sanitaire qui touche principalement les commerçants et les artisans, la garantie au cœur du sujet est la perte d'exploitation qui a amené à la révision, pour la plupart des assureurs, des clauses liées à la couverture en cas de pandémie. Or, la pandémie a également des effets sous-jacents sur les autres garanties de la multirisque des professionnels. Les différentes mesures mises en place par le gouvernement depuis le mois de mars 2020 ont changé les habitudes des français et des commerçants avec des incidences sur la sinistralité de l'ensemble des garanties, et notamment sur les garanties couvrant les dommages matériels.

La principale mission de l'équipe tarification d'une compagnie d'assurances est la mise à jour du tarif en fonction des évolutions de l'activité. Le tarif se décompose en deux parties, une partie destinée à couvrir le risque lui-même : la prime pure, à laquelle s'ajoute les chargements permettant de couvrir les frais de l'assureur et une marge de risque. Les travaux présentés dans ce mémoire se concentrent exclusivement sur la prime pure, c'est-à-dire la partie correspondant au risque.

La spécificité du cycle inversé de production en assurance nécessite de déterminer ce que chaque contrat devrait couter en moyenne avant sa vente. L'estimation du meilleur coût attendu d'un contrat repose sur l'estimation du coût probable des sinistres par profils de risque. Ce coût est déterminé par le produit de deux facteurs, d'une part le pourcentage de risque de survenance de sinistres : la fréquence et d'autre part le montant du sinistre moyen attendu par segment de risques homogènes : le coût moyen. Les modèles statistiques utilisés, les modèles linéaires généralisés : GLM, aident à capter l'aspect aléatoire, et s'appuient sur ce qui s'est réellement produit grâce à des données collectées dans le passé.

Toutes les sources d'informations internes et externes (Open Data) sont précieuses pour améliorer la segmentation des risques. Avoir une bonne vision prospective des risques souscrits est d'autant plus importante dans le contexte économique présent. Trois axes d'amélioration ont été identifiés dans les modèles de tarification actuels : la révision de la classification des activités, la refonte du zonier et la performance de la modélisation des sinistres incendie atypiques.

L'émergence de la pandémie, en raison des réductions de mobilité sociale, soulève un certain nombre de problèmes quant à l'application du cadre standard de modélisation, spécifiquement sur la stabilité des phénomènes à modéliser en 2020. Comment prendre en compte dans les modèles une année qui se détache du passé, marquée par une suite de restrictions susceptibles de se reproduire dans les années à venir ?

L'objectif de cette étude est de capter les impacts de la pandémie dans la modélisation de la prime pure du produit multirisque professionnelle (MRP) des artisans et commerçants d'Allianz. La démarche consiste à introduire dans les modèles des données liées à la Covid afin d'obtenir des niveaux de prédictions de sinistralité sur les périodes dites « standards », c'est-à-dire sans restriction, ainsi que sur les différentes périodes de restrictions : confinement total ou partiel, couvre-feu, … Pour chacune des étapes de la modélisation, après avoir évalué l'impact de la pandémie sur les données à modéliser, une méthode est proposée le cas échéant, pour adapter les modèles classiques d'analyse de prime pure.

Une approche traditionnelle est retenue pour la modélisation des attritionnels : modèle fréquence, coût moyen dont l'effet des variables explicatives sur le risque à expliquer est modélisé par un algorithme de type

GLM. En revanche, pour l'analyse des sinistres atypiques plus volatils et dont l'effet est plus difficile à capter, l'approche GLM est comparée à un modèle de *Gradient boosting*, dans le but de mieux segmenter la réallocation de sinistres atypiques par profils de risque. Bien que peu mis en pratique par les compagnies d'assurances, les modèles de *Machine Learning* ont fait leur apparition dans les mémoires de tarification depuis quelques années. Ces modèles sont potentiellement meilleurs pour identifier les facteurs explicatifs de la sinistralité.

Les travaux menés sont illustrés par une application sur les deux garanties principales de la multirisque professionnelle d'Allianz : l'incendie et le dégât des eaux, permettant ainsi de traiter une garantie d'intensité et une garantie de fréquence. Pour des raisons de confidentialité, les indicateurs d'origine et les résultats sont transformés. Les données sont traitées sous SAS Entreprise Guide et la modélisation utilise les logiciels Towers Watson (Emblem, Classifier, Radar) ainsi que Python.

Cette étude se décompose en quatre parties. Dans la première partie, le contexte et la problématique liés aux travaux sont détaillés dans le but de mettre en lumière la méthodologie et les apports de ce mémoire. Afin d'apporter une justification mathématique, la deuxième partie expose les fondamentaux et principes théoriques qui sous-tendent ces travaux. Les deux dernières parties s'attachent à la mise en pratique. La troisième partie présente les données et les traitements nécessaires pour constituer la base en entrée des modèles. Et la quatrième partie est consacrée à la modélisation et aux résultats.

1. Le cadre de l'étude

Le premier chapitre expose le contexte général, dans le but de mieux comprendre le problématique. Une première section décrit le marché de l'assurance des professionnels, et plus particulièrement le produit multirisque d'Allianz réservé aux commerçants et aux artisans, traité dans ce mémoire. Ces informations mettent en lumière les caractéristiques de la multirisque professionnelle qui en font un marché concurrentiel. Pour mieux comprendre la problématique, une deuxième section présente le contexte économique actuel lié la Covid, et la nécessité d'adapter les modèles de tarification.

1.1. La multirisque professionnelle

1.1.1. Généralités sur l'assurance des professionnels

L'assurance peut être définie au travers de deux notions : une juridique celle du contrat d'assurance et une technique celle de l'opération d'assurance.

Le contrat d'assurance est « une convention par laquelle, en contra partie d'une prime, l'assureur s'engage à garantir le souscripteur en cas de réalisation d'un risque aléatoire prévu au contrat. »

L'assurance est « une opération par laquelle un assureur organise en mutualité une multitude d'assurés exposés à la réalisation de certains risques et indemnise ceux d'entre eux qui subissent un sinistre grâce à la masse commune des primes collectées. »

L'assurance multirisque professionnelle (MRP) est un contrat global et complet qui vise à couvrir les risques portés par un professionnel. Ainsi, en contrepartie du versement d'une prime annuelle, un professionnel sécurise son activité. La MRP s'adresse à des structures de taille, de chiffre d'affaires, de statut et de secteur d'activité variés. Elle concerne aussi bien les professions libérales que les entrepreneurs du secteur médical, les commerçants, les artisans, les exploitants agricoles ou encore les associations.

Ce mémoire se concentre plus spécifiquement sur la multirisque professionnelle des **commerçants et des artisans**, et couvre des métiers hétérogènes. Un artisan n'est pas un commerçant et deux commerçants peuvent être très différents entre eux : un fleuriste n'a ni stock ni matériel, alors qu'un seul four de boulanger peut valoir 70 000 €. Les profils de risques sont très diversifiés, les situations d'exposition au risque sont variables selon l'activité : un boulanger n'aura pas les mêmes risques qu'un coiffeur, un taxi ou un antiquaire. Une activité saisonnière pourra nécessiter une adaptation en fonction des périodes de forte activité…

Quel que soit le secteur d'activité : opticien, boulanger, pâtissier, restaurateur, gérant d'un magasin de prêt à porter, ou encore artisan, un professionnel est soumis à des risques : incendie dans des locaux, vol de matériel, Responsabilité Civile engagée par un tiers, perte d'exploitation en cas d'interruption de l'activité.

Les risques couverts sont spécifiques à chaque secteur d'activité. Le principe de la multirisque est de proposer un large choix de garanties pouvant convenir à diverses professions :

- Des garanties pour couvrir les **dommages** aux locaux professionnels et leur contenu (mobilier, matériel, archives, fonds et valeurs). Parmi elles, les deux garanties étudiées dans ce mémoire : le **dégât des eaux** et l'**incendie** en font partie.
- Des garanties de **responsabilité civile** liées à l'activité, en cas de préjudice causé à d'autres personnes (clients, employés, voisins) dans l'exercice de l'activité. Par exemple, un incendie dans les locaux qui se propage au bâtiment voisin, un client qui se blesse dans une boutique, des articles vendus qui occasionnent des dommages ….
- Des garanties **financières** pour couvrir la perte de revenus. La garantie perte d'exploitation est une garantie essentielle pour les commerçants. Un incendie peut entrainer la fermeture des locaux et l'arrêt temporaire de l'activité. Cette garantie indemnise le manque à gagner subi dans ce cas. Son but est de reconstituer la marge brute réalisée en l'absence de sinistres et de couvrir les frais fixes pendant la période d'inactivité. Elle peut également indemniser en plus les pertes de nature

financière : intérêts ou indemnités de retard. Le professionnel a le choix de souscrire une perte d'exploitation en cas de dommages directs ou de dommages indirects.
- Des garanties complémentaires, comme la protection juridique pour chercher une solution amiable en cas de litiges avec des fournisseurs, salariés …

La garantie catastrophes naturelles est automatiquement souscrite quand une garantie dommage est souscrite. Un évènement catastrophe naturelle concerne l'inondation et la sécheresse, les lieux et la date sont définis par arrêté publié au Journal officiel. En France métropolitaine, la prime de cette garantie est fixe par la loi à 12% des primes dommages.

L'assurance multirisque professionnelle n'est pas obligatoire. Toutefois, elle est régie par le Code Civil qui détermine les conditions d'engagement de la responsabilité civile professionnelle d'une entreprise. La Responsabilité Civile professionnelle est obligatoire pour les professions réglementées, c'est-à-dire soumises à un cadre législatif et réglementaire particulier. Il s'agit entre autres des professionnels de santé, des professionnels du droit, des professionnels du tourisme et des professionnels du conseil.

Allianz propose un produit multirisque pour les commerçants et les artisans, dénommé Profil pro.

1.1.2. Le produit multirisque Allianz : Profil pro

➢ Profil pro au sein d'Allianz

• Présentation d'Allianz

Le **Groupe Allianz** créé en 1890 en Allemagne est le premier assureur européen devant Axa. Le groupe implanté dans plus de 70 pays, couvre les besoins d'assurances de 86 millions de clients et offre des services sur chacun des segments d'activité : vie, santé, dommages, gestion d'actifs et banque. La marque est placée première au monde dans le classement Interbrand[1].

Avec un chiffre d'affaires en 2019 de 142 milliards d'euros en hausse de 7,6% par rapport à 2018, le groupe Allianz n'échappe pas à la crise liée à la Covid-19, et affiche un chiffre d'affaires en baisse de 1,3 % en 2020 (140 Md€). L'impact de la crise sanitaire est estimé à 1,3 Md€ et s'accompagne d'une baisse de la plupart des indicateurs clés : le bénéfice net de 6,8 Md€ a diminué de 14% sur un an et le bénéfice d'exploitation à 10,8 Md€ a perdu 9%. Le ratio de solvabilité en baisse depuis 2018 passe de 229% en 2018, à 212% en 2019 pour atteindre 207% en 2020.

En **France**, Allianz est né de la fusion de plusieurs compagnies d'assurances avec à l'origine le rachat d'AGF en 1998 qui devient Allianz France en 2009. Son chiffre d'affaires est de 12,88 Md€ dont de 4,6 Md€ générés par l'assurance dommages.

La branche assurance dommage comporte plusieurs segments de marché : particuliers, entreprises, professionnels …. Le mémoire porte sur le segment de marché des professionnels qui représente 13% du chiffre d'affaires IARD d'Allianz France, dont près de la moitié dédiée aux risques « Incendie et Risques Divers des professionnels » : IRD Professionnels.

[1] Interbrand : filiale du groupe Omnicom, est un des leaders mondiaux du conseil en stratégie et design de marques.

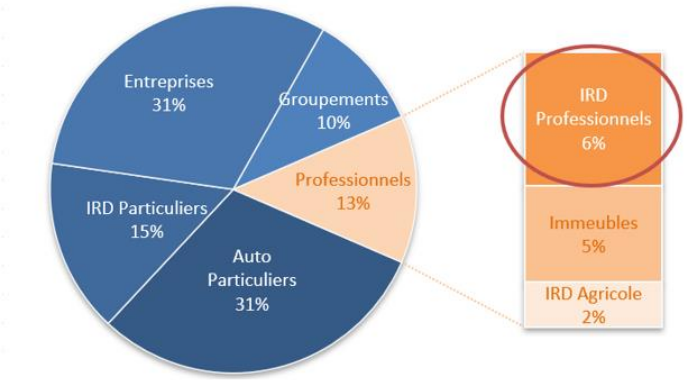


Figure 1-1 - Chiffre d'affaires sur les branches d'activité IARD d'Allianz France (source Allianz)

Le marché des professionnels se compose de trois segments dont la répartition en pourcentage du chiffre d'affaires IARD est : IRD Professionnels (6%), Immeubles (5%) et IRD Agricole (2%). L'étude porte sur le produit **Profil Pro**, principal produit de la ligne d'activité **IRD Professionnels**. Dédié aux commerçants et artisans, ce produit souscrit en agences ou par des courtiers, représente 47% des primes et 38% des polices de l'IRD Professionnels.

- Le poids des garanties du produit Profil Pro

Profil Pro s'adresse aux commerces de détail, activités de restauration et entreprises artisanales, avec le large choix de garanties typique d'une multirisque professionnelle. Les contrats peuvent être souscrits avec un panel de garanties dommages et responsabilité civile : multirisque, ou uniquement avec la garantie responsabilité civile : mono RC. La majorité des contrats sont souscrites en multirisque (88% du portefeuille), les souscriptions en mono RC ne représentent que 12% du portefeuille.

La répartition par garantie des contrats Profil Pro souscrits en multirisque est la suivante :

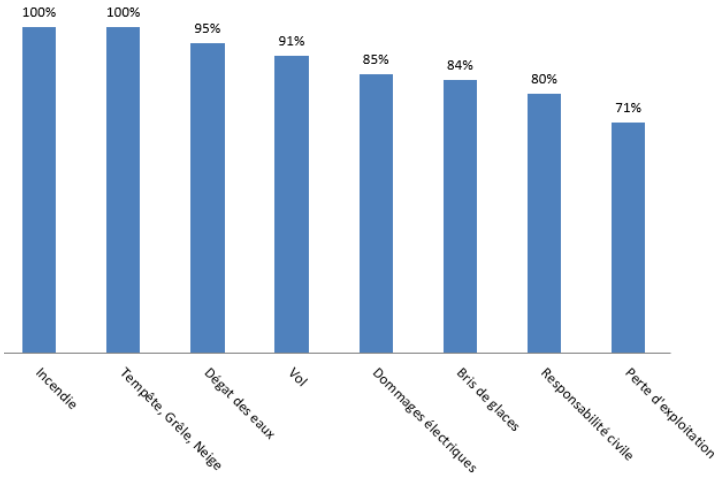


Figure 1-2 – Pourcentage de nombre de contrats sur les garanties principales

Les garanties **incendie** et **dégât des eaux** représentent les deux garanties les plus souscrites. Cela s'explique par le fait que, dans l'offre Allianz, la garantie incendie est obligatoire en multirisque (ainsi que la garantie tempête, grêle, neige), et qu'elle est très souvent complétée par un socle de garanties couvrant les dommages aux locaux dont la plus importante est le dégât des eaux, suivie du vol, des dommages électriques et des bris de glaces. D'autres garanties optionnelles plus spécifiques telles que la responsabilité civile liée à l'activité ou la perte d'exploitation peuvent être souscrites. A noter que le taux de souscription de 71% sur la garantie perte d'exploitation est supérieur à celui du marché proche de 50 %.

Ces garanties couvrent des risques de natures différentes qui se reflètent dans les niveaux de prime pure (sinistralité moyenne par contrat) :

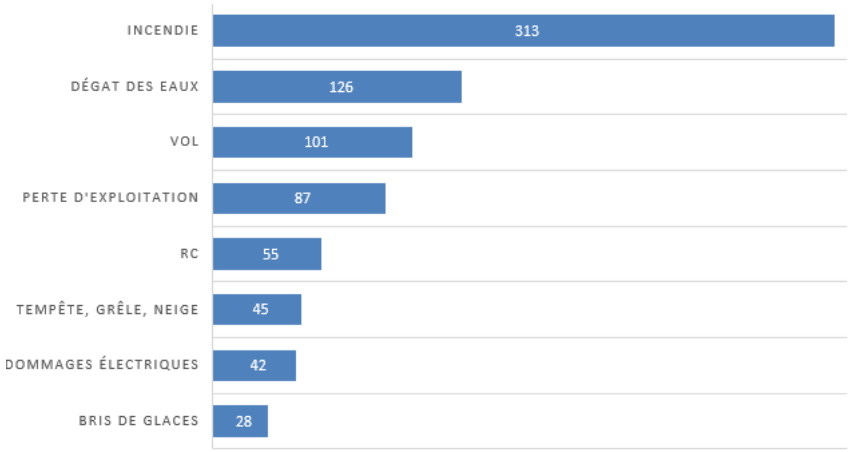


Figure 1-3 – Prime pure par garantie

La garantie incendie ressort avec une prime pure de 313 €, soit 2,5 fois plus élevée que celle de la garantie dégât des eaux dont la prime pure de 126 €. En effet, l'incendie couvre des risques pouvant amener à la destruction totale des locaux et du contenu d'un professionnel (et de ses voisins). Ces risques de faible fréquence peuvent atteindre les milliers d'euros. L'incendie est une garantie dite d'intensité. Contrairement à l'incendie, la garantie dégât des eaux couvre des risques d'intensité plus faible mais de fréquence plus élevée. Une fuite d'eau dans un local arrive plus fréquemment qu'un incendie

Hormis son large choix de garanties, le produit Profil pro présente des caractéristiques spécifiques de souscription.

➢ Profil Pro: la multirisque des commerçants et des artisans

• Les limites de souscription

L'activité n'est pas la seule condition d'entrée, la superficie, le contenu et le chiffre d'affaires ne doivent pas dépasser un certain niveau :

- superficie maximum 3 000 m² ;
- contenu maximum 3 000 000 € ;
- chiffre d'affaires maximum 5 000 000 €.

Ces limites s'appliquent à toutes les activités, exceptées : les glaciers, les traiteurs, les bars-café, les brasseries, les activités de restauration et les activités de tourisme, accueils de personnes âgées et autres hébergements.

• L'offre

Le produit Profil pro propose deux offres : une offre plus classique (Class pro) et une offre packagée (Qualité pro).

L'offre **Class Pro** permet au professionnel de choisir librement ses garanties avec une approche tarifaire garantie par garantie. Tandis que **Qualité Pro** est une offre simplifiée au niveau de la souscription et du tarif. Elle propose une couverture complète avec la possibilité de limiter le vol aux détériorations immobilières (vandalisme). Cette offre est réservée sous certains prérequis à un ensemble de professions déterminées et réparties en trois familles : commerces alimentaires, habillement et autres commerces de la rue. Le but de cette simplification est de permettre à un plus grand nombre d'agents non spécialisés dans le domaine des professionnels, de travailler sur ce marché.

Pour chacune de ces deux offres, est proposé en option un ensemble de « compléments » indissociables permettant d'atteindre un niveau de garanties « haut de gamme ». Cette option appelée « **complément plus** » couvre des événements supplémentaires. Par exemple, l'option prend en charge, pour la garantie incendie, l'effondrement des bâtiments à la suite d'un glissement de terrain accidentel non déclaré catastrophes naturelles, et pour la garantie dégât des eaux, les ruptures, fuites ou débordements de canalisations enterrées.

Un large choix de **franchises** est proposé, soit à partir d'un montant prédéterminé : de 0 à 3 000 €, soit à partir d'un niveau de réduction tarifaire, de 0 à 30 %. Un avantage spécifique est réservé aux affaires nouvelles souscrites avec une franchise au moins égale à 380 €. En l'absence de sinistre (hors catastrophes naturelles) pendant les deux premières années, la franchise est diminuée de moitié après cette période. Cet avantage est nommé « **franchise dégradable** ».

L'offre et le niveau de franchise caractérisent le niveau de risque souscrit, avec un impact sur le tarif. Ces deux caractéristiques font parties des variables explicatives de la sinistralité.

Les conditions d'entrée et la structure de l'offre sont importantes pour pouvoir évaluer le risque, mais la « colonne vertébrale » du produit est la nomenclature qui regroupe l'ensemble des activités.

➢ La nomenclature

• La structure des activités

La nomenclature regroupe l'exhaustivité des activités pouvant être souscrites sur le produit Profil Pro, près de **400 activités** y sont répertoriées. Le découpage assez fin des activités, propre à Allianz, permet de bien spécifier le risque associé. Par exemple, l'activité « Bar-Café » est distinguée en fonction de la présence ou non de bureau de tabac. Chaque activité est identifiée par un code et un libellé, définis par Allianz, toutefois la correspondance avec la nomenclature d'activité française (NAF) est également renseignée : un **code activité Allianz** peut être associé à plusieurs codes NAF.

Les activités du produit Profil Pro sont regroupées en douze rubriques dont le poids des contrats en portefeuille est représenté ci-dessous :

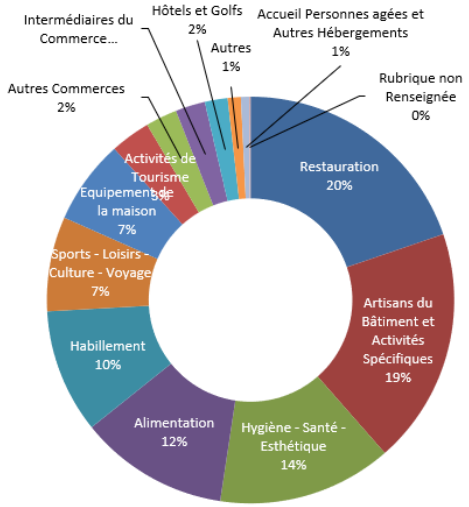


Figure 1-4 – Répartition des contrats par rubrique d'activité (source : Allianz)

La restauration et les artisans du bâtiment représentent à elles deux près de 40% du nombre de contrats en portefeuille.

La nomenclature stocke pour chaque activité, des informations liées à la souscription mais également des données tarifaires, par réseau de distribution et par garantie.

• Les informations de souscription

Un certain nombre d'informations propres à la souscription sont renseignées dans la nomenclature. Pour chaque activité sont spécifiés :

- l'éligibilité aux offres Class Pro / Qualité Pro ;
- les interdits de souscription ;
- les obligations de visites de risques ;
- les clauses : dispositions particulières permettant de clarifier les conditions d'exécution du contrat. Par exemple, la clause « Présence de chambres frigorifiques ou à atmosphère contrôlée de plus de 300 m^3 », pour spécifier la capacité totale des chambres frigorifiques à ne pas dépasser.

• Les classes de risque

Chaque activité appartient à une classe de risque sur chacune des garanties suivantes : incendie (Inc), perte d'exploitation (PE), dégât des eaux (DDE), vol, responsabilité civile (RC) et protection juridique (PJ). Les classes de risques ont été déterminées par les experts métiers, et regroupent les activités selon leur niveau croissant d'exposition au risque : de 1 à 6 en incendie et de 1 à 3 en dégât des eaux. Elles sont généralement identiques selon le réseau de distribution : agent, courtage. Ces classes sont tarifaires et leur attribution est revue tous les ans lors des revalorisations : certaines activités peuvent être reclassées à la hausse après constatation de mauvais résultats techniques.

Ci-dessous un extrait de la nomenclature sur des activités des rubriques « Alimentation » et « Esthétique » :

Code	Activité	Sous-activité	Classes tarifaires											
			Agences						Courtage					
			Inc	PE	DDE	Vol	RC	PJ	Inc	PE	DDE	Vol	RC	PJ
01-04-2	Boulangerie, Pâtisserie (artisan)	avec four électrique	5	3	2	2	1	1	5	3	2	2	1	1
01-05-1	Vente de pain ou de viennoiseries sans fabrication, avec ou sans terminal de cuisson (détaillant)		3	3	2	2	2	1	3	3	2	2	2	1
05-01-1	Coiffeur	(exercice en salon)	1	1	2	2	1	2	1	1	2	2	1	2
05-01-2	Coiffeur	à domicile	6	6	3	5	1	1	6	6	3	5	1	1

Figure 1-5 – Extrait de classes de risque de la nomenclature

Un boulanger qui possède un four pour la fabrication de ses produits est plus exposée au risque incendie qu'un simple distributeur de pain et de viennoiserie sans fabrication. La classe de risque incendie d'un boulanger, pâtissier ressort à 5 dans la nomenclature, alors qu'elle n'est que de 3 pour la vente sans fabrication de pain. Un autre exemple, celui du coiffeur dont l'exposition au risque est plus élevée pour une activité exercée à domicile plutôt qu'en salon. Exercer une activité chez un particulier nécessite de couvrir les dommages qui peuvent être causés au domicile de la personne.

Avec près de 400 activités présentes dans la nomenclature du produit Allianz, l'assurance multirisque des professionnels couvre un vaste secteur d'activité avec un large choix de garantie adapté à un grand nombre de professionnels. Cette diversification d'activité offre un grand nombre de clients potentiels et représente un atout pour les assureurs. Le marché des professionnels possède d'autres avantages qui en font un secteur de plus en plus prisé et concurrentiel.

1.1.3. Un marché attractif pour les assureurs

Les professionnels représentent une cible particulièrement intéressante pour les assureurs : masse assurable importante, avec des besoins d'assurance qui s'étendent au-delà de leur activité. Une segmentation adéquate du tarif est indispensable pour se démarquer de la concurrence et évaluer les risques au plus juste.

➢ Des secteurs d'activités diversifiés

D'après les dernières statistiques de l'Insee sur les locaux commerciaux, à fin 2017 la France comptait 300 000 points de vente dans le commerce de détail et l'artisanat commercial. En moyenne, ces commerces génèrent un chiffre d'affaires de 1,2 million d'euros, occupent une surface de vente de 240 m² et emploient cinq personnes à temps pleins.

• Des caractéristiques propres à chaque secteur

D'après des informations transmises par l'Insee, les caractéristiques des commerces : chiffre d'affaires, surface de vente, effectif et nombre de points de vente diffèrent selon le secteur d'activité et se répartissent de la manière suivante :

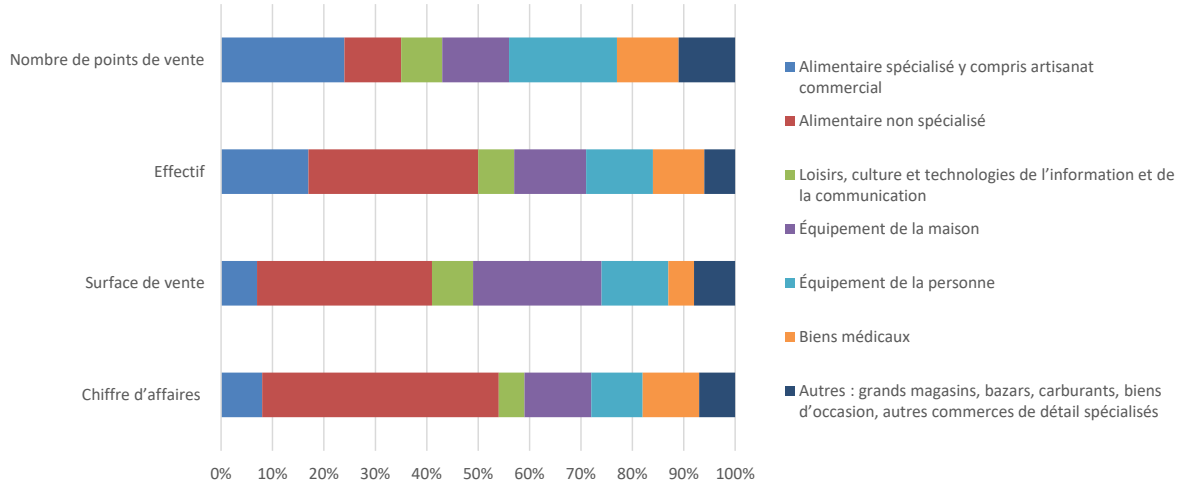


Figure 1-6 – Caractéristiques des commerces de détail et de l'artisanat (source Insee)

La figure ci-dessus montre les disparités entre les types de commerces. Un tiers des points de vente est occupé par des **commerces alimentaires**. Parmi eux, l'alimentaire non spécialisé (supérette, supermarché, hypermarché) avec peu de points de vente (11 %), contribue le plus aux surfaces commerciales et au chiffre d'affaires. En comparaison, l'alimentaire spécialisé (boulangerie, boucherie, maraîcher, chocolatier) dispose de deux fois plus de points de vente (24 %), mais génère six fois moins de chiffre d'affaires (8 %).

Des divergences ressortent également dans le commerce **non alimentaire**. L'équipement de la personne (habillement, chaussure, maroquinerie, hygiène-beauté, horlogerie-bijouterie) compte le plus grand nombre de points de vente (21 %) pour 13 % des emplois, 13 % des surfaces et seulement 10 % du chiffre d'affaires. L'équipement de la maison (électroménager, textile, quincaillerie, meubles) dispose de surfaces commerciales particulièrement importantes (25 %), alors que ce secteur ne représente que 13 % des points de vente et du chiffre d'affaires et 14 % des emplois. À l'inverse, les commerces de biens médicaux (pharmacies) possèdent des surfaces de vente beaucoup plus petites (5 %) pour des parts de points de vente, d'emplois ou de chiffre d'affaires équivalentes (11 %).

Les commerces alimentaires et non alimentaires présentent des caractéristiques très différentes en termes de superficie, chiffre d'affaires, effectifs, nombre de points de vente. Cette diversité des activités complexifie le tarif, et nécessite de segmenter les activités par groupes de risques homogènes.

Ces distinctions entre commerces influent sur la localisation et la fréquentation. Les grandes surfaces se situent généralement en périphérie des grandes villes tandis que les commerces de petite taille sont plus souvent localisés en centre-ville.

• Des emplacements stratégiques

L'emplacement géographique fluctue selon les secteurs d'activité. Certains commerces s'implantent naturellement à proximité de leur clientèle, d'autres privilégies les zones moins peuplées avec plus de surface.

La surface commerciale des secteurs des biens médicaux et de l'alimentaire spécialisé est fortement liée à la population de la commune. Pour les biens médicaux, les autorisations d'implantation de pharmacies dépendent directement de la population. À l'opposé, pour l'équipement de la maison, les magasins de ce secteur s'implantent généralement dans des zones moins peuplées pour disposer de surfaces plus grandes, et couvrent une clientèle sur des zones qui s'étendent au-delà de la commune d'implantation.

La surface des magasins dépend également de l'aire urbaine. Les grandes villes comme Paris, Lyon, Marseille disposent de plus de surfaces pour l'équipement de la personne. Dans les communes en périphérie, les surfaces de l'alimentaire spécialisé sont proportionnellement moins importantes.

La localisation influe sur l'exposition au risque. Un professionnel situé en zone rurale est plus isolé que dans une grande ville, avec une distance aux pompiers et une vitesse d'intervention plus élevées en cas d'incendie. En revanche le risque de vol est plus élevé. Une bonne segmentation par zone géographique (zonier) est indispensable pour un bon tarif.

➢ Un client enclin au multi-équipement

Un professionnel qui veut assurer son activité possède l'avantage pour l'assureur d'avoir une double casquette, celle de professionnel et de particulier. Le professionnel qui souscrit une multirisque a de fortes chances de confier d'autres contrats à son assureur : l'assurance de véhicules, la prévoyance du dirigeant et des salariés, mais aussi tout ce qui relève du privé (habitation, auto …). Un professionnel a couramment cinq ou six contrats chez le même assureur.

Un des enjeux pour un assureur est d'augmenter son ratio de multi-équipement, c'est à dire le nombre de contrats d'assurance détenus par un client. En effet, un client multi-équipé est moins volatile.

➢ Stratégie de conquête des assureurs et bancassureurs

Les grands acteurs de l'assurance dommage affichent depuis quelques années un intérêt grandissant pour le marché des professionnels. A titre d'exemple, Axa a racheté en 2018 l'assureur et réassureur XL GROUP spécialisé dans l'assurance dommages des PME.

Par ailleurs, la crise sanitaire actuelle, avec les prêts garantis par l'état, a renforcé la relation entre les banquiers et leurs clients professionnels. Ce contact avec les clients est une opportunité pour les bancassureurs qui ont fait leur entrée sur le marché des professionnels depuis plusieurs années. Les deux premiers bancassureurs sur le marché du dommage des professionnels : le Crédit Mutuel (avec 1% de part de marché) et le Crédit Agricole renforcent leur stratégie de développement sur ce secteur. Les bancassureurs profitent de ce contexte propice pour proposer de nouvelles offres comme l'expliquent ABADIE A. (2021) ou KARAYAN R. (2020) dans des articles de L'argus de l'assurance.

➢ Un secteur en évolution

Le secteur des pros évolue avec des pratiques qui changent et des risques qui se transforment. Les points de retrait de colis et la vente à distance se développent chez les petits commerçants, de nouvelles activités comme les « food trucks » apparaissent.

Les contrats et les garanties doivent s'adapter à ces nouveaux risques : « Les contrats sont en évolution continuelle, en fonction de l'activité des pros, remarque Nadine Durand, experte des risques des professionnels au sein d'Allianz France. Par exemple, la manière d'appréhender les garanties sur le matériel informatique a

évolué. Auparavant, les ordinateurs portables faisaient l'objet de garanties spécifiques, souvent restrictives. Désormais, leur protection est en général incluse dans celle du matériel informatique. » (L'argus de l'assurance, « Risques d'entreprise : la multirisque, produit d'appel des pros », Séverine Charon, 14/11/2019).

➢ Un secteur rentable … jusqu'en 2019

D'après les données de France Assureurs, le secteur des multirisques des artisans, commerçants et prestataires de services (ACPS) représente 1,827 Md€ de cotisations en 2020, soit près de 27 % des cotisations des assurances de dommages aux biens des professionnels.

Sur les multirisques ACPS, l'année 2020 est particulièrement atypique avec des cotisations en baisse de 0,7 % par rapport à 2019, et une forte croissance de la sinistralité (fréquence et coût moyen).

	2016	2017	2018	2019	2020
Fréquence moyenne annuelle					
Tous sinistres	**106 ‰**	**104 ‰**	**111 ‰**	**100 ‰**	**128 ‰**
dont Incendie	10 ‰	9 ‰	10 ‰	9 ‰	9 ‰
dont TGN	4 ‰	7 ‰	7 ‰	6 ‰	5 ‰
Coût moyen (€)					
Tous sinistres	**3 555**	**3 500**	**3 785**	**3 970**	**5 440**
dont Incendie	14 560	16 140	16 645	19 605	15 215
dont TGN	3 795	3 435	4 150	4 855	3 540

Figure 1-7 - Fréquence et coût moyen des multirisques ACPS (source : France Assureurs)

L'année 2020 impactée par la crise sanitaire est marquée par une fréquence des sinistres en hausse de 28 % et un coût moyen en hausse de 37%. De tels niveaux de fréquence et de coût moyen n'ont jamais été observés depuis 2016. Ces hausses ne sont portées ni par l'incendie ni par les évènements climatiques (TGN : tempête, grêle, neige) dont les fréquences sont assez stables et dont les coûts moyens sont en baisse par rapport à 2019.

La hausse de la sinistralité se reflète dans le S/P (ratio sinistre à primes) :

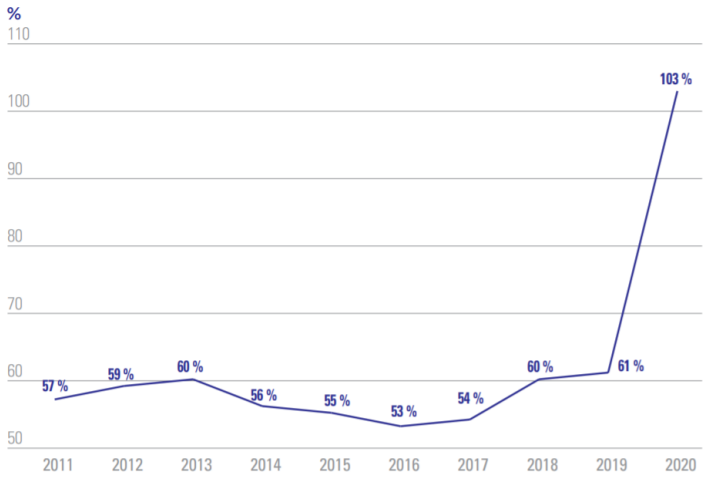


Figure 1-8 - Ratio sinistres à primes des multirisques ACPS (source : France Assureurs)

Ce secteur d'activité particulièrement rentable jusqu'en 2019 avec un S/P de l'ordre de 60 %, voit ses résultats fortement se dégrader en 2020. Avec 103 % de S/P, la multirisque des artisans, commerçants et prestataires de services jusque-là rentable, ressort déficitaire en 2020.

1.2. Problématique

1.2.1. Un contexte économique tendu : la Covid-19

Pour faire face à l'épidémie de Covid-19, l'année 2020 a connu des périodes successives de restriction avec la mise en place de différentes mesures : confinement, couvre-feu. La première période de restriction, le confinement à la suite du décret n°2020-293 du 23 mars 2020, a été la plus impactante pour les commerçants. L'impact de cet incident diffère par catégories d'activités et par régions.

➢ L'impact du premier confinement par secteur d'activité

Sur la base des établissements du commerce de détail et de l'artisanat commercial à fin 2017 en France, l'Insee a répertorié les points de ventes touchés par le décret : sur 300 000, 136 000 sont soumis à l'obligation de fermeture, soit 45 %.

La part des points de vente fermés dépend du secteur d'activité :

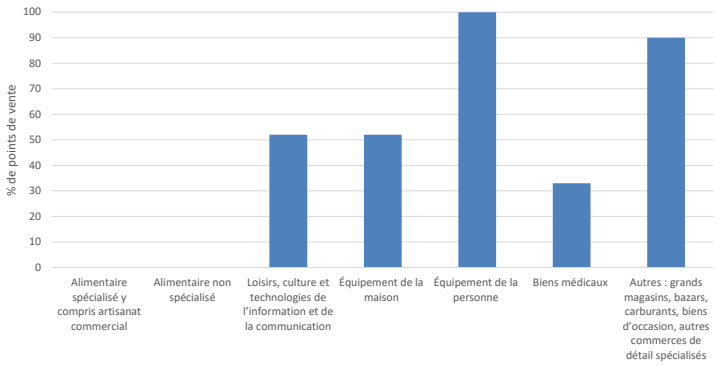


Figure 1-9 - Part des secteurs concernés par le décret du 23 mars 2020 (source : Insee)

Les secteurs d'activités n'ont pas tous été impactés de la même manière. La continuité de l'activité n'était autorisée que pour les commerces de détail alimentaire spécialisés et non spécialisés, tandis que dans le secteur de l'équipement de la personne tous les commerces étaient contraints de fermer.

➢ L'impact du premier confinement par région

La part de la surface commerciale fermée varie également selon les départements :

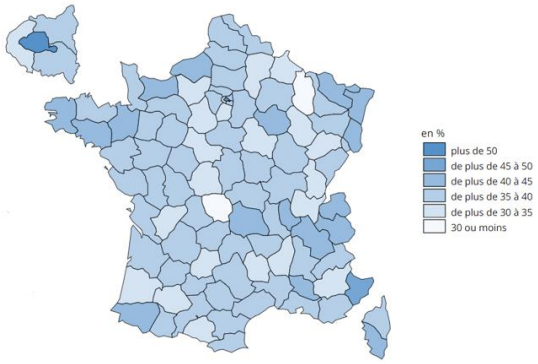


Figure 1-10 - Surface commerciale concernée par le décret du 23 mars 2020 (source : Insee)

Les départements sont impactés en fonction de la composition des commerces : plus de 50 % de la surface commerciale à Paris est concernée par le décret contre moins de 30 % dans la Creuse.

La première mesure mise en place a été le premier confinement a été. Face à l'ampleur de la pandémie, cette mesure a été très restrictive et a surtout touché les commerces n'étant pas considérés comme de première nécessité, situés dans les grandes villes. Les différentes mesures qui ont suivi ont été certes moins restrictives, cependant elles ont eu des répercussions sur les habitudes des français et des commerçants, avec un effet sur l'ensemble du secteur.

1.2.2. L'impact des mesures de restrictions

➢ Sur le marché des professionnels

D'après les sources de l'Insee, les créations de commerces en France était en croissance de 9,3% en 2020. Malgré la crise sanitaire, l'année 2020 a enregistré un bon niveau de croissance : la chute du nombre de création lors du premier confinement a été rattrapée par un rythme soutenu des créations jusqu'à la fin de l'année.

Bien qu'en 2020, le nombre de faillites des entreprises soit en baisse de 9% grâce aux aides de l'état, au gel de cessation de paiement et à la fermeture des tribunaux de commerce pendant le confinement, l'entreprise d'assurance-crédit Euler Hermes[2] prévoit une hausse de défaillance des entreprises de 32% en 2021. Les assureurs pourraient être confrontés à des risques de **non-paiements** ou retards de paiement des primes.

La crise sanitaire pourrait également s'accompagner d'une augmentation des déclarations **frauduleuses** chez les professionnels : hausse des réclamations de restaurants concernant la détérioration d'aliments à la suite de pannes du système de réfrigération, augmentation des déclarations de cambriolage des magasins de détails, fausses déclarations de dégât des eaux dans les entrepôts qui n'arrivent pas à écouler leur stock de marchandises invendues en raison des confinements.

Une autre répercussion de la crise sanitaire est la **détérioration de l'image** des assureurs liée au mécontentement des assurés dont la **perte d'exploitation** n'a pu être indemnisée. En effet, dans le cadre de la Covid-19, la garantie perte d'exploitation n'intervient que rarement. Tout d'abord parce que la perte d'exploitation sans dommages est beaucoup moins souscrite que la perte d'exploitation consécutive à des dommages. Or aucun dommage matériel n'est lié à l'arrêt d'une activité en cas d'épidémie. Ensuite, parce que l'épidémie, phénomène systématique et généralisé, apparait dans pratiquement tous les contrats comme étant non assurable. Certains assureurs cherchent de **nouvelles solutions assurantielles.** Dans un article des Echos, POULLENNEC S. (2020) évoque la collaboration de l'assureur Generali avec l'UMIH (Union des Métiers et des Industries de l'Hôtellerie) pour mettre en place une nouvelle offre d'assurance multirisque professionnelle.

Dans ce contexte, les assureurs ont accepté de ne **pas augmenter** en 2021 les **cotisations** des contrats d'assurance multirisque professionnelle pour les secteurs d'activités les plus **touchés par la crise** (secteurs de l'hôtellerie, des cafés, de la restauration, du tourisme, de la culture, du sport et de l'événementiel). Toutefois, la pandémie touche l'ensemble des activités avec un impact direct sur la sinistralité. Ce gel des primes a des répercussions sur la politique tarifaire des compagnies d'assurance comme l'évoque ROBERT T. (2020) sur Assurlandpro.

➢ Sur la sinistralité du produit Profil pro

Les commerçants ont dû s'adapter aux mesures successives de confinement et de couvre-feu : moins de présence dans les locaux, plus de vente en ligne …. Ces changements de comportement ont un impact sur la sinistralité observée pendant ces périodes de restrictions. Cet effet se constate notamment sur les deux garanties principales du produit Profil pro : l'incendie et le dégât des eaux.

[2] Euler Hermes : entreprise d'assurance-crédit, filiale du groupe Allianz.

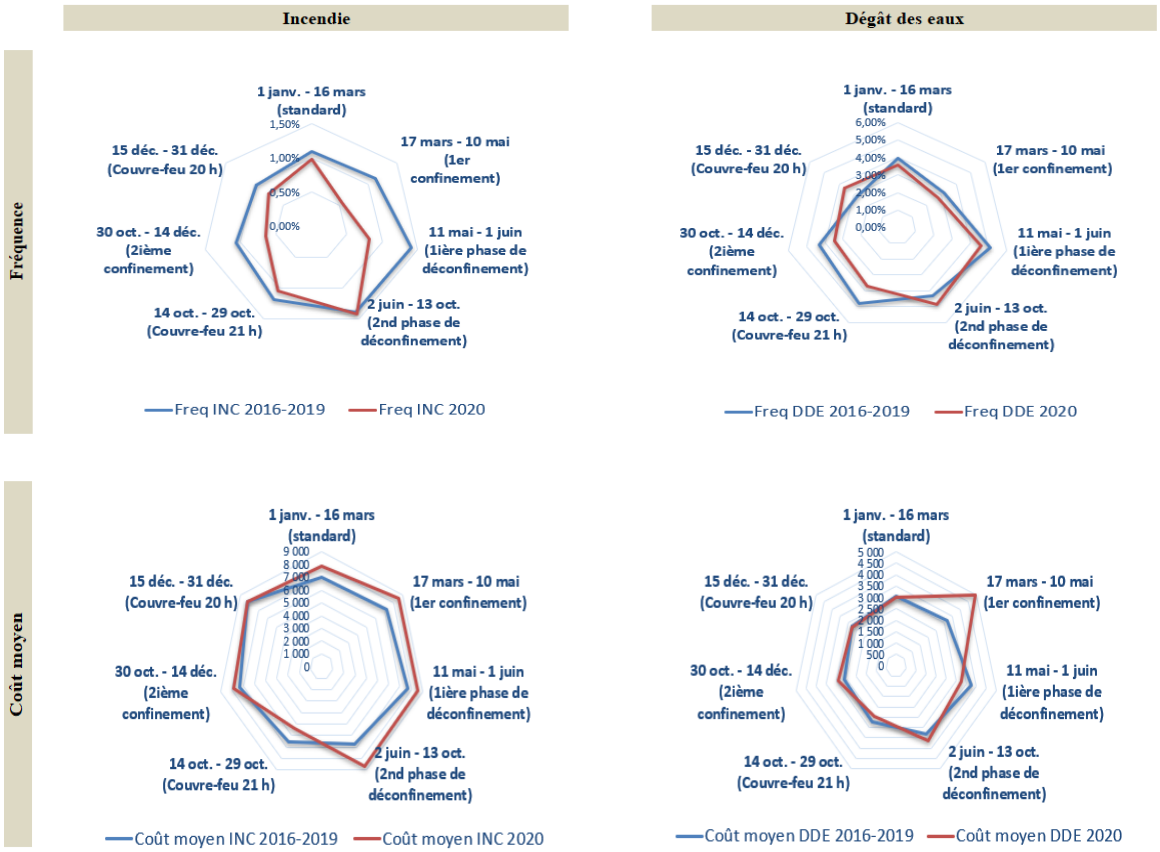


Figure 1-11 - Variation des fréquences et coûts moyens écrêtés par périodes de restrictions

Pour visualiser l'impact de la pandémie sur la sinistralité observée sur ces deux garanties, les indicateurs de sinistralité : fréquence et coût moyen écrêté à 50 k€, de la survenance 2020 sont comparés à la moyenne des années 2016 à 2019, vus avec un délai de vieillissement identique : fin mars de l'année N+1. La comparaison est réalisée sur chacune des périodes affectées par une restriction sanitaire, et sur la période sans restriction dite « standard ».

Les effets constatés des périodes de restriction sur la sinistralité 2020 sont une baisse de la fréquence et une hausse des coûts moyens. En incendie, la baisse de la fréquence est plus importante sur les périodes sans activité : premier et deuxième confinement. Certes, les restrictions engendrent une baisse de la fréquence mais les périodes de reprise peuvent avoir l'effet inverse. C'est ce qui est observé sur la fréquence dégât des eaux avec une hausse de la fréquence lors de la seconde phase de déconfinement et la période de couvre-feu à 20 h. Cette hausse apparait sur deux périodes de reprise d'activité après une période de confinement. Une reprise d'activité dans des locaux inoccupés pendant plusieurs semaines peut impacter l'état de la tuyauterie. Concernant le coût moyen, l'effet le plus marquant est observé sur la fréquence dégât des eaux lors du premier confinement. En effet, cette période a été la plus contraignante en termes de restrictions, les commerçant n'allant que rarement dans leurs locaux, la découverte plus tardive des sinistres entraine une dégradation plus importante.

En 2020, la fréquence et le coût moyen des deux garanties principales du produit Profil pro se sont sensiblement écartées des tendances observées les années précédentes. Les écarts constatés, à la hausse ou à la baisse, sont propres au risque couvert par chacune des garanties et diffèrent d'une mesure de restriction à l'autre. Ces chocs inattendus chamboulent le pouvoir prédictif de la tarification qui repose sur l'estimation de la prime pure (coût probable des sinistres). Les modèles de tarification traditionnels sont-ils pour autant remis en cause ?

1.2.3. L’adaptation des modèles de prime pure

Les **modèles linéaires généralisés** (GLM), du fait de leur transparence et leur interprétabilité sont encore aujourd’hui la norme en termes de tarification. Dans ce contexte de crise, les tensions économiques et concurrentielles obligent à les optimiser et à les adapter.

Ce mémoire se place dans la cadre de la refonte en 2021, des modèles de prime pure du produit Profil pro, multirisque professionnelle dédiée aux commerçants et aux artisans. L’un des principaux enjeux est d’identifier une segmentation optimale qui permet de définir des profils de risques homogènes. Cette segmentation est indispensable pour se distinguer de la concurrence. L’activité et le lieu géographique sont deux critères essentiels dans la segmentation du tarif du produit Profil Pro. Cependant, ils ne peuvent être intégrés tels quels dans la modélisation. Avec plus de quatre cents activités couvertes par ce produit, certaines cases tarifaires n’auraient que très peu d’effectifs, et l’estimation obtenue à partir des observations empiriques ne serait pas significative. Quant au lieu géographique, les seules données fiables et normalisées dans les bases de données sont le code postal et le code Insee. Cette maille présente une granularité trop fine et un manque d’information géographique pour être exploitée telle quelle. Pour remédier à ces problèmes, les informations liées à l’activité et au lieu géographique seront définies par des variables dites « latentes ». Une **variable latente** est une variable qui ne peut pas être mesurée directement et qui n’est pas observée mais qui est à l’origine d’autres variables observées. Elle est issue d’algorithmes ou de modélisations statistiques et elle synthétise plusieurs variables observées.

La modélisation de la prime pure, définie comme le produit de la fréquence et du coût moyen, requière de distinguer les sinistres de montant important (sinistres atypiques), des sinistres de faible montant et de fréquence élevée (attritionnels). Le faible volume des sinistres **atypiques** rend leur segmentation plus difficile à déterminer. Pour les modéliser, des algorithmes de ***Machine Learning*** sont comparés aux GLM classiques.

Le contexte actuel de pandémie remet en cause la stabilité dans le temps des modèles statistiques qui s’appuient habituellement sur un historique de périodes annuelles. Cette périodicité est retenue pour prendre en compte les éventuelles évolutions du risque au cours de la vie du contrat (modification du montant de capital assuré, changement d’activité …), et pouvoir rattacher la sinistralité à la bonne vision du risque. A cause de l'évolution dans le temps des mesures de restrictions adoptées en 2020 avec des durées et des étendues différentes, la vision par année d’accident habituellement retenue dans les modèles n’est plus adaptée. Une vision annuelle ne permet pas de capter les changements de comportement qui ont fait suite aux différentes mesures de restriction.

Une solution pourrait être de passer d'une vision annuelle à une vision plus granulaire comme le mois d’accident. Cette option est utile pour détecter rapidement les principales tendances à modéliser, mais dans une perspective de modélisation, le mois d’accident n'a pas de réelle corrélation avec le phénomène sous-jacent à analyser, et est plutôt lié à l'application des restrictions de mobilité et leurs implications. De plus, un déroulement mensuel des données s'accompagne d'une augmentation conséquente de la taille de la base de données qui peut influencer l'efficacité des modèles.

Une autre possibilité est de dérouler les données en utilisant un **indicateur des restrictions** et en le reliant aux impacts sur la mobilité sociale. Cette solution, plus appropriée notamment du point de vue de la modélisation, est retenue. L’indicateur sélectionné correspond aux **périodes de restrictions** de mobilité dont la granularité s’appuie sur la source officielle du site de l’ECDC (European Centre for Disease Prevention and Control : Centre européen de prévention et de contrôle des maladies). Ces périodes sont détaillées dans le Chap.3 § 1.5.

Après cette présentation du contexte et de la problématique, la solution proposée pour y répondre demande la mise en application de méthodes décrites dans le chapitre suivant.

2. Les aspects théoriques

Ce chapitre expose les aspects mathématiques des méthodes appliquées dans ce mémoire. La première section présente les modèles de fréquence et de coût moyen, base de la modélisation de la prime pure qui nécessite la distinction entre sinistres attritionnels et sinistres atypiques pour ne pas perturber la loi de distribution des sinistres. La deuxième section quant à elle, s'intéresse aux méthodes classiques de modélisation : les modèles linéaires généralisés (GLM). Ensuite, la troisième partie s'attache à la l'identification du risque géographique et aux étapes nécessaires à la mise en place d'un zonier. La dernière section est consacrée aux méthodes de *Machine Learning* utilisée d'une part pour la classification des activités et d'autre part pour la régression logistique en comparaison du GLM.

2.1. Les modèles fréquence, coût moyen

2.1.1. Généralités

L'objectif des modèles de fréquence et de coût moyen est d'identifier les critères pouvant avoir une incidence sur la sinistralité future. Les assurés sont regroupés en classes de risques homogènes afin de pouvoir appliquer un tarif commun à tous les assurés au sein d'une même classe.

En raison de la loi des grands nombres, la sinistralité future doit se rapprocher de la somme attendue d'un grand nombre de sinistres indépendants, relativement plus prévisible que celle d'un individu. C'est pourquoi l'analyse de la sinistralité s'intéresse à la charge des sinistres engendrés par un groupe d'assurés dont les risques sont relativement homogènes et non contrat par contrat. La charge aléatoire globale écrêtée $\boldsymbol{S}$ s'écrit alors comme la somme sur le nombre aléatoire **N** de sinistres, des montants écrêtés aléatoires $\mathbf{Y_i}$ de chaque sinistre, d'où le nom de modèle **collectif** :

$$S = \sum_{i=1}^{N} Y_i \, .$$

En appliquant le théorème de Wald (comme détaillé dans le cours de BARADEL N. (2020) sur la théorie du risque), la prime pure : espérance du coût des sinistres, se décompose selon un modèle de **fréquence** et un modèle de **coût** :

$$\mathbb{E}[S] = \mathbb{E}[N]\,\mathbb{E}[Y].$$

Les variables N et les $(Y_i)_{i\geq 1}$ sont supposés indépendantes, et les $(Y_i)_{i\geq 1}$ identiquement distribuées. Modéliser le nombre de sinistres indépendamment de leur montant permet de les expliquer par des critères tarifaires propres à chacun. De plus, la connaissance empirique du nombre de sinistres est beaucoup plus stable que celle de leur coût : le montant final peut être connu plusieurs années après la survenance du sinistre, tandis que les sinistres sont généralement tous déclarés dans les deux ans qui suivent la survenance.

En multirisque professionnelle, chaque assuré est susceptible d'avoir plusieurs sinistres sur une année. En supposant que le nombre de sinistres suit une loi de Poisson, la charge probable S^j de l'assuré j peut s'écrire sous la forme d'un modèle collectif :

$$S^j = \sum_{i=1}^{N^j} Y_i^j.$$

La prime pure s'écrit :

$$\mathbb{E}[S^j] = \mathbb{E}[N^j]\,\mathbb{E}[Y^j].$$

La sinistralité attendue varie selon les caractéristiques x^j propres à chaque assuré. Pour des assurés appartenant à des classes de risques homogènes, sous condition d'indépendance de la fréquence et des coûts des sinistres, le modèle revient à modéliser l'espérance conditionnelle du nombre de sinistres $\mathbb{E}[N^j / X^j = x^j]$ et l'espérance conditionnelle du coût unitaire $\mathbb{E}[Y^j / X^j = x^j]$ à l'intérieure de chaque classe de risque.

Les statistiques fournissent un ensemble d'outils pour estimer la relation $f(X)$ entre une variable réponse Y et une ou plusieurs variables indépendantes X. Les modèles les plus utilisés en tarification sont les Modèles Linéaires Généralisés (**GLMs** : Generalized Linear Models) introduits dans les sciences actuarielles dans les années 90 (Brockman et Wright, 1992; Haberman et Renshaw, 1996; Murphy et al., 2000). Les GLMs particulièrement appropriés pour estimer ces espérances conditionnelles, ont l'avantage d'être facilement interprétables et directement utilisable pour la tarification.

Ces modèles offrent la possibilité de segmenter le portefeuille tout en évaluant le risque relatif de chaque segment. Une segmentation trop fine conduit à des tarifs quasiment individualisés et non mutualisés, et demande de disposer d'un volume de données suffisant pour appliquer la loi des grands nombres et avoir une volatilité faible de l'estimateur. Un compromis est à faire entre la précision du tarif et la significativité statistique.

La construction de la grille tarifaire nécessite de définir des classes de risques suffisamment homogènes. La problématique consiste à savoir comment segmenter la population assurable, c'est-à-dire quels sont les critères tarifaires à retenir pour constituer la grille tarifaire. Pour modéliser la prime pure, plusieurs types d'informations sont disponibles dans les bases de données. Tout d'abord, les données liées au **risque** :

- l'activité : permet d'identifier les risques spécifiques auxquels est soumis le professionnel ;
- les caractéristiques du professionnel à assurer : superficie, valeur du contenu en euros, chiffre d'affaires… ;
- la localisation du professionnel ;
- les besoins de celui-ci : garanties du contrat, niveau de couverture souscrit ;
- la nature et la valeur des biens à assurer.

La collecte des données ne se limite pas à celles du risque. Des informations internes propres aux **clients**, le nombre de contrats détenu par le client au sein d'Allianz par exemple, sont également très utiles pour mieux comprendre le comportement du client. Mais aussi, de nombreuses données **externes** disponibles en libre accès grâce au développement de l'Open Data, comme les données géographiques apportent une source d'information sur le lieu géographique.

Les risques couverts par la multirisque professionnelle peuvent être de grande ampleur, particulièrement en incendie, et se chiffrer à plusieurs millions d'euros. Ces sinistres rendent les estimations très volatiles, et leur effet doit être réduit en écrêtant leur coût à partir d'un certain **seuil**. La détermination de ce seuil garantie par garantie est le sujet de la section suivante.

2.1.2. Le choix des seuils d'écrêtement des sinistres atypiques

Un sinistre atypique est un sinistre qui se produit avec une faible fréquence et un coût important. C'est un risque homogène et reproductible au fil du temps qui peut être modélisé à partir d'un historique de données internes. Le local d'un professionnel qui prend feu est un exemple de sinistre atypique.

La présence de sinistres atypiques dans les données impacte les modèles pour trois raisons principales :

- la queue de distribution des coûts empiriques des sinistres est plus lourde que prévue et se caractérise par peu de sinistres avec un coût très important ;
- les sinistres dans la queue de distribution sont aussi très volatiles d'une année à l'autre et rend leur estimation plus difficile ;
- certains segments de risques très spécifiques peuvent être plus touchés par les sinistres atypiques, créant un biais dans les estimations.

C'est pourquoi les coûts des sinistres atypiques doivent être retraités des données avant la modélisation. Deux méthodes sont possibles : écrêter le coût lorsqu'il dépasse un certain seuil, ou séparer les sinistres selon leur coût. La méthode d'écrêtement a été retenue dans ce mémoire.

➢ La distribution des sinistres

Les sinistres atypiques sont des sinistres avec une faible fréquence et un coût moyen important. Une manière simple d'identifier le seuil des sinistres atypiques est de trouver le bon compromis entre un pourcentage faible du nombre de sinistres au-dessus du seuil et un pourcentage élevé de leur coût au-dessus du seuil.

L'idée est assez simple et repose sur la distribution empirique des pertes, c'est-à-dire l'histogramme des coûts de sinistres observés. L'analyse s'appuie sur un certain nombre de KPIs directement liés à la distribution des sinistres. Plus précisément, les couûs observés sont discrétisés et pour chaque tranche les KPI suivants sont calculés en considérant comme seuil la borne supérieure de la tranche :

- le nombre de sinistres en-dessous et au-dessus du seuil ;
- le coût total en-dessous et au-dessus du seuil ;
- le pourcentage de sinistres en-dessous et au-dessus du seuil ;
- le pourcentage du coût en-dessous et au-dessus du seuil ;
- le coût moyen des sinistres en-dessous et au-dessus du seuil ;
- le coût total si les sinistres étaient capés au seuil, c'est-à-dire :

$$\text{seuil} \times (\text{nombre de sinistres au dessus du seuil}) + \text{coût des sinistres sous le seuil};$$

- le coût moyen si les sinistres étaient capés au seuil, c'est-à-dire :

$$\text{coût total si les sinistres étaient capés au seuil} / \text{nombre total de sinistres};$$

- l'excess (ou sur-crête) :

$$\text{coût total des sinistres} - \text{coût total si les sinistres étaient capés au seuil};$$

- l'excess moyen (sur-crête moyenne) :

$$\text{excess} / \text{nombre total de sinistres}.$$

Pour déterminer graphiquement le seuil adéquate, des graphiques du nombre de sinistres et du pourcentage de coût sous le seuil sont créés. Un bon choix de seuil correspond au coût pour lequel les courbes des percentiles de nombre de sinistres et de coût des sinistres commencent à montrer un changement structurel, comme des tendances qui s'aplatissent. Le seuil d'écrêtement permet ainsi de discriminer les sinistres peu nombreux mais avec un coût important qui peuvent fausser l'analyse.

La présence de sinistres atypiques est identifiée par une loi de distribution à queue lourde modélisable par une loi de Pareto généralisée. Le seuil d'écrêtement peut alors être déterminé à l'aide de la théorie des valeurs extrêmes.

➢ La théorie des valeurs extrêmes

Selon le théorème de **Pickands-Balkema-De Haan**, la queue d'une distribution, indépendamment de la tendance de la distribution sous-jacente, peut toujours être modélisée par une loi de Pareto généralisée (GPD) caractérisée par la fonction de distribution suivante :

$$G_{(\xi,\mu,\sigma)}(x) = \begin{cases} 1 - \left(1 + \xi \dfrac{x-\mu}{\sigma}\right)^{-\frac{1}{\xi}} \; si\; \xi \neq 0. \\ 1 - exp\left(-\dfrac{x-\mu}{\sigma}\right) \; si\; \xi = 0. \end{cases}$$

Avec $x \geq \mu$ quand $\xi \geq 0$, $\mu \leq x \leq \mu - \sigma/\xi$ quand $\xi < 0$, et où

- $\xi \in (-\infty, +\infty)$ est le paramètre **de forme,**
- $\sigma \in (0, +\infty)$ est le paramètre **d'échelle**,
- $\mu \in (-\infty, +\infty)$ est le paramètre **de localisation**.

Le cas le plus courant où $\xi > 0$, correspond aux lois à queues épaisses (Pareto de type I) dont la distribution tend très lentement vers 0 comme une loi puissance. Si $\xi = 0$, la distribution est exponentielle et tend très rapidement vers 0. Et enfin, le cas où $\xi < 0$, concerne les lois de Pareto de type II à support borné.

Trois méthodes visuelles exploitant les propriétés d'une GPD sont utilisées pour définir le seuil adéquat :

- le **paramètre de forme de la GPD** : si la queue de distribution peut être modélisée par une GPD, alors l'estimation du paramètre de forme doit être stable ;

- la fonction **moyenne des excès** qui pour une GPD est une fonction linéaire en fonction du seuil u :

$$e(u) = \mathbb{E}[X - u | X > u] = \frac{\sigma + \xi u}{1 - \xi} ;$$

- le test de **Kolmogorov-Smirnov (KS) :** test statistique non paramétrique permettant de comparer les données avec une distribution connue et de comprendre si la distribution est identique à une GPD.

Une fois le seuil d'écrêtement défini pour chacune des garanties, la charge d'un sinistre se décompose alors en deux parties :

- la charge **attritionnelle**, charge écrêtée par sinistre qui indique le montant de la charge en dessous du seuil : charge écrêtée = min (charge, seuil) ;
- et la **sur-crête** équivaut au montant de la charge en dessous du seuil : max (0, charge – seuil).

La modélisation de la prime pure nécessite de séparer la charge des sinistres attritionnelle de la charge des sinistres atypiques, et de leur appliquer des modèles adéquates. La section suivante présente la méthode classique de modélisation : le GLM.

2.2. Le modèle linéaire généralisé

Le modèle linéaire généralisé (GLM) généralise le modèle de régression en introduisant la non-linéarité au travers d'une fonction de lien, sans s'appuyer sur une hypothèse de normalité. Le GLM, plus flexible qu'un modèle de régression traditionnel permet de modéliser des phénomènes dont la distribution suit une loi de famille exponentielle. L'utilisation des GLMs en tarification est décrite dans l'ouvrage de OHLSSON E., JOHANSSON B. (2010).

2.2.1. Du modèle linéaire gaussien au GLM

La régression est utilisée pour modéliser la relation entre les variables :

- une variable réponse aléatoire $\mathbf{Y}$ $(Y_{1,\dots,}Y_n)$, variable à expliquer : le nombre de sinistres pour modéliser la fréquence et la charge pour modéliser le coût moyen ;
- les prédicteurs $(\mathbf{X_1}, \mathbf{X_2}, \dots \mathbf{X_p})$, variables explicatives, non aléatoires et mesurables sans erreur : la surface, le contenu, le capital, l'activité, etc.

Les modèles linéaires généralisés sont une extension des modèles linéaires classiques. Si les $Y_{1,\dots,}Y_n$sont des variables aléatoires normales indépendantes de moyenne $\boldsymbol{\mu}$, le modèle linéaire gaussien s'écrit comme la somme de deux composantes :

Une composante **déterministe** :

$$E[Y_i] = \mu_i = \beta_0 + \sum_{j=1}^{p-1} x_{ij}\beta_j.$$

Où x_{ij} est la valeur pour l'individu i de la variable explicative j, et les β_j $(\beta_0, \beta_1, \beta_2, \dots \beta_{p-1})$ sont les paramètres du modèle ;

Une composant **aléatoire** : la variance des erreurs, admise constante et indépendante.

Le modèle linéaire repose sur l'hypothèse forte que le terme d'erreur suit une loi normale et de même variance. Dans la pratique, en tarification IARD, un tel modèle ne peut pas convenir, pour deux raisons principales.

La première est que la distribution de la variable à expliquer n'est pas compatible avec une variable aléatoire symétrique normalement distribuée avec une variance fixe. Considérer la variable à expliquer, et son erreur, comme suivant une loi normale suppose une distribution symétrique autour de la moyenne, alors que les erreurs peuvent être fortement asymétrique. Par exemple, la distribution du coût des sinistres ne présente pas de queue à gauche mais une queue droite significativement épaisse, et est donc asymétrique, plutôt comparable à

une densité Gamma. Par ailleurs, les erreurs peuvent ne pas avoir de valeurs négatives ou décimales, comme dans le cas du comptage. Par exemple, pour la distribution du nombre de sinistres, seules des valeurs entières et positives ou nulles doivent être mesurées, alors que le modèle linéaire simple peut malgré tout prédire des valeurs décimales ou négatives. De plus, les données présentent généralement une variance supérieure à la moyenne.

L'autre raison est que, dans le modèle linéaire, les variables prédictives ont un effet linéaire sur la variable à expliquer qui se traduit par les coefficients de régression, or ces effets ne sont pas linéaires en réalité. Par exemple, le nombre de sinistres ne change pas linéairement avec la surface du local professionnel à assurer.

Ces problèmes peuvent être résolus si la variable à expliquer n'est pas normale mais appartient à une **famille exponentielle**, c'est-à-dire en utilisant les modèles linéaires généralisés. L'idée est d'utiliser une transformation mathématique sur la variable à expliquer Y grâce à une fonction appelée « **fonction de lien** », notée **g**. La fonction de lien utilisée tient compte de la véritable distribution des erreurs. Les déviations aléatoires obéissent à une distribution autre que normale, par exemple une loi de Poisson pour des données de comptage.

La normalité et la constance de la variance ne sont plus nécessaires dans les modèles linéaires généralisés, toutefois, l'hypothèse d'indépendance reste une caractéristique essentielle. Une seconde hypothèse sur la structure de l'erreur est l'existence d'un terme d'erreur unique dans le modèle.

2.2.2. Les trois composantes du GLM

Trois composantes caractérisent les modèles linéaires généralisés.

La composante aléatoire

La composante **aléatoire** est définie par la distribution de probabilité de la variable réponse **Y**. La distribution des variables aléatoires $Y_1, \dots, Y_n$ appartient à une famille exponentielle, et sa fonction de densité peut s'écrire sous la forme :

$$f_y(y_i, \theta_i, \phi) = \exp\left(\frac{y_i\,\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi)\right).$$

Les fonctions spécifiques a(·), b(·) et c(·) sont telles que : a(.) est une fonction non nulle, dérivable sur $\mathbb{R}$, b(.) est une fonction trois fois dérivable sur $\mathbb{R}$ et sa dérivée première est inversible et c(.) est une fonction définie sur $\mathbb{R}^2$. Avec $a_i(\phi) = \phi/\omega_i$ et $c = c(y_i, \phi/\omega_i)$ avec ω_i un poids connu pour chaque observation i.

Une distribution de famille exponentielle est une famille a deux paramètres (**θ** et **ϕ**), qui a deux propriétés.

La distribution est complètement spécifiée en fonction de sa moyenne et variance. Le paramètre **θ** est lié à la moyenne, et le paramètre **ϕ**, appelé paramètre de dispersion, est lié à la variance et ne dépend pas de i.

L'espérance de Y s'écrit : $\mu_i = E[Y_i] = b'(\theta_i)$.

La variance de Y s'écrit : $Var[Y_i] = b''(\theta_i) \times a(\phi)$.

La variance de la réponse **Y** est fonction de sa moyenne : la variance change lorsque la moyenne change (hétéroscédasticité). La fonction de variance **V** fournit un lien entre la valeur prédite et la variance des valeurs observées (y_i) :

$$Var[Y_i] = \phi . V(\mu_i)/\omega_i,$$

où V() est la fonction variance, ϕ le paramètre de dispersion et ω_i le poids de chaque observation.

La composante déterministe

La composante **déterministe** est définie comme une combinaison linéaire des variables explicatives observées. Cette composante systématique du modèle attribue à chaque observation i un prédicteur linéaire :

$$\eta_i = \beta_0 + \sum_{j=1}^{p-1} x_{ij}\beta_j.$$

Les paramètres β_j sont estimés par la méthode du « maximum de vraisemblance » (et non par la méthode des moindres carrés, comme dans le modèle linéaire classique).

La fonction de lien

Une troisième composante permet de connecter les deux premières. La **fonction de lien**, fonction différentiable et monotone notée g, fait le lien entre l'espérance μ_i de la composante aléatoire Y_i et le prédicteur linéaire :

$$\eta_i = g(\mu_i) = \beta_0 + \sum_{j=1}^{p-1} x_{ij}\beta_j.$$

La valeur de η est obtenue en transformant μ par la fonction de lien,

$$g(\mathbb{E}[Y_i]) = g(\mu_i).$$

La fonction de lien qui associe la moyenne μ_i au paramètre naturel est appelée fonction de lien canonique. Dans ce cas,

$$g(\mu_i) = \theta_i = \beta_0 + \sum_{j=1}^{p-1} x_{ij}\beta_j.$$

La fonction de lien retenue est la fonction logarithmique définie par :

$$g]0,1] \to \mathbb{R}$$
$$g(x) = ln(x)$$

Cette fonction de lien permet d'avoir des modèles multiplicatifs :

$$\ln(\mathbb{E}[Y_i]) = \beta_0 + \sum_{j=1}^{p-1} x_{ij}\beta_j = 1p - 1x_{ij}\beta j \iff \mathbb{E}[Y_i] = exp\left(\beta_0 + \sum_{j=1}^{p-1} x_{ij}\beta_j\right) = exp(\beta_0) \times exp(x_{i1}\beta_1) \times exp(x_{i2}\beta_2) \times \cdots .$$

Le modèle multiplicatif assure que les coefficients calculés soient tous positifs, ce qui écarte la possibilité d'avoir une prime pure négative. De plus, il permet de voir facilement l'effet de chaque modalité d'un critère de tarification sur la prime de référence : pour tout changement de x_{ij}, $\mathbb{E}[Y_i]$ augmente de $\left(exp(\beta_j) - 1\right)$ % avec un niveau de base tel que $\hat{\mu} = \prod exp(\beta_x)$.

Le tableau ci-dessous récapitule les paramètres utilisés dans les modèles de fréquence, coût moyen et propension :

Modèle	Loi	Nom du lien	Fonction de lien (g)	Réponse	Poids
Fréquence	Poisson	Log	$g(\mu) = \ln(\mu)$	Nombre de sinistres	Exposition
Coût moyen	Gamma	Log*	$g(\mu) = \ln(\mu)$	Charge (coût) des sinistres	Nombre de sinistres
Propension	Binomiale	Logit	$g(\mu) = \ln(\frac{\mu}{1-\mu})$	Nombre de sinistres atypiques	Nombre de sinistres

* Le lien canonique est la fonction réciproque : $g(\mu) = -\frac{1}{\mu}$. Le log est utilisé pour avoir une structure multiplicative comme pour la fréquence.

Tableau 2.1 – Paramètres du GLM

2.2.3. La sélection des variables explicatives

➢ La méthode mRMR

La méthode « mRMR » (*minimum Redundancy and Maximum Relevance* : redondance minimale et pertinence maximale) est utilisée pour classer les variables selon leur importance et mesurer l'association entre variables. Cette approche repose sur la pertinence et sur la redondance des variables.

L'**importance** des variables est mesurée par la **F-statistique**. En considérant que chaque observation a un poids spécifique, par exemple l'exposition pour la fréquence, la formule pour chaque variable s'écrit :

$$F(var) = \frac{\sum_i w_{i.}(\bar{Y}_{i.} - \bar{Y})^2/(K-1)}{\sum_i \sum_j w_{ij} * (\bar{Y}_{ij} - \bar{Y})^2/(\sum_j w_{i.} - K)}.$$

Avec :

$\bar{Y}$: la fréquence ou le coût moyen global
$\bar{Y}_{i.}$: la fréquence ou le coût moyen de la modalité i de la variable
$\bar{Y}_{ij}$: la fréquence ou le coût moyen de l'observation j de la modalité i de la variable
$w_{i.}$: le poids de la modalité i de la variable
w_{ij} : le poids de l'observation j de la modalité i de la variable
K : le nombre de modalités de la variable

La **redondance** entre deux variables var_1 et var_2 est évaluée grâce au **V de Cramer** :

$$V(var_1, var_2) = \sqrt{\frac{\chi^2(var1,var2)/w_{..}}{\min(K-1,R-1)}}.$$

$$\text{Où } \chi^2(var_1, var_2) = \sum_{ij} \frac{\left(w_{ij} - \frac{w_{i.} * w_{j.}}{w_{..}}\right)^2}{\frac{w_{i.} * w_{j.}}{w_{..}}}.$$

Avec :

w_{ij} : le poids de l'observation de la modalité i de la variable 1 et de la modalité j de la variable 2
$w_{i.}$: le poids de l'observation de la modalité i de la variable 1
$w_{j.}$: le poids de l'observation de la modalité j de la variable 2
$w_{..}$: le poids total
K : le nombre de modalités de la variable 1
R : le nombre de modalités de la variable 2

La formule du score mRMR s'écrit comme la combinaison de la F-statistique et du V de Cramer :

$$mRMR_{score}(var_i) = F(var_i) * \left(1 - \frac{\sum_s V(var_i, var_s)}{|S|}\right),$$

où S est l'ensemble des variables sélectionnées.

➢ La méthode ascendante : *forward stepwise*

Les variables identifiées comme étant explicatives sont introduites dans le modèle par la méthode ascendante : ***forward stepwise.*** Les variables rangées par ordre d'importance sont ajoutées une à une dans le modèle. La validation de la sélection des variables se fait par étape.

Un modèle de bonne qualité décrit correctement les valeurs observées. La qualité de l'ajustement d'un modèle GLM peut être mesurée par la **déviance** D. Elle consiste à comparer le modèle étudié au modèle saturé (modèle possédant autant de paramètres que d'observations et estimant donc exactement les données : la moyenne de la variable est définie par l'observation elle-même) en mesurant l'écart en termes de log-vraisemblance $\mathcal{L}$ entre les deux modèles :

$$D = 2(\ln(\mathcal{L}_{saturé}) - \ln(\mathcal{L}_{estimé}))$$

L'objectif est de minimiser D. Une déviance positive et faible est signe de bonne qualité d'un modèle. Cette statistique suit asymptotiquement une loi du Chi-2 à n – p - 1 degrés de liberté.

Cependant, ajouter une variable au modèle aboutit forcément à un gain d'information, et implique une baisse de la déviance. Un autre critère de significativité est utilisé : l'**AICc** (Critère d'Information d'Akaike corrigé). Ce dernier est une correction de l'AIC pour les échantillons de petite taille. L'AIC est une mesure statistique de comparaison de modèles imbriqués qui permet de pénaliser un modèle en fonction du nombre de paramètres. L'AIC se calcule selon la formule suivante :

$$\mathrm{AIC} = 2\mathrm{k} - 2\ln(\mathcal{L}).$$

Avec k le nombre de paramètres du modèle et $\mathcal{L}$ le maximum de la log-vraisemblance. L'AIC pénalise la déviance du modèle avec 2 fois le nombre de paramètres et permet de favoriser les modèles les plus parcimonieux, avec peu de paramètres. L'AICc intègre une pénalité supplémentaire pour les paramètres additionnels. Pour un échantillon de taille n, l'AICc s'écrit :

$$\mathrm{AICc} = \mathrm{AIC} + \frac{2k(k+1)}{n-k-1}.$$

L'utilisation de l'AICc plutôt que l'AIC évite de sur-ajuster le modèle en sélectionnant un trop grand nombre de paramètres. Ainsi, chaque nouvelle variable est ajoutée uniquement si l'AICc diminue par rapport au modèle précédent (i.e. sans la variable). Chaque variable sélectionnée doit conduire à une optimisation de l'AICc, une minimisation, le meilleur modèle étant celui avec l'AICc le plus faible.

La statistique du $\boldsymbol{\chi^2}$ permet également de valider l'ajout d'une nouvelle variable X en comparant les modèles avant et après l'intégration de la variable. Sous l'hypothèse nulle H_0 : la variable X n'est pas influente dans le modèle, la statistique $\mathrm{S} = -2\ln\left(\frac{\text{vraisemblance du modèle sans X}}{\text{vraisemblance du modèle avec X}}\right)$ suit asymptotiquement une loi du χ^2 à m degrés de liberté, avec m le nombre de modalités de X . Si P[S $\leq$ seuil à 95 % d'une χ^2à m degrès de libertés] > 5 %, alors les deux vraisemblances sont proches et l'apport de la variable X dans le modèle n'est pas significatif.

➢ Le test de significativité des variables

La significativité de chaque variable introduite et son impact éventuel sur la significativité des autres variables déjà incluses sont vérifiés par le **test de Wald**. Le test de Wald permet de tester la significativité des variables du modèle. Pour une variable donnée, l'hypothèse nulle $\mathrm{H}_0: \beta_i = 0$ est testée, la statistique du test suit asymptotiquement une loi du χ^2. L'erreur de première espèce est contrôlée et seules les variables dont la p-value est inférieure à 5% sont retenues.

2.2.4. L'optimisation du modèle

➢ L'identification des interactions

L'interaction se produit lorsque l'effet d'une variable varie selon le niveau d'une autre variable. En d'autres termes, l'interaction existe lorsque les effets d'une variable dépendent de la valeur d'autres variables. L'inclusion d'une interaction dans le modèle montre l'existence d'une interdépendance entre deux variables. L'interaction explique l'effet conjoint des deux variables A et B. Cela améliore l'estimation, car un autre effet est ajouté aux effets simples : $estimation = A + B + A * B$.

L'interaction concerne l'effet des facteurs sur le risque et n'est pas liée à la corrélation de l'exposition entre deux variables.

➢ La simplification des variables

La simplification permet de réduire le nombre de paramètres dans le modèle, tout en conservant le message essentiel du modèle.

- Le regroupement de modalités

Si les variables retenues sont toutes significatives, ce n'est pas forcément le cas pour chacune des modalités. Les estimateurs par modalités sont analysés pour savoir si une modalité est significative ou non.

Le logiciel Emblem affiche en sortie du GLM les relativités de chaque modalité des variables retenues dans le modèle. La fiabilité statistique de chaque paramètre estimé est testée au travers de **l'intervalle de confiance**, c'est-à-dire de la plage déterminée en ajoutant et en retranchant 2 erreurs standard de la relativité. Si zéro appartient à l'intervalle de confiance, cela signifie que la relativité n'est pas statistiquement différente du niveau de base.

L'**erreur standard** exprimée en pourcentage, permet d'évaluer le degré de relation entre deux modalités d'une même variable. Une valeur élevée de cette statistique pour deux modalités indique peu de différence entre elles. Ces modalités doivent être regroupées si elles sont contigües. Le regroupement se fait entre modalités avec des relativités proches.

De même, une modalité avec une exposition trop faible (inférieure à 3%) sera regroupée avec une autre modalité dont l'estimateur est proche, en tenant compte de l'ordre si la variable est ordonnée.

- L'ajustement des variables continues

Le logiciel Emblem n'accepte que des variables catégorielles en entrée, c'est pourquoi les variables continuent ont été discrétisées. Mais les variables catégorielles initialement continues peuvent être assimilées à des variables continues en ajustant une **courbe polynomiale** ou une **spline**. Une courbe spline est une série de fonctions avec chaque fonction définie sur un intervalle délimité par deux points spécifiés appelés nœuds : ajustement de polynômes lisse continus par morceaux entre des nœuds. Le nombre de nœud ne doit pas être trop grand pour ne pas conduire à un modèle sur-paramétré.

L'utilisation d'une courbe pour simplifier une variable ne peut se faire que si la variable démontre une tendance qui peut être approximée par une courbe.

➢ La stabilité dans le temps

Pour qu'un modèle soit pertinent, sa capacité prédictive ne doit pas varier d'une année à l'autre. La stabilité des tendances dans le temps est évaluée en appliquant une interaction entre chaque variable retenue dans le modèle et la variable année. Les variables dont les estimateurs ne sont pas stables dans le temps ne peuvent pas être utilisées.

2.2.5. L'implémentation dans Emblem

L'un des principaux avantages du logiciel Emblem est de pouvoir visualiser grâce à une interface graphique, le comportement de la variable réponse et des coefficients estimés en fonction de chaque variable sélectionnée.

Les variables sont ajoutées manuellement et une par une. A chaque ajout de variable, les caractéristiques suivantes sont analysées.

La **qualité de l'ajustement** : pour chaque variable ajoutée, la moyenne des observations pour chaque modalité doit être suffisamment proche de la courbe moyenne des prédictions.

Les **intervalles de confiance** : pour chaque coefficient estimé, les intervalles de confiance sont construits par une approximation de la loi Normale. Des intervalles de confiance des coefficients des modalités d'une variable qui ne se chevauchent pas est signe que la variable a un effet significatif. Plus les bornes supérieures et inférieures sont proches, plus l'estimation est stable. Par ailleurs, le test de Wald permet de vérifier que les coefficients sont significativement non nuls.

La **déviance** et **l'AICc** : la déviance est un indicateur qui repose sur la log-vraisemblance et qui permet d'identifier la manière dont le modèle considéré dévie du modèle saturé. Une variable qui apporte de

l'information permet de diminuer la déviance du modèle. L'AICc qui s'écrit en fonction de la déviance introduit une pénalisation sur la complexité du modèle.

Le logiciel Emblem permet également de regrouper facilement les modalités rares d'une variable pour éviter les effets de bord, et de lisser les coefficients des variables quantitatives pour introduire un lien non linéaire entre la prédiction et la variable. Les données manquantes sont considérées comme une modalité à part entière. Si la volumétrie des données manquantes le permet, elles sont rapprochées de la variable dont le coefficient est le plus proche, sinon leur coefficient est mis au mis au niveau de base.

Comme l'ajout des variables se fait manuellement, les corrélations éventuelles sont détectées pour chaque variable ajoutée : Emblem signale que la matrice de design X n'est plus identifiable. En cas de corrélation, seule la variable qui améliore au mieux la qualité de la prédiction est gardée.

2.2.6. La validation du modèle

L'adéquation du modèle : l'analyse des résidus

L'analyse des **résidus** permet de s'assurer de la cohérence du modèle choisi et de juger la pertinence de son ajustement, en vérifiant les hypothèses sur le terme d'erreur $\varepsilon = (\varepsilon_1, \dots \varepsilon_n)$ et la présence de points aberrants qui auraient une influence négative sur l'estimation des paramètres du modèle.

Les hypothèses sur le terme d'erreur sont :

- $E(\varepsilon) = 0$, en moyenne le modèle est bien spécifié ;
- $E(\varepsilon^2) = \sigma^2$, la variance de l'erreur est constante (homoscédasticité) ;
- $E(\varepsilon_i, \varepsilon_j) = 0$, les erreurs sont non-corrélés ;
- $Cov(\varepsilon, X) = 0$, l'erreur est indépendante de la variable explicative ;
- ε suit une loi normale $\mathcal{N}(0, \sigma^2)$.

Sous ces hypothèses, les résidus sont définis comme la distance entre l'observation réelle et la valeur prédite par le modèle :

$$\varepsilon_i = y_i - E(y_i) = y_i - \hat{y}_i.$$

Les résidus de Pearson et de déviance sont souvent utilisés pour diagnostiquer les modèles linéaires généralisés (GLM).

Les **résidus de Pearson** sont définis comme les distances standardisées entre les réponses observées et attendues :

$$r_{p_i} = \frac{y_i - \hat{y}_i}{\sqrt{Var(\hat{y}_i)}}.$$

En pratique, dans le cas d'un modèle de fréquence, les résidus « **crunched »** sont analysés. En effet, ces résidus calculés sur des groupes et non sur des données individuelles sont adaptés aux modèles de fréquence pour lesquels il y a deux nuages de points (selon la présence ou non d'un sinistre). Les résidus « crunched » sont définis par :

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

Les **résidus de déviance** sont définis comme la racine carrée signée (plus ou moins) des contributions individuelles à la déviance du modèle (c'est-à-dire la différence entre les log-vraisemblances des modèles saturés et ajustés). Ces résidus sont équivalents aux résidus de Pearson en termes d'interprétation, et sont définis par :

$$r_{D_i} = signe(y_i - \hat{y}_i) \sqrt{\frac{d_i}{(1 - h_{ii})}}.$$

Où d_i représente la contribution de l'observation i à la déviance D telle que $D = \sum d_i$, et h_{ii} est le $i^{éme}$terme de la diagonale de la matrice $H = W^{\frac{1}{2}} (X'WX)^{-1} W^{\frac{1}{2}}$ avec W la matrice diagonale dont le $i^{éme}$terme est $w_i = \frac{1}{Var(\hat{y}_i)(g'(\hat{y}_i))^2}$.

La somme des carrés des résidus est dans les deux cas, asymptotiquement, un Chi-2 à n – p – 1 degrés de liberté.

Les résidus sont analysés graphiquement en regardant la forme du nuage de points en fonction des valeurs prédites :

- les résidus doivent être centrés autour de zéro pour valider l'hypothèse de linéarité sur laquelle se base le modèle. Si ce n'est pas le cas, les résidus par rapport à chacune des variables explicatives doivent être analysés pour trouver la ou les variables responsables de cette non-linéarité ;
- les résidus ne doivent pas présenter une allure particulière pour vérifier l'hypothèse d'homoscédasticité ;
- un point isolé sur l'axe des ordonnées est signe d'un fort résidu (valeur extrême) qu'il convient de retirer pour ne pas biaiser les coefficients du modèle.

La performance du modèle : la courbe de gains

La courbe de gains est une aide visuelle utilisée pour évaluer le pouvoir prédictif de différents modèles où l'exposition cumulée, classée par valeurs prédites (la plus élevée en premier), est tracée en fonction de la réponse cumulée. Pour les modèles de fréquence par exemple, la part cumulée du nombre de sinistres observé est en ordonnée et la part cumulée d'exposition triées dans l'ordre croissant de nombre de sinistres prédits se trouve en abscisse.

Les modèles peuvent être évalués, par rapport au modèle moyen, par la vitesse à laquelle le nombre réel de réponses s'accumule. Dans un modèle prédictif, les contrats avec les fréquences observées les plus élevées doivent contribuer à la gauche du graphique, car les observations sont classées par valeur prédite. En d'autres termes, si le modèle est prédictif, les valeurs prédites élevées correspondent aux valeurs observées élevées. La logique est la même pour un modèle coût moyen en adaptant les données.

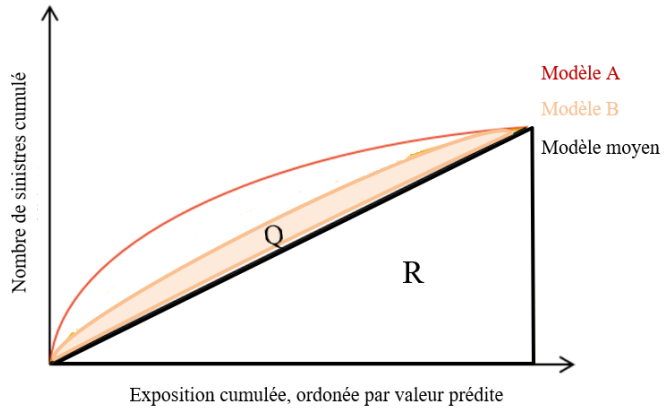


Figure 2-1 - Courbe de gains

Dans le graphique de la figure ci-dessus, en comparant le modèle A et le modèle B, le modèle A semble plus prédictif que le modèle B car la courbe du premier est toujours au-dessus de celle du second

Pour mesurer le pouvoir prédictif d'un modèle à partir de la courbe de gain, le coefficient de Gini est utilisé. Le coefficient de Gini est une mesure de l'aire sous la courbe (AUC). Plusieurs définitions du coefficient de Gini existent. Celle retenue est : $Gini = (2 \times AUC) - 1$ avec $AUC = Q + R$ dans le cas du modèle B de la

figure ci-dessus. La ligne droite sur le graphique correspond au modèle aléatoire (moyen) et a un coefficient de Gini de 0. Le coefficient de Gini peut prendre des valeurs comprises entre -1 et 1, plus le coefficient de Gini est positif et proche de 1, plus le modèle est prédictif.

L'analyse des corrélations

L'analyse des corrélations permet de s'assurer de la stabilité d'un modèle. La matrice de corrélations montre la corrélation entre toutes les paires de paramètres du modèle. Pour deux paramètres X_1, X_2 d'écarts types σ_1/σ_2, elle est liée à la matrice de variance / covariance par la formule ci-dessous :

$$\rho = \frac{\mathrm{Cov}(X_1, X_2)}{\sigma_1/\sigma_2}.$$

Le coefficient de corrélation est compris entre -1 et +1, et s'interprète de la manière suivante :

- 1 : parfaitement corrélé positivement
- > 0, < 1 : corrélé positivement
- 0 : non corrélé
- > -1, < 0 : corrélé négativement
- -1 : parfaitement corrélé négativement.

Les GLM permettent de prédire la sinistralité par profils de risque en s'appuyant sur un historique de données. Le pouvoir prédictif et la performance des modèles sont quantifiables, toutefois, la qualité de l'information qui alimente les modèles est primordiale. Les données géo démographiques qui apportent une source d'information sur le risque géographique en font partie. Ces données sont synthétisables au sein d'un zonier. La section suivante s'intéresse au traitement des données géographiques au travers de la création d'un zonier.

2.3. La théorie du zonier

En multirisque professionnelle, de nombreux risques relèvent de facteurs géographiques. Par exemple, les indicateurs géographiques comme la distance aux pompiers ou la densité peuvent avoir une incidence sur les incendies. D'où l'utilité de définir des zones géographiques par garantie.

D'un point de vue technique, une solution triviale pour définir les zones géographiques serait d'introduire les codes postaux comme variable dans les modèles GLM, mais plusieurs problèmes se posent :

- les GLMs dans Emblem ne peuvent pas gérer des facteurs avec plus de 255 niveaux distincts, or le nombre de codes postaux est supérieur à 6 000 ;
- les GLMs ne gèrent pas de façon automatique la formation de groupes de risques suffisamment homogènes ;
- les estimations pour les communes avec peu d'exposition ne sont pas fiables. Un minimum d'exposition est nécessaire pour que les effets soient statistiquement significatifs ;
- les GLMs ne tiennent généralement pas compte des corrélations spatiales.

La solution est d'introduire dans les GLMs les indicateurs géographiques à la place des codes postaux : modèle **pré-zonier**, puis de les remplacer par une variable unique regroupant l'ensemble des informations géographiques : modèle **post-zonier**.

2.3.1. La méthode

La méthodologie appliquée dans ce mémoire fait référence aux pratiques mises en place par le Groupe Allianz.

En considérant un modèle GLM dont le code postal est omis au profit des variables géographiques, la sinistralité observée (fréquence et coût moyen) est expliquée par deux types de variables : des données « standard », propres au risque et au client (surface, capitaux, activité, âge de l'entreprise …), et des données

géographiques (densité de population, revenu moyen, taux de criminalité …). Cependant, la sinistralité observée n'est jamais complètement captée par le modèle, le résidu : écart entre l'observé et l'estimé représente la partie qui reste inexpliquée.

Le résidu du modèle se décompose en deux parties :

- la première est une composante d'erreur systématique due aux informations géographiques omises, c'est-à-dire aux informations perdues lors du remplacement du code postal par les variables géographiques. Dans la suite, cette partie est intitulée « **effet géographique inconnu** » ;
- la seconde est le véritable **bruit aléatoire** associé à la nature statistique des données. Par définition, seul l'échantillonnage peut le réduire.

Le modèle avant zonier peut se schématiser de la façon suivante :

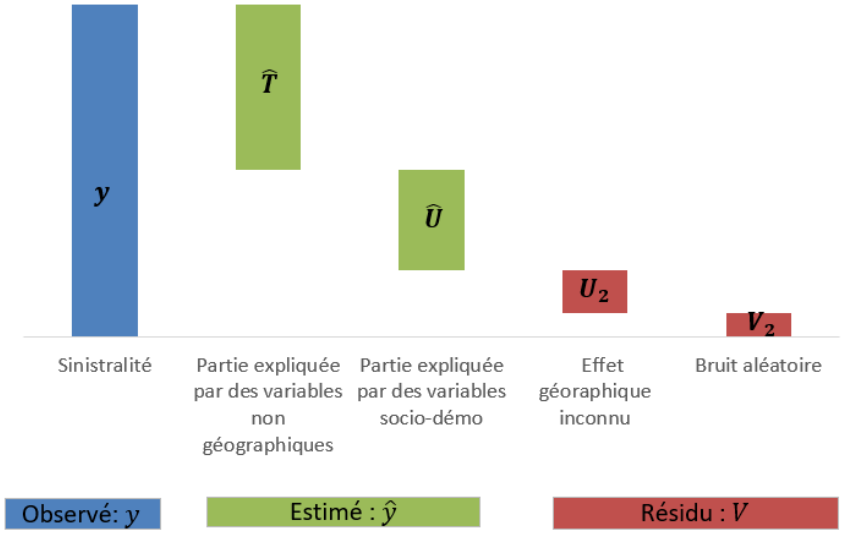


Figure 2-2 – Schéma du modèle avant zonier

Les statistiques prises en compte tant pour la fréquence que le coût moyen sont :

- l'observé : $y = \mu * T * U * V$;
- l'estimé : $\hat{y} = \hat{\mu} * \widehat{T} * \widehat{U}$;
- le résidu : $R = \frac{\text{Observé}}{\text{Estimé}} = \frac{y}{\hat{y}_{\text{avant zonier}}} = \frac{\mu * T * U * V}{\hat{\mu} * \widehat{T} * \widehat{U}} = V \simeq U_2 * V_2$.

Remplacer le code postal par un ensemble de variables corrélées, mais simples, génère une perte d'informations et une erreur systématique dans le modèle. L'idée de base du zonier est de pouvoir identifier dans le résidu du modèle avant zonier, ce qui provient de l'effet géographique inconnu : U_2, du véritable bruit aléatoire : V_2.

L'effet géographique inconnu U_2 est isolé par lissage des résidus du GLM pré-zonier. Un algorithme de lissage est appliqué aux résidus en considérant une corrélation entre communes voisines. L'effet géographique global est obtenu en combinant les résidus lissés à l'effet géographique connu.

L'effet géographique combiné est transformé en une variable discrète : « zone technique », par une méthode de clustering utilisant la méthode de Ward. Cette nouvelle variable, zone technique, combinaison des deux effets territoriaux connus et inconnus, peut ainsi être intégrée dans le model GLM en substitution des variables géographiques explicatives du modèle.

Les effets géographiques systématiques sont identifiés à partir des résidus du modèle avant zonier en appliquant une technique de correction spatiale. Le processus de correction spatiale comprend trois phases : la standardisation, le lissage et la validation.

2.3.2. La construction d'un zonier

➢ La standardisation

Les statistiques utilisées : observé, estimé et résidu, peuvent être différentes par codes postaux en raison du mix de variables « standard » (i.e. non géographiques) par contrat. Pour supprimer les effets non géographiques, une standardisation des statistiques est nécessaire avant d'analyser l'effet géographique inconnu. La standardisation est le processus de construction d'une réponse dans laquelle l'effet relatif aux variables non géographiques est retraité de sorte que les effets restants soient tous liés à la zone géographique.

Pour calculer l'effet résiduel de chaque code postal, la standardisation des données observées est effectuée au niveau de l'observation individuelle. Les valeurs estimées correspondantes sont également standardisées en forçant les variables explicatives non géographiques du modèle avant zonier au niveau de base. Les réponses résultantes sont agrégées au niveau du code postal en tenant compte des différents poids (w) de l'ensemble d'échantillons.

Soit Q une quantité à standardiser et k un code postal, l'agrégation des données standardisées par code postal peut se faire de deux manières :

- Ratio des valeurs moyennes : réponse standardisée pour l'exposition

$$Q^{std} = \frac{\langle Q \rangle_k}{\langle \hat{T} \rangle_k} = \frac{\sum_{\{i \in k\}} Q_i \, w_i}{\sum_{\{i \in k\}} w_i} \Big/ \frac{\sum_{\{i \in k\}} \hat{T}_i \, w_i}{\sum_{\{i \in k\}} w_i} = \frac{\sum_{\{i \in k\}} Q_i \, w_i}{\sum_{\{i \in k\}} \hat{T}_i \, w_i}.$$

- Moyenne du ratio : réponse standardisée

$$Q^{std} = \left\langle \frac{Q}{\hat{T}} \right\rangle_k = \frac{\sum_{\{i \in k\}} \frac{Q_i}{\hat{T}_i} w_i}{\sum_{\{i \in k\}} w_i}.$$

Les statistiques standardisées sont pour la fréquence et le coût moyen :

- L'estimé standardisé : $\mathrm{std}(\hat{y}) = \hat{\mu} * \hat{U}$.
- L'observé standardisé : $\mathrm{std}(y) = \frac{y}{T} = \mu * U * V$.

➢ Le lissage

Le lissage spatial est une méthode d'apport de connaissances sur les codes postaux environnants pour améliorer les estimations. Cette technique permet d'ajuster la valeur d'un code postal en tenant compte de la réponse des voisins, dans le but d'éliminer le bruit aléatoire.

La technique de lissage spatial utilisée est basée sur la proximité des codes postaux, où chaque code est supposé se comporter de la même manière que ses voisins. Cette technique repose sur l'hypothèse qu'une zone peu ou pas sinistrée présente un risque similaire à celui des zones voisines ou proches. Cette hypothèse n'est pas toujours vérifiée dans la pratique : les zones contigües peuvent parfois être de nature assez différente, et les rivières et les voies ferrées peuvent séparer des zones qui présentent des risques sous-jacents différents bien qu'elles soient proches.

Le lissage par proximité est un modèle Bayésien qui peut s'écrire sous la forme :

$$P(\theta/Y) \propto P(Y/\theta) \, . \, P(\theta),$$

$\theta = u_k$: paramètre inconnu correspondant à l'effet géographique inconnu.

La réponse agrégée par codes postaux, supposée suivre une distribution Poisson pour la fréquence et Gamma pour le coût moyen, peut s'écrire :

$$y_k \sim f(w_k \mathrm{Offset}_k e^{u_k}, \phi).$$

$$\Downarrow$$

$$P(Y/\theta) \rightarrow P(y_k / u_k).$$

Avec,

- y_k : la réponse (nombre de sinistres ou coût des sinistres) agrégée par code postal k,
- f : Gamma ou Poisson,
- w_k: le poids (exposition ou nombre de sinistres) agrégée par code postal k,
- $Offset_k$: la valeur ajustée issu du modèle avant zonier,
- ϕ : le paramètre d'échelle de la distribution.

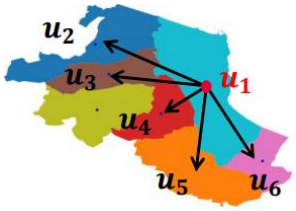
La représentation a priori de la corrélation spatiale suit une distribution gaussienne avec une moyenne qui représente la moyenne des u_k sur les codes postaux adjacents, et un écart type fixe $\boldsymbol{\sigma}$ qui joue le rôle de **paramètre de lissage** :

$$\mathrm{u_k} \sim \mathcal{N}(\overline{\mathrm{u_k}}, \sigma).$$

$$\Downarrow$$

$$\mathrm{P}(\theta) \longrightarrow \mathrm{P}(\mathrm{u_k}).$$

La moyenne des voisins assure le lissage :



$$\overline{\mathrm{u_1}} = \frac{1}{5} \sum_{\mathrm{k=2}}^{6} \mathrm{u_k}.$$

Figure 2-3 – Lissage par moyenne des voisins

Représentation de σ :

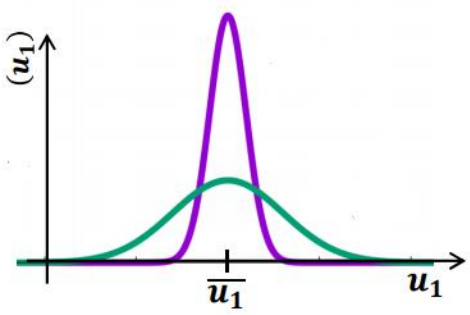


Figure 2-4 – Représentation de l'écart type : paramètre de lissage

Un $\boldsymbol{\sigma}$ petit indique une faible probabilité que l'effet géographique dévie de sa moyenne, donc le lissage est efficace.

Un tableau des voisins répertorie les codes adjacents les uns aux autres.

➢ La validation du lissage

L'hyperparamètre pour un niveau de lissage optimal est choisi en appliquant une méthode d'apprentissage/validation. L'hyperparamètre de lissage est $1/\sigma$, et la métrique utilisée est la moyenne des écarts au carré (MSE : Mean Squared Errors). Le MSE représente la moyenne sur l'ensemble des codes postaux

k, de la différence au carré entre les statistiques observées et lissées pour un hyperparamètre de lissage S_i, pondérée par le poids w_k, et s'écrit sous la forme :

$$MSE_i = \frac{\sum_k (Obs_k - Smoothed(S_i)_k)^2 w_k}{\sum_k w_k}.$$

La base de données de modélisation utilisée pour le modèle Emblem avant zonier (échantillon qui représente 80% de la base de données globale) est divisée en deux échantillons : apprentissage (60%) et validation (20%). Pour plusieurs niveaux de lissage S_i, le lissage est ajusté sur l'échantillon d'apprentissage puis est testé sur la base de validation sur laquelle le MSE est calculé. Le meilleur hyperparamètres de lissage est celui minimisant le MSE.

Pour connaitre l'effet géographique global, le produit doit être fait entre l'effet inconnu lissé et l'effet connu issu des modèles GLM avant zonier. L'effet global s'obtient facilement en multipliant la quantité lissée par l'estimé standardisé.

2.3.3. La classification des codes postaux

Grâce au lissage, une estimation de l'effet géographique global est disponible pour chaque code postal. Mais une variable avec autant de valeurs distinctes que de codes postaux ne peut pas directement être utilisée dans Emblem.

La méthode de Ward est utilisée pour regrouper dans un même cluster les codes postaux avec un effet géographique similaire. Elle consiste à réunir des clusters dont le regroupement impacte le moins possible l'inertie interclasse. La distance entre deux classes est celle de leurs barycentres au carré, pondérée par les effectifs des deux clusters. Cette méthode de clustering a été retenue car elle tend à regrouper de petites classes entre elles. Chaque cluster représente une zone de la nouvelle variable « zone technique » qui est ensuite intégrée dans le GLM.

2.3.4. La correction spatiale

Avant d'utiliser le zonier dans la modélisation, l'effet attendu du lissage sélectionné doit être validé et testé sur l'échantillon de validation. Pour ce faire, le résidu lissé est appliqué au modèle ajusté.

Le pouvoir prédictif des modèles avant zonier a été validé, avec notamment l'adéquation de la moyenne globale avec la moyenne observée. L'application des résidus lissés à ces modèles, va changer la moyenne globale ajustée. La première étape consiste donc à calculer le biais introduit en appliquant le résidu lissé, afin de pouvoir procéder à un ajustement.

Tout d'abord, l'ajustement à appliquer aux valeurs ajustées est calculé comme la moyenne pondérée du rapport de l'observé sur l'ajusté x le résidu lissé, de sorte que la moyenne du portefeuille correspond toujours au résultat moyen observé.

Ensuite, le facteur de correction spatiale (SCF : « Spatial Correction Factor ») est calculé pour chaque code postal en appliquant l'ajustement aux résidus lissés. Ce facteur est ensuite appliqué aux résultats estimés et aux résidus correspondants, donnant les résultats modélisés spatialement ajustés requis pour les tests de validation. Les valeurs ajustées spatialement corrigées, pondérées par l'exposition pour la fréquence sont regroupées, sur la base d'un regroupement de Ward. Le nombre de groupes retenu pour le regroupement est choisi en fonction de la taille de l'échantillon afin que chaque groupe soit représenté par une exposition suffisante.

2.3.5. Le modèle post-zonier

Après intégration de la variable « zone technique » en remplacement des variables géographiques, les estimations du GLM post-zonier se schématisent sous la forme :

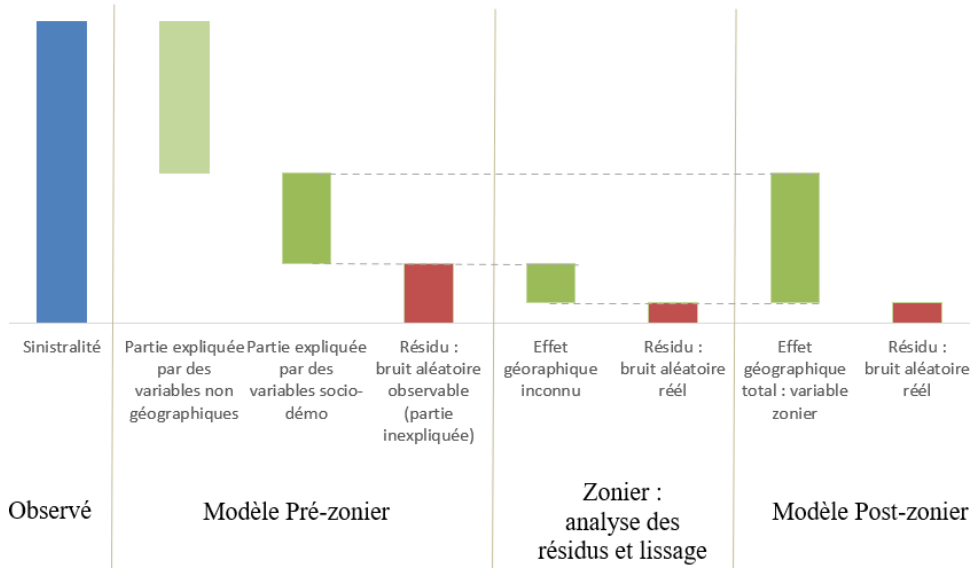


Figure 2-5 – Passage des modèles pré-zonier aux modèles post-zonier

Les avantages de la variable « zone technique » dans le GLM sont :

- une réduction de la partie inexpliquée par le modèle : le résidu, et de ce fait une amélioration de la performance du GLM ;
- une réduction du nombre de variables explicatives : toutes les variables géographiques sont résumés dans la variable « zone technique » ;
- les zones sont regroupées en tenant compte de la corrélation spatiale existant dans les données.

Cette méthode présente comme inconvénient une perte d'information contenue dans les codes postaux, et nécessite la prise en compte de méthodes de corrélation spatiale au-delà des GLMs mais a l'avantage de réintroduire dans le modèle, avec une grande probabilité, la notion de proximité spatiale.

Le zonier constitue un regroupement du risque géographique en zones homogènes. De même que l'information contenue dans les codes postaux ne peut être utilisée telle quelle dans les modèles, les activités doivent être regroupées en groupes de risques homogènes : classification, avant d'être introduites dans la modélisation. Cette classification des activités est réalisée à l'aide de méthodes de *Machine Learning*. Par ailleurs, les modèles de régression classiques de type GLM pour modéliser les sinistres atypiques peuvent être moins performants que des méthodes plus modernes de type *Machine Learning*. La section suivante présente les méthodes de classification et de régression de *Machine Learning* utilisées dans le mémoire.

2.4. Les méthodes de *Machine Learning*

2.4.1. Les différents types d'apprentissage

➢ Apprentissage « non supervisé » ou « supervisé »

Les méthodes de *Machine Learning* se regroupent en deux grandes familles : l'apprentissage non supervisé et supervisé.

En apprentissage **non supervisé,** l'algorithme cherche à identifier les similarités et distinctions au sein de données qui ne sont pas annotées. Il reçoit en entrée des observations brutes de variables aléatoires x_1, x_2, x_3, x_4, ... et cherche à obtenir la relation avec des variables latentes structurelles : $x_i \rightarrow y_i$. Cette méthode vise à regrouper les variables ayant des caractéristiques communes, sans chercher à expliquer une variable réponse. Elle est utilisée pour mieux comprendre un jeu de données ou pour identifier des comportements similaires. L'analyse en composante principale (ACP) et le clustering en sont des exemples.

En apprentissage **supervisé**, le modèle est entrainé sur des données annotées de leurs sorties : un label, une classe cible ou une variable réponse leur est déjà associé. L'algorithme doit être capable, une fois entraîné, de prédire la cible sur de nouvelles données non annotées. Il reçoit des données annotées $(x_1, y_1), (x_2, y_2), (x_3, y_3)$, ... dans l'objectif de prédire la sortie sur de nouvelles observations : $x^* \rightarrow y^*$. L'apprentissage supervisé est utilisé pour tout problème dont les observations peuvent être annotées de la cible souhaitée en sortie. Les modèles s'appuyant sur des arbres (Random Forest, Boosting, …) sont des méthodes d'apprentissage supervisé.

➢ Régression ou classification

Une distinction pour choisir l'algorithme de Machine Learning adapté au besoin est le type de données attendu en sortie : valeur **continue** (un nombre) ou **discrète** (une catégorie). Le premier cas est associé à un modèle de **régression**, et le second à une **classification**. La différence entre classification et régression linéaire est illustrée ci-dessous :

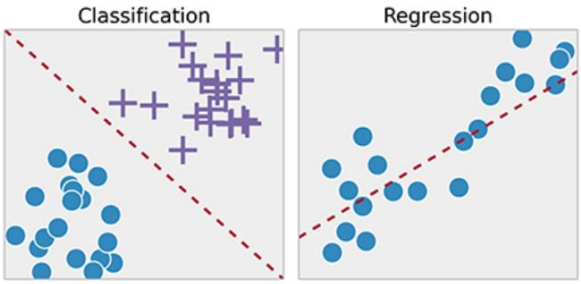


Figure 2-6 – Différence entre classification et régression linéaire (source : https://openclassrooms.com)

2.4.2. Deux méthodes de classification non-supervisée : ACP et CAH

L'analyse en composantes principales (ACP) a pour objectif d'expliquer les similarités entre n individus par leur proximité en fonction de k variables quantitatives les caractérisant, tout en identifiant les corrélations entre variables. Cependant, cette méthode ne permet de classifier les individus de façon fiable. Pour classifier des individus : identifier des groupes d'individus au sein des plans factoriels sur plusieurs dimensions, l'utilisation de méthodes spécifiques de classification comme la classification ascendante hiérarchique (CAH) est nécessaire.

La classification utilisée pour établir des **typologies** est complémentaire de l'analyse en composante principale qui met en évidence la **structure** des données. L'ACP permet de visualiser la structure des corrélations et d'obtenir de nouvelles variables non corrélées présentant une information plus grande qu'une simple variable. Cependant, les visualisations ne sont réalisables que sur des plans, ce qui limite la perception précise des phénomènes.

➢ L'analyse en composante principale : ACP

L'ACP analyse un tableau de données **X** comportant en lignes les valeurs des **p** caractéristiques d'un même individu et en colonne les **n** valeurs prises par les individus pour une même variable :

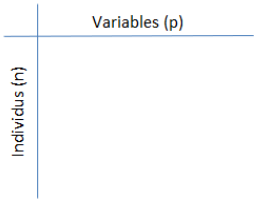


Figure 2-7 – Schéma des données en entrée de l'ACP

L'objectif est d'explorer le nuage des **n** individus décrits par **p** variables quantitatives et de transformer des variables très corrélées en nouvelles variables décorrélées les unes des autres. Le principal objectif est d'identifier les variables les plus importantes et de les représenter de manière simple. L'information contenue dans la base de données est résumée en un certain nombre de variables synthétiques appelées : **composantes principales**. Ensuite, ces données sont projetées sur l'hyperplan le plus proche afin d'avoir une représentation simple des données.

La représentation géométrique du nuage de points est donnée par la coordonnée des points de chaque variable sur les p axes $(\mathbf{e_1}, \mathbf{e_2}, \cdots, \mathbf{e_p})$ du repère canonique de $\mathbb{R}^p$. Afin de visualiser un nuage dans un espace de dimension p, ce nuage est projeté dans des plans de faible dimension. La réduction de dimension s'accompagne d'une perte d'informations. Tout l'enjeu de l'analyse en composantes principales est de réduire la dimension des données tout en conservant un maximum d'information. Le but étant de déterminer les sous-espaces de projection qui permettent de distinguer les ressemblances et les oppositions entre les individus, et qui **maximise la variance** des individus projetés. Les individus sont dotés de poids $\mathrm{p_i}$ qui vérifient $\sum_{i=1}^{n} p_i$ pour éventuellement donner plus d'importance à certains individus dans l'analyse.

La matrice de variance-covariance du nuage correspond à sa matrice d'inertie. Ainsi, le sous-espace optimal est celui qui maximise l'inertie expliquée par ce sous-espace, et minimise l'inertie résiduelle autour de ce sous-espace. En notant **D** la matrice diagonale des poids associés aux individus, la matrice d'inertie du nuage **V** s'écrit sous la forme :

$$\mathrm{V} = \mathrm{X}'\mathrm{D}\,\mathrm{X}.$$

Le problème consiste à trouver les axes $(\mathrm{u}_1, \mathrm{u}_2, \cdots, u_\mathrm{p})$ d'un nouveau repère de manière à maximiser l'inertie expliquée par la droite de direction u_1, le plan $(\mathrm{u}_1, \mathrm{u}_2)$ et tous les sous-espaces de dimension k $(\mathrm{u}_1, \mathrm{u}_2, \cdots, u_\mathrm{k})$.

Comme l'inertie expliquée par un axe de direction u vaut $\boldsymbol{u}'\,\mathbf{V}\,\mathbf{u}$ et que V est symétrique défini-positive, le repère optimal est le repère composé des vecteurs propres $(\mathrm{u}_1, \mathrm{u}_2, \cdots, u_\mathrm{p})$ de V associés au valeurs propres $(\lambda_1, \lambda_2, \cdots, \lambda_\mathrm{p})$ rangées par ordre décroissant. Les nouvelles variables sont ensuite obtenues en projetant les points du nuage sur les nouveaux axes générés par les vecteurs propres de V. Ces nouvelles variables sont appelées **composantes principales** et ont l'avantage d'être hiérarchisées et non corrélées deux à deux. Les valeurs propres représentent l'inertie du nuage expliquée par les axes, et les variances des composantes principales.

Les résultats sont représentés par deux types de graphiques : le cercle des corrélations des variables et la carte factorielle des individus. La représentation des variables diffère de celle des observations : les observations sont représentées par leurs projections, mais les variables sont représentées par leurs corrélations.

Le **cercle des corrélations des variables** : le lien entre les composantes principales et les variables d'origine se fait facilement, puisque les composantes principales sont des combinaisons linéaires des variables de départ. Ce lien s'interprète graphiquement à l'aide de la projection du nuage des variables sur le plan des composantes principales qui permet de visualiser les corrélations entre les variables de départ et les composantes principales. Plus les variables sont proches du cercle, plus elles sont bien représentées et plus l'angle entre deux variables est faible, plus elles sont corrélées. La mesure du cosinus de l'angle formé entre deux variables est égale au coefficient de corrélation linéaire entre deux variables. Le cercle de corrélation des variables montre les relations entre **toutes** les variables :

- les variables positivement corrélées sont regroupées ;
- les variables négativement corrélées sont positionnées sur les côtés opposés de l'origine du graphique (quadrants opposés) ;
- la distance entre les variables et l'origine mesure la qualité de représentation des variables. Les variables qui sont loin de l'origine sont bien représentées par l'ACP.

La **carte factorielle des individus :** l'information intéressante pour les individus est la distance entre les points. La carte des individus permet de représenter les individus dans le plan factoriel. Plus une coordonnée est proche de 0, moins l'axe correspondant est significatif c'est à dire que l'individu participe de moins en moins à la structure mise en évidence par l'axe.

Deux méthodes sont utilisées pour sélectionner les axes principaux nécessaires pour obtenir le meilleur système de représentation. Ces méthodes s'appuient sur le calcul des valeurs propres et de leur inertie qui permet de déterminer la proportion d'information contenue dans un plan.

Le **« scree plot »** proposé par Cattell en 1966 et 1977 repose sur l'analyse de la courbe de décroissance des valeurs propres. L'idée est de détecter les changements de structure identifiées par des cassures appelées « coudes ». En 1966, Cattell conseil de sélectionner uniquement les axes avant le coude, puis en 1977, il préconise de les intégrer. En réalité, tout dépend de la valeur du coude : si elle est faible l'axe peut être négligé, en revanche si elle est élevée, l'axe doit être sélectionné.

L'inertie expliquée par les axes correspond à la courbe des valeurs propres cumulées, en pourcentage. Un coude y est également visible mais la partie postérieure au coude est horizontale car elle doit indiquer un apport négligeable des axes restants

Les composantes principales issues de l'ACP deviennent les nouvelles variables descriptives des individus sur lesquelles s'appuie la CAH pour former des groupes d'individus homogènes.

➢ La classification ascendante hiérarchique : CAH

• L'algorithme de la CAH

La classification ascendante hiérarchique (CAH) est une méthode de classification qui permet de :

- **organiser** l'information ;
- regrouper des individus qui se **ressemblent** en fonction de plusieurs critères et opposer ceux qui se **différencient** ;
- former une **partition** des données étudiées ;
- associer à chaque classe un type **généralisant** tous les éléments de la classe.

La classification est une méthode de **discrétisation**. Elle divise l'ensemble des individus en sous-ensembles exhaustifs et disjoints : chaque classe constituée est non-vide, et un individu appartient à une classe et à une seule (l'intersection de deux classes est vide).

La classification ascendante hiérarchique (CAH) a pour but de constituer des groupes (classes) d'individus homogènes en leur sein et différenciés les uns des autres, au regard de variables les décrivant. Cette technique est dite ascendante car au départ, chaque individu est pris individuellement, puis à chaque étape de regroupement, la cause de regroupement de certains individus et leur distinction par rapport aux autres est analysée. Le regroupement des individus analysés est progressif, selon leur degré de ressemblance jusqu'à l'obtention d'une unique classe les regroupant tous.

La CAH procède en plusieurs étapes pour regrouper les individus. Chaque étape correspond à un niveau :

- Première étape (premier niveau) : les **n** individus sont caractérisés par leurs valeurs sur **p** variables. Tous les individus sont progressivement regroupés deux à deux, pour former une classe, en fonction de leur ressemblance.
- Deuxième étape (deuxième niveau) : les deux classes qui se ressemblent le plus sont regroupées.
- Troisième étape et suivantes (niveaux suivants) : répétition des regroupements de classes jusqu'à obtenir une classe unique regroupant l'ensemble des individus.

Deux critères de classifications sont à définir en amont de la CAH : le critère de **ressemblance** et le critère d'**agrégation**. Ces deux critères sont utiles pour formaliser à quel point deux individus sont proches les uns des autres, à quel point une observation est proche d'un cluster et à quel point deux classes sont proches les uns des autres.

Le critère de **ressemblance**, défini par une distance, mesure la proximité ou l'éloignement des individus, afin de regrouper ensemble les individus les plus proches. S'agissant de variables quantitatives, la **distance euclidienne** est retenue comme critère de ressemblance.

Après le calcul de la distance entre les individus, le critère **d'agrégation** détermine la façon dont les classes vont se former : comment les individus vont s'agréger puis comment ils vont s'agréger à des classes

déjà formées. Plusieurs critères d'agrégation sont possibles : plus proche voisin, diamètre maximum, distance moyenne, Ward. Le critère retenu est le critère de **Ward**, le plus souvent utilisé en classification. Le critère de Ward, également appelé critère de variance (inertie) recherche une partition qui minimise la variance interne des classes et maximise la variance entre les classes. Ainsi, chaque classe est formée d'individus se ressemblant le plus possible entre eux et se démarquant le plus des autres. Ce critère est basé sur les calculs de la variance.

En notant :

- $I(N)$ l'inertie totale du nuage de points : mesure de la dispersion des individus autour du centre de gravité G du nuage ;
- $I(C)$ l'inertie de la classe C: dispersion des individus de la classe C autour de son centre de gravité G_c ;
- $I(Q)$ l'inertie d'une partition Q : dispersion des centres de gravité G_c des classes formant cette partition autour du centre de gravité G du nuage ;
- m_i : la masse relative de l'individu i.

La formule de la variance totale s'écrit :

$$I(N) = \sum_{i=1}^{n} m_i \, d^2(i, G).$$

$$I(N) = I(Q) + \sum I(C).$$

La variance totale est égale à l'inertie inter classes plus la somme des inerties intra-classe. La décomposition de l'inertie du nuage de points permet de mesurer le rôle que joue la différence entre un individu et une classe déjà constituée dans l'inertie totale du nuage de points. Les classes d'une partition sont d'autant plus homogènes et différentes les unes des autres que l'inertie intra-classe est faible par rapport à l'inertie interclasse. Ce critère d'agrégation consiste à regrouper les deux classes qui augmentent le moins l'inertie intra-classe (qui diminuent le moins l'inertie interclasse). La distance entre deux classes est mesurée par la différence entre l'inertie interclasse de la partition avant et après le regroupement de ces deux classes.

- Le nombre de classes optimal

Bien que le résultat d'une classification non-supervisée soit assez subjectif, la performance de l'algorithme dépend du nombre de classes. Le nombre optimal de classes est analysé au travers de trois mesures.

Le **coefficient de silhouette** renseigne sur la **forme des groupes** (homogénéité et séparation). Pour un individu donné x, le coefficient de silhouette $s(x)$ évalue que cet individu appartient au bon groupe : assez poche des individus de son groupe et assez éloigné des autres individus. La distance moyenne de x à tous les autres individus de son groupe C_k est calculée : $a(x) = \frac{1}{|C_k|-1} \sum_{u \in C_k, u \neq x} d(u, x)$, ainsi que la plus petite valeur que pourrait prendre $a(x)$, si x était assigné à un autre cluster : $b(x) = \min_{l \neq k} \frac{1}{C_l} \sum_{u \in C_l} d(u, x)$. Si x est rattaché au bon groupe, alors $a(x) < b(x)$. Le coefficient de silhouette s'écrit : $s(x) = \frac{b(x)-a(x)}{max(a(x),b(x))}$. Le coefficient de silhouette pour un échantillon est donné par la moyenne du coefficient de silhouette sur l'échantillon. Le coefficient est compris entre -1 pour une classification incorrecte et +1 pour une classification dense et des classes bien séparées. Un coefficient autour de zéro indique des classes qui se chevauchent.

L'indice de **Calinski-Harabasz** également connu sous le nom de critère de ratio de variance, est également utilisé pour évaluer le nombre de classes. Il est défini comme le rapport de la moyenne de dispersion entre les classes et de la dispersion intra-classes, où la dispersion correspond à la somme des distances au carré.

Pour un échantillon E de taille n avec k regroupements, soient (B_k) la trace de la matrice de dispersion inter classes et (W_k) la trace de la matrice de dispersion intra-classes définies par :

$$W_k = \sum_{q=1}^{k} \sum_{x \in C_q} (x - c_q)(x - c_q)^T,$$

$$B_k = \sum_{q=1}^{k} n_q \, (x - c_q)(x - c_q)^T.$$

Avec C_q les points de la classe q, c_q le centre de la classe q, c_E le centre de E, et n_q le nombre de points dans la classe q.

Le score de Calinski-Harabasz en fonction de k regroupements s'écrit :

$$s = \frac{tr(B_k)}{tr(W_k)} \times \frac{n-k}{k-1}.$$

Plus ce score est élevé, mieux les classes sont définies (plus elles sont denses et bien séparées).

L'indice de **Davies-Bouldin** est un autre indicateur qui s'intéresse à la « similitude » moyenne entre les classes, où la similitude est une mesure qui compare la distance entre les classes avec la taille des classes elles-mêmes.

Pour $i = 1, \ldots, k$ classes, en posant s_i la distance moyenne entre chaque point de la classe i et le centroïde de ce cluster, et d_{ij} la distance entre le centroïde de la classe i et celui de la classe j, alors la similitude entre la classe i et la classe j est définie par R_{ij} telle que :

$$R_{ij} = \frac{s_i + s_j}{d_{ij}}.$$

R_{ij} ne peut être négative et l'utilisation de la distance centroïde limite la métrique de distance à l'espace euclidien. L'indice de Davies-Bouldin pour k classes, s'écrit alors :

$$DB = \frac{1}{k}\sum_{i=1}^{k} \max_{i \neq j} R_{ij}.$$

Plus l'indice de Davies-Bouldin est proche de 0, meilleure est la partition.

- Les mesures de performance de la CAH

Un des critères à prendre en compte pour valider une classification est la **stabilité** : si le nombre de groupes retenu correspond à la structure naturelle des données, la classification est plus stable. La stabilité peut être évaluée en comparant la classification avec celle lancée sur un échantillon (sous-ensemble différent de données). La performance s'intéresse à ce que les mêmes individus appartiennent au même groupe, que ce groupe soit le premier, le deuxième ou le k-ième. L'indice de **Rand** est la mesure utilisée pour apprécier la concordance de deux partitions. Cet indice correspond à la proportion de paires d'éléments qui sont conjointement groupées ou conjointement séparées. Soit deux partitions P_1 et P_2 de E, avec :

- a, le nombre de paires d'éléments de E groupés dans P_1et dans P_2 ;
- b, le nombre de paires d'éléments de E groupés dans P_1mais séparés dans P_2 ;
- c, le nombre de paires d'éléments de E groupés dans P_2 mais séparés dans P_1 ;
- d, le nombre de paires d'éléments de E séparés dans P_1 et dans P_2.

La somme $a + d$ représente la consistance entre les deux partitions, tandis que la somme $b + c$ représente le désaccord entre les deux partitions. L'indice de Rand (RI) s'écrit :

$$RI(P_1, P_2) = \frac{a + d}{a + b + c + d}.$$

Dans le cas où un grand nombre de groupes sont prédits, l'indice de Rand peut être artificiellement gonflé. En effet, les paires d'individus appartenant à des groupes différents seront nombreuses, et deux individus étiquetés différemment auront de grandes chances d'être dans deux clusters différents. L'indice de Rand ajusté (ARI : Adjusted Rand Index) corrige cet effet en normalisant l'indice de Rand(RI). Soit $\mathbb{E}(RI)$, l'espérance de la valeur de l'indice de Rand (partitionnement au hasard), alors l'indice ajusté d'écrit :

$$ARI = \frac{RI - \mathbb{E}(RI)}{max(RI) - \mathbb{E}(RI)}.$$

Cet index ajusté, compris entre 0 et 1, est proche de 0 pour un clustering aléatoire et égal à 1 pour un clustering correspondant exactement à la partition initiale.

2.4.3. Des méthodes d'apprentissage supervisée

➢ L'arbre de décision

L'**arbre de décision**, appelé Arbre de Régression et de Classification (**CART**), est une application de méthode d'apprentissage supervisé utilisée pour la régression ou la classification. Capable d'atteindre une grande précision tout en étant interprétable, la méthode CART a été introduite par Leo Breiman en 1984. Le but principal est d'expliquer une variable Y à partir de variables continues ou discrètes $X_1, X_2, \ldots, X_p$. La variable Y peut prendre des valeurs **quantitatives** : arbre de **classification**, ou **numériques** : arbre de **régression**. Dans ce mémoire, des arbres de régression sont utilisés pour les modèles de propension et de coût moyen des sinistres atypiques.

- La construction de l'arbre

Un arbre binaire est une construction hiérarchique dont le sommet est appelé **racine**. Les traits qui partent en descendant de la racine sont des **branches** qui joignent des **nœuds.**

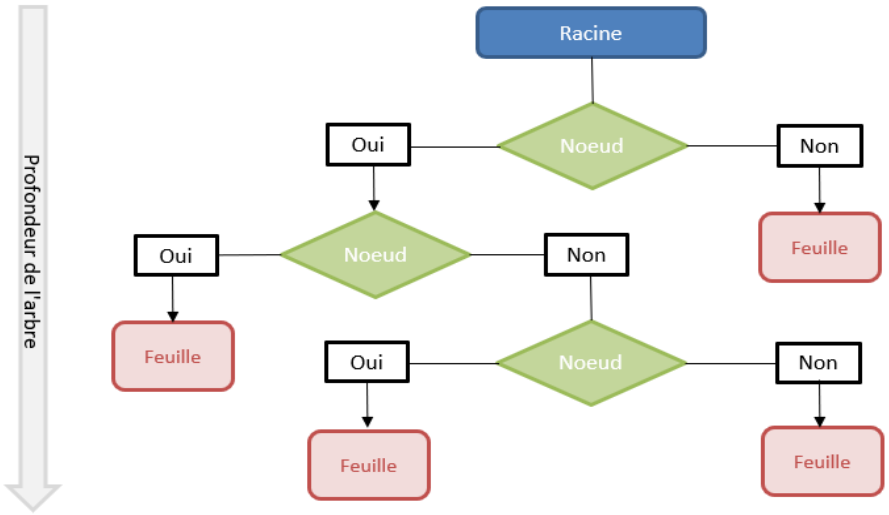


Figure 2-8 – Illustration schématique d'un arbre CART

Un nœud représente l'intersection d'un ensemble de règles : ensemble des données de départ sur lesquelles les **règles de fractionnement** sont appliquées. Ces règles visent à segmenter la population en différents sous-groupes disjoints en fonction des variables explicatives. Un nœud se caractérise par :

- des règles d'arrêt : la profondeur de l'arbre, le nombre minimum d'individus pour envisager un fractionnement, etc. Ces règles sont spécifiées au niveau des **hyperparamètres** du modèle ;
- des conditions de fractionnement qui reposent sur des critères mathématiques. Chaque règle de fractionnement met en jeu une seule variable parmi $X_1, X_2, \ldots, X_p$, mais une même variable peut être utilisée pour définir plusieurs conditions de fractionnement. L'ordre des variables sélectionnées indique leur importance.

Les **feuilles** sont les nœuds en bas de l'arbre qui ne sont plus divisibles. La réponse moyenne dans la feuille représente l'estimation de la variable réponse. Un arbre se lit de la racine (nœud principal) vers les feuilles (nœuds terminaux).

Cas d'un arbre de classification

Pour la classification, la performance consiste à comparer les classes réelles aux classes prédites et permettent d'interpréter les probabilités prédites des classes.

La performance d'un modèle de classification se visualise sous forme de **matrice de confusion** qui représente sous forme de tableau les prédictions du modèle par rapport aux vrais labels. Elle donne une vue d'ensemble des prédictions justes et des prédictions fausses : chaque ligne de la matrice de confusion représente les instances de la classe réelle et chaque colonne représente les instances de la classe prédite.

Dans le cas de données déséquilibrées avec une clase fortement majoritaire, la fonction de perte tente d'optimiser des métriques telles que le taux de bonnes prédictions en ne tenant pas compte de la distribution des données. Ce taux est affecté par le déséquilibre des classes, en effet, il est élevé même si une grande partie des points de la classe minoritaire sont mal classifiés. Les classes minoritaires sont traitées comme des valeurs aberrantes de la classe majoritaire et l'algorithme d'apprentissage génère simplement un classifieur trivial qui classe chaque exemple dans la classe majoritaire. Le modèle peut sembler performant alors qu'en réalité, il reflète la surreprésentation de la classe majoritaire.

L'objectif est de pouvoir identifier la classe minoritaire. Pour cela, des métriques beaucoup moins influencées par la classe majoritaire, issues de la matrice de confusion sont utilisées : la précision, le rappel (« recall »), et le F_β-score.

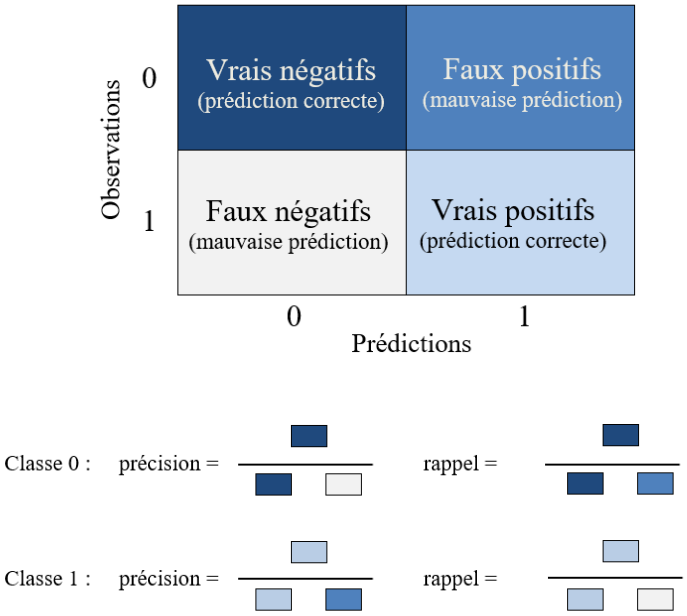


Figure 2-9 – Matrice de confusion

La **précision** ou valeur prédictive positive est le rapport entre les vrais positifs et tous les positifs. :

$$\text{précision} = \frac{\#\text{ vrais positifs}}{\#\text{ vrais positifs} + \#\text{ faux positifs}}$$

Une précision élevée signifie que la majorité des prédictions positives du modèle sont des positifs bien prédits. Autrement dit, plus la précision est forte, moins le modèle se trompe sur les positifs.

Le **rappel** ou sensibilité correspond au pourcentage de positifs bien prédits par le modèle :

$$\text{rappel} = \frac{\#\text{ vrais positifs}}{\#\text{ vrais positifs} + \#\text{ faux négatifs}}$$

Un rappel élevé signifie que le modèle ne rate aucun positif. Néanmoins cela ne donne aucune information sur sa qualité de prédiction des négatifs. Autrement dit, plus le rappel est fort, plus le modèle repère les positifs.

Pour chacune des classes, la précision et le rappel s'interprètent de la manière suivante :

- précision et rappel élevé : la classe est bien gérée par le modèle ;

- précision élevée et rappel faible : la classe n'est pas bien détectée mais lorsqu'elle l'est, le modèle est très fiable ;
- précision faible et rappel élevé : la classe est bien détectée, mais inclut également des observations de l'autre classe ;
- précision et rappel faible : la classe n'est pas bien gérée.

La précision et le rappel pris séparément ne permettent pas d'évaluer un modèle déséquilibré. Si le modèle ne prédit jamais les positifs, la précision est élevée, au contraire, si le modèle prédit tout le temps les positifs, le rappel est élevé.

Le **F_β-score** combine précision et rappel en calculant la moyenne harmonique pondérée de la précision et du rappel :

$$F_\beta - \text{score} = \frac{(1+\beta)^2 \times \text{précision} \times \text{rappel}}{\beta^2 \times \text{précision} + \text{rappel}}.$$

Le F_β-score compris entre 0 et 1 est un bon indicateur de performance d'un modèle déséquilibré. Le paramètre β détermine le poids du rappel dans le score combiné :

- β >1, plus d'importance est apportée au rappel (autrement aux faux négatifs) ;
- β <1, plus d'importance est apportée à la précision (autrement dit aux faux positifs) ;
- β =1 : F1-score, autant d'importance est apportée à la précision qu'au rappel .

Cas d'un arbre de régression

Les arbres sont construits en divisant l'échantillon d'origine de façon récursive selon des règles simples, de façon à minimiser une fonction de perte.

Dans le cas d'un arbre de régression, en posant les variables qualitatives parmi $X_1, X_2, \cdots, X_p$, $j \in \{1, \cdots, p\}, c \in \mathbb{R}$ et

- $\bar{y}_{j,c,gauche}$ la moyenne des valeurs de Y pour les individus vérifiant $X_j < c$,
- $\bar{y}_{j,c,droit}$ la moyenne des valeurs de Y pour les individus vérifiant $X_j \geq c$,
- $f_{gauche}(j, c)$ la fonction de perte pour les individus vérifiant $X_j < c$,
- $f_{droit}(j, c)$ la fonction de perte pour les individus vérifiant $X_j \geq c$,

pour un nœud donné, la perte commise en séparant les individus selon $X_j < c$ ou $X_j \geq c$ est donnée par :

$$E(j, c) = f_{gauche}(j, c) + f_{droit}(j, c).$$

La condition de fractionnement consiste à minimiser cette perte. Elle correspond à $X_{j*} \geq c_*$ (ou à $X_{j*} < c_*$) où j_* et c_* rendent minimale la perte $E(j, c)$. La variable X_{j*} est appelée caractère de coupure et c_* valeur seuil.

Si un ou plusieurs variables parmi $X_1, X_2, \cdots, X_p$ sont qualitatives, la condition $X_j \geq c$ devient $X_j = m$, où m désigne une des modalités de la variable. Ainsi, les individus sont divisés en 2 groupes : $X_j = m$ ou $X_j \neq m$.

- Les limites de l'arbre de décision

L'arbre de décision est sensible à l'échantillon initial, ce qui le rend peu flexible et instable. Les variables sont analysées successivement et les nœuds sont élaborés de façon enchaînée, de ce fait, si les données changent avec le temps, l'arbre doit être reconstruit. La construction de l'arbre dépend fortement de l'ordre des variables sélectionnées ce qui peut nuire au pouvoir prédictif du modèle.

Par ailleurs, la dépendance aux données sur laquelle l'arbre est calibré entraîne des risques de surapprentissage. L'arbre s'imprègne de toutes les caractéristiques des données parcourues, et peut retenir un point exceptionnel et le considérer comme normal, générant ainsi un biais. Ce problème se rencontre surtout en présence d'un grand nombre de nœuds ou de feuilles.

Deux sources d'erreurs empêchent les algorithmes d'apprentissage supervisé de généraliser au-delà de l'échantillon d'apprentissage : le **biais** et la **variance**. Le **biais,** erreur entre l'observé et la prédiction peut provenir d'un manque de relations pertinentes entre les données en entrée et les sorties (**sous-apprentissage**), lié à des erreurs dans les hypothèses de l'algorithme d'apprentissage. La **variance** est l'erreur provenant de la sensibilité aux petites fluctuations dans les données d'apprentissage. Elle est liée à la généralisation du modèle sur d'autres données. Un modèle avec une variance trop élevée peut amener à décrire le bruit aléatoire dans les données d'apprentissage au lieu des sorties prévues (**surapprentissage**).

La solution pour corriger ces erreurs est **l'agrégation de modèles** qui permet de remonter les variables les plus déterminantes dans l'explication de la variable cible, et de sélectionner le critère qui est redondant sur l'ensemble des arbres construits en mesurant sa contribution dans la construction de la variable d'intérêt dans chaque arbre. Parmi les méthodes de combinaison d'arbres existantes, le **Boosting** a été retenu car il permet de **réduire le biais** des modèles à forte variance. L'algorithme de Boosting est une méthode ensembliste qui s'appuie sur un apprentissage séquentiel : les arbres sont construits de manière consécutive dans le but de résoudre l'erreur de l'arbre précédent.

➢ Le Gradient Boosting Machine (GBM)

Le processus du **GBM** vise à minimiser l'erreur entre les prédictions et les observations. L'algorithme de ***Gradient boosting*** se construit sur un premier arbre très basique et peu efficace qui correspond à la moyenne des observations. L'écart entre la prédiction (la moyenne, pour le premier arbre) et l'attendu est calculé. La particularité de l'algorithme de *Gradient boosting* est d'essayer de prédire à chaque étape cet écart, appelé **résidu**. Le second arbre est alors entrainé pour prédire le premier résidu.

L'objectif du *Gradient boosting* est de s'écarter du modèle de la moyenne pour se rapprocher de la réalité. Pour cela, les prédictions du second arbre sont multipliées par un coefficient inférieur à 1 appelé learning rate (λ). Ce coefficient permet de réduire la taille des pas qui le rapproche de l'attendu dans le but d'augmenter la précision. La création des arbres suivants suit le même modèle :

- les résidus sont calculés à partir des dernières prédictions,
- le nouvel arbre est entrainé pour prédire ces résidus,
- les prédictions du nouvel arbre sont multipliées par un facteur inférieur à 1.

Le modèle final est simplement la combinaison de tous les arbres de régression individuels.

L'algorithme ***LightGBM*** est une version particulière de l'algorithme de *Gradient boosting* dont la spécificité réside dans la division de l'arbre par les feuilles et non par niveaux. Habituellement, dans les algorithmes de *Gradient boosting*, d'une itération à l'autre, toutes les feuilles de l'arbre sont divisées en un nouveau lot de feuilles pour créer un nouveau niveau ou étage. A l'inverse, le *LightGBM* ne prend pas les feuilles du dernier niveau mais choisit une feuille dans l'arbre pour la diviser. La croissance de l'arbre se fait verticalement : *leaf-wise tree growth* et non horizontalement : *level-wise tree growth*.

Schématiquement, l'algorithme *LightGBM* peut s'illustrer de la manière suivante par rapport aux autres algorithmes de *boosting* :

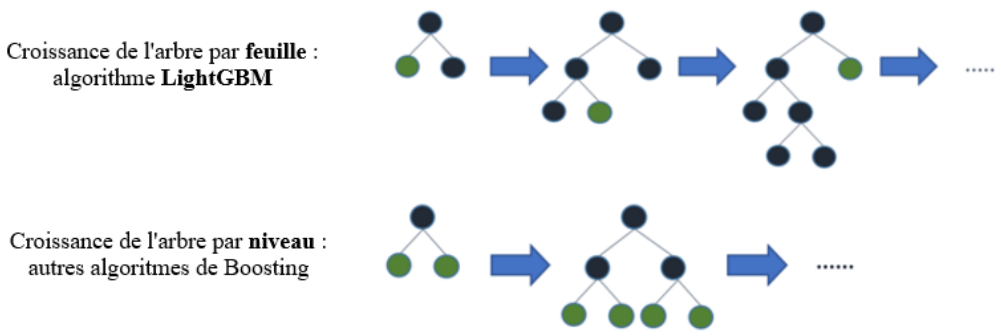


Figure 2-10 – Illustration de l'algorithme *LightGBM* (source : MANDOT P. (2017) *What is LightGBM*)

Le principal avantage de la croissance de l'arbre par les feuilles est sa rapidité d'exécution, en revanche le surapprentissage est fréquent avec ce type de méthode. Une manière d'éviter le surapprentissage est de définir les hyperparamètres optimaux du modèle.

➢ Les hyperparamètres

L'un des avantages du *LightGBM* est son nombre important de paramètres qui permet une configuration optimale.

Pour obtenir de bons résultats avec une méthode de croissance par les feuilles (*LightGBM*), les paramètres les plus importants à ajuster sont :

- **num_leaves** : le nombre de feuilles est le paramètre principal pour contrôler la complexité du modèle. Pour un nombre de feuilles fixées, l'arbre d'un algorithme de développement par feuille est généralement plus profond qu'un développement par niveau. Le nombre de feuilles doit être inférieur à 2^{max_depth}.

- **max_depth** : ce paramètre est utilisé pour limiter la profondeur. La profondeur des arbres permet d'étudier les interactions entre les variables. Toutefois, plus un arbre est profond plus il surapprend les données d'apprentissage.

- **learning rate** : paramètre qui permet d'avoir un apprentissage plus lent et plus fin en contrôlant la vitesse de convergence de l'algorithme. Le choix du taux dépend du nombre d'arbres choisi : plus il est petit, plus le nombre d'itération optimal est grand.

Le temps d'apprentissage d'un *LightGBM* augmente avec le nombre de nœuds ajoutés. Certains paramètres peuvent être utilisés pour **contrôler le nombre de nœuds** par arbre :

- **min_gain_to_split** : lors de l'ajout d'un nouveau nœud, *LightGBM* choisit le point de fractionnement qui a le gain le plus important. Le gain est essentiellement la réduction de perte générée par l'ajout d'un nœud. De trop petites améliorations de la fonction de perte peuvent être onsidérées comme non significatives. Augmenter ce paramètre est utile pour définir un seuil minimal d'amélioration apporté par un nœud, et réduire le temps d'entraînement.

- **min_data_in_leaf** et **min_sum_hessian_in_leaf** : le nombre minimum de données par feuille est un paramètre très important pour éviter le surajustement. Sa valeur optimale dépend du nombre d'échantillons d'apprentissage et du nombre de feuilles. Ce paramètre et la somme minimale de la dérivée seconde de la fonction objectif (Hessian) pour les observations d'une feuille permettent d'éviter que l'algorithme *LightGBM* n'ajoute des nœuds avec très peu d'observations et difficilement généralisables.

D'autres paramètres sont utilisés pour **limiter le nombre d'arbres** :

- **num_iterations** : ce paramètre contrôle le nombre d'arbres. Réduire ce paramètre permet de réduire les temps d'apprentisage. Toutefois, si le paramètre num_iterations est diminué, le paramètre learning_rate doit être augmenté pour améliorer la performance.

- **early_stopping_rounds** : ce paramètre permet d'arrêter le processus si la performance du modèle ne s'améliore pas pendant un certain nombre d'arbres.

Les paramètres ci-dessous permettent d'utiliser moins de **données** :

- **feature_fraction** : ce paramètre permet de ne sélectionner qu'un pourcentage de variables à prendre en compte au début de la construction de chaque arbre. Cela réduit le nombre total de fractionnements à évaluer pour ajouter chaque nœud.

- **bagging_freq** et **bagging_fraction** : ces paramètres permettent d'échantillonner au hasard les données d'apprentissage. Ce processus d'apprentissage sur plusieurs échantillons aléatoires sans

remise est appelé « bagging ». Le paramètre bagging_freq défini sur un entier supérieur à 0 contrôle la fréquence d'itération à laquelle un nouvel échantillon est tiré, et le paramètre bagging_fraction défini sur une valeur > 0,0 et < 1,0 contrôle la taille de l'échantillon (pourcentage des données d'entrainement utilisées).

Mais le tout premier paramètre à définir est la fonction de perte. L'apprentissage du *LightGBM* se base sur l'erreur du modèle définie par la **fonction de perte**. Cette dernière permet la comparaison de différents modèles et le **choix optimal des hyperparamètres** en sélectionnant le modèle qui minimise l'erreur commise dans les prédictions.

La fonction de coût et la fonction de perte

La fonction de coût quantifie l'erreur entre les valeurs prédites et les valeurs attendues. Cette fonction permet de mesurer les performances du modèle d'apprentissage pour un ensemble de données. La fonction de perte a un rôle similaire pour un seul exemple d'apprentissage. La fonction de coût est la moyenne des fonctions de perte de l'ensemble d'apprentissage.

Le choix de la fonction de coût se fait en fonction des données et du problème adressé. La distribution de la variable réponse conditionne la fonction de coût :

- Pour le modèle de propension : **régression logistique**, la fonction de coût retenue est le **log loss** qui permet la pénalisation des mauvaises prédictions.

$$\text{LogLoss} = -\frac{1}{N}\sum_{i=1}^{N} y_i \log\big(p(y_i)\big) + (1 - y_i)\log\big(1 - p(y_i)\big).$$

- Pour le modèle du coût moyen des sinistres atypiques dont la distribution suit une loi Gamma, la fonction de coût est la **déviance Gamma** :

$$\text{GammaDeviance} = 2 \times \left(\frac{\hat{y}}{y} - \frac{\hat{y} - y}{\hat{y}}\right).$$

Une fois la fonction de coût calculée, une fonction de **régularisation** lui est appliquée. La régularisation L2 ou Ridge régression ajoute à la fonction de perte une pénalité équivalente à la somme des coefficients au carré multiplié par une constante : **lambda L2.** L'objectif de la régularisation est de sous-ajuster le modèle afin d'éviter le surapprentissage.

L'optimisation des hyperparamètres

La recherche des hyperparamètres optimaux est indispensable pour avoir un modèle avec une bonne généralisation. Le package **Optuna** est utilisé pour l'ajustement des hyperparamètres. Ce dernier permet d'effectuer un processus d'optimisation visant à trouver le meilleur ensemble d'hyperparamètres. Optuna se base sur une étude (« study ») qui représente l'optimisation, composée d'évaluations uniques d'une fonction « objectif », et sur des essais (« trials »), obtenus en faisant varier les paramètres cibles sur un espace de recherche déterminé.

L'une des principales forces d'Optuna est de proposer différentes stratégies concernant l'exploration de l'espace des hyperparamètres dans le but de rendre le processus d'optimisation plus efficace. Plus précisément, différentes méthodes d'échantillonnage sont définies dans le package Samplers. Dans ce mémoire, la méthode d'échantillonnage par défaut : **TPESampler** est retenue. Cette méthode s'appuie sur l'algorithme « Tree-structured Parzen Estimator » (TPE), basé sur une approche **bayésienne** : à chaque essai, pour chaque paramètre et indépendamment des autres, il prend la valeur la plus performante par rapport à un modèle probabiliste, en gardant trace des valeurs précédemment tirées.

Le processus général d'optimisation d'Optuna suit les étapes suivantes :

- définition d'une fonction « objectif » et spécification, pour chaque paramètre à optimiser, du type et de la plage des valeurs à rechercher ;

- sélection d’une méthode d’échantillonnage et initialisation de l'étude ;
- évaluation de la fonction « objectif » pour chaque essai.

Enfin, l'ensemble des hyperparamètres les plus performants obtenus par ce processus est utilisé pour réentraîner le modèle sur l’ensemble des données.

Les notions apportées dans ce chapitre présentent les méthodes appliquées dans le cadre de ce mémoire : les modèles fréquence, coût moyen et la détermination du seuil d’écrêtement, les principes du modèle linéaire généralisé (GLM) classiquement utilisé pour modéliser la prime pure, la création d’un zonier pour déterminer le risque géographique et les méthodes de *Machine Learning* retenues pour la classification des activités et pour la modélisation des sinistres atypiques. Une étape préalable à leur mise en œuvre est la constitution de la base de données. Le chapitre suivant décrit les sources de données utilisées et leur traitement.

3. La constitution de la base de données

Tout l'enjeu des modèles réside dans le choix et la fiabilité des données. Ce chapitre est consacré à la présentation des données. La base principale est la base des contrats, appelée base portefeuille car elle contient les caractéristiques des risques en portefeuille. Le périmètre de modélisation concerne le portefeuille du produit multirisque professionnel d'Allianz : Profil pro, qui couvre les **commerçants** et les **artisans** uniquement en **monosite** (un seul site pour la production et la commercialisation), sur la **France métropolitaine**. Toutes les activités présentes dans la nomenclature (près de 400 activités) sont retenues. Et, les deux réseaux de distribution : **agents** et **courtiers** sont pris en compte. A cette base vient s'ajouter des données internes sur les **clients** mais aussi des informations externes sur les entreprises issues de la base **Sirene** et sur le risque géographique. L'évaluation de l'impact de la pandémie demande une adaptation appropriée de la préparation des données avec l'intégration d'indicateurs pour identifier les périodes de restriction.

Pour estimer le coût espéré par profil de risque, les données descriptives du risque sont rapprochées du montant et du nombre de sinistres observés dans le passé. L'analyse est présentée sur les deux principales garanties dommages de la multirisque : l'**incendie** et le **dégât des eaux**, permettant d'illustrer les résultats sur une garantie d'intensité et une garantie de fréquence. Les sinistres dégât des eaux sont connus et évalués rapidement, généralement dans l'année, tandis que les sinistres incendie peuvent mettre plusieurs années avant d'être clôturés surtout dans le cas de sinistres importants. Pour tenir compte de ces contraintes et avoir un volume suffisant de sinistres atypiques, un historique de 5 ans : **2016** à **2020** est retenu.

Un certain nombre de retraitement et de mise en forme sont nécessaires avant de joindre les données risques à la sinistralité associée. Ce chapitre se compose de deux sections. La première décrit les données utilisées : source, fiabilité, retraitements, etc. La seconde s'intéresse aux étapes nécessaires à la création de la base utilisée en entrée des modèles.

3.1. La qualité et le retraitement des données

Les erreurs de saisie et de stockage des informations dans les systèmes d'information peuvent amener à des incohérences dans les données. Des contrôles de qualité et des retraitements sont réalisés sur les données extraites.

3.1.1. Les données descriptives du risque

Les données descriptives du contrat et du risque sont stockées dans une base appelée **portefeuille**. Cette base, spécifique au produit, donne la situation de chaque contrat ainsi que les caractéristiques du risque à chaque fin de mois. Pour le mémoire, toutes les visions mensuelles de **janvier 2016 à fin décembre 2020** de la base portefeuille dédiée au produit Profil Pro ont été concaténées, afin de récupérer la situation exacte du risque au moment du sinistre.

Les données sur le **contrat** sont : le numéro du client, le numéro du contrat, sa date de création et de résiliation ainsi que la date d'échéance. Y sont également restituées, les garanties choisies ainsi que la prime annuelle associée, les annexes et options, et le niveau de franchise.

Les données descriptives du **risque** regroupent l'ensemble des informations demandées à la souscription. Pour couvrir les caractéristiques de l'ensemble des activités, près de cent critères sont stockés dans la base portefeuille. Parmi les plus importants se trouvent les montants de capitaux garantis, le contenu, la superficie, le chiffre d'affaires et le code Insee de la commune du lieu du risque. Y sont également stockées des données plus spécifiques comme la capacité des chambres froides pour les restaurants, le nombre de chambres pour les hôtels, le capital périodique pour les activités avec des variations saisonnières comme les fleuristes et les chocolatiers. Les mesures de prévention incendie : extincteurs, électricité contrôlée, thermographie, et de protection contre le vol : alarme, système de détection d'intrusion sont aussi indiquées.

Sur les données de la base **portefeuille** (contrats et risques), les vérifications suivantes sont effectuées :

- le numéro de contrat permet d'identifier chaque contrat de manière unique ;
- la date de début de souscription est bien inférieure à la date de fin ;
- la date de résiliation est inférieure ou égale à la date de fin de contrat ;
- les montants de primes annuelles sont bien positifs ;
- l'indexation des capitaux.

Le tableau ci-dessous présente le nombre de contrats par année d'exercice :

ANNEE D'EXERCICE	NOMBRE DE CONTRATS
2016	117 697
2017	120 563
2018	121 936
2019	123 316
2020	122 909
TOTAL	606 421

Tableau 3.1- Nombre de contrats dans la base portefeuille

Sur les **606 421 contrats** extraits de la base portefeuille pour l'analyse, aucune incohérence n'est ressortie sur les données, toutefois un retraitement est effectué sur les **capitaux** assurés. Ces derniers, déclarés à la souscription, ne sont jamais réévalués dans le système de gestion. Un retraitement permet de les indexer en appliquant le rapport entre l'indice FFB[3] 2020 et celui à la date de souscription :

$$\text{Capitaux assurés vus 2020} = \text{capitaux assurés à la souscription} \times \frac{\text{indice FFB 2020}}{\text{indice FFB à la date de souscription}}$$

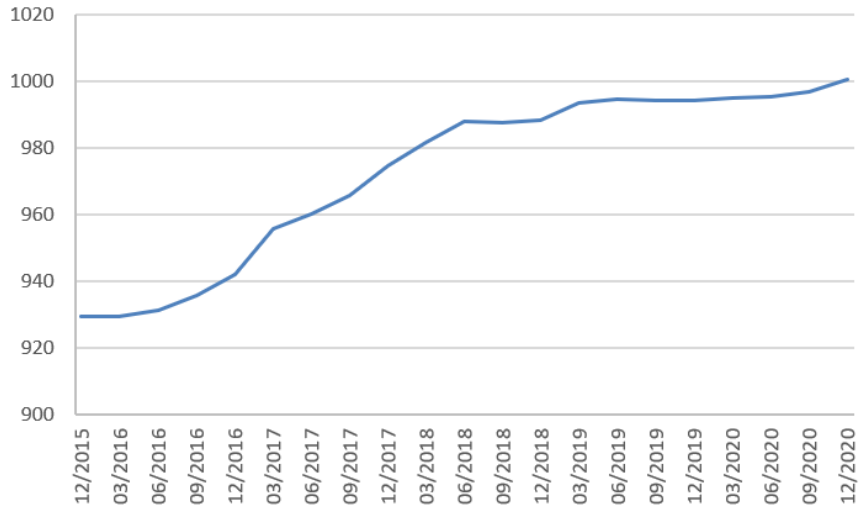


Figure 3-1 - Evolution trimestrielle de l'indice FFB du coût de la construction

[3] FFB : l'indice FFB du coût de la construction est calculé par la Fédération Française du Bâtiment, à partir du prix de revient d'un immeuble de rapport de type courant à Paris. Il enregistre les variations de coût des différents éléments qui entrent dans la composition de l'ouvrage. Ce calcul ne prend pas en compte la valeur des terrains.

L'indice ffb est passé de 930 à 1 000 entre décembre 2015 et décembre 2020, soit une évolution de 8%. D'où l'importance de réévaluer le montant des capitaux assurés depuis la date de souscription afin de connaitre le coût du capital garanti au moment du sinistre.

Outre les contrôles de cohérence des données, des **transformations** sont nécessaires pour récupérer de l'information qui n'est pas directement disponible dans les données brutes. Les deux principales transformations sont la création de critères de **durées**, d'**âge** ou d'**ancienneté** à partir de dates. Par exemple l'ancienneté du contrat, de l'entreprise à partir de leur date respective de création. Mais aussi, des transformations plus complexes comme la **transformation logarithmique** afin de synthétiser de manière plus efficace l'information. En rapprochant les valeurs extrêmes, la transformation logarithmique permet d'obtenir une distribution moins étendue et de corriger une asymétrie dans les données d'origine. La transformation logarithmique est appliquée aux valeurs de sinistre maximum possible **(SMP)** en fonction du contenu ou de la superficie.

3.1.2. Les données sinistres

La base des sinistres est commune à tous les produits IARD (Incendie, Accidents et Risques Divers), et comporte une ligne par garantie sinistrée : un sinistre peut toucher différentes garanties et avoir plusieurs lignes. Toutes les fins de mois, une vision est restituée sur l'historique des survenances.

Pour le mémoire, la situation à **fin mars 2021** a été retenue (dernière vision des sinistres disponible au moment de la création de la base de données pour les modèles), avec une sélection sur le produit Profil Pro et sur les **survenances 2016 à 2020**.

Les montants renseignés sur chaque ligne se réfèrent à une garantie sinistrée. Y sont indiqués :

- les paiements : le montant déjà payé au client ;
- les provisions : le montant attendu restant à payer avant la clôture de la garantie ;
- les frais encourus pendant le traitement du sinistre (les frais juridiques par exemple) ;
- les recours : les montants reçus en cas de présence d'un tiers responsable par exemple. Les recours ne sont pas signés dans la base de données.

Le coût total est calculé comme la somme des paiements (y compris les frais) et des provisions, à laquelle les recours sont déduits.

Les modèles de risque requièrent un enregistrement par sinistre. La vue par garantie est transformée en une vue par sinistre, en transposant les montants par garanties en colonnes et en les agrégeant par sinistre.

Le numéro de contrat rattaché au sinistre est également indiqué, permettant de faire le lien avec les données risques.

Sur les données **sinistres**, les contrôles de cohérence sont les suivants :

- le numéro de sinistre permet d'identifier chaque sinistre de manière unique ;
- le numéro de contrat associé à un sinistre existe bien dans la base contrat et permet de le rattacher à celui-ci ;
- les montants de règlement des sinistres sont positifs ;
- le montant des provisions des sinistres clos est nul ;
- le changement de convention CIDRE/IRSI a bien été appliqué.

Sur l'ensemble des sinistres, seuls 26 ont un numéro de **contrat sans correspondance** dans la base risque à cause d'une mauvaise affectation de produit. Ces sinistres sont exclus de l'analyse.

Le nombre de sinistres extrait pour l'analyse se présente de la manière suivante :

	INCENDIE				DEGAT DES EAUX			
ANNEE	Nombre de sinistres	Poids des clos	Poids des sinistres CIDRE - IRSI	Poids de la charge CIDRE - IRSI	Nombre de sinistres	Poids des clos	Poids des sinistres CIDRE - IRSI	Poids de la charge CIDRE - IRSI
2016	1 226	97,8%	-	-	4 251	98,1%	19,3%	7,0%
2017	1 324	95,8%	-	-	4 317	96,6%	17,5%	5,9%
2018	1 510	92,8%	0,3%	0,04%	5 494	93,6%	18,5%	6,9%
2019	1 563	74,1%	1,0%	0,1%	4 804	79,1%	26,2%	10,6%
2020	1 221	39,0%	0,2%	1,4%	4 499	41,4%	26,4%	11,2%
TOTAL	**6 844**	**80,4%**	**0,3%**	**0,3%**	**23 365**	**81,9%**	**12,3%**	**8,5%**

Tableau 3.2- Répartition des sinistres par année

La répartition des sinistres met en avant un volume de sinistres bien supérieur en dégât des eaux (23 365 sinistres) qu'en incendie (6 844 sinistres) mais avec un coût moyen nettement inférieur (2 874€ en dégât des eaux et 21 823 € en incendie).

L'ampleur et la responsabilité des sinistres incendie expliquent le délai de clôture légèrement plus long sur cette garantie : 80,4% de sinistres **clos** contre 81,9% en dégât des eaux. Tous les sinistres clos ont bien un montant de provision à zéro.

L'assurance des professionnels est soumise à une convention établie entre assureurs pour faciliter le traitement des sinistres : la convention **CIDRE** (Convention d'Indemnisation Directe et de Renonciation à Recours en dégât des Eaux), remplacée depuis le **1er juin 2018** par la convention **IRSI** (convention d'Indemnisation et de Recours des Sinistres Immeuble). Le rôle de l'assureur est de gérer le sinistre du local de son assuré, en organisant notamment la recherche de fuite. En assurance multirisque professionnelle, la convention s'applique en cas de sinistres **dégât des eaux** ou **incendie** d'un montant **inférieur à 1 600 € HT** sur **les locaux professionnels et commerciaux.** L'impact est visible sur la garantie dégât des eaux avec une prise en charge de sinistres qui augmente depuis 2018. Mais, la grande nouveauté de la convention IRSI est l'application aux petits sinistres incendie. Dans la réalité, les sinistres incendies inférieurs à 1 600 € sont rares : 0,3 % des sinistres incendie ont été ouverts avec une convention IRSI depuis 2018. Afin d'identifier ce changement de convention dans les modèles, une variable « période_convention » est créée avec la modalité « CIDRE » si la période d'analyse est antérieure au 1er juin 2018 et « IRSI » la période est postérieure ou égale au 1er juin 2018.

Par ailleurs, des retraitements ont été effectuées. Les sinistres **sans-suite**, avec un **coût négatif** ou un coût **très faible (<10€)** sont exclus de l'analyse. Sont définis comme sans-suite des sinistres clos dont le coût est porté uniquement par des frais d'expert ou dont le montant des règlements, des provisions et des recours sont nuls. Ces sinistres ne sont pas retenus, car ils ne sont pas pris en charge par l'assureur. Les sinistres avec un coût négatif sont considérés comme une anomalie. En effet, en multirisque professionnels la charge (règlements + provisions – recours) ne peut être négative, car les recours sont calculés au coût réel (vétusté déduite) et ne sont pas forfaitaires. Le tableau ci-dessous reprend le nombre, le coût et le poids de ces retraitements sur les deux garanties analysées :

	Nombre				Coût en €		
	Sans-suite	Coût < 0	Coût = 0	Coût ∈]0;10]	Sans-suite	Coût < 0	Coût ∈]0;10]
Incendie	933	103	110	16	194 128	-172 558	46
	13,63%	1,50%	1,61%	0,23%	0,13%	-0,12%	0,00%
Dégats des eaux	4 329	159	65	24	668 860	-321 588	54
	18,53%	0,68%	0,28%	0,10%	1,00%	-0,48%	0,00%

Tableau 3.3- Sinistres retraités en nombre et en coût

Les sinistres sans-suite représentent 13,63 % des sinistres incendie et 18,53 % des sinistres dégât des eaux, avec un coût négligeable. Les prendre en compte viendrait fausser la fréquence en la surestimant. Seuls les sinistres qui entrainent une charge pour l'assureur sont pris en compte dans la modélisation.

Les sinistres ouverts au **forfait d'ouverture** sont également identifiés par un top afin de les retraiter pour ne pas perturber la distribution des coûts moyens (les sinistres ouverts au forfait sont retraités uniquement pour l'estimation du coût moyen mais pas de la fréquence). Les graphiques ci-dessous illustrent l'impact des forfaits d'ouverture sur la distribution des coûts des sinistres incendie et dégât des eaux :

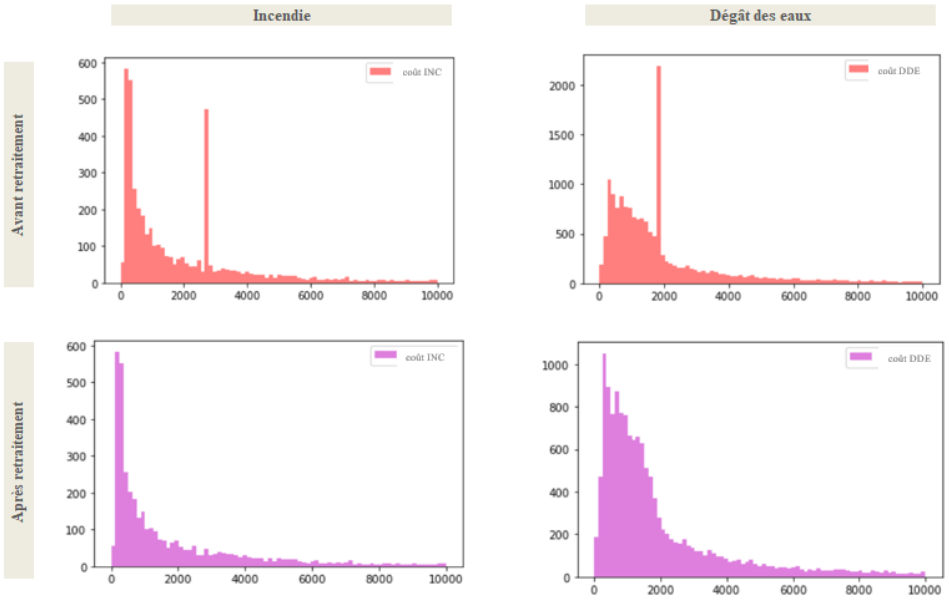


Figure 3-2 - Distribution du coût des sinistres avant et après retraitement des forfaits d'ouverture

En **incendie**, sur 6 844 sinistres, 445 ont un coût équivalent à celui du forfait d'ouverture identifié à **2 710 €.** Tandis qu'en **dégât des eaux** 1 823 sinistres sur 23 365 sont au forfait d'ouverture de **1 880 €.** Le pic formé par ces sinistres est très nettement visible dans la distribution des coûts.

3.1.3. Les données clients

La base des clients Allianz contient pour chaque client, particulier ou professionnel, les données permettant de l'identifier : numéro client, nom, date de naissance, adresse, etc. ainsi que des informations sur l'ensemble des contrats souscrits chez Allianz, quelle que soit la branche d'activité : IARD, santé, vie. Des informations sur la rentabilité du client : valeur client, contrats en surveillance, contentieux, montant des primes par branche d'activité, etc.

Pour les clients professionnels, la base client est complétée par des données **externes** issues de la base **Sirene**[4] et de l'entreprise d'assurance-crédit **Euler Hermes.** La base Sirene apporte des renseignements sur l'entreprise : numéro SIREN (système d'identification du répertoire des entreprises) qui sert à identifier l'entreprise en tant qu'entité, numéro SIRET (système d'identification du répertoire des établissements) qui identifie chaque établissement de l'entreprise, nombre de salariés, date de création, informations sur le dirigeant, etc. Alors que l'entreprise Euler Hermes fournit la note sur le risque de faillite de l'entreprise.

La base client est mise à jour tous les mois. La situation mensuelle des clients est rapprochée de celle des contrats par le numéro du client.

[4] La base Sirene rassemble des informations économiques et juridiques sur 30 millions d'établissements, appartenant à tous les secteurs d'activité.

Sur les données provenant de la base **client**, les principaux contrôles sont :

- la cohérence du numéro du client avec celui de la base portefeuille ;
- l'alimentation du code Siret.

Le code Siret est indispensable pour récupérer des informations complémentaires sur l'entreprise dans la base Sirene. Normalement, le Siret est demandé à la souscription. Si le client ne peut pas le fournir (lors de la création d'une entreprise par exemple), le code est récupéré par la suite dans la base Sirene à partir des renseignements transmis par le client (nom, adresse de l'entreprise, etc.).

3.1.4. Les données géographiques

Le portefeuille n'est pas géocodé. L'information géographique disponible dans les bases au niveau du contrat est le code de la commune (code Insee élaboré par l'Institut national de la statistique et des études économiques, il y a près de 35 000 codes Insee en France). Le code Insee enregistré est normalisé et contrôlé à la souscription. Seul des contrats monosites sont souscrits sous le produit Profil pro : un contrat ne peut être rattaché qu'à un seul site.

L'information interne n'étant pas suffisante pour spécifier le risque lié au lieu géographique, des données climatiques sont récupérées du service des risques climatiques d'Allianz et des données externes, disponibles en open data, sont collectées via les sites suivants :

- https://www.insee.fr : site de l'Insee qui collecte, produit, analyse et diffuse des informations sur l'économie et la société françaises.
- https://www.data.gouv.fr : plateforme ouverte de données publiques françaises.
- OpenStreetMap (OSM) : projet collaboratif de cartographie en ligne qui vise à constituer une base de données géographiques libre du monde, en utilisant le système GPS et d'autres données libres.

Les informations sont recueillies à la **maille Insee** (code commune) **ou département**. Les grandes catégories de données retenues sont les suivantes :

- **Socio Démographique** : données sur la population, les ménages, l'éducation, le logement, etc.
- **Revenu** : données sur l'impôt 2019, le niveau de revenu et de pauvreté.
- **Commune** : données sur les aires urbaines.
- **Criminalité** : données sur les crimes et délits enregistrés par la gendarmerie et la police.
- **Occupation des terres** : données sur la répartition des communes en termes d'occupation des terres.
- **Prix de l'immobilier** : prix moyen des ventes immobilières et coût de la construction.
- **Pompiers et Gendarmerie** : données sur les interventions en cas d'incendie ou de secours, ainsi que l'adresse des pompiers et des postes de police.
- **Climatique** : score inondation, sécheresse et zonier tempête, grêle, vent.

Le détail des sources des données géographique est donné en annexe.

Les données collectées sont transformées sous forme d'indicateurs calculés à la maille département et/ou Insee. En tout, près de 75 indicateurs sont créés. Par exemple, l'indicateur pourcentage de personnes âgées (> 64 ans) est calculé à partir des données suivantes :

$$\%\ \text{de personnes âgées } (> 64\ \text{ans}) = \frac{\text{Nombre de personnes âgées de 65 à 79 ans} + \text{Nombre de personnes de plus de 80 ans}}{\text{Population totale}}$$

Ces indicateurs fournissent des précisions sur la richesse, le niveau de vie, l'isolement, l'exposition aux risques de vol et d'incendie, les délais d'intervention des secours, les coûts de reconstruction de chaque commune. Ces facteurs influent sur l'exposition au risque en termes de fréquence ou de coût moyen. La commune d'implantation d'un professionnel est un critère important pour l'estimation de son risque (cf chap 1, § 1.3).

Sur les données géographiques, le contrôle se fait sur la cohérence des codes Insee entre ceux renseignés dans la base portefeuille et ceux provenant des données collectées en open data. Sur 3,74% de la base portefeuille, le code Insee est manquant ou n'a pas de correspondance avec celui des données géographiques. Dans ces cas de figure, la jointure de la base portefeuille avec les indicateurs géographiques est réalisée au niveau du département (les indicateurs géographiques étant calculés à la maille Insee et département).

3.1.5. Les données Covid

Le critère « Période_Covid » est défini sur les périodes suivantes :

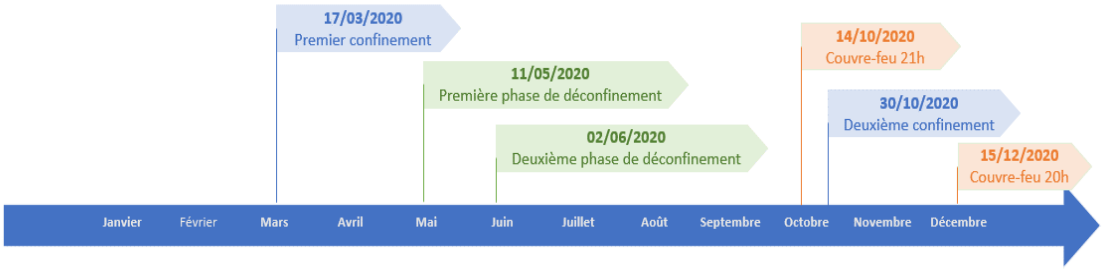


Figure 3-3 - Périodes de restriction (source : ECDC)

Ce critère divise les données dans le temps en fonction de la période de validité des restrictions de mobilité locales. Cependant, le déroulement des données selon le critère « Période_Covid » ne supprime pas la nécessité du découpage habituel en années d'accident. En effet, deux périodes avec des restrictions équivalentes peuvent être différentes en raison d'autres effets. La variable année d'accident est croisée avec ce nouvel indicateur pour créer le critère « **Année x Période_Covid** » : 2016-Standard, 2017-Standard, 2018-Standard, 2019-Standard, 2020-Standard, 2020-Premier_confinement, 2020-Première_phase_déconfinement, 2020-Deuxième_phase_déconfinement, etc.

Les commerces et les établissements n'ont pas tous été affectés de la même manière par les différentes mesures de restriction. Par exemple, les commerces de première nécessité ont pu rester ouvert pendant les périodes de confinement. Les rubriques d'activité du produit Profil pro sont regroupées afin de distinguer des groupes d'activités selon l'impact des restrictions. La variable **Groupe_Activité** est définie de la manière suivante :

Rubrique Profil pro	Groupe_Activité
01, 05	Alimentation, Santé
08	Commerces de gros
14	Accueil de personnes âgées et autres hébergements
12, 13	Hôtels, golfs
09	Artisans du bâtiment
02	Restaurants
03, 04, 07	Autres commerces de détail

Figure 3-4 - Regroupement des rubriques d'activité liées aux périodes de restrictions

Pour chaque contrat en portefeuille, le groupe d'activité est identifié selon la rubrique à laquelle appartient l'activité. Le graphique ci-dessous présente sur les garanties incendie et dégât des eaux, les fréquences et coûts moyens écrêtés par groupe d'activité en 2020 comparés à la moyenne de 2016 à 2019, vus à fin mars de l'année N+1 :

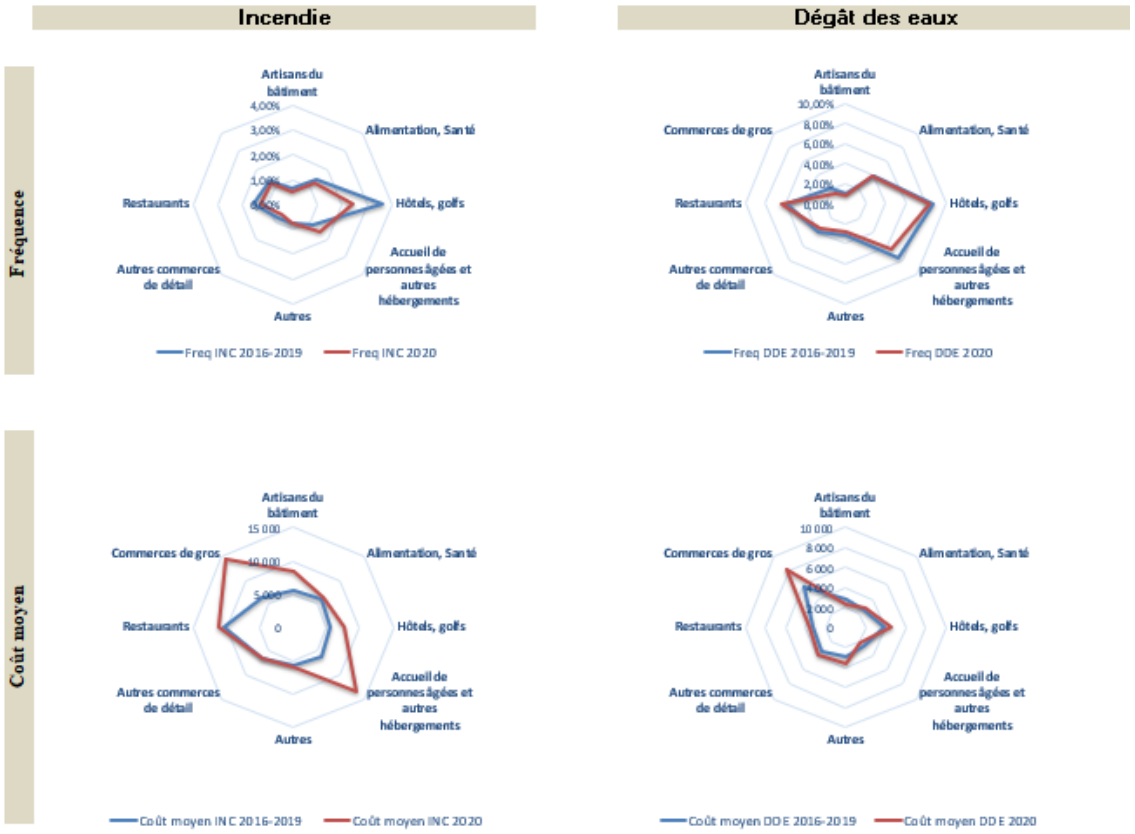


Figure 3-5 - Evolution 2016-2019/2020 de la fréquence et du coût moyen écrêté par groupe d'activité

La figure ci-dessus montre en 2020 une baisse de la fréquence incendie sur les hôtels (3,7 % en moyenne entre 2016 et 2019 contre 2,4 % en 2020), une des activités la plus touchée par l'arrêt de l'activité. En incendie, le coût moyen écrêté passe du simple au triple en 2020 sur les commerces de gros et les établissements d'hébergement. Plus de sinistres incendie importants sont constatés sur ces deux types d'activité. Le coût moyen des dégâts des eaux évolue également sur les commerces de gros (hausse de près de 40 %). Cette évolution du coût des sinistres peut être signe de déclaration frauduleuse.

En complément, un autre critère lié à la réduction de la mobilité est mis en place à partir de données « Google ». Depuis février 2020, Google met à disposition des chiffres et des tableaux de bord permettant de suivre l'évolution de la **fréquentation** par **département** sur **six types de lieux** : commerces et loisirs, alimentation et pharmacies, parcs, transports en commun, lieu de travail et lieux de résidence. « Ces rapports indiquent dans quelle mesure la fréquentation des différents lieux et la durée sur place varient par rapport à une référence propre au jour de la semaine concerné. Cette référence est la valeur de la médiane, pour un jour donné, calculée sur la période de cinq semaines comprise entre le 3 janvier et le 6 février 2020 ». Ces informations actualisées tous les jours sont accessibles via le lien : https://www.google.com/covid19/mobility/. La variable **Variation_mobilité_commerces** est construite à partir de l'évolution de la fréquentation journalière des **commerces et loisirs** par département. La catégorie de lieux retenue : commerces et loisirs (restaurants, cafés, centres commerciaux, etc.) est la plus représentative du portefeuille étudié. L'évolution de la fréquentation sur le secteur alimentation et santé (pharmacies, parapharmacies) est faible : ces commerces de première nécessité sont très peu impactés par les restrictions sanitaires. C'est pourquoi la variation de mobilité sur ce type d'activité n'est pas prise en compte. Les données Google par jour sont trop volatiles pour être exploitées telles quelles. Elles sont moyennées sur les différentes périodes de restriction.

En comparant les moyennes de variation de mobilité par périodes, de grandes différences ressortent par département :

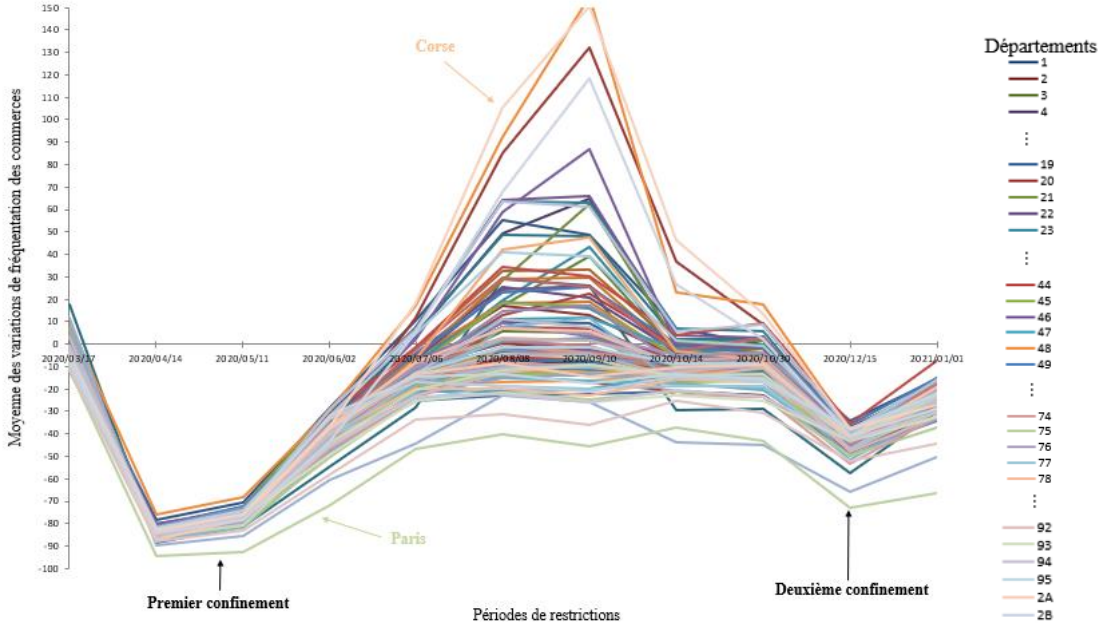


Figure 3-6 - Fréquentation des commerces par période et par département (source : Google)

La variation de fréquentation des commerces diffère visiblement selon la période de restriction et le département. La principale différence est observée durant l'été, avec de grandes variations de mobilité (100%, 150%, …) sur les départements touristiques du sud de la France, tandis que les plus faibles variations s'observent dans la région parisienne désertée par les touristes étrangers.

Pour chaque contrat en portefeuille, la variable **variation_mobilité_commerces** est identifiée en fonction de la période de restriction et du département. Le graphique ci-dessous présente les fréquences 2020 du portefeuille en incendie et dégât des eaux par variations de mobilité des commerces :

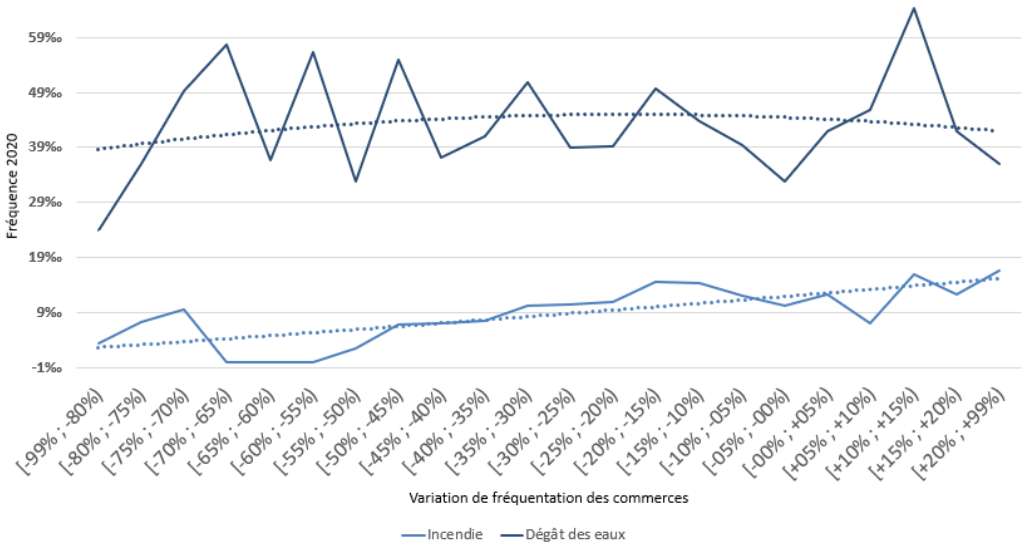


Figure 3-7 - Fréquence incendie et dégât des eaux par variation de fréquentation des commerces

L'évolution de la fréquence incendie par variation de mobilité a une tendance monotone et discriminante, contrairement à la fréquence dégât des eaux qui ne présente pas de tendance monotone.

L'ensemble des données utilisées proviennent des systèmes d'information ou de sources externes. Elles peuvent être codifiées pour faciliter le stockage, ce qui peut les rendre difficiles à interpréter. Le formatage des données permet de faciliter les analyses. Une bonne distribution pour une variable donnée permet de bien segmenter les données sous-jacentes, condition nécessaire pour une bonne modélisation.

3.1.6. Le formatage des données

Les variables catégorielles comportant des modalités avec très peu d'exposition ou des variables numériques avec des pics de concentration autour d'une valeur, entrainent généralement une perte d'information dans les modèles. Une mauvaise distribution pour une variable donnée se caractérise par une concentration autour d'une modalité. La réponse moyenne (fréquence ou coût moyen) devient alors représentative d'un portefeuille moyen. L'utilisation d'une telle variable dans les modèles n'apporte pas d'intérêt, car elle n'aide pas à créer une segmentation efficace, d'où l'utilité d'un formatage adéquate.

➢ La discrétisation des variables continues

Le logiciel Emblem utilisé pour la modélisation n'accepte que des variables catégorielles en entrée dont le nombre de modalités est limité à 250. La discrétisation des variables continues est nécessaire en amont de la modélisation. Un « banding » ou découpage en tranche, est utilisé pour catégoriser ces variables. La méthode consiste à diviser la variable d'origine en catégories distinctes qui peuvent être de largeur régulière ou non. Dans ce dernier cas, la distribution de données se fait via une transformation non linéaire. Le découpage en percentiles en est un exemple. Tout découpage implique une perte d'informations car le nombre de modalités du critère est réduit.

La discrétisation des variables continues est réalisée sous Python. Pour une variable numérique X_i, l'histogramme de la distribution pondérée par l'exposition donne une première indication sur le découpage à appliquer à la variable formatée pour améliorer la qualité du modèle. Par exemple, si la distribution est asymétrique vers la gauche, la variable formatée suivra une distribution normale.

Pour **chaque variable continue :** les indicateurs géographiques, les variables « contenu », « chiffre d'affaires », « superficie » … deux variables formatées sont créées, une avec une **distribution normale** et une avec une **distribution uniforme**. Les N observations de la variable X_i sont triées : $X_{i,1} <..< X_{i,j} <..< X_{i,N}$, et à chaque observation est associé un niveau p_j correspondant à la part de l'exposition représentée par l'ensemble des valeurs inférieures (exposition cumulée) :

$$p_j = \frac{\sum_{k=1}^{j} w_k}{\sum_{k=1}^{N} w_k} \text{ avec } w_k \text{ l'exposition de la } k^{ième} \text{ observation.}$$

L'espace est décomposé en n segments : $\alpha_1, \dots, \alpha_{n-1}$ avec $\alpha_{n-1} = 1 - \alpha_1$. Avec une loi normale $N(0,1)$, les niveaux de quantiles $F(\alpha_j)$ sont plus resserrés sur les bords :

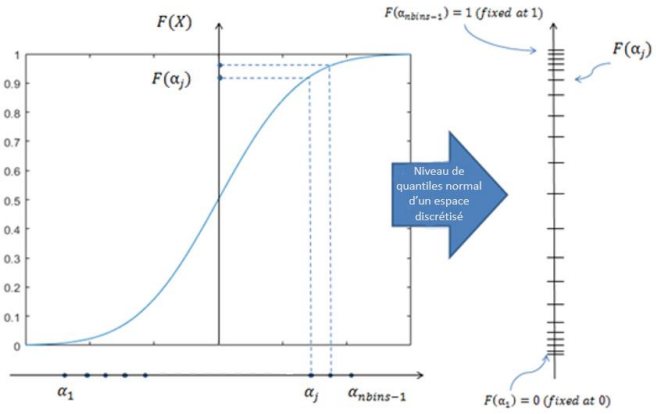


Figure 3-8 – Discrétisation par une loi normale

Avec une loi uniforme $U(0,1)$, les niveaux de quantiles $F(\alpha_j)$ sont réguliers :

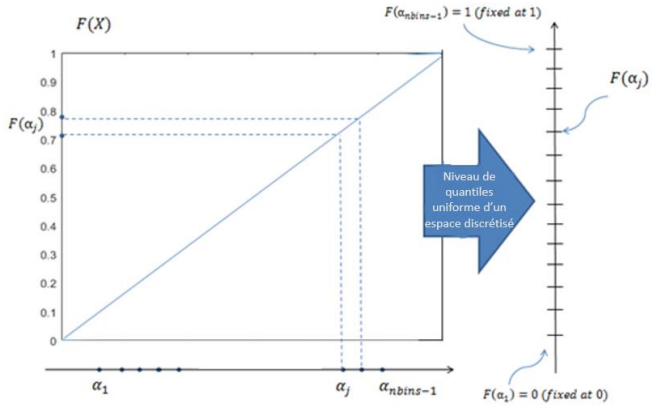


Figure 3-9 – Discrétisation par une loi uniforme

Pour chaque niveau $F(\alpha_j)$ le poids p_k est défini tel que : $F(\alpha_j) \in]p_k, p_{k+1}[$ et la valeur $X_{i,k+1}$ au bord de la $j^{ième}$ tranche lui est associée.

Pour illustrer les banding gaussien et uniforme, ci-dessous la répartition de la variable géographique « % de personnes âgées (> 64 ans) » :

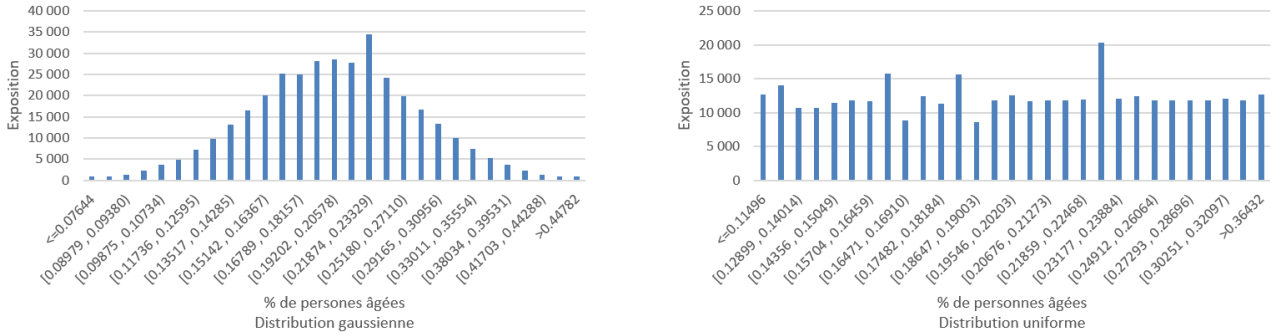


Figure 3-10 – Banding gaussien et uniforme de la variable « % de personnes âgées »

La distribution gaussienne affiche des tranches très concentrées au milieu et avec très peu d'exposition aux extrémités, tandis que la distribution uniforme permet d'avoir des tranches uniformément réparties.

➢ Le formatage des données catégorielles

Les variables catégorielles sont formatées pour une meilleure interprétation par un renommage des modalités ou un regroupement adéquat des modalités : une bonne connaissance des données permet d'identifier les modalités ayant la même signification et de les regrouper.

➢ L'encodage des erreurs ou valeurs manquantes

Pour chaque format créé, deux modalités sont ajoutées : une modalité « other » pour gérer des modalités non identifiées ou aberrantes, et une modalité « NA » (Not Avalable) pour identifier les valeurs manquantes.

Les données formatées ne viennent pas écraser les données brutes, mais seulement changer leur affichage, de manière à apporter plus de flexibilité en cas de modifications des formats dans le futur.

3.2. La construction de la base de données

Les données qui alimentent les modèles proviennent de sources internes et externes. Ces données comportent un maximum d'informations descriptives du risque et de la sinistralité associée. Pour pouvoir les restituer dans une base commune exploitable pour la modélisation, un certain nombre de traitements sont nécessaires.

3.2.1. Le calcul des expositions

Les contrats Profil Pro sont des contrats annuels renouvelés tous les ans par tacite reconduction. La base portefeuille comporte une ligne avec les caractéristiques de chaque contrat vu en début de mois. Les visions mensuelles sont concaténées par période de souscription, période pendant laquelle toutes les caractéristiques du contrat sont les mêmes. Une année de souscription, couvre généralement deux années calendaires, sauf en cas de résiliation ou d'avenant. Dans un modèle de risque, le but est d'estimer le coût sur la période de couverture de chaque contrat, uniquement sur la base de données historiques. Or, sur une année de souscription, une partie des données utiles pour estimer les coûts futurs n'est pas connue au moment de l'analyse. Se ramener à une année calendaire repose sur l'hypothèse que ce qui s'est produit dans le passé peut être reconduit dans le futur.

Pour passer d'une vision par année de souscription en année calendaire, chaque contrat couvrant plusieurs années est coupé en deux enregistrements : un nouvel enregistrement est créé à la fin de chaque année calendaire. Par ailleurs, les contrats en cours au début de la période d'analyse ont pu être souscrits avant la date de début d'analyse et peuvent se clôturer après la date de fin de la période d'analyse. Les dates de début et de fin de contrat sont alors forcées à celles de l'analyse.

Pour prendre en compte les différentes périodes de restrictions sanitaires en 2020, une maille plus fine que l'année calendaire est retenue pour le calcul des expositions. Les expositions sont calculées par **Année x Période_Covid**.

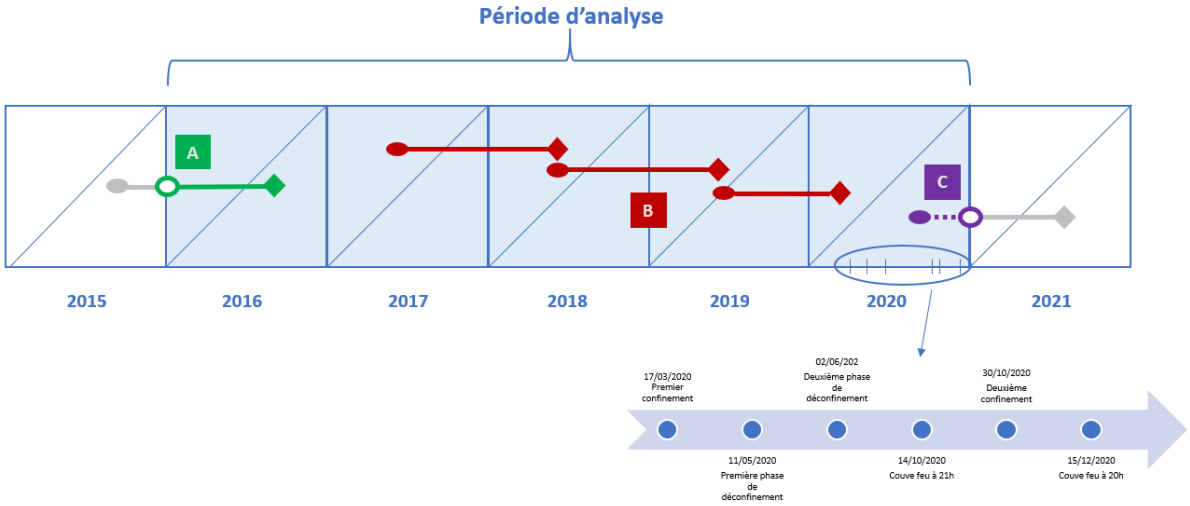


Figure 3-11 – Schéma de calcul des expositions

Pour les trois cas A, B et C ci-dessus, le retraitement des dates de début et de fin de période pour passer d'une vision année de souscription à année calendaire croisée avec les périodes Covid est le suivant :

- Cas A : date de création avant la date d'analyse

Numéro de contrat	Date début souscription	Date fin souscription
XXXXX	01/09/2015	01/09/2016

Numéro de contrat	Date de début de période	Date de fin de période
XXXXX	01/01/2016	01/09/2016

- Cas B : année de souscription sur plusieurs années calendaires

Numéro de contrat	Date début souscription	Date fin souscription	Date résiliation
YYYYY	01/06/2017	01/06/2018	
YYYYY	01/06/2018	01/06/2019	
YYYYY	01/06/2019	01/06/2020	10/04/2020

Numéro de contrat	Date de début de période	Date de fin de période
YYYYY	01/06/2017	31/12/2017
YYYYY	31/12/2017	01/06/2018
YYYYY	01/06/2018	31/12/2018
YYYYY	31/12/2018	01/06/2019
YYYYY	01/06/2019	31/12/2019
YYYYY	31/12/2019	17/03/2020
YYYYY	17/03/2020	10/04/2020

- Cas C : date de résiliation après la période d'analyse

Numéro de contrat	Date début souscription	Date fin souscription
ZZZZZ	01/10/2020	01/10/2021

Numéro de contrat	Date de début de période	Date de fin de période
ZZZZZ	01/10/2020	14/10/2020
ZZZZZ	14/10/2020	30/10/2020
ZZZZZ	30/10/2020	15/12/2020
ZZZZZ	15/12/2020	31/12/2020

Figure 3-12 – Identification des dates de début et de fin de période

L'objectif final étant de calculer un tarif technique pour les contrats futurs en utilisant les caractéristiques au moment de la souscription, et de le comparer avec le tarif commercial, tous les champs qui dépendent du temps, l'ancienneté du contrat par exemple, sont calculés à partir de la date de renouvellement (ou de souscription). Ainsi, la prime payée correspond aux caractéristiques de l'enregistrement.

Le split des enregistrements par année calendaire engendre plusieurs lignes par année pour un même contrat, et des observations avec des périodes de couverture distinctes. Pour se ramener à une vision annuelle, chaque enregistrement est pondéré par sa durée appelée exposition. L'exposition correspond au nombre de jours couvert par chaque enregistrement pondéré par le nombre de jours dans l'année :

$$\text{Exposition} = \frac{\min(\text{date de résiliation}, \text{date de fin de période}) - \text{date de début de période}}{\text{Nombre de jours de la période}}.$$

Le tableau ci-dessous présente les expositions calculées par année croisée avec les périodes Covid sur les garanties incendie et dégât des eaux :

Périodes	Nombre de contrats	Exposition Incendie	Exposition Dégât des eaux
2016 Standard	117 697	88 173	83 334
2017 Standard	120 563	90 040	85 409
2018 Standard	121 936	91 704	87 303
2019 Standard	123 316	92 164	88 032
2020 Standard	110 023	18 230	17 431
2020 1er confinement	107 713	14 866	14 215
2020 1re phase de déconfinement	106 506	5 531	5 290
2020 2nd phase de déconfinement	112 613	34 052	32 586
2020 Couvre-feu 21 h	107 554	4 062	3 889
2020 2ième confinement	109 524	11 646	11 151
2020 Couvre-feu 20 h	107 481	4 311	4 129

Tableau 3.4- Expositions par périodes

L'exposition est toujours inférieure au nombre de contrats, car tous les contrats ne restent pas une année civile entière en portefeuille (entrées, sorties en cours d'année). Par ailleurs, l'exposition est plus importante sur l'incendie que sur le dégât des eaux, car l'incendie est une garantie obligatoire sur ce produit, et est de ce fait plus souvent souscrite.

Le retraitement de la base portefeuille en base par périodes permet l'affectation des sinistres à la situation exacte du risque au moment de leur survenance. Un autre retraitement doit être apporté au niveau des sinistres pour identifier les sinistres atypiques, supérieurs à un certain seuil, des sinistres attritionnels. Plusieurs méthodes sont appliquées pour déterminer les seuils d'écrêtement par garantie.

3.2.2. La détermination des sinistres atypiques

➢ La distribution des sinistres

Un seuil est défini par garantie pour discriminer les sinistres peux nombreux avec un coût important qui peuvent fausser l'analyse. Ce seuil est déterminé à l'aide des graphiques de distribution du nombre de sinistres par tranches avec le pourcentage de sinistres et de coût au-dessus du seuil pour chaque tranche. Un bon candidat pour le seuil est le montant pour lequel les percentiles de nombre de sinistres et de coût commencent à montrer un changement structurel.

Le premier graphique analysé est le ratio entre le pourcentage de coût et le pourcentage de sinistres au-delà du seuil, avec en par tranche de sinistres abscisse la borne supérieure de chaque tranche (correspondants aux différents seuils) :

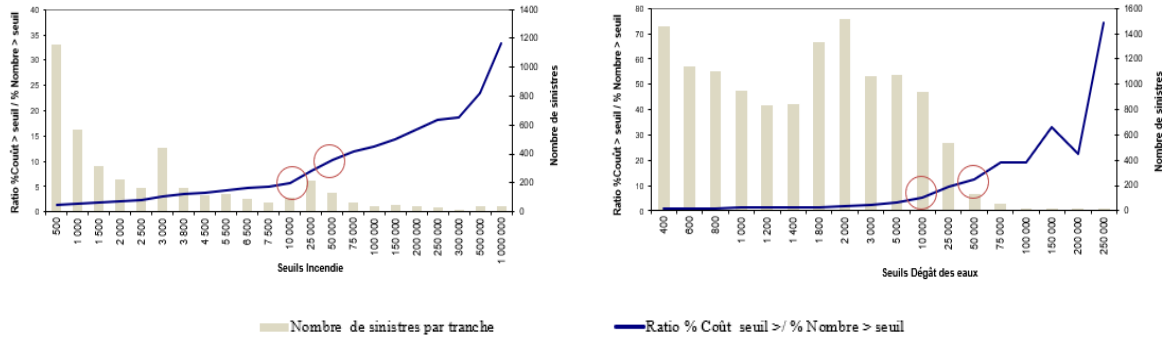


Figure 3-13 - Ratio du pourcentage de coût sur le pourcentage de nombre de sinistres au-delà du seuil

La courbe représente le coût sur le nombre de sinistres au-delà du seuil. La présence de cassures permet d'identifier les seuils de sinistres atypiques. Sur les deux garanties analysées, les deux premières cassures apparaissent au seuil de 10 k€ de 50 k€. Les graphes distinguant le pourcentage de nombre et de coût permettent d'affiner ce choix :

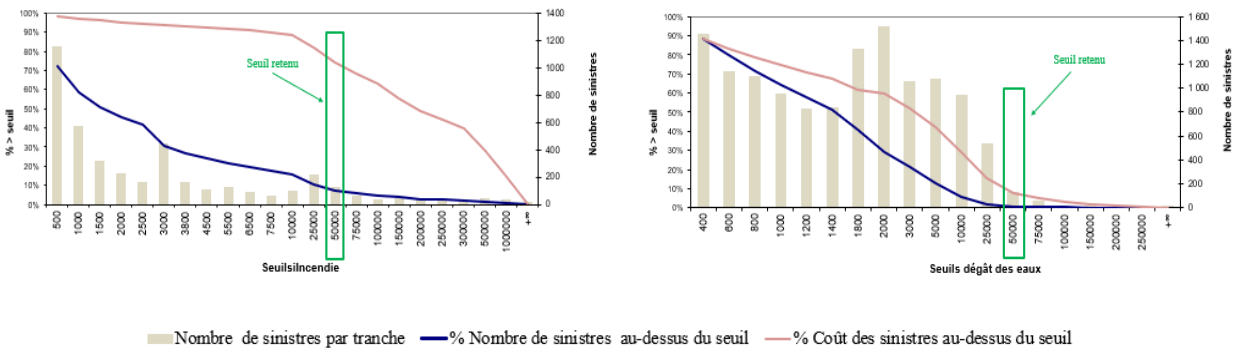


Figure 3-14 - Pourcentage de sinistres et de coût au-dessus du seuil par garantie et par année

L'écart important entre la courbe du pourcentage de nombre (courbe bleue) et celle du pourcentage de côut (courbe rouge) sur la garantie incendie caractérise une garantie d'intensité : le coût des sinistres au-dessus du seuil décroit beaucoup moins vite que sur la garantie dégât des eaux. Sur la garantie **incendie**, un chagement de tendance est observé à 10 k€ tant sur le coût que sur le nombre de sinistres au-dessus du seuil. Cependant pour garder un maximum de sinistres et assurer la stabilité de la modèlisation des attritionnels (sinistres en-dessous du seuil), le deuxième seuil à 50 k€, à partir duquel la tendance à la baisse du nombre de sinistres au-dessus du seuil se stabilise, est retenu. Pour la garantie **dégât des eaux**, un plateau du nombre de sinistres et une cassure de la baisse du coût au-delà du seuil de **50 k€** constribuent à retenir ce seuil.

Afin de vérifier si l'année 2020, impactée par la Covid, a des repercussions sur le choix du seuil des sinistres atypiques, les pourcentages de nombre de sinistres et de coût au-delà du seuil de 50k€ sur l'historique 2016-2020 sont comparés à 2016-2019 :

	HISTORIQUE 2016-2020		HISTORIQUE 2016-2019	
GARANTIE	Pourcentage de sinistres > 50 000 €	Coût > 50 000 €	Pourcentage de sinistres > 50 000 €	Coût > 50 000 €
INCENDIE	7,3%	73,4%	7,1%	71,1%
DEGAT DES EAUX	0,7%	7,6%	0,6%	7,9%

Tableau 3.5– Poids de la sur-crête en nombre et en montant

L'année 2020 ne modifie quasiment pas le pourcentage de nombre et de coût au-delà du seuil. Avec plus de 70% des coût au-dessus du seuil pour seulement 7% des sinistres, la distribution des sinistres de la garantie incendie se distingue de celle du dégât des eaux : une grande partie de la charge incendie est portée par un petit nombre de sinistres.

Les graphiques ci-dessous présentent le nombre de sinistres atypiques (> 50 k€) et le coût moyen de la sur-crête par années et par périodes de restrictions sur chacune des deux garanties :

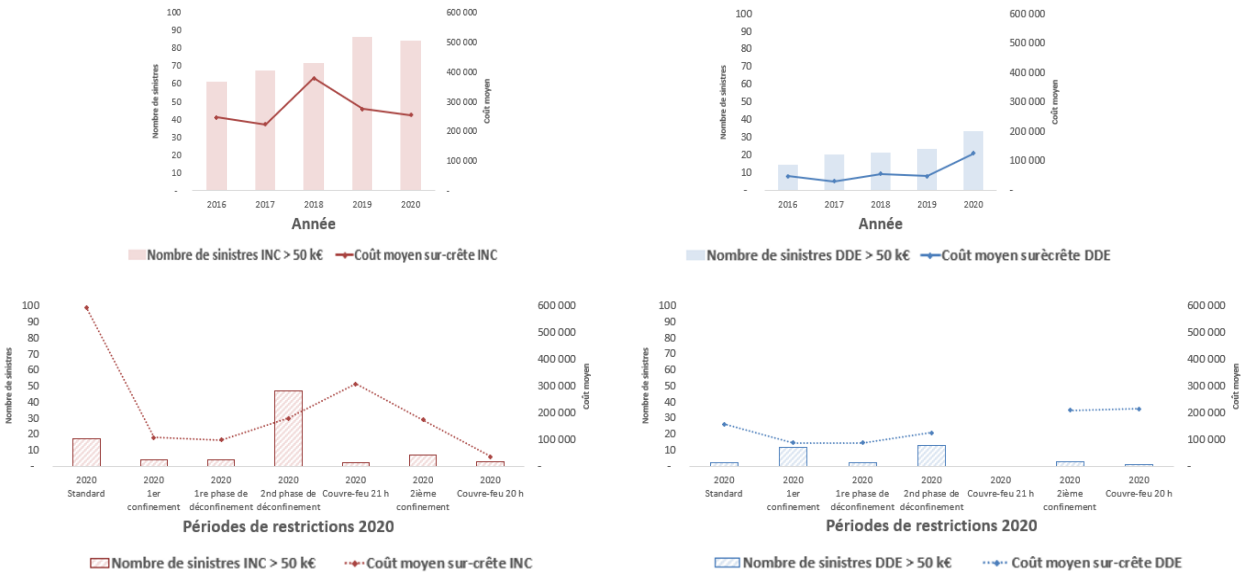


Figure 3-15 – Nombre de sinistres atypiques (> 50 k€) et coût moyen de la sur-crête par périodes

Le nombre de sinistres **incendie** atypiques est de l'ordre de 70 par an et un coût moyen de la sur-crête qui oscille autour de 250 k€. L'année 2018 est marquée par deux incendies d'un montant particulièrement important de 3,8 M€ sur une brasserie avec bureau de tabac et de 5,5 M€ sur un centre de golf. Les années les plus récentes : 2019 et 2020 se différencient par un nombre de 85 sinistres atypiques, plus élevé par rapport aux années antérieures. En zoomant sur les périodes de l'année 2020, la seconde phase de déconfinement, avec la reprise de l'activité, porte la majorité des sinistres atypiques de l'année. Le coût moyen du début l'année, avant

la mise en place des restrictions sanitaires, est quant à lui impacté par un sinistre incendie de 5,9 M€ sur une activité de « bar à sushis avec ou sans livraison de plats préparés avec matériel de cuisson ».

Sur la garantie **dégât des eaux**, l'année 2020 se distingue des autres années. Le nombre de sinistres supérieurs à 50 k€ est de 33 contre une vingtaine les années antérieures, et le coût moyen de la sur-crête s'élève à 125 k€ comparé 45 k€ en moyenne les années précédentes. Les volumes par périodes de restrictions sont trop faibles pour pouvoir apporter des conclusions par types de mesures, toutefois cette tendance à la hausse en 2020 s'explique des dégâts des eaux détectés plus tardivement du fait du manque de présence dans les locaux.

Sur la garantie dégât des eaux, le volume de sinistres atypiques (au-dessus du seuil) est insuffisant pour être modélisé : la sur-crête est répartie sur l'ensemble des assurés. En revanche, pour la garantie incendie, le nombre de sinistres atypiques peut être modélisé par un modèle de propension. Sur cette garantie, le choix du seuil d'écrêtement est également déterminé par la théorie des valeurs extrêmes afin de valider le seuil fixé par l'analyse de la distribution des sinistres.

➢ La théorie des valeurs extrêmes

La distribution des sinistres a permis de déterminer un seuil d'écrêtement de 50 k€ sur les deux garanties : incendie et dégât des eaux. Mais, le volume de sinistres atypiques nécessaires pour les modéliser n'est suffisant que sur la garantie incendie. La théorie des valeurs extrêmes, utilisée pour prévoir des évènements dépassant un certain seuil, est appliquée uniquement sur la garantie incendie.

Un **QQ-Plot** permet tout d'abord de vérifier que la distribution est bien à queue épaisse. Le QQ-Plot est un diagramme de probabilité utilisé pour évaluer si les données suivent une probabilité théorique donnée en traçant leurs quantiles. En supposant que les coûts des sinistres suivent une loi Gamma, le QQ-Plot de la garantie incendie ci-dessous, compare les coûts observés aux quantiles d'une loi Gamma :

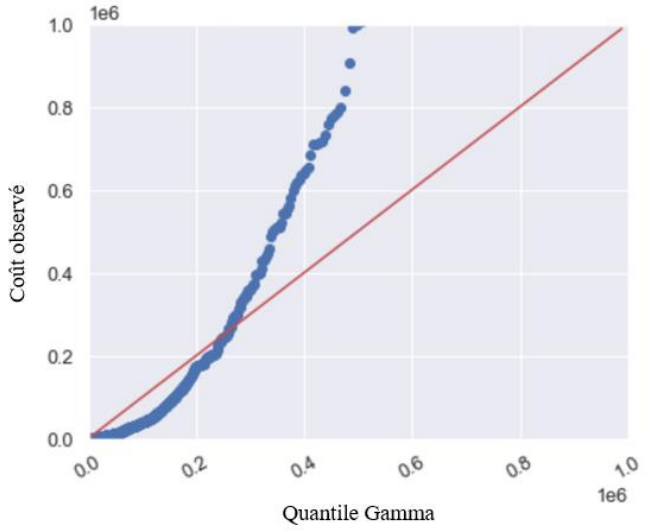


Figure 3-16 - QQ-plot de loi Gamma des coûts des sinistres incendie

La distribution observée s'écarte de la distribution théorique. La distribution des coûts des sinistres incendie ne suit pas une loi Gamma, mais une loi à queue épaisse. Plus la distribution observée est au-dessus de la bissectrice, plus un écrêtement des sinistres atypiques est nécessaire : les sinistres importants doivent être séparés des sinistres plus petits. Les sinistres au-dessus d'un certain seuil sont supposés suivre une loi de Pareto généralisée.

Les trois méthodes visuelles s'appuyant sur les propriétés de la GPD (distribution de Pareto généralisée) : le **paramètre de forme de la GPD**, la fonction **moyenne des excès** et le test de **Kolmogorov-Smirnov (KS)** sont analysées pour vérifier l'adéquation d'un seuil à 50 k€.

Le graphique de l'estimation du **paramètre de forme** ξ en fonction du seuil permet d'identifier visuellement le niveau de seuil à retenir :

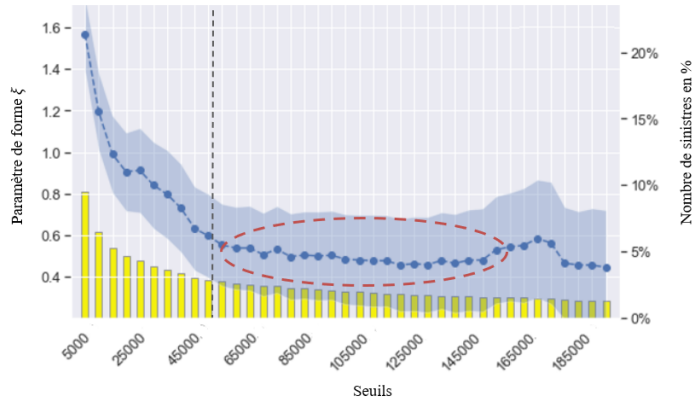


Figure 3-17 - Paramètre de forme ξ par seuils sur la garantie incendie

Le seuil optimal est le premier pour lequel le paramètre de forme ξ commence à être stable, c'est-à-dire d'après le graphe ci-dessus, entre 50 k€ et 55 k€.

La **fonction moyenne des excès** observée est calculée par seuils pour obtenir le graphique ci-dessous :

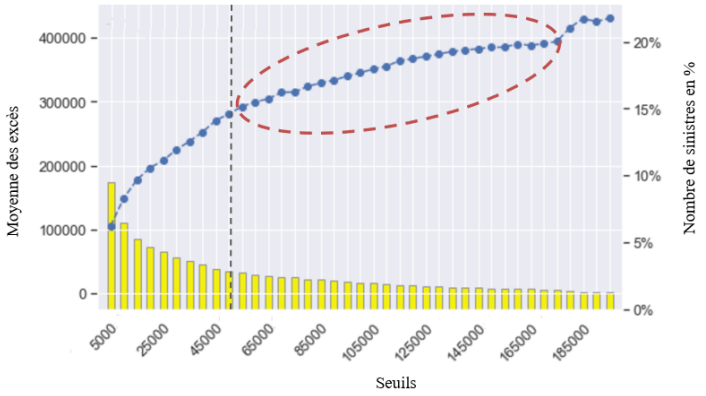


Figure 3-18 - Moyenne des excès par seuils sur la garantie incendie

Le seuil optimal est celui pour lequel la moyenne des excès commence à devenir linéaire. Un seuil de 50 k€ est donc approprié d'après la figure ci-dessus.

Et enfin, le graphique de la KS (**Kolmogorov-Smirnov)** en fonction du seuil donne :

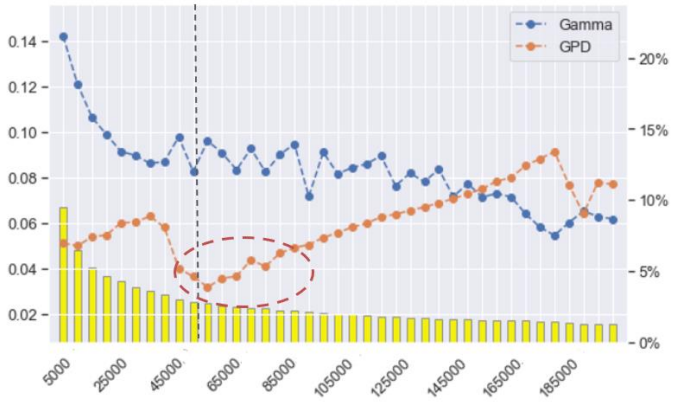


Figure 3-19 - KS statistique par seuils sur la garantie incendie

Le seuil optimal est celui pour lequel la KS statistique de la GPD atteint son minimum. D'après la figure ci-dessus, le seuil optimal se situe entre 50 k€ et 55 k€.

Les méthodes graphiques de la théorie des valeurs extrêmes confortent le choix du seuil de sinistres incendie atypiques à 50 k€.

Le seuil de sinistres atypiques fixé à 50 k€ pour les deux garanties permet d'écrêter les sinistres, et d'identifier les sinistres atypiques dont le coût est supérieur au seuil. Cette distinction est nécessaire pour modéliser les sinistres attritionnels d'une part et les sinistres atypiques d'autre part. Une autre contrainte pour constituer la base pour les modèles, est le nombre de modalités par critère limité à 255 par le logiciel de tarification. Cette limite impose de regrouper les activités en groupes homogènes de risque avant de les introduire dans les modèles.

3.2.3. La classification des activités

L'activité est un critère prépondérant dans la tarification du produit Profil pro. Avec près de quatre cents activités référencées dans la nomenclature, le tarif ne peut être segmenté à une maille aussi fine : le manque d'observation ne permettrait pas la fiabilité des méthodes statistiques, et le logiciel de modélisation utilisé (Emblem) n'accepte pas plus de 255 modalités par critère. L'objectif est de regrouper les activités à partir de critères les caractérisant. Les regroupements se font en deux temps : l'ACP permet d'analyser les activités en fonction des critères choisis, puis la CAH regroupe des individus en classe de risques homogènes à partir des axes de l'ACP. Les groupes ainsi constitués sont appelés classe « techniques » et se distinguent des classe « commerciales » déterminées à dire d'expert.

➢ La création de la base par activités

Afin d'analyser les activités sur un ensemble de critères numériques qui leurs sont propres, une base de données est construite avec en **ligne** les **activités** et en **colonnes** les **variables** agrégées par activité. Les activités sont analysées par garantie selon les variables suivantes :

- « CA » : le chiffre d'affaires ;
- « Cap*Gar* » : le capital garanti ;
- « Surface » : la surface ;
- « Effectif » : le nombre de salariés dans l'entreprise ;
- « AgeEnt » : l'âge de l'entreprise ;
- « Note » : la note de risque de faillite (plus la note est basse plus le risque de faillite est grand) ;
- « Cl1*Gar* » : la classe commerciale de la garantie (déterminée à dire d'expert) ;
- « ActSensi » : les activités dites sensibles (avec des clauses de préventions spécifiques) ;
- « CdPole1 » : le réseau de distribution agent (les profils de risque souscrit en agence sont différents de ceux souscrits en courtage). Sachant qu'il n'y a que deux réseaux de distribution : agent et courtage, les risques qui ne sont pas souscrits en agence sont souscrits par des courtiers ;
- « OffreQ » : le choix de l'offre : Qualité pro (l'offre Qualité pro étant réservée aux commerces alimentaires, habillement et autres commerces de la rue). Les autres risques sont souscrits sur l'offre Class pro ;
- « DAI » : la présence d'un système de détection automatique d'incendie ;
- « RIA » : l'installation de robinets d'incendie armés ;
- « ElecCtrl » : l'électricité contrôlée ;
- « Extinc » : la présence d'extincteurs mobiles vérifiés ;
- « Thermo » : la présence de contrôle par thermographie infrarouge ;
- « CCSprink » : le local est situé dans un centre commercial avec sprinkler.

Les variables par contrat sont agrégées par activité par une moyenne pondérée par l'exposition. La matrice de corrélation permet une première analyse des variables :

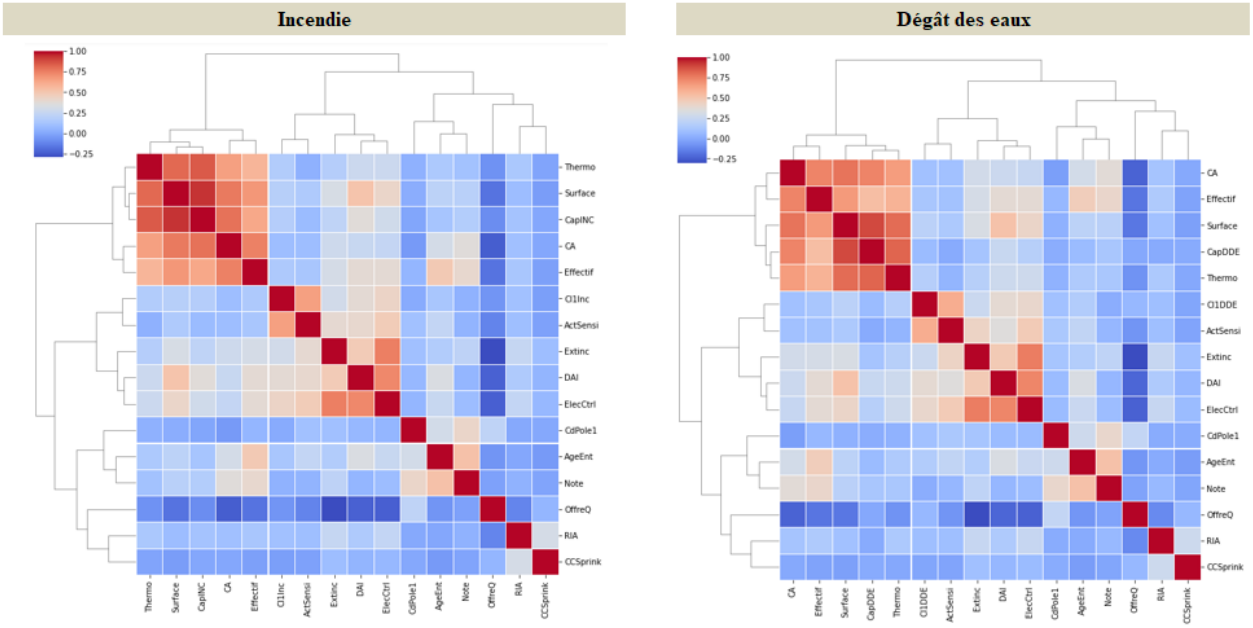


Figure 3-20 - Matrice de corrélation des variables caractéristiques des activités

Les matrices de corrélations des deux garanties sont très proches. Elles mettent en évidence la forte corrélation des variables de taille de l'entreprise (le chiffre d'affaires, la surface, le capital souscrit et l'effectif) entre elles et avec la présence de contrôle par thermographie infrarouge. La classe commerciale est corrélée aux activités « sensibles » (qui présentent un risque plus élevé), et plus l'entreprise est ancienne plus sa stabilité financière est bonne.

L'étude de la corrélation entre tous les couples de variables ne permet pas d'appréhender les relations existantes entre elles. La méthode de l'Analyse en Composantes Principales (ACP) est une méthode d'analyse des données en tenant compte de leur caractère pluridimensionnel.

➢ Les liaisons entre variables et les ressemblances entre activités : ACP

Pour analyser les activités en fonctions des critères chiffre d'affaires, surface, effectif …, une ACP est appliquée à la base de données par activités. Le but de l'ACP est de condenser l'information contenu dans la base par activités par une analyse des corrélations linéaires entre les variables et une visualisation graphique des distances entre les activités. Les variables retenues possèdent des unités de valeur sont très différentes. Pour rendre les données comparables et indépendantes de l'unité, elles sont dans un premier temps centrées et réduites.

L'ACP permet de représenter les variables de manière indépendante. Les variables sont représentées graphiquement par le cercle des corrélations. Ce cercle correspond à la projection des variables initiales sur un plan à deux dimensions dont les axes sont constitués de deux composantes principales (combinaisons linéaires des variables initiales).

Le graphique ci-dessous représente le cercle de corrélation sur les deux premiers axes :

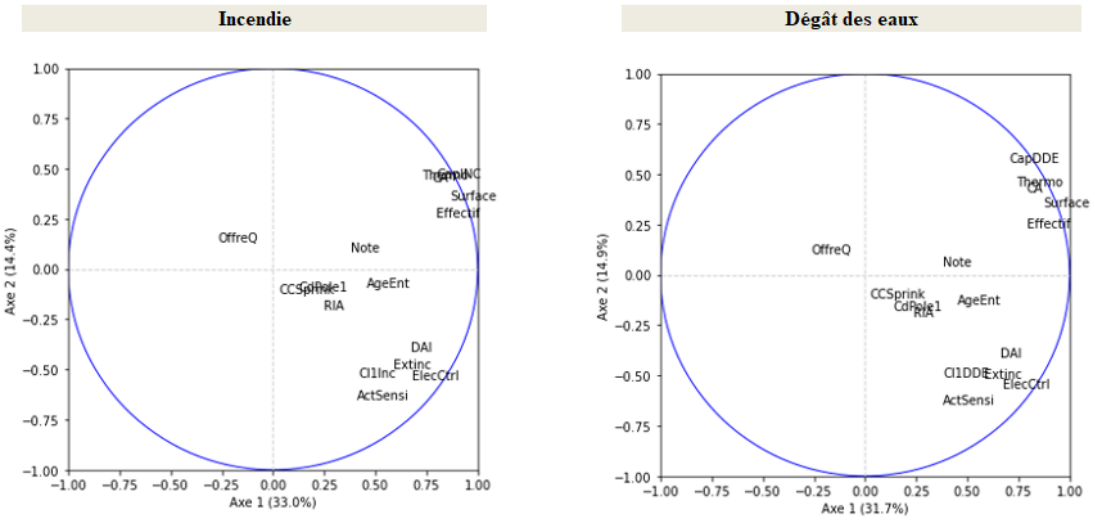


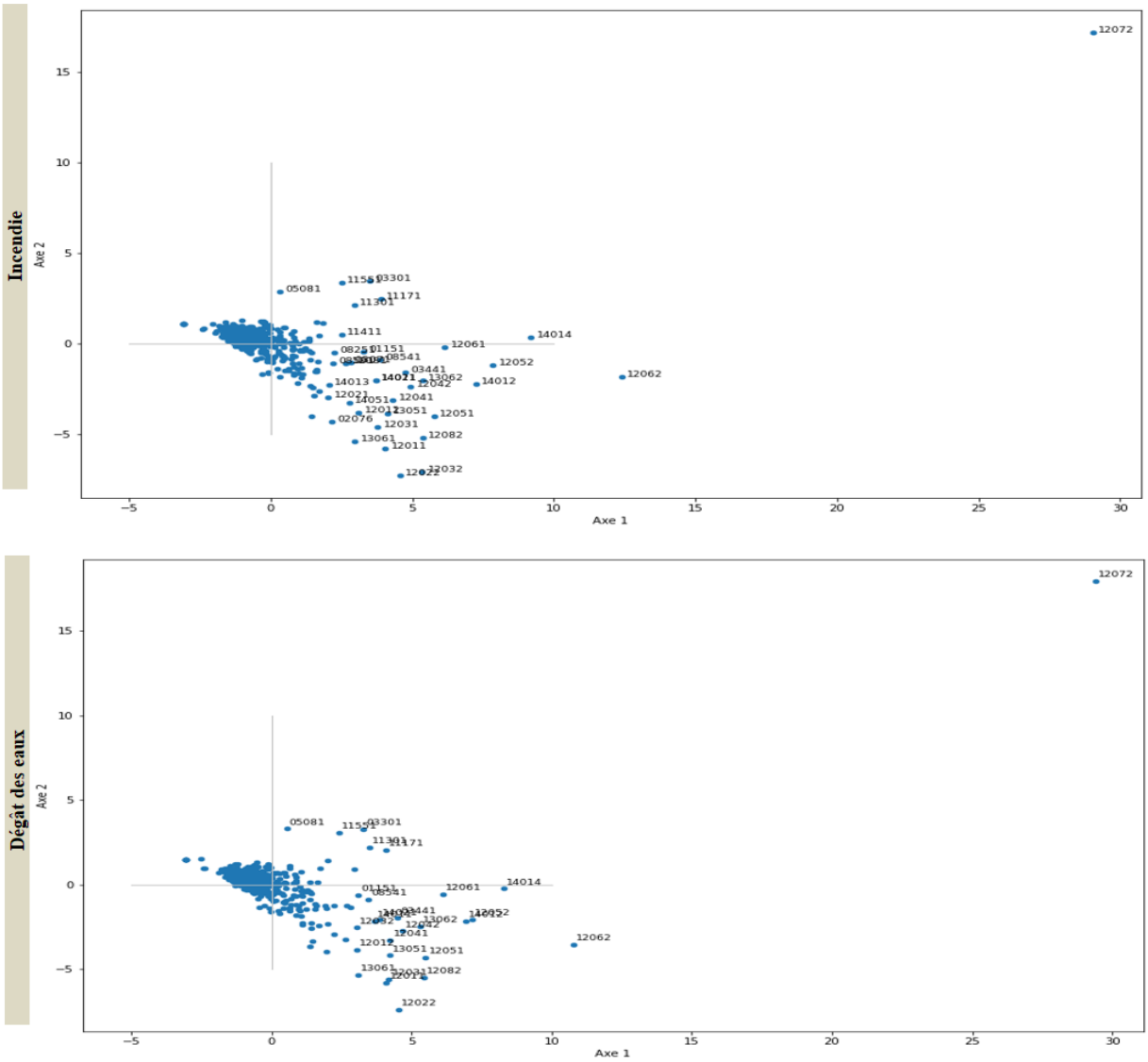
Figure 3-21 - Cercles de corrélations sur les axes 1 et 2

La première composante principale est celle qui résume le mieux les informations contenues dans la base par activités. La deuxième apporte un pourcentage inférieur mais complémentaire d'information, et ainsi de suite. Le cercle de corrélations représente la première composante (axe horizontal) et la seconde (axe vertical). La somme des pourcentages d'explication des deux composantes renseigne sur le taux d'information restitué par ces deux axes. D'après la figure ci-dessus, en incendie, la première composante résume 33 % et la seconde 14,4 % de l'information. Les deux premiers axes résument 47,4 % de l'information présente dans la base. En dégât des eaux, les deux premiers axes représentent 46,6 % de l'information.

Les variables renseignées à l'intérieur du cercle indiquent leur corrélation avec les composantes principales et conduisent à préciser la signification axes. Une variable proche du centre apporte peu d'information, une variable proche du cercle de corrélation (forte corrélation) et proche d'un axe contribue fortement à la formation de l'axe. Les angles inter-variables, en partant de l'origine, renseignent sur les corrélations entre elles. Deux variables séparées par un angle aigue par rapport à un axe sont corrélées positivement, les angles droits reflètent l'indépendance, et deux variables symétriquement opposées par rapport au centre sont significativement négativement décorrélées. Sur les garanties incendie et dégât des eaux, le premier axe apporte de l'information sur le chiffre d'affaires, la surface, le capital, l'effectif et la présence de contrôle par thermographie infrarouge qui sont des variables très corrélées entre elles. Tandis que le deuxième axe distingue les activités « sensibles » avec une clase commerciale élevée et la présence de moyens de prévention.

L'ACP permet aussi de calculer les coordonnées des activités sur les axes, leurs contributions à la dispersion selon chacun de ces axes et les cosinus carres. Les coordonnées permettent de réaliser des graphiques des activités. L'un des avantages de l'ACP est de pouvoir représenter les activités sur un plan à deux dimensions et identifier les tendances.

Le graphique des individus sur les deux premiers axes (représenté sur la page suivante) fait ressortir quelques points. Les palaces avec restaurant ont des caractéristiques différentes du fait qu'ils ressortent du groupe. Les hôtels ne sont pas tous regroupés ensemble : les hôtels cinq étoiles caractérisés par une surface, un capital et un effectif importants se distinguent des autres hôtels. De même pour les établissements d'accueil et d'hébergements qui selon leurs spécificités se rapprochent de l'axe 1 ou 2.



01151 Supermarche, surface de vente a dominante alimentaire n'excedant pas 2 500 m2
02076 Restaurant sans activite de pizzeria
03301 Vente de produits de telephonie et accessoires y compris telephones portables
03441 Magasin de bricolage et d'equipement de la maison
05081 Pharmacie
08251 Commerce de gros de cd et dvd
08541 Commerce de gros de produits de verrerie (a usage industriel ou d'emballage)
08591 Commerce de gros de materiaux de construction, d'appareils sanitaires et de chauffage vente aux professionnels et particuliers
11171 Selon libelle activite complementaire
11301 Selon libelle activite complementaire
11411 Selon libelle activite complementaire
11551 Selon libelle activite complementaire
12011 Hotel sans etoile sans restaurant
12012 Hotel sans etoile avec restaurant
12021 Hotel 1 etoile sans restaurant
12022 Hotel 1 etoile avec restaurant
12031 Hotel 2 etoiles sans restaurant
12032 Hotel 2 etoiles avec restaurant
12041 Hotel 3 etoiles sans restaurant
12042 Hotel 3 etoiles avec restaurant
12051 Hotel 4 etoiles sans restaurant
12052 Hotel 4 etoiles avec restaurant
12061 Hotel 5 etoiles sans restaurant
12062 Hotel 5 etoiles avec restaurant
12072 Palace avec restaurant
12081 Centre de golf sans hotel sans exploitation directe de pro-shop
12082 Centre de golf sans hotel avec exploitation directe de pro-shop
13051 Auberge de jeunesse ou centre international de sejour touristique
13061 Etablissement prive d'accueil collectif de mineurs de 4 a 14 ans en sejour de vacances
13062 Etablissement prive d'accueil collectif de mineurs de 4 a 17 ans en sejour de vacances
13071 Centre de thalassotherapie sans hebergement sans personnel medical ou paramedical salarie
14011 Etablissement prive d'hebergement pour personnes agees sans personnel medical ou paramedical salarie
14012 Etablissement prive d'hebergement pour personnes agees avec personnel medical ou paramedical salarie
14013 Etablissement prive d'hebergement pour personnes agees sans personnel medical ou paramedical salarie, avec assurance responsabilite des residents
14014 Etablissement prive d'hebergement pour personnes agees avec personnel medical ou paramedical salarie, avec assurance responsabilite des residents
14021 Residence-services seniors privee, sans personnel medical ou paramedical salarie
14051 Residence etudiante privee

Figure 3-22 - Graphes des individus sur les axes 1 et 2

Le pourcentage d'inertie des deux premiers axes, de l'ordre de 45 %, représente la part de l'information initiale conservée après projection dans le plan défini par ces deux axes. Le nombre d'axes à retenir est évalué en étudiant la courbe de décroissance des valeurs propres (« scree plot ») et l'évolution de l'inertie expliquée par les axes :

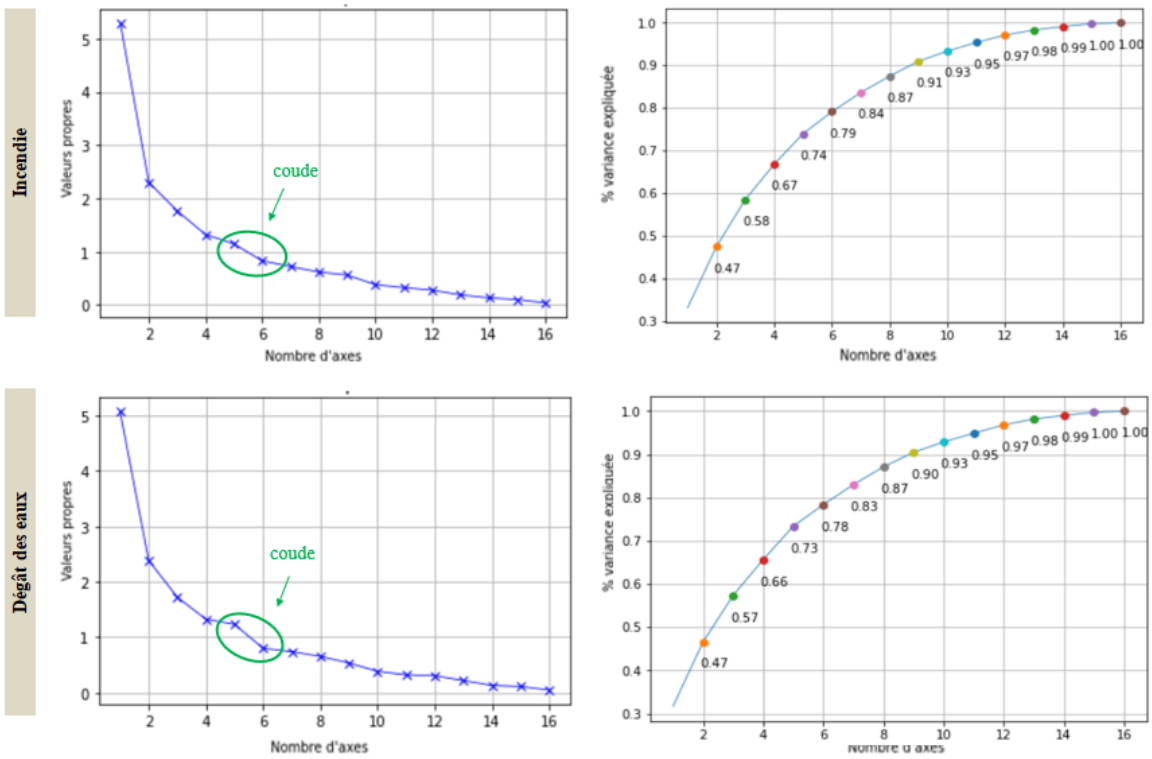


Figure 3-23 - Scree plot et pourcentage de variance expliquée

Le « scree plot » montre pour les garanties incendie et dégât des eaux, un coude entre 5 et 6 axes. Pour préciser la lecture du « scree plot », le graphique décrivant l'évolution de l'inertie expliquée par les axes indique que le gain informationnel après le sixième axe est faible pour les deux garanties. L'objectif étant de conserver un maximum d'information, une sélection de 6 axes pour l'incendie (79 % de l'inertie expliquée) et le dégât des eaux (78 % de l'inertie expliquée) est retenu.

L'ACP a permis de créer des axes, les composantes principales, synthétisant plusieurs informations. Une composante principale est une combinaison linéaire de plusieurs variables, telle que la variance (ou l'inertie) du nuage autour de cet axe soit maximale. Les composantes principales issues de l'ACP deviennent les nouvelles variables décrivant les activités et sont utilisées pour classifier les activités à l'aide de la classification ascendante hiérarchique (CAH).

➢ Le regroupement des activités : CAH

La classification ascendante hiérarchique (CAH) a pour but de constituer des groupes d'activités homogènes et différenciés les uns des autres, en s'appuyant sur les composantes principales de l'ACP. Elle s'effectue par fusions successives de groupes déjà existants ou d'activités encore isolées. La méthode choisie pour regrouper les classes est le critère de Ward, qui consiste à sélectionner à chaque étape le regroupement de classes dont l'augmentation d'inertie interclasse est maximum.

La distance Euclidienne est retenue comme mesure de distance entre activités, et la méthode de Ward est utilisée comme indice d'agrégation. Cette métrique et cette méthode sont appliquées pour calculer les distances des clusters, en commençant par n échantillons individuels, puis en fusionnant à chaque itération les deux clusters qui ont la plus petite distance.

Le dendrogramme permet de visualiser sous forme d'arbre, les regroupements successifs constitués par l'algorithme :

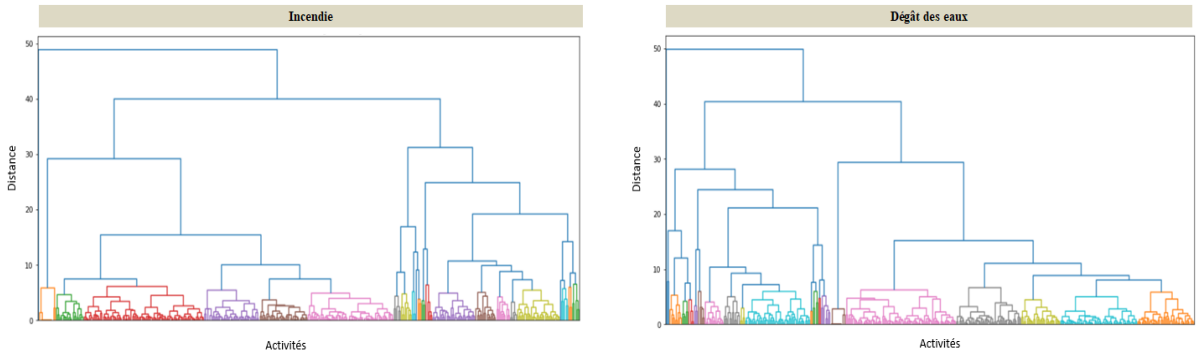


Figure 3-24 - Dendrogrammes

Les lignes horizontales du dendrogramme représentent les fusions de clusters, et les lignes verticales indiquent la distance entre deux classes, c'est-à-dire la distance « à combler » pour former une nouvelle classe. La hauteur de coupe de l'arbre détermine le nombre de classes. Elle est considérée comme pertinente si elle se trouve entre deux nœuds dont les hauteurs sont assez éloignées, ce qui revient à couper l'arbre où les branches sont assez longues. Le nombre de classes ne doit être ni trop petit : risque que les classes ne soient pas homogènes, ni trop grand : les classes ne se différencieraient pas suffisamment.

Le nombre de classes est déterminé par une analyse visuelle du dendrogramme, mais également par plusieurs métriques. Trois métriques sont utilisées pour choisir le nombre optimal de classes, le coefficient de **silhouette**, l'indice de **Calinski-Harabasz** et l'indice de **Davies-Bouldin :**

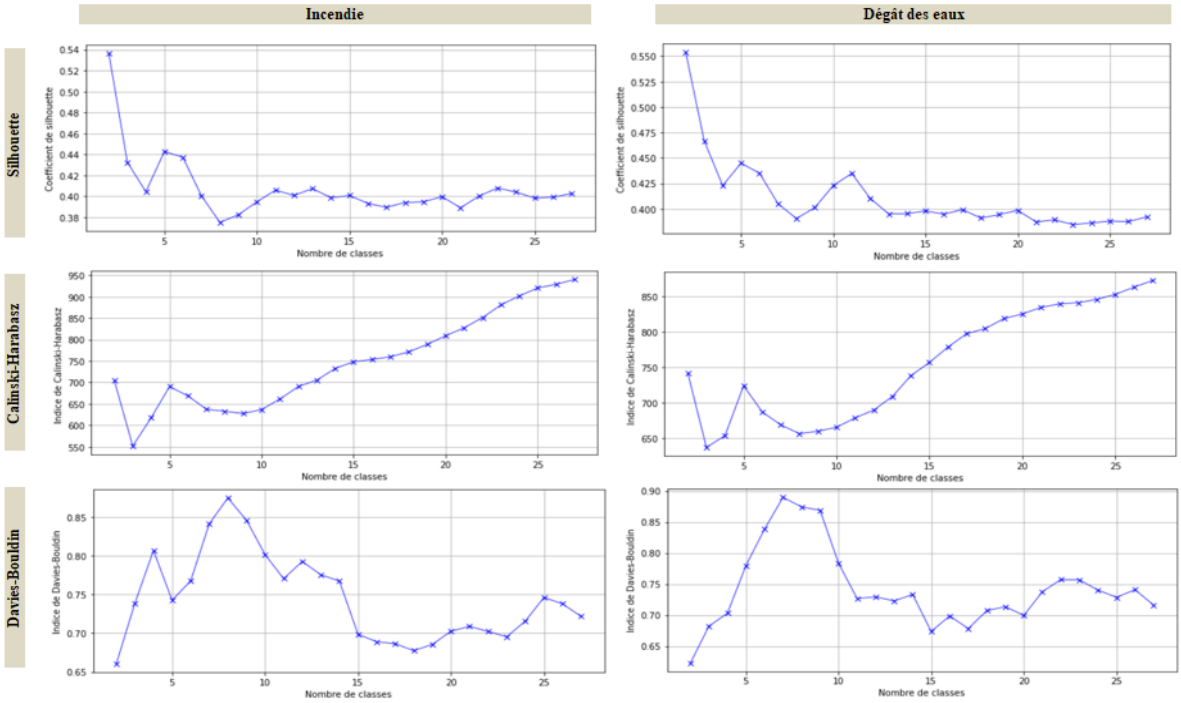


Figure 3-25 - Indicateurs du nombre de classes optimal

Le nombre de classes optimal doit correspondre à un coefficient de silhouette proche de 1, un indice de Calinski-Harabasz élevé et un indice de Davies-Bouldin bas. D'après la figure ci-dessus, un nombre de 5 classes est retenu pour l'incendie et le dégât des eaux.

Les dendrogrammes simplifiés coupés à 5 classes avec en abscisse le nombre d'activités par groupes ou le code de l'activité pour les singletons, sont représentés ci-dessous :

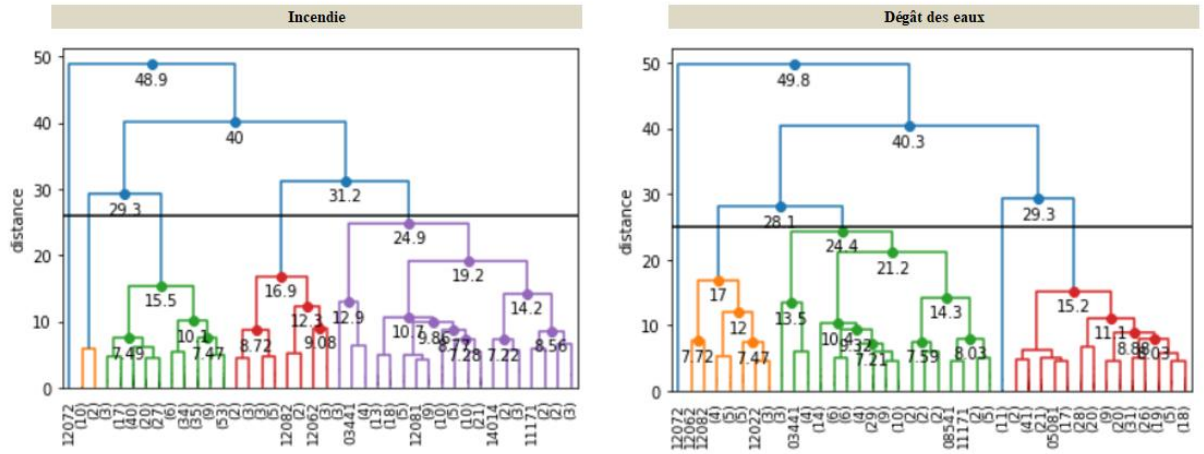
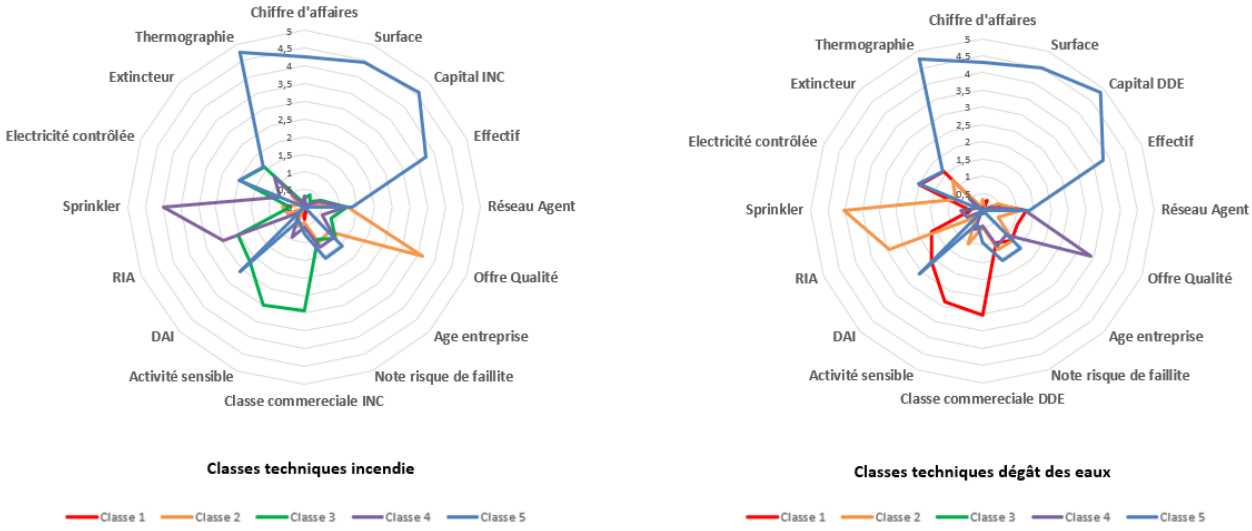


Figure 3-26 - Dendrogrammes simplifiés coupés à 5 classes

Visuellement, la longueur des branches conforte un découpage en 5 classes. Les palaces avec restaurant (code activité 12072) se dissocient des autres activités en ressortant dans un groupe à part.

Les moyennes conditionnelles des classes selon les variables initiales permettent d'interpréter les regroupements :



Principales caractéristiques	Classe INC	Classe DDE
Petite taille, sans moyens préventions	1	3
Offre Qualité pro (commerces de la rue : alimentaires, habillement, etc.)	2	4
Activités « sensibles », classe commreciale élevée, avec moyens de prévention	3	1
Présence de Sprinkler et RIA	4	2
Palaces avec Restaurant	5	5

Figure 3-27 – Interprétation des classes techniques d'activité

Les classes peuvent s'interpréter de la façon suivante :

- les activités de petite taille et sans moyens de préventions : classe 1 en incendie et classe 3 en dégât des eaux ;
- les activités souscrites avec l'offre Qualité pro (principalement les commerces alimentaires, habillement et autres commerces de la rue) : classe 2 en incendie et classe 4 en dégât des eaux;
- les activités « sensibles » dont la classe commerciale est plus élevée. Il s'agit principalement des hôtels et des établissements d'hébergement : classe 3 en incendie et classe 1 en dégât des eaux ;
- les activités avec présence de Sprinkler et RIA, le plus souvent présentes dans les centres commerciaux : classe 4 en incendie et classe 2 en dégât des eaux ;
- les palaces avec restaurants : activité qui se distingue des autres de par ses caractéristiques en terme de chiffre d'affaires, de surface et de capital. Cette activité est représentée par la classe 5 en incendie et en dégât des eaux.

Plus le regroupement des activités est « naturel », plus il est stable. La stabilité de la classification est évaluée grâce à l'indice de Rand.

➢ La stabilité des regroupements

Le nombre de classes a été déterminé sur un échantillon représentant 80% des données 2016 à 2019. Pour évaluer la stabilité de la classification, la partition retenue est comparée avec celle lancée sur l'échantillon des 20% restants. Par ailleurs, dans le but de valider la stabilité dans le temps de la classification mais également de voir si l'année 2020 touchée par la Covid se différencie, la classification est comparée à une classification sur l'année 2020. L'indice de Rand ajusté est utilisé comme mesure de comparaison des partitions : un indice proche de 1 indique une classification stable.

Le tableau ci-dessous donne les valeurs obtenues pour l'indice de Rand ajusté sur les deux comparaisons :

	INDICE DE RAND	
GARANTIE	Base 2016-2019 Echantillon 80 % / Échantillon 20 %	Base 2016-2019/ Base 2020
INCENDIE	0,71	0,67
DEGAT DES EAUX	0,76	0,73

Tableau 3.6 - Indice de Rand sur la classification des activités

Les valeurs de l'indice de Rand ajusté valide la stabilité de la classification. Cette classification constitue la nouvelle variable « Classe ». Pour s'assurer que cette variable latente, créée à partir d'autres variables, apporte de l'information complémentaire et n'est pas redondante avec les données dont elle est issue, sa corrélation avec les données qui ont servi à la créer est testée. Les coefficients de corrélation de la variable « Classe » avec les autres variables sont représentés ci-dessous :

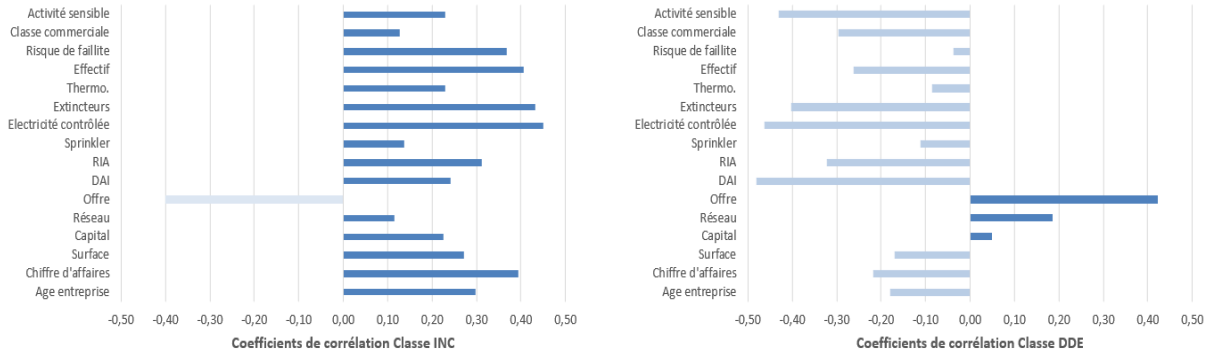


Figure 3-28 – Coefficients de corrélation des variables Classe

Les coefficients de corrélation ne font apparaitre aucune corrélation forte de la variable Classe avec les variables dont elle est issue. Cette nouvelle variable est alors intégrée dans la base de données qui alimente les modèles, en rapprochant le code activité de la base de modélisation à celui de la classification.

Pour chacun des deux garanties étudiées, les classes sont identiques pour les modèles de fréquence et de coût moyen. Le rôle du GLM sera d'estimer des niveaux tarifaires différents entre les classes du modèles fréquence et celles du modèle de coût moyen.

3.2.4. La constitution de la base finale

➢ La jointure des données

Les données caractéristiques du risque découpées en périodes d'exposition sont le socle de la base. A ces données sont rattachés les sinistres dont les montants sont distingués entre la partie écrêtée (attritionnel) et la sur-crête (sinistres atypiques). Puis, les données clients sont rapprochées et enfin les indicateurs géographiques sont joints avec le code de la commune du risque (code Insee).

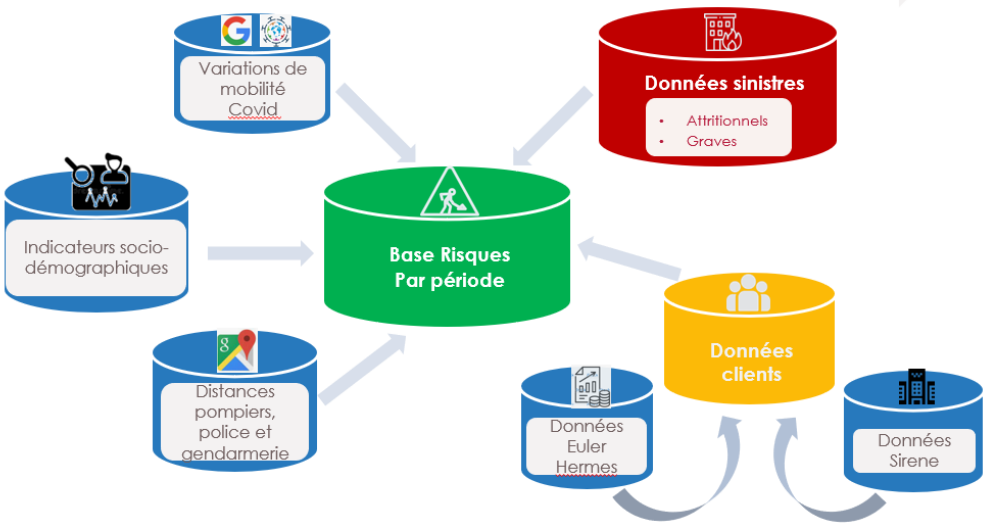


Figure 3-29 – Schéma de la construction de la base de données

➢ Le découpage de la base : modélisation/validation/test

Par définition, un modèle est complètement optimisé pour les données à l'aide desquelles il a été créé. L'erreur est minimisée sur ces données, alors qu'elle est plus élevée sur des données que le modèle n'a pas encore utilisées. Afin d'évaluer la qualité d'un modèle, la meilleure approche est de séparer la base de données en échantillons distincts :

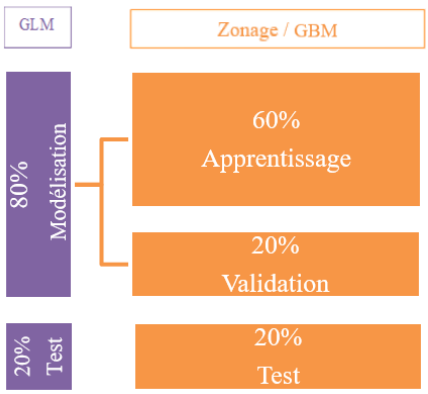


Figure 3-30 - Echantillonnage de la base de données

L'échantillon de « modélisation » utilisé pour déterminer les variables prédictives du GLM est le plus grand (80 % de la base). Cet échantillon sert également à estimer le niveau de lissage du zonier. Il est séparé en un échantillon d'apprentissage (60 % de la base) pour évaluer les paramètres optimaux du lissage et du GBM, et en un échantillon de validation (20 % de la base). La performance du modèle GLM final qui intègre le zonier est testée sur de nouvelles données encore jamais exploitées : l'échantillon de test du GLM (20 % de la base). De même, la performance du modèle GBM est évaluée sur les 20 % de données n'ayant pas servies pour l'apprentissage et la validation des hyperparamètres.

Chaque échantillon doit être représentatif de la base globale, notamment au niveau de la distribution des sinistres, en nombre et en coût. La division de la base en échantillons ne peut se faire de manière déterministe, sinon le modèle serait biaisé. A cause de la présence de sinistres atypiques, un échantillonnage aléatoire pourrait produire une répartition de la sinistralité différente dans chacun des échantillons, et donner deux échantillons très différents : le coût moyen des sinistres pourrait être hétérogène entre les deux échantillons et la base totale.

Au lieu de réaliser un seul jeu d'échantillons, un algorithme de bootstrap est appliqué afin de créer plusieurs jeux d'échantillons à partir de la base globale. Pour chacune d'eux l'écart la somme des carrés des écarts (SSE) des KPIs : fréquence (« freq ») et coût moyen (« sev » pour sévérité) entre la base globale et de la base de test sont calculés.

Pour chaque garantie, sur chacun des N jeux d'échantillons créés, le **SSE** s'écrit :

$$((Freq_{60} - Freq_{Tot})/Freq_{Tot})^2 + ((Freq_{20} - Freq_{Tot})/Freq_{Tot})^2 + ((Freq_{20} - Freq_{Tot})/Freq_{Tot})^2$$
$$+((Sev_{60} - Sev_{Tot})/Sev_{Tot})^2 + ((Sev_{20} - Sev_{Tot})/Sev_{Tot})^2 + ((Sev_{20} - Sev_{Tot})/Sev_{Tot})^2.$$

L'objectif de la méthode Boostrap est d'obtenir le meilleure seed qui minimise le SSE, sachant que le seed détermine la séquence de nombres aléatoires générés. Le jeu d'échantillon avec la distance minimale est retenu.

Une fois la base de données construite en rapprochant les profils de risques aux sinistres, les différentes méthodes qui permettent de modéliser la prime pure peuvent être appliquées. Les modèles sont construits sur un échantillon d'apprentissage, un échantillon de validation est utilisé pour les modèles avec des paramètres à estimer et les performances des modèles sont évaluées en généralisant sur l'échantillon de test.

4. La modélisation

Ce chapitre présente la mise en œuvre des méthodes utilisées pour modéliser la prime pure sur chacune des deux garanties analysées : l'incendie et le dégât des eaux.

La première section s'intéresse à la modélisation des sinistres attritionnels avec l'application de modèles GLMs de fréquence et de coût moyen. Les variables retenues ainsi que les mesures de performance et de validation des modèles y sont exposées. La deuxième section est consacrée à la modélisation des sinistres atypiques spécifiques à la garantie incendie. Les résultats obtenus par la méthode classique GLM sont comparés à une méthode de M*achine Learning* : *LightGBM*. Dans la troisième et dernière section, la prime pure estimée à partir de la modélisation des sinistres attritionnels et atypiques est analysée. Pour compléter la modélisation du risque, une règle d'identification de profil de clients plus enclins aux sinistres atypiques est déterminée à l'aide de méthodes de *Machine Learning*. Et des évolutions du tarif actuel sont proposées au niveau des activités et des zones géographiques après une mise en regard du tarif commercial avec la prime pure estimée.

4.1. La modélisation des attritionnels

La modélisation des sinistres attritionnels est réalisée à l'aide de GLMs classiquement utilisés en tarification. Cet algorithme possède l'avantage d'être facilement implémenté d'un point de vue informatique et explicable. Le logiciel Emblem de chez Towers Watson est utilisé pour la modélisation.

Les différentes périodes de restrictions mises en place par le gouvernement en 2020 ont eu un impact sur les habitudes des français et sur l'activité des commerces. Afin d'évaluer les conséquences de chaque mesure : confinement, couvre-feu sur la sinistralité des commerces, l'**interaction** de chaque variable avec la variable **Année x Période_Covid** est testée. Les tendances sont analysées sur chaque critère afin d'identifier celles qui ne sont pas identiques par période. Si pour une variable, les tendances diffèrent sur certaines périodes de restriction, l'interaction avec Année x Période_Covid est rajoutée en regroupant les périodes ayant des tendances identiques.

La mise en place d'un zonier requiert des premiers modèles intégrant les variables géographiques qui seront par la suite remplacées par la variable zonier.

4.1.1. Les modèles pré-zonier

➢ Les modèles de fréquence

Le modèle de fréquence est supposé Poissonnien. Pour valider l'hypothèse d'une loi de Poisson, l'espérance et la variance de la fréquence observée sont calculées sur l'échantillon de modélisation, à partir des données ci-dessous :

Garantie	Nombre de sinistres	Exposition	E[X]	V[X]
Incendie	3 310	285 370	1,16 %	1,15 %
Dégât des eaux	10 318	271 112	3,81 %	3,66 %

Tableau 4-1 - Fréquence modélisée

L'espérance et la variance sont très proches, ce qui conforte le choix d'une loi de Poisson pour le modèle de fréquence.

• Les variables explicatives

La base de données utilisée pour la modélisation comporte près de **100 variables**. Une pré-sélection de **50 variables** classées par ordre d'importance est effectuée par la méthode mRMR (redondance minimale et pertinence maximale). La pertinence est choisie comme la valeur maximale de l'information mutuelle entre les variables prises individuellement et la variable à expliquer. La sélection des meilleures variables, en utilisant

uniquement la pertinence ne retourne pas nécessairement la solution optimale car les variables peuvent être redondantes entre elles. La combinaison de ces deux propriétés permet de sélectionner les variables qui améliorent la prédiction.

La connaissance du produit et des garanties permet d'apporter un regard critique sur cette première sélection de variables. Les variables sont ajoutées une à une de façon manuelle en optimisant les critères d'ajustement et de prédiction.

Les variables retenues dans les modèles de fréquence

Le gain d'AIC (Critère d'Information d'Akaike) est évalué pour chaque variable ajoutée dans le modèle. La baisse d'AIC par variables explicatives ajoutée dans les modèles de fréquence des garanties incendie et dégât des eaux est présenté sur les graphiques ci-dessous :

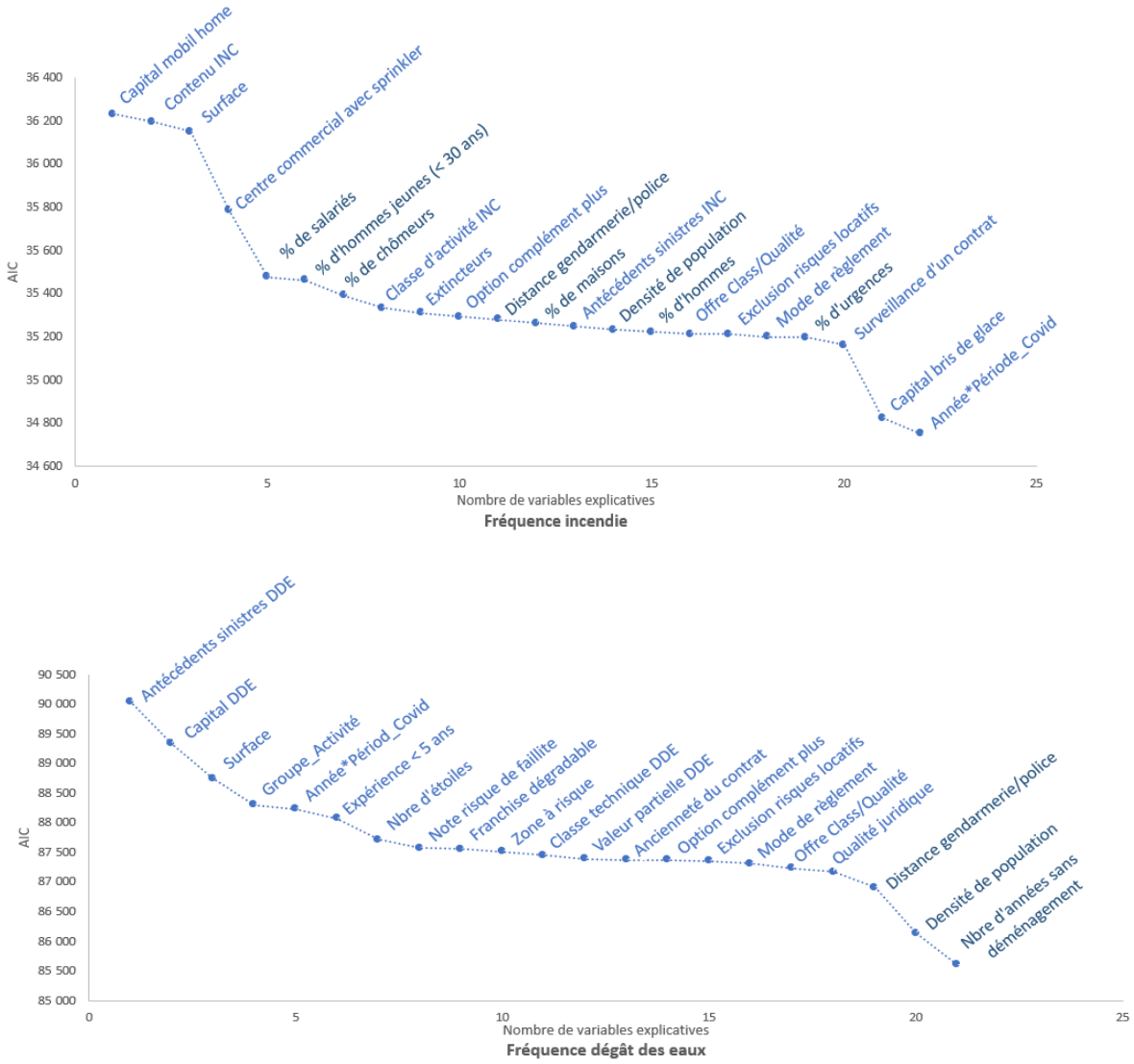


Figure 4-2 – Evolution de l'AIC des modèles de fréquence par variables explicatives

Les variables de taille de risque : capital, superficie se retrouvent dans les modèles de fréquence des deux garanties : incendie et dégât des eaux. D'autres variables sont plus caractéristiques au risque couvert par la garantie. Les moyens de prévention : extincteurs, sprinkler font parties des variables explicatives de la fréquence incendie. Les variables géographiques (en bleu plus foncé sur les graphiques) apportent un gain d'AIC plus important dans le modèle de la garantie dégât des eaux.

Afin de capter les effets des différentes périodes de restrictions sanitaires, l'interaction de chacune de ces variables avec la variable Année x Période_Covid est testée. Les interactions suivantes sont retenues :

- Modèle fréquence incendie : Surface x Année x Période_Covid.
- Modèle fréquence dégât des eaux : Franchise dégradable x Année x Période_Covid et Option complément plus x Année x Période_Covid (cf Chap. 1 § 1.2.2.2 pour les notions de franchise dégradable et de complément plus).

Les graphiques de tendances sur les principales variables explicatives sont présentés en annexe. Seule l'analyse des quelques variables prédictives de la fréquence est détaillée ci-dessous.

➢ Les variables caractéristiques du risque

Année x Période_Covid

La variable Année x Période_Covid est toujours intégrée dans les modèles, elle permet de capter une tendance dans l'évolution de la sinistralité non liée à un phénomène structurel, et de distinguer les différents effets des périodes de restrictions.

Pour simplifier le modèle et réduire le nombre de modalités, les périodes avec des niveaux de prédiction proches ou avec de faibles volumes sont regroupés :

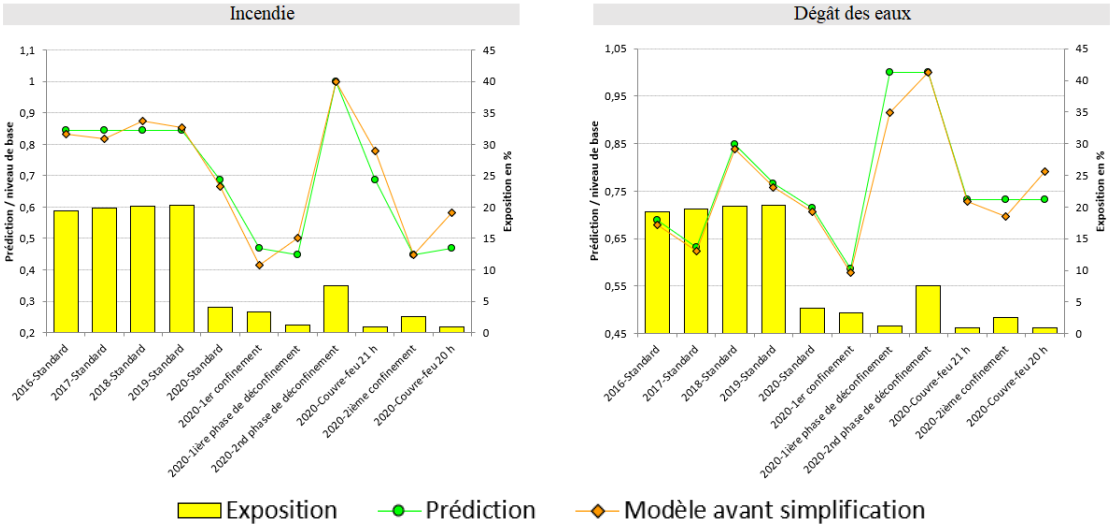


Figure 4-3 - Fréquence par périodes

Les tendances de fréquence par périodes sont différentes entre l'incendie et le dégât des eaux. En incendie, les années 2016 à 2019 ont des fréquences très proches et peuvent être regroupées, ce qui n'est pas le cas en dégât des eaux où les tendances sur ces quatre années se distinguent les unes des autres. Par ailleurs, en dégât des eaux, les deux phases de déconfinement à la suite du premier confinement sont les périodes dont la fréquence prédite est la plus élevée.

Les capitaux assurés

Le capital assuré est une des principales variables explicatives des modèles. Trois formats de discrétisation sont appliqués sur cette variable : Gamma, gaussien et uniforme. Les graphiques ci-dessous représentent la fréquence DDE observée sur chacune des distributions, et le format retenu dans le modèle :

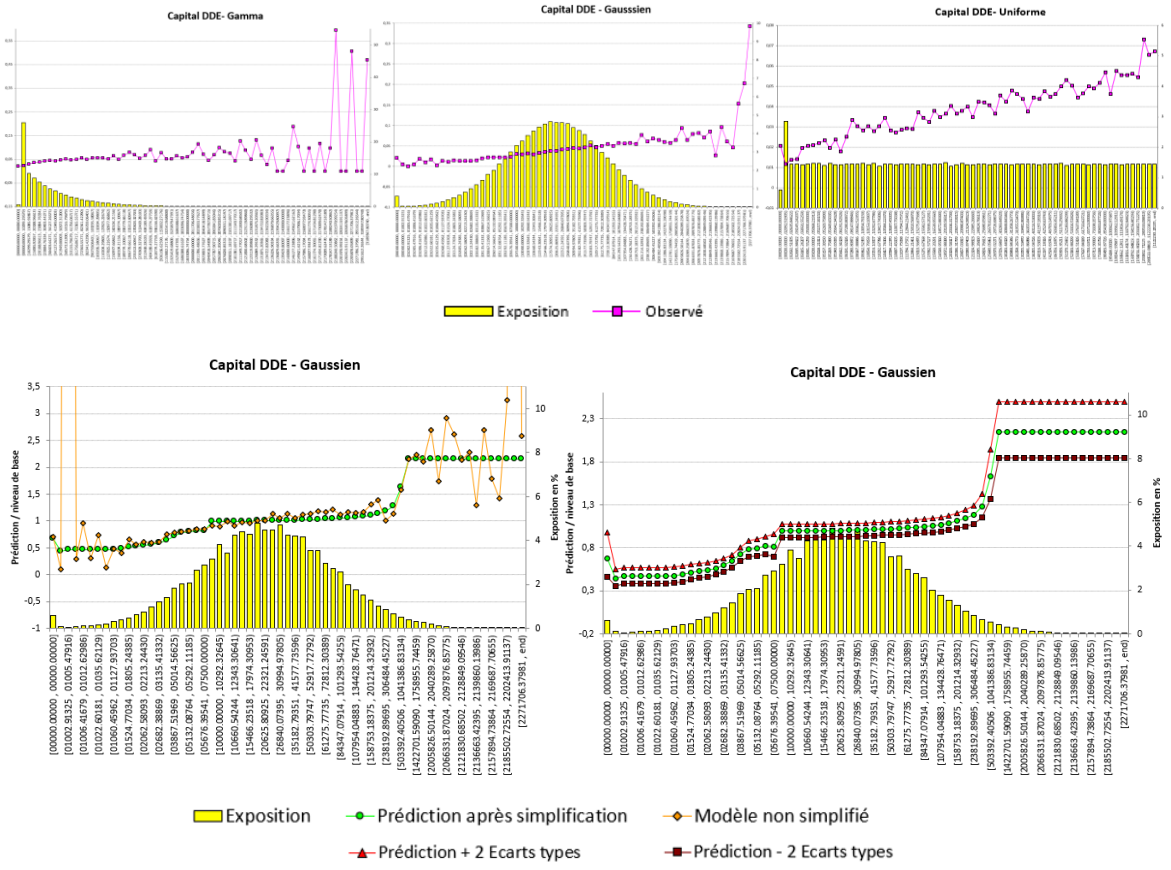


Figure 4-4 - Fréquence DDE observée et estimée par niveau de capital

Le modèle de fréquence dégât des eaux retient la distribution gaussienne du capital. Cette variable est modélisée par un lissage à l'aide de deux polynômes et d'un regroupement des dernières tranches dont les volumes sont trop faibles pour avoir un bon niveau prédictif (écart type plus important). La tendance monotone et le pouvoir discriminant du capital sont visibles : la fréquence d'un capital de 25 k€ est 1,6 fois plus élevée que celle d'un capital de 3,5 k€. L'augmentation de la fréquence avec le capital assuré se constate également en incendie. En effet, plus la quantité de biens d'un professionnel augmente plus ces biens peuvent être exposés à un dégât des eaux ou à un incendie.

La surface

La surface est l'une des principales variables explicatives de la fréquence. L'intersection de la surface avec les périodes : Année x Période_Covid, fait ressortir une fréquence incendie qui se distingue selon les périodes, alors que la tendance de la fréquence dégât des eaux est plutôt identique par périodes :

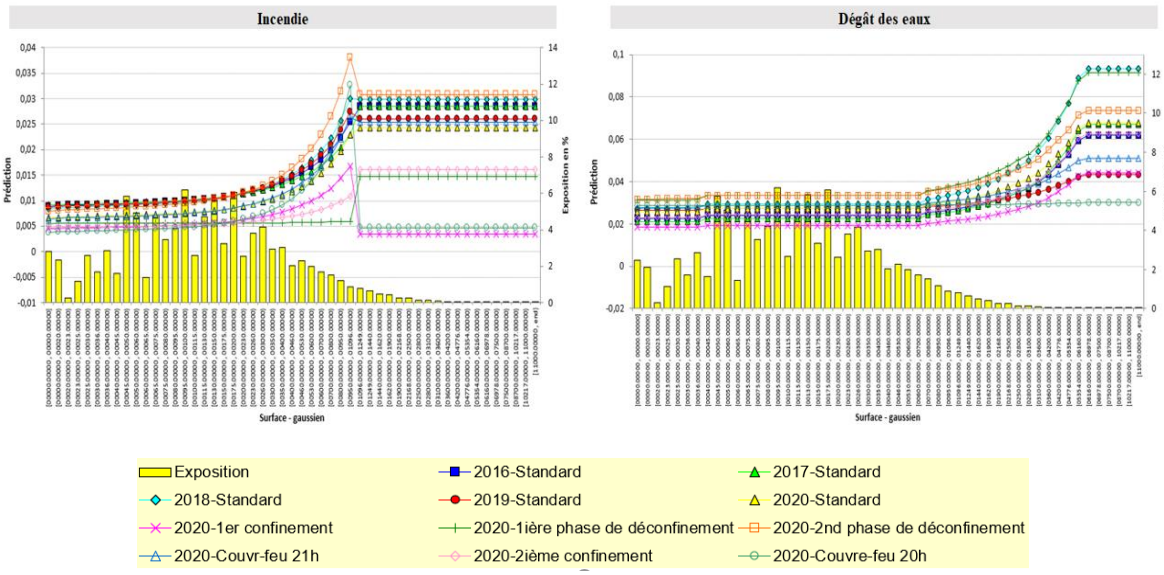


Figure 4-5 - Fréquence par surface x Année x Période_Covid

Sur les grandes surfaces, la fréquence incendie est plus basse lors du premier confinement et du couvre-feu à 20 h, alors qu'elle a tendance à augmenter lors des autres périodes. Les périodes sont regroupées en deux : le premier confinement avec le couvre-feu à 20 h, et l'ensemble des autres périodes. L'interaction de la surface avec ces deux groupes de périodes est retenue.

La fréquence augmente avec la surface : un local de 700 m² a deux fois plus de sinistre incendie qu'un local de 65 m². L'évolution de la fréquence avec la surface est assez intuitive : plus la surface est grande, plus elle possède d'éléments susceptibles de provoquer un incendie. Ce qui est également vrai en dégât des eaux : une plus grande surface implique plus de canalisation et de point d'eau. Toutes les variables de taille susceptibles d'augmenter le nombre de points d'eau engendrent une augmentation de la fréquence.

➢ Les variables géographiques

Toutes les variables géographiques ont été créées avec deux formats : une discrétisation par une distribution gaussienne et une autre par une distribution uniforme. La distribution gaussienne plutôt qu'uniforme est généralement retenue, ce qui n'est pas surprenant étant donné que les données sous-jacentes sont généralement normales et que la distribution normale facilite l'identification des tendances aux extrêmes, où se situent la plupart des effets.

La **densité** de population est le critère géographique le plus prédictif de la fréquence, sept autres critères expliquent le risque géographique en incendie et trois en dégât des eaux. Le graphique ci-dessous représente la prédiction de la fréquence dégât des eaux par densité de population :

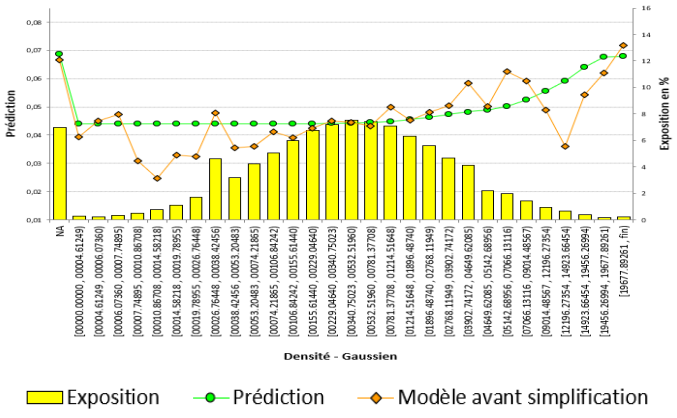


Figure 4-6 - Fréquence dégât des eaux par densité de population

La tendance **monotone** et le **pouvoir discriminant** de la fréquence dégât des eaux par densité a été lissée par un polynôme du premier degré avec un regroupement des trois dernières tranches qui n'ont pas assez d'exposition pour une prédiction stable. D'après la figure ci-dessus, la fréquence dégât des eaux augmente significativement à partir de 5 000 habitants par km², c'est-à-dire dans les villes les plus peuplées.

Les variables géographiques retenues dans les modèles serviront par la suite à créer les zoniers par garantie.

- La validation des modèles

La significativité des variables retenues dans les modèles est contrôlée par l'évolution de la déviance et du critère **AICc**. L'apport des variables géographiques est présenté dans le tableau ci-dessous :

	DEVIANCE			AICc			χ^2
GARANTIE	Modèle sans socio-démo	Modèle final avec socio-démo	Delta	Modèle sans socio-démo	Modèle final avec socio-démo	Delta	
INCENDIE	37 286,54	37 222,68	-63,86	35 313,13	35 278,47	-34,66	0,0%
DEGAT DES EAUX	96 060,73	95 294,74	-765,99	87 345,13	87 254,86	-90,27	0,0%

Tableau 4.1- Déviance, AICc et χ^2 des modèles de fréquence avant zonier

La baisse de la déviance et de l'AICc après l'ajout des variables géographiques, et la statistique du χ^2 proche de 0 % qui compare les modèles avec et sans variables géographiques valident l'importance de ces variables dans les modèles.

Le pouvoir prédictif des modèles est évalué en appliquant les coefficients tarifaires estimés sur l'échantillon de modélisation (80 % de la base), sur un nouvel échantillon, l'échantillon de test (20 % de la base), qui n'a pas servi à la modélisation. Le graphique résultant affiche les prédictions regroupées en valeurs absolues ascendantes. Au sein de chaque groupe, la valeur moyenne pondérée des données observées et des prédictions sur l'échantillon de test sont calculées et tracées :

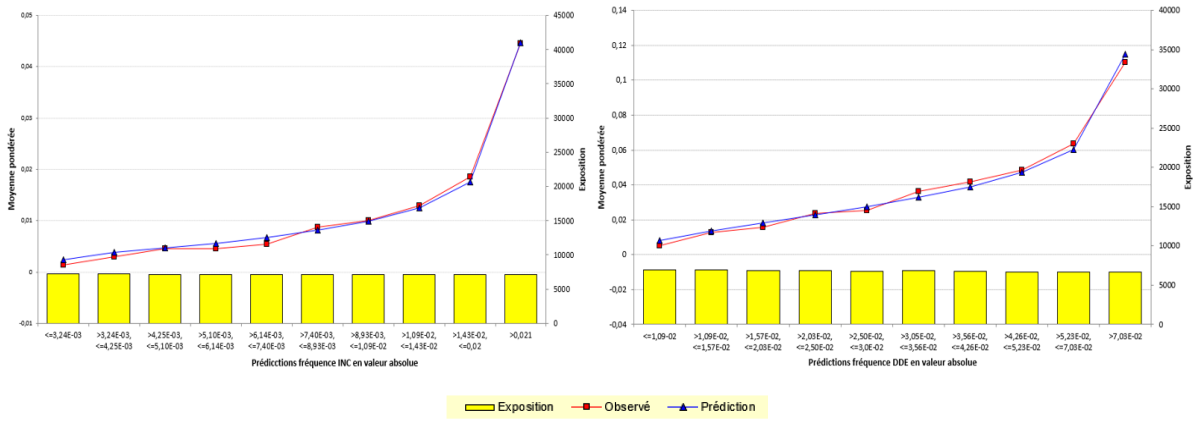


Figure 4-7 - Validation des modèles de fréquence sur l'échantillon de test

D'après la figure ci-dessus, l'observé de la base de test coïncide raisonnablement bien avec la prédiction. Les modèles de fréquence incendie et dégât des eaux sont prédictifs.

L'examen des résidus est une étape cruciale de l'analyse statistique pour identifier les écarts entre les modèles et les données, et évaluer la qualité d'ajustement globale des modèles. L'analyse résiduelle sert à diagnostiquer la qualité globale de l'ajustement et l'adéquation des modèles. Les « crunched » résidus sont tracés en fonction des valeurs prédites :

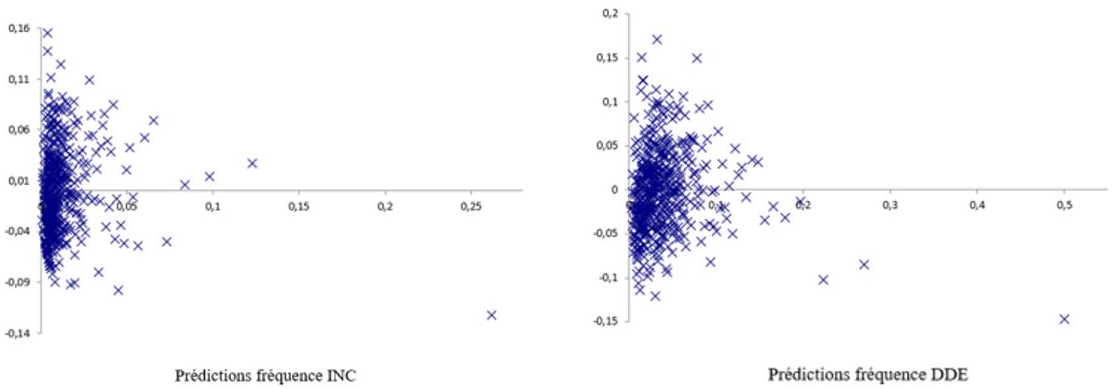


Figure 4-8 - Résidus « crunched » des modèles de fréquence incendie et dégât des eaux

L'absence de tendance significative et la structure des points centrés autour de l'axe des abscisses indiquent une bonne adéquation des modèles.

➢ Les modèles de coût moyen

Le coût moyen correspond au **coût écrêté** des sinistres sur le nombre de sinistres, après **retraitement des sinistres au forfait d'ouverture**. Le tableau ci-dessous présente le coût moyen observé sur l'échantillon modélisé :

Garantie	Nombre de sinistres	Coût écrêté	Coût moyen
Incendie	3 062	22 385 179	7 311
Dégât des eaux	11 804	37 090 874	3 144

Tableau 4.2- Coût moyen modélisé

Les lois usuelles pour modéliser le coût sont la loi log-normale ou la loi Gamma. La loi Gamma, habituellement utilisée car équivalente à la loi de Poisson en continue, est retenue. Pour valider ce choix, la distribution des coûts est comparée à celle des lois log-normale et Gamma.

Sous Python, une modélisation automatique de la distribution des coûts observés est réalisée sur chacune des deux garanties. La méthode « fit » de la librairie « scipy » appliquée sur le jeu de données permet d'identifier les paramètres optimaux par maximum de vraisemblance pour les deux lois testées. L'adéquation entre la distribution des coûts observés et les loi log-normal et Gamma, sur la garantie incendie, est représentée sur le graphique ci-dessous :

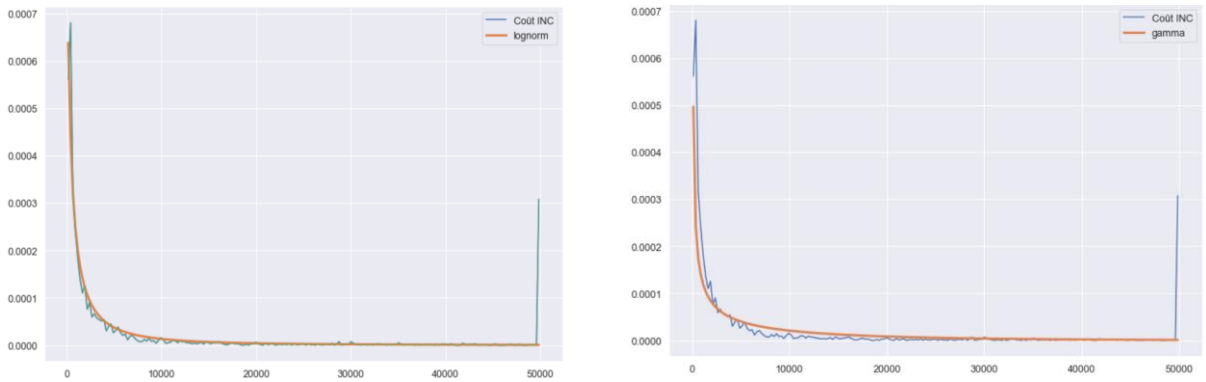


Figure 4-9 - Distribution des coûts incendie écrêtés, log-normale (à gauche) et Gamma (à droite)

La loi qui ajuste au mieux les données est celle dont la densité est la plus proche de l'observé. La distance entre les deux courbes est mesurée en calculant la somme de écarts au carrés (SSE : « sum of squared errors »). Les résultats des SSE sont donnés dans le tableau ci-dessous :

Garantie	SSE Log-normal / Observé	SSE Gamma / Observé
Incendie	1,65E-07	3,36E-07
Dégât des eaux	4,78E-08	6,11E-07

Tableau 4.3 - SSE entre l'observé et la courbe des lois log-normale et Gamma

Que ce soit pour la loi log-normale ou la loi Gamma, les écarts avec l'observé sont très faibles. La distribution log-normal est un peu plus proche de l'observé, cependant par habitude et parce que la loi Gamma est l'équivalent en continue de la loi de Poisson, la loi Gamma est retenue pour les modèles des coûts moyens.

- Les variables explicatives

Selon une démarche similaire aux modèles de fréquence, les coûts moyens des sinistres incendie et dégât des eaux sont estimés avec les critères explicatifs suivants :

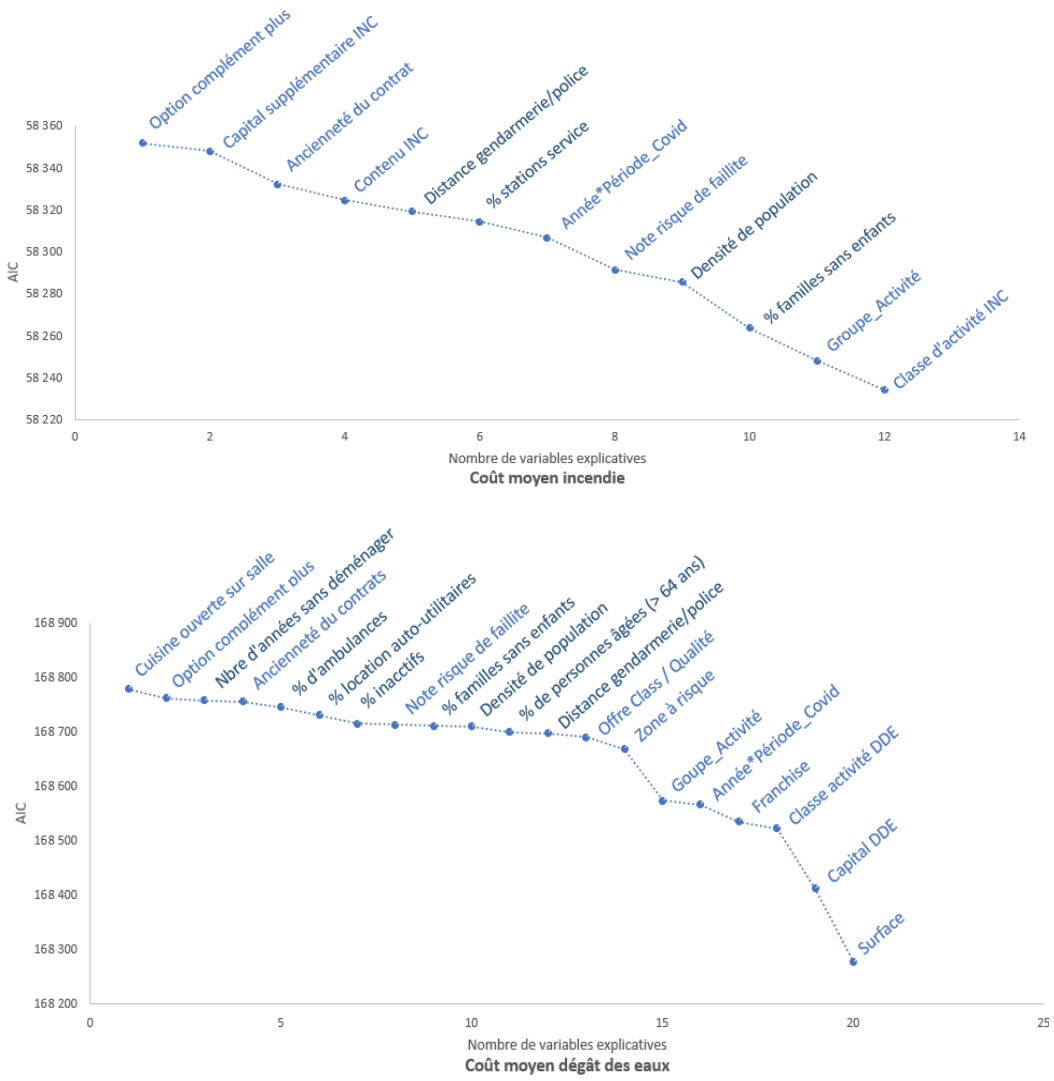


Figure 4-10 – Evolution de l'AIC des modèles de coûts moyens par variables explicatives

Le coût moyen incendie est plus difficile à prédire, le nombre de variables explicatives est moins important par rapport au modèle dégât des eaux. Les critères de taille du risque présents dans les modèles de fréquence se retrouvent dans les modèles du coût : les capitaux, la surface avec une interprétation similaire.

Après analyse de l'interaction de ces variables avec la variable Année x Période_Covid, aucune interaction n'est retenue pour le modèle coût moyen incendie, et le croisement des variables Cuisine ouverte sur salle, Offre class/qualité et Zone à risque avec Année x Période_Covid est sélectionné dans le modèle dégât des eaux.

Afin de visualiser l'effet des périodes de restrictions sur le coût moyen, les prédictions par périodes (Année x Période_Covid) sont représentées ci-dessous :

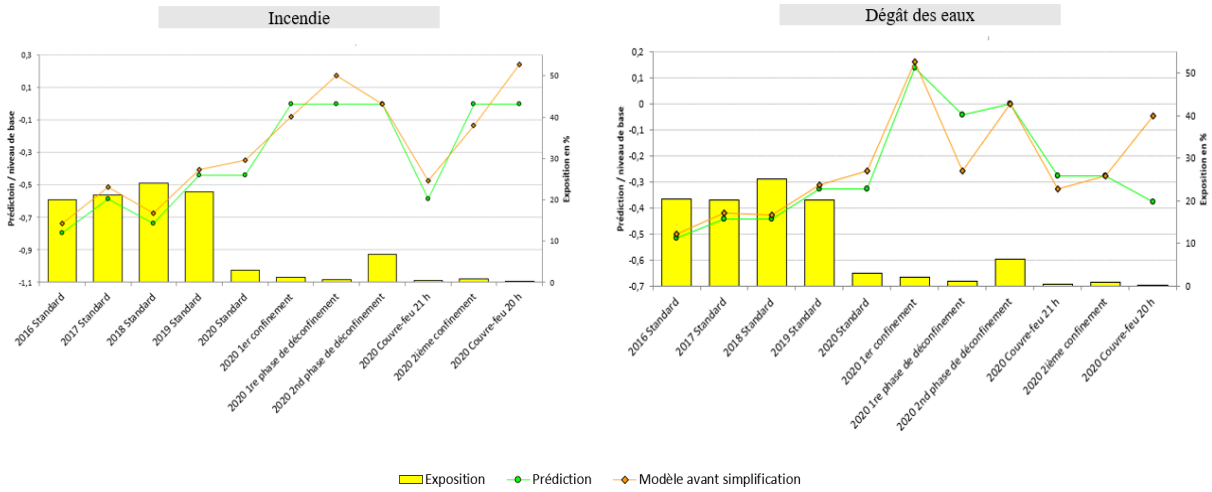


Figure 4-11 – Coût moyen par périodes

Les périodes avec trop peu d'exposition et des tendances similaires sont regroupées. Les périodes du premier confinement et de déconfinement affichent des coûts moyens plus élevés qu'en période « standard ».

- La validation des modèles

L'ajout de chaque variable dans un modèle est contrôlé par son impact sur l'**AICc** et la déviance du modèle. L'apport des variables géographiques dans les modèles coût moyen est présentée dans le tableau ci-dessous :

	DEVIANCE			AICc			χ^2
GARANTIE	Modèle sans socio-démo	Modèle final avec socio-démo	Delta	Modèle sans socio-démo	Modèle final avec socio-démo	Delta	
INCENDIE	8 487,92	8 295,093	-192,83	58 435,68	58 351,06	-84,62	0,0%
DEGAT DES EAUX	13 440,84	13 186,87	-253,96	168 606,4	168 803,0	-196,67	0,0%

Tableau 4.4 - Déviance, AICc et χ^2 des modèles coût moyen avant zonier

La significativité des variables géographiques dans les modèles de coût moyen est validée par la baisse de la déviance et de l'AICc du modèle après leur intégration.

Le pouvoir prédictif des modèles est validé sur la base de test. Les graphiques ci-dessous affichent les coûts moyens prédits (en bleu) par rapport aux coûts moyens observé sur 20 % de la base :

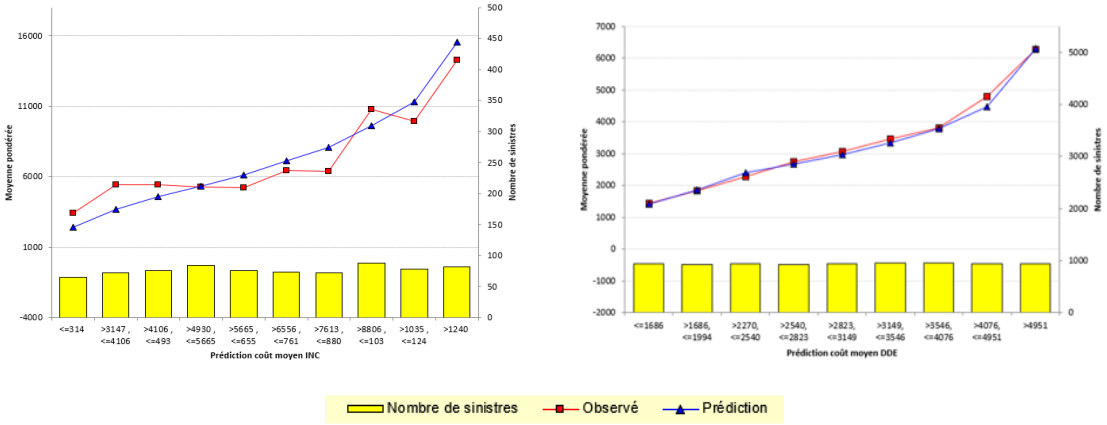


Figure 4-12 - Validation des modèles coût moyen sur l'échantillon de test

En incendie, des écarts apparaissent entre la courbe des points observés et celle des prédictions (graphique de gauche). Cependant, ce niveau d'écart reste tolérable par rapport au faible nombre de sinistres sur cette garantie. En revanche, les deux courbes sont très proches en dégât des eaux (graphique de droite) et valident le pouvoir prédictif du modèle.

L'analyse graphique des résidus déviance standardisés représentés en fonction des valeurs prédites est un autre moyen de valider les modèles.

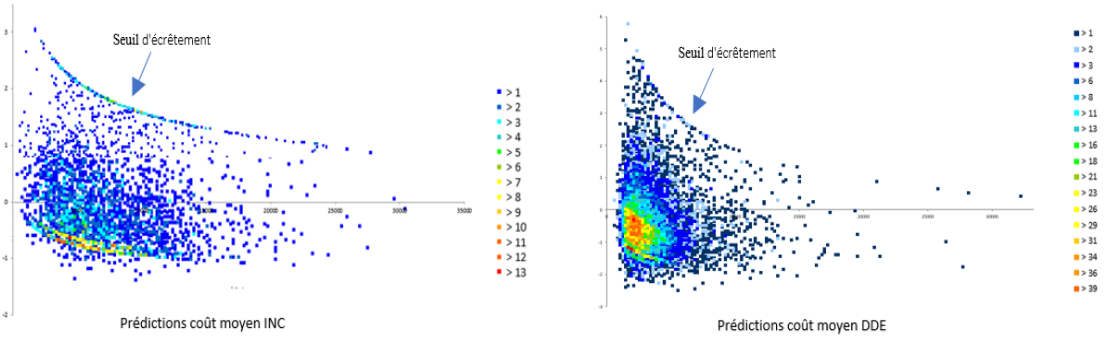


Figure 4-13 - Résidus déviance standardisés des modèles coût moyen incendie et dégât des eaux

Les résidus sont assez bien répartis autour de l'axe des abscisses, sans présence de points aberrants. La tendance observée est liée à l'écrêtement du coût des sinistres.

L'information présente dans les critères géographiques sélectionnés dans les modèles est synthétisée au sein d'une seule variable : le zonier.

4.1.2. La construction des zoniers

Le zonier est construit à partir des résidus des modèles validés à l'étape précédente. Le logiciel Radar de Towers Watson en lien avec l'algorithme de lissage de l'outil Classifier de Towers Watson est utilisé pour la construction des zoniers.

➢ La standardisation des indicateurs

Les variables géographiques intégrées dans les modèles déterminés à l'étape précédente sont, par définition du GLM, décorrélés des variables tarifaires non géographiques : risque, client. La construction des zoniers consiste à analyser, par codes postaux, les informations géographiques issues de ces modèles.

Pour chaque code postal, trois composantes sont disponibles :

- l'effet non-géographique ;
- l'effet géographique expliqué par les variables géographiques ;
- l' effet résiduel.

Pour calculer l'**effet résiduel de chaque code postal**, les données observées doivent être « standardisées » pour supprimer les effets non géographiques, de sorte que les effets restants soient tous liés au code postal. La standardisation est effectuée au niveau de chaque observation. Les valeurs prédites standardisées correspondantes doivent également être calculées. Elles sont définies comme les prédictions des modèles précédents (incluant les variables géographiques), avec les variables non-géographiques forcées au niveau de base. Les statistiques utilisées, à la fois pour les modèles de fréquence et les modèles de coût moyen, sont les suivantes :

- la **prédiction** : les valeurs attendues données par les modèles de risque ;
- la **prédiction standardisée** : les valeurs ajustées du modèle de risque dans lesquelles les critères non géographiques fixés à leurs niveaux de base ;
- le **facteur de standardisation** : calculé à partir des deux valeurs précédentes

$$\text{Facteur de standardisation} = \text{Estimé standardisé}/\text{Estimé}\,;$$

- l'**observé** : le nombre de sinistres pour le modèle de fréquence et le coût des sinistres pour le modèle de coût moyen ;
- l'**observé standardisé** : rapport de l'observé sur le facteur de standardisation ;
- le **résidu** : observé / prédiction. Une mesure multiplicative de la variation des valeurs observées par rapport aux valeurs prédites attendues.

La maille d'analyse retenue pour le lissage est le code postal et non le code Insee : maille des indicateurs géographiques. En effet, un premier test de lissage au niveau de la commune a permis de constater que ce niveau était trop granulaire pour que l'algorithme de lissage soit efficace.

➢ Le lissage des résidus

Chacune de ces statistiques est agrégée au niveau du code postal, cependant, les résidus pour chaque code postal peuvent donner des différences importantes entre codes voisins. L'application d'un lissage permet d'ajuster la valeur d'un code postal en tenant compte de la réponse des voisins. La technique utilisée est basée sur l'approche bayésienne, où le lissage est local et prend en compte le risque des codes postaux immédiatement voisins, c'est-à-dire adjacents. Le lissage identifie les effets systématiques des résidus. La loi de distribution sous-jacente : Poisson pour la fréquence et Gamma pour le coût moyen, ainsi que le niveau de lissage doivent être spécifiés comme paramètre du lissage. Pour sélectionner le niveau de lissage approprié, les résidus sont examinés avec différents niveaux allant de 50 à 150.

Le graphique ci-dessous un exemple de lissage des résidus du modèle de fréquence dégât des eaux :

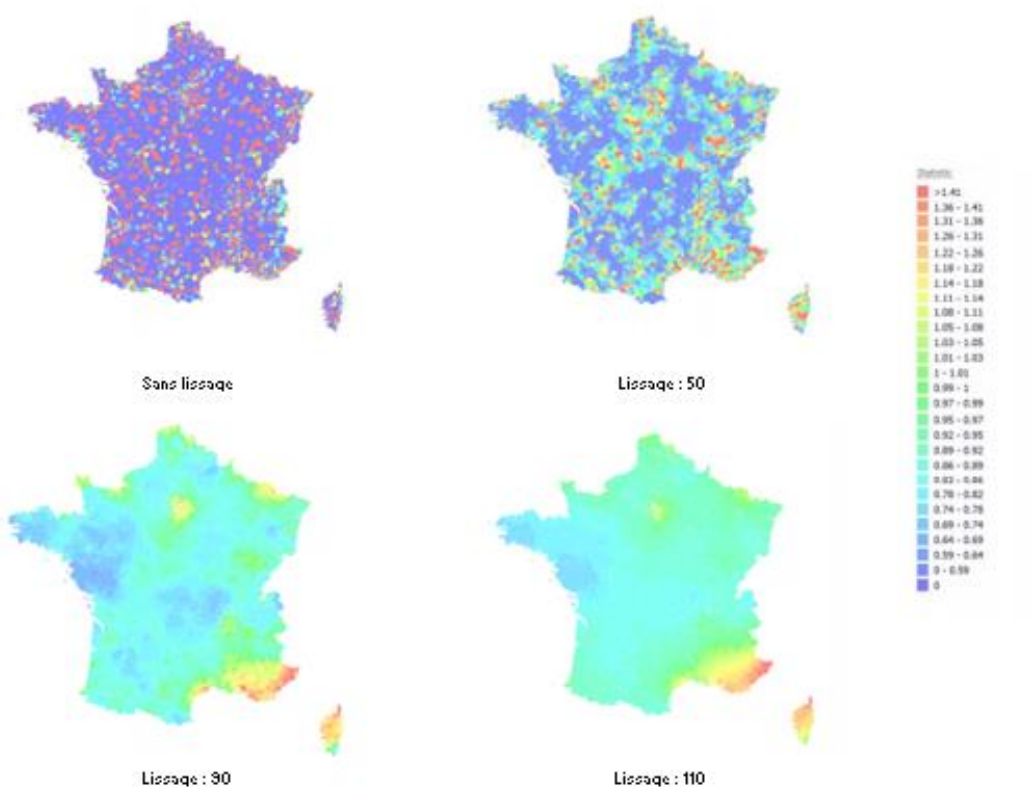


Figure 4-14 - Lissage des résidus de la fréquence dégât des eaux

Le meilleur niveau de lissage semble être 90. Au niveau 50, la carte montre encore du bruit tandis qu'à 110, la différenciation commence à se perdre.

Pour effectuer des tests de validation du niveau de lissage, la base a été scindée en deux échantillons : apprentissage et validation. La pertinence du niveau de lissage a également été analysé selon la stabilité des résidus au cours des années d'observation. Pour cela, le jeu de données d'apprentissage a été segmenté en fonction des périodes d'analyse, avec la possibilité d'accorder moins d'importance aux données plus anciennes en appliquant des poids plus faibles.

En dernier lieu, l'erreur quadratique moyenne (MSE : mean of squared errors) est calculée sur l'échantillon de validation pour chaque niveau de lissage (pour éviter tout risque de surajustement). Cette statistique mesure la qualité de l'ajustement de manière quantitative, et donne une indication du pouvoir prédictif.

Les tracés de l'erreur quadratique moyenne sur les quatre modèles étudiés sont représentés sur les graphs ci-dessous :

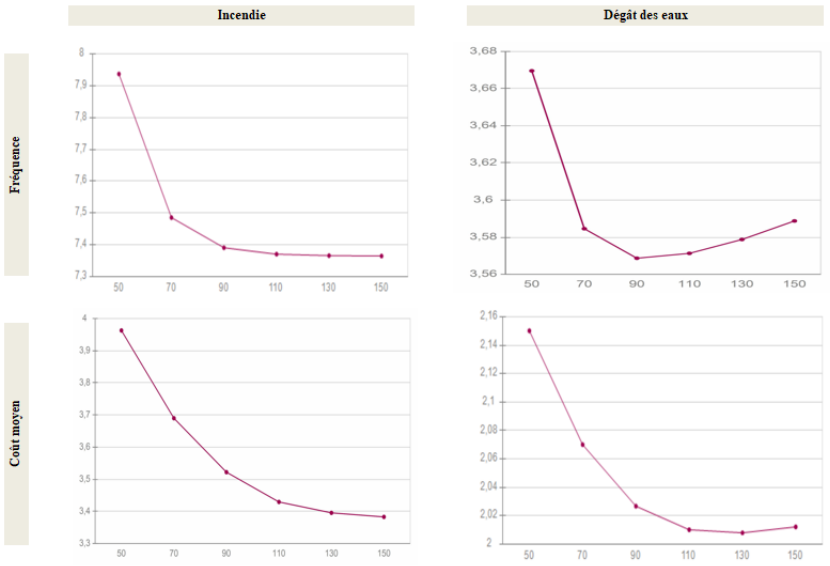


Figure 4-15 - Moyenne des écarts au carré par niveau de lissage

Dans le cas de la fréquence dégât des eaux, le niveau de lissage optimal est clairement défini à 90 où le MSE est minimal. En revanche, pour les trois autres modèles, le minimum n'est pas clairement défini. La « règle du coude » est adoptée en prenant comme optimal le niveau de lissage à partir duquel le MSE cesse de diminuer de manière significative : 90 pour la fréquence incendie et 110 pour les deux modèles du coût moyen.

➢ La validation des zoniers avant intégration dans les modèles

L'estimé standardisée combiné au résidu lissé sont regroupés selon la méthode de Ward. La précision et le pouvoir prédictif du zonier peuvent être rapidement validés avant leur intégration dans les modèles en traçant le graphique de la moyenne pondérée des valeurs estimées corrigées de la correction spatiale : SCF (c'est-à-dire la moyenne prédite) sur l'échantillon d'apprentissage et la moyenne pondérée des résultats observés (normalisés pour les effets non géographiques) sur l'échantillon de validation.

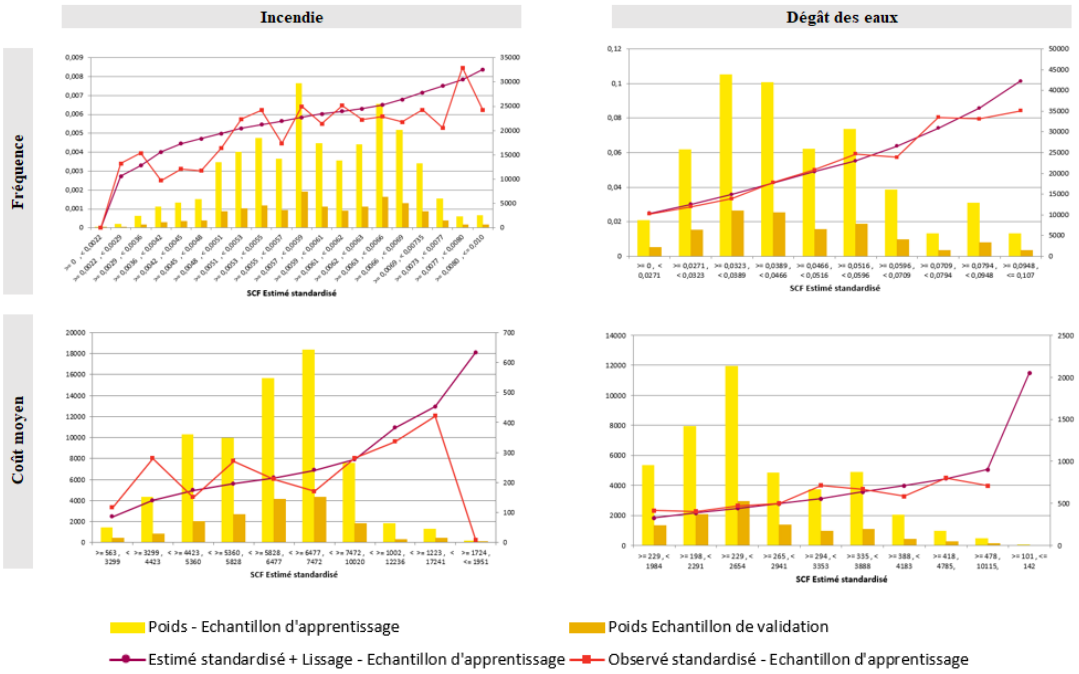


Figure 4-16 - Validation des zoniers

Sur la garantie dégât des eaux, le zonier (en violet) correspond bien à l'observé standardisé de l'échantillon de validation (en rouge). En revanche, en incendie, les écarts sont plus importants.

Afin d'analyser l'impact du lissage, l'effet seul des variables géographiques : la prédiction standardisée, va être comparé dans les modèles, à l'effet géographique global combinant la prédiction standardisée et l'effet géographique résiduel (les résidus lissés). Ces deux statistiques calculées par codes postaux sont regroupées en zones, et sont dénommées par la suite « zonier non lissé » et « zonier lissé ». Afin d'avoir un bon compromis entre granularité et robustesse compte tenu des volumes de données, un regroupement en 50 zones est implémenté avec la méthode de Ward.

L'intégration de chacun des deux zoniers : non lissé et lissé, dans les modèles GLM en remplacement des variables géographiques est testée. L'analyse des perturbations des prédictions va permettre d'évaluer lequel des deux zoniers est le plus performant pour chaque modèle.

4.1.3. L'intégration des zoniers dans les modèles

Les zoniers (non lissés et lissés) créés par garantie pour la fréquence et le coût moyen sont testés dans les modèles. Toutes les variables géographiques des modèles avant zonier sont retirées, puis chaque modèle est dupliqué en un modèle avec la variable « zonier non lissé » retenue et un autre avec la variable « zonier lissé » :

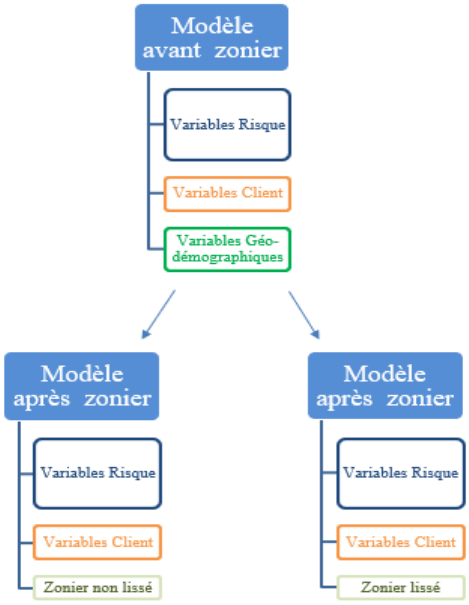


Figure 4-17 – Modèle avant/après zonier

Une série de tests graphiques et statistiques sont analysés pour comparer les deux modèles après zonier et décider quel zonier sera retenu : le zonier non lissé ou lissé.

Le pouvoir prédictif des modèles avec zonier

Comme pour toute nouvelle variable intégrée dans un modèle, la tendance des prédictions, le pouvoir discriminant et la stabilité dans le temps sont observés. Pour choisir le meilleur modèle, le pouvoir prédictif du modèle avec zonier non lissé et comparé à celui avec zonier lissé. La **comparaison** des modèles est réalisée sur l'**échantillon de test** et se fait à l'aide d'un graphique. Sur ce graphique, la différence en pourcentage entre les prédictions pour chacun des deux modèles est calculée pour chaque enregistrement. Cette différence est décomposée en tranches affichées sur l'axe des abscisses. Le poids des enregistrements de chaque tranche est représenté par un diagramme en bâton, et les moyennes pondérées des données observées et des prédictions de chacun des modèles pour chaque tranche sont représentées par des courbes.

Le graphique de comparaison des modèles de fréquence et de coût moyen pour chacune des deux garanties est illustré ci-dessous :

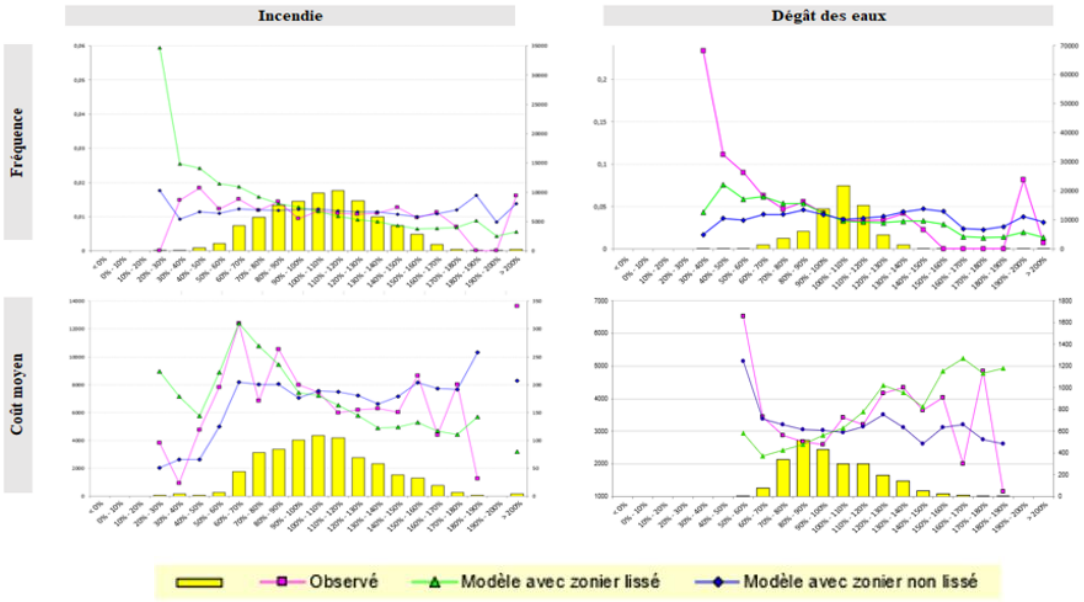


Figure 4-18 - Comparaison des modèles avec zonier lissé et non lissé

Pour chaque graphique, le modèle dont la courbe est la plus proche des données observées peut être considéré comme le modèle le plus prédictif. En observant les graphiques ci-dessous, trois cas de figure se présentent :

- le modèle avec le **zonier lissé** apparaît comme étant le plus prédictif. C'est le cas pour le **dégât des eaux** en **fréquence** et en **coût moyen** où la courbe du modèle avec zonier lissé (courbe verte) est globalement plus proche de l'observé (courbe rose) que la courbe du modèle avec zonier non lissé (courbe bleue) ;
- le modèle avec le **zonier non lissé** apparaît comme étant le plus prédictif. C'est le cas du modèle de **fréquence incendie** ;
- **aucun** des modèles ne semble satisfaisant. C'est le cas du modèle **coût moyen incendie** pour lequel ni la tendance des prédictions du modèle avec zonier lissé ni celle avec zonier non lissé ne suit raisonnablement l'observé.

Afin de valider l'observation faite sur le modèle coût moyen incendie, le pouvoir prédictif des modèles avec zonier (non lissé et lissé) est comparé à celui du modèle avant zonier :

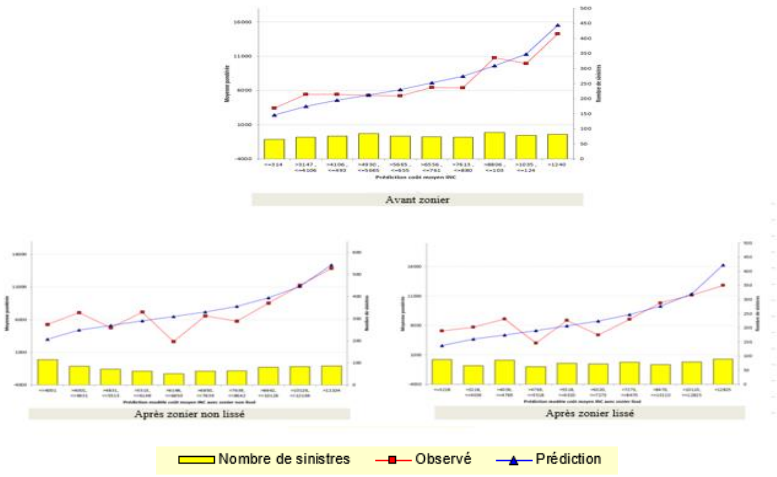


Figure 4-19 - Pouvoir prédictif des modèles coût moyen incendie avant et après zonier

Le regroupement des variables géographiques en une seule variable « zonier » entraîne une perte d'information qui impact le pouvoir prédictif du modèle coût moyen incendie. Le faible volume de sinistres sur ce dernier le rend beaucoup plus sensible aux changements que les autres modèles. Aucun effet du risque géographique n'est retenu dans le modèle coût moyen incendie.

Pour les autres modèles, l'apport du zonier en remplacement des variables externes est mesuré par le coefficient de Gini. Ce dernier est calculé avant et après zonier pour les trois modèles dans lesquels les variables externes ont été remplacées par un zonier (fréquence incendie et dégât des eaux, et coût moyen dégât des eaux) :

	Coefficient de Gini		
Modèles	Avant zonier	Après zonier	Delta
INC fréq.	0,4957	0,5055	0,0098
DDE fréq.	0,4119	0,4424	0,0305
DDE Coût moyen	0,2418	0,2795	0,0377

Tableau 4.5 - Coefficients de Gini avant et après zonier

L'augmentation du coefficient de Gini montre que le zonier améliore le pouvoir prédictif et la segmentation des modèles. Cette amélioration est plus particulièrement marquée sur le modèle coût moyen dégât des eaux dont les courbes de gain sont représentées sur le graphique ci-dessous :

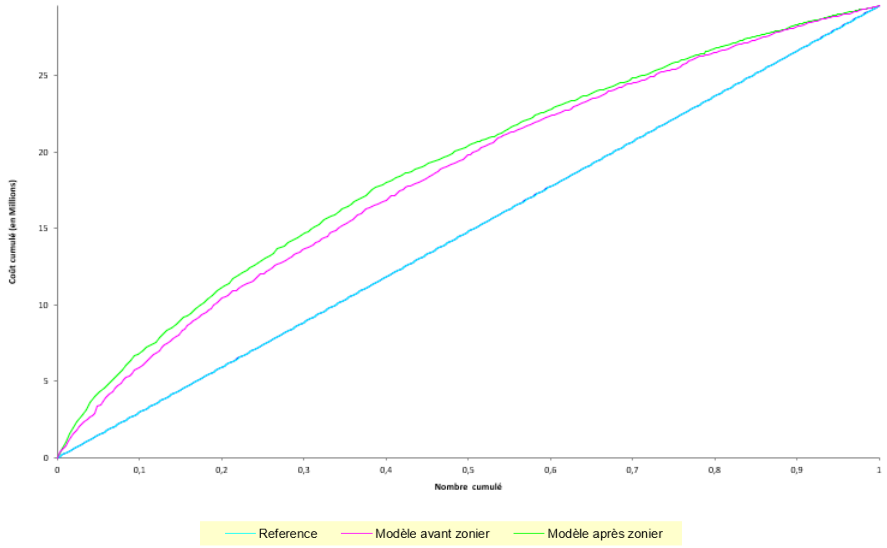


Figure 4-20 - Courbes de gain dégât des eaux coût moyen

D'après le graphique ci-dessus, la courbe de gain du modèle après zonier est toujours au-dessus de celle du modèle avant zonier. Le modèle après zonier est plus prédictif que la modèle avant zonier.

La stabilité dans le temps de la variable zonier

Pour s'assurer de la stabilité dans le temps de la nouvelle variable « zonier » introduite dans les modèles, l'interaction avec la variable « Année x Période_Covid » est analysée.

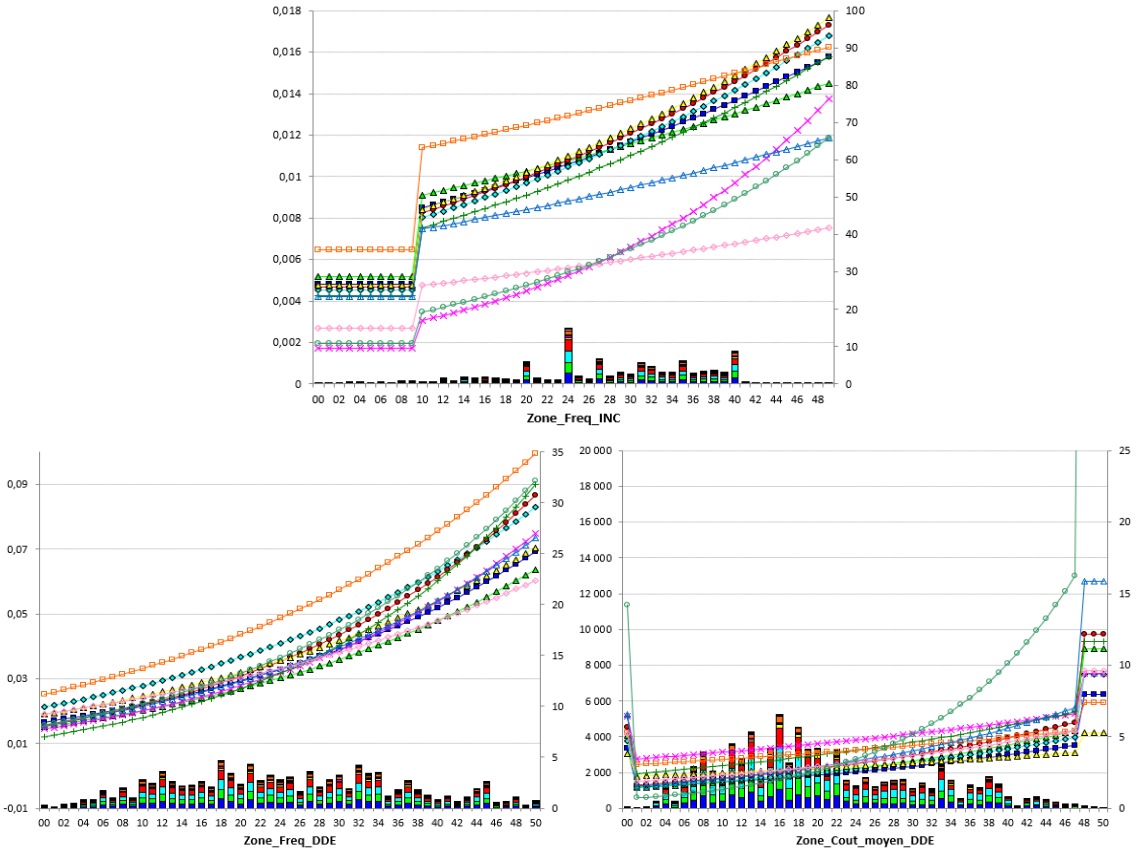


Figure 4-21 – Stabilité du zonier dans le temps

Les prédictions de la variable zonier montrent des tendances identiques quelles que soient les périodes, tant sur le modèle fréquence incendie que sur les modèles fréquence et coût moyen dégât des eaux.

L'analyse des corrélations

Une ultime vérification est faite sur la corrélation entre toutes les paires de paramètres du modèle afin d'éviter de trop fortes corrélations qui peuvent avoir un impact sur la stabilité du modèle.

La matrice de corrélation du modèle fréquence incendie est représentée ci-dessous :

Matrice de corrélations

	OFFER CLASS QUALITY (QualitePro)	OPTION COMPLEMENT PLUS (Yes)	MODE REGLEMENT (00)	EXCLUSION DES RISQUES LOCATIFS (Yes)	CODE EXISTENCE OPPOSANTS (Yes)	EXTINCT (No)	CENTRE COMM SPRINKLE (Yes)	ANT SIN GLOBAL Inc (NB - NbSin = 1)	ANT SIN GLOBAL Inc (NB - NbSin >= 2, RE - NbSin = 2)	ANT SIN GLOBAL Inc (RE - NbSin = 1)	ANT SIN GLOBAL Inc (RE - NbSin = 3)	ANT SIN GLOBAL Inc (RE - NbSin >= 4)	CAPI MOBILHOME (Yes)	SURVEILLANCE D UN CONTRAT (Yes)	CLASSE TEC INC (1)	CLASSE TEC INC (2)	CLASSE TEC INC (3)	CLASSE TEC INC (4)	CLASSE TEC INC (5)	Year*Covid_Period (2020-Standard+Curfew9PM)	Year*Covid_Period (Lockdown202001-202002-202005+Curfew8PM)	Indexed Fire content gaussian (1)	GLASS BREAKAGE CAPITAL (OPoly(1))	surface gaussian (OPoly(1))	Zone Inc Curve 1 (OPoly(1))	Surface*YEARxCOVID_PERIOD (2).surface gaussian
OFFER CLASS QUALITY (QualitePro)	1,00																									
OPTION COMPLEMENT PLUS (Yes)	0,01	1,00																								
MODE REGLEMENT (00)	0,05	0,04	1,00																							
EXCLUSION DES RISQUES LOCATIFS (Yes)	0,01	0,01	-0,02	1,00																						
CODE EXISTENCE OPPOSANTS (Yes)	0,02	-0,01	0,02	-0,01	1,00																					
EXTINCT (No)	0,03	0,04	-0,03	0,04	-0,02	1,00																				
CENTRE COMM SPRINKLE (Yes)	0,00	0,00	0,00	-0,03	0,00	0,01	1,00																			
ANT SIN GLOBAL Inc (NB - NbSin = 1)	0,00	0,00	0,00	0,00	0,01	0,02	0,00	1,00																		
ANT SIN GLOBAL Inc (NB - NbSin >= 2, RE - NbSin = 2)	0,01	-0,02	-0,01	0,01	-0,02	0,00	0,00	0,02	1,00																	
ANT SIN GLOBAL Inc (RE - NbSin = 1)	0,03	-0,02	-0,01	0,01	-0,03	-0,01	0,01	0,03	0,13	1,00																
ANT SIN GLOBAL Inc (RE - NbSin = 3)	0,00	-0,02	-0,01	0,00	-0,01	0,00	0,00	0,01	0,07	0,07	1,00															
ANT SIN GLOBAL Inc (RE - NbSin >= 4)	0,00	-0,01	-0,04	0,01	-0,03	-0,01	0,00	0,01	0,10	0,09	0,06	1,00														
CAPI MOBILHOME (Yes)	0,12	0,18	-0,05	0,04	-0,06	0,13	0,00	0,00	-0,17	-0,11	-0,11	-0,18	1,00													
SURVEILLANCE D UN CONTRAT (Yes)	0,04	-0,01	0,01	0,00	-0,04	-0,05	0,00	0,01	-0,09	-0,07	-0,08	-0,08	0,03	1,00												
CLASSE TEC INC (1)	0,01	0,00	0,00	0,00	0,01	-0,02	0,00	0,00	0,00	0,01	0,00	0,00	0,04	0,01	1,00											
CLASSE TEC INC (2)	0,14	0,02	0,01	-0,01	0,00	0,11	0,00	0,00	0,01	0,00	0,01	0,01	0,20	0,00	0,05	1,00										
CLASSE TEC INC (3)	0,09	0,13	0,02	0,02	-0,04	0,21	0,00	-0,02	0,00	-0,02	0,00	-0,01	0,32	0,03	0,04	0,21	1,00									
CLASSE TEC INC (4)	0,01	-0,03	0,02	0,00	-0,05	0,26	-0,02	-0,01	-0,02	-0,03	-0,01	-0,01	0,23	0,05	0,05	0,28	0,25	1,00								
CLASSE TEC INC (5)	-0,34	0,05	-0,09	0,01	-0,07	-0,06	0,00	-0,06	-0,07	-0,08	-0,07	-0,03	0,06	0,07	0,08	0,07	0,08	0,13	1,00							
Year*Covid_Period (2020-Standard+Curfew9PM)	-0,02	-0,01	0,02	0,00	0,00	0,05	0,00	0,01	0,00	-0,01	0,00	0,01	0,01	-0,01	0,00	0,00	0,00	0,01	-0,01	1,00						
Year*Covid_Period (Lockdown202001-202002-202005+Curfew8PM)	-0,03	-0,01	0,02	0,00	0,00	0,05	0,00	0,00	0,00	-0,01	-0,01	0,00	0,01	-0,01	0,00	0,00	0,01	0,02	-0,01	0,05	1,00					
Indexed Fire content gaussian (1)	0,00	-0,01	0,00	0,00	-0,01	0,00	0,00	0,00	0,00	0,00	0,00	0,00	-0,01	-0,02	0,03	0,00	0,00	-0,02	-0,06	0,00	0,00	1,00				
GLASS BREAKAGE CAPITAL (OPoly(1))	-0,24	-0,08	0,01	0,00	-0,03	0,12	0,00	-0,01	-0,02	-0,02	-0,03	-0,03	-0,05	0,03	0,10	0,00	-0,03	-0,08	0,31	0,01	0,01	-0,09	1,00			
surface gaussian (OPoly(1))	0,27	-0,02	-0,10	-0,01	-0,04	0,15	0,01	-0,02	-0,10	-0,07	-0,06	-0,09	-0,22	-0,05	-0,03	-0,02	-0,29	0,16	-0,03	0,00	0,03	0,00	-0,10	1,00		
Zone Inc Curve 1 (OPoly(1))	-0,02	-0,01	0,06	0,00	0,01	0,00	-0,01	-0,01	-0,02	-0,02	-0,01	-0,01	0,08	-0,01	0,01	0,00	0,03	0,04	-0,03	0,00	0,00	0,00	0,01	-0,06	1,00	
Surface*YEARxCOVID_PERIOD (2).surface gaussian (OPoly(1))	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,01	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	-0,08	0,00	0,00	-0,07	0,00	1,00

Figure 4-22 – Matrice de corrélation du modèle fréquence incendie

Dans le modèle fréquence incendie, les coefficients de corrélation ont un niveau acceptable compris entre -0,3 et 0,3. La corrélation négative la plus forte, de -0,34 entre la classe technique incendie n°5 et l'offre Qualité, s'interprète par une présence plus importante de commerces de rue dans la classe d'activité n°5.

La mesure de la performance et du surapprentissage (*overfitting*) des modèles

Les modèles GLMs sont entraînés sur l'échantillon de modélisation. Le but est de permettre au GLM de tirer des généralités des prédictions réalisées sur cet échantillon et de pouvoir les appliquer à de nouvelles données : l'échantillon de test. Les performances du modèle sont testées sur un échantillon de données différent de celui utilisé pour l'entraîner.

Pour mesurer la performance et la capacité des modèles à généraliser les tendances apprises, la différence de déviance entre le modèle sans variable sélectionnée et le modèle retenu est comparée entre la base de modélisation et la base de test :

Modèles	Echantillon	Déviance Modèle retenu	Déviance Modèle sans variables	Différence en %
INC fréquence	Modélisation	41 379,98	45 468,36	-8,99
	Test	10 448,06	11 471,64	-8,92
DDE fréquence	Modélisation	10 9310,33	118 569.,4	-7,81
	Test	27 010,30	29 338,38	-7,93
INC Coût moyen	Modélisation	10 423,97	11 193,88	-6,87
	Test	2 739,19	2 888,43	-5.17
DDE Coût moyen	Modélisation	16 458,81	19 116,81	-13,90
	Test	4 283,95	4 901,41	-12,60

Tableau 4.6 - Déviances bases de test et de modélisation

Sur les modèles de fréquence, les performances sur l'échantillon de test sont très proches de celles de l'échantillon de modélisation. Sur les modèles de coût moyen, les performances sur l'échantillon de test sont inférieures à celles de l'échantillon de modélisation, mais restent toutefois acceptables pour conclure à l'absence de surapprentissage.

Les modèles GLMs ainsi validés sur l'incendie et le dégât des eaux permettent d'identifier les critères explicatifs de la fréquence et du coût moyen et mettent en avant les disparités par profils de risques. Ces modèles caractérisent uniquement des sinistres de faible montant survenant fréquemment. Les sinistres dits atypiques sont associés à des montants pouvant atteindre plusieurs milliers d'euros. Ces sinistres importants ne se trouvent que sur la garantie incendie et nécessitent un traitement spécifique.

4.2. La modélisation des sinistres incendie atypiques

La garantie incendie couvre des risques dont les coûts pour un assureur sont bien plus grands que sur les autres garanties. En effet, un incendie peut amener à la destruction totale du local commercial de l'assuré mais peut également toucher les locaux voisins. Ces sinistres peu fréquents peuvent porter à eux seuls une grosse partie de la charge comme le montre le tableau ci-dessous :

	Sinistres incendie	dont sinistres atypiques > 50 k€	Poids des sinistres atypiques
Nombre	5 084	369	7,3%
Charge	136 775 629	100 387 462	73,4%

Tableau 4.7 - Poids des sinistres incendie atypiques

En incendie, 7 % des sinistres représentent 73 % de la charge. Le nombre des sinistres incendie atypiques est modélisé par un modèle de propension : estimation du pourcentage de sinistres atypiques parmi l'ensemble des sinistres incendie. La différenciation du coût de ces sinistres par profil de risque est quant à elle plus difficile à déterminer.

4.2.1. Le modèle de propension des sinistres incendie atypiques

La propension représente le pourcentage de sinistres atypiques sur l'ensemble des sinistres incendie. C'est la probabilité qu'un individu soit placé dans un groupe d'exposition aux sinistres incendie atypiques étant donné un ensemble de caractéristiques observées. L'approche traditionnelle pour générer un score de propension est la régression logistique. La présence d'un sinistre atypique (supérieur ou égal à 50 k€) parmi les sinistres incendie est la variable à expliquer (réponse : sinistre incendie atypique, poids : nombre de sinistres incendie) et les caractéristiques sur lesquelles sont comparés les groupes sont les variables explicatives. Le modèle logistique génère un nombre de 0 à 1 (score de propension) pour chaque observation. Ensuite, les scores de propension similaires sont agencés de manière à obtenir des groupes dans lesquels chaque individu a la même probabilité d'être exposé à un sinistre incendie atypique.

La base de données pour la modélisation des sinistres incendie atypiques correspond à la base utilisée pour la modélisation des attritionnels sur laquelle seuls les sinistres incendie supérieur à 0 sont retenus.

➢ La régression logistique avec un modèle GLM

La régression logistique s'inscrit dans la série des modèles linéaires généralisés. L'objectif ici est d'estimer la probabilité qu'un sinistre soit atypique conditionnellement au profil du risque sous-jacent : la distribution suit une loi binomiale. L'avantage de la régression logistique est qu'elle introduit une fonction de lien logit qui permet de réaliser la régression linéaire dans ce cas.

• Les variables explicatives des sinistres incendie atypiques

Comme pour les modèles de fréquence et de coût moyen, une démarche pas à pas est appliquée pour sélectionner les variables. Le modèle retenu comporte 9 variables. L'évolution de l'AIC en fonction des variables sélectionnées est représentée ci-dessous :

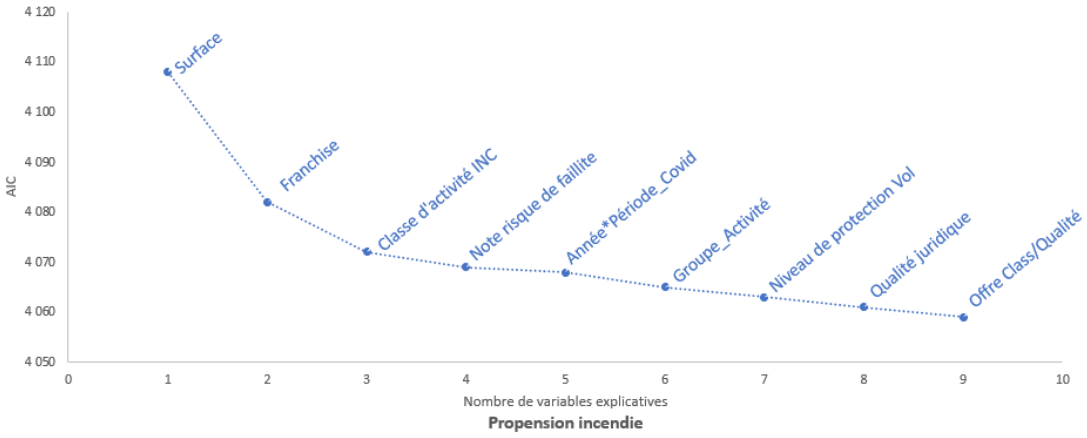


Figure 4-23 –Gain d'AIC du modèle de propension des sinistres incendie atypiques

Chacune des variables retenues est croisée avec la variable Année x Période_Covid pour s'assurer de leur stabilité par période. Aucune interaction n'est retenue, leur résultat est stable par période. Les effets des principales variables retenues sont exposés en annexe. Quelques-unes sont interprétées ci-dessous à titre d'illustration :

L'année croisée avec les périodes Covid

Le poids des sinistres incendie atypiques varie selon les années :

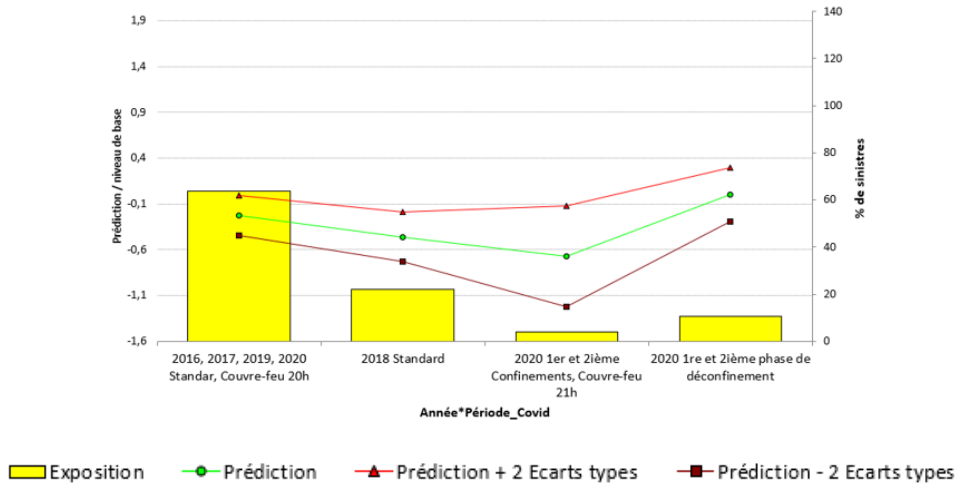


Figure 4-24 - Prédiction des sinistres incendie atypiques par Année x Période_Covid

Les deux périodes de confinement sont moins impactées par les sinistres incendie atypiques. La fermeture ou la baisse d'activité de certains commerces, et la diminution des déplacements de la population pendant ces périodes expliquent un nombre moins important de sinistres incendie atypiques. L'effet inverse se constate lors des phases de déconfinement qui s'accompagne d'une augmentation de sinistres incendie atypiques à la suite d'une reprise d'activité à plein temps après une période de baisse d'activité.

L'activité

Deux variables liées aux activités sont retenues : le groupe d'activité et la classe technique incendie. Chacune de ces variables apporte un niveau d'information complémentaire sur la prédiction de sinistres incendie atypiques.

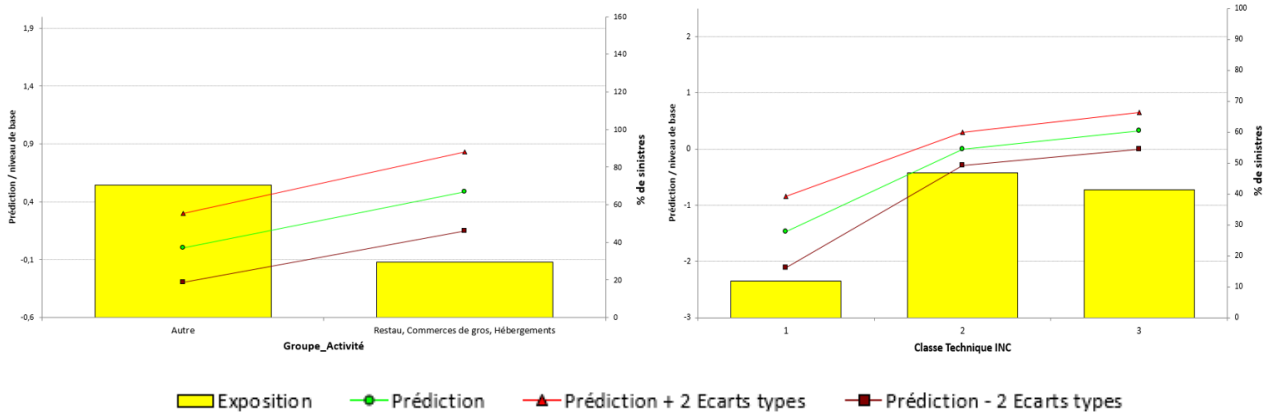


Figure 4-25 - Prédiction des sinistres incendie atypiques par critères d'activités

Les restaurants, commerces de gros et les établissements d'hébergements sont plus exposés aux sinistres incendie atypiques. La superficie et les capitaux assurés sur ces activités sont généralement plus élevés que sur les autres activités. De plus, les restaurants du fait de la présence d'une cuisine sont des activités plus exposées aux sinistres incendie de grande ampleur.

Les classes techniques qui pour rappel sont définies selon les caractéristiques des activités (superficie, chiffre d'affaires, effectif, note de risque de faillite …) sont regroupées en 3 classes : une classe très peu exposée aux sinistres incendie atypiques et deux classes plus exposées.

Le niveau de protection vol

La présence de moyens de protections contre le vol est une variable explicative de la survenance de sinistres incendie atypiques :

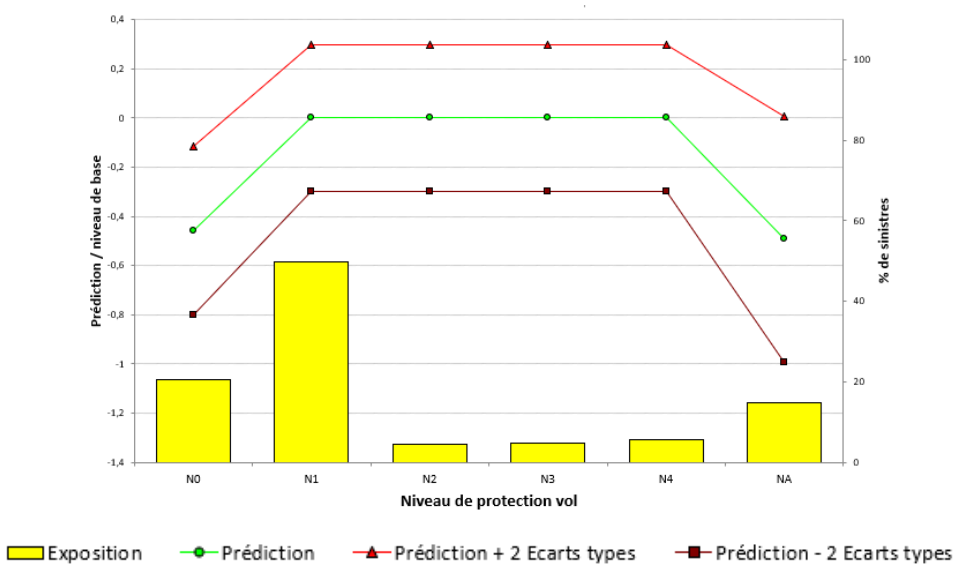


Figure 4-26 – Prédiction des sinistres incendie atypiques par niveau de protection vol

La question posée initialement en vol reflète des niveaux de risque différents en incendie. La survenance de sinistres incendie atypiques est plus importante chez les professionnels qui mettent en place des moyens pour se protéger du vol, signe d'une activité plus risquée.

Le niveau de franchise

Les contrats souscrits avec un niveau de franchise élevé sont plus exposés aux sinistres incendie atypiques :

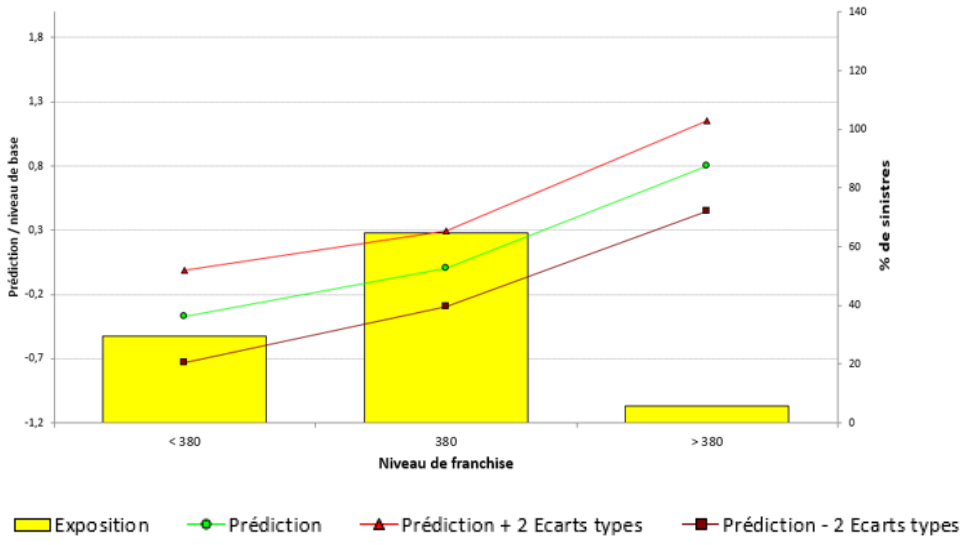


Figure 4-27 – Prédiction des sinistres incendie atypiques par niveau de franchise

Le montant de franchise compris entre 0 € et 3 000 € a été regroupé en 3 niveaux : inférieur, égal ou supérieur à 380 €. Les montants de franchises, faibles par rapport au coût d'un sinistre atypique, caractérisent des comportements différents des assurés. Par ailleurs, un niveau de franchise élevé est souvent lié à un niveau de capital garanti important, et plus le capital garanti est grand, plus l'indemnité en cas d'incendie peut dépasser le seuil des sinistres atypiques.

Pour l'ensemble des variables retenues, le test de Wald est satisfaisant, la p-value est nettement inférieure à 5 % :

Variables	Wald p-value
Année_Période_Covid	0,9823%
Groupe_Activité	0,0084%
Classe technique INC	0,0095%
Offre Class/Qualité	0,0139%
Qualité juridique	0,0797%
Surface	0,806%
Niveau de protection vol	0,0149%
Note de risque de faillite	0,0036%
Franchise	0,0000%

Tableau 4.8 - P-value du modèle de propension incendie

L'importance d'une variable explicative est mesurée par l'envergure de ses beta. Le calcul de l'intervalle entre le beta le plus grand et le plus petit (βmax - βmax) permet de classer les variables retenues dans le modèle par ordre d'importance :

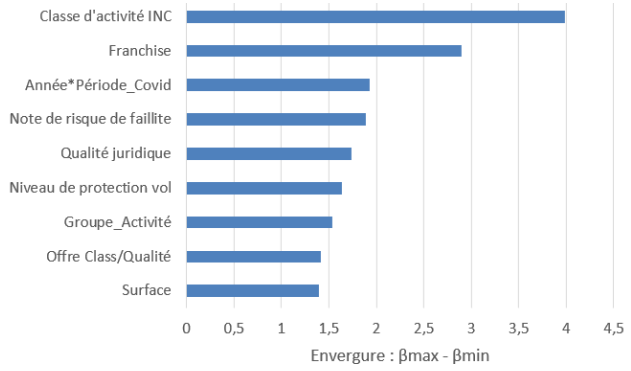


Figure 4-28 – Importance des variables du modèle GLM de propension incendie

La classe d'activité incendie ressort avec le plus fort pouvoir discriminant. Les caractéristiques de l'activité jouent un rôle important dans la détermination de sinistres incendie atypiques. La période : Année x Période_Covid, arrive en troisième position des variables les plus importantes, en effet l'aspect aléatoire des sinistres de plus grande ampleur contribue à avoir certaines périodes plus touchées que d'autres par ces sinistres.

- La performance statistique du modèle

La performance du modèle est évaluée par la qualité de l'ajustement et le pouvoir de prédiction. La validation de l'ajustement est effectuée en comparant l'observé et l'estimé sur l'échantillon de modélisation. Le pouvoir prédictif du modèle est vérifié en appliquant les coefficients tarifaires estimés sur l'échantillon de test.

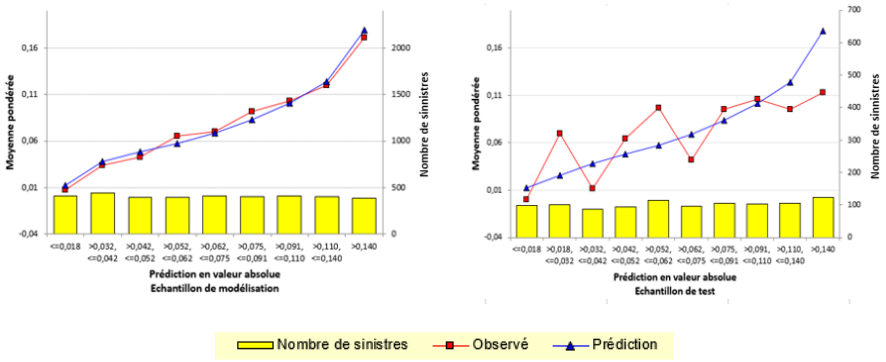


Figure 4-29 – Qualité de l'ajustement et pouvoir prédictif du modèle de propension incendie

L'ajustement sur l'échantillon de modélisation est cohérent avec l'observé. Le faible volume de données sur la base de test montre une prédiction plus éloignée de l'observé. Toutefois l'observé est équitablement réparti de part et d'autre des prédictions, ce qui conforte le pouvoir prédictif du modèle.

Le pouvoir de prédiction est évalué avec le coefficient de Gini calculé sur la base de test :

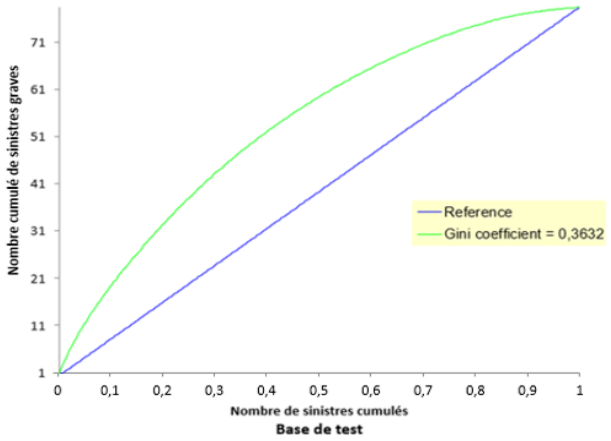


Figure 4-30 – Coefficient de Gini du modèle GLM de propension incendie

Le coefficient de Gini de 0,3632 traduit une segmentation satisfaisante.

La régression logistique suppose que toutes les variables soient linéaires et additives sur l'échelle logarithmique. Par ailleurs, pour créer des groupes d'exposition aux sinistres atypiques comparables, toutes les variables qui peuvent avoir un impact sur l'appartenance à un groupe d'exposition doivent être identifiées. Or, plus le nombre de variables inclues dans le modèle augmente, plus le nombre de degrés de liberté est important et plus la puissance statistique diminue. L'incapacité à identifier toutes les variables potentielles peut entraîner un déséquilibre entre les groupes d'exposition.

Dans le but d'améliorer la qualité du modèle de propension, une approche via un algorithme d'apprentissage statistique (*Machine Learning*) est testée. L'utilisation d'un algorithme de *Gradient boosting* (*LightGBM*) peut être une solution pour créer des groupes d'exposition comparables.

➢ La régression logistique avec un modèle *LightGBM*

L'algorithme *Light Gradient Boosting Machine* (*LightGBM*) est un algorithme de *Machine Learning* qui génère entre autres des scores de propension. Ce modèle utilise des arbres de régression multiples qui capturent toutes les relations complexes qui peuvent exister entre l'affectation de l'exposition aux sinistres atypiques et les variables explicatives.

L'algorithme *LightGBM* a été retenu pour sa vitesse d'apprentissage et son efficacité plus élevée par rapport aux autres algorithmes de *Machine Learning*. Le modèle *LightGBM* possède de nombreux paramètres, ce qui lui permet de produire des arbres beaucoup plus complexes en suivant une approche de division par feuille plutôt qu'une approche par niveau qui est le principal facteur permettant d'obtenir une plus grande précision. Cependant, cela peut parfois conduire à un surapprentissage qui peut être évité en définissant les hyperparamètres optimaux comme la profondeur maximale ou le nombre de feuilles de l'arbre.

Deux aspects sont propres au *LightGBM*. Le premier est que l'algorithme fonctionne avec sa propre structure de base de données qui s'obtient à partir d'un DataFrame en utilisant la fonction lightgbm.Dataset. Le deuxième est le score d'initialisation à renseigner qui permet d'initialiser le modèle à partir de la réponse moyenne. Cela évite que l'algorithme apprenne avec une constant inappropriée. Une fois les données prêtes, le modèle peut être entraîné.

- L'optimisation des paramètres

Un premier modèle ***LightGBM*** de **régression logistique** ('objective' : binary, 'metric' : binary_logloss) est entrainé et validé avec des **paramètres par défaut** (la fonction lightgbm.train prend en entrée la base d'apprentissage : 60 % des données, la base de validation : 20 % des données et une série d'hyperparamètres). Ensuite, les **hyperparamètres** sont optimisés via le package **Optuna**. Le processus d'optimisation avec Optuna vise à trouver le meilleur ensemble d'hyperparamètres en les faisant varier dans une plage restreinte, dans le but de minimiser la métrique logloss. Le tableau ci-dessous reprend les paramètres appliqués par défaut dans le premier modèle et les paramètres après optimisation :

Hyperparamètres	Modèle par défaut	Modèle optimal
lambda_l2	0,1	1,9968
min_sum_hessian_in_leaf	-	9,5790
max_depth	8	4
min_gain_to_split	-	0,5196
feature_fraction	0,8	0,8764
bagging_fraction	0,75	0,7778
learning_rate	0,01	0,0066
num_leaves	64	9
LogLoss	0,2522	0,2405

Tableau 4.9 - Optimisation des hyperparamètres du *LightGBM* de propension incendie

L'optimisation des hyperparamètres permet de limiter la profondeur des arbres, le nombre de feuilles, et d'appliquer un paramètre de régularisation (lambda L2) plus élevé favorisant la généralisation et limitant ainsi les risques de surapprentissage.

Optuna fournit plusieurs fonctions via le module « optuna.visualization » pour analyser visuellement les résultats du processus d'optimisation. La fonction plot_param_importances() est utilisée pour visualiser l'importance des paramètres :

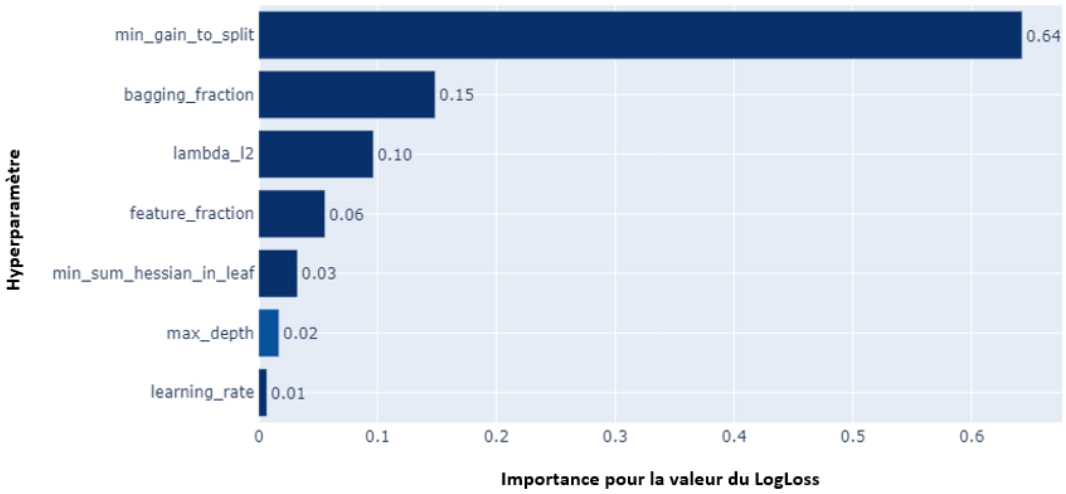


Figure 4-31 – Importance des hyperparamètres du *LightGBM* de propension incendie

Le gain apporté par chaque division est le paramètre qui a le plus d'influence sur l'optimisation de la valeur de la métrique retenue.

L'objectif de la fonction d'optimisation implémentée est de minimiser le logloss, cependant, un autre ensemble de valeurs est intéressant à analyser : le nombre d'arbres (itérations) et le nombre moyen de divisions qui contiennent des informations sur la complexité du modèle. Par exemple, dans le cas de deux arbres différents avec des prédictions similaires, travailler avec le plus petit, c'est-à-dire celui avec un nombre de divisions inférieur est plus pertinent. Le compromis entre la **complexité** : nombre total de divisions (nombre d'arbres multiplié par le nombre moyen de divisions) et la **qualité** (logloss) pour chaque essai du processus d'optimisation est présenté ci-dessous :

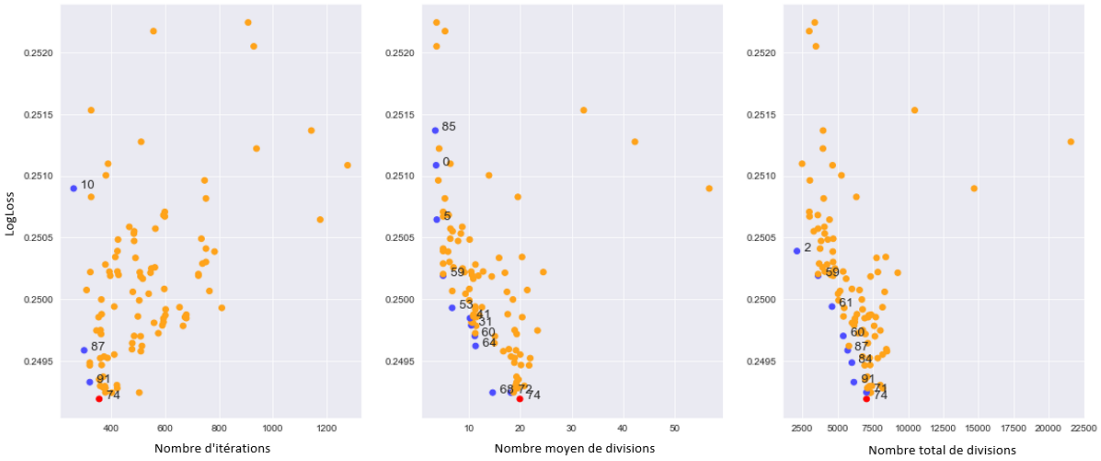


Figure 4-32 – Compromis complexité vs qualité du modèle

Le choix doit se porter sur l'un des points sur la frontière efficiente, à savoir ceux pour lesquels il n'y a pas d'autres points caractérisés à la fois par une meilleure valeur de la métrique et une complexité minimale (nombre total de divisions). Le point retenu correspond au meilleur essai (essai 74) identifié par un point rouge sur la frontière efficace. Le modèle avec les hyperparamètres optimaux est entrainé.

- L'importance des variables

L'importance d'une variable est mesurée par les gains totaux des fractionnements qui utilisent cette variable. Le graphique ci-dessous représente les 14 variables principales par ordre d'importance de gain :

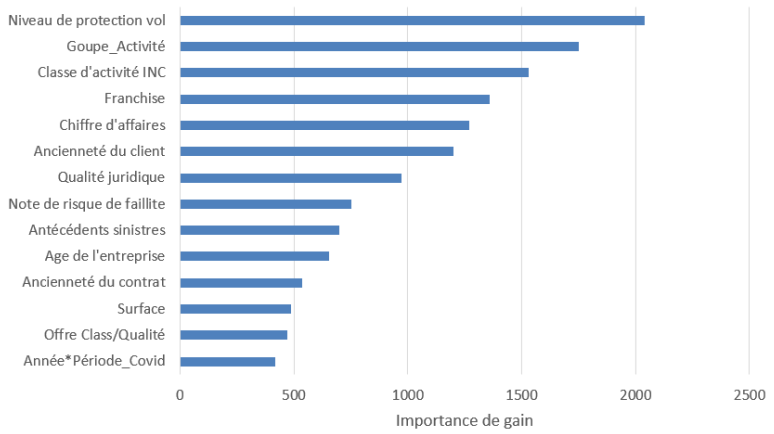


Figure 4-33 – Importance des variables du modèle de propension *LightGBM*

Le niveau de protection contre le vol, le groupe d'activité et la classe d'activité sont les trois variables les plus importantes du modèle. Elles renseignent sur la valeur des biens stockés et le type d'activité. La protection vol, question posée initialement en vol reflète des niveaux de risque différent en incendie. La franchise qui ressort également comme une des variables explicatives de la propension des sinistres incendie atypiques caractérise des comportements différents des assurés. Les variables de taille : capital, chiffre d'affaires et surface ressortent aussi comme explicatives de sinistres incendie atypiques. Certaines années peuvent être plus impactées que d'autres par les sinistres importants, d'où l'importance de la variable Année x Période_Covid.

- La performance du modèle

La qualité du modèle est évaluée en traçant l'observé et les prédictions obtenues sur l'échantillon de **test**. Le graphique du modèle initial avec les paramètres par défaut est comparé à celui du modèle après ajustement des paramètres :

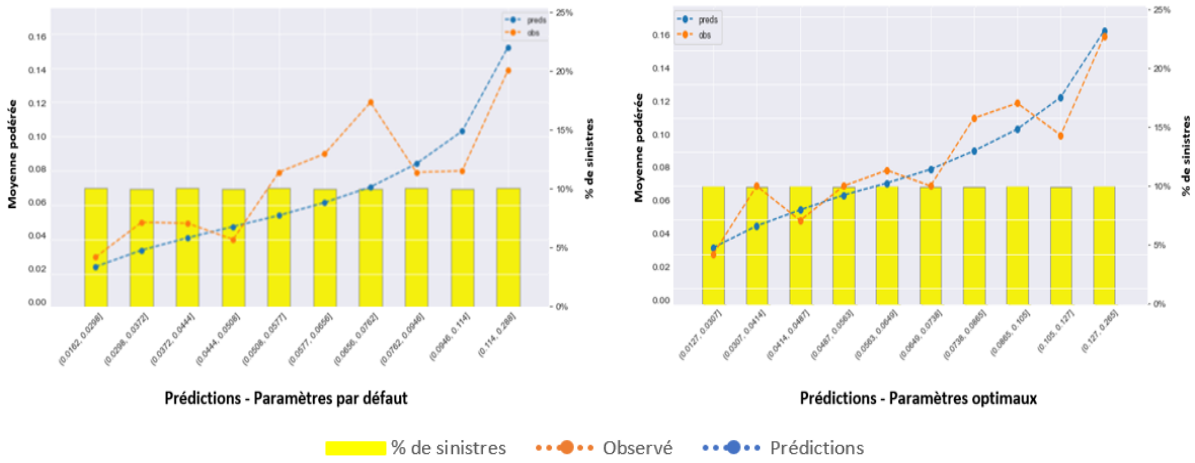


Figure 4-34 – Prédictions du score de propension avant et après ajustement des hyperparamètres

En comparant la courbe de prédiction du modèle optimal avec le modèle par défaut, la prédiction du modèle optimal (courbe bleue) est plus proche de l'observé (courbe rouge).

Le coefficient de Gini avant et après optimisation est présenté dans le tableau ci-dessous :

	Modèle par défaut	Modèle optimal	Gain
Coefficient de Gini	0,4101	0,4237	0,0136

Tableau 4.10 - Gini avant et après optimisation des hyperparamètres

Le coefficient de Gini gagne plus d'un point après ajustement des hyperparamètres. Le gain sur la performance du modèle n'est pas négligeable, cela signifie que la performance dépend du choix des hyperparamètres.

Pour s'assurer de la stabilité du modèle, le coefficient de Gini de la base de test est comparé à celui de la base d'apprentissage :

	Base d'apprentissage	Base de test	Delta
Coefficient de Gini	0,4252	0,4237	-0,0015

Tableau 4.11 - Gini base de test et d'apprentissage

Le coefficient de Gini de 0,4252 obtenu sur la base d'apprentissage est proche de celui de la base de test qui vaut 0,4237. Cette baisse très faible du coefficient de Gini valide la stabilité du modèle.

➢ La comparaison GLM vs *LightGBM*

Le modèle de propension de type GLM est comparé au modèle *LightGBM* sur deux critères : le choix des variables et la performance.

L'importance de variables est évaluée en fonction du pouvoir discriminant (intervalle des beta) pour le GLM et selon le gain pour le *LightGBM*. Afin de pouvoir les comparer, ces deux métriques rapportées à leur moyenne, sont représentées sur le graphique ci-dessous pour chacune des variables retenues :

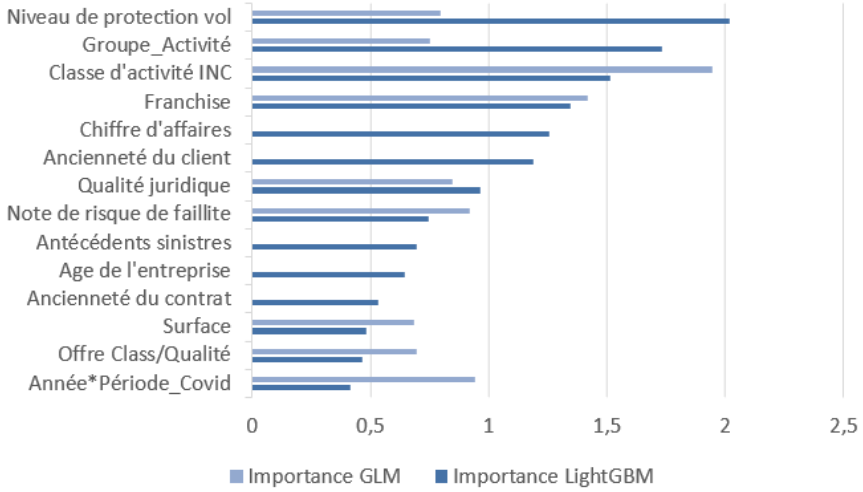


Figure 4-35 –Importance des variables des modèle GLM et *LightGBM* de propension incendie

Les 9 variables retenues dans le GLM se retrouvent parmi les variables les plus importantes du *LightGBM*, bien que leur ordre d'importance soit différent. Les variables permettant de distinguer les effets du Covid : Année x Période_Covid et Groupe_activité sont identifiées dans les deux méthodes. En revanche, le *LightGBM* sélectionne plus de variables que le GLM. Les variables chiffre d'affaires, antécédents sinistres et d'ancienneté ont été ajoutées dans le modèle *LightGBM*. Cet algorithme prend en compte plus de variables pour identifier la survenance de sinistres incendie atypiques, et permet une distinction plus fine des groupes de risque exposés aux sinistres importants.

Le second critère pour comparer les deux algorithmes est le coefficient de Gini qui permet d'évaluer la performance des modèles. Le tableau ci-dessous reprend les coefficients de Gini calculés sur la base de test pour chacun des modèles :

Modèle	Coefficient de Gini		
	GLM	*LightGBM*	Delta Gini *LightGBM* - GLM
Propension incendie	0,3632	0,4237	0,0605

Tableau 4.12 - Comparaison Gini GLM/*LightGBM*

Le modèle *LightGBM* améliore le coefficient de Gini de 6 points, ce qui n'es pas négligeable en termes de performance. Ce modèle est retenu pour estimer la propension des sinistres incendie atypiques.

4.2.2. Le modèle coût moyen des sinistres incendie atypiques

➢ La distinction entre sinistres incendie atypiques de tendance et exceptionnels

La sur-crête représente pour chaque sinistre supérieur à 50 k€, la partie de la charge supérieure au seuil de 50 k€ déterminé selon la théorie des valeurs extrêmes. La distribution de la sur-crête est identifiée en traçant les QQ plot de différentes lois.

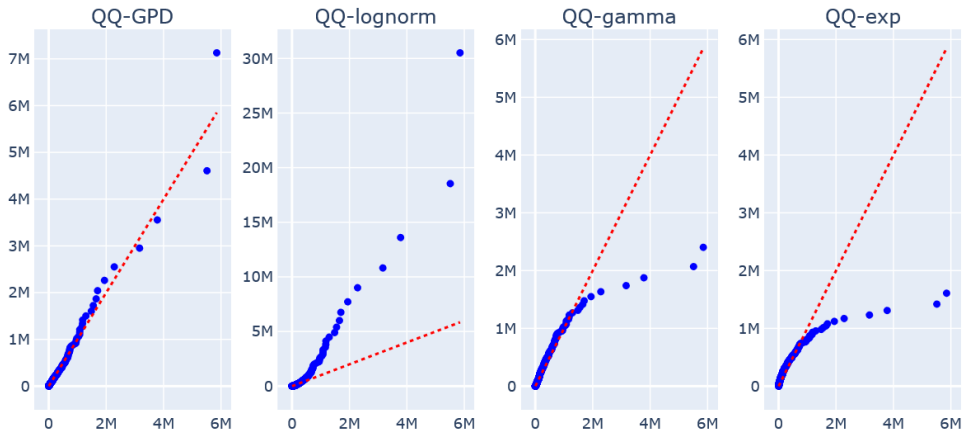


Figure 4-36 – QQ plot de la sur-crête

D'après les graphiques ci-dessus, la distribution de la sur-crête suit une loi de Pareto Généralisée (GPD), loi à queue lourde. Cependant, le QQ-plot de la loi Gamma montre que le début de la distribution suit une loi Gamma.

La sur-crête allant de 50 k€ à plusieurs millions d'euros est très vaste, l'idée est de la dissocier en deux, avec :

- d'une part la portion de la sur-crête modélisable par une loi Gamma représentant les sinistres incendie atypiques **de tendance**;
- d'autre part, les sinistres incendie atypiques dits **exceptionnels** trop peu nombreux pour être modélisables et dont le coût sera mutualisé sur l'ensemble des sinistres atypiques.

Le seuil retenu pour distinguer les sinistres incendie atypiques de tendance des sinistres atypiques exceptionnels est déterminé en analysant la distribution de la sur-crête :

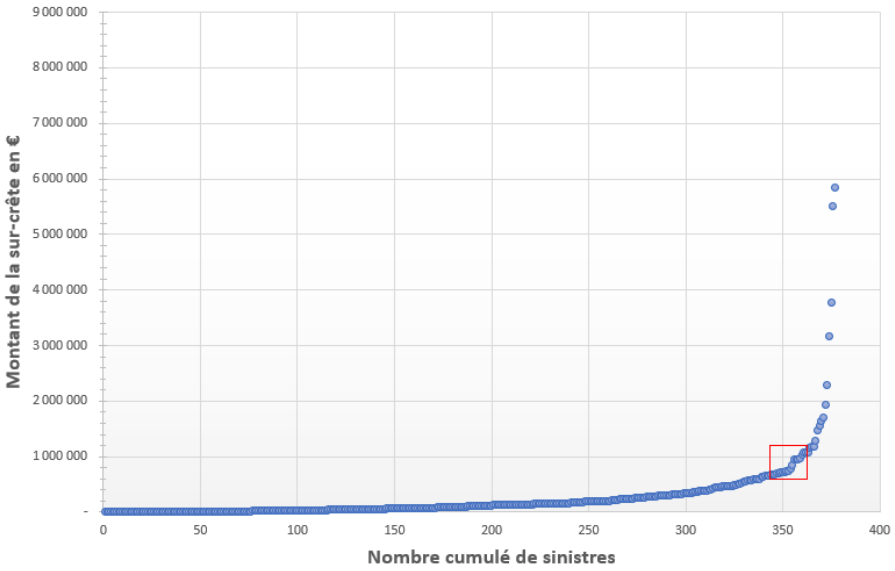


Figure 4-37 – Distribution de la sur-crête

La distribution de la sur-crête montre une cassure de la tendance à partir de 750 k€. Ce seuil est retenu pour séparer les sinistres incendie atypiques de tendance des sinistres atypiques exceptionnels. Un **second seuil d'écrêtement** est défini à 750 k€. La sur-crête au-dessus de 750 k€ concerne **25 sinistres** pour un montant de **25,4 M€.**

➢ La modélisation des sinistres incendie atypiques de tendance

Les sinistres incendie atypiques de tendances dont le montant de la sur-crête est supérieur à 50 k€ et capé à 750 k€ suivent une loi Gamma :

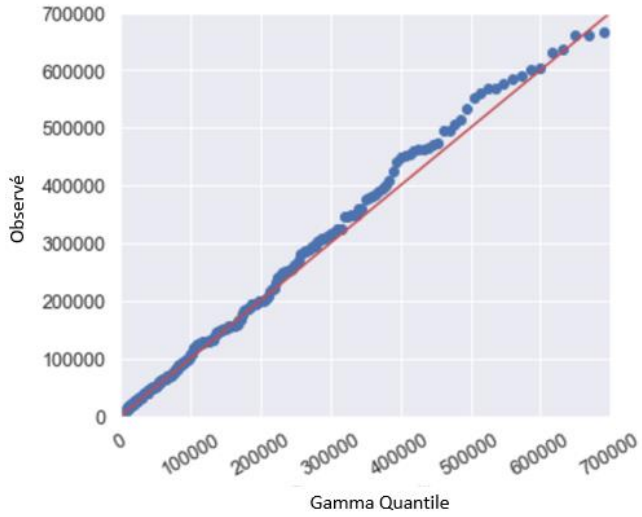


Figure 4-38 – QQ plot sur-crête 50 k€ - 750 k€

Pour être cohérent avec la méthode utilisée pour le score de propension, le coût moyen des sinistres incendie atypiques de tendance est modélisé par un algorithme de régression ***LightGBM*** qui suit une loi Gamma et dont la fonction de perte est la deviance Gamma ('objective' : Gamma, 'metric' : Gamma). Après ajustement des paramètres en utilisant le package Optuna, la courbe des prédictions par rapport à l'observé sur l'échantillon de test est la suivante :

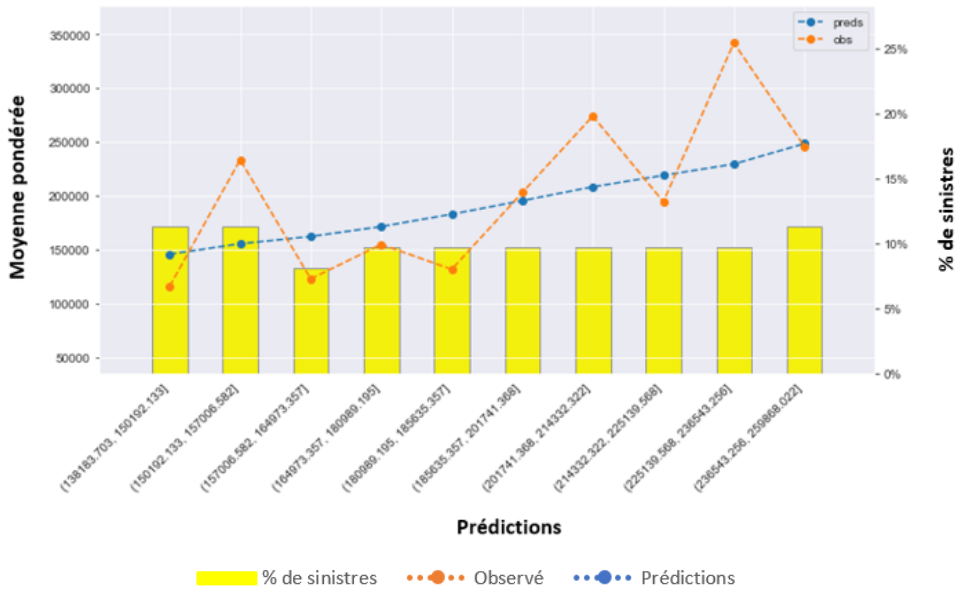


Figure 4-39 – Prédiction du coût moyen des sinistres incendie atypiques de tendance

Le coût moyen observé des sinistres compris entre 50k€ et 750 k€ (en rouge) est assez est bien réparti de part et d'autre de la courbe des prédictions (en bleu).

Les variables du modèle par importance de gain sont les suivantes :

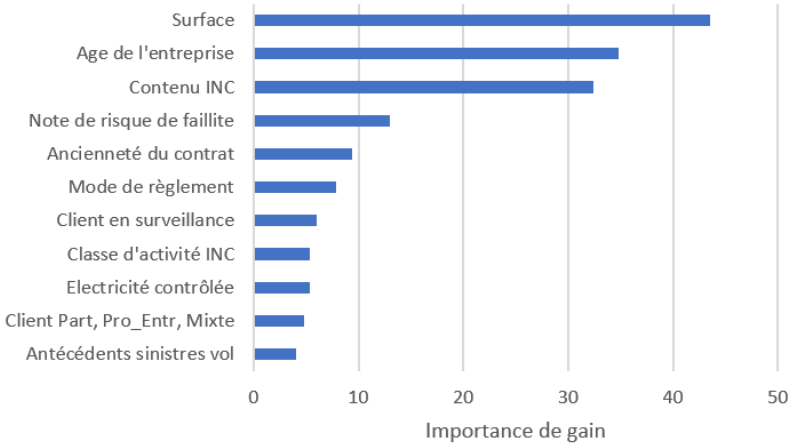


Figure 4-40 – Importance des variables du modèle *LightGBM* de coût moyen des sinistres incendie atypiques

La surface apparait comme la variable la plus significative du coût moyen des sinistres incendie atypiques. Plus les locaux commerciaux sont grands, plus le coût de reconstruction est élevé en cas d'incendie. Le capital assuré ressort également comme variable importante : le montant d'indemnisation est lié au capital assuré. La variable temporelle Année x Période_Covid est identifiée par le modèle pour segmenter le coût moyen des sinistres incendie atypiques. D'autres variables liées au client comme la multi-détention de contrats, ou relatives aux moyens de prévention apportent de l'information.

Afin de valider la stabilité du modèle, le coefficient de de Gini de la base d'apprentissage est comparé à celui de la base de test :

	Coefficient de Gini		
Modèle	Apprentissage	Test	Delta Gini
LightGBM Coût moyen incendie	0,383	0,377	-0,006

Tableau 4.13 - Gini base de test et d'apprentissage

Le coefficient de Gini de la base d'apprentissage de 0,383 est très proche de celui de la base de test. Ce faible écart valide la stabilité du modèle.

4.3. La prime pure et les outils d'aides à la décision

4.3.1. La prime pure

Une fois les modèles de fréquence, coût moyen des attritionnels et des sinistres atypiques validés, la prime pure par garantie est déterminée de la manière suivante :

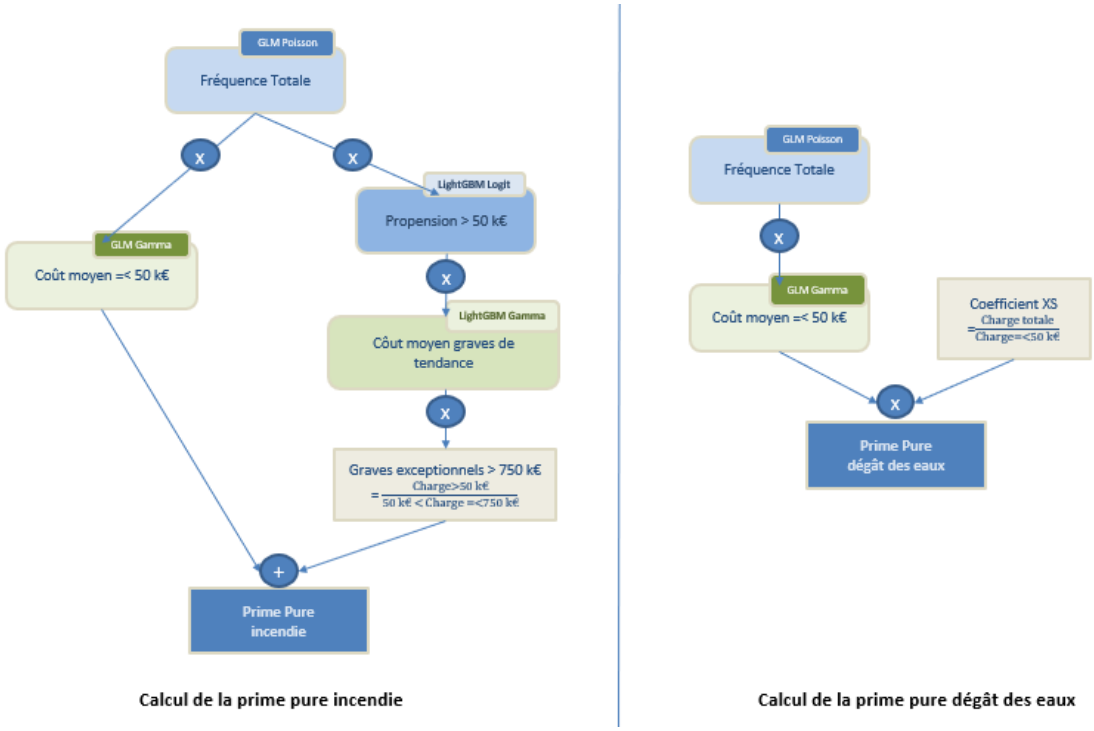


Figure 4-41 – Détermination de la prime pure

Le calcul de la prime pure incendie se distingue de celui du dégât des eaux : pour la garantie incendie, le seuil d'écrêtement retenu a permis de modéliser une partie de la sur-crête, tandis que pour la garantie dégât des eaux, un coefficient de sinistres atypiques permet de répartir le coût de la sur-crête.

Les graphiques ci-dessous représentent la prime pure estimée comparée à celle observée. Cette comparaison est présentée par périodes et sur l'un des principaux critères de tarification, la surface.

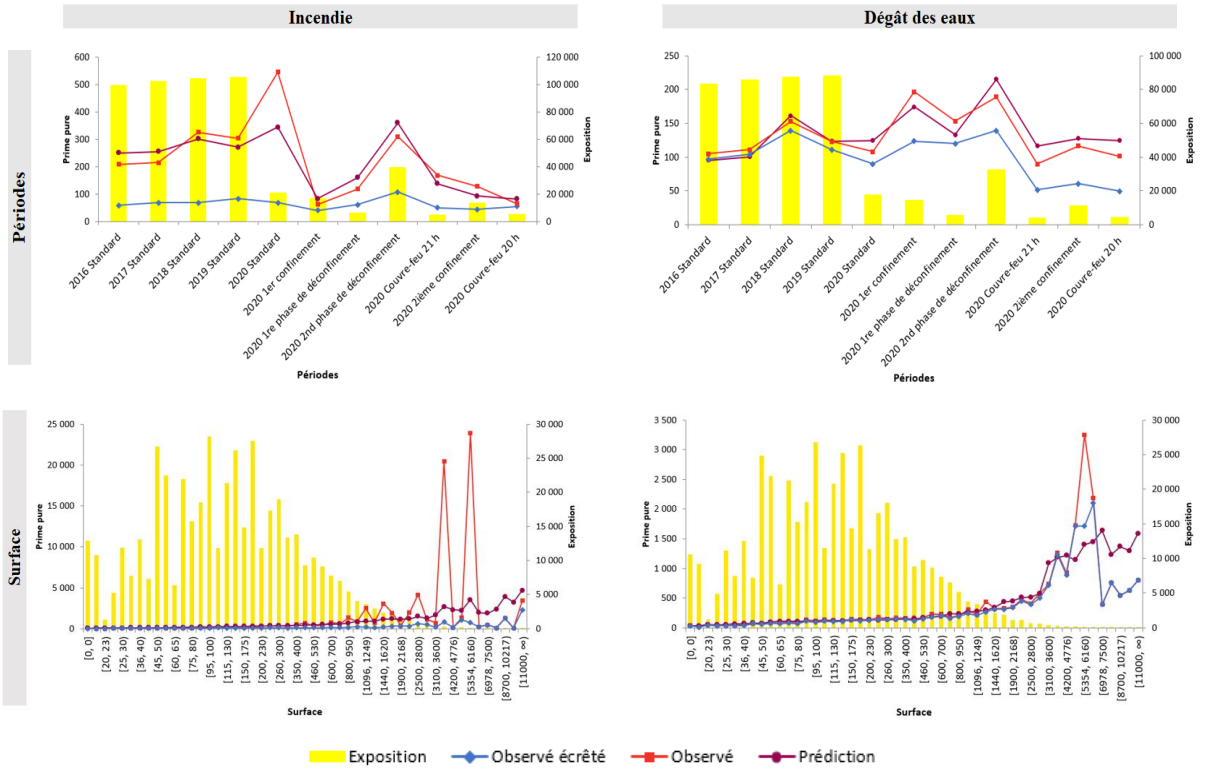


Figure 4-42 – Prime pure par périodes et surface

Sur les graphiques de prime pure par périodes, le gap important en incendie entre l'écrêté et la prime pure totale met en avant le poids des sinistres atypiques sur cette garantie. L'estimation de la prime pure par période permet d'identifier les tendances de la sinistralité selon les différentes mesures de restrictions. En incendie, la sinistralité est plus faible pendant les périodes de restrictions liées à une baisse de l'activité, et a tendance à augmenter en périodes de déconfinement lorsque l'activité reprend. Sur la garantie dégât des eaux, les périodes de confinements et de déconfinements s'accompagnent d'une augmentation du coût moyen des sinistres, tandis que les périodes de couvre-feu semblent moins impacter la sinistralité. Cette distinction permet de poser des hypothèses sur la situation de la pandémie dans l'année à venir et d'appliquer un coefficient de sinistralité en fonction des hypothèses retenues.

Sur les graphiques de prime pure par surface, la prime pure estimée est en phase avec l'observée sur les tranches où le volume d'exposition est important. Sur les tranches avec une faible exposition, l'espérance des sinistres est inférieure au montant réel de la sinistralité dû à l'effet de la mutualisation des sinistres atypiques. Toutefois, pouvoir détecter les profils de risque avec une probabilité de sinistres incendie atypiques supérieure à la moyenne est intéressant pour formuler des règles de sélection à la souscription et améliorer la rentabilité du produit.

4.3.2. La mise en place de règles de souscription de sinistres incendie atypiques

Une règle d'identification de segments de clients avec une forte propension à avoir des sinistres incendie atypiques peut être définie à partir d'arbre de décision.

La méthode appliquée est la suivante :

- un arbre de décision est paramétré et entrainé pour modéliser la propension ;
- les feuilles où la propension moyenne est supérieure à la propension au niveau du portefeuille sont identifiées et les règles qui les définissent sont extraites ;
- chaque règle peut être vue comme un modèle de classificateur binaire : si l'observation répond aux critères de la règle, le profil peut être classé comme sujet à des sinistres incendie atypiques ;
- la règle qui maximise la classification : le F_β-score est sélectionnée ;

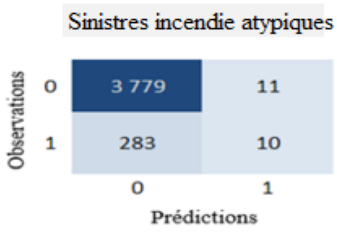
- la règle sélectionnée est combinée avec les autres ;
- la règle combinée qui maximise le F_β-score représente la règle de sélection des profils avec une forte probabilité de sinistres incendie atypiques.

Cet algorithme permet de déterminer la règle de détermination des profils d'assurés avec une propension de sinistres incendie atypiques plus élevée que la moyenne :

Règle de profils sinistres incendie atypiques
Groupe_Activité = Restauration Et Chiffre d'affaires > 226 186 € Et Age_entreprise < 6 ans Et Prévention restaurant = Oui

Tableau 4.14 - Profils d'assurés enclins aux sinistres atypiques

La matrice de confusion qui permet de détecter si un sinistre est atypique ou non par application de cette règle est la suivante :



	Précision	Rappel	β	F_β-score
Sinistre atypique	0,476	0,034	0,1	0,422

Figure 4-43 – Matrice de confusion et métriques de la règle de profils de sinistres incendie atypiques

La précision mesure les sinistres incendie correctement identifiés comme étant atypique parmi tous les sinistres incendie topés atypiques, tandis que le rappel renseigne sur le nombre de sinistres correctement identifiés comme atypiques parmi les sinistres incendie qui sont réellement atypiques. Le F_β-score, moyenne harmonique de la précision et du rappel, combine ces deux notions. Le beta, paramètre de l'algorithme, permet d'apporter plus de poids à la précision ou au rappel. Le F_β-score est optimisé pour un beta égal à 0,1. Une plus grande importance est accordée à la précision. Un assureur ne peut pas perdre des clients sur la base de la prédiction d'un modèle qui les détermine à tort plus exposés aux sinistres incendie atypiques que la moyenne des assurés.

La mise en place de règles de souscription pour identifier les profils d'assurés avec une plus forte propension aux sinistres incendie atypiques complète l'information apportée par les modèles de prime pure. Par ailleurs, la modélisation du risque par classes d'activités permet d'identifier les activités dont la prime pure n'est pas en phase avec le niveau de tarif actuellement commercialisé.

4.3.3. La révision des classes d'activités

Les modèles ont permis de déterminer des niveaux tarifaires par classes d'activités. Les coefficients tarifaires des classes d'activités créées avec la classification ACP/CAH (classes « techniques ») sont comparés à ceux des classes déterminées à dire d'expert (classes « commerciales ») actuellement utilisées dans le tarif commercial.

La comparaison du niveau tarifaire technique et commercial est réalisée par activité, en rattachant chaque activité au coefficient tarifaire de sa classe. Après normalisation des coefficients techniques et commerciaux

par rapport au coefficient moyen pondéré par l'exposition, les activités avec les écarts les plus importants sont représentés dans les graphiques ci-dessous :

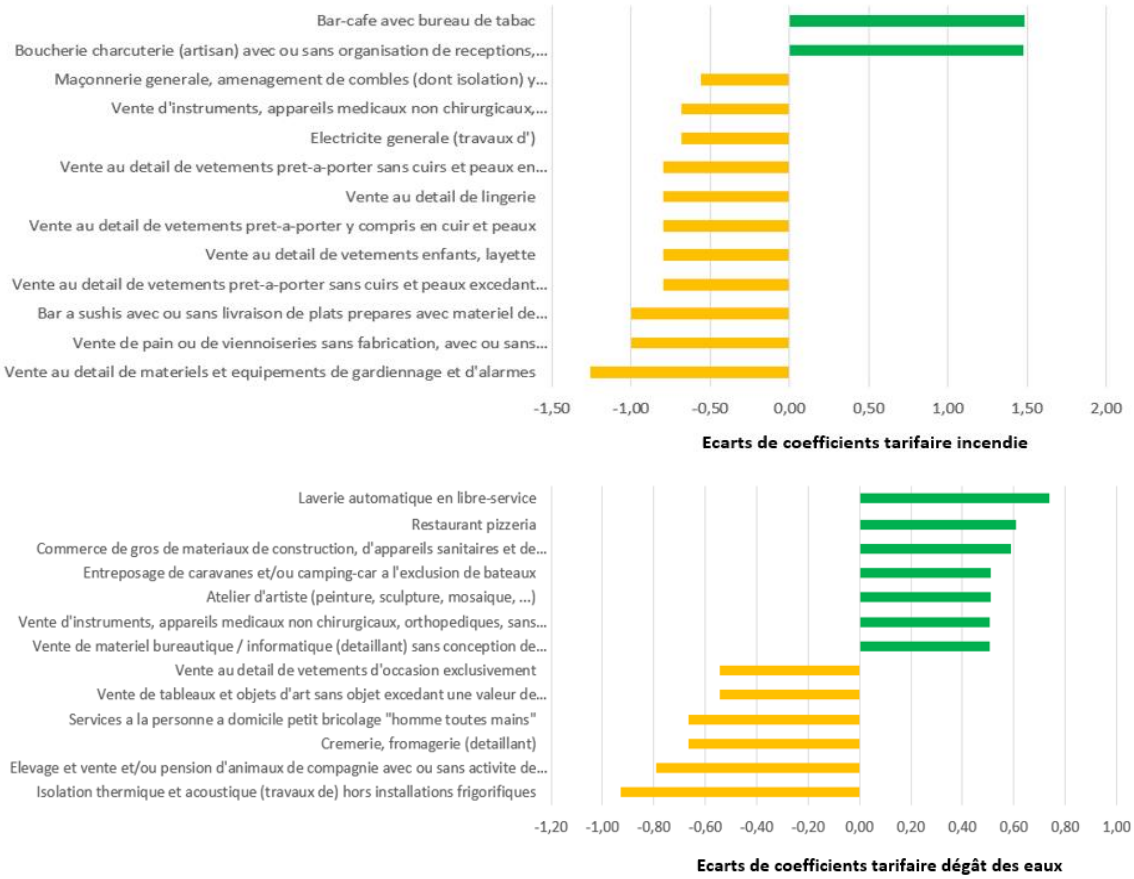


Figure 4-44 – Principaux écarts tarifaires techniques vs commercial par activités

Les écarts positifs montrent une sur-tarification par rapport à la sinistralité projetée, tandis que les écarts négatifs affichent une sous-tarification.

Les bars avec bureau de tabac et les boucheries ressortent avec une sur-tarification en incendie, de même pour les laveries automatiques et les pizzerias en dégât des eaux. Toutefois, certaines de ces activités présentant un risque plus élevé ont fait l'objet de conditions de souscriptions qui se sont durcies au cours du temps avec un impact positif sur la sinistralité.

En revanche, des activités comme la vente au détail de vêtement, les bars à sushi, la vente de pain sans fabrication et les services à la personne apparaissent sous-tarifées, et peuvent faire l'objet d'une revalorisation tarifaire lors de la révision annuelle des tarifs.

La modélisation de la prime pure permet d'apporter des ajustements du tarif commercial sur certaines activités, mais également sur les zones géographiques déterminées par le zonier.

4.3.4. La révision des zones géographiques

Les 50 zones géographiques constituées à partir du zonier (zones « techniques ») sont regroupées en 7 pour être rapprochées des 7 zones utilisées dans le tarif commercial (zones « commerciales »). Ce rapprochement est réalisé à partir des coefficients tarifaires et de l'exposition.

Les cartes ci-dessous affichent les zones commerciales et techniques incendie et dégât des eaux sur l'ensemble des communes de la France :

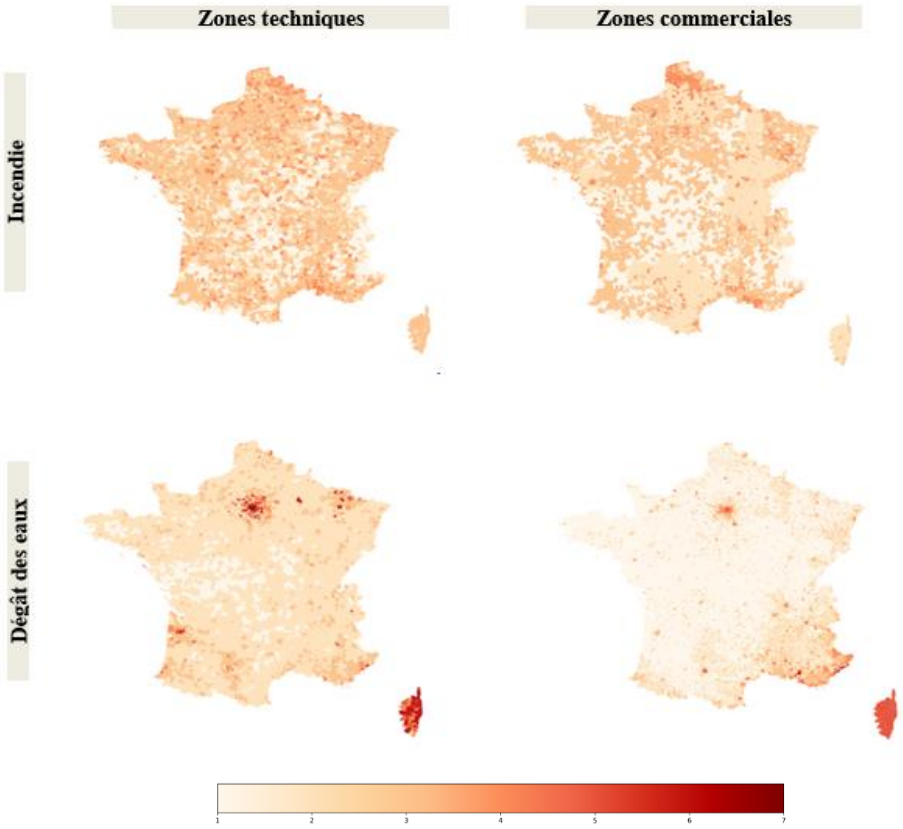


Figure 4-45 – Zones techniques et commerciales par communes

Les cartes montrent que le zonier technique a un pouvoir discriminant plus fort que le zonier commercial. Afin de quantifier les écarts entre zones techniques et commerciales, la répartition des expositions par nombre de zones en écart est représentée sur le graphique ci-dessous :

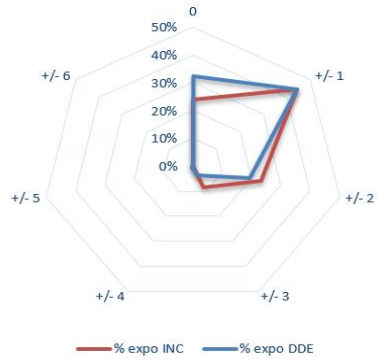


Figure 4-46 – Pourcentage d'expositions par écarts de zones techniques vs commerciales

Le graphique ci-dessus montre que 25 % des contrats ont une zone technique incendie équivalente à leur zone commerciale, et 32% en dégât des eaux. Des écarts allant jusqu'à plus ou moins 3 zones sont constatées. Pour tendre à se rapprocher des zones techniques, des réaffectations de zones sont proposées pour les communes avec des écarts de +/- 3 zones entre les visions techniques et commerciales.

La modélisation de la prime pure met en avant les différences de sinistralité par profils de risque. En complément, une règle définie grâce à un arbre de décision permet de sélectionner les profils avec une propension aux sinistres incendie atypiques plus élevée que la moyenne. La prime pure modélisée permet de déterminer des niveaux tarifaires par classes d'activité et zones géographiques utilisés pour réajuster le tarif commercial actuel.

CONCLUSION

Le but de ce mémoire était de modéliser la prime pure d'une multirisque professionnelle (MRP) dans un contexte de Covid et d'obtenir une prédiction de la sinistralité par profils de risques en fonction des périodes de restriction. L'analyse sur les deux garanties principales : l'incendie et le dégât des eaux, a permis de modéliser un risque de sévérité et un risque de fréquence et de montrer que les effets de la crise sanitaire sont différents selon les garanties.

L'enjeu était de trouver une source de données permettant de distinguer dans les modèles l'impact des différentes mesures de restriction sur l'année 2020. L'analyse des données officielles Google sur les variations de mobilité a permis de déterminer six périodes de restrictions : 2 mesures de confinement, 2 phases de déconfinement lors du 1er confinement et 2 mesures de couvre-feu. Grâce à ces données, une nouvelle variable temporelle : croisement de l'année avec ces périodes, a été utilisée pour segmenter l'exposition au risque. Par ailleurs, une variable regroupant les rubriques d'activités en fonction de l'impact des mesures de fermetures a pu être constituée. Par exemple, l'alimentation et la santé, commerces de première nécessité pouvant rester ouverts en période de confinement ont été regroupées. L'interaction dans les modèles, de ces nouvelles variables avec les autres variables a permis de différencier l'effet de la pandémie au sein de chaque model.

Toutes les étapes d'une modélisation de prime pure ont été détaillées avec un focus sur la segmentation des activités et du risque géographique.

En amont de la modélisation, la constitution de la base de données a nécessité un regroupement des activités en classes homogènes et l'écrêtement de la charge sinistre pour ne pas perturber leur distribution.

Les activités, trop nombreuses pour pouvoir être exploitées directement par le logiciel de tarification Emblem, ont été regroupées en 5 classes. Ces classes ont été déterminées par une ACP et une CAH sur la base de données caractérisant les activités : surface, chiffre d'affaires, effectif (nombre de salariés), etc. La classification a permis de distinguer les activités de petite taille, des activités avec des moyens de préventions importants, des commerces de rues, des activités présentes dans les centres commerciaux, mais surtout de faire ressortir les palaces avec restaurant qui se différencient nettement des autres activités. L'indice de Rand a montré que cette classification était stable dans le temps.

Pour distinguer les sinistres attritionnels (sinistres de fréquence) des sinistres atypiques (sinistres d'intensité), la distribution des sinistres et la théorie des valeurs extrêmes ont permis de déterminer un seuil d'écrêtement de la charge à 50 k€. Sur la garantie dégât des eaux, la sur-crête concerne 0,7 % des sinistres et représente 7,6 % de la charge. La crise sanitaire a eu un impact défavorable sur le nombre et coût moyen de sinistres atypiques. Toutefois, le faible nombre de sinistres atypiques de cette garantie ne permet pas de les modéliser, et la sur-crête a été mutualisée sur l'ensemble des assurés. La garantie incendie quant à elle a la spécificité des couvrir des risques pouvant atteindre plusieurs millions d'euros. La charge est portée par un petit nombre de sinistres : 7,3 % des sinistres couvrent 73,4 % de la charge. Ce choix de seuil d'écrêtement a permis de modéliser les sinistres incendie atypiques.

La modélisation de la prime pure est dissociée en deux parties : les modèles de sinistres attritionnels et les modèles de sinistres atypiques.

Pour modéliser les sinistres attritionnels, les modèles GLMs de fréquence (loi de Poisson) et de coût moyen (loi Gamma) ont été réalisés en deux temps : avant et après zonier.

Dans un premier temps, les variables explicatives ont été identifiées, avec quelques variables se comportant différemment selon les périodes de restrictions : la surface pour la fréquence incendie, la franchise dégradable et l'option complément plus pour la fréquence dégât des eaux, l'option cuisine ouverte sur salle, l'offre et l'option zone à risque pour le coût moyen dégât des eaux.

Parmi les variables explicatives retenues, certaines sont des variables géographiques, captant ainsi l'effet géographique connu. Pour chacun des modèles, l'analyse des résidus par codes postaux a permis par lissage,

après neutralisation de l'effet non-géographique, de capter l'effet géographique inconnu. Une méthode de lissage local qui prend en compte le risque des codes postaux immédiatement voisins, c'est-à-dire adjacents, a été appliquée. L'effet géographique résiduel a été ensuite combiné à l'effet géographique connu issu du GLM. L'effet géographique total ainsi constitué par codes postaux a été regroupé en 50 zones géographiques homogènes dans le but de créer une nouvelle variable « zone technique » pour la fréquence et le coût moyen. Cette variable a été introduite dans les modèles GLMs initiaux en remplacement des variables géographiques. L'impact de l'intégration du zonier sur le pouvoir prédictif des modèles a amené à ne pas retenir de zonier dans le modèle coût moyen incendie.

Les sinistres incendie atypiques ont été modélisés par un modèle de propension (pourcentage de sinistres incendie atypiques parmi les sinistres incendie), et un deuxième niveau d'écrêtement a été appliqué afin de pouvoir modéliser une partie du coût moyen.

Dans le but d'identifier au mieux les profils de risque, deux méthodes de régression logistique ont été testée pour modéliser la propension : une méthode classique de type GLM et une méthode de *Machine Learning* de type *Gradient boosting : LightGBM*. L'amélioration de 6 points du coefficient de Gini et les variables retenus avec le *LightGBM* ont conclu à sélectionner ce modèle pour estimer la propension des sinistres incendie atypiques.

Pour modéliser le coût moyen des sinistres incendie atypiques, un second seuil d'écrêtement a été déterminé à 750 k€, permettant de modéliser la partie de la sur-crête comprise entre 50 k€ et 750 k€ par une loi Gamma. Afin de rester cohérent avec le modèle de propension, un modèle *LightGBM* a été retenu pour estimer le coût moyen des sinistres incendie atypiques. La partie supérieure à 750 k€ a été mutualisée sur l'ensemble des sinistres incendie atypiques.

Sur chacune des garanties étudiées, l'ensemble des modèles ont été consolidés pour avoir une estimation de la prime pure. Pour illustrer les résultats, l'estimation de la prime pure a été comparée à la sinistralité observée par périodes et sur un des principaux critères en MRP : la surface. Les tendances par périodes de restrictions ont bien été captées par les modèles. En incendie, la sinistralité a tendance à baisser avec les restrictions et à augmenter en période de déconfinement avec la reprise de l'activité. En dégât des eaux, les périodes de confinement et de déconfinement ont un impact négatif sur la sinistralité tandis que les périodes de couvre-feu ne semblent pas avoir de répercussions. Sur certains profils de risque, notamment ceux avec une exposition très faible, la prime pure estimée est ressortie très inférieure à l'observée du fait de la mutualisation d'une partie des sinistres atypiques. Une analyse par un arbre de décision a permis de générer une règle de sélection des profils les plus exposés aux sinistres atypiques. Cette règle est destinée aux souscripteurs comme aide à la sélection des risques. Par ailleurs, les résultats de ces travaux ont conduit à faire des propositions d'évolution du tarif commercial au niveau des activités et des zones géographiques.

Le cadre de ce mémoire s'arrête à l'estimation de la prime pure. La prochaine étape est le calcul du tarif avec l'ajout des chargements mais surtout son calibrage. L'intégration de données liées à la Covid dans les modèles de prime pure a permis d'estimer une sinistralité selon les différents types de restriction : confinement, déconfinement, couvre-feu. Grâce à ces distinctions, le tarif peut être calibré en fonction d'hypothèses sur la situation de la pandémie pour les années à venir.

La limite de ce mémoire est le manque de recul sur l'évolution de la crise sanitaire au moment où l'étude a été réalisée. Sur les cinq années d'historique retenues, seule l'année la plus récente était impactée par des restrictions, ce qui ne lui donne que peu de poids par rapport à un historique sans Covid. En 2022, la situation est loin d'être terminée et continue à évoluer. Une mise à jour des modèles avec plus de recul et l'ajout d'années supplémentaires touchées par la crise sanitaire est nécessaire pour généraliser les effets sur la sinistralité des changements de comportements liés aux restrictions.

Bibliographie

ABADIE A. (2021) Dommages aux professionnels : les réseaux se préparent à l'après-crise. L'argus de l'assurance.

BARADEL N. (2020) Théorie du risque. Support de cours.

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

KARAYAN R. (2020) Multirisque professionnelle : les bancassureurs accélèrent sur les pros et l'entreprise. L'argus de l'assurance.

MANDOT P. (2017) *What is LightGBM, How to implement it? How to fine tune the parameters?*

OHLSSON E., JOHANSSON B. (2010) *Non-life Insurance Pricing with Generalized Linear Models*. Berlin: Springer.

POULLENNEC S. (2020) Hôteliers et restaurateurs promeuvent une assurance, mais qui ne couvre pas le Covid. Les Echos.

ROBERT T. (2020) Gel des primes d'assurance : les agents généraux et courtiers doivent réorganiser les contrats. Assurlandpro.

Documentation Allianz France sur la théorie des zoniers.

https://www.ecdc.europa.eu/sites/default/files/document/response_graphs_data_2021-01-14.csv (site du Centre européen de prévention et de contrôle des maladies), consulté le 20 novembre 2020.

https://www.franceassureurs.fr/wp-content/uploads/VF-Donnees-cles-2020.pdf (site de France Assureurs, données clés 2020 de l'assurance française), consulté le 18 janvier 2022.

https://www.google.com/covid19/mobility/ (rapport sur l'impact de la COVID-19 sur les déplacements des habitants), consulté le 3 février 2021.

https://www.insee.fr/fr/statistiques/ (site de l'Insee, statistiques et études sur les commerces de détail), consulté le 7 juin 2021.

https://openclassrooms.com/fr/courses/4011851-initiez-vous-au-machine-learning/4020611-identifiez-les-differents-types-dapprentissage-automatique (site de cours en ligne), consulté le 10 février 2021.

Annexes

Annexe1-Source des données géographiques

Type de données		Lien	Date de collecte	Date de publication	Granularité
Socio Démographique	Activités des résidents	https://www.insee.fr/fr/statistiques/4799323	2017	19/10/2020	INSEE
	Couples - Familles - Ménages	https://www.insee.fr/fr/statistiques/4799268	2017	19/10/2020	INSEE
	Diplômes - Formation	https://www.insee.fr/fr/statistiques/4799252	2017	19/10/2020	INSEE
	Population	https://www.insee.fr/fr/statistiques/4799309	2017	19/10/2020	INSEE
	Logement	https://www.insee.fr/fr/statistiques/4799305	2017	03/11/2020	INSEE
Revenu	Indicateurs sur les revenus et la pauvreté	https://www.insee.fr/fr/statistiques/4507225	2017	16/06/2020	INSEE
	Impôt 2019 (revenu 2018)	https://www.data.gouv.fr/fr/datasets/l-impot-sur-le-revenu-par-collectivite-territoriale/	2018	01/10/2020	INSEE
Commune	Aires urbaines	https://www.insee.fr/fr/information/2115011	2017	26/02/2020	INSEE
Criminalité	Crimes et délits enregistrés par la gendarmerie et la police	https://www.data.gouv.fr/fr/datasets/crimes-et-delits-enregistres-par-les-services-de-gendarmerie-et-de-police-depuis-2012/	2018 à 2020	29/01/2020	DEPT
Occupation des terres	Répartition des communes en termes d'occupation des terres	https://www.statistiques.developpement-durable.gouv.fr/corine-land-cover-0	2018	27/12/2018	INSEE
Prix de l'immobilier	Prix moyen des ventes de maisons et d'appartements	https://www.data.gouv.fr/fr/datasets/prix-moyen-au-m2-des-ventes-de-maisons-et-dappartements-par-commune-en-2019/	2019	01/01/2019	INSEE
	Coût de la reconstruction au m²	Callon	2018		DEPT
Pompiers & Gendarmerie	Interventions effectuées par les services d'incendie et de secours	https://www.data.gouv.fr/fr/datasets/interventions-realisees-par-les-services-d-incendie-et-de-secours/#_	2019	21/10/2020	DEPT
	Adresse de la caserne des pompiers (et du poste de police)	OpenStreetMap	2020	31/12/2020	XY
Climatique	Score (sécheresse, inondation) & zonier (tempête, grêle, vent)	Allianz : équipe des risques climatiques			

Annexe2 -Tendances des variables explicatives de la fréquence des sinistres attritionnels

La Classe d'activité

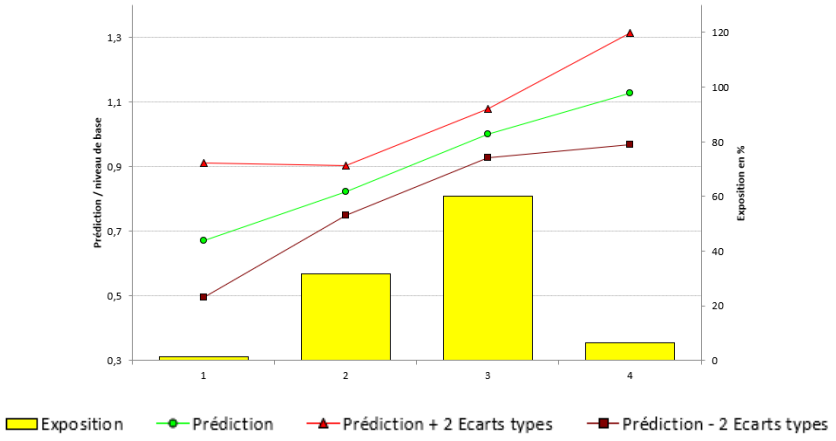


Figure A-1 - Fréquence DDE par classe d'activité

Le Groupe d'activité

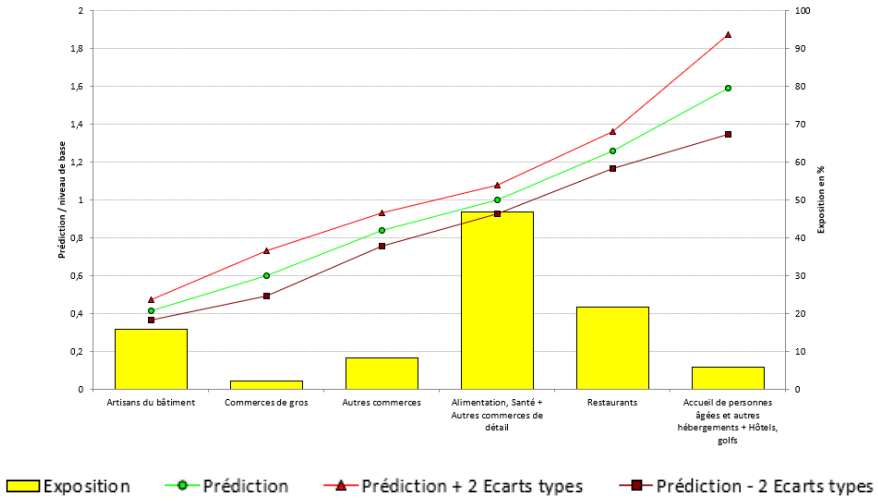


Figure A-2 - Fréquence DDE par groupe d'activités

Les antécédents sinistres

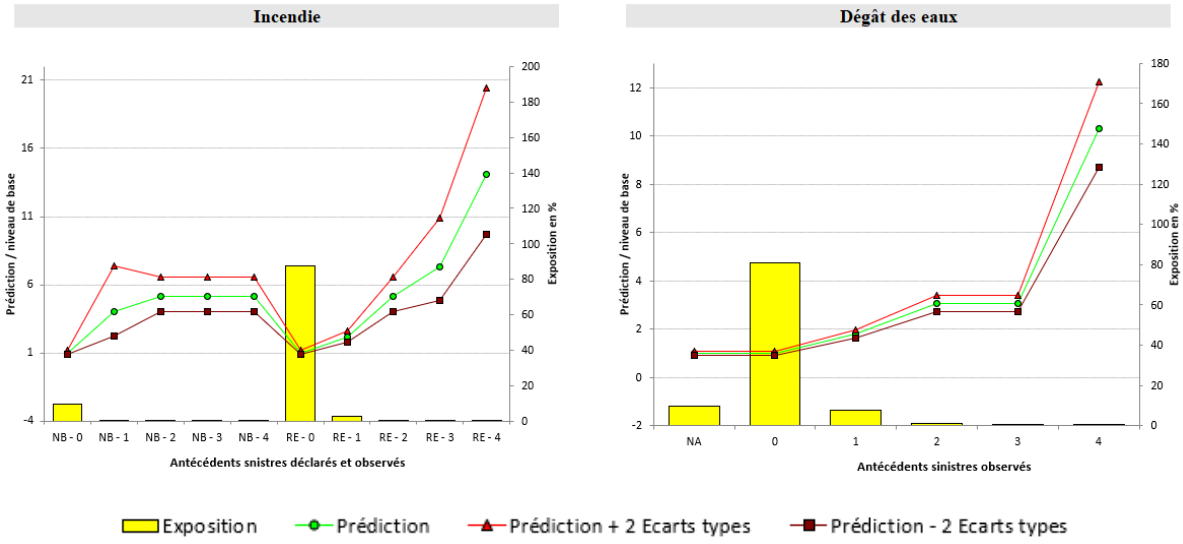


Figure A-3 - Fréquence par antécédents sinistres

Les moyens de prévention : extincteurs, local situé dans un centre commercial équipé de sprinklers

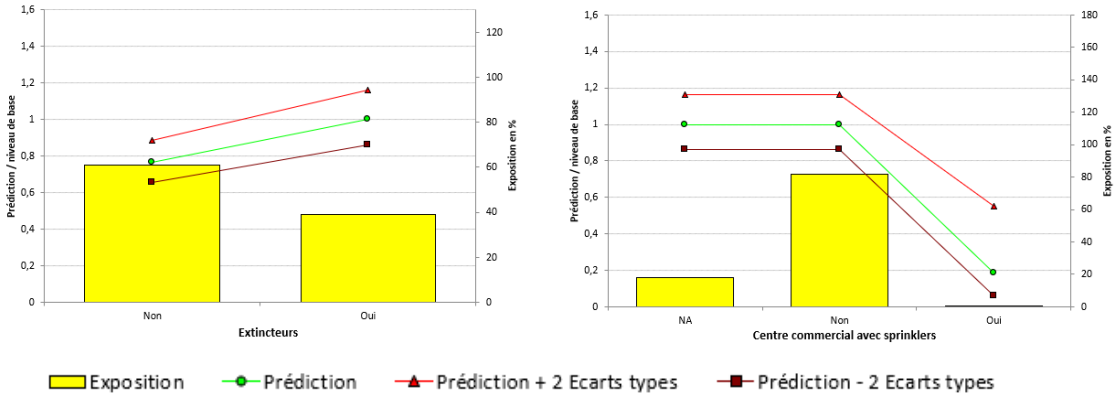


Figure A-4 - Fréquence incendie par présence d'extincteurs et de centre commercial avec sprinklers

La qualité juridique

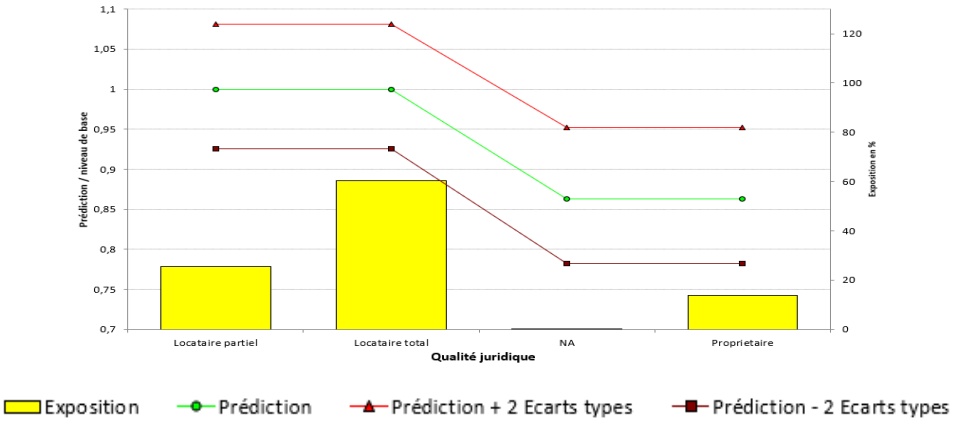


Figure A-5 - Fréquence DDE par qualité juridique

Client avec un contrat en surveillance

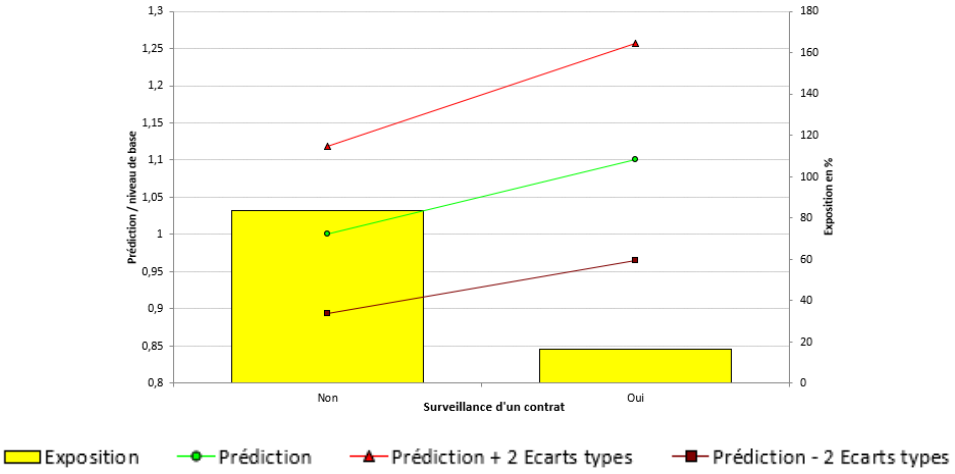


Figure A-6 - Fréquence incendie selon la présence d'un contrat en surveillance

La note de risque de faillite

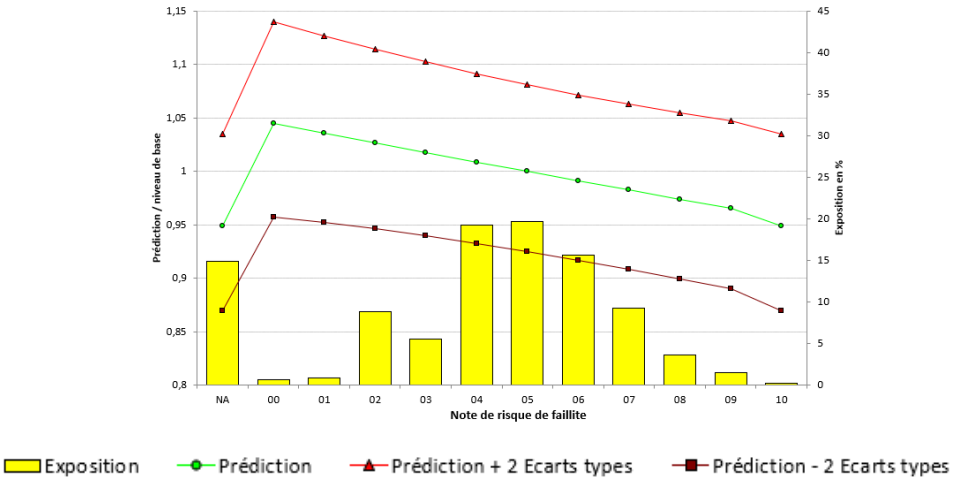


Figure A-7 - Fréquence dégât des eaux par note de risque de faillite

L'ancienneté du contrat

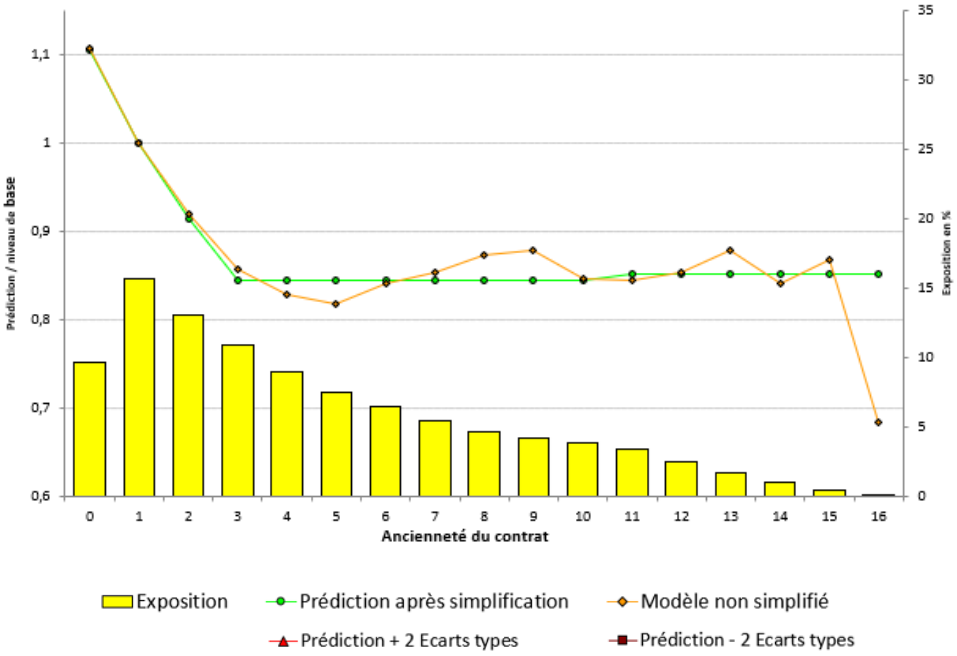


Figure A-8 - Fréquence dégât des eaux par ancienneté du contrat

Annexe3 -Tendance des variables explicatives du modèle de propension incendie

L'offre

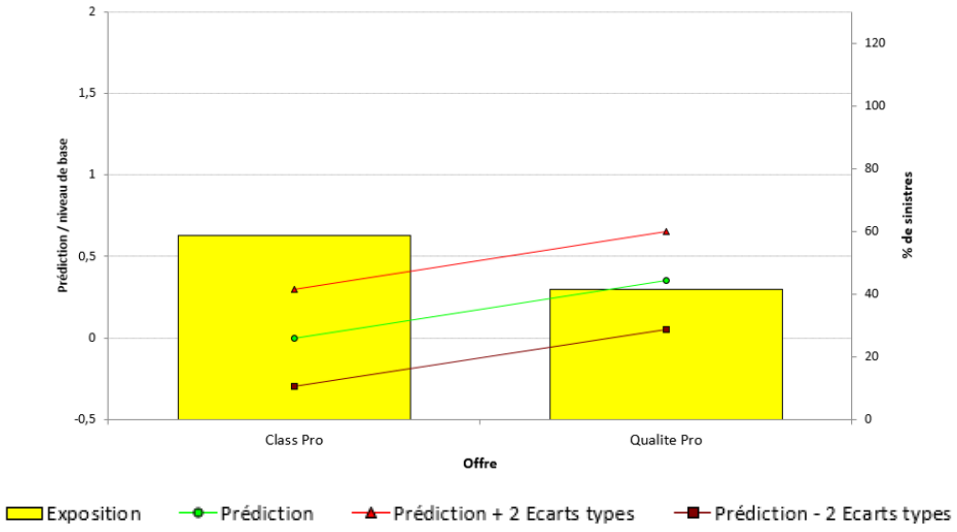


Figure A-9 - Prédiction des sinistres incendie atypiques par type d'offre

La qualité juridique

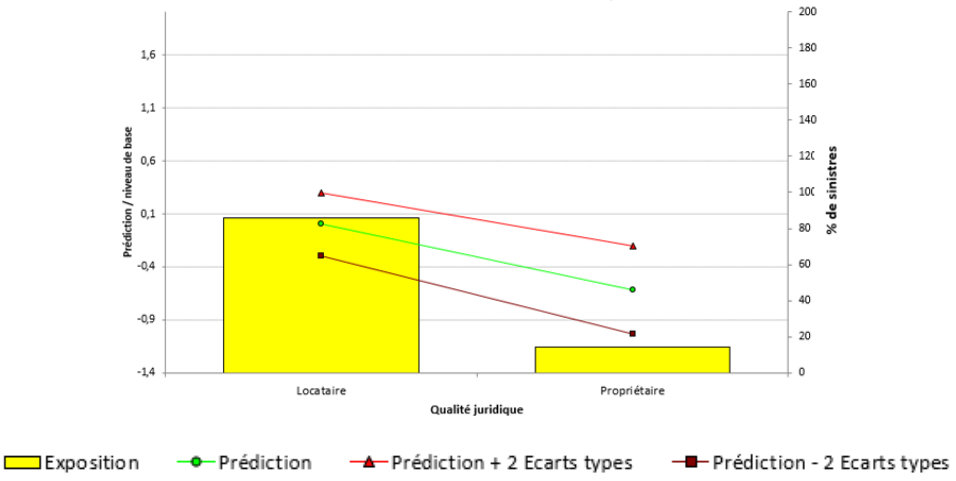


Figure A-10- Prédiction des sinistres incendie atypiques par qualité juridique

La superficie

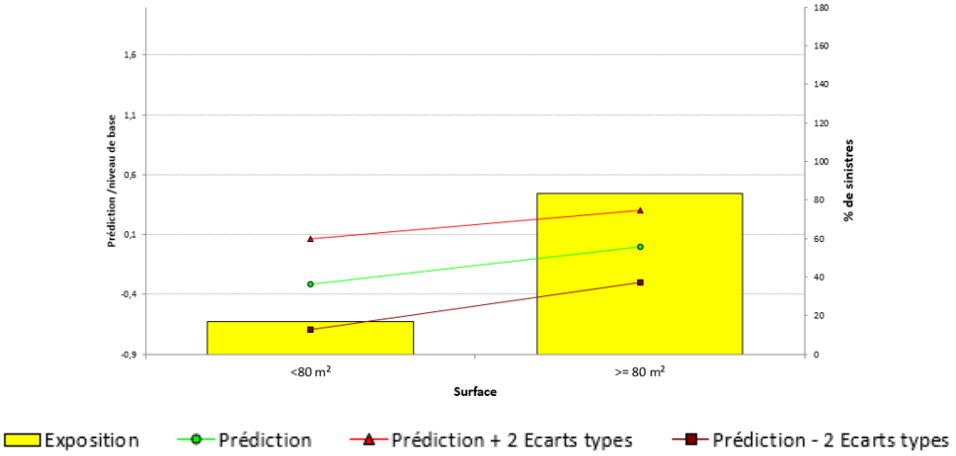


Figure A-11– Prédiction des sinistres incendie atypiques par superficie

La note de risque de faillite

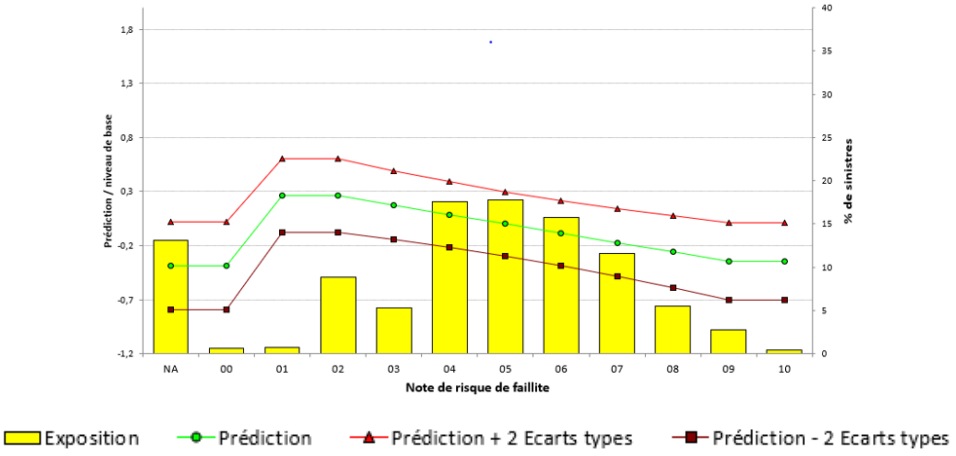


Figure A-12 – Prédiction des sinistres incendie atypiques par note de risque de faillite