Institut des ACTUAIRES

IRM | institut du risk management

Mémoire présenté devant l'Institut du Risk Management
pour la validation du cursus à la formation d'Actuaire
de l'Institut du Risk Management
et l'admission à l'Institut des Actuaires
le

Par : **Sébastien PERRIN**
Titre : **Refonte des ELR sur la garantie Dégât des eaux pour le produit Habitation**

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

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

Membres présents du jury de l'Institut des Actuaires

Entreprise :Axa France
Nom : François Luu
Signature :

Membres présents du jury de l'Institut du Risk Management

Directeur de mémoire en entreprise :

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)

Secrétariat :

Bibliothèque :

Signature du responsable entreprise

Signature du candidat

Résumé

Le marché de l'assurance Habitation étant en constante évolution en termes de compétitivité et de besoins clients, les assureurs se doivent de proposer un tarif de plus en plus fin et adapté aux risques des assurés. L'objectif de ce mémoire est d'effectuer une refonte de la modélisation du risque principal en assurance Habitation sur le périmètre des appartements, qui correspond au dégât des eaux. Cela, afin d'obtenir ainsi un tarif plus compétitif tout en effectuant un compromis entre la mutualisation (répartition du coût d'un sinistre au sein d'une classe la plus homogène possible) et la segmentation visant à minimiser l'anti-sélection des risques.

En ce sens, une fois la phase de traitement des données effectuée (constitution de la base de modélisation, définition du seuil de sinistres graves, vieillissement des sinistres), cette étude se déroulera en trois étapes principales.

Tout d'abord, la prime pure sera tarifée en utilisant deux approches mathématiques différentes : les modèles linéaires généralisés sous leurs déclinaisons pénalisées (Lasso, Ridge et Elastic Net) ainsi que des méthodes d'agrégation d'arbres (*Random Forest* et *Gradient Boosting*).

Puis ensuite sera pris en compte le signal porté par la segmentation géographique via la création de zoniers à différentes mailles. Cette étape est nécessaire au perfectionnement de la tarification en prime pure en assurance Habitation. Ainsi, il peut être intéressant de tenter de capter la contribution géographique du risque par un zonier, défini comme la correspondance entre une zone et un coefficient tarifaire.

Enfin, un suivi de la rentabilité selon le type de contrats (affaires nouvelles ou non) par segments commerciaux sera effectué en utilisant différents indicateurs.

Mots-clés : *MultiRisques Habitation, Dégât des eaux, Loi de Pareto, Prime pure, Modèles linéaires généralisés, Pénalisations, Random Forest, Gradient Boosting, Zonier, Lissage spatial, Crédibilité, Rentabilité.*

Abstract

The household insurance market is constantly evolving in terms of competitiveness and customer needs, so insurers must offer an increasingly fine price and adapted to the risks of policyholders. As part of the continuous updating of the household product, the aim of this dissertation is to propose a price revision for the main guarantee, on the perimeter of the apartments, in terms of the number of claims : the water damage guarantee. This, in order to obtain a more competitive price while making a compromise between pooling (distribution of the cost of a claim within a class that is as homogeneous as possible) and segmentation aimed at minimizing the anti-selection of risks.

In this sense, once the data processing phase has been completed (constitution of the modeling base, definition of the threshold for large claims, aging of claims), this study will be carried out in several main parts.

First, the pure premium will be priced using penalized generalized linear models (Lasso, Ridge and Elastic Net) as well as Machine Learning (Random Forest and Gradient Boosting).

Then, the signal carried by the geographical segmentation via the creation of zones with different levels will be taken into account. Taking geographical risk into account is one of the steps necessary to perfect pure premium pricing. Thus, it can be interesting to model the geographical contribution of the risk by a zone manager, defined as the correspondence between a zone and a tariff coefficient.

Finally, profitability monitoring according to the type of contracts (new business or not) by commercial segment will be carried out using different indicators.

Keywords : *Household insurance, Water damage, Pareto distribution, Pure premium, Penalized Generalized linear models, Random Forest, Gradient Boosting, Zoning, Spatial smoothing, Credibility, Profitability.*

Remerciements

Parmi les personnes ayant permis la réalisation de ce mémoire, je souhaite tout d'abord remercier Romain TOESCA pour son encadrement régulier, sa disponibilité pour répondre à mes interrogations et ses nombreux conseils qui ont alimenté ma réflexion.

Je tiens également à remercier Alexis BERNANOSE pour l'aide précieuse et les riches échanges qu'il a pu m'apporter tout au long de cette rédaction.

Mes remerciements également à Mohamed HALIMI, qui a eu la patience de partager ses connaissances techniques tout au long de la formation du CEA, et à Charles PARTINGTON pour sa disponibilité et sa bienveillance au quotidien.

Table des matières

Introduction

Ce mémoire a été réalisé au sein de l'équipe Actuariat Tarification Habitation dans le cadre d'une refonte tarifaire du produit d'assurance **M**ulti**R**isques **H**abitation (dite MRH). Cette assurance est destinée à couvrir l'habitation et son contenu mais également la responsabilité civile des occupants envers un tiers. Les principales garanties sont le bris des vitres, les catastrophes naturelles, les évènements climatiques, le dégât des eaux, l'incendie ou encore le vol. Le marché de l'assurance Habitation est très concurrentiel et ce phénomène s'est intensifié avec l'arrivée de nouveaux acteurs tels que les bancassureurs et la mise en place de la loi Hamon en janvier 2015. Désormais, grâce à cette loi, les assurés ont la possibilité de résilier un contrat d'assurance à partir de l'échéance de la première année, et de souscrire auprès d'un assureur concurrent afin d'obtenir un tarif plus attractif. Cela a impacté fortement la dynamique de production du produit MRH.

Dans ce contexte, il n'est pas possible pour l'assureur de décider d'une hausse globale de la prime des clients du portefeuille. Afin de ne pas dégrader les résultats, mais également afin de maintenir la compétitivité de l'entreprise, le tarif du produit doit être revu régulièrement via une mise à jour des données de modélisation et l'application de nouvelles techniques actuarielles de tarification et de segmentation. Cette segmentation consiste à considérer que le risque fluctue selon le contrat, ce qui implique une prime d'assurance différente. C'est ainsi un moyen efficace de lutter contre le phénomène d'anti-sélection.

Le produit d'assurance Habitation étudié est basé sur une approche de modélisation de prime pure (prime requise pour faire face à la sinistralité espérée du portefeuille), effectuée lors de son lancement en 2016 avec une distinction entre la fréquence et le coût moyen sur les principales garanties de couverture. Cette revue des modèles permet de créer un tarif dissociant la prime technique (prime pure sécurisée par des chargements additifs et multiplicatifs permettant de faire face à la charge de sinistre des assurés et aux différents frais) de la prime commerciale (prime proposée au client permettant d'assurer un minimum de rentabilité à la compagnie d'assurances tout en garantissant un certain niveau de conversion).

Dans cette étude, **l'objectif est de revoir la prime pure de la garantie Dégât des eaux**, sur le périmètre des appartements. L'enjeu est d'obtenir la vision la plus juste possible du risque du portefeuille d'assurés. La première partie présentera les caractéristiques du marché MRH ainsi que les principes de la tarification. Dans une seconde partie, la base de modélisation résultant de la fusion de différentes sources d'information sera présentée. Dans la continuité, une description détaillée des actions effectuées sur les données, telles que l'écrêtement et le vieillissement des sinistres sera également présentée.

Dans une troisième partie, le risque sera modélisé en utilisant les modèles linéaires généralisés, et en particulier leurs déclinaisons pénalisées. Ces premiers résultats seront comparés à ceux obtenus via des méthodes d'apprentissage statistique que sont le *Gradient Boosting Machine* et le *Random Forest*.

Le modèle linéaire généralisé est, en effet, le modèle le plus utilisé en assurances Dommages, mais une approche *Gradient Boosting* peut représenter une bonne alternative car il permet de faire face à la quantité massive de variables disponibles pour réaliser l'étude.

Puis ensuite sera abordée la thématique du zonier. Il existe des variables qui traduisent l'environnement géographique où évolue le contrat. En effet, cet aspect du risque est très important en assurance Habitation et il est primordial de le prendre en compte dans le calcul du tarif. Pour cela, une nouvelle méthodologie de zonier, pour chacune des composantes de la prime pure, sera présentée à partir d'une approche résiduelle et une fois les zones de risque définies à l'aide d'outils de classification, celles-ci seront lissées en utilisant une technique basée sur la théorie de la crédibilité.

Enfin, pour clore l'étude, un suivi de la rentabilité par segments commerciaux sera effectué, en distinguant les affaires nouvelles du reste du portefeuille, et permettra de voir les potentielles actions à mener sur la prime commerciale pour améliorer la marge générée.

— Chapitre 1 —

Objectifs de l'étude et données utilisées

1.1 L'Assurance MultiRisques Habitation

1.1.1 Présentation du marché IARD

Le marché IARD - **I**ncendie, **A**ccidents et **R**isques **D**ivers - couvre les dommages que peuvent subir les biens des clients. Deux grands types de couverture sont présents sur ce marché : l'assurance Automobile et l'assurance Habitation. Ils représentent près de 60% des cotisations perçues au titre de l'assurance IARD. A titre d'illustration, en 2020, le montant perçu sur les contrats Habitation s'élevait à 11,7 milliards d'euros [1].

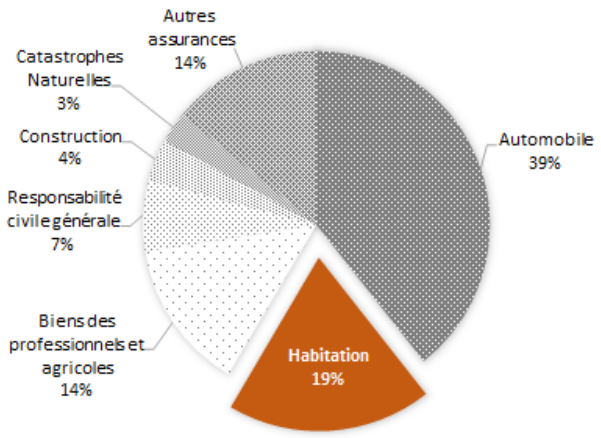


FIGURE 1.1 – Poids des différents types d'assurance en IARD - 2020

1. Chiffres issus de la synthèse des données clés 2020 publiée par France Assureurs

Les principales garanties proposées dans un produit Habitation sont les suivantes :

- Assistance ;
- Attentats et actes de terrorisme ;
- Bris de vitres ;
- Dégât des eaux ;
- Évènements climatiques et Catastrophes naturelles : la garantie Évènements climatiques couvre les dommages causés par la tempête, la neige, la grêle et l'inondation. Lorsqu'un arrêté ministériel est publié au Journal Officiel suite à un phénomène naturel, la garantie Catastrophes naturelles est alors appliquée ;
- Incendie ;
- RC (**R**esponsabilité **C**ivile) : la garantie RC couvre les dommages causés par l'assuré, ses proches ou ses animaux dans le cadre de sa vie privée ;
- Vol.

Après la présentation du marché de l'assurance Habitation, une attention particulière sera effectuée sur l'importance d'AXA sur ce segment de marché.

1.1.2 L'offre AXA

L'étude est basée sur les deux produits d'assurance Habitation les plus représentés du portefeuille. Le socle de garanties inclues dans ces produits correspond aux garanties principales en MRH mentionnées précédemment. A cela peuvent s'ajouter des options telles que la casse des appareils nomades, les dommages aux appareils électriques ou encore la protection juridique.

L'offre d'AXA est proposée à partir des canaux de distribution traditionnels (agents généraux et courtiers), mais également depuis 2017 sur le Web. La présence d'AXA sur ce nouveau réseau de distribution est nécessaire pour s'adapter au comportement des clients qui sont de plus en plus présents sur Internet et qui comparent davantage les offres. Cette concurrence tarifaire accrue s'illustre par l'évolution à la baisse du portefeuille. Entre 2014 et 2020, bien que le chiffre d'affaires soit resté stable autour du milliard d'euros, le nombre de contrats a, quant à lui, diminué de près de 10% malgré une hausse du nombre de logements de 7,3% sur la même période[2].

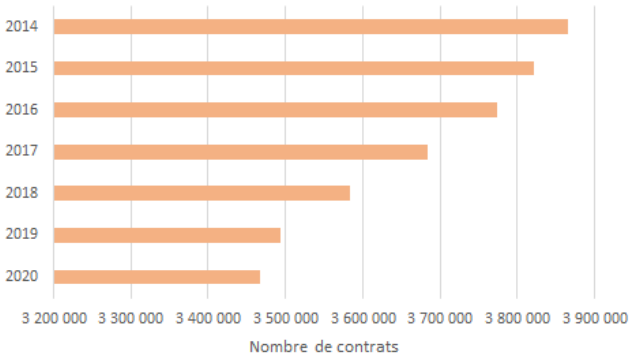


FIGURE 1.2 – Évolution du nombre de contrats en portefeuille entre 2014 et 2020

La segmentation des risques demeure donc primordiale pour AXA afin d'améliorer l'adéquation du tarif à chaque profil de risque pour ainsi conserver les assurés présents en portefeuille mais également pour attirer de nouveaux clients.

2. Source : INSEE (**I**nstitut **N**ational de la **S**tatistique et des **É**tudes **É**conomiques), INSEE focus numéro 217 paru le 08/12/2020

1.2 Généralités sur les principes de tarification

La tarification d'un produit d'assurance est un processus complexe nécessitant de respecter un certain nombre de contraintes. Dans cette section, les grandes étapes de la conception d'un tarif d'assurance, les contraintes présentes dans la tarification ainsi que les composantes de la prime commerciale vont être présentées.

1.2.1 Principes de segmentation et de mutualisation

Le concept de la tarification s'appuie sur la segmentation qui consiste à séparer les assurés selon différents critères de façon à ce que les risques au sein de chaque classe soient les plus homogènes possibles. Et ce afin de lutter contre le phénomène d'anti-sélection. Si le même tarif était appliqué à l'ensemble des clients, les profils les plus risqués décideraient de s'assurer, contrairement aux risques les plus faibles qui auraient une prime trop élevée et décideraient de s'engager auprès de la concurrence. Cela aurait pour conséquence une dégradation rapide du résultat. L'exemple ci-dessous permet d'illustrer la nécessité de segmenter le portefeuille d'assurés.

Soient deux assureurs ayant les caractéristiques suivantes :

- L'assureur X décide de ne pas segmenter son portefeuille ;
- L'assureur Y segmente son portefeuille en fonction de la zone de risque uniquement.

Chaque assureur propose un contrat avec les mêmes garanties dont le tarif est le suivant :

	Zone de risque élevée	Zone de risque faible
Assureur X	500 €	500 €
Assureur Y	700 €	300 €

TABLE 1.1 – Primes proposées par les assureurs X et Y pour deux profils de risque

L'assureur X propose la même prime d'assurance quelle que soit la zone de risque. Ainsi, il fait un profit avec les clients faiblement risqués et une perte avec les autres assurés. La prime perçue par les bons risques sert, en partie, à couvrir les sinistres des mauvais risques. L'assureur Y propose, quant à lui, deux tarifs différents en fonction du profil de risque. Ainsi, les profils risqués auront une prime supérieure du fait de leur exposition au risque. Les conséquences de la présence de segmentation dans le portefeuille sont les suivantes :

- Les assurés étant dans une zone de risque faible décideront de s'assurer auprès de l'assureur Y ;
- L'assureur X, proposant un tarif unique, concentrera la plupart des assurés étant dans une zone de risque élevée.

La mise en concurrence de ces deux assureurs entraîne des flux d'assurés d'une compagnie à l'autre. L'assureur choisissant de ne pas segmenter son portefeuille se retrouvera dans une position qui pourrait conduire à la ruine. Cependant, la segmentation présente des limites. En effet, lorsque celle-ci est trop fine, c'est-à-dire lorsque le tarif tend à devenir individualisé, l'estimation de la charge moyenne de sinistre est alors biaisée conduisant à de mauvais résultats pour l'assureur. Il convient donc de rester prudent et de ne pas segmenter à l'extrême. Le principe de mutualisation s'applique alors au sein de chacune des classes.

La mutualisation des risques consiste à répartir le coût d'un sinistre au sein d'une classe la plus homogène possible. Néanmoins, ce principe repose sur le fait que les sinistres survenus soient iid (**i**ndépendants et **i**dentiquement **d**istribués).

Pour qu'un assureur ne soit pas face à une situation de perte, l'inégalité suivante doit être vérifiée :

$$\sum_i P_i \geq \sum_i \mathbb{E}(X_i) + F_i. \tag{1.1}$$

avec P_i la prime versée par l'assuré i, X_i le coût d'un sinistre pour ce même assuré i et F_i les frais de l'assureur liés à l'assuré i.

1.2.2 Différents types de prime

1.2.2.1 Définition et construction de la prime pure

Pour aboutir à la prime effectivement payée par le client, il existe plusieurs étapes dont la première consiste à calculer l'espérance du risque, appelée **prime pure**. L'inversion du cycle de production en assurance impose à l'assureur de déterminer au préalable cette prime pure qui permet de couvrir totalement la charge de sinistres, pour une police donnée sur une période d'assurance donnée. Pour modéliser la sinistralité, le modèle collectif va être utilisé pour distinguer la fréquence de sinistres par assuré d'une part et le coût des sinistres d'autre part (*Sauveplane, 2019*, **[10]**). Cette séparation n'est pas que mathématique. En effet, le nombre de sinistres (et donc la fréquence associée) est essentiellement lié au comportement de l'assuré tandis que le coût est plutôt expliqué par les caractéristiques du bien. Il y a de plus un décalage structurel entre les deux notions : le coût n'est connu qu'au terme du développement final du sinistre alors que la survenance est connue dès la déclaration à l'assureur.

La détermination de la prime pure par l'approche des modèles de fréquence et de coût moyen consiste à estimer l'espérance de la valeur totale des sinistres survenus au cours de l'exercice considéré pour l'assuré i, notée $\mathbb{E}(X_i)$. La variable aléatoire X_i peut alors être décomposée de la manière suivante :

$$X_i = \sum_{j=1}^{N_i} C_{i,j}, \tag{1.2}$$

avec N_i le nombre de sinistres survenus et $C_{i,j}$ le coût du sinistre j de l'assuré i.

Pour estimer $\mathbb{E}(X_i)$, la fréquence de sinistres et le coût moyen sont définis :

- La fréquence de sinistres de l'assuré i correspond au ratio :

$$f_i = \frac{\text{Nombre de sinistres}}{\text{Exposition}}. \tag{1.3}$$

  L'exposition (appelée aussi année-police) désigne la durée durant laquelle le risque est couvert. Par exemple, un contrat souscrit le 1^{er} juillet de l'année N et toujours en cours au 31 décembre N a une exposition égale à 0,5 année-police. Si un sinistre survenait sur ce contrat au cours de cette période de couverture, la fréquence observée serait de 200% car il s'agit d'une moyenne annualisée : un sinistre sur six mois équivaut, en terme de risque, à en moyenne deux sinistres par an.

- Le coût moyen de l'assuré i correspond au ratio :

$$c_i = \frac{\text{Charge totale des sinistres}}{\text{Nombre de sinistres}}. \tag{1.4}$$

Cette approche fréquence-coût repose sur deux hypothèses :

- Les charges des sinistres individuels sont des variables aléatoires iid ;
- Le nombre total des sinistres est indépendant de leur coût.

La première hypothèse est intuitivement vérifiée par le fait que le montant d'un sinistre n'a pas lieu de dépendre du montant d'un sinistre précédent. Pour valider la seconde hypothèse, il faut étudier la dépendance entre la fréquence et le coût afin d'observer une éventuelle corrélation.

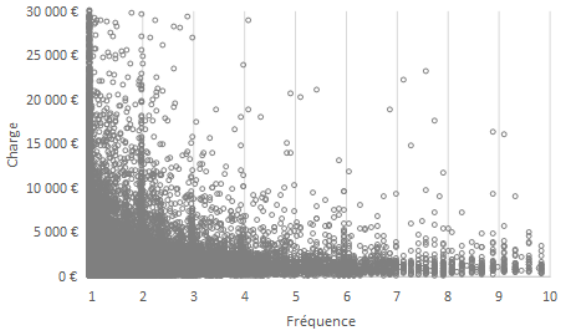


FIGURE 1.3 – Représentation du croisement de la fréquence et de la charge des sinistres

La répartition des points, observée sur la figure 1.3, est uniforme et aucune tendance ne se dégage dans la forme du nuage de points. Pour évaluer de manière plus rigoureuse le lien entre ces deux variables, plusieurs mesures de dépendance vont être calculées :

- Le coefficient de corrélation de Pearson est défini comme :
$$r(X_1, X_2) = \frac{\mathbb{E}(X_1, X_2) - \mathbb{E}(X_1)\mathbb{E}(X_2)}{\sqrt{\mathbb{V}(X_1)\mathbb{V}(X_2)}}. \tag{1.5}$$
Plus la valeur de $r(X_1, X_2)$ est proche de 0, plus les variables X_1 et X_2 sont indépendantes.

- Le Tau de Kendall repose sur la notion de paires discordantes et concordantes : deux paires d'observations (x_i, y_i) et (x_j, y_j) sont concordantes si $(x_i - x_j)(y_i - y_j) > 0$, discordantes si $(x_i - x_j)(y_i - y_j) < 0$. Cet indicateur est défini comme :
$$\tau(X_1, X_2) = \mathbb{P}[(X_1 - X_1^{'})(X_2 - X_2^{'}) > 0] - \mathbb{P}[(X_1 - X_1^{'})(X_2 - X_2^{'}) < 0], \tag{1.6}$$
où $(X_1^{'}, X_2^{'})$ est un couple indépendant et de même distribution que (X_1, X_2). Tout comme pour le coefficient de corrélation de Pearson, si la valeur de $\tau(X_1, X_2)$ est proche de 0, alors les variables X_1 et X_2 sont indépendantes.

Les résultats obtenus, à l'aide du logiciel R, sont les suivants :

Mesure de dépendance	Valeur
Coefficient de Pearson	0,0137
Tau de Kendall	0,0074

TABLE 1.2 – Mesures de dépendance

Les valeurs prises par ces deux mesures sont faibles et très proches de 0. Ainsi, il est pertinent de considérer qu'il y a bien indépendance entre la fréquence et le coût. La seconde hypothèse de l'approche fréquence-coût est validée. Sous ces hypothèses vérifiées, $\mathbb{E}(X_i)$ peut s'écrire de la manière suivante :

$$\begin{aligned}\mathbb{E}(X_i) = \mathbb{E}[\mathbb{E}(X_i|N_i)] = \mathbb{E}[\mathbb{E}[\sum_{j=1}^{N_i} C_{i,j}|N_i]] &= \mathbb{E}[\mathbb{E}(N_i C_i|N_i)] \text{ car les } \mathrm{C_{i,j}} \text{ sont iid,} \\ &= \mathbb{E}[N_i \mathbb{E}(C_i|N_i)], \\ &= \mathbb{E}[N_i \mathbb{E}(C_i)] \text{ par indépendance de } \mathrm{C_i} \text{ et } \mathrm{N_i}, \\ &= \mathbb{E}(N_i)\mathbb{E}(C_i). \end{aligned} \tag{1.7}$$

Ainsi, l'espérance de la charge totale de sinistres est égale au produit des espérances du nombre et du coût. De ce fait, la prime pure de l'assuré i sera alors :

$$PP_i = \frac{\text{Charge totale des sinistres}}{\text{Exposition}} = \text{Fréquence de sinistres} \times \text{Coût moyen} = f_i \times c_i. \tag{1.8}$$

Cette séparation de la modélisation en un modèle de la fréquence et en un modèle de coût a été motivée par plusieurs raisons :

- Les variables sélectionnées dans les deux modèles peuvent être différentes, autrement dit, les facteurs expliquant la fréquence ne sont peut-être pas les mêmes que ceux qui expliquent le coût ;
- Les modèles de fréquence sont habituellement plus stables que les modèles de coût. En effet, dans les modèles de coût, le nombre d'observations est moins important car seuls les contrats présentant au moins un sinistre peuvent être utilisés dans le but d'étudier la loi des montants des sinistres.

1.2.2.2 De la prime pure chargée à la prime commerciale

Une fois la prime pure définie, la seconde étape dans le processus de construction de la prime finale consiste à ajouter les composantes suivantes dans cet ordre :

- Les chargements pour frais qui permettent de financer les différents coûts supportés par l'assureur, non directement liés à la sinistralité. Ces frais sont liés aux coûts d'acquisition, à la gestion et à l'administration des contrats, mais également à la gestion des sinistres, au commissionnement du réseau et aux frais de réassurance ;
- Les taxes, pouvant varier selon les garanties, régies par le Code des Assurances et versées à des fonds nationaux ou fonds de garanties ;
- Un chargement dit de sécurité généralement intégré pour pallier à la volatilité de la sinistralité afin de réduire la probabilité de ruine de l'assureur ;
- La marge comprenant une composante destinée à la rémunération des fonds propres demandée par les actionnaires et une composante bénéficiaire pour l'assureur sur le produit.

Finalement, cette prime technique chargée est ajustée en fonction de la politique commerciale pour aboutir à la prime commerciale. Cette dernière prime représente le prix que doit payer un assuré pour bénéficier d'une couverture d'un risque en cas de sinistre. Le tarif a ainsi été calculé sans connaître le montant de l'indemnité qui sera versé en cas de sinistre.

1.3 Objectifs de l'étude

Les travaux présentés dans ce mémoire s'inscrivent dans le cadre d'une refonte complète du tarif du produit Habitation, afin d'obtenir une vision plus juste et plus récente des risques au sein du portefeuille. Il a été décidé de construire un modèle de prime pure résultant d'une combinaison d'un modèle de fréquence et d'un modèle de coût. L'analyse des coûts de sinistres est sensiblement plus complexe que celle de la fréquence. Là où tous les individus sont utilisés pour la modélisation du nombre de sinistres, seuls les contrats sinistrés doivent être considérés lors de l'estimation du coût moyen, ce qui limite le nombre d'observations.

Le choix de la garantie étudiée dans cette étude s'est porté sur **le dégât des eaux** qui couvre les risques suivants[3] :

- Les dommages provoqués à l'intérieur des bâtiments assurés par :
  - La fuite, la rupture ou le débordement des canalisations intérieures, des chéneaux, des gouttières et de tous les appareils à effet d'eau (installation de chauffage, lave-linge, lave-vaisselle, baignoire, lavabo, aquarium, *etc.*) ;
  - Les infiltrations au travers des toitures, ciels vitrés, terrasses, balcons, façades et murs extérieurs ;
  - La rupture accidentelle ou le débordement exceptionnel d'égouts, non dû à un événement climatique ;
  - Le gel des conduites des appareils de chauffage et des appareils à effet d'eau.
- Les dommages matériels causés par les pompiers ;
- Les dommages subis sur les bâtiments, les embellissements et aménagements immobiliers assurés, consécutifs à un dégât des eaux dû à la faute d'un tiers identifié.

Cette décision de n'étudier que la dégât des eaux a été justifiée par le fait qu'il s'agit de la cause de sinistres la plus importante en assurance Habitation. Et pour éviter d'introduire une hétérogénéité dans les données en mélangeant différents types de biens, **seul le segment des appartements sera conservé**. En effet, ce type d'habitation représente plus de la moitié des sinistres déclarés pour cause de dégât des eaux (soit 56%, figure 1.4). La base de modélisation utilisée pour l'étude contiendra près de 120 000 sinistres, représentant une charge globale à fin 2020 de plus de 166 M€.

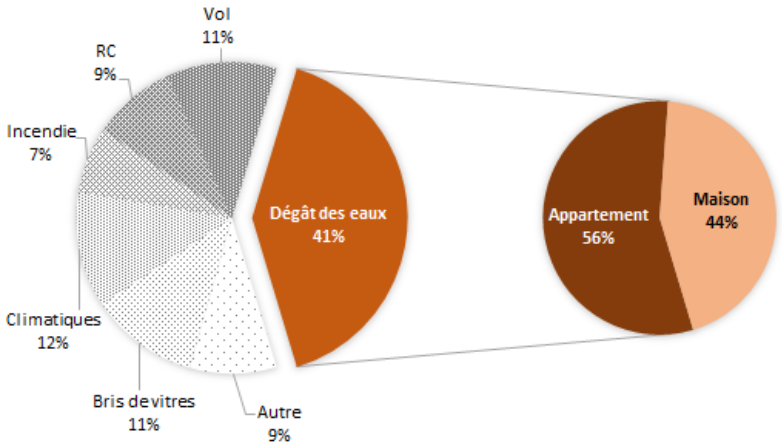


FIGURE 1.4 – Répartition des sinistres selon leur typologie

L'étude aura pour objectif de déterminer autant que possible un modèle lisible et parcimonieux. Il est, en effet, du rôle de l'actuaire de pouvoir échanger de façon transparente sur le contenu des modèles avec différents interlocuteurs tels que les agents généraux ou d'autres équipes en interne.

Il est donc possible maintenant de présenter le portefeuille d'étude et la construction de la base de modélisation à partir des données disponibles.

3. Source : Conditions générales AXA

1.4 Construction de la base de modélisation

Afin de pouvoir effectuer une modélisation pertinente, il est nécessaire de construire une base de données robuste et la plus complète possible. Celle-ci doit contenir les informations des contrats Habitation sur plusieurs années, dans le but d'obtenir un maximum de caractéristiques concernant le contrat, le client associé et les éventuels sinistres survenus. Ce grand nombre de variables explicatives sera réduit après analyse des statistiques descriptives.

1.4.1 Base Contrats par image de risque

Les bases Contrats donnent toutes les caractéristiques de l'habitation disponibles (nombre de pièces, type d'habitation, type d'occupation, ancienneté du logement, *etc.*) et les caractéristiques de souscription (exposition, cotisation émise, garanties et options souscrites). Ces bases associent les variables de risque à un numéro de contrat. L'objectif est de construire une base de risque qui comporte autant de lignes que de risques en portefeuille durant la période considérée, et non autant de lignes que de numéros de contrat. Pour s'assurer d'une plus grande stabilité et fiabilité de la modélisation, le périmètre d'étude intégrera l'ensemble des contrats ayant été au moins un jour en vigueur sur la période du 01/01/2017 au 31/12/2019. Par ailleurs, les risques atypiques de type Mobil-homes ou Propriétaires non occupants feront l'objet d'une modélisation séparée qui ne sera pas abordée dans ce mémoire.

Lors de la souscription, le contrat d'assurance détaille les caractéristiques du risque assuré. Or, ce contrat peut être modifié dans le temps à mesure que le risque évolue : ajout ou suppression de garanties, changement de franchise, modification des informations déclarées, déménagement. Ces modifications génèrent un nouveau fait de production appelé Remplacement. L'assuré conserve le même numéro de contrat, mais les caractéristiques du risque diffèrent. Pour la construction de la base de données avec l'ensemble des risques présents dans le portefeuille, les risques avant et après remplacement seront traités comme étant deux risques différents, nécessitant de créer deux lignes, ou images de risque, distinctes.

Le principe de la construction de la base de modélisation est détaillé ci-dessous pour expliciter la façon dont a été construite la base par image de risque (*Halimi, 2017*, [**5**]).

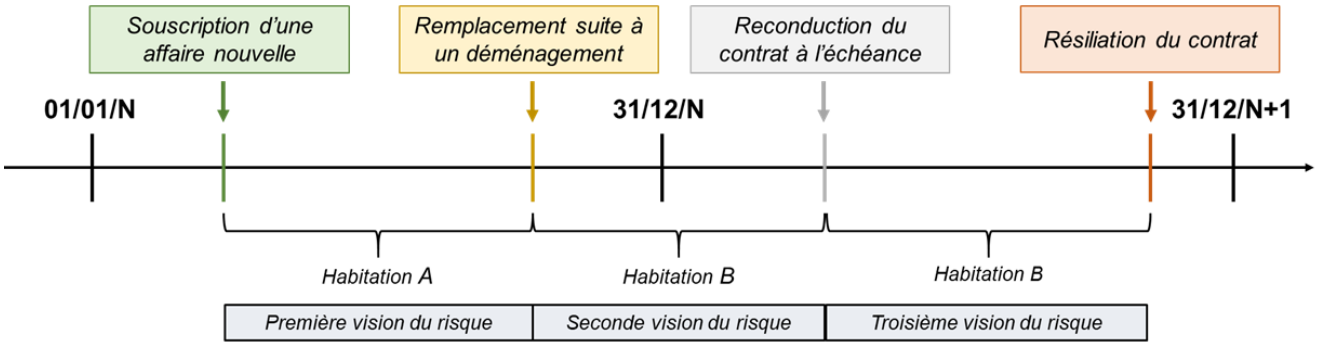


FIGURE 1.5 – Approche par image de risque et par exercice pour un contrat donné

Pour une année d'exercice donnée, chaque vision mensuelle des contrats est récupérée. Le tableau suivant illustre, pour quelques contrats qui ont comme seule caractéristique l'adresse, les images enregistrées dans la base.

Contrat	Date de souscription	Date de remplacement	Date de résiliation	Adresse	Vision
01	20/01/N			1 rue A	30/11/N
01	20/01/N			1 rue A	31/12/N
02	08/08/N			2 rue B	30/11/N
02	08/08/N	08/12/N		3 rue C	31/12/N
03	28/10/N			4 rue D	30/11/N
03	28/10/N		22/12/N	4 rue D	31/12/N

TABLE 1.3 – Vision mensuelle de la base des contrats

Cependant, à ce stade de l'étude, chaque ligne correspond à la vision mensuelle d'un contrat et non à une image de risque. Afin d'obtenir la base souhaitée, il a fallu procéder aux étapes suivantes (*Bernanose, 2020*, [**1**]) :

- Tri de la table par numéro de contrat, date de remplacement et de résiliation ;
- Dédoublonnage de la table par le biais de la concaténation du numéro de contrat et de la date de remplacement. Ce dédoublonnage a pour effet d'identifier chaque version du contrat et de conserver uniquement une ligne par image de risque ;
- Détermination de la date de début de couverture du risque qui correspond à la date la plus récente entre les dates de début d'exercice, d'affaire nouvelle et de remplacement ;
- Détermination de la date de fin de couverture du risque qui correspond à la date la plus ancienne entre les dates de fin d'exercice, de résiliation et de remplacement.

Le résultat final obtenu est une base contenant une unique image par risque pour un exercice donné.

Contrat	Adresse	Date de début de couverture du risque	Date de fin de couverture du risque
01	1 rue A	20/01/N	31/12/N
02	2 rue B	08/08/N	08/12/N
02	3 rue C	08/12/N	31/12/N
03	4 rue D	28/10/N	22/12/N

TABLE 1.4 – Vision par image de risque

Néanmoins, il subsiste une limite quant à cette méthode de construction de base. En effet, si plusieurs remplacements sont effectués sur un contrat durant le même mois, seule la dernière vision mensuelle sera conservée. Cependant, ce cas de figure est rare et le fait de ne pas prendre en compte ces visions de risque n'impactera pas le reste de l'étude. Ce point pourrait être corrigé en créant une base image journalière et non mensuelle. Mais la structure du flux interne de remontées d'informations ne permet pas actuellement de constituer une telle base. La base par image de risque étant constituée, après utilisation du logiciel SAS, il faut maintenant obtenir les informations liées à la sinistralité pour chaque observation.

1.4.2 Base Sinistres

Les bases Sinistres utilisées sont créées par année de survenance. Il est possible de connaître la ou les garanties concernées par le sinistre ainsi que les flux comptables observés (paiements, provisions et recours). Le périmètre d'étude de la base Contrats est repris à l'identique, soit une prise en compte des sinistres survenus entre le 01/01/2017 et le 31/12/2019, ainsi qu'une exclusion de ceux rattachés aux contrats couvrant des risques de type Mobil-homes ou Propriétaires non occupants. De même, les sinistres disposant d'une charge nulle ou négative ont été écartés dans une logique de nettoyage de la base de données.

Tous les sinistres pris en compte dans l'étude, et ce quelle que soit l'année de survenance, sont observés à une vision fixe du 31/12/2020. Cela permet ainsi d'avoir une année de développement pour les sinistres survenus en 2019, ce qui est, dans la plupart des cas, suffisant pour une garantie à développement court comme le dégât des eaux. Quant aux sinistres survenus en 2017 et 2018, cette vision plus tardive permet d'obtenir un nombre de sinistres clos plus important, et donc une charge observée la plus proche possible de la vision finale du sinistre.

Par ailleurs, ce choix de vision a été également justifié par la volonté de réconcilier les données issues de ces bases Sinistres avec celles d'autres équipes, notamment l'équipe Comptes, dans le but d'assurer une exactitude des données qui vont être utilisées pour la modélisation. En effet, le nombre et la charge de sinistres obtenus par cette équipe sont considérés comme des valeurs de référence. Et la vision au 31/12/2020 correspondait à la date de dernière évaluation des sinistres effectuée par l'équipe Comptes.

1.4.3 Autres bases internes utilisées

Deux autres sources de données internes vont permettre d'enrichir la base actuelle comprenant les informations du contrat et des éventuels sinistres survenus.

Tout d'abord, les bases Clients donnent des renseignements sur la situation personnelle et professionnelle de l'assuré (âge, statut marital, nombre d'enfants, nombre de contrats détenus au sein de la compagnie d'assurances). Ces informations pourront permettre d'observer des impacts liés au comportement et au niveau de vie de l'assuré.

Une autre source de données va être utilisée et sera particulièrement utile pour la thématique du zonier, abordée à la fin de l'étude : les bases Adresses. L'objectif est de joindre les attributs géographiques de chaque adresse relative aux contrats présents dans la base de modélisation à partir des informations suivantes :

- Adresse, code postal et commune associées au risque assuré. Il existe 34 881 communes en France métropolitaine[4] ;
- Code IRIS - **I**lots **R**egroupés pour l'**I**nformation **S**tatistique : il s'agit d'un découpage du territoire en mailles de taille homogène. Il en existe 48 606 en France métropolitaine[5] ;
- Coordonnées géographiques avec une projection conique de type Lambert-93 (projection utilisée pour représenter seulement une partie du globe en minimisant les déformations à l'échelle de la France) ;
- Précision du géocodage : 4 pour l'adresse exacte, 3 pour le numéro de rue approché, 2 pour le centroïde de la voie, 1 pour le centroïde de la ville et 0 en cas d'erreur.

4. Source : INSEE, Recensement de la population paru le 28/12/2020
5. Source : INSEE, Découpage infra-communal paru le 13/07/2021

1.4.4 Données externes

Les différentes sources de données présentées jusqu'à présent ne contenaient que des informations propres à la compagnie d'assurances. Pour enrichir la base de modélisation, il a été décidé d'incorporer des variables externes dans l'étude. Ne disposant pas de données payantes fournies par un prestataire externe, les informations récupérées proviennent de différentes sources en accès libre :

- La base gouvernementale DVF (**D**emandes de **V**aleurs **F**oncières) recense l'ensemble des transactions immobilières sur les cinq dernières années. Par soucis de cohérence avec les autres bases, seule la période 2017-2019 a été conservée. Cela représente toutefois 1,7 million de transactions sur des biens de type Appartement. Les données contenues sont issues des actes notariés et des informations cadastrales. Il est donc possible de connaître, entre autres, le prix et la surface de chaque parcelle vendue. Or, ne disposant pas de la référence de parcelle pour les contrats détenus en portefeuille, les données ont été agrégées au niveau de la commune pour pouvoir être exploitées ;
- Les bases Logements issues du recensement de la population et publiées par l'INSEE aux mailles Communes et IRIS apportent des informations sur le nombre de logements, de résidences principales ou secondaires, la répartition des biens selon le nombre de pièces ou l'ancienneté de construction ;
- La base Salaires et revenus d'activité, issue de la base globale DADS (**D**éclaration **A**nnuelle des **D**onnées **S**ociales) publiée par l'INSEE, permet d'avoir accès, pour toutes les communes de plus de 2 000 habitants, au salaire horaire moyen par catégorie socioprofessionnelle et par tranche d'âge. L'utilisation de cette donnée va permettre de voir, entre autres, s'il existe ou non un lien entre le coût d'un sinistre et le coût de la main d'oeuvre d'un artisan en cas de travaux de réparation ;
- Les données mensuelles des différentes stations météorologiques de Météo France apportent des informations sur les paramètres atmosphériques mesurés tels que la température, l'humidité, la direction et la force du vent, la pression atmosphérique, ainsi que le niveau de précipitations. Cette dernière variable est particulièrement intéressante dans le cadre de l'étude de la garantie Dégât des eaux car celle-ci couvre également les dégâts causés par les infiltrations d'eau au niveau des toitures, murs et terrasses provenant de l'extérieur.

Cet ajout de données va, par ailleurs, permettre de contourner la problématique d'asymétrie d'information à laquelle sont souvent confrontés les assureurs lorsque seules des informations déclaratives sont utilisées pour la tarification.

Les données externes étant disponibles à différents niveaux géographiques, leur répartition peut être analysée en les visualisant sur une carte (après utilisation du module *leaflet* du logiciel R).

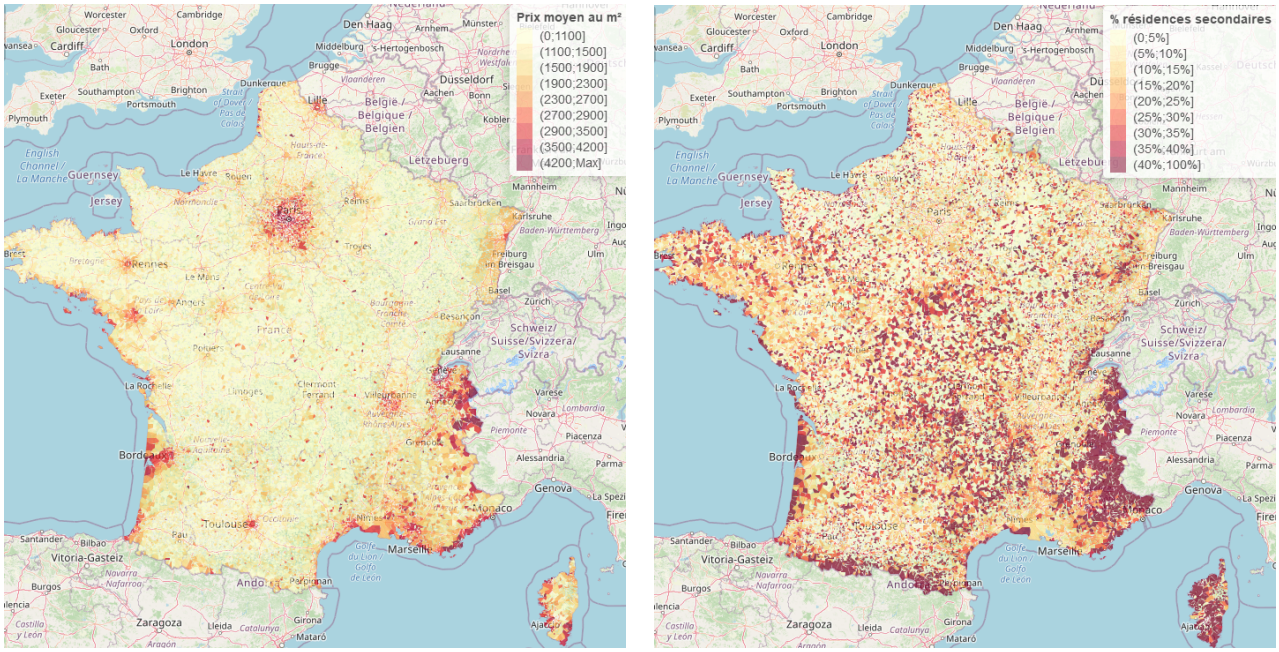


FIGURE 1.6 – Répartition par commune du prix au m^2 (gauche) et de la part de résidences secondaires (droite)

Il est possible de remarquer que l'information apportée par ces données n'est pas uniforme car les zones géographiques ayant un prix au m^2 élevé correspondent aux bassins de vie des principales villes de France en termes de population (Paris, Lyon, Marseille ou encore Bordeaux) ou aux régions touristiques du quart Sud-Est (Savoies et pourtour méditerranéen). Concernant la part de résidences secondaires sur la commune, le Sud de la façade Atlantique ainsi que les Pyrénées, la région PACA (**P**rovence **A**lpes **C**ôte d'**A**zur) et la Corse sont les zones où celle-ci est la plus élevée avec en général plus de 30% des logements inoccupés une partie de l'année.

Une fois la jointure entre ces différentes sources de données réalisée, la base de données qui va être utilisée pour la modélisation est composée de près de dix millions d'observations. Cependant, il est nécessaire d'effectuer au préalable un certain nombre de retraitements, principalement sur les données Sinistres, pour éviter une discontinuité dans la distribution de la sinistralité.

— Chapitre 2 —

Préparation des données

2.1 Gestion des sinistres graves

Dans le cadre de la tarification, une hypothèse classique est celle selon laquelle le portefeuille est constitué de risques similaires. Un problème pour que cette hypothèse soit vérifiée est le poids important des sinistres dits graves qui ne sont pas mutualisables. Les sinistres observés vont donc être écrêtés et mutualisés. L'objectif est de trouver un seuil d'écrêtement, c'est-à-dire un seuil de charge de sinistre, au-delà duquel les sinistres seront considérés comme graves. A noter qu'il existe d'autres méthodes pour gérer ces sinistres graves, telles que les modèles de propension qui visent à effectuer une modélisation séparée. Mais leur utilisation est traditionnellement limitée aux garanties d'intensité (incendie, responsabilité civile, sécheresse).

2.1.1 Écrêtement des sinistres

L'écrêtement va consister à repérer un changement dans la queue de la distribution de la sinistralité. Les sinistres attritionnels (forte fréquence et montants faibles) vont être distingués de ceux considérés comme graves (faible fréquence et montants très élevés). Cela permet d'éviter que les sinistres ayant une charge élevée biaisent les résultats des modèles qui seront effectués. La charge associée aux sinistres graves est ensuite mutualisée uniformément sur l'ensemble des contrats sinistrés, la sur-crête correspondant au montant de charge au-delà d'un seuil à définir :

$$\text{Charge}_{\text{mutualisée}} = \text{Charge}_{\text{écrêtée}} \left(1 + \frac{\text{Sur-crête totale}}{\text{Sous-crête totale}}\right). \tag{2.1}$$

Le tableau suivant illustre de manière simplifiée le processus d'écrêtement avec un seuil de sur-crête α compris entre 5 000 € et 50 000 € et un coefficient β calculé en utilisant la formule précédente :

Charge initiale	Charge écrêtée[6]	Sur-crête	Charge mutualisée
500 €	500 €		500 € $\times(1+\beta)$
5 000 €	5 000 €		5 000 € $\times(1+\beta)$
50 000 €	α	50 000 € - α	$\alpha \times (1+\beta)$
55 500 €	5 500 € + α	50 000 € - α	55 500 €

Table 2.1 – Construction de la charge mutualisée

6. Appelée également Sous-crête

Ainsi, une fois la sur-crête mutualisée, il n'y a plus de distinction entre les sinistres. Il est donc nécessaire maintenant de déterminer le seuil permettant de juger de la gravité d'un sinistre.

2.1.2 Détermination du seuil de sinistres graves

La détermination du seuil va être effectuée en utilisant la théorie des valeurs extrêmes (*Thomas, 2020*, [**11**]), les méthodes statistiques traditionnelles n'étant pas adaptées lorsque les données sont peu nombreuses. L'objectif est de trouver le seuil optimal séparant les classes de sinistres attritionnels et graves. La distribution de Pareto généralisée (ou GPD pour ***G**eneralized **P**areto **D**istribution*) est pertinente pour modéliser des dépassements de seuil. Pour une variable aléatoire X, les dépassements $(X-u)$, au-delà d'un certain seuil u, suivent une $GPD_{\sigma,\gamma}(x)$, où σ est un paramètre positif dit d'échelle et γ le paramètre de forme. Plus la valeur du paramètre de forme est élevée, plus il y aura de valeurs extrêmes dans la distribution. La fonction de répartition d'une $GPD_{\sigma,\gamma}(x)$ est pour $x > 0$:

$$\forall x \in \mathbb{R}, \mathbb{P}(X - u \geq x | X > u) = \begin{cases} 1 - \left(1 + \gamma \frac{x}{\sigma}\right)^{-1/\gamma} & \text{si } \gamma > 0, \\ 1 - \exp(-\frac{x}{\sigma}) & \text{si } \gamma = 0. \end{cases} \tag{2.2}$$

Afin de définir la valeur du seuil, la stabilité des paramètres qui correspondent à certaines propriétés des GPD ainsi que la fonction de dépassement moyen des excès seront analysées, tout comme l'estimateur de Hill. Cette sélection du seuil constitue une difficulté. En effet, u doit être suffisamment élevé pour que l'approximation GPD soit valide, sans que cela ne conduise à avoir un faible nombre de données pour estimer les paramètres du modèle. Le seuil doit être choisi de façon à faire un arbitrage, traditionnel en statistiques, entre le biais et la variance.

La première méthode appliquée est l'analyse des paramètres des GPD. Ces lois présentent une propriété de stabilité par seuil qui implique que si les dépassements $(X-u)$ suivent une $GPD_{\sigma_u,\gamma}(x)$, alors pour tout seuil $v > u$, les dépassements $(X-v)$ suivent également une GPD de paramètres σ_v et γ. Seul le paramètre d'échelle diffère et est une fonction linéaire du seuil : $\sigma_v = \sigma_u + \gamma(v-u)$, le paramètre γ étant identique pour tout u. L'analyse de la stabilité du paramètre d'échelle s'effectue de manière graphique. L'estimation de ce paramètre est donc représentée pour plusieurs seuils, en incluant les intervalles de confiance à 95%. Le seuil retenu est celui correspondant à la plus petite valeur de u pour laquelle le paramètre est stable.

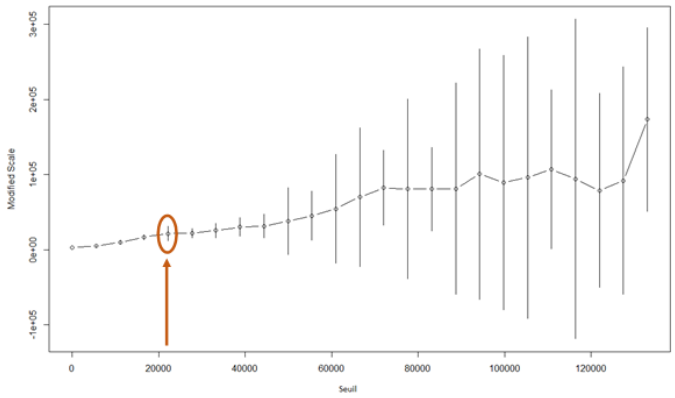


FIGURE 2.1 – Stabilité du paramètre d'échelle

Le paramètre d'échelle devient stable pour un seuil légèrement supérieur à 20 000 €, la valeur exacte étant de 20 333 €. Il convient donc de comparer cette première valeur avec d'autres méthodes de définition de seuil.

La seconde méthode utilisée est celle de la fonction de dépassement moyen des excès. Cette fonction, linéaire par rapport à v si l'approche GPD est valide, est définie comme :

$$e(v) = \mathbb{E}[X - v | X > v] = \frac{\sigma + \gamma(v - u)}{1 - \gamma}, \text{ pour } v > u \text{ et } \gamma < 1. \tag{2.3}$$

L'estimateur de $e(v)$ est :

$$e_n(v) = \frac{1}{N_v} \sum_{i=1}^{N_v} (X_i - v)_+ \text{ et } N_v \text{ le nombre de données supérieures à } v. \tag{2.4}$$

Si la variable aléatoire suit une loi de Pareto généralisée pour un seuil donné, alors le graphique doit être approximativement linéaire au-delà de ce seuil. En pratique, le seuil est déterminé via le graphique des excès à la moyenne en exploitant la linéarité de la fonction de dépassement moyen pour la GPD.

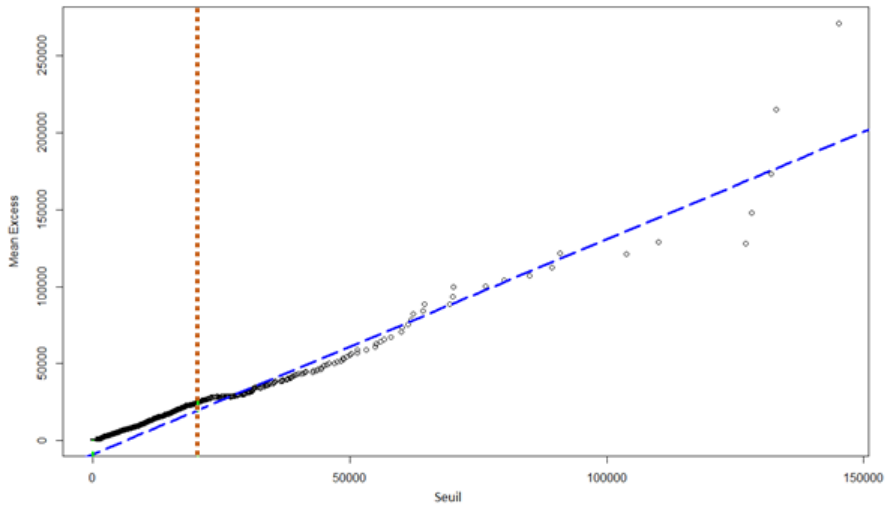


FIGURE 2.2 – Dépassement moyen en fonction du seuil

La valeur retenue correspond au moment où les observations s'alignent à la droite bleue représentant la droite de régression linéaire du nuage de points. Elle est égale à 20 468 €.

Une dernière méthode va être utilisée pour définir le seuil : il s'agit de l'estimateur de Hill qui n'est utilisable que pour les distributions à queue lourde, c'est-à dire où le paramètre de forme γ est positif :

$$\hat{\gamma}_H(k) = \frac{1}{k} \sum_{j=1}^{k} \ln \left(\frac{X_{(i)}}{X_{(k+1)}} \right), \tag{2.5}$$

où $X_{(i)}$ est la statistique d'ordre associée à l'échantillon $X_1, ..., X_n$ et k représente un nombre d'excès inférieur ou égal à n qui est le nombre d'observations. Le graphique de l'estimateur de Hill représente la valeur de l'estimateur en fonction de l'indice k de la statistique d'ordre, soit l'estimateur construit à partir des observations supérieures ou égales à $X_{(k)}$. L'objectif est de trouver une zone où l'estimateur est stable et semble robuste. Le plus petit seuil u appartenant à cette zone est alors défini comme optimal.

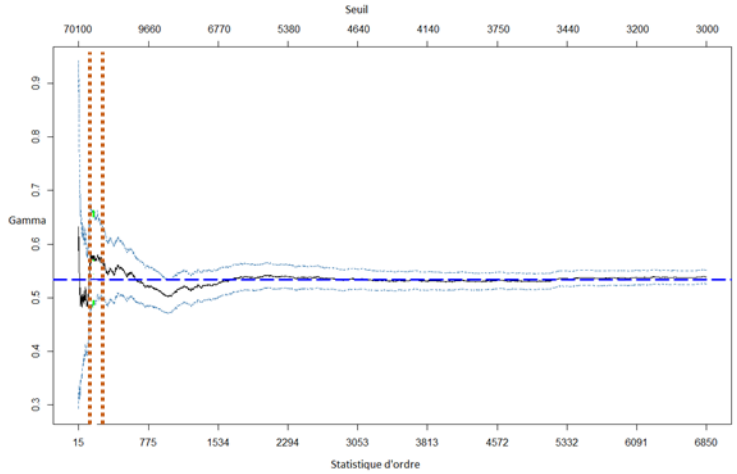


FIGURE 2.3 – Estimateur de Hill

Une zone de stabilisation de l'estimateur est observée assez tôt dans l'échantillon, proche de la 150^e statistique d'ordre. Le seuil associé est de 20 207 €. Il est à noter que les seuils retenus pour chacune des trois méthodes sont très proches. Cependant, une étape de vérification d'adéquation à une GPD est nécessaire. Pour le seuil le plus élevé obtenu (20 468 € par la méthode de la fonction de dépassement moyen des excès, arrondi par la suite à 20 500 €), les paramètres σ et γ vont être estimés en utilisant le maximum de vraisemblance. Ensuite, pour évaluer la qualité d'ajustement à une GPD, un graphique Quantile-Quantile va être généré. Ce graphique permet de tester visuellement l'adéquation d'une famille de lois à des données. Il s'agit d'un nuage de points ayant pour abscisse les quantiles théoriques d'une GPD (calculés à partir des paramètres estimés), et pour ordonnée les quantiles empiriques de la distribution observée. Si l'adéquation est vérifiée, les quantiles de la famille de lois testées et les quantiles de l'échantillon sont linéairement liés.

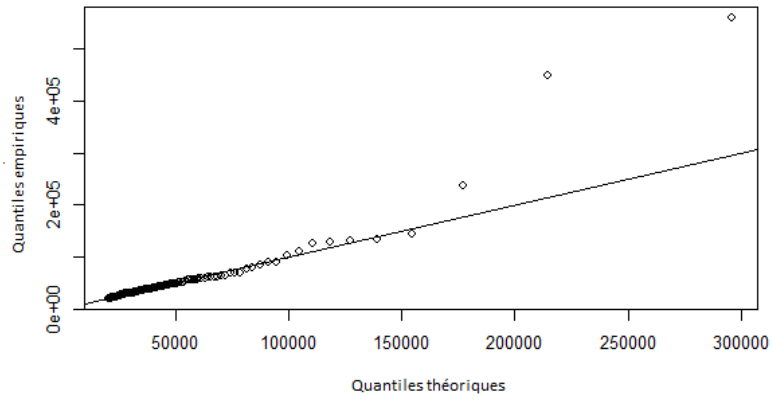


FIGURE 2.4 – Graphique Quantile-Quantile pour une loi GPD

Graphiquement, la queue de distribution semble bien être en adéquation avec une loi de Pareto généralisée. Il a donc été décidé de retenir ce seuil comme étant la séparation entre la charge attritionnelle et grave des sinistres de la base d'étude. 0,4% des contrats sinistrés ont alors une charge considérée comme grave (la charge maximale s'élevant à 560 000 €). En comparaison, la charge moyenne observée s'élève à 1 428 €. Le montant cumulé des sinistres dont la charge est au moins égale au seuil représente 4,3% de la charge totale de la sinistralité associée à la garantie Dégât des eaux sur les appartements.

Le seuil ayant été défini, il est possible maintenant de calculer la charge mutualisée en utilisant les notions de sous-crête et sur-crête. La charge totale sur la garantie étudiée s'élève à 166,24 M€, dont 159,03 M€ de sous-crête et 7,21 M€ de sur-crête. Le coefficient de mutualisation déduit de ces deux valeurs est égal à 1,0453. Par conséquent, la charge de chacun des sinistres va être multipliée par cette quantité et ainsi l'impact des sinistres graves sera neutralisé sur la base d'étude.

2.2 Vieillissement des sinistres

Lorsqu'un sinistre survient, une évaluation forfaitaire lui est affectée en fonction de sa nature lors de l'ouverture du dossier. Celle-ci sera ensuite revue à la baisse ou à la hausse après passage d'un expert. Ainsi, le montant total du coût du sinistre (somme des paiements versés à l'assuré et des provisions de paiements futurs à verser à laquelle est déduit le montant des recours récupérés) peut évoluer au cours du temps. Pour cela, un vieillissement de ce coût va être effectué. Cela consiste à estimer les évolutions du coût des sinistres au cours du temps. Pour obtenir la charge finale des sinistres, la méthode de Chain Ladder va être utilisée en étudiant des triangles de charge qui présentent l'évolution de la charge cumulée d'une année comptable à une autre pour chaque année de survenance. Cette méthode permet de projeter des valeurs observées jusqu'à extinction de tous les mouvements des sinistres, c'est-à-dire jusqu'à l'ultime. Les notations utilisées dans la suite de la présentation de la méthode de Chain Ladder sont les suivantes :

- i : année de survenance des sinistres ;
- j : année de développement des sinistres, soit la $j^{ième}$ année après la survenance du sinistre ;
- $X_{i,j}$: charge versée lors de la $j^{ième}$ année de développement pour les sinistres survenus lors de l'année i ;
- $C_{i,j} = \sum_{l=0}^{j} X_{i,l}$: charge cumulée lors des j premières années de développement et associée à l'année de survenance i.

Le développement de la charge pour chaque année de survenance peut être résumé dans le tableau ci-dessous :

i / j	0	1	...	j	...	*n-1*	n
1	$C_{1,0}$	$C_{1,1}$	...	$C_{1,j}$	...	$C_{1,n-1}$	$C_{1,n}$
2	$C_{2,0}$	$C_{2,1}$	...	$C_{2,j}$	...	$C_{2,n-1}$	
...	...	...	...	...	...		
i	$C_{i,0}$	$C_{i,1}$	...	$C_{i,j}$			
...	...	...	...				
n-1	$C_{n-1,0}$	$C_{n-1,1}$					
n	$C_{n,0}$						

TABLE 2.2 – Triangle de charges cumulées

La méthode de Chain Ladder consiste à déterminer la diagonale inférieure en estimant les montants non connus et repose sur les hypothèses suivantes :

- Aucun sinistre ne peut être encore ouvert après l'année de développement n ;
- Les versements cumulés sont indépendants entre les années de survenance ;
- La répartition des paiements est supposée constante dans le temps. Ce dernier point peut se réécrire comme suit :

$$\frac{C_{1,j+1}}{C_{1,j}} = \frac{C_{2,j+1}}{C_{2,j}} = ... = \frac{C_{i,j+1}}{C_{i,j}} = ... = \frac{C_{n-1,j+1}}{C_{n-1,j}} = \frac{C_{n,j+1}}{C_{n,j}} = \text{constante}\,. \tag{2.6}$$

Le triangle de charges se complète en utilisant ces proportions appelées facteurs de développement pour chaque année de développement j :

$$\hat{f}_j = \frac{\sum_{i=1}^{n-j-1} C_{i,j+1}}{\sum_{i=1}^{n-j-1} C_{i,j}}. \tag{2.7}$$

Une fois ces facteurs calculés, il est possible d'obtenir une estimation de la charge cumulée finale :

$$\hat{C}_{i,n} = \hat{C}_{i,n-1} \times \hat{f}_{n-1} = \hat{C}_{i,n-2} \times \hat{f}_{n-1} \times \hat{f}_{n-2} = ... = C_{i,n-j} \times \hat{f}_{n-j} \times ... \times \hat{f}_{n-2} \times \hat{f}_{n-1}. \tag{2.8}$$

Pour utiliser cette technique de provisionnement de la charge de sinistres, il est préférable que l'historique utilisé pour le développement des sinistres soit stable et profond. En effet, sans cela, la charge finale pourrait ne pas être estimée de façon optimale. L'historique de sinistres entre 2005 et 2019 va être utilisé pour développer la charge ultime des sinistres de la période 2017-2019 de la base d'étude. Tous ces sinistres seront vus à fin 2020 afin de bénéficier de la vision la plus tardive possible.

Afin de gagner en précision, il a été décidé de séparer le développement des sinistres attritionnels et graves. Cette distinction est d'autant plus importante que le temps d'évaluation de la charge du sinistre dépend fortement de sa sévérité. Ainsi, le seuil de 20 500 € retenu dans la section précédente va être utilisé.

Les facteurs de développement obtenus d'une année de survenance à l'autre pour les sinistres du périmètre Dégât des eaux sur les appartements, avec une distinction attritionnels-graves, sont les suivants :

	Année de développement									
Type de sinistres	1	2	3	4	5	6	7	8	...	14
Attritionnels	1,070	1,011	**1,002**	1,000	1,000	1,000	1,000	1,000	...	1,000
Graves	1,161	1,062	1,074	1,034	1,036	1,043	1,037	**1,005**	...	1,000

TABLE 2.3 – Facteurs de développement pour les sinistres attritionnels et graves

Les sinistres attritionnels ont une charge stable après trois années de développement, les facteurs étant égaux à 1 ensuite. Concernant les sinistres graves, la stabilisation de la charge est beaucoup plus lente, celle-ci n'intervient qu'à partir de la huitième année de développement. Cela s'explique par une charge beaucoup plus volatile, qui nécessite donc plus de temps pour tendre vers la charge finale estimée. Il est à remarquer que la charge affectée lors de l'ouverture des sinistres est sous-estimée par rapport à celle observée un an après. Cet écart se chiffre à respectivement 7% et 16,1% pour les sinistres attritionnels et graves. Une cause possible est un coût forfaitaire d'ouverture trop faible qui pourrait donc, par conséquent, être amélioré. Maintenant que les facteurs de développement ont été calculés, il est possible de définir une charge finale estimée pour chaque année de survenance ainsi qu'un coefficient de passage entre la charge observée en 2020 et la charge ultime :

	Charge au 31/12/2020		Charge à l'ultime		Facteurs de développement	
Année	Attritionnels	Graves	Attritionnels	Graves	Attritionnels	Graves
2017	49,163 M€	1,593 M€	49,359 M€	2,051 M€	1,004	1,287
2018	47,174 M€	1,496 M€	47,881 M€	2,045 M€	1,015	1,367
2019	52,961 M€	1,823 M€	57,516 M€	2,893 M€	1,086	1,587

TABLE 2.4 – Charge à l'ultime et Facteurs de développement par année de survenance étudiée

Les facteurs de développement obtenus vont être appliqués aux sinistres non clos selon leur catégorie (attritionnelle ou grave). A titre d'illustration, la charge des sinistres graves va être multipliée par 1,367 pour l'année de survenance 2018.

Une actualisation de la charge des sinistres aurait également pu être faite afin de prendre en compte l'inflation. En effet, il est raisonnable de penser qu'un sinistre ayant eu lieu en 2017 n'aurait plus le même coût en 2019. Pour pallier à ce problème, la charge aurait pu être indexée à partir de l'indice publié par la FFB (**F**édération **F**rançaise du **B**âtiment). Mais cela aurait conduit à augmenter artificiellement le montant des sinistres, ce qui n'était pas souhaité. En conséquence, il a été décidé de considérer l'année de survenance du sinistre comme une variable explicative qui sera conservée dans la base de modélisation et qui portera l'information liée à la notion d'inflation.

2.3 Nettoyage des données

La première étape de nettoyage de la base de données a été de supprimer l'ensemble des variables de type identificateur (numéro de contrat, de client, nom et prénom de l'assuré, coordonnées de contact) afin de respecter le cadre RGPD (**R**ègleme**n**t **G**énéral sur la **P**rotection des **D**onnées[7]). Par ailleurs, les variables temporelles utilisées pour la constitution de la base par image de risque (dates d'affaire nouvelle, de remplacement et de résiliation) ont été écartées, car elles étaient superflues dans un contexte de modélisation.

Avant de commencer l'étude, il faut s'assurer que la base de données soit homogène. Les variables conservées doivent être correctement renseignées (absence de valeurs aberrantes ou dénuées de sens) et il doit y avoir aussi peu que possible de données manquantes. En effet, la donnée est l'élément clé de la modélisation. Seule une variable comportait un nombre important de valeurs non renseignées (près de 2%) : la catégorie socio-professionnelle du client. Sachant que cette variable était une variable catégorielle, il a été décidé d'ajouter une modalité supplémentaire correspondant aux valeurs non renseignées. De cette manière, ces observations pourront être prises en compte dans la modélisation au lieu d'être écartées. Il a également été décidé d'effectuer certaines modifications de format sur les variables :

- Regroupement de modalités pour éviter que leur nombre soit trop élevé et que la volumétrie de contrats dans chacune des classes soit trop faible (exemple du nombre de pièces) ;
- Discrétisation de toutes les variables continues (âge du client, montant de capital assuré ou d'objets de valeurs) afin de tenter de capter de potentiels effets non-linéaires dans la distribution des données. La méthode utilisée pour le découpage est l'approche par quantiles. Il s'agit de construire des classes ayant le même nombre d'observations ;
- Création de variables binaires pour chaque modalité des variables catégorielles, afin de les gérer comme des variables quantitatives.

Une dernière étape est nécessaire avant de pouvoir commencer la modélisation. Il s'agit d'étudier la corrélation entre les variables de la base d'étude. Lors de la décision dans le choix des variables à intégrer dans le modèle, l'étude des corrélations est importante car la présence de plusieurs variables corrélées est à éviter pour conserver l'interprétabilité du modèle. L'indicateur de corrélation du V de Cramer est retenu et correspond à une mesure de l'association entre deux variables catégorielles basée sur la statistique du khi-deux de Pearson. Les valeurs prises sont comprises entre 0 (aucune association entre les variables) et 1 (complète association).

7. Règlement entré en application le 25 Mai 2018, Source : Ministère de l'Économie et des Finances

Cet indicateur est défini de la façon suivante :

$$V_{Cramer} = \sqrt{\frac{\chi^2/n}{[\min(k,l)-1]}} \text{ et } \chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{l} \frac{(n_{i,j} - \frac{n_{i.}n_{.j}}{n})^2}{\frac{n_{i.}n_{.j}}{n}}, \tag{2.9}$$

avec :

- k et l le nombre de modalités des variables X et Y ;
- $n_{i,j}$ le nombre d'observations ayant la modalité i pour la variable X et la modalité j pour la variable Y ;
- $n_{i.}$ le nombre d'observations ayant la modalité i pour la variable X ;
- $n_{.j}$ le nombre d'observations ayant la modalité j pour la variable Y ;
- n le nombre d'observations total.

La représentation graphique des corrélations (appelée corrélogramme) a été effectuée à l'aide du logiciel R (modules *vcd* et *corrplot*) et est présentée ci-dessous :

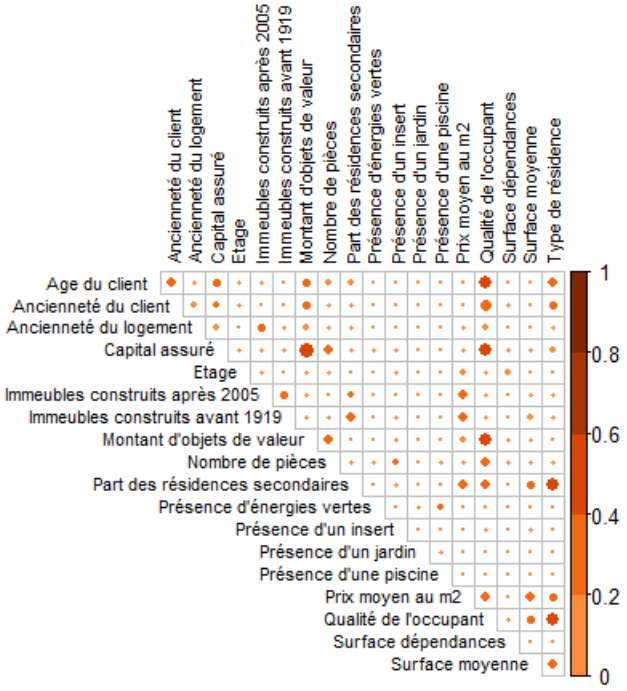


FIGURE 2.5 – Corrélogramme des variables avec la méthode du V de Cramer

Plus la taille du cercle est grande, plus la liaison entre les deux variables observées est importante. Les quatre liaisons les plus élevées sont les suivantes[8] :

- Le capital assuré et le montant d'objets de valeur [V de Cramer = 0,46] ;
- Le type de résidence et la part de résidences secondaires dans la zone géographique du bien assuré [V de Cramer = 0,44] ;
- La qualité de l'occupant (locataire ou propriétaire) et le type de résidence [V de Cramer = 0,43] ;
- La qualité de l'occupant et le capital assuré [V de Cramer = 0,41].

Les variables de la base d'étude n'ayant pas un lien fort entre elles, il n'y aura donc pas lieu de créer d'interactions qui ajouteraient une complexité supplémentaire. La base de données ayant été nettoyée, reformatée et contrôlée, il est donc possible maintenant de décrire plus en détails les variables qui serviront à la modélisation.

8. Le détail complet des valeurs de corrélation est présenté en Annexe I.

2.4 Statistiques descriptives

Le but est de bien appréhender la donnée disponible et d'obtenir des intuitions sur les relations entre certaines variables explicatives et les variables réponses que sont la fréquence et le coût moyen pour la garantie Dégât des eaux sur le périmètre des appartements. Une première analyse est d'étudier l'évolution de ces deux variables sur chacune des années d'observation de la base.

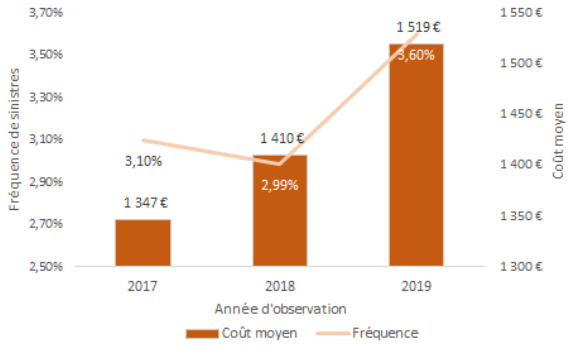


FIGURE 2.6 – Évolution de la fréquence et du coût moyen sur la période 2017-2019

Comme le montre la figure 2.6, la fréquence de sinistres a connu en 2019 une hausse de 0,5 point par rapport à 2017 et 2018. Par ailleurs, sur la période 2017-2019, le coût moyen d'un sinistre Dégât des eaux sur un risque Appartement a augmenté de près de 13% pour atteindre 1 519 €. Ces deux indicateurs illustrent une dégradation de la sinistralité, ce qui justifie une revue des modèles de tarification. Il peut être intéressant, maintenant, de regarder la variation de la fréquence et de la charge par rapport à différentes variables qui caractérisent le risque assuré.

FIGURE 2.7 – Évolution de la fréquence et du coût moyen par rapport au nombre de pièces (haut gauche), à la surface des dépendances (haut droite), au capital déclaré (bas gauche) et à l'âge du client (bas droite)

Pour le premier graphique, l'élément pris en compte est le nombre de pièces de l'appartement. C'est une des principales variables utilisées lors de la tarification d'un contrat, quelle que soit la garantie observée. Il semble que le nombre de pièces du bien assuré impacte de manière significative le fait qu'un sinistre se produise et que son coût soit élevé. En effet, l'augmentation d'une pièce entraîne en moyenne une hausse d'un point de la fréquence de sinistre. Quant à la charge moyenne de sinistre, elle reste constante pour les risques disposant de moins de quatre pièces (1 320 € environ), mais augmente fortement par la suite (pour dépasser 2 500 €).

Ces résultats se retrouvent, dans une moindre mesure cependant, pour le capital assuré. Plus le montant de biens déclarés est élevé, plus la fréquence l'est. Au niveau du coût moyen, une certaine stabilité est observée jusqu'au montant de 30 000 € de capitaux déclarés, puis une nette hausse apparaît pour atteindre une valeur de près de 2 000 € pour la dernière tranche de capitaux (montant supérieur à 35 000 €).

Le fait d'avoir une dépendance semble jouer également sur la sinistralité, mais de manière beaucoup moins marquée que pour les deux précédentes variables présentées. La fréquence de sinistres est quasi identique entre les profils sans dépendances (3,21%) et ceux disposant d'une dépendance de plus de 100 m^2 (3,26%) . Par ailleurs, la présence d'une dépendance augmente entre 50 € et 100 € en moyenne (selon la surface déclarée) le coût d'un sinistre.

En étudiant l'évolution de la sinistralité en fonction de l'âge, il ressort que les assurés de moins de 30 ans ont un profil de risque faible comparé au reste de la base de données. La fréquence observée fluctue sans qu'une réelle tendance ne se dessine. Elle décroît entre 30 et 50 ans puis augmente de manière significative sur les dernières tranches d'âge. Quant au coût moyen, il augmente sur les premières tranches d'âge pour ensuite se stabiliser puis alterne les hausses et les baisses à partir de 55 ans.

Ce portefeuille d'étude est le support sur lequel va être opérée la démarche de modélisation présentée dans le chapitre suivant.

— Chapitre 3 —

Détermination de la prime pure à l'aide des GLM

Dans cette partie, l'étude détaillée de la garantie **Dégât des eaux sur le périmètre des appartements** sera présentée. L'objectif est de créer un modèle prédictif de la sinistralité future en utilisant des **M**odèles **L**inéaires **G**énéralisés, plus couramment appelées GLM (***G**eneralized **L**inear **M**odels*).

3.1 Cadre théorique

Pour une meilleure compréhension des GLM, les principaux résultats du modèle linéaire gaussien vont être rappelés.

3.1.1 Modèle linéaire gaussien

Un modèle linéaire a pour but de pouvoir exprimer une certaine variable aléatoire Y en fonction de plusieurs variables explicatives $X_1, ..., X_p$ supposées iid :

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{i,j} + \epsilon_i, \tag{3.1}$$

avec :

- $x_{i,j}$ la réalisation de la variable explicative X_j pour l'observation i ;
- β_j le paramètre inconnu mais non-aléatoire associé à la variable explicative X_j ;
- ϵ_i l'erreur de modélisation (exprimant l'information manquante) pour l'observation i, qui est une variable aléatoire suivant une loi normale de paramètres 0 et σ^2 : le modèle est bien spécifié en moyenne et la variance est une valeur constante pour tous les résidus (homoscédasticité). Il est, par ailleurs, supposé que les erreurs sont indépendantes entre elles.

Le modèle linéaire suppose une normalité dans la distribution des données, ce qui n'est pas le cas en assurance notamment. En effet, cela supposerait d'avoir des valeurs négatives pour des sinistres, ce qui ne correspond pas à la réalité des données. Les modèles linéaires généralisés vont donc être utilisés afin de prendre en considération la distribution des données. L'intérêt du modèle linéaire généralisé est qu'il permet de s'affranchir de cette hypothèse de normalité des observations du modèle linéaire gaussien et de l'étendre à la famille exponentielle.

3.1.2 Modèle linéaire généralisé

Dans le cadre des modèles linéaires généralisés, la distribution des Y_i n'est pas nécessairement normale mais doit être de la famille exponentielle (*Lopez, 2018*, [**8**]). Une distribution appartient à la famille de dispersion exponentielle si sa fonction de densité peut être écrite sous la forme :

$$f_{\theta,\phi}(y) = \exp\left(\frac{y\theta - a(\theta)}{\phi} + c_\phi(y)\right), \tag{3.2}$$

avec :

- θ le paramètre réel, appelé aussi paramètre naturel ;
- ϕ le paramètre de dispersion (strictement positif) ;
- $a(\theta)$ est de classe C^2 et convexe ;
- $c_\phi(y)$ ne dépend pas de θ.

Pour une variable aléatoire Y dont la densité est de la forme exponentielle, alors :

$$\mu = \mathbb{E}(Y) = a^{'}(\theta) \text{ et } \mathbb{V}(Y) = a^{''}(\theta)\phi. \tag{3.3}$$

Chacune des lois de la famille exponentielle possède une fonction de lien spécifique, dite fonction de lien canonique, permettant de relier l'espérance μ au paramètre naturel θ. En effet, un GLM suppose une relation plus générale entre les variables explicatives X et la variable réponse Y et peut s'écrire sous la forme générale :

$$\mu = \mathbb{E}(Y) = g^{-1}(X^T\beta), \tag{3.4}$$

où g représente une fonction de lien monotone et dérivable.

Ce type de modèles possède trois composantes :

- Une composante stochastique qui précise que les observations sont des variables aléatoires indépendantes avec une densité appartenant à la famille de dispersion exponentielle ;
- Une composante systématique qui attribue à chaque observation un prédicteur linéaire

$$\eta_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{i,j}. \tag{3.5}$$

- La troisième composante connecte les deux précédentes. L'espérance μ_i de Y_i est liée au prédicteur linéaire η_i par une fonction de lien telle que $\eta_i = g(\mu_i)$.

Pour estimer les différents paramètres $\beta_0, \beta_1, ..., \beta_p$ du modèle, il va falloir utiliser les notions de maximum de vraisemblance et de log-vraisemblance. Dans le cas des modèles exponentiels, la log-vraisemblance est donnée par :

$$l(\beta, \phi; y) = \sum_{i=1}^{n} \frac{Y_i\theta_i - a(\theta_i)}{\phi} + c_\phi(Y_i). \tag{3.6}$$

Pour maximiser cette quantité, la dérivée doit être annulée :

$$\frac{\partial l(\beta, \phi; y)}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial}{\partial \beta_j}\left(\frac{Y_i\theta_i - a(\theta_i)}{\phi} + c_\phi(Y_i)\right) = 0. \tag{3.7}$$

Les p équations se résolvent alors numériquement par une descente de gradient.

Le coefficient β_0 observé dans la formule 3.5 est appelé *intercept* et représente la classe de référence. Pour les autres coefficients β, Si $\beta_j > 0$ (respectivement < 0), alors une observation avec la modalité X_j aura un profil plus risqué (respectivement moins risqué) que la classe de référence.

3.1.3 Pénalisations

L'idée de la pénalisation est de sélectionner uniquement les variables qui contribuent le plus à l'explication de la variable réponse. Or, les modèles linéaires généralisés classiques prennent en compte toutes les variables et modalités, c'est à l'utilisateur de définir des seuils de significativité pour les écarter. Mais un nombre de variables élevé implique souvent une variance du modèle forte, et de ce fait, un sur-apprentissage des données. Ce sur-apprentissage peut être traduit comme le fait que le modèle ajuste correctement les données sur lesquelles il a été créé, mais très mal dès lors que les données sont nouvelles. Cela entraîne une perte du pouvoir de prédiction. De plus, réduire le nombre de variables permet d'améliorer l'interprétation des modèles. Trois méthodes consistant à ajouter une pénalité relative à la complexité du modèle afin de favoriser la parcimonie vont être présentées par la suite (*Lopez, 2018*, [**8**]).

La première procédure de pénalisation utilisée se nomme **LASSO** (***L**east **A**bsolute **S**hrinkage and **S**election **O**perator*). Elle utilise la notion de norme L_1. Ce terme supplémentaire va donc contraindre le programme d'optimisation à modifier la valeur des coefficients. Il s'agit d'augmenter volontairement le biais du modèle pour en réduire la variance, contribuant ainsi à rendre le modèle plus robuste. L'estimateur associé dans le cas d'une régression LASSO est le suivant :

$$\beta_{LASSO} = \arg\min_{\beta} \sum_{i=1}^{n} (Y_i - \beta' X_i)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \text{ avec } \lambda \in \mathbb{R}^+. \tag{3.8}$$

Il est pertinent de remarquer que si le coefficient de pénalisation λ vaut 0, alors l'expression de l'estimateur correspond à celle obtenue dans le cadre d'une régression non pénalisée. Ce paramètre λ est donc très important puisqu'il définit le poids attribué à la pénalité. Plus il est élevé, plus la pénalisation est forte et plus les coefficients estimés sont proches de zéro. Le calibrage est généralement effectué à l'aide d'une technique de validation croisée, qui sera présentée dans la suite de l'étude. L'avantage de cette méthode est de produire un résultat parcimonieux où un certain nombre de coefficients aura été forcé à 0. Mais des limites peuvent apparaître si le nombre de variables retenu est trop important et se rapproche du nombre d'observations.

La seconde procédure de pénalisation se nomme **Ridge** et utilise la notion de norme L_2. L'estimateur associé dans le cas d'une telle régression est le suivant :

$$\beta_{RIDGE} = \arg\min_{\beta} \sum_{i=1}^{n} (Y_i - \beta' X_i)^2 + \lambda \sum_{j=1}^{p} |\beta_j|^2 \text{ avec } \lambda \in \mathbb{R}^+. \tag{3.9}$$

L'avantage de cette méthode est de produire un estimateur explicite qui va permettre d'améliorer le conditionnement de la matrice des observations X. Mais il n'y aura qu'un lissage des coefficients, aucun d'entre eux ne sera annulé. La régression Ridge ne solutionne pas le problème des modèles en grande dimension.

Enfin, la dernière procédure est une combinaison des deux précédentes et se nomme **Elastic Net**. L'estimateur associé dans ce cas est le suivant :

$$\beta_{EN} = \arg\min_{\beta} \sum_{i=1}^{n} (Y_i - \beta' X_i)^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} |\beta_j|^2, \tag{3.10}$$

avec $\lambda_1 = \alpha\lambda, \lambda_2 = (1-\alpha)\lambda$ et α compris entre 0 et 1. Ce dernier paramètre permet de définir l'équilibre entre les régressions Ridge et LASSO. En effet, pour $\alpha = 1$, l'équation correspond à celle de l'approche LASSO, et pour $\alpha = 0$, il s'agit de Ridge.

Dans la suite de l'étude, **ces modèles pénalisés seront préférés aux modèles linéaires classiques** pour effectuer les différentes modélisations.

3.1.4 Choix de la loi de la distribution et de la fonction de lien

Les modèles GLM, qu'ils soient pénalisés ou non, sont des modèles paramétriques où une distribution est définie en amont, c'est-à-dire la structure des données à modéliser. La densité de la loi choisie au sein de la famille exponentielle doit décrire au mieux la structure des données. En ce sens, une distribution Gamma sera retenue pour les modèles représentant le coût des sinistres (les sinistres négatifs ou nuls ayant été écartés de la base d'étude) et une loi de Poisson pour ceux représentant le nombre de sinistres.

Pour chacun des modèles, la fonction de lien utilisée sera logarithmique, car cela permet d'obtenir une approche multiplicative qui prend en compte les effets des facteurs de risque de façon proportionnelle. Il faut cependant bien s'assurer que ces deux lois retenues appartiennent à la famille exponentielle :

- Pour une loi de Poisson de paramètre λ et pour tout y $\in \mathbb{N}$:

$$f_\lambda(y) = \frac{\exp(-\lambda)\lambda^y}{y!} = \exp(y\ln(\lambda) - \lambda - \ln(y!)) = \exp\left(\frac{y\theta - \exp(\theta)}{\phi} - \ln(y!)\right)$$
$$= \exp\left(\frac{y\theta - a(\theta)}{\phi} + c_\phi(y)\right), \tag{3.11}$$

  avec $\theta = \ln(\lambda)$, $\phi = 1$, $a(\theta) = \lambda = \exp(\theta)$ et $c_\phi(y) = -\ln(y!)$.

- Pour une loi Gamma de paramètres α et β et pour tout y $\in \mathbb{R}^+$:

$$f_{\alpha,\beta}(y) = \frac{\beta}{\alpha}\Gamma(\alpha)y^{\alpha-1}\exp(-\beta y) = \exp((\alpha-1)\ln(y) - \beta y + \alpha\ln\beta - \ln\Gamma(\alpha)).$$

  En posant $\mu = \mathbb{E}(Y) = \frac{\alpha}{\beta}$ et $\nu = \alpha$, la densité peut se réécrire :

$$f_{\mu,\nu} = \frac{1}{\Gamma(\nu)}(\frac{\nu y}{\mu})^\nu \exp(\frac{\nu y}{\mu})\frac{1}{y} = \exp\left(\frac{y\theta - a(\theta)}{\phi} + c_\phi(y)\right), \tag{3.12}$$

  avec $\theta = -\frac{1}{\mu}$,$\phi = \frac{1}{\nu}$, $a(\theta) = -\ln(-\theta)$.

Ces deux lois appartiennent donc bien à la famille exponentielle.

3.1.5 Indicateurs de performance

Il existe plusieurs indicateurs qui permettent de mesurer la performance d'un modèle :

- La déviance permet de quantifier la qualité de la régression. Le modèle estimé est comparé au modèle dit saturé ou parfait. Le calcul de la déviance se fait par la comparaison de la log-vraisemblance (LV) entre le modèle retenu et le modèle dit saturé (modèle possédant autant de paramètres que d'observations et estimant donc exactement les données) :

$$\text{Déviance} = 2 \times log(LV_{max} - LV). \tag{3.13}$$

- L'indice de Gini fournit une mesure de la qualité de la segmentation du modèle. Il est calculé à partir de la fonction représentée par la courbe de Lorenz (développée initialement pour permettre de mesurer les inégalités de richesse au sein d'une population, mais qui peut être transposée à une donnée de répartition statistique quelconque).

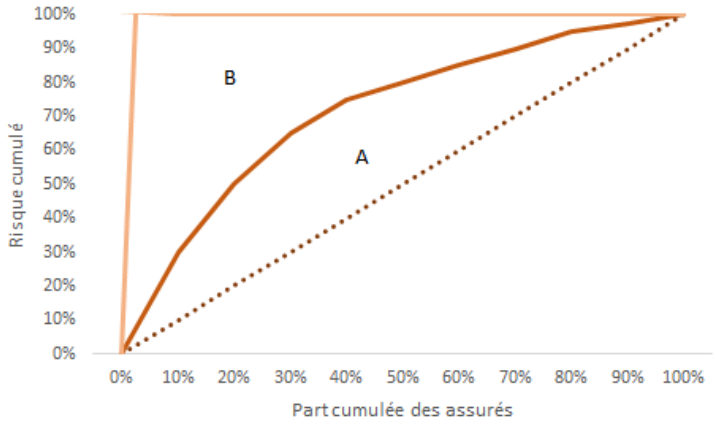


FIGURE 3.1 – Courbe de Lorenz et indice de Gini

Sur la figure 3.1, la première bissectrice représente un cas d'égalité parfait où il y a une mutualisation égale du risque sur l'ensemble des observations. La zone A correspond à l'aire entre la courbe de Lorenz et la bissectrice et la zone B à l'aire au-dessus de la courbe de Lorenz. L'indice de Gini est défini de la manière suivante :

$$\text{Gini} = \frac{A}{A+B}. \tag{3.14}$$

- La RMSE (pour ***R**oot **M**ean **S**quared **E**rror*) correspond à la racine de l'erreur quadratique moyenne et est calculée à partir des résidus :

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}. \tag{3.15}$$

- Les résidus de Pearson se basent sur les résidus observés $r_i = y_i - \hat{y}_i$ en les normalisant :

$$r_i^{Pearson} = \frac{y_i - \hat{y}_i}{\sqrt{var(\hat{y}_i)}}. \tag{3.16}$$

- L'erreur totale permet d'évaluer la qualité globale d'ajustement du modèle :

$$\text{Erreur totale} = \frac{\sum \text{Valeurs prédites} - \sum \text{Valeurs réelles}}{\sum \text{Valeurs réelles}}. \tag{3.17}$$

Une fois ce cadre théorique présenté, il est possible maintenant d'étudier les résultats de la modélisation obtenus à l'aide du logiciel R (modules *caret* et *h2o*).

3.2 Modélisation de la fréquence

Cette section va présenter les résultats de modélisation pour la prédiction du nombre annuel de sinistres au titre de la garantie Dégât des eaux, sur le périmètre des appartements.

3.2.1 Mise en oeuvre de la modélisation

Avant de commencer la modélisation de la fréquence de sinistres, il faut tenir compte d'une caractéristique que possèdent les données. En effet, toutes les images de risque ne sont pas nécessairement observées sur une période d'un an. Ainsi, il ne faut pas modéliser de la même façon des risques qui auraient été exposés sur des durées différentes. Par conséquent, l'hypothèse de linéarité du risque par rapport à l'exposition a été faite. Afin de prendre en considération cet effet pour expliquer la variable aléatoire N modélisant le nombre de sinistres, il est possible d'intégrer l'exposition e du contrat dans la régression qui utilise un modèle poissonnien et une fonction de lien logarithmique. En effet, comme l'espérance de la variable considérée devient λe, la régression s'écrit alors :

$$\mathbb{E}(N|X,e) = e \times (\exp(\beta_0) \times \exp(\beta_1 X_1) ... \times \exp(\beta_p X_p)) = \exp(x^T\beta + \ln e). \tag{3.18}$$

Ainsi, la prise en compte de l'exposition de l'image de risque consiste à ajouter une nouvelle variable explicative dont le coefficient β est connu et constant à 1. Cette variable supplémentaire se nomme *offset*. Avant de procéder à la modélisation, une première étape consiste à scinder la base de données en deux échantillons : une base dite d'apprentissage qui contiendra 80% des données de la base initiale et une base dite de validation qui contiendra les 20% restants.

L'affectation des données entre ces deux bases a été effectuée de manière aléatoire. La modélisation sera calibrée sur la base d'apprentissage et lorsque les paramètres auront été fixés, ils seront appliqués sur la base de validation afin de s'assurer de la robustesse du modèle. Cette dernière base permet d'évaluer objectivement l'erreur réelle de prédiction du modèle.

Pour l'évaluation des modèles, la méthode de validation croisée des *K-folds* va être effectuée sur la base d'apprentissage. Cette base va être divisée en K échantillons de taille égale. Le modèle sera calibré sur les (K-1) échantillons et testé sur le $K^{ème}$. A l'issue de cette étape, une première estimation est obtenue pour les coefficients β associés à chaque variable explicative du modèle et noté $\hat{\beta}$. L'opération est répétée K fois et ainsi pour chaque coefficient, K estimations différentes sont obtenues. L'estimation finale est calculée comme la moyenne des différentes valeurs obtenues. Ainsi, en augmentant le nombre d'observations utilisées pour estimer β et en l'estimant par une moyenne de plusieurs itérations, la variance de l'estimateur est diminuée. La valeur de K retenue pour l'étude sera de 10.

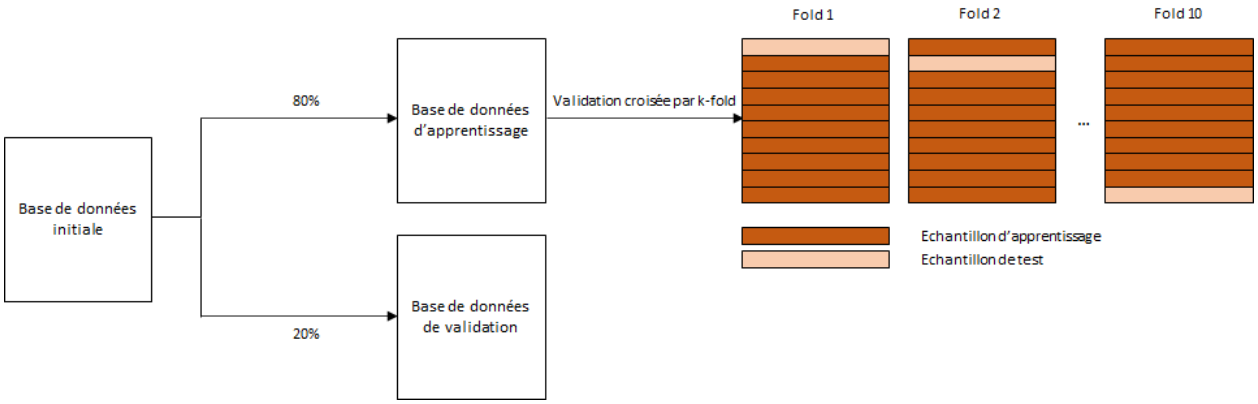


FIGURE 3.2 – Récapitulatif du découpage des bases

Une fois les bases de modélisation préparées, il est nécessaire de passer à l'optimisation des paramètres pour chacun des modèles de régression pénalisée afin d'obtenir la prédiction la plus précise et éviter le sur-apprentissage.

La métrique d'évaluation permettant de sélectionner le λ optimal par validation croisée est la déviance de Poisson, définie comme :

$$D_{Poisson} = -2 \sum_{i=1}^{N} \left\{ y_i \ln(\frac{y_i}{\hat{y}_i}) - (y_i - \hat{y}_i) \right\}. \tag{3.19}$$

Concernant, la technique Elastic Net, l'optimisation des différents paramètres s'effectue par le biais d'une grille de recherche (ou *grid search*). C'est une méthode d'optimisation qui va permettre de tester une série de paramètres et de comparer les performances pour en déduire le meilleur paramétrage possible. Une grille de paramètres pour λ et α va donc être définie. A une valeur de α fixée ($\alpha \in [0; 1]$), une grille de valeurs de λ est définie.

Pour chaque valeur de λ, la base d'entraînement est découpée de manière aléatoire en K $folds$. Par validation croisée, une moyenne d'évaluation de la prédiction des modèles créés pour chaque valeur de λ peut être calculée. La valeur optimale de λ créant le modèle ayant la meilleure prédiction peut alors être sélectionnée. La figure 3.3 illustre le principe. Le processus est répété autant de fois qu'il y a de valeurs différentes pour α. Le couple optimal (α, λ) retenu sera celui qui possède l'indicateur de mesure d'erreur le plus faible. Ces paramètres définis sur la base d'apprentissage sont ensuite appliqués sur la base de validation afin de juger de la qualité de prédiction en utilisant différents indicateurs (RMSE et Gini).

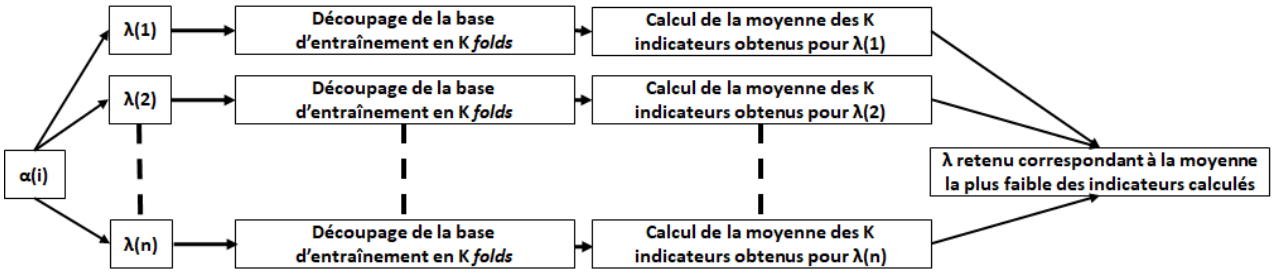


FIGURE 3.3 – Processus d'optimisation des paramètres pour les GLM Elastic Net

La grille retenue pour α dans cette étude correspond à 101 valeurs allant de 0 à 1, chacune incrémentée par un pas de 0,01. La phase d'optimisation étant terminée, les résultats obtenus vont être présentés.

3.2.2 Comparaison des modèles obtenus

La figure 3.4 représente, pour chaque modèle pénalisé, la valeur du paramètre $\ln(\lambda)$ ainsi que le nombre de modalités sélectionnées en fonction du critère de la déviance de Poisson.

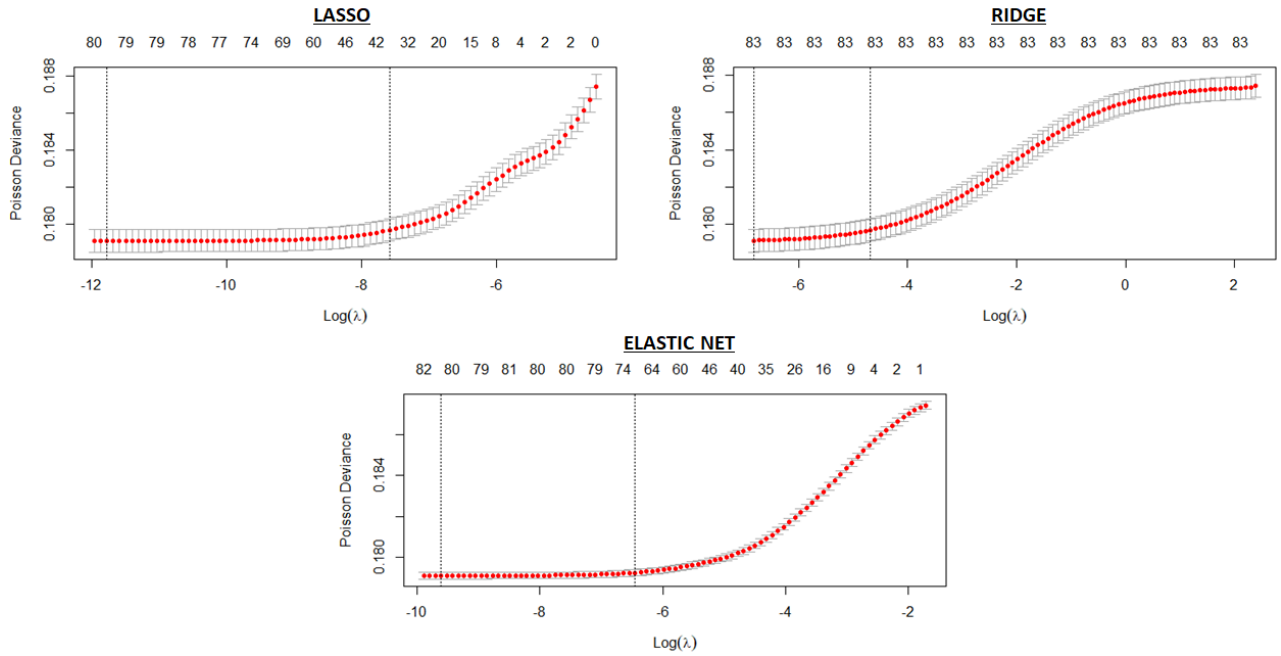


FIGURE 3.4 – Déviance de Poisson et nombre de variables retenues pour chaque modèle

Deux droites verticales sont présentes sur les graphiques :

- La première est celle associée à la plus faible valeur du logarithme de λ qui minimise l'erreur de déviance et notée λ^{min} ;
- La seconde est celle associée à la plus petite valeur du logarithme de λ pour laquelle l'erreur est majorée par l'écart-type de λ^{min}, notée λ^{1se}.

En pratique, le modèle obtenu avec λ^{min} apparaît parfois complexe et a tendance à sur-apprendre. Ici, cette approche retiendrait plus de 80 modalités de variables (valeur présente sur l'axe des abscisses en haut). A l'inverse, λ^{1se} correspond à un modèle bien plus simple, puisque davantage pénalisé, dont la précision est comparable à celle du meilleur modèle. Il s'agit d'une approche qui privilégie la parcimonie.

Modèle	$\ln(\lambda)$	λ	Modalités retenues
LASSO	-7,586	$5{,}07 \times 10^{-4}$	37
Ridge	-4,679	$9{,}29 \times 10^{-3}$	83
Elastic Net	-6,447	$1{,}58 \times 10^{-3}$	68

TABLE 3.1 – Résultats de la pénalisation - Fréquence

Pour la régression LASSO, l'effet de la norme L_1 dans la pénalisation est bien visible car cela permet de retirer du modèle plus de la moitié des modalités. A l'inverse, en conservant l'ensemble des modalités, la pénalisation Ridge permet de maximiser la précision, tout en effectuant un lissage des coefficients, mais rend l'interprétation des modèles plus complexe.

Enfin, l'approche Elastic Net se distingue par un compromis entre les deux méthodes précédentes : une sélection de modalités est effectuée (pour une valeur optimale de α égale à 0,6), mais pas de manière aussi marquée que pour la régression LASSO.

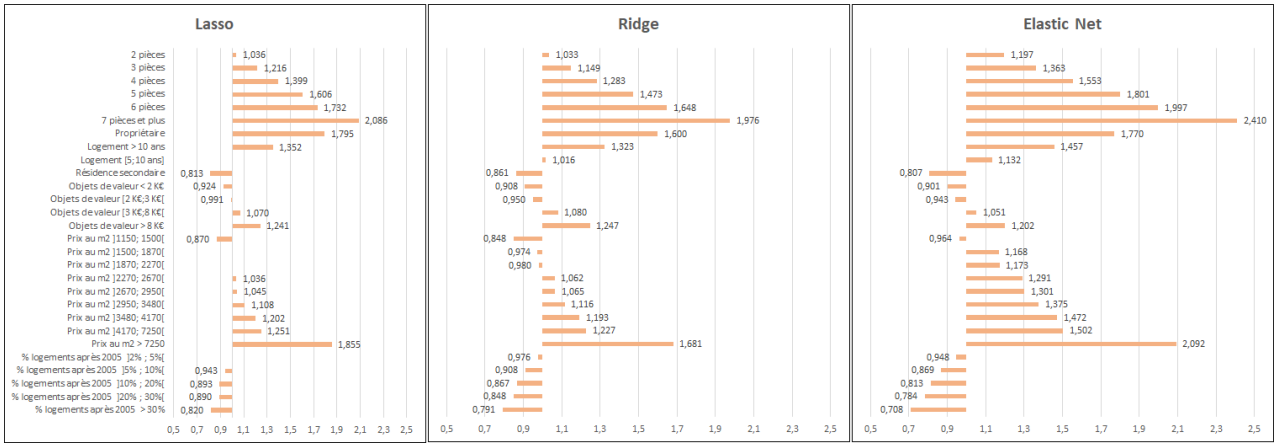


FIGURE 3.5 – Coefficients associés aux principales modalités pour chaque GLM pénalisé - Fréquence

La figure ci-dessus permet de comparer la valeur des coefficients des principales modalités pour chacune des pénalisations. Les résultats énoncés dans le tableau 3.1 se retrouvent visuellement : la régression LASSO n'a pas défini de coefficients pour toutes les modalités, contrairement aux approches Ridge et Elastic Net dans une moindre mesure. Cependant, les tendances prédites sont similaires d'une méthodologie à une autre.

Le risque lié au dégât des eaux (sur le périmètre des appartements) croît de manière significative avec le nombre de pièces. Par exemple, un appartement de quatre pièces possède une fréquence de sinistres entre 28% et 55% plus élevée (selon les méthodes) qu'un appartement d'une pièce. Cette corrélation positive se retrouve également pour le montant d'objets de valeur déclaré lors de la souscription. Posséder plus de 8 000 € d'objets de valeur entraîne, à caractéristiques de risque équivalentes, une hausse jusqu'à 25% de la probabilité d'avoir un sinistre. De même, le fait d'être propriétaire ou d'assurer un bien de plus de dix ans sont des facteurs de risque. A l'inverse, le fait que le bien soit destiné à un usage de résidence secondaire influe positivement sur le risque en le diminuant de plus de 10%. Enfin, plusieurs variables externes ont été retenues dans le modèle. Plus le prix moyen au m^2 dans la commune est élevé, plus le risque associé l'est également. Par exemple, un bien se situant dans une zone où le prix au m^2 est de 3 000 € aura un risque de sinistre au moins 10% plus élevé que pour un même bien se situant dans une zone à 2 000 €.

Au global, les variables les plus influentes sur l'apparition de la sinistralité sont les suivantes par ordre d'importance : le nombre de pièces, le prix moyen au m^2 dans la commune (ou l'arrondissement le cas échéant), la qualité de l'occupant (locataire ou propriétaire), l'ancienneté du logement, le type de résidence (principale ou secondaire), la part de logements construits après 2005 ainsi qu'avant 1945 dans la zone du bien assuré.

Le tableau 3.2 fournit quelques métriques d'évaluation afin de juger de la qualité des différents modèles. Ces derniers semblent robustes car les résultats sont stables entre les bases d'apprentissage et de validation. De plus, ils ne diffèrent peu en terme de qualité de prédiction et de segmentation. Sachant que le modèle LASSO possède un nombre limité de variables et que la valeur de Gini est la plus élevée des trois modélisations, ce modèle sera préféré aux deux autres. Il est à noter que l'indicateur RMSE n'est pas facilement interprétable car il est appliqué sur des fréquences (prédites et observées), et par conséquent, les résultats sont très faibles.

	Base d'apprentissage		Base de validation	
Modèle	RMSE	Gini	RMSE	Gini
LASSO	**0,2067**	**34,81%**	**0,2067**	**34,74%**
Ridge	0,2069	33,53%	0,2069	33,44%
Elastic Net	0,2068	34,28%	0,2068	34,16%

TABLE 3.2 – Métriques d'évaluation des GLM - Fréquence

3.2.3 Validation du modèle

Pour valider la qualité du modèle LASSO retenu, plusieurs étapes sont à effectuer :

- Analyse de la stabilité des indicateurs de validation du modèle ;
- Vérification que les résidus du modèle sont centrés ($E(\epsilon) = 0$). Il s'agit en effet d'une hypothèse de base des modèles linéaires généralisés. En parallèle, une représentation graphique sous forme d'un nuage de point est effectuée afin de détecter une éventuelle tendance ;
- Comparaison des fréquences observées et prédites ;
- Analyse de la stabilité des coefficients dans le temps.

Le graphique ci-dessous permet de constater que l'indicateur de Gini est stable sur chacun des sous-échantillons de la base de validation, les différentes courbes de Lorenz étant superposées. Le modèle ne fluctue pas lorsque de nouvelles données sont utilisées, preuve d'une absence de sur-apprentissage.

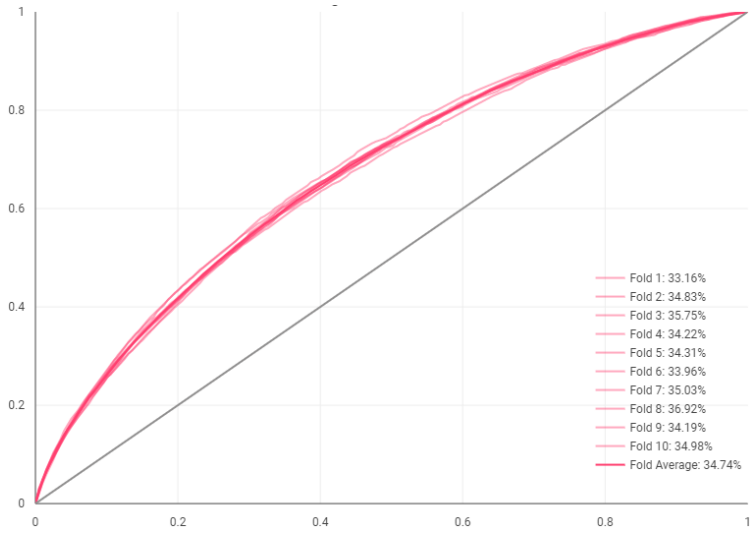


FIGURE 3.6 – Variation de Gini en fonction du découpage des données

Une fois cette étape de stabilité validée, il est possible de passer à l'étude des résidus du modèle. L'étude des résidus est indispensable pour la validation du modèle. Cela permet de valider la pertinence des choix de distributions et de fonction lien.

La moyenne des résidus obtenue sur les bases d'apprentissage et de validation est de respectivement -0,0058 et -0,0062. Les résidus semblent donc centrés, mais une analyse graphique va cependant être effectuée pour confirmer ce point. Le nuage des résidus standardisés en fonction des valeurs ajustées doit être centré autour de 0 et ne pas présenter de tendance, ce qui implique une erreur de modélisation faible.

A la lecture des deux nuages de points ci-dessous, aucune tendance ne semble se dessiner et les points sont répartis de manière symétrique autour de 0. Les résultats obtenus sur les résidus sont donc satisfaisants.

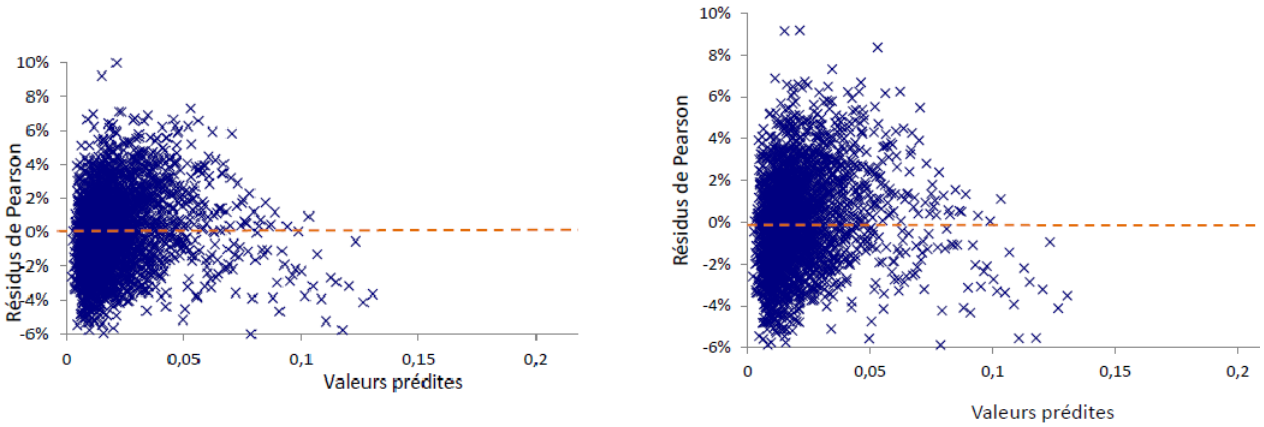


FIGURE 3.7 – Résidus de Pearson sur les bases d'apprentissage (gauche) et de validation (droite)

Il est nécessaire maintenant de comparer les fréquences observées et prédites sur la base de validation. La validation par quantiles permet d'estimer la performance de prédiction du modèle. Cette méthode consiste à construire des quantiles sur la fréquence prédite issue du modèle, puis de tracer les fréquences observées et prédites. Si le modèle est performant, la fréquence prédite doit converger vers la fréquence observée.

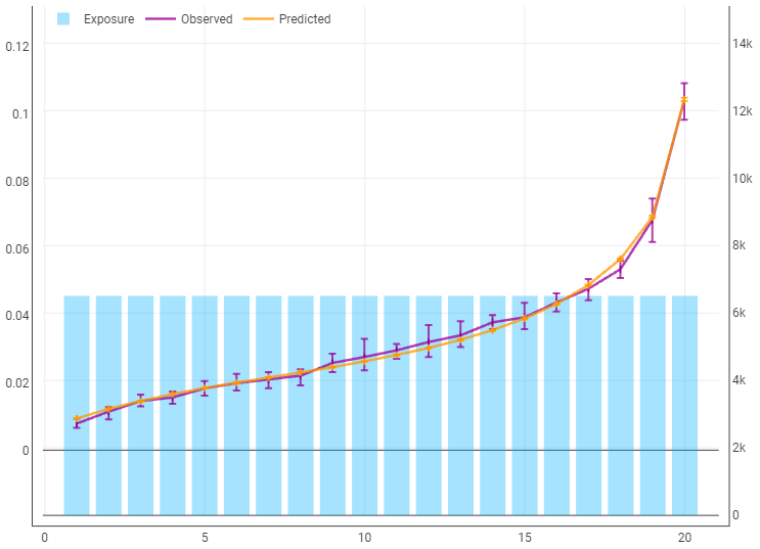


FIGURE 3.8 – Comparaison des fréquences observées et prédites par quantiles

Cette validation visuelle est satisfaisante puisque la fréquence prédite est globalement similaire à la fréquence observée (confirmé numériquement par des valeurs de respectivement 3,36% et 3,33%).

Une dernière étude est d'analyser la stabilité des paramètres dans le temps du fait que les données utilisées pour la modélisation sont sur plusieurs d'années d'observation. En effet, l'objectif étant de prédire la sinistralité dans le futur quelle que soit la période, la modélisation ne doit pas être instable dans le temps. Seule l'analyse sur la variable la plus importante du modèle (qui est le nombre de pièces) est présentée dans ce mémoire, mais la vérification a été effectuée sur l'ensemble des variables significatives retenues.

Sur le graphique 3.9, pour chaque modalité de la variable sont représentées l'exposition (où chaque strate correspond à une année différente) ainsi que les fréquences observée et prédite pour chaque année entre 2017 et 2019. Il est tout d'abord possible de remarquer que l'exposition est constante sur la période, et ce quel que soit le nombre de pièces observé. Concernant les fréquences observées annuelles, elles sont très proches des fréquences prédites.

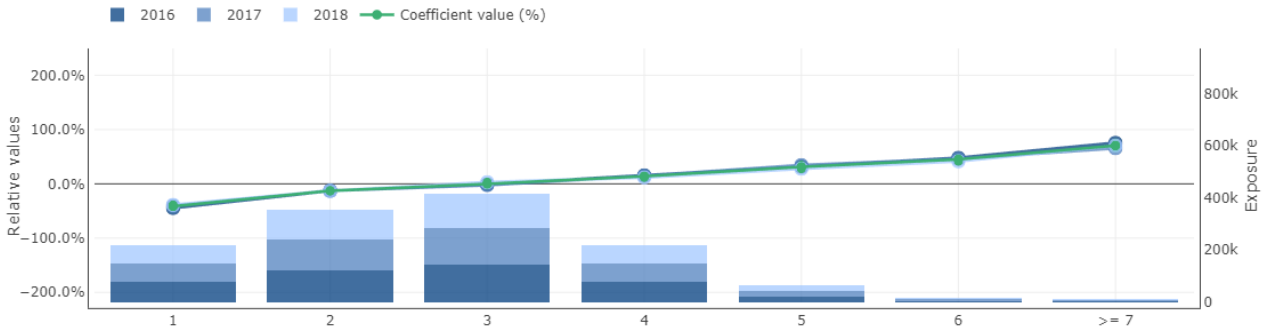


FIGURE 3.9 – Stabilité temporelle des coefficients

L'ensemble de ces tests a permis de valider les résultats générés par le modèle LASSO retenu (indicateurs stables, résidus homogènes) et d'exclure la présence de sur-apprentissage. Il est donc maintenant possible de passer à la modélisation de la seconde composante de la prime pure qu'est le coût moyen.

3.3 Modélisation du coût moyen

Les résultats de modélisation obtenus pour la prédiction de la charge annuelle de sinistres au titre de la garantie Dégât des eaux, sur le périmètre des appartements, vont pouvoir maintenant être présentés.

3.3.1 Comparaison des modèles obtenus

Dans cette section, seules les images de risque pour lesquelles une charge de sinistre est associée sont conservées. Cependant, tout comme pour la fréquence, il faut tenir compte d'une caractéristique que possèdent les données. En effet, la charge doit être divisée par le nombre de sinistres survenus. Afin de prendre en considération cet effet, cette dernière variable a été considérée comme une variable explicative *offset* dont le coefficient β est connu et constant à 1.

La figure 3.10 représente, pour chaque modèle pénalisé, la valeur du paramètre $\ln(\lambda)$ en fonction du critère de la déviance résiduelle.

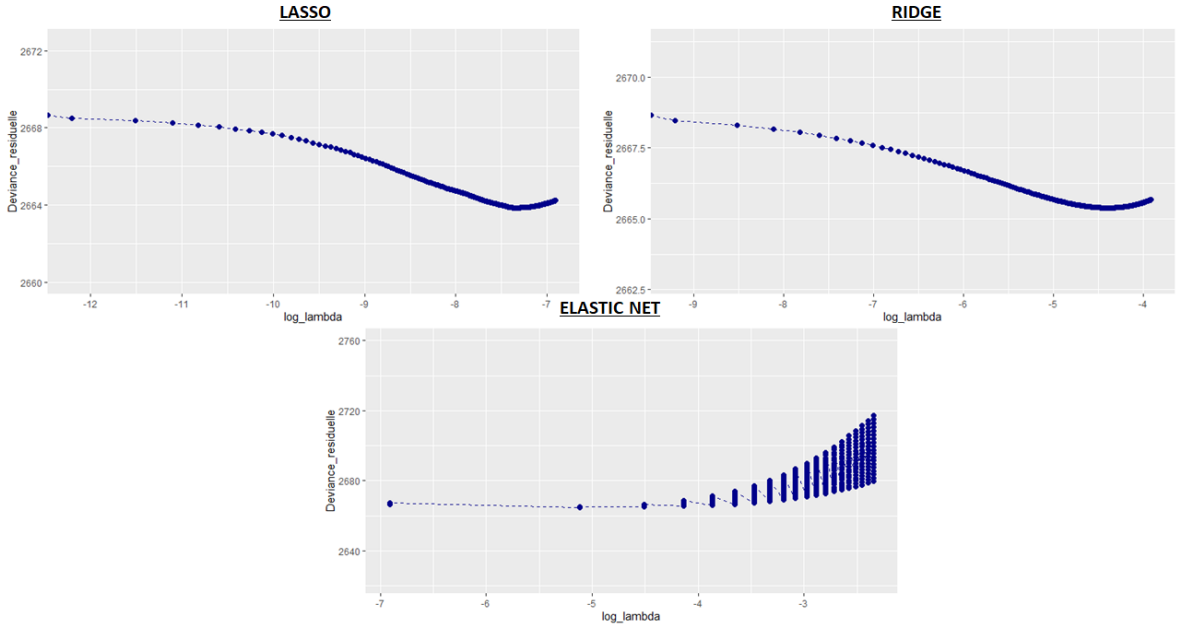


FIGURE 3.10 – Déviance résiduelle pour chaque modèle

La valeur retenue sera celle associée à la valeur la plus faible de déviance résiduelle. Le tableau ci-dessous permet de préciser les informations apportées par les graphiques (valeur exacte du paramètre λ, nombre de modalités retenues) :

Modèle	$\ln(\lambda)$	λ	Modalités retenues
LASSO	-7,323	$6{,}60 \times 10^{-4}$	56
Ridge	-4,398	$1{,}23 \times 10^{-2}$	83
Elastic Net[9]	-5,116	$6{,}00 \times 10^{-3}$	62

TABLE 3.3 – Résultats de la pénalisation - Coût moyen

Les coefficients obtenus par les différents modèles peuvent maintenant être étudiés et comparés. Tout comme pour la modélisation de la fréquence, les tendances prédites pour le coût moyen sont similaires d'une méthodologie à une autre. Par exemple, pour le nombre de pièces, le risque lié au dégât des eaux en terme de charge décroît, à caractéristiques de risque équivalentes, entre les risques disposant d'une seule pièce et ceux en disposant de trois (-3%) puis ensuite augmente de manière significative.

Ces fluctuations se retrouvent également pour d'autres variables telles que le montant d'objets de valeur ou de capitaux déclaré lors de la souscription. En effet, le fait de posséder jusqu'à 2 000 € d'objets de valeur est associé à un risque plus faible que le fait de ne pas en avoir. Cela peut paraître contre-intuitif mais s'explique par le fait que les modèles effectués sont dits non-contraints, c'est-à-dire qu'ils ne sont pas contraints par des critères commerciaux. Au-delà de ce seuil de 2 000 €, une augmentation continue du coût moyen est observée. De même, le fait d'être propriétaire ou d'assurer un bien destiné à un usage de résidence secondaire sont considérés comme des facteurs de risque.

9. La valeur de α associée est de 0,095

Enfin, plus le prix moyen au m^2 est élevé, plus le coût d'un sinistre l'est également. Au global, les variables les plus influentes sur la charge de la sinistralité sont par ordre d'importance : le nombre de pièces, le prix moyen au m^2, la qualité de l'occupant (locataire ou propriétaire), le type de résidence (principale ou secondaire), la part de logements construits avant 1945 dans la zone du bien assuré ainsi que le montant d'objets de valeur déclaré à la souscription.

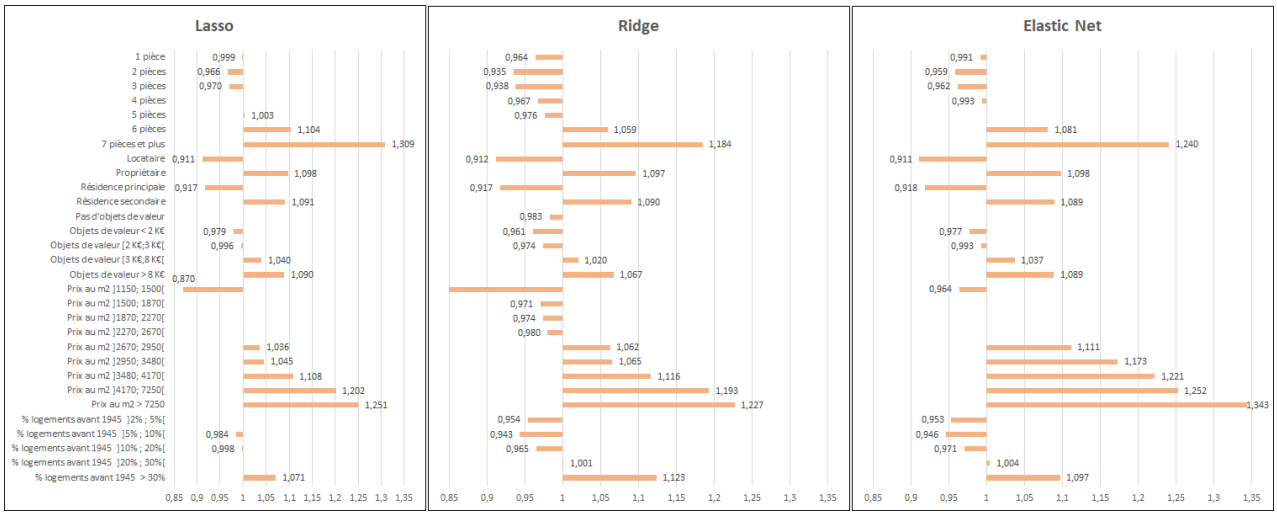


FIGURE 3.11 – Coefficients associés aux principales modalités pour chaque GLM pénalisé - Coût moyen

Il est donc maintenant nécessaire d'étudier les métriques d'évaluation afin de sélectionner le meilleur modèle sur la composante du coût moyen.

	Base d'apprentissage		Base de validation	
Modèle	RMSE	Gini	RMSE	Gini
LASSO	1 741,48	12,08%	1 761,46	11,90%
Ridge	1 742,46	11,63%	1 762,02	11,01%
Elastic Net	**1 739,22**	**12,25%**	**1 750,25**	**11,93%**

TABLE 3.4 – Métriques d'évaluation - Coût moyen

Les résultats obtenus sont très proches d'un modèle à l'autre. Cependant, il a été décidé de retenir le modèle Elastic Net car il possède la RMSE la plus faible (que ce soit sur la base d'apprentissage ou de validation) ainsi que le Gini le plus élevé. Enfin, ce modèle ne sera guère plus complexe que celui proposé par le modèle LASSO (62 modalités contre 56, comme indiqué dans le tableau 3.3). Les performances globales de ce modèle vont maintenant être étudiées.

3.3.2 Validation du modèle

Comme énoncé précédemment dans la section 3.2.3, plusieurs étapes d'analyse sont à effectuer pour valider le modèle retenu. Les résultats obtenus sont les suivants :

- Les résidus sont centrés et symétriques autour de 0 ;

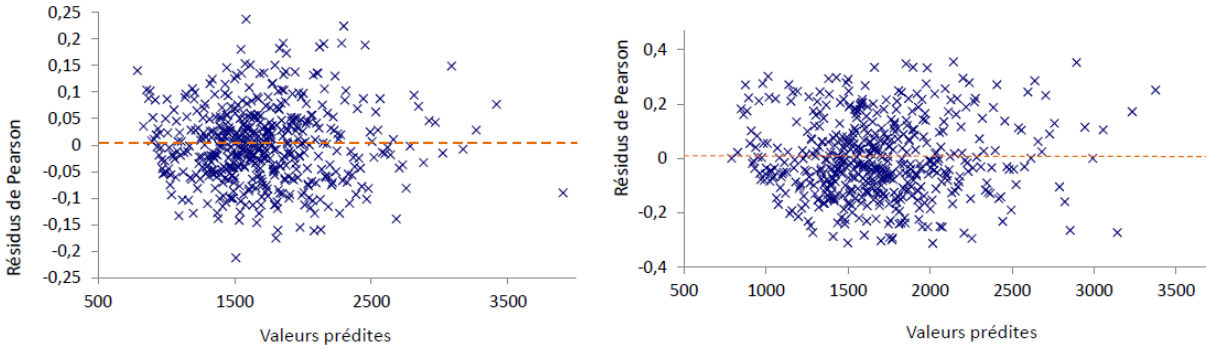


FIGURE 3.12 – Résidus de Pearson sur les bases d'apprentissage (gauche) et de validation (droite)

- L'indicateur de Gini est stable sur chacun des sous-échantillons de la base de validation. De plus, le coût moyen prédit est relativement proche du coût moyen observé, et ce quel que soit le quantile étudié. Enfin, la stabilité des paramètres pour le montant de capital déclaré a été représentée graphiquement. A l'exception d'un écart en 2018 sur la tranche]16 000 € ;19 000 €], les coefficients sont similaires d'une année à l'autre, tout comme l'exposition.

FIGURE 3.13 – Stabilité du Gini (haut gauche), Comparaison du coût moyen observé et prédit par quantiles (haut droite) et Stabilité temporelle des coefficients (bas)

L'ensemble de ces tests a permis de valider les résultats générés par le modèle Elastic Net retenu (indicateurs stables, résidus homogènes) et d'exclure à nouveau la présence de sur-apprentissage.

3.4 Synthèse

Afin de modéliser la prime pure de la garantie Dégât des eaux sur le périmètre des appartements, il a été décidé de retenir, dans cette section, un modèle LASSO pour la fréquence de sinistres et un modèle Elastic Net pour la charge moyenne de sinistres. Les variables retenues (par ordre d'importance) proviennent des informations liées au contrat (noir) ou de données externes (marron) :

Importance	Modèle de fréquence	Modèle de coût moyen
1.	Nombre de pièces	Nombre de pièces
2.	Prix moyen au m^2 dans la commune	Prix moyen au m^2 dans la commune
3.	Qualité de l'occupant	Qualité de l'occupant
4.	Ancienneté du logement	Type de résidence
5.	Type de résidence	Part de logements construits avant 1945
6.	Part de logements construits après 2006	Montant d'objets de valeur
7.	Part de logements construits avant 1945	-

TABLE 3.5 – Variables les plus importantes

Il est intéressant de remarquer que le risque associé au dégât des eaux est, au global, caractérisé par les mêmes facteurs, que ce soit pour la fréquence ou le coût moyen. En effet, le nombre de pièces, le prix moyen au m^2, le fait que l'assuré soit locataire ou propriétaire ainsi que les informations concernant l'ancienneté du bien sont les variables explicatives de la sinistralité. Il est à noter que si ces mêmes variables arrivent à segmenter de manière relativement correcte la fréquence (indice de Gini de près de 35%), cela n'est cependant pas le cas pour le coût moyen (indice de Gini de 12% seulement).

L'utilisation des données externes étant nouvelle, il a été décidé d'effectuer les mêmes modèles en les écartant pour pouvoir mesurer leur impact sur la qualité de prédiction et de segmentation des modèles. Au niveau de la fréquence, leur apport permet d'augmenter le Gini de 4,76 points (soit une hausse de 16%). Quant au coût moyen, l'amélioration est moindre mais reste tout de même significative (+ 0,96 point, soit +8%).

Ces premiers résultats obtenus à partir des modèles linéaires pénalisés vont être maintenant comparés avec ceux obtenus par des méthodes d'apprentissage statistique.

— Chapitre 4 —

Utilisation de méthodes d'apprentissage statistique

Deux méthodologies d'apprentissage statistique vont être analysées : les forêts aléatoires plus communément appelées *Random Forests* et le ***G**radient **B**oosting **M**achine* (GBM). Contrairement à la statistique classique qui requiert de formuler des hypothèses sur la structure et la distribution des données, la théorie de l'apprentissage statistique ne formule qu'une seule hypothèse : les données à prédire sont générées de façon iid. Ainsi, le but est de construire un algorithme qui va apprendre à prédire la valeur de Y en fonction des valeurs explicatives X : $Y = f(X) + \epsilon$, où ϵ est un bruit centré, avec dans le cadre d'une régression, $f(X) = \mathbb{E}[Y|X = x]$. La base de modélisation utilisée sera différente de celle utilisée pour les GLM : les variables continues telles que le montant de capitaux ou l'âge du client ne seront pas modifiées en variables catégorielles.

4.1 Cadre théorique

Pour une meilleure compréhension des techniques du *Random Forest* et du GBM, le principe général des arbres de décision CART (***C**lassification **A**nd **R**egression **T**rees*) va être présenté dans un premier temps.

4.1.1 Arbres de décision

Les arbres de décision qui regroupent les arbres de classification et de régression sont des outils non paramétriques de segmentation (*Boyer, 2020*, [**2**]). L'objectif est de détecter des critères permettant de répartir les observations en deux classes. L'algorithme utilisé sélectionne tout d'abord la variable qui permet d'avoir deux sous-ensembles distincts les plus homogènes possibles et les plus éloignés l'un de l'autre puis il choisit la façon optimale de découper par rapport à cette variable. De ce noeud initial, appelé noeud racine, naissent deux branches correspondant aux deux réponses possibles. L'arbre va orienter l'observation dans l'une des deux branches, laquelle engendrera soit un nouveau noeud (et donc un nouveau test conditionnellement au test précédent) soit un noeud final (appelé feuille).

Ainsi, à chaque noeud, un arbre binaire est construit et à chaque étape, une variable explicative est sélectionnée et une nouvelle branche est créée. Le choix optimal de la variable X ainsi que du seuil de séparation à chaque noeud dépend de la réduction maximale de l'impureté (mesurée par la variance au sein des noeuds dans le cadre d'une régression et par l'indice de Gini dans le cadre d'une classification).

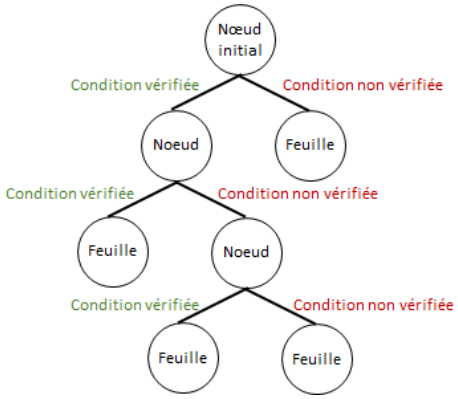


FIGURE 4.1 – Schéma simplifié d'un arbre de décision

La construction d'un arbre de décision peut se décomposer en trois étapes :

- Initialisation : la partition contient une seule prédiction, constituant le noeud initial ;
- Expansion : l'arbre est agrandi jusqu'à ce que le critère d'arrêt soit atteint. Il peut s'agir du nombre d'éléments dans chaque feuille ou du nombre total de feuilles ;
- Élagage : les branches de l'arbre qui n'améliorent pas la qualité de prédiction sont supprimées.

La principale qualité des arbres de décision est qu'ils sont très facilement lisibles et interprétables en un ensemble de règles simples. En revanche, ils sont instables et les prédictions calculées varient beaucoup.

4.1.2 Random Forest

Afin de gagner en efficacité et en précision, le partitionnement est parfois répété un grand nombre de fois. La solution générale s'écrit alors comme la combinaison des réponses de chacun. Le *Random Forest* repose sur le principe de *Bagging* qui consiste à agréger plusieurs modèles entre eux dans le but d'obtenir un seul résultat prédictif (*Hastie, 2013*, [**6**]). Une forêt aléatoire est donc un ensemble d'arbres de décisions pour lesquels la base de données utilisée de chaque arbre est constituée aléatoirement. Cette méthode permet de corriger le manque de robustesse dans le cas où un seul arbre de régression est utilisé pour la prédiction et ainsi de proposer une meilleure stabilité et fiabilité des résultats, en palliant au sur-apprentissage, principal défaut des arbres de décision. Le fait de construire un grand nombre d'arbres permet de réduire considérablement la variance de l'estimation et d'obtenir des estimateurs plus consistants. Le résultat de l'utilisation des arbres de décision est le découpage d'une variable aléatoire en différentes classes homogènes. Pour chacune d'entre elles, la prédiction sera la moyenne des prédictions observées au sein de celle-ci. Il existe deux paramètres à fixer :

- Nombre minimal d'observations par feuille : si ce paramètre est très petit, la prédiction sera extrêmement fine sur la base d'apprentissage, mais difficile à généraliser sur la base de validation (situation de sur-apprentissage). A l'inverse, si ce paramètre est choisi trop grand, la prédiction ne sera pas suffisamment segmentée ;
- Nombre de variables considérées aléatoirement pour construire l'arbre : ne prendre uniquement qu'un échantillon des variables explicatives pour chaque arbre permet réellement de les décorréler. En effet, si une variable explicative explique très fortement la variable réponse Y, alors la majorité des arbres construits l'utiliseront pour découper l'espace dans le premier noeud. Ainsi, la plupart des arbres se ressembleront, et les prédictions de ces derniers seront fortement corrélées. Par conséquent, plus ce paramètre est grand, plus le risque de sur-apprentissage est important.

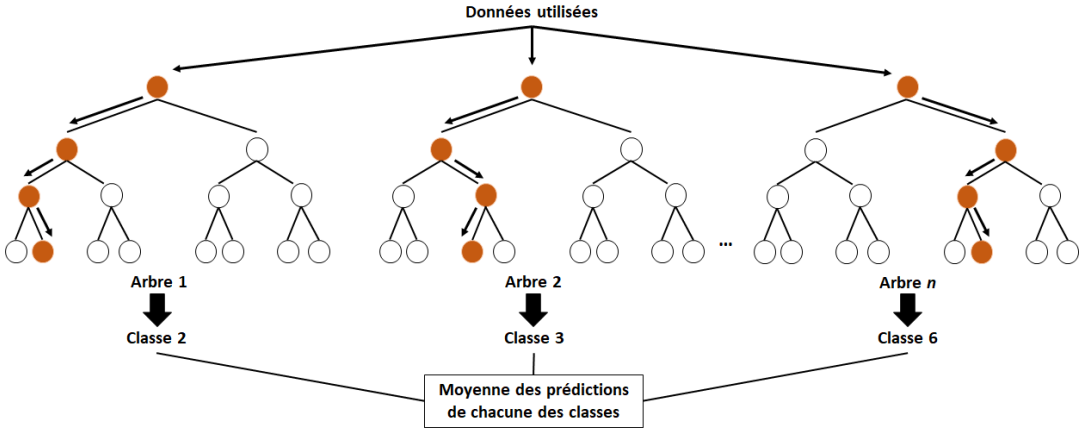


FIGURE 4.2 – Schéma simplifié d'un Random Forest

Une *grid search* va être utilisée sur l'ensemble de toutes les combinaisons de valeurs possibles des paramètres afin de trouver la valeur optimale de chacun des paramètres telle que le couple associé minimisera l'erreur de prédiction (basée sur l'indicateur de la RMSE).

4.1.3 Gradient Boosting Machine

Le *Gradient Boosting Machine* (*Friedman, 2001*, [**3**]) repose sur le principe de *Boosting*, basé sur une méthode de descente de gradient, qui consiste à utiliser plusieurs modèles entre eux dans le but d'obtenir un seul résultat prédictif. L'objectif est de construire une séquence de modèles de telle sorte qu'à chaque nouvelle étape, le nouveau modèle apparaisse comme une meilleure solution que le précédent. Pour avoir une amélioration de la prédiction à chaque étape, le *Boosting* affecte un poids plus important aux individus pour lesquels la valeur a été mal prédite à l'étape précédente. Le réajustement des poids à chaque étape permet une meilleure prédiction des valeurs difficiles.

Le GBM optimise ainsi les performances d'une série de modèles avec un pouvoir prédictif faible afin de créer un modèle robuste. Généralement, les modèles de prédiction faible (modèles à peine plus efficaces qu'un tirage aléatoire) utilisés sont des arbres de décision CART. Le but du *Gradient Tree Boosting* est de réaliser une succession d'arbres de décision où chaque arbre est construit sur l'erreur résiduelle du précédent.

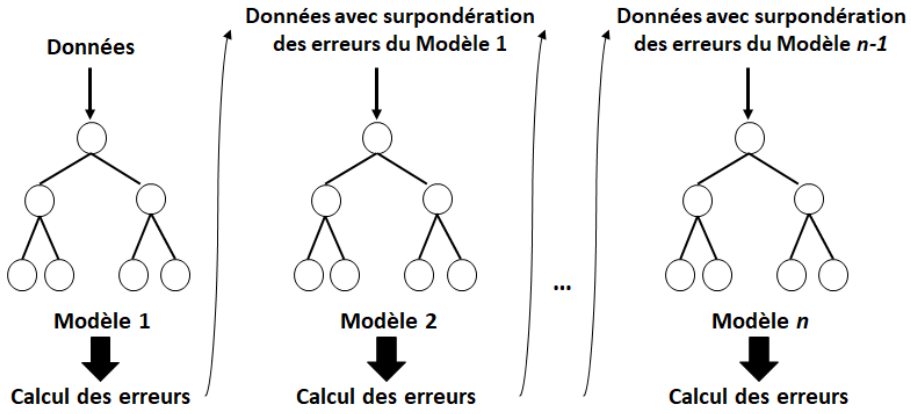


FIGURE 4.3 – Schéma simplifié d'un GBM

Afin d'avoir un modèle le plus performant possible, tout en évitant le sur-apprentissage, il existe plusieurs paramètres influençant l'apprentissage des données (nommés hyper-paramètres) à fixer pour le GBM :

- Nombre d'arbres générés : correspond au nombre d'itérations effectuées par l'algorithme. Le fait d'augmenter le nombre d'itérations conduit à une diminution de l'erreur et à une stabilisation de la prédiction. Cependant, un nombre d'arbres trop grand risque de conduire à un sur-apprentissage ;
- Profondeur de l'arbre : plus l'arbre possède de noeuds, meilleure sera la prise en compte d'interactions entre les différentes variables. Mais une profondeur importante est également source de sur-apprentissage ;
- Taux d'apprentissage (appelé aussi *shrinkage*) : permet de réguler la contribution de chaque arbre et de retarder la vitesse d'apprentissage de l'algorithme. Une faible valeur conduit à un apprentissage plus long et nécessitant plus d'arbres pour atteindre un niveau de performance optimal. A noter qu'un faible taux d'apprentissage permet d'obtenir de meilleurs performances de modélisation ;
- Pourcentage de variables considérées aléatoirement pour construire l'arbre : contrairement au *Random Forest* où un nombre de variables est défini, dans le GBM il s'agit d'une part des variables retenues ;
- Nombre minimal d'observations par feuille : notion identique à celle présentée pour le *Random Forest*.

Une grille de recherche va également être utilisée pour définir la combinaison optimale d'hyper-paramètres.

Une fois ce cadre théorique présenté, il est possible de passer à l'étape de modélisation effectuée ici à l'aide du logiciel R (modules *randomForest* et *xgboost*).

4.2 Modélisation de la fréquence

4.2.1 Mise en oeuvre de la modélisation

La grille de paramètres suivante a été testée pour le *Random Forest* avec une fonction de perte quadratique :

- Le nombre minimal d'observations par feuille est compris entre 1% et 5% du nombre total d'observations ;
- La base de modélisation comprenant 30 variables, il a été décidé de retenir arbitrairement au maximum dix d'entre elles afin de construire les arbres.

Pour cette méthode d'agrégation d'arbres, le risque de sur-apprentissage n'augmente pas en fonction du nombre d'arbres générés. Néanmoins, afin d'obtenir des résultats robustes avec un temps d'exécution correct, l'étude a été effectuée avec 400 arbres.

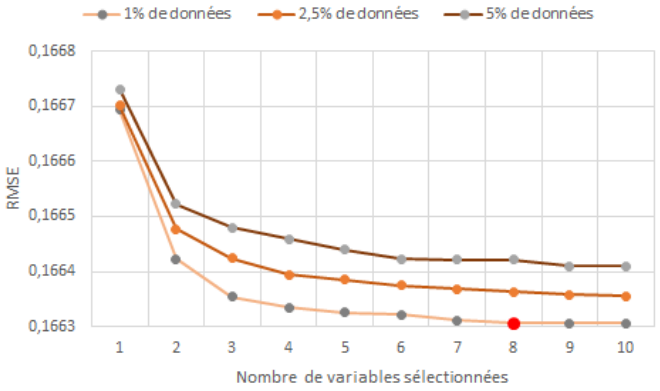


FIGURE 4.4 – Optimisation des paramètres du *Random Forest* - Fréquence

Le graphique ci-dessus permet de comparer l'évolution de l'indicateur RMSE en fonction du nombre de variables sélectionnées aléatoirement pour construire les arbres et du nombre d'observations minimal par feuille. Le couple de paramètres qui minimise le critère d'erreur RMSE (matérialisé par le point rouge sur le graphique) est obtenu pour 1% des observations par feuille et huit variables retenues. L'optimisation des paramètres étant finalisée pour le *Random Forest*, une démarche similaire va être effectuée pour le GBM (qui utilise également une fonction de perte quadratique) :

- Le nombre d'arbres généré est compris entre 50 et 400 par pas de 50 ;
- La profondeur de l'arbre est comprise entre 2 et 6 noeuds ;
- Le taux d'apprentissage est compris entre 1% et 10% ;
- La part de variables utilisées est comprise entre 10% et 100% par pas de 10% ;
- Le nombre minimal d'observations par feuille est paramétré de la même manière que pour le *Random Forest*, soit entre 1% et 5%.

La grille étant plus conséquente, il a été décidé d'effectuer l'optimisation de manière séquentielle. Un ou deux paramètres sont étudiés simultanément, puis une fois la valeur optimale définie, celle-ci est fixée et d'autres paramètres sont variés. Les deux premiers paramètres analysés sont le taux d'apprentissage et la profondeur de l'arbre. Leur grille de recherche est présentée sur le premier graphique groupé ci-dessous. La RMSE est minimale (valeur de 0,16630) pour un taux d'apprentissage de 0,1, soit 10%, et une profondeur d'arbre de six noeuds.

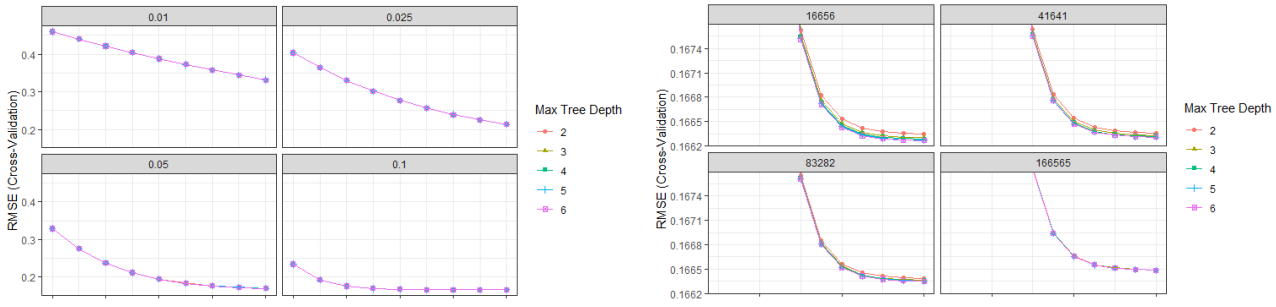


FIGURE 4.5 – Optimisation des paramètres du GBM - Fréquence - 1/2

Une fois ces deux paramètres fixés, le nombre minimal d'observations par feuille a été étudié. La valeur optimale d'un point de vue de la minimisation des erreurs est 16 656, ce qui correspond à 1% des observations utilisées. Ce résultat est similaire à celui observé sur le *Random Forest*, ce qui montre une stabilité malgré différentes méthodes utilisées. Avec ce couple de trois paramètres optimisés, la RMSE a diminué pour être égale à 0,16627. Il est donc possible maintenant d'optimiser le nombre d'arbres générés ainsi que la part de variables retenues pour la construction du modèle. Plus le nombre d'arbres est important, plus la RMSE est faible, et ce quel que soit la part de variables retenues. La valeur optimale est de conserver 40% des variables, soit 12, à chaque construction d'arbre et de répéter ce processus 400 fois (ce dernier paramètre étant également similaire à celui obtenu par la méthode du *Random Forest*). Cette double optimisation de paramètres a permis à nouveau de diminuer la RMSE par rapport à l'étape précédente (0,16622).

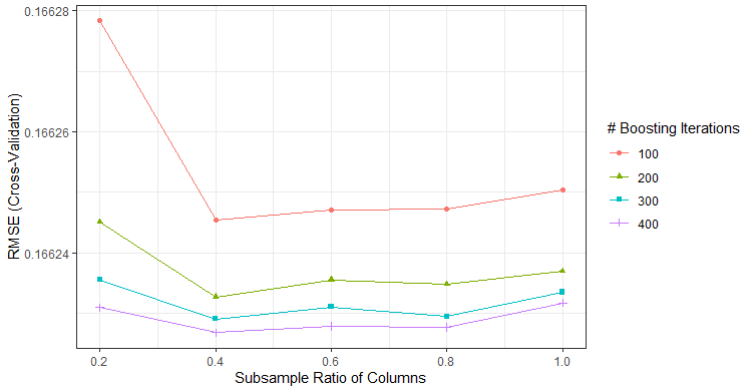


FIGURE 4.6 – Optimisation des paramètres du GBM - Fréquence - 2/2

La phase d'optimisation étant terminée, les résultats obtenus vont maintenant être présentés.

4.2.2 Comparaison des modèles obtenus

Pour comparer des modèles d'apprentissage statistique entre eux, il est possible de regarder certains indicateurs comme le Gini ou la RMSE, mais également l'importance relative des variables retenues. Cela consiste à distinguer les variables qui sont les plus utilisées pour construire les arbres et noeuds.

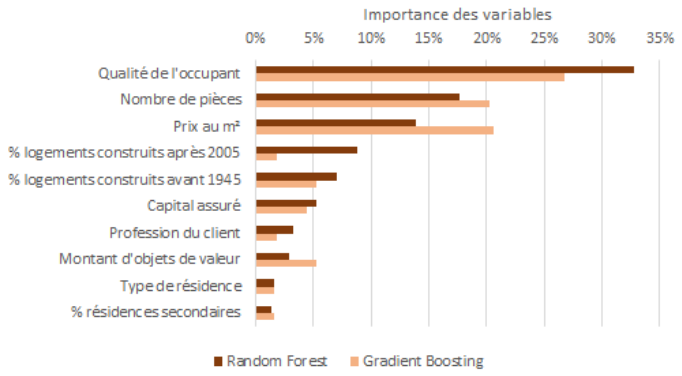


FIGURE 4.7 – Importance des variables - Fréquence

Les principales variables discriminantes sont relativement similaires entre le *Random Forest* et le *Gradient Boosting*. En effet, la qualité de l'occupant est utilisée dans plus d'un arbre sur quatre pour séparer l'espace de modélisation, et ce quelle que soit la méthode sélectionnée.

Il peut être intéressant de remarquer que si certaines variables ont une importance similaire entre les deux approches (nombre de pièces, capital assuré, profession ou type de résidence), d'autres sont plus volatiles. Par exemple, le montant d'objets de valeur est une variable bien plus significative dans l'approche GBM (quatrième variable avec une importance de 5,23%) que dans celle du *Random Forest* (huitième variable avec une importance de 2,85%). Enfin, il est à noter que la profession du client ressort comme variable importante alors qu'elle était absente des résultats pour les modèles linéaires pénalisés. Concernant la qualité de prédiction, le tableau suivant synthétise les résultats :

	Base d'apprentissage		Base de validation	
Modèle	RMSE	Gini	RMSE	Gini
Random Forest	0,16631	33,85%	0,16752	33,87%
GBM	**0,16622**	**34,88%**	**0,16643**	**34,82%**

TABLE 4.1 – Métriques d'évaluation - Fréquence

Les valeurs obtenues par les différents indicateurs sont stables entre les bases d'apprentissage et de validation et sont très légèrement supérieures à celles présentées dans la section des GLM. L'approche GBM est celle qui présente les meilleurs résultats : gain d'un point de Gini et RMSE plus faible ($1{,}09 \times 10^{-3}$). Ce modèle sera donc retenu pour modéliser la fréquence dans cette section de *Machine Learning*. Il est donc maintenant possible de passer à l'étape de la modélisation du coût moyen.

4.3 Modélisation du coût moyen

4.3.1 Mise en oeuvre de la modélisation

Du fait que seules les images de risque pour lesquelles une charge de sinistre non nulle et positive sont conservées dans cette section, le volume de données est moindre. Par conséquent, cela permet d'avoir une grille de recherche plus large tout en conservant un temps d'exécution acceptable. La grille de paramètres suivante a donc été testée pour le *Random Forest* avec un nombre d'arbres égal à 400 et une fonction de perte toujours quadratique :

- Le nombre minimal d'observations par feuille est compris entre 1% et 10% du nombre total d'observations ;
- La construction des arbres prendra en compte jusqu'à 15 variables, soit la moitié du nombre total.

A l'aide du graphique ci-dessous, le couple de paramètres qui minimise le critère d'erreur RMSE est obtenu pour 1% des observations par feuille et seulement trois variables.

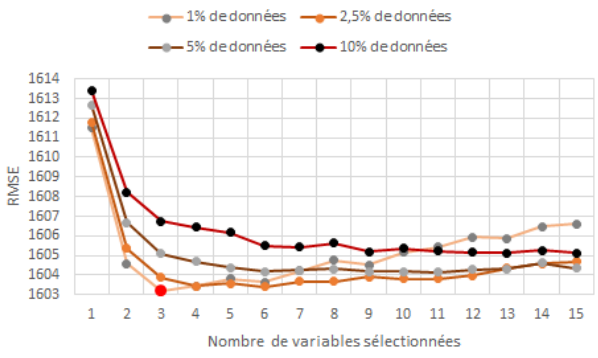


FIGURE 4.8 – Optimisation des paramètres du *Random Forest* - Coût moyen

Concernant le GBM, la grille de paramètres suivante a été testée, avec également une fonction de perte quadratique :

- Le nombre d'arbres générés est compris entre 50 et 400 par pas de 50 ;
- La profondeur de l'arbre est comprise entre 2 et 7 noeuds ;
- Le taux d'apprentissage est compris entre 1% et 15% ;
- La part de variables utilisées est comprise entre 10% et 100% par pas de 10% ;
- Le nombre minimal d'observations par feuille est paramétré de la même manière que pour le *Random Forest*, soit entre 1% et 10%.

En étudiant la figure 4.9, la combinaison qui minimise la RMSE est un taux d'apprentissage de 10%, une profondeur d'arbre de deux noeuds et un nombre minimal d'observations par feuille égal à 1% du nombre total.

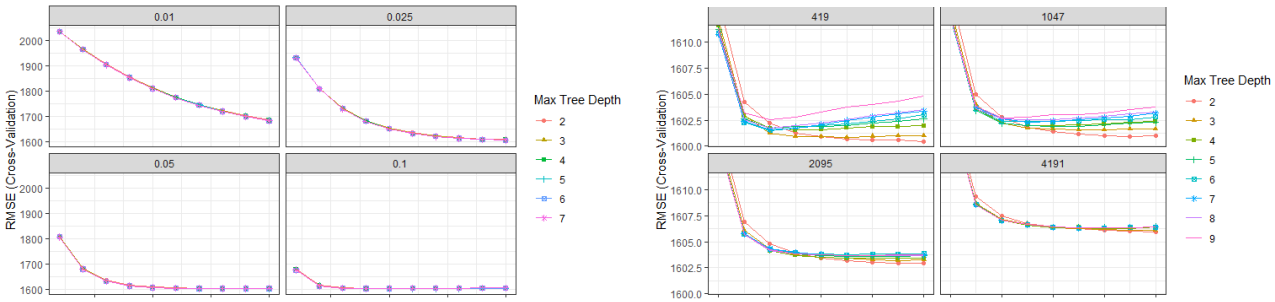


FIGURE 4.9 – Optimisation des paramètres du GBM - Coût moyen - 1/2

Il est donc possible maintenant d'optimiser le nombre d'arbres générés ainsi que la part de variables retenues pour la construction du modèle. La valeur optimale est de conserver 20% des variables à chaque construction d'arbres et de répéter ce processus 130 fois. Il est à noter que la valeur de ce dernier paramètre est bien plus faible que celle utilisée pour la modélisation de la fréquence, alors qu'il y a moins de données.

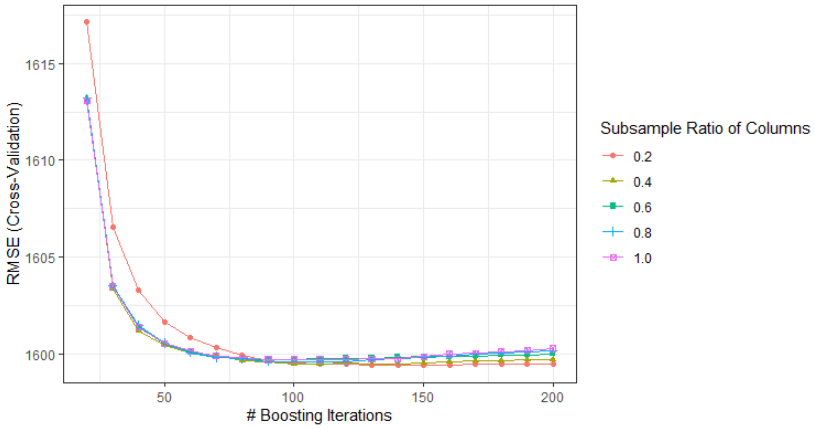


FIGURE 4.10 – Optimisation des paramètres du GBM - Coût moyen - 2/2

4.3.2 Comparaison des modèles obtenus

Contrairement au modèle de fréquence, les résultats obtenus en terme d'importance des variables sont très proches entre les deux méthodes. Seul le prix au m^2 est utilisé de manière différente pour segmenter l'espace des variables. Par ailleurs, il est à noter que trois variables (âge et profession du client ainsi que surface moyenne dans la zone du bien assuré) ressortent comme variables importantes alors qu'elles étaient absentes des résultats pour les modèles linéaires pénalisés.

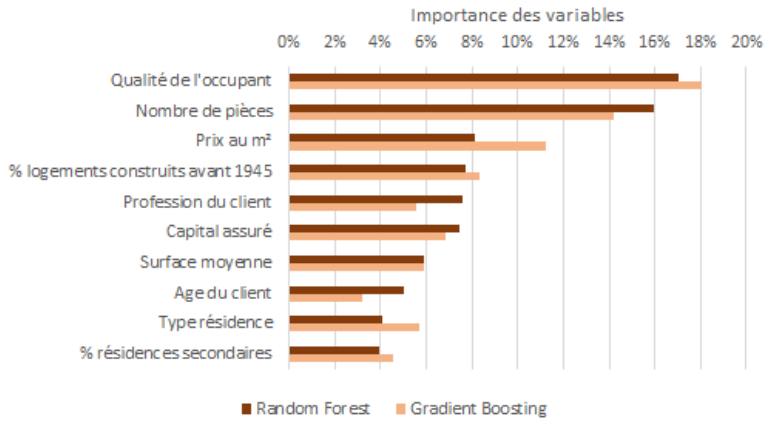


FIGURE 4.11 – Importance des variables - Coût moyen

Concernant la qualité de prédiction, le tableau ci-dessous synthétise les résultats. Tout d'abord, il est à noter que les modèles effectués selon les deux approches sont stables entre les bases d'apprentissage et de validation. Et, comme pour la fréquence, l'approche GBM possède les meilleurs indicateurs de sélection de modèle mais les écarts sont ici nettement plus faibles entre les deux approches.

	Base d'apprentissage		Base de validation	
Modèle	RMSE	Gini	RMSE	Gini
Random Forest	1 603,23	12,27%	1 675,47	12,12%
GBM	**1 598,13**	**12,34%**	**1 645,03**	**12,18%**

TABLE 4.2 – Métriques d'évaluation - Coût moyen

4.4 Synthèse de modélisation

A la vue des résultats présentés dans le tableau 4.3, l'approche GBM s'est révélée être plus performante que le *Random Forest*, et ce qu'il s'agisse de modéliser la composante fréquence ou coût moyen. Cela semble logique dans la mesure où le GBM est beaucoup plus paramétré et que chaque itération générée tient compte des erreurs effectuées précédemment. Il va être donc nécessaire de sélectionner le modèle qui servira de référence pour la suite de l'étude, entre les approches GLM pénalisés et GBM.

	Fréquence		Coût moyen	
Modèle	RMSE	Gini	RMSE	Gini
GLM pénalisé	0,2067	34,74%	1 750,25	11,93%
GBM	0,1664	34,82%	1 645,03	12,18%

TABLE 4.3 – Comparaison des métriques d'évaluation entre les GLM et GBM sur la base de validation

Le modèle GBM possède de meilleurs indicateurs d'ajustement et de segmentation que l'approche pénalisée des GLM (LASSO pour la fréquence et Elastic Net pour le coût moyen). Cependant, les écarts observés n'ont pas été jugés suffisamment importants, en particulier en terme de gain de Gini, pour que cette méthode plus complexe soit retenue. En effet, la prédiction du GBM revient à trouver le noeud final de l'arbre dans lequel est affecté le contrat. Or, pour une faible variation de la fréquence ou de la charge, la prédiction peut varier fortement si celle-ci influe sur un noeud peu profond dans l'arbre.

De plus, l'interprétation de chaque variable n'est pas évidente, au contraire des modèles GLM. Cette complexité rend l'explication du contenu des différents modèles bien plus difficile auprès d'interlocuteurs divers, ce qui est pourtant un des principaux rôles de l'actuaire. Par exemple, il a été montré qu'avec l'approche GLM, le risque augmentait en fonction du nombre de pièces, à caractéristiques de risque équivalentes. Le tarif sera donc croissant avec cette variable, et ce tout le temps. Or, cela pourrait ne pas être le cas en utilisant un GBM, ce qui serait contre-intuitif d'un point de vue commercial, tant pour le client que pour le distributeur. **L'approche utilisant les GLM pénalisés sera donc retenue dans la suite de l'étude**.

Une fois les modèles finaux sélectionnés, la prime pure de chaque contrat peut être obtenue en multipliant entre eux les coefficients issus des modèles de fréquence et de coût moyen. La valeur de Gini associée à ce modèle agrégé s'élève à 43,64%. Au global, le graphique ci-dessous permet de visualiser que la plupart des images de risque ont une prime pure estimée comprise entre 16,10 € et 34,40 €.

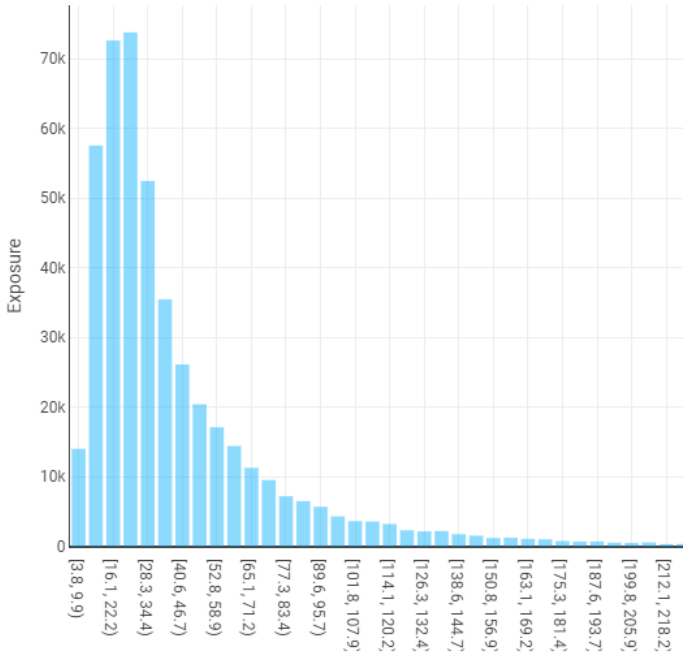


FIGURE 4.12 – Répartition des montants de prime pure prédits

Quant à l'écart entre les primes pures estimées et observées, il s'élève à -3,07%. Cela signifie que le modèle sous-estime le montant de prime pure de 3%. L'analyse des écarts selon les principaux segments commerciaux va permettre d'affiner le jugement sur la qualité de ce premier modèle global.

Segment commercial	Poids des segments	Écart observé
Locataire	74,9%	-4,01%
Propriétaire	25,1%	-2,08%
Locataire 1 pièce	16,8%	+0,54%
Locataire 2 pièces	21,9%	-0,32%
Locataire 3 pièces	22,9%	-6,21%
Locataire 4 pièces et plus	13,3%	-6,29%
Propriétaire 1 pièce	2,1%	-17,87%
Propriétaire 2 pièces	5,7%	-9,14%
Propriétaire 3 pièces	8,2%	-1,36%
Propriétaire 4 pièces et plus	9,1%	+1,98%
Global		-3,07%

TABLE 4.4 – Écart de prime pure par segment commercial

Que ce soit pour la notion de locataire ou de propriétaire, la prime pure modélisée est sous-estimée par rapport à la prime pure observée (de respectivement -4% et -2%). Mais en affinant les segments observés, il est possible de remarquer que plus le nombre de pièces augmente, moins la précision est correcte sur le périmètre des locataires. Par exemple, les biens de moins de trois pièces (qui représentent plus de la moitié des contrats de ce segment) ont un écart quasi-nul. Mais les résultats se dégradent ensuite. A l'inverse des locataires, le modèle gagne en précision pour les propriétaires lorsque le nombre de pièces augmente. Les écarts sont très importants pour les appartements d'une pièce (près de 18% d'écart) puis diminuent fortement par la suite pour devenir faibles sur les biens de trois pièces et plus (2/3 des contrats sur le périmètre des propriétaires).

Ce premier modèle peut sembler disposer de quelques lacunes à la vue de ces résultats, mais il est à rappeler que jusqu'à présent, seules les données récoltées principalement à la souscription ont été utilisées. Il peut être intéressant maintenant d'ajouter une variable supplémentaire relative à l'environnement géographique de l'habitation assurée.

— Chapitre 5 —

Modélisation du signal géographique

5.1 Cadre théorique

5.1.1 Méthodologie de construction

Dans cette section, l'hypothèse suivante va être exposée : le nombre de sinistres ainsi que leur charge ne s'explique pas uniquement à partir des composantes relatives au bien assuré et au client. Une composante géographique doit être incluse. En effet, les données géographiques sont explicatives du risque du fait que le bien assuré possède un emplacement fixe. L'objectif est de déterminer cette nouvelle information sous forme d'un zonier (correspondance entre une zone géographique et un coefficient représentatif du risque) puis d'étudier son incorporation dans la structure du risque, ceci afin d'ajuster au mieux le tarif.

La nouvelle méthodologie de construction, proposée dans ce mémoire, est découpée en plusieurs étapes :

- Analyse de l'écart entre la variable modélisée (fréquence ou coût moyen) et la variable réelle observée (approche résiduelle). Ce résidu est supposé être constitué d'un signal géographique et de bruit :

$$\text{Variable prédite} = \text{Variable observée} + \text{bruit},$$
$$\text{où Variable observée} = \text{Variable observée non géographique} + \text{Variable observée géographique}.$$

- Agrégation des résidus à un niveau géographique donné ;
- Constitution de zones par un découpage des résidus ;
- Homogénéisation des zones par un lissage géographique ;
- Introduction de la variable zonier dans le modèle GLM pénalisé initial afin de devenir une variable géographique tarifaire. Les coefficients du modèle de départ auront été fixés, de telle sorte qu'ils n'évoluent pas en ajoutant une variable supplémentaire. Le but de cette manipulation est de fixer la part d'information expliquée par les variables internes pour ne pas qu'elle soit biaisée par l'ajout d'une autre variable dans le modèle ;
- Affectation d'un coefficient à chaque zone ajustable permettant d'adapter le tarif au niveau de risque.

5.1.2 Différentes mailles géographiques

La précision d'un zonier dépend de la granularité de la maille géographique sélectionnée. Il en existe trois principales en assurance Habitation :

- Maille INSEE : le territoire est découpé par rapport au code INSEE (code à cinq chiffres résultant de la concaténation du code département à deux chiffres ou lettres et du code commune à trois chiffres), correspondant à un niveau de précision à la commune ou à l'arrondissement pour les villes en disposant. Il en existe 34 881 en France métropolitaine ;
- Maille IRIS (**I**lots **R**egroupés pour l'**I**nformation **S**tatistique) : le territoire est découpé en quartiers regroupant environ 2 000 habitants. Parmi les 34 881 codes INSEE recensés, 33 040 ont un découpage IRIS identique au découpage INSEE. En revanche, les 1 841 codes INSEE restants sont découpés en plusieurs zones, au nombre de 15 566. Au total, il en existe 48 606 ;
- Maille Voronoï (appelée également micro-zonier) : le territoire est découpé en polygones dits de Voronoï qui correspondent à une zone géographique autour de chaque adresse observée en portefeuille. Tous les points contenus dans le polygone de Voronoï sont alors plus proches de cette adresse que de n'importe quelle autre. Ainsi, pour les zones denses, les cellules seront plus petites et plus nombreuses, tandis que dans les zones où il y a moins de contrats observés, leur superficie sera plus importante. La segmentation géographique est donc plus marquée. Au total, il en existe plus de 600 000 (certains polygones ayant été fusionnés entre eux pour des raisons pratiques).

  Le schéma ci-dessous permet d'illustrer la construction d'un polygone. Le point jaune symbolise une adresse de risque et est entouré d'un polygone, affiché en rouge. Chaque adresse dans ce polygone est plus proche du point jaune que de tout autre point d'échantillonnage (matérialisé par les points bleus). Après la création des polygones, les voisins d'un point sont définis comme tout autre point d'échantillonnage dont le polygone partage une bordure avec le point initial sélectionné.

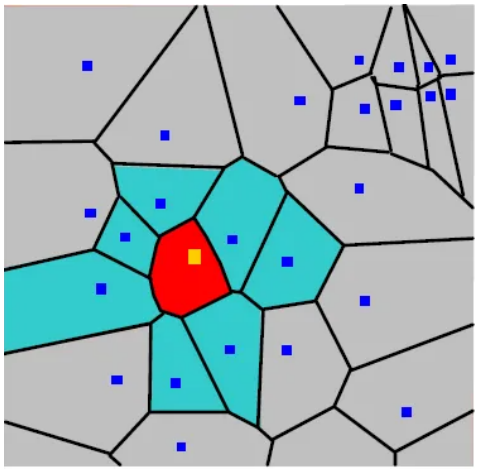


FIGURE 5.1 – Polygones de Voronoï
[10]

10. Source : sigterritoires.fr/index.php/analyse-exploratoire-des-donnees-pour-la-geostatistiqueles-diagrammes-de-voronoi/

Les modèles de prime pure actuellement implémentés en production utilisent pour les adresses les mieux géocodées (adresse exacte ou numéro de rue approché, soit 65% des contrats) un maillage de type Voronoï. Le niveau de précision des zoniers réalisés avec cette démarche est élevé mais la question de la robustesse se pose. En effet, bien qu'une forme de lissage ait été introduite dans le zonier, la prédiction du résidu pour un polygone donné est basée seulement sur un contrat (pas nécessairement sinistré). Ce type de zonier nécessite également d'avoir une méthodologie alternative en cas d'adresse inconnue. Près d'un tiers des adresses sont rejetées de la mécanique Voronoï pour être tarifées à la maille INSEE. Cette utilisation parallèle de zoniers à plusieurs mailles peut également poser des problèmes vis à vis des clients. Par exemple, au sein d'une même rue, deux adresses peuvent être tarifées en Voronoï avec des tarifs significativement différents. Et pour une troisième adresse inconnue, mais dans la même rue, un autre référentiel sera utilisé et donnera un tarif pouvant être sans lien avec les deux précédents.

L'objectif de la création de ces nouveaux zoniers, qui s'inscrit dans la refonte globale de la garantie Dégât des eaux, est de tenter d'obtenir une segmentation du risque similaire à celle obtenue par la méthode du Voronoï en production, tout en se restreignant à une maille moins fine. Seules les approches INSEE et IRIS seront retenues par la suite.

5.2 Définition des résidus et création des zones de risque

Pour chaque image de risque présente dans la base de modélisation (qui a été séparée en une base d'apprentissage et une base de validation pour éviter tout risque de sur-apprentissage), les informations sur les fréquences prédites et observées (ainsi que pour les charges moyennes) sont disponibles. Par la suite, pour alléger la lecture, seule la fréquence de sinistres sera mentionnée dans cette section, mais la méthodologie est identique pour la création du zonier sur le coût moyen. La première étape dans la construction du zonier va consister à agréger les données en fonction de la maille choisie (INSEE ou IRIS). Puis ensuite le résidu va être calculé comme étant la différence entre la fréquence moyenne prédite et la fréquence moyenne observée. De ce fait, chaque centroïde de découpage INSEE ou IRIS est associé à une valeur de résidus correspondant à l'agrégation des résidus dont les adresses appartiennent à la zone considérée. Le centroïde de cette zone fait référence à un point fictif situé à l'intérieur du polygone et dont les coordonnées correspondent au centre de celui-ci. Il est à noter que le résidu peut être inconnu si aucun contrat n'est présent dans la maille considérée.

Il faut maintenant découper ce résidu en classes. En effet, il existe plusieurs dizaines de milliers de communes et il ne peut être envisagé d'inclure un tel facteur comme variable ordinaire d'un modèle linéaire généralisé. Il convient alors d'effectuer des regroupements à l'aide de techniques de segmentation. Chaque classe représentera alors une région du zonier. Deux méthodes de découpage vont être présentées : CAH (**C**lassification **A**scendante **H**iérarchique) et *K-means* (*Jollois, 2010*, [**7**]).

5.2.1 Classification Ascendante Hiérarchique

Il s'agit d'une méthode de classification itérative dont le principe est le suivant pour n observations :

- Partition initiale en n classes, chaque observation représentant une classe ;
- Calcul de la dissimilarité entre les observations ;
- Regroupement d'observations permettant de minimiser un critère d'agrégation donné, créant ainsi une classe ;
- Calcul de la dissimilarité entre cette classe et les observations restantes en utilisant le critère d'agrégation. Puis sont regroupées les observations ou les classes minimisant ce critère. Le processus continue jusqu'à ce que toutes les classes d'observations soient regroupées.

L'objectif de ces regroupements est de maximiser l'inertie inter-classes (deux classes considérées doivent être très différentes) et de minimiser l'inertie intra-classe (les observations au sein d'une même classe doivent être très proches). Pour une partition des données regroupant les n observations en k classes $[A_1, ..., A_k]$, ces notions d'inertie sont définies de la manière suivante :

$$\begin{aligned} \text{Inertie inter-classes} : I_B &= \sum_{j=1}^{k} m_j d^2(G_{Aj}, G_E), \\ \text{Inertie intra-classe} : I_W &= \sum_{j=1}^{k} \sum_{i \in A_j} d^2(n_i, G_{Ai}), \\ \text{Inertie totale} : I_T &= I_B + I_W, \end{aligned} \tag{5.1}$$

avec :

- m_j le poids de la classe A_j ;
- G_{Aj} le barycentre de la classe A_j ;
- $d^2(,)$ la distance calculée entre deux sous-ensembles ;
- n_i le $i^{ème}$ individu de la classe A_j.

La métrique utilisée pour mesurer l'écart entre les observations sera la distance euclidienne. Concernant la mesure entre les classes, il existe plusieurs stratégies d'agrégation de classes (saut minimum, moyen, complet ou maximum, méthode des centroïdes par exemple) mais la méthode de Ward est celle qui est la plus utilisée. Cette agrégation sélectionne le regroupement de l'étape suivante qui provoque la plus petite augmentation d'inertie intra-classe. La notion de distance utilisée ici est la distance entre les barycentres de deux classes élevée au carré et pondérée par l'effectif de ces classes :

$$D_{Ward}(A_i, A_j) = \frac{m_i m_j}{m_i + m_j} d^2(G_{Ai}, G_{Aj}). \tag{5.2}$$

Les deux classes (A_i, A_j) qui se regroupent sont celles qui minimisent le critère de Ward. Cette agrégation est une optimisation à chaque étape et donc ne garantit pas une optimalité globale. Mais, dans le cadre de cette étude, il sera fait l'hypothèse que la partition trouvée à partir du regroupement optimal local à chaque étape est proche de la partition optimale globale. En pratique, les étapes de regroupement à chaque partition sont représentées graphiquement par un arbre hiérarchique appelé dendrogramme.

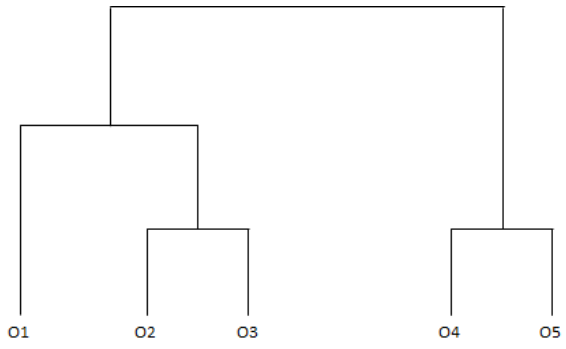


FIGURE 5.2 – Dendrogramme

Cet arbre synthétise les différents rapprochements de classes. Les branches de l'arbre correspondent à la distance de Ward. Il est notamment utile pour déterminer le nombre final de classes.

Dans l'illustration précédente, il est possible de repérer facilement un découpage en deux ou trois classes distinctes. Ce choix est un compromis entre un nombre de classes souhaité et ce qu'indique la classification. En effet, si la distance de Ward est importante pour le regroupement de cette étape, ceci indique que les classes ne possèdent pas vraiment de ressemblance, et donc, que la partition n'est pas adéquate. De plus, sélectionner un nombre de classes trop important retire l'intérêt de faire une classification, dont le but initial est de réduire le nombre de modalités.

Il existe également des indicateurs permettant de définir de manière plus rigoureuse le nombre de classes (*Jollois, 2010*, [**7**]) :

- Le R^2 est le rapport de l'inertie inter-classes sur l'inertie totale. C'est le pourcentage de l'inertie obtenue par les classes qui peut être interprété comme une mesure de l'homogénéité des classes. La classification est jugée pertinente si ce rapport est le plus proche possible de 1, tout en ayant un nombre de classes limité. A noter qu'un R^2 nul signifie qu'il n'y a qu'une seule classe, donc aucune segmentation. Et à l'inverse, un R^2 de 1 est le résultat d'un découpage où chaque classe comprend une unique observation ;
- Le R^2 semi-partiel représente la perte d'inertie inter-classes en fusionnant deux classes, ce qui est équivalent à la baisse de l'indicateur R^2 lors du passage d'un segmentation en k classes à $k-1$ classes ;
- Le pseudo-F mesure la séparation entre les k classes et compare l'homogénéité d'une partition de k classes à une partition en $k-1$ classes pour n observations. L'objectif est de maximiser cet indicateur qui est défini de la manière suivante :

$$\text{pseudo-F} = \frac{R^2}{k-1} \times \frac{n-k}{1-R^2}. \tag{5.3}$$

5.2.2 K-means

La méthode des *K-means* est une technique statistique de partitionnement permettant de diviser des données en groupes homogènes. Pour un paramètre k donné et défini *a priori*, k classes vont être créées sous contrainte de minimiser la distance entre les points dans chacune d'entre elles. L'affectation des observations va être corrigée itérativement afin que les classes deviennent plus homogènes et qu'elles contiennent des résidus de plus en plus proches. Les différentes étapes sont :

- Définition aléatoire de k points initiaux qui correspondent aux barycentres des classes ;
- A chaque affectation d'une observation à sa classe la plus proche, les barycentres sont recalculés ;
- L'étape précédente est répétée jusqu'à ce que l'inertie intra-classe ne diminue plus de manière significative ou que le nombre d'itérations soit atteint.

Cependant, il est à noter que les classes obtenues par cette méthode dépendent du choix aléatoire initial des barycentres. Deux partitions aléatoires sur les mêmes données peuvent amener à deux classifications finales différentes. Afin de réduire ce risque sur les résultats, un grand nombre de tirages aléatoires (100 dans cette étude) a été réalisé pour s'assurer de la robustesse de la classification. Par ailleurs, ne disposant pas d'une valeur précise du nombre de classes souhaité, celui-ci a été varié entre 2 et 20. L'analyse d'indicateurs de partitionnement va permettre de définir le nombre k optimal, ce qui revient à se ramener à un cadre décisionnel similaire à celui présenté pour la CAH.

Les indicateurs retenus dans l'approche des *K-means* seront le R^2, le pseudo-F (utilisés également pour la CAH) ainsi que l'indice de Calinski-Harabasz. Ce dernier doit être à maximiser et est défini comme suit pour n observations découpées en k classes :

$$\text{CH} = \frac{(n-k)I_B}{(k-1)I_W}, \tag{5.4}$$

avec I_B et I_W les inerties respectivement inter-classes et intra-classe.

A noter que d'autres indicateurs d'aide à la classification tels que le CCC (***C**ubic **C**lustering **C**riterion*), le pseudo T^2 ou l'indice de Dunn existent, mais ils ne seront pas présentés et utilisés dans ce mémoire.

Une fois les notions de découpage des données présentées, il est maintenant possible de passer à l'analyse des résultats obtenus sur les résidus du modèle de fréquence à la maille INSEE.

5.2.3 Application sur le modèle de fréquence - maille INSEE

Les résultats obtenus à partir de la méthode de la CAH, après utilisation du logiciel R, vont tout d'abord être présentés. Les différents graphiques ci-dessous permettent de repérer le nombre de classes optimal en fonction de l'indicateur considéré.

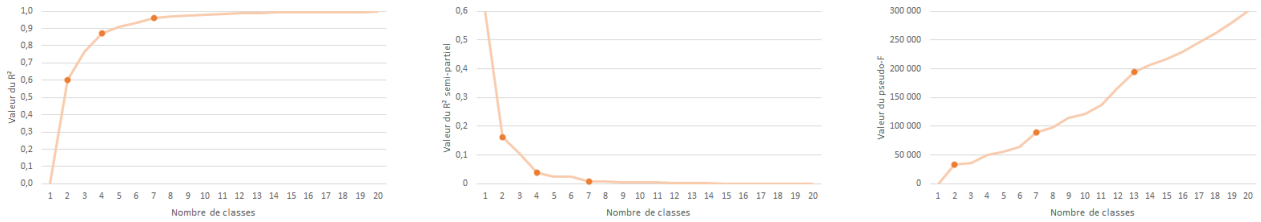


FIGURE 5.3 – Nombre de classes selon l'indicateur du R^2, du R^2 semi-partiel et du pseudo-F

Concernant l'indicateur du R^2, trois cassures (appelées coudes) sont distinctes pour une segmentation en 2, 4 et 7 classes (matérialisées par les points sur la courbe). Les résultats obtenus sont identiques pour l'indicateur du R^2 semi-partiel. Enfin, pour le pseudo-F, bien qu'il faille en théorie le maximiser, il faut, en pratique, repérer les hausses nettes générant des pics, visibles pour 2, 7 et 12 classes.

Au global, en synthétisant ces premiers résultats, il existe quatre possibilités de découpages des résidus (2, 4, 7 et 12 classes) qui correspondront aux zones de risque. Mais, sachant que l'objectif d'un zonier est de discriminer au mieux les risques, un découpage avec un très faible nombre de classes ne sera pas pertinent et l'effet géographique sera uniformisé. Par conséquent, il a été décidé de considérer que la meilleure segmentation obtenue à partir de la méthode de la CAH correspond à la valeur commune la plus élevée, soit sept zones de risque.

Il peut être également pertinent d'analyser visuellement le dendrogramme obtenu afin de vérifier ces premiers résultats.

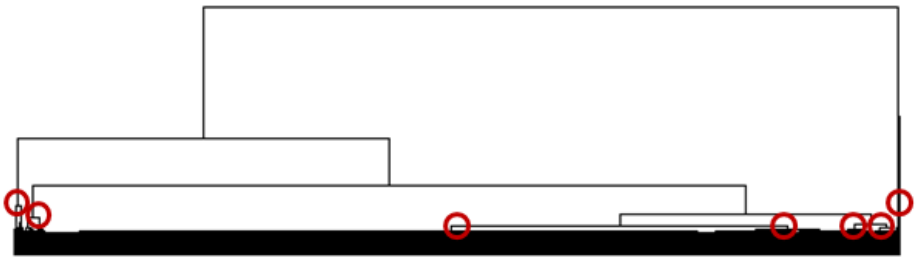


FIGURE 5.4 – Dendrogramme de la CAH sur les résidus du modèle de fréquence - Maille INSEE

Le graphique est moins lisible que celui utilisé pour la présentation de la CAH du fait d'un nombre d'observations bien plus important. Cependant, il est possible de repérer sept groupes de données dont les noeuds de séparation sont matérialisés en rouge.

Ces premiers résultats vont maintenant être comparés à ceux obtenus à partir de la méthode des *K-means*, après utilisation également du logiciel R.

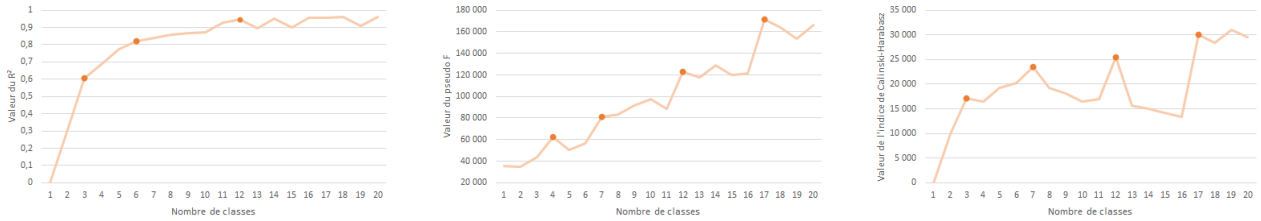


FIGURE 5.5 – Nombre de classes selon l'indicateur du R^2, du pseudo-F et de Calinski-Harabasz

Concernant le R^2, un certain nombre de cassures sont visibles et les trois premières correspondent à un découpage en 3, 6 et 12 classes. Les résultats obtenus sont similaires pour l'indicateur du pseudo-F avec des pics atteints pour 4, 7, 12 voire 17 classes (cette dernière étant associée à la valeur maximale de l'indicateur). Enfin, avec l'étude de l'indice de Calinski-Harabasz, quatre découpages peuvent également être retenus : 3, 7, 12 et 17 classes. Au global, quel que soit l'indicateur observé, le nombre de classes proposé est stable. Il est à noter que les découpages optimaux proposés par cette méthode des *K-means* conduisent à un nombre de classes plus élevé qu'avec la CAH. Au final, il a été décidé de retenir le découpage des *K-means* avec le plus grand nombre de classes commun aux indicateurs observés, soit 12 classes (la segmentation équivalente proposée par la CAH générait certains groupes avec une faible exposition). La classification obtenue est la suivante :

Classe	1	2	3	4	5	6	7	8	9	10	11	12
INSEE	7 763	14	775	657	52	221	32	80	5	88	12 265	191
Contrats	3,3%	9%	2,3%	17%	2,7%	17%	10,1%	2,2%	6,5%	15,6%	12,2%	2,1%

TABLE 5.1 – Caractéristiques des classes obtenues

Il faut tout d'abord préciser que les découpages présentés ne sont pas encore ordonnés : il n'est pas possible d'établir à partir de ce tableau que telle classe est plus risquée que telle autre. La segmentation obtenue ne génère pas de classes avec une exposition très faible (il y a au minimum 2% d'exposition). Une fois les zones triées par résidu moyen, il est possible de représenter la distribution sur une carte.

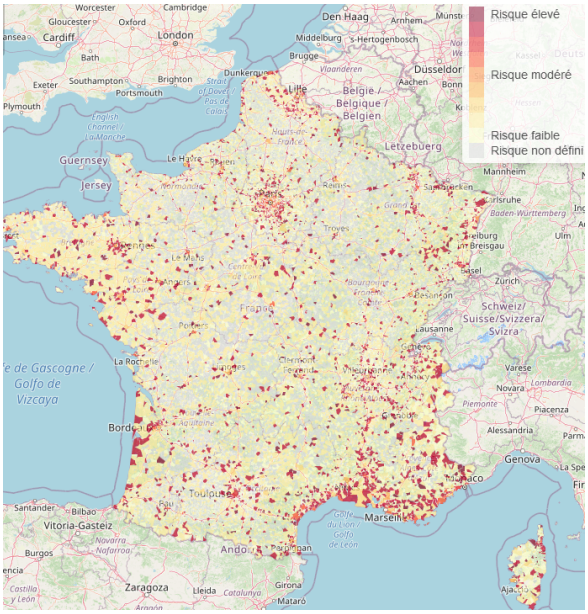


FIGURE 5.6 – Segmentation géographique à la maille INSEE - Fréquence

La lecture de cette carte permet tout d'abord de repérer les principales agglomérations qui ressortent en rouge, cela signifiant qu'à caractéristiques assurées équivalentes la zone est plus risquée. Il faut également préciser que les zones neutres qui ressortent de la carte correspondent à des communes pour lesquelles il n'y a aucune exposition de risque.

5.2.4 Résultats obtenus pour les autres modèles

Pour les autres modèles effectués (Fréquence à la maille IRIS et Coût moyen aux mailles INSEE et IRIS), seuls les résultats finaux seront présentés afin d'éviter d'alourdir la lecture. Cependant, les résultats détaillés sont présents en Annexe II. Le tableau ci-dessous synthétise le nombre de classes à retenir selon la méthode de classification et l'indice choisis.

	CAH		K-means		
Modèle	R^2	Pseudo-F	R^2	Pseudo-F	Indice de Calinski
Fréquence - IRIS	9	9	4 ou 8	7, 12 ou 16	5 ou 10
Coût moyen - INSEE	5 ou 7	7 ou 13	6 ou 12	5, 11 ou 15	10 ou 13
Coût moyen - IRIS	5 ou 7	7 ou 13	6 ou 12	6 ou 15	6, 10 ou 13

TABLE 5.2 – Nombre de classes proposé selon la méthode de classification et l'indicateur

Premier constat, les résultats sont quasi-identiques entre les mailles INSEE et IRIS pour le coût moyen. Quelle que soit la granularité choisie, les résidus seront découpés en 13 classes en utilisant la méthode des *K-means*. Concernant le modèle Fréquence-IRIS, les propositions de découpage diffèrent sensiblement de celles obtenues à la maille INSEE (où pour rappel la valeur finale retenue est de 12). Par ailleurs, aucune valeur ne ressort majoritairement, chaque indicateur ayant des recommandations différentes. Il a donc été décidé de regarder la répartition des observations au sein des classes selon les différentes valeurs de découpage. Au final, le découpage proposé par la CAH en neuf classes a été retenu car c'était le seul qui permettait d'obtenir une segmentation des observations sans avoir de classes sous-représentées.

L'étape de création des zones de risque étant finalisée pour les différentes approches retenues, l'homogénéisation géographique via un lissage peut maintenant être expliquée.

5.3 Lissage des résidus

En représentant la répartition des résidus découpés pour le modèle de coût moyen par exemple, il est possible de remarquer qu'ils ne sont pas distribués de manière aléatoire, des groupes se dessinent. Il y a donc bien un effet géographique non expliqué par le modèle initial retenu. Mais deux problèmes se posent. Tout d'abord, à l'intérieur de ces zones, les valeurs des résidus sont parfois très hétérogènes entre deux découpages voisins, ce qui risque d'engendrer des sauts tarifaires plus ou moins marqués. Par ailleurs, toutes les zones géographiques ne sont pas représentées dans la base de modélisation. En effet, il existe des communes ou des quartiers IRIS pour lesquels aucun sinistre Dégât des eaux sur un appartement n'est survenu. La carte ci-dessous permet d'illustrer ce dernier point : même en région parisienne, un certain nombre de zones ne dispose pas de résidus car aucun sinistre n'a été recensé sur la période d'observation.

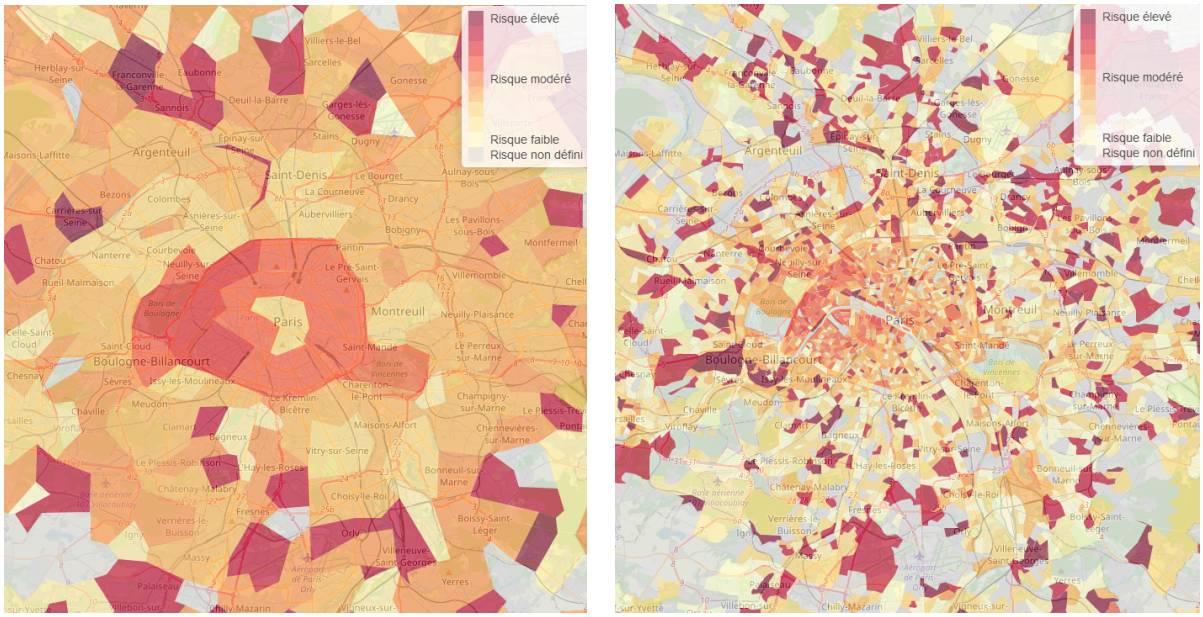


FIGURE 5.7 – Carte des résidus segmentés - Modèle de coût moyen aux mailles INSEE et IRIS - Région parisienne

Sans effectuer aucun lissage géographique, les variations tarifaires pourraient être importantes et la problématique actuelle concernant la méthodologie du Voronoï existerait toujours. De même, les zones n'ayant pas d'observations, et donc pas de résidus, ne pourraient pas être associées à un tarif. Cependant, si le lissage devient important, alors le risque géographique sera uniformisé et l'information portée serait supprimée.

Le lissage, qui va être effectué dans cette étude à l'aide du logiciel R, est basé sur la théorie de la crédibilité. Cela consiste à lisser le risque d'un point à partir des caractéristiques de risque des zones environnantes et à le pondérer en fonction de l'exposition et de la distance de ces mêmes zones. La fonction de distance assigne une influence plus grande aux communes plus proches (effet décroissant avec l'éloignement des communes voisines). Cependant, avant d'entrer plus en détails sur le processus de lissage, les principes liés à la crédibilité vont être présentés.

5.3.1 Théorie de la crédibilité

La théorie de la crédibilité est une technique mathématique qui a pour objectif de proposer une tarification adaptée aux groupes à partir des données historiques (*Gorrand, 2020*, [**4**]). L'objectif est de conserver un équilibre entre mutualisation et segmentation du risque. Parmi les différents modèles de crédibilité, le plus utilisé est le modèle de Bühlmann-Straub. Ce modèle est une généralisation du modèle de Bühlmann tenant compte de l'exposition au risque des assurés au cours du temps et se base sur les hypothèses suivantes :

- L'espérance et la variance de la variable $X_{i,t}$, représentant le montant de sinistres (dans la définition générale) pour l'assuré i l'année t, doivent exister ;
- $\mathbb{E}(X_{i,t}|\theta_i) = m(\theta_i)$, soit une hypothèse de stabilité dans le temps avec θ_i le risque intrinsèque de l'individu i non observable ;
- $Cov(X_{i,t}; X_{i,t'}|\theta_i) = 0$ pour t $\neq t'$;
- $\mathbb{V}(X_{i,t}|\theta_i) = \frac{\sigma^2(\theta_i)}{p_{i,t}}$, avec $p_{i,t}$ le poids du risque i à la date t ;
- $\forall i, \forall t, (\theta_i; X_{i,t}) \perp (\theta_j; X_{j,t})$, soit une hypothèse d'indépendance entre les différents risques.

Le facteur de crédibilité est, quant à lui, défini de la manière suivante :

$$c_i = \frac{\sum_{t=1}^{n} p_{i,t}}{\sum_{t=1}^{n} p_{i,t} + k}, \tag{5.5}$$

avec k nommé le coefficient d'information qui correspond au rapport entre les variances intra-groupe (incertitude qui reste malgré l'historique) et inter-groupes (incertitude qui disparaît avec l'historique) et qui est égal à :

$$k = \frac{\mathbb{E}[\mathbb{V}(X|\theta)]}{\mathbb{V}[\mathbb{E}(X|\theta)]}. \tag{5.6}$$

Plus cette valeur est faible, plus l'information captée est pertinente et, par conséquent, plus une tarification *a posteriori* basée sur l'expérience est justifiée. L'expression de la prime tenant compte de la crédibilité peut s'écrire de la manière suivante :

$$\begin{aligned} \Pi_i &= (1 - c_i) \times \mu + \frac{c_i}{\sum_{t=1}^{n} p_{i,t}} \sum_{t=1}^{n} p_{i,t} \times X_{i,t}, \\ &= (1 - c_i) \times \text{prime } a\ priori + c_i \times \text{prime d'expérience.} \end{aligned} \tag{5.7}$$

Une fois la théorie de la crédibilité présentée dans le cadre de la tarification, celle-ci va être mise en place dans une problématique de lissage géographique.

5.3.2 Adaptation de la théorie de la crédibilité au lissage spatial

L'objectif dans cette section est de définir la valeur finale du coefficient associée à une zone géographique en fonction de la valeur initiale de celle-ci et de celles des zones environnantes. Le lissage doit donc s'effectuer sur deux niveaux. Tout d'abord, deux zones voisines doivent avoir des coefficients proches. De plus, une zone doit impacter sa voisine seulement si cette dernière a des caractéristiques proches.

La classe finale va donc s'écrire sous la forme suivante :

$$\text{classe}_{finale} = c_i \times \text{classe}_{initiale} + (1 - c_i) \times \frac{\sum_{v=1}^{n} f(v) \times \text{classe}_v}{\sum_{v=1}^{n} f(v)}. \tag{5.8}$$

avec :

- c_i le coefficient de crédibilité de la zone i ;
- $\text{classe}_{initiale}$ la classe de risque attribuée initialement à la zone en question après la segmentation ;
- classe_v la classe de risque attribuée initialement à la zone voisine ;
- f_v une fonction de pondération des voisins permettant de définir les voisins les plus proches et semblables de la zone considérée.

Concrètement, cela signifie que la classe finale sera égale à la classe initiale éventuellement corrigée si celle-ci diffère de celles des zones voisines à caractéristiques équivalentes. Et pour déterminer les plus proches voisins de la zone considérée, il faut utiliser une notion de distance et plusieurs méthodes existent :

- La distance euclidienne permet de calculer la distance entre deux centres de maille à partir de leurs latitudes et longitudes :

$$d_E = \sqrt{(\lambda_B - \lambda_A)^2 + (\Psi_B - \Psi_A)^2} \times k. \tag{5.9}$$

- La distance de Haversine permet également de calculer cette distance mais en prenant en compte également la sphéricité de la Terre :

$$d_H = 2r \arcsin \sqrt{\sin^2\left(\frac{\Psi_B - \Psi_A}{2}\right) + \cos(\Psi_A)\cos(\Psi_B)\sin^2\left(\frac{\lambda_B - \lambda_A}{2}\right)}. \tag{5.10}$$

avec :

- Ψ_A et Ψ_B les latitudes en radians des zones A et B ;
- λ_A et λ_B les longitudes en radians des zones A et B ;
- k le facteur permettant d'être à l'échelle des kilomètres (pour la distance euclidienne)[11] ;
- r le rayon de la Terre (pour la distance de Haversine).

Sachant que le calcul de distance est utilisé pour définir les plus proches voisins d'une zone INSEE ou IRIS, il a été décidé de retenir l'approche euclidienne. En effet, bien que la méthode d'Haversine soit plus précise du fait de la prise en compte de la sphéricité terrestre, celle-ci est bien plus complexe. A noter qu'il a été vérifié que pour des calculs de proximité, comme c'est le cas dans cette étude, les résultats obtenus entre les deux techniques sont similaires. Pour construire le coefficient de crédibilité c_i propre à chaque point géographique, deux hypothèses ont été retenues (*Toesca, 2020*, [**12**]). Tout d'abord, les zones les plus denses en terme d'exposition ne doivent pas être influencées car leur volumétrie leur permet d'éviter le sur-apprentissage. Et à l'inverse, les zones à faible densité doivent être comparées aux zones environnantes pour éviter des fluctuations. Un découpage géographique est considéré comme suffisamment robuste si l'exposition associée est supérieure au quantile 99,5% de la base de modélisation. Ce seuil a été retenu car il est couramment utilisé pour des notions liées aux valeurs extrêmes (*Value-at-Risk*, cadre réglementaire Solvabilité II). La formule de calcul pour le paramètre c_i est la suivante :

$$c_i = min\left(1; \frac{\text{exposition}_i^3}{\text{quantile}(99,5\%)^3}\right). \tag{5.11}$$

11. A noter qu'un mile marin (équivalent à 1 852 mètres) est égal à $\frac{k}{60}$, soit une minute d'arc

Cette écriture mathématique a été retenue après avoir effectué des simulations sur la forme de la fonction et la valeur de l'exposant. Parmi toutes les fonctions qui ont été testées, la formule 5.11 est celle présentant les meilleures caractéristiques en termes de lissage.

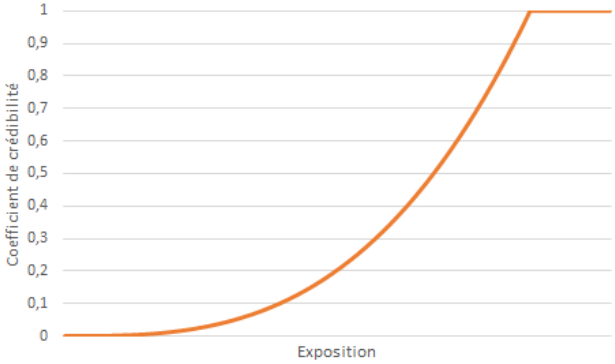


FIGURE 5.8 – Fonction de crédibilité

En effet, de par la forme convexe de la fonction retenue, il faut dépasser un certain seuil d'exposition pour que la zone considérée commence à avoir de l'importance. Par ailleurs, la confiance accordée à la zone croît de plus en plus vite avec l'exposition. L'hypothèse sous-jacente est que l'exposition de la zone est liée à la population de la ville.

La seconde composante du lissage qu'est la fonction de pondération des voisins va pouvoir maintenant être détaillée. L'hypothèse est faite que deux voisins doivent avoir une note semblable s'ils sont proches respectivement en termes de distance géographique et de niveau d'exposition (plus le polygone du voisin est proche - respectivement semblable - de celui étudié, plus sa note doit être impactante). La fonction de pondération des voisins va donc comprendre deux composantes différentes à modéliser et s'écrire sous la forme :

$$f(\text{voisin}) = f(\text{distance}) \times f(\text{exposition}). \tag{5.12}$$

Usuellement, les fonctions servant à mesurer la proximité entre des points sont de la forme $\frac{1}{\sqrt{x}}$, avec x représentant la distance entre la zone étudiée et le voisin considéré. Ainsi, plus la distance augmente, plus vite la fonction tend vers zéro. Comme pour la fonction de crédibilité, différents exposants ont été testés et par souci de cohérence, l'écriture suivante $\frac{1}{\sqrt[3]{x}}$ a été retenue pour modéliser la fonction de pondération des voisins.

Concernant la seconde composante qu'est la fonction d'exposition, afin d'accorder le même poids aux deux composantes et ainsi éviter des biais, cette fonction sera de la même famille de lois que pour la distance. Ainsi, elle est de la forme :

$$f(\text{exposition}) = \frac{1}{\sqrt[3]{|x-m|+1}}, \tag{5.13}$$

avec :

- x l'exposition du polygone voisin ;
- m l'exposition du polygone considéré.

A noter, le fait d'ajouter une unité au dénominateur permet de gérer les cas où ni le polygone observé ni son voisin ne possèdent d'exposition. Sans cette précaution, il y aurait une division par zéro qui rendrait le calcul impossible. Par ailleurs, contrairement à la fonction de distance qui aura toujours la même représentation, la forme de la fonction d'exposition sera différente selon le polygone considéré du fait d'une translation selon le nombre de contrats associé.

Une fois les deux coefficients calculés (dont les fonctions associées sont représentées ci-dessous), ceux-ci sont multipliés entre eux pour donner une valeur à la fonction de pondération des voisins.

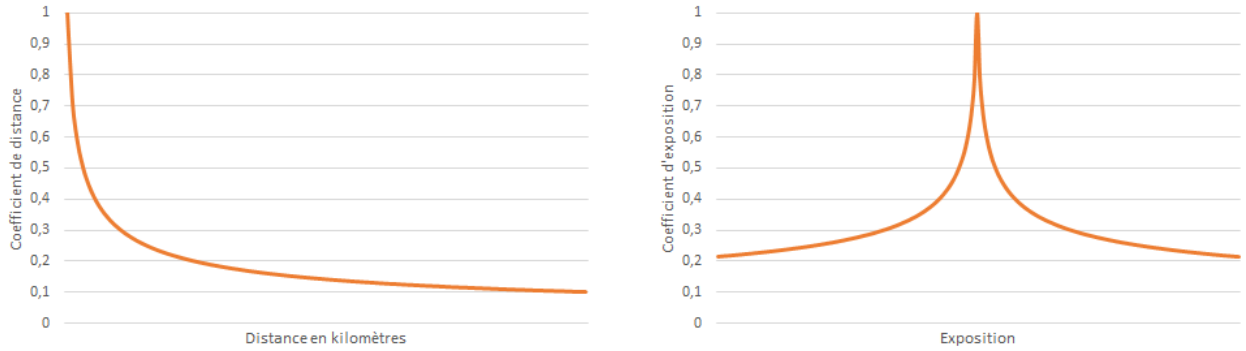


FIGURE 5.9 – Fonctions de distance et d'exposition composant la fonction de pondération des voisins

Toutes les composantes de la fonction de lissage ayant été présentées, il est possible d'analyser les résultats obtenus pour les zoniers Fréquence et Coût moyen.

5.4 Résultats obtenus et impacts sur la modélisation GLM

A l'issue du lissage effectué, tous les polygones géographiques ont une valeur de classe. Les zones qui ne disposaient d'aucune exposition voient leur classe dépendre uniquement des classes associées aux zones voisines. Enfin, celles ayant une exposition supérieure au quantile 99,5% de la base de modélisation ne voient pas la valeur de leur classe modifiée. En analysant la carte du zonier Fréquence ci-dessous[12], il est ainsi possible de remarquer que les tâches colorées (relatives le plus souvent à un seul sinistre sur un seul contrat dans une zone peu dense) ont grandement disparu, sans pour autant que cela n'affecte les zones risquées que sont les principales agglomérations françaises.

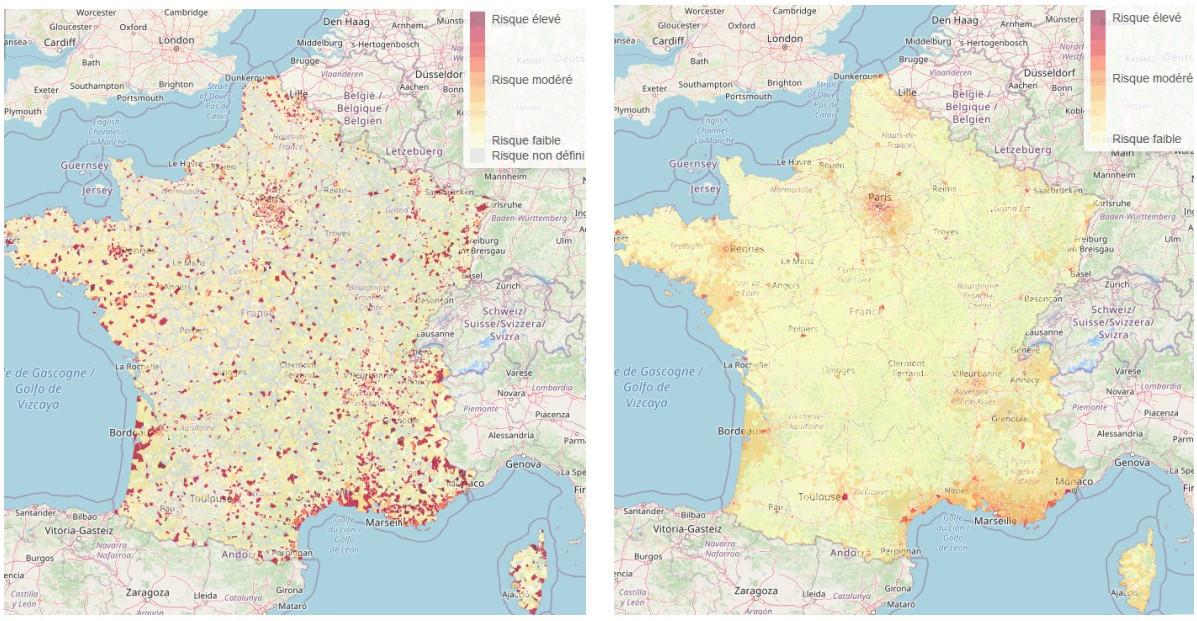


FIGURE 5.10 – Zonier Fréquence à la maille INSEE avant et après lissage

12. Les cartes pour les autres zoniers sont présentées en Annexe III.

Maintenant que toutes les zones sont associées à une classe de risque, il est possible d'injecter cette nouvelle variable dans les modèles GLM pénalisés retenus. Il faut bien préciser que pour ces nouveaux modèles, les coefficients obtenus précédemment vont être inchangés et resteront fixés. Cela va permettre ainsi de mesurer précisément l'effet du zonier dans la modélisation. Le tableau suivant synthétise les principaux indicateurs de performance en fonction du modèle considéré.

	Fréquence		Coût moyen	
Modèle	RMSE	Gini	RMSE	Gini
Sans données externes ni zonier	0,2078	29,98%	1 775,77	10,97%
Avec données externes	0,2067	34,74%	1 750,25	11,93%
Avec données externes et zonier INSEE	**0,2065**	**35,21%**	**1 543,54**	**12,09%**
Avec données externes et zonier IRIS	0,2065	35,24%	1 548,31	12,19%
Avec zonier INSEE	0,2066	35,05%	1 740,03	11,96%
Avec zonier IRIS	0,2066	35,09%	1 610,03	12,04%

TABLE 5.3 – Indicateurs de performance pour les deux composantes de la prime pure

Il est possible de remarquer qu'ajouter de l'information géographique permet d'améliorer grandement la performance des modèles (hausse de l'indice de Gini et baisse de la RMSE), que ce soit à partir de données externes ou d'un zonier. Cependant, le gain réalisé en ajoutant simultanément ces deux éléments n'est pas extrêmement important, celui-ci s'élevant à quelques dixièmes de points de Gini. Par ailleurs, bien que la maille IRIS soit plus fine que la maille INSEE, les résultats ne diffèrent guère en termes d'indicateurs. Il ne semble donc pas pertinent de conserver l'approche IRIS qui ne solutionnerait pas certaines des problématiques du modèle actuel en Voronoï. Il faudrait toujours un zonier à la maille INSEE pour les adresses inconnues ou géocodées de manière incorrecte. Le découpage INSEE est donc retenu.

Plusieurs questions se posent alors : Est-ce envisageable d'effectuer un modèle de prime pure satisfaisant sans qu'aucun zonier ne soit intégré ? Est-il, au contraire, préférable de conserver une approche traditionnelle en ne disposant que d'un zonier comme élément géographique ? En effectuant un modèle incluant des données externes mais aucun zonier, cela pourrait permettre de simplifier le processus de mise à jour tarifaire (*Pariente, 2017*, [**9**]). En effet, il n'est donc plus nécessaire d'effectuer un certain nombre d'étapes complexes pour définir les zones de risque (calcul des résidus, agrégation, segmentation puis lissage).

Cependant, plusieurs problématiques se posent avec l'utilisation de données externes. Tout d'abord, il a pu être noté, en effectuant cette étude, qu'un travail important de mise en forme et de nettoyage des données est nécessaire avant de pouvoir les utiliser. Par ailleurs, il existe un réel sujet sur l'accès futur à ces mêmes données. Par exemple, des données publiques utilisées actuellement peuvent très bien être en accès restreint lors de prochaines études, voire ne plus être publiées. Au contraire, en optant pour un modèle ne comprenant pas de données externes mais un zonier, la maintenance des modèles sera toujours possible et la question de l'accès aux données ne se pose pas. De plus, le fait d'utiliser uniquement un zonier permet d'éviter d'avoir des fluctuations tarifaires importantes entre deux zones de risque. En effet, les classes de risque ont fait l'objet d'un lissage au contraire des données externes.

La carte ci-dessous montre que l'information géographique portée par les données externes est bien plus fluctuante du fait du caractère brut des données. Il faudrait effectuer également un lissage afin de réduire cette volatilité.

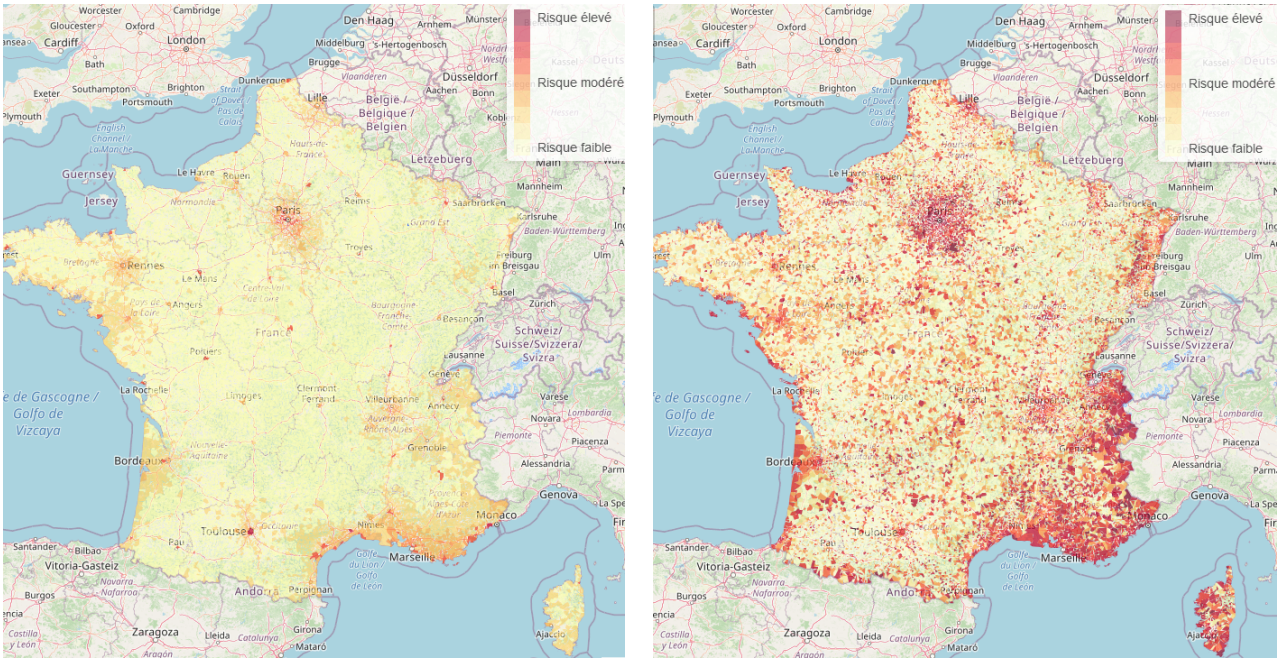


FIGURE 5.11 – Information géographique portée par le zonier ou les données externes

Afin de sélectionner le modèle le plus pertinent entre les trois propositions (zonier INSEE seul, données externes seules, combinaison des deux), les erreurs de prédiction obtenues ont été comparées à celles générées par le modèle initial comprenant uniquement des données externes.

	Écart observé		
Segment commercial	Données externes	Zonier INSEE	Combinaison des deux
Locataire	-4,01%	-4,35%	-2,56%
Propriétaire	-2,08%	+0,06%	-1,98%
Locataire 1 pièce	+0,54%	+1,98%	+1,37%
Locataire 2 pièces	-0,32%	-0,08%	-0,28%
Locataire 3 pièces	-6,21%	-7,24%	-4,45%
Locataire 4 pièces et plus	-6,29%	-6,98%	-3,72%
Propriétaire 1 pièce	-17,87%	-16,85%	-9,82%
Propriétaire 2 pièces	-9,14%	-6,48%	-9,22%
Propriétaire 3 pièces	-1,36%	+0,38%	-2,24%
Propriétaire 4 pièces et plus	+1,98%	+4,26%	+2,03%
Global	-3,07%	-2,30%	-2,16%

TABLE 5.4 – Écart de prime pure par segment commercial

Quelque soit l'approche considérée, le modèle sous-estime toujours la sinistralité, mais l'écart tend à se réduire par rapport au modèle initial. Cependant, en analysant plus en détails les principaux segments commerciaux, des différences significatives apparaissent entre les modèles retenus.

En utilisant uniquement des données externes, les écarts étaient dus aux segments minoritaires en portefeuille qui n'étaient pas correctement modélisés (locataires de trois pièces et plus ou propriétaires de moins de trois pièces). Lorsque le zonier INSEE est utilisé seul, l'amélioration de cette erreur globale (-2,30% contre -3,07%) n'est due qu'au segment des propriétaires (écart quasi nul au global malgré des valeurs de près de 17% sur certains segments), au détriment du profil majoritaire en portefeuille que sont les locataires (dégradation de 0,3 point de l'écart). Enfin, en disposant d'un modèle complet, les écarts observés sont bien plus pertinents. En effet, l'erreur est réduite de près de moitié sur le segment principal des locataires et est en légère amélioration pour les propriétaires par rapport au modèle initial. Par ailleurs, il est possible de remarquer que toutes les erreurs qui étaient importantes ont diminué significativement (gain de 2,5 points sur le segment des locataires de quatre pièces et plus, gain de 8 points sur les propriétaires d'une pièce).

Après avoir pris en compte l'ensemble des résultats présentés précédemment, pour refondre la prime pure de la garantie Dégât des eaux (sur le périmètre des appartements), **le modèle final retenu sera celui combinant les deux types d'informations géographiques** (données externes et zonier à la maille INSEE). Contrairement à l'approche traditionnelle qui vise à n'utiliser des variables externes que dans la modélisation du résidu du modèle servant à construire le zonier, le risque géographique est, ici, appréhendé par le zonier et les données externes.

Il est intéressant de noter que pour des modèles équivalents en terme de qualité de segmentation et d'erreur de modélisation, les résultats affinés diffèrent nettement. Le modèle incluant uniquement un zonier aurait pu être également sélectionné d'après les valeurs des indicateurs de modélisation. Mais l'analyse des segments commerciaux a montré que les résultats obtenus n'étaient pas pertinents. Une extension possible pour améliorer ce modèle pourrait être de créer une variable dite croisée construite à partir de la qualité de l'occupant et du nombre de pièces (malgré le fait que la corrélation entre ces deux variables ne soit pas importante[13]). En effet, il est courant qu'un locataire occupe un bien avec un nombre de pièces moins important qu'un propriétaire, et donc cette nouvelle variable pourrait capter une tendance actuellement non prise en compte.

Une dernière étape est de comparer les performances du modèle retenu à celles du modèle actuellement en production (à la maille Voronoï). Sur la composante fréquence, le nouveau modèle est légèrement inférieur à celui actuellement implémenté dans les systèmes informatiques bien que les mailles géographiques considérées soient très différentes (passage d'une tarification à l'adresse avec un micro-zonier à une tarification à la commune). Concernant le coût moyen, la valeur du Gini est équivalente entre les deux modèles.

	Maille	Fréquence	Coût moyen
Modèle actuel	Voronoï	36,57%	12,14%
Nouveau modèle	INSEE	35,61%	12,12%

Table 5.5 – Valeurs de Gini pour le modèle en production et le modèle retenu

13. Le V de Cramer associé s'élève à 0,26, voir Annexe I.

Le modèle retenu dans cette étude possède les caractéristiques suivantes :

- La segmentation du risque du nouveau modèle est similaire à celui implémenté en production ;
- L'approche opérationnelle a été grandement simplifiée du fait qu'il n'existera plus qu'un seul zonier en production contre deux actuellement. Par ailleurs, la problématique liée aux adresses inconnues a été solutionnée ;
- L'approche commerciale a été améliorée en limitant les fluctuations tarifaires entre des zones proches. En effet, le passage a une maille INSEE a permis, de supprimer les écarts au sein d'une même commune. Et en effectuant un lissage prenant en compte les caractéristiques des zones environnantes, les écarts tarifaires entre des communes voisines ont ainsi été limités.

Ce nouveau modèle est donc légitime pour remplacer le modèle actuel. Sachant que ces travaux liés au dégât des eaux s'inscrivent dans une logique de refonte complète du tarif des principales garanties du produit Habitation, il peut être intéressant d'effectuer, dans une logique d'ouverture, un état des lieux de la rentabilité des contrats, selon les segments commerciaux avec l'ensemble des primes pures actuellement utilisées dans le tarif.

5.5 Suivi de la rentabilité du produit Habitation

5.5.1 Indicateurs de suivi de la rentabilité

Deux principaux indicateurs sont couramment utilisés pour suivre la rentabilité du portefeuille :

- Le ratio de sinistralité, appelé également ratio **S**inistres à **P**rimes ou S/P, est un ratio de perte calculé afin de vérifier la rentabilité d'un portefeuille ou d'un segment. La formule de calcul de ce ratio est :

$$\text{S/P} = \frac{\text{Montant des sinistres}}{\text{Montant des primes acquises}}. \tag{5.14}$$

  Cet indicateur permet de savoir si le tarif couvre le risque. Lorsque celui-ci est supérieur à 1, cela indique que le montant des primes n'est pas suffisant pour absorber les sinistres.

- Le ratio de sinistralité attendu, appelé plus couramment ***E**xpected **L**oss **R**atio* ou ELR, reflète la rentabilité du contrat en comparant la prime pure modélisée par l'assureur à la prime réellement payée par l'assuré. Il est défini de la manière suivante :

$$\text{ELR} = \frac{\text{Montant des primes pures}}{\text{Montant des primes commerciales hors taxes}}. \tag{5.15}$$

  L'ELR est une prédiction du S/P, comme l'indique le caractère attendu de son nom. En effet, la prime pure est une estimation de la charge que devra payer l'assureur en cas de sinistre. Cet indicateur permet de mesurer une sous ou sur-tarification éventuelle d'un contrat par rapport à son risque. Un ELR supérieur à 100% reflète un contrat pour lequel la prime reçue ne permet pas à l'assureur de couvrir les sinistres. Ce contrat sera donc sous-tarifé.

L'objectif de tout assureur est d'obtenir des indicateurs de S/P et ELR faibles ainsi qu'une convergence de l'ELR vers le S/P. Cependant, bien que couramment utilisé, l'ELR ne permet pas d'avoir une vision réelle de la rentabilité d'un segment ou d'un portefeuille. En effet, les montants comparés ne sont pas de la même grandeur. La prime pure est une vision technique du risque, sans aucune prise en compte d'un aspect commercial, à l'inverse de la prime commerciale.

Cette dernière inclut, outre une marge et des rabais ou majorations, une couverture d'un certain nombre de charges telles que :

- les frais d'acquisition correspondant aux frais engagés par la compagnie lorsqu'un prospect souscrit un contrat ;
- les frais de gestion administratifs utilisés pour couvrir les charges liées aux fournitures, aux salaires des employés et aux frais d'entretien ou de location des locaux ;
- les commissions utilisées pour rémunérer les apporteurs d'affaires (agents, courtiers voire salariés de la compagnie), que ce soit lors de la souscription de nouveaux contrats ou pour le maintien de leur portefeuille de contrats ;
- les frais de réassurance correspondant aux frais engagés par la compagnie pour se couvrir contre des risques extrêmes et transférer, le cas échéant selon le type de traité, une partie des sinistres auprès d'un réassureur ;
- les frais de gestion des sinistres.

Afin d'affiner le suivi de la rentabilité, il faut donc ajouter ces différentes composantes à l'ELR. Dans ce cas, l'indice se nomme ECR (***E**xpected **C**ombined **R**atio*) et est défini comme :

$$\text{ECR} = \frac{\text{Montant des primes pures} + \text{Montant des frais}}{\text{Montant des primes commerciales hors taxes}}. \tag{5.16}$$

Un ECR supérieur à 100% reflète une situation de perte pour l'assureur, le tarif vendu étant inférieur au risque estimé auquel vient s'ajouter les frais.

A partir de cet indicateur ECR, il a été possible de définir un ELR dit d'équilibre (noté ELR_{EQ}) qui correspond à la valeur d'ELR permettant de ne pas avoir de pertes, tout en tenant compte des frais.

$$\text{ELR}_{\text{EQ}} = 1 - \frac{\text{Montant des frais}}{\text{Montant des primes commerciales hors taxes}}. \tag{5.17}$$

Une fois les indicateurs de rentabilité définis, il est possible de présenter les résultats obtenus pour les principaux segments commerciaux.

5.5.2 Résultats obtenus

Afin d'être cohérent avec l'ensemble du mémoire qui n'aborde que le périmètre des appartements, l'analyse des segments ne se concentrera uniquement que sur ce type de biens. Par ailleurs, les valeurs présentées dans cette section ont été volontairement modifiées dans un souci de confidentialité, sans que cela n'affecte cependant la cohérence des résultats ni les conclusions qui en découleraient. La rentabilité des affaires nouvelles va, dans un premier temps, être analysée.

Segment commercial	ELR	ECR	ELR_{EQ}	Poids des frais d'acquisition
Locataire	58,1%	187,8%	-29,7%	63,1%
Propriétaire	79,6%	185,5%	-6,0%	57,5%
Locataire 1 pièce	60,1%	214,3%	-54,2%	66,8%
Locataire 2 pièces	57,8%	192,8%	-35,0%	64,1%
Locataire 3 pièces	57,9%	183,0%	-25,1%	62,2%
Locataire 4 pièces et plus	57,7%	169,2%	-11,5%	59,1%
Propriétaire 1 pièce	64,1%	208,0%	-43,9%	63,1%
Propriétaire 2 pièces	77,7%	200,6%	-22,9%	60,4%
Propriétaire 3 pièces	78,7%	189,6%	-10,9%	58,4%
Propriétaire 4 pièces et plus	82,7%	173,8%	8,9%	54,0%
Global	64,2%	187,2%	-23,0%	61,7%

TABLE 5.6 – Rentabilité par segment commercial - Affaires nouvelles

En étudiant tout d'abord les valeurs prises par l'ELR, la rentabilité semble *a priori* satisfaisante étant donné que la prime commerciale couvre assez largement la charge de sinistres attendue. Par exemple, pour le segment des propriétaires avec au moins quatre pièces (qui correspond au segment avec la valeur d'ELR la plus élevée), la sinistralité ne représente que 82,7% du montant de la prime payée par le client. Or, lorsque les différents frais sont ajoutés, l'hypothèse de rentabilité est nettement remise en question.

En effet, quelque soit le segment observé, l'indicateur ECR est largement supérieur à 100%, ce qui signifie que la prime payée par le client ne permet pas de couvrir la sinistralité et les frais associés au contrat. Ces résultats dégradés se confirment par l'étude de l'ELR d'équilibre qui est ici négatif, preuve des pertes subies par la compagnie. Une valeur négative indique que, même dans une hypothèse d'absence totale de sinistralité, la prime payée ne permet pas de couvrir uniquement les frais et par conséquent le segment ne sera jamais rentable.

Bien que les conclusions déduites de cette première analyse ne soient pas positives, elles ne sont pas, pour autant, surprenantes. Cela s'explique tout d'abord par un environnement concurrentiel fort, intensifié avec l'arrivée de nouveaux acteurs tels que les bancassureurs et la mise en place de la loi Hamon en janvier 2015, qui impose aux entreprises d'assurances de proposer des tarifs toujours plus attractifs pour capter de nouveaux prospects, quitte à renier sur la rentabilité. Cette situation de perte peut également s'expliquer par des montants de frais trop élevés, en particulier sur les frais d'acquisition qui ne concernent que les affaires nouvelles.

Ces derniers se présentent en général sous la formes de forfaits en euros qui viennent augmenter sensiblement la prime commerciale. Ils représentent plus de la moitié des frais, et ce quel que soit le segment observé. Pour améliorer la rentabilité des affaires nouvelles, il faudrait donc revoir plus en détails la structure des frais car augmenter le tarif n'est pas possible à la vue de la concurrence sur le marché.

Il peut donc maintenant être intéressant d'étudier la rentabilité des contrats en portefeuille (hors affaires nouvelles) afin de voir si certains segments permettent de générer de la ressource pour l'entreprise.

Segment commercial	ELR	ECR	ELR_{EQ}
Locataire	42,7%	86,4%	56,3%
Propriétaire	50,5%	93,8%	56,6%
Locataire 1 pièce	44,3%	90,5%	53,8%
Locataire 2 pièces	44,2%	88,9%	55,3%
Locataire 3 pièces	43,0%	86,6%	56,4%
Locataire 4 pièces et plus	40,4%	82,5%	58,0%
Propriétaire 1 pièce	38,2%	86,1%	52,1%
Propriétaire 2 pièces	49,0%	94,8%	54,2%
Propriétaire 3 pièces	50,6%	94,7%	55,7%
Propriétaire 4 pièces et plus	52,2%	93,7%	58,4%
Global	46,9%	90,4%	56,5%

TABLE 5.7 – Rentabilité par segment commercial - Hors affaires nouvelles

Contrairement au segment des affaires nouvelles, la rentabilité est, ici, bien présente. En effet, quelque soit le segment observé, l'indicateur ECR est inférieur à 100%, ce qui indique qu'une fois les sinistres et les frais payés, il reste un montant qui correspond à la marge. Par exemple, tous segments étudiés confondus, l'ECR de 90,4% signifie que sur 100 € de prime perçue, les différentes charges représentent 90,40 € et donc que 9,60 € permettent d'alimenter la marge. Ces résultats se confirment par l'étude de l'ELR d'équilibre qui est, cette fois, toujours supérieur à l'ELR de base. Cela permet ainsi de voir que, quelque soit le segment considéré, l'entreprise dispose d'un coussin de sécurité si les résultats et/ou la sinistralité venaient à se dégrader.

Les résultats associés au portefeuille de contrats déjà présents depuis plusieurs années peuvent s'expliquer par différentes raisons. Tout d'abord, chaque année à l'échéance lors du renouvellement du contrat, la prime payée par le client est revue plus ou moins fortement à la hausse selon la politique de majorations annuelles. Cela permet ainsi à la compagnie d'assurances de redresser un tarif initial très attractif mais qui ne représente pas le risque réel, et ainsi de le faire converger vers le tarif cible correspondant.

Par ailleurs, une fois la première année écoulée, les frais d'acquisition disparaissent, ce qui représente une charge importante en moins à comptabiliser. Par conséquent, en augmentant le volume des primes tout en diminuant le montant des frais, il parait logique que la rentabilité s'améliore.

Maintenant que la rentabilité des deux catégories de contrats (affaires nouvelles ou non) a été analysée, il est possible de passer au suivi du portefeuille global regroupant tous les contrats.

Segment commercial	ELR	ECR	ELR_{EQ}
Locataire	44,8%	100,1%	44,6%
Propriétaire	51,9%	98,4%	53,6%
Locataire 1 pièce	47,4%	114,5%	32,8%
Locataire 2 pièces	46,4%	105,6%	40,8%
Locataire 3 pièces	44,8%	98,6%	46,2%
Locataire 4 pièces et plus	42,2%	91,4%	50,8%
Propriétaire 1 pièce	39,4%	92,0%	47,4%
Propriétaire 2 pièces	50,5%	100,5%	50,0%
Propriétaire 3 pièces	52,1%	100,0%	52,1%
Propriétaire 4 pièces et plus	53,5%	97,3%	56,2%
Global	48,5%	99,2%	49,3%

TABLE 5.8 – Rentabilité par segment commercial - Portefeuille

En tenant compte de tous les contrats en portefeuille, il est possible de constater qu'au global, celui-ci est tout juste rentable sur le segment des appartements (0,8% de marge générée). Mais l'analyse des segments permet de voir des disparités, l'ECR pouvant être supérieur à 100%. En effet, les propriétaires engendrent plus de ressources (+1,6%) que les locataires, qui sont eux juste à l'équilibre. Par ailleurs, plus le nombre de pièces est élevé, plus la rentabilité est présente. Cela est particulièrement visible sur le périmètre des locataires où la rentabilité s'améliore de 16 points entre les biens assurés d'une pièce ou de trois pièces.

Cette amélioration de la rentabilité selon le nombre de pièces s'explique de plusieurs façons. Tout d'abord, la prime pure croît avec le nombre de pièces. Par conséquent, les revalorisations tarifaires annuelles à l'aide de majorations ont plus d'impacts et permettent de faire converger la prime plus rapidement vers la valeur cible lorsque le bien assuré est plus grand. De plus, la notion de duration (qui correspond à la durée de présence moyenne en portefeuille) joue grandement sur la rentabilité. Par exemple, la duration des locataires d'une pièce est inférieure à cinq ans, quand celle des locataires de trois pièces dépasse les sept années. Non seulement, les locataires d'une pièce ont une prime faible, mais la duration n'est pas suffisamment élevée pour que les hausses annuelles sur quelques années permettent de réduire les pertes. Autre point, pour des raisons commerciales, des réductions importantes sont accordées aux étudiants (dans une logique de fidélisation et de multi-détention future). Et cette cible de prospects représente principalement des locataires de petites pièces, ce qui explique également que ces segments sont peu rentables.

Pour clore cette section et ce mémoire, l'analyse du portefeuille, sur le segment des appartements, a permis de montrer qu'il y a bien une rentabilité du fait d'une marge globale de 0,8%. Cependant, certains segments pourraient voir leur prime augmenter pour réduire les pertes. Quant aux nouvelles souscriptions qui ne sont pas rentables immédiatement, il parait difficile de maintenir un apport net positif (différence entre les affaires nouvelles et les résiliations) en proposant des tarifs plus élevés. Il faudrait alors effectuer une analyse de la sensibilité au prix pour améliorer la rentabilité sous contrainte de volume en augmentant les tarifs des prospects peu élastiques à cette hausse tarifaire.

Conclusion

Dans le but d'effectuer une refonte tarifaire de la garantie Dégât des eaux (en particulier sur le périmètre des appartements), ce mémoire a détaillé différents types de modélisation permettant ainsi de fournir une meilleure segmentation du risque. Une fois la base d'étude constituée (couvrant la période 2017-2019), celle-ci comporte pour chaque image de risque, un grand nombre d'informations liées au logement, à l'assuré, à la sinistralité, mais également provenant de bases de données externes publiques. Cette base a été traitée de façon à ce que la modélisation soit la plus précise possible en effectuant, entre autres, une gestion des sinistres graves, un développement de la charge des sinistres non clos, le retrait des sinistres sans suite ainsi que le retraitement des valeurs manquantes.

Il a donc été possible par la suite de modéliser les deux composantes de la prime pure, que sont la fréquence et le coût moyen. Cette séparation permet d'attribuer à chaque modèle les variables explicatives qui lui sont propres. La modélisation a été effectuée en utilisant des modèles linéaires généralisés (sous leurs déclinaisons pénalisées) et des méthodes d'apprentissage statistique basées sur l'agrégation d'arbres que sont les forêts aléatoires et le *Gradient Boosting*. Cependant, l'intérêt d'avoir recours à ces dernières méthodes plus sophistiquées eu égard au faible gain de précision obtenu par rapport au degré de complexité de la mise en oeuvre de cette technique n'est pas pertinent d'un point de vue opérationnel. Les modèles linéaires obtenant des résultats pertinents en terme d'efficacité technique (observée à partir de la stabilité d'indicateurs comme l'indice de Gini, la RMSE ou encore la déviance), ils ont donc été conservés pour refondre la prime pure. Il est à noter que ces modèles retenus intègrent pour la première fois des variables externes qui ont permis d'expliquer un effet que les critères tarifaires du portefeuille à eux seuls n'auraient pas pu décrire.

Ensuite, la notion de signal géographique a tenté d'être captée par la création d'un zonier. Afin de ne pas constituer un zonier présentant des corrélations avec les facteurs explicatifs de la sinistralité retenus dans l'équation tarifaire, cette approche repose sur la volonté d'expliquer la sinistralité en isolant la composante géographique du reste. Le critère zone est alors fondé uniquement sur l'effet résiduel, supposé lié au moins en partie à l'effet géographique. Deux types de mailles géographiques moins fines que celle utilisée actuellement en production ont été retenues (commune/arrondissement et îlot IRIS) afin de tenter de pallier aux problèmes de stabilité survenus lors de la construction du précédent micro-zonier pour la garantie Dégât des eaux sur les appartements. La démarche de construction peut se résumer en trois étapes majeures : agrégation, au niveau commune ou IRIS, des résidus issus du modèle linéaire retenu, classification des résidus afin de créer les zones tarifaires via les méthodes de CAH ou des *K-means*, lissage des classes définies. Concernant la création de classes, les résultats ont montré une certaine difficulté à définir un équilibre entre un nombre de classes pertinent et le degré de qualité de la classification. D'autres techniques auraient pu être explorées telles que la classification mixte (qui combine les deux approches utilisées) pour améliorer les résultats et limiter les problèmes d'exposition au sein des classes.

L'enjeu stratégique du zonier, qui est présent traditionnellement pour les garanties principales en assurance Habitation, est ici remis en cause par ce mémoire. En effet, l'intégration d'un critère de segmentation géographique ne permet pas d'améliorer nettement la performance du modèle en termes d'indice de Gini ou d'erreur de modélisation si celui-ci comprend déjà des données externes, et ce quelle que soit la maille retenue. Par ailleurs, les résultats obtenus dans ce mémoire ont montré que les performances d'un modèle en terme de qualité de prédiction et de segmentation étaient similaires entre l'utilisation d'un zonier ou de données externes.

Cependant, l'étude réalisée a permis de mettre en évidence la nécessité d'utiliser des données externes pour capter plus d'informations sur l'environnement géographique dans lequel évolue le contrat. La volumétrie et la variété de ces données disponibles pourraient ainsi permettre de bien mieux appréhender la connaissance du risque géographique d'un portefeuille. En effet, dans cette étude, seules des données publiques à une maille relativement grossière ont été utilisées. Mais en exploitant des bases de données payantes fournies par des prestataires externes, il serait possible d'avoir accès à des informations extrêmement détaillées à une maille très fine, comme cela semble être promis par la technologie LIDAR (***LI****ght* ***D****etection* ***A****nd* ***R****anging* ou détection par laser), et ainsi se passer définitivement d'un zonier. Cela pourrait permettre également de simplifier le processus de souscription en ayant un certain nombre d'informations pré-remplies (surface habitable, présence d'un jardin ou d'une piscine par exemple). Par ailleurs, il est également à préciser que les modèles finaux retenus sont dits techniques. Par conséquent, ils devront encore faire l'objet de retraitements commerciaux afin de prendre en compte la politique de compétitivité tarifaire de l'entreprise.

Enfin, ces travaux effectués dans ce mémoire vont avoir plusieurs utilités finales. Tout d'abord, le modèle de prime pure retenu sera utilisé pour suivre la rentabilité du produit Habitation via les indicateurs de l'ELR et de l'ECR. Et en fonction des résultats obtenus, les majorations annuelles seront adaptées afin de permettre une plus rapide convergence vers l'ELR d'équilibre en tenant compte de la duration moyenne. Ce modèle, une fois contraint, servira également à mettre à jour la prime commerciale payée par le client.

Synthèse

L'objectif de ce mémoire est d'effectuer une refonte de la modélisation du risque principal en assurance Habitation sur le périmètre des appartements, qui correspond au dégât des eaux. Cette étude présente le processus de construction de la base de modélisation, la création de différents modèles pour la fréquence et le coût moyen ainsi que la création d'un zonier à différentes mailles géographiques.

La base de modélisation a été structurée par image de risque sur la période 2017-2019, par le biais de la jointure de différentes sources de données internes : contrats, sinistres ou clients. En parallèle, un travail de recherche complémentaire a été réalisé afin de récupérer des variables externes correspondant à des données géographiques à différentes mailles, relatives à l'environnement du contrat et provenant de différentes sources publiques. Il est à noter qu'il s'agit, au sein de l'équipe, de la première refonte des modèles de prime pure où ce genre de données est intégré.

Cette étape de création de la base d'étude doit être effectuée soigneusement car la qualité des données est un enjeu primordial pour la modélisation. Afin que la modélisation ne soit pas perturbée par des valeurs élevées et que les observations ayant un coût important soient gérées spécifiquement, l'utilisation de la théorie des valeurs extrêmes a permis de définir un seuil de séparation : toute charge au-dessus de ce seuil est plafonné à ce montant et le surplus est ventilé sur l'ensemble des sinistres. Cela permet ainsi d'éliminer les fluctuations liées à ces sinistres élevés et ainsi d'homogénéiser les données. Par ailleurs, la période d'observation étant récente, tous les sinistres observés n'ont pas de charge finale close et définitive. Or, l'objectif est d'obtenir une prime pure en accord avec la sinistralité observée. Ainsi, pour solutionner ce problème, la charge observée a été développée en utilisant la méthode des cadences dite de Chain-Ladder.

Une fois que les données nécessaires à la modélisation ont été nettoyées et regroupées si besoin, certains retraitements ont été effectués dans le but d'obtenir une base aussi propre que possible. Les variables quantitatives ont été discrétisées et deviennent donc qualitatives, afin d'identifier plus facilement des catégories de risques au sein du portefeuille d'étude et de prendre en compte certains effets non-linéaires. Par ailleurs, les variables explicatives utilisées nécessitent d'être non-corrélées. Par conséquent, une étude des corrélations en utilisant le V de Cramer a été effectuée et aucun groupement de variables significatif n'est ressorti.

Lorsque la base de données a été finalisée, il a été possible par la suite de modéliser les deux composantes de la prime pure, que sont la fréquence et le coût moyen à partir de la théorie des modèles linéaires généralisés, dans un premier temps. Ils sont structurés de telle manière que leur composante aléatoire et leur composante déterministe sont reliées par une relation fonctionnelle, appelée aussi fonction de lien. Dans ce mémoire, ces modèles ont été étudiés sous leurs déclinaisons pénalisées (Lasso, Ridge et Elastic Net). Les trois variables les plus influentes, que ce soit sur l'apparition de la sinistralité ou sur le montant du sinistre lorsqu'il survient, sont le nombre de pièces, le prix moyen au m^2 dans la commune (ou l'arrondissement le cas échéant) et la qualité de l'occupant (locataire ou propriétaire).

Il est à noter que l'intégration de données externes était pertinente car certaines des variables ajoutées font partie des variables les plus discriminantes (exemple du prix moyen au m^2). Le modèle générant les meilleurs résultats sur la composante fréquence est l'approche pénalisée Lasso. Au niveau du coût moyen, il s'agit cette fois de l'approche Elastic Net. Ces premiers résultats ont été comparés à ceux obtenus par des méthodes d'agrégation d'arbres simples que sont les forêts aléatoires et le *Gradient Boosting*. Cependant, ces techniques d'apprentissage statistique nécessitent d'optimiser un grand nombre de paramètres. Le processus d'optimisation est effectué à l'aide d'un *grid search* qui consiste à définir un paramètre en fonction d'un critère à minimiser ou maximiser. Ces deux approches de *Machine Learning* ont des résultats semblables. En effet, l'importance des variables et les métriques d'évaluation sont assez proches. Comme pour les modèles linéaires, les trois critères les plus importants sont le nombre de pièces, le prix moyen au m^2 et la qualité de l'occupant.

A la vue des résultats, il semble plus pertinent de retenir les modèles pénalisés pour estimer la prime pure. Effectivement, face aux méthodes d'agrégation d'arbres, les indicateurs de performance sont légèrement inférieurs, mais la relative simplicité d'implémentation et d'interprétation par le biais de l'analyse des coefficients compense ces très faibles écarts.

Disposant des modèles finaux, il est maintenant possible de commencer la construction du zonier (traitement du signal géographique lors de la création du modèle de tarification). L'hypothèse est faite que malgré l'ajout des variables externes qui portent une part d'information géographique vu qu'elles caractérisent l'environnement du bien assuré, il existera toujours, à caractéristiques de risque équivalentes, des effets liés à la zone du risque que le modèle GLM ne sera pas capable de capter. L'approche retenue est une approche résiduelle.

$$\begin{aligned}\text{Risque} = &\text{ Effet non géographique capté}\\ &+ \text{Effet géographique capté} + \text{Effet géographique non capté}\\ &+ \text{Bruit},\end{aligned}$$

avec :

- L'effet non géographique capté correspondant aux variables tarifaires usuelles ;
- L'effet géographique capté correspondant aux variables externes ;
- L'effet géographique non capté correspondant au zonier calculé sur les résidus ;
- Le bruit correspondant à l'erreur finale de modélisation supposée ne contenir aucune tendance géographique.

La première étape dans la construction d'un zonier est d'agréger les résidus (correspondant à l'erreur de modélisation) obtenus lors de la modélisation de la fréquence ou du coût moyen des sinistres à une maille géographique donnée (INSEE ou IRIS). Une fois cette étape effectuée, les résidus sont segmentés selon deux méthodes usuelles : la méthode de classification hiérarchique ascendante avec la distance de Ward pour effectuer les regroupements de classes et la méthode de partitionnement des *K-means*. La difficulté de ces méthodes est de faire le bon choix au niveau du nombre optimal de classes. Pour ce faire, plusieurs indicateurs sont présentés, comme le R^2, le pseudo-F ou encore l'indice de Calinski-Harabasz.

Puis ces classes de résidus obtenues doivent être lissées pour éviter une volatilité trop importante du tarif. Pour cela, une nouvelle technique basée sur la théorie de la crédibilité a été développée. Le coefficient de crédibilité associé tient compte de l'exposition dans la zone considérée. Si la zone possède un grand nombre de contrats, alors seule l'information portée par celle-ci sera conservée. Dans le cas contraire, il y aura un lissage en tenant compte des caractéristiques de l'ensemble des polygones géographiques environnants avec une influence décroissante en fonction de la distance et de la dissimilarité de l'exposition.

L'étape de lissage est particulièrement importante car en lissant trop peu, l'hétérogénéité dans les zones est toujours présente et il y a un risque de sur-apprentissage. Et à l'inverse, en lissant de manière trop importante, la segmentation géographique est supprimée en égalisant les résidus entre eux, ce qui engendre une perte de précision. Enfin, cette variable zonier est réintégrée dans le modèle GLM initial pour que les coefficients associés aux zones soient définis, les autres coefficients précédemment calculés restant inchangés. Cela permet ainsi d'observer l'impact, toutes choses égales par ailleurs, de la variable zonier sur les données.

Différentes comparaisons entre les modèles GLM suivants ont été effectuées afin de choisir le modèle le plus pertinent et parcimonieux :

- Modèle initial avec données récupérées à la souscription et données provenant de sources externes (M1) ;
- Modèle basé sur M1 auquel est ajouté le zonier créé à la maille INSEE (M2 INSEE) ;
- Modèle basé sur M1 auquel est ajouté le zonier créé à la maille IRIS (M2 IRIS) ;
- Modèle initial avec données récupérées à la souscription, sans données externes, mais avec le zonier créé à la maille INSEE (M3 INSEE) ;
- Modèle initial avec données récupérées à la souscription, sans données externes, mais avec le zonier créé à la maille IRIS (M3 IRIS).

Tout d'abord, les modèles IRIS ne possèdent pas de meilleurs indicateurs de modélisation que les modèles INSEE. Utiliser un niveau de granularité plus fin n'apporte donc qu'un risque plus élevé de sur-apprentissage pour des performances similaires. Par ailleurs, la présence simultanée des données externes et du zonier n'améliore pas de manière importante la segmentation du risque. Il a, cependant, été décidé de comparer le modèle initial M1 aux modèles M2 et M3 afin d'analyser en détails leurs caractéristiques.

Finalement, le modèle retenu est bien le modèle complet M2 avec zonier INSEE et données externes. Une telle démarche est novatrice, car peu courante en assurance Habitation où il est usuel d'intégrer un zonier à un modèle initial ne comprenant aucune donnée externe. Une étude plus approfondie sur la qualité de prédiction du nouveau modèle a été mise en place en comparant les primes pures modélisées et observées par segments commerciaux. Il en ressort que le modèle sous-estime légèrement la sinistralité (de l'ordre de 2%), mais les principaux segments restent cependant bien ajustés. La qualité principale de ce modèle retenu est d'obtenir une qualité de modélisation similaire au modèle actuellement implémenté en production, tout en évitant les problèmes d'instabilité et de sur-apprentissage propres au micro-zonier. L'étude montre ainsi l'importance des variables externes dans la tarification, car lorsqu'elles sont ajoutées aux variables tarifaires usuelles, elles permettent d'obtenir un gain d'information important, ce qui amène à une amélioration de la qualité des modèles. Par ailleurs, l'étude révèle que l'utilisation des variables externes permet d'obtenir un modèle relativement aussi performant qu'avec un modèle composé d'un zonier.

Enfin, un suivi de la rentabilité des différents segments commerciaux a permis d'avoir plusieurs résultats. Tout d'abord, les contrats récemment souscrits ne génèrent aucun profit, conséquence d'un tarif proposé très bas pour faire face à la concurrence et d'un montant de frais d'acquisition probablement trop élevé. Mais à une vision globale du portefeuille, la compagnie arrive à générer de la marge, bien que pour certains segments, tels que les locataires de petites surfaces, certaines actions doivent être mises en place pour améliorer la situation.

Synthesis

The objective of this dissertation is to update the modeling of the main risk in Household insurance, which is water damage on the perimeter of apartments. This study presents the process of building the modeling base, the creation of different models for the frequency and the average cost as well as the creation of a zone with different geographical levels.

The modeling base was structured by risk image over the 2017-2019 period, by combining different internal data sources : contracts, claims or customers. At the same time, additional research work was carried out in order to retrieve external variables corresponding to geographic data with different levels, relating to the contract environment and coming from different public sources. This is the first update of pure premium models where this kind of data is integrated.

This step of creating the database must be carried out carefully because the quality of the data is a key issue for modeling. So that the modeling is not disturbed by outliers and that the observations having an extreme cost are managed specifically, the use of the theory of the extreme values made it possible to define a separation threshold : any load above is capped at this amount and the surplus is broken down across all claims. This thus makes it possible to eliminate the fluctuations linked to these extreme disasters and thus to homogenize the data. In addition, since the observation period is recent, not all observed claims have a final closed and definitive charge. However, the objective is to obtain a pure premium in accordance with the observed loss experience. Thus, to solve this problem, the observed load was developed using the so-called Chain-Ladder rate method.

Once the data necessary for the modeling were cleaned and grouped together if necessary, certain restatements were carried out in order to obtain a base as clean as possible. The quantitative variables have been discretized and therefore become qualitative, in order to more easily identify risk categories within the study portfolio and to take into account certain non-linear effects. Moreover, the explanatory variables used need to be uncorrelated. Therefore, a correlation study using Cramer's V was used and no significant variable groupings emerged.

When the database was finalized, it was subsequently possible to model the two components of the pure premium, which are the frequency and the average cost, starting from the theory of generalized linear models, as a first step. They are structured in such a way that their random component and their deterministic component are linked by a functional relation, also called a link function. In this dissertation, these models have been studied under their penalized variations (Lasso, Ridge and Elastic Net). The three most influential variables, whether on the appearance of the loss experience or on the amount of the loss when it occurs, are the number of rooms, the average price per m^2 in the city (or the district if applicable) and the type of the occupant (lessee or home owner). It should be noted that the integration of external data was relevant because some of the added variables are among the most discriminating variables (example of the average price per m^2). The model generating the best results on the frequency component is the penalized Lasso approach. In terms of average cost, this time it is the Elastic Net approach.

These first results were compared with those obtained by simple tree aggregation methods, namely random forests and Gradient Boosting. However, these statistical learning techniques require the optimization of a large number of parameters. The optimization process is carried out using a grid search which consists in defining a parameter according to a criterion to be minimized or maximized. The two approaches have similar results. Indeed, the importance of the variables and the evaluation metrics are quite similar. As with linear models, the three most important variables are the number of rooms, the average price per m^2 and the quality of the occupant. In view of the results, it seems more relevant to retain the penalized models to estimate the pure premium. Indeed, compared to tree aggregation methods, the performance indicators are slightly lower, but the relative simplicity of implementation and interpretation through the analysis of the coefficients compensates for these very small differences.

Having the final models, it is now possible to start the construction of the zone (processing of the geographical signal during the creation of the pricing model). The assumption is made that despite the addition of the external variables which carry a part of geographical information since they characterize the environment of the insured home, there will always be, all things equal otherwise, effects linked to the zone of the risk that the GLM model will not be able to pick up. The approach adopted is a residual approach.

$$\begin{aligned}\text{Risk} = &\text{ Non-geographical captured effect}\\ &+ \text{Geographical captured effect} + \text{Geographical uncaptured effect}\\ &+ \text{Noise},\end{aligned}$$

with :

- The captured non-geographic effect corresponding to the usual tariff variables ;
- The geographic effect captured corresponding to the external variables ;
- The uncaptured geographic effect corresponding to the zone calculated on the residues ;
- The noise corresponding to the final modeling error assumed to contain no geographic trend.

The first step in the construction of a zoning is to aggregate the residuals (corresponding to the modeling error) obtained during the modeling of the frequency or the average cost of claims at a given geographic grid (INSEE or IRIS level) . Once this step has been carried out, the residuals are segmented according to two usual methods : the ascending hierarchical classification method with the Ward distance to perform the groupings of classes and the K-means partitioning method. The difficulty with these methods is to make the right choice in terms of the number of classes. To do this, several tests are presented, such as the R^2, the pseudo-F or the Calinski-Harabasz index.

Then, these classes of residues obtained must be smoothed out to avoid excessive price volatility. For this, a new technique based on the theory of credibility has been developed. The associated credibility coefficient takes into account the exposure in the zone considered. If the zone has a large number of contracts, then only the information carried by it will be kept. Otherwise, there will be a smoothing taking into account the characteristics of all the geographic polygons with a decreasing influence depending on the distance and the dissimilarity of the exposure. The smoothing step is particularly important because by smoothing too little, heterogeneity in zones is always present and there is a risk of overfitting. And on the contrary, by smoothing too much, the geographical segmentation is removed by equalizing the residuals between them, which generates a loss of precision. Finally, this zone variable is reintegrated into the initial GLM model so that the coefficients associated with the zones are defined, the other previously calculated coefficients remaining unchanged. This makes it possible to observe the impact, all other things being equal, of the zoning variable on the data.

Different comparisons between the following GLM models were made in order to choose the most relevant and parsimonious model :

- Initial model with data retrieved at subscription and data from external data (M1) ;
- Model based on M1 to which is added the zonier created at the INSEE level (M2 INSEE) ;
- Model based on M1 to which is added the zonier created at the IRIS level (M2 IRIS) ;
- Initial model with data retrieved at subscription, without external data, but with the zone created at the INSEE level (M3 INSEE) ;
- Initial model with data retrieved at subscription, without external data, but with the zone created at the IRIS level (M3 IRIS).

First of all, IRIS models do not have better modeling indicators than INSEE models. Using a finer level of granularity therefore only brings a higher risk of overfitting for similar performance. In addition, the simultaneous presence of external data and the zoning does not significantly improve the risk segmentation given the complexity required for the construction of geographical risk zones. However, it was decided to compare the initial M1 model with models M2 and M3 in order to analyze their characteristics.

The model retained is indeed the complete model M2 with INSEE zoning and external data. Such an approach is innovative, because it is very common in Household insurance to create a zoning variable and add after to the initial model which include no external data. A more in-depth study on the prediction quality of the new model was carried out by comparing the pure premiums modeled and observed by commercial segment. It emerges that the model slightly underestimates the loss experience (of the order of 2%), but the main segments nevertheless remain well adjusted. The main quality of this selected model is to obtain a modeling quality similar to the model currently implemented in production, while avoiding the problems of instability and overfitting specific to this kind of zoning. The study thus shows the importance of external data in pricing, because when they are added to the usual pricing variables, they make it possible to obtain a significant gain in precision, which leads to an improvement in the quality of the models. In addition, the study reveals that the use of external data makes it possible to obtain a model that is relatively as efficient as with a model composed of a zoning.

Finally, monitoring the profitability of the various commercial segments has produced several results. First of all, the contracts recently taken out do not generate any profit, a consequence of a very low price offered to face the competition and an amount of acquisition costs that is probably too high. But with a global vision of the portfolio, the company manages to generate margin, although for certain segments, such as tenants of small surfaces, certain actions must be taken to improve the situation.

Bibliographie

[1] Bernanose A. (2020) *Modélisation du risque Incendie en assurance MultiRisques Habitation, Mémoire d'actuariat, ISUP*

[2] Boyer C., Sangnier M. (2020) *Machine Learning. Cours de seconde année - Certificat d'Expertise Actuarielle, Institut du Risk Management*

[3] Friedman J. (2001) *Greedy function approximation : a gradient boosting machine, The Annals of Statistics, volume 29, pages 1189 - 1232*

[4] Gorrand R. (2020) *Assurance dommage : Tarification a priori et a posteriori. Cours de seconde année - Certificat d'Expertise Actuarielle, Institut du Risk Management*

[5] Halimi M. (2017) *Réactualisation des méthodes classiques de tarification en IARD. Mémoire d'actuariat, ENSAE*

[6] Hastie T., James G., Tibshirani R. et Witten D. (2013) *An Introduction to Statistical Learning With Applications in R, Springer-Verlag New York Inc.*

[7] Jollois F.-X. (2010) *Méthodes de classification. Cours de seconde année - DUT STID, Université Paris Descartes*

[8] Lopez O. (2018) *Économétrie de l'assurance. Cycle de perfectionnement - Certificat d'Expertise Actuarielle, Institut du Risk Management*

[9] Pariente J. (2017) *Modélisation du risque géographique en assurance Habitation. Mémoire d'actuariat, Université Paris Dauphine*

[10] Sauveplane P. (2019) *Théorie du risque et réassurance. Cours de première année - Certificat d'Expertise Actuarielle, Institut du Risk Management*

[11] Thomas M. (2020) *Théorie des valeurs extrêmes. Cours de seconde année - Certificat d'Expertise Actuarielle, Institut du Risk Management*

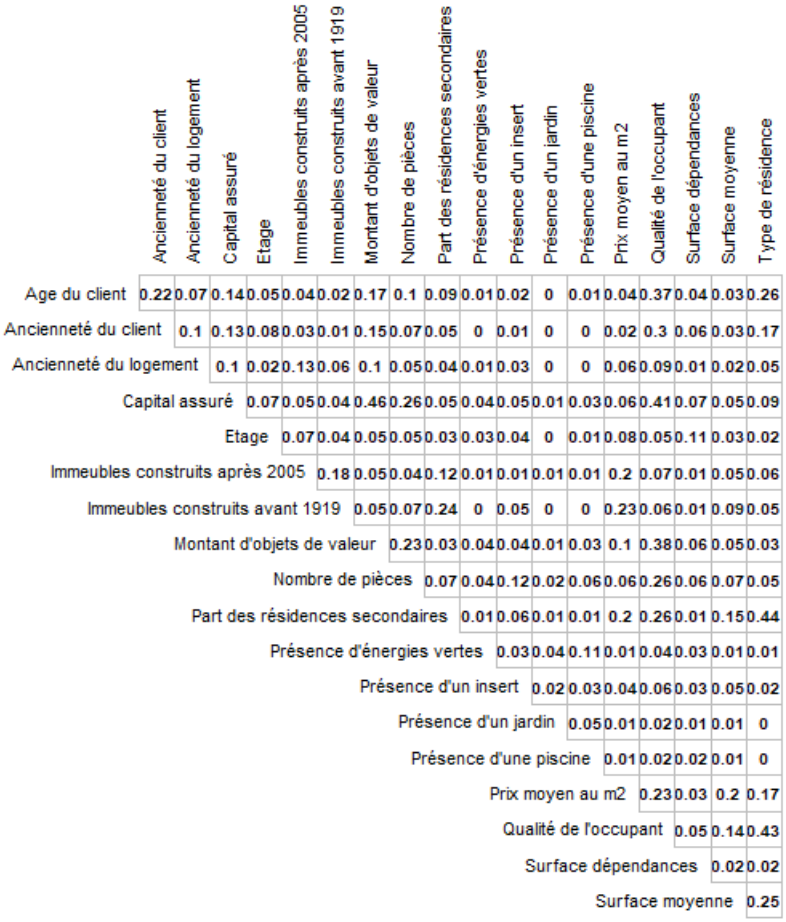
[12] Toesca R. (2020) *Méthode GEOREV. Présentation interne, AXA*

Table des figures

Liste des tableaux

Annexes

Annexe I. Corrélogramme détaillé



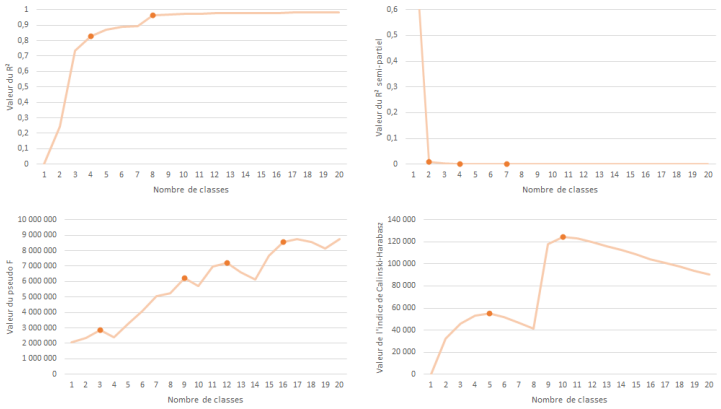
	Ancienneté du client	Ancienneté du logement	Capital assuré	Etage	Immeubles construits après 2005	Immeubles construits avant 1919	Montant d'objets de valeur	Nombre de pièces	Part des résidences secondaires	Présence d'énergies vertes	Présence d'un insert	Présence d'un jardin	Présence d'une piscine	Prix moyen au m2	Qualité de l'occupant	Surface dépendances	Surface moyenne	Type de résidence
Age du client	0.22	0.07	0.14	0.05	0.04	0.02	0.17	0.1	0.09	0.01	0.02	0	0.01	0.04	0.37	0.04	0.03	0.26
Ancienneté du client	0.1		0.13	0.08	0.03	0.01	0.15	0.07	0.05	0	0.01	0	0	0.02	0.3	0.06	0.03	0.17
Ancienneté du logement	0.1			0.02	0.13	0.06	0.1	0.05	0.04	0.01	0.03	0	0	0.06	0.09	0.01	0.02	0.05
Capital assuré	0.07				0.05	0.04	0.46	0.26	0.05	0.04	0.05	0.01	0.03	0.06	0.41	0.07	0.05	0.09
Etage	0.07				0.04	0.05	0.05	0.03	0.03	0.04	0	0.01	0.08	0.05	0.11	0.03	0.02	
Immeubles construits après 2005	0.18					0.05	0.04	0.12	0.01	0.01	0.01	0.01	0.2	0.07	0.01	0.05	0.06	
Immeubles construits avant 1919	0.05						0.07	0.24	0	0.05	0	0	0.23	0.06	0.01	0.09	0.05	
Montant d'objets de valeur	0.23							0.03	0.04	0.04	0.01	0.03	0.1	0.38	0.06	0.05	0.03	
Nombre de pièces	0.07								0.04	0.12	0.02	0.06	0.06	0.26	0.06	0.07	0.05	
Part des résidences secondaires	0.01									0.06	0.01	0.01	0.2	0.26	0.01	0.15	0.44	
Présence d'énergies vertes	0.03										0.04	0.11	0.01	0.04	0.03	0.01	0.01	
Présence d'un insert	0.02											0.03	0.04	0.06	0.03	0.05	0.02	
Présence d'un jardin	0.05												0.01	0.02	0.01	0.01	0	
Présence d'une piscine	0.01													0.02	0.02	0.01	0	
Prix moyen au m2	0.23														0.03	0.2	0.17	
Qualité de l'occupant	0.05															0.14	0.43	
Surface dépendances	0.02																0.02	
Surface moyenne	0.25																	

Le corrélogramme présenté ci-dessus permet d'avoir la valeur du V de Cramer pour l'ensemble des croisements de variables deux à deux. Aucune variable n'est fortement corrélée à une autre : le maximum étant de 0,46 (entre le capital assuré et le montant d'objets de valeur).

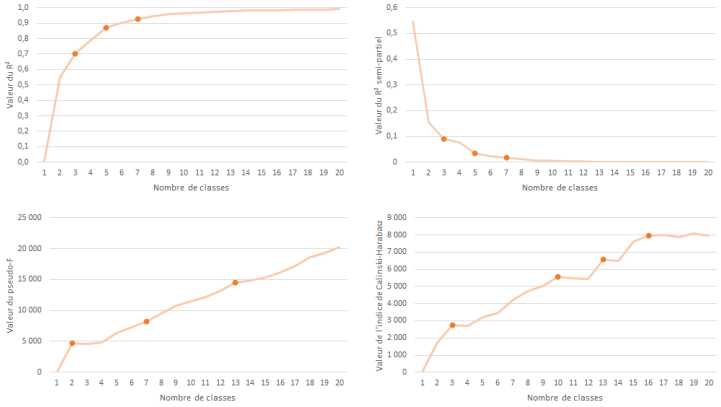
Annexe II. Indicateurs de classification détaillés

Les résultats du nombre de classes optimal proposés par les différents indicateurs n'ayant été présentés dans le corps du mémoire que pour le modèle de Fréquence INSEE, il est possible de retrouver le détail pour les autres composantes dans cette annexe.

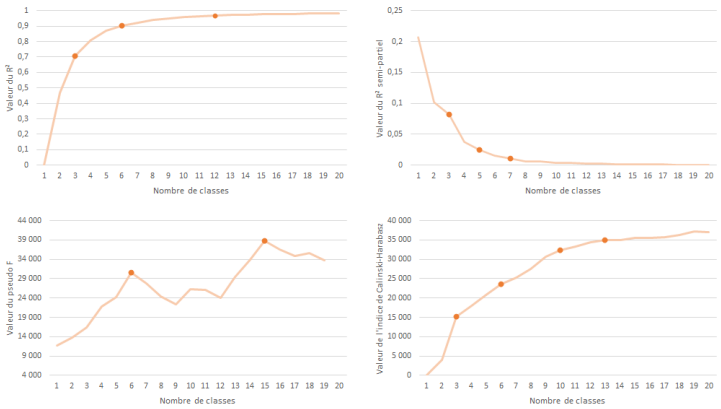
Fréquence - maille IRIS



Coût moyen - maille INSEE



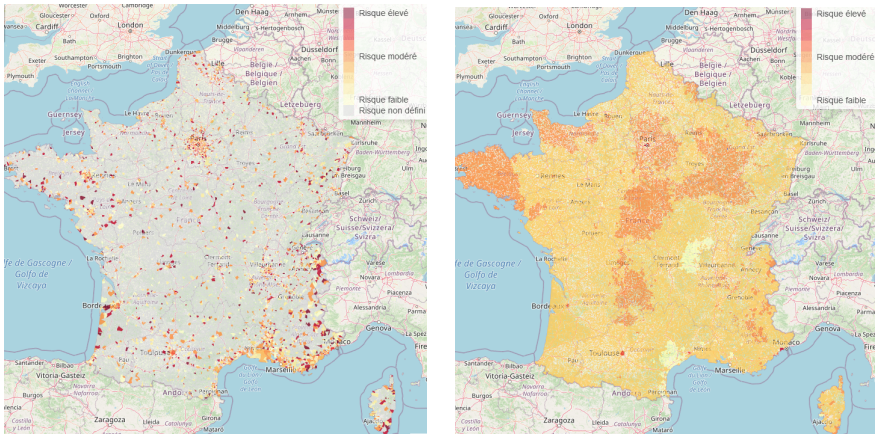
Coût moyen - maille IRIS



Annexe III. Zoniers détaillés

Le résultat du lissage des zones de risque n'ayant été présenté dans le corps du mémoire que pour le modèle de Fréquence INSEE, il est possible de retrouver le détail pour les autres composantes dans cette annexe.

Coût moyen - maille INSEE



Fréquence - maille IRIS - Région Île-de-France

Coût moyen - maille IRIS - Région Île-de-France