Mémoire présenté le : 14/09/2021

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

Par : DESHAYES Pierre

Titre **Pilotage technique de la sinistralité Incendie d'un produit Habitation par modélisation de la prime pure extrême**

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

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

Membre présents du jury de l'Institut des Actuaires signature

Membres présents du jury de l'ISFA

Entreprise : COVEA

Nom : *Ewen WILCZYK*

Signature :

Directeur de mémoire en entreprise :

Nom : Mathieu RIOULT

Signature :

Invité

Nom :

Signature :

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

Signature du responsable entreprise

Signature du candidat

Sommaire

Résumé

En assurance Habitation destinée aux particuliers, la prime pure Incendie représente une part conséquente de la prime pure totale. La rentabilité de cette garantie est fortement dégradée par des sinistres d'intensité forte mais avec une fréquence faible. Il est donc primordial de comprendre et de piloter cette sinistralité selon les profils de risque. Pour ce faire, il est nécessaire de caractériser ces sinistres par un seuil. On s'appuie sur la théorie des valeurs extrêmes afin de déterminer un seuil au-delà duquel le sinistre est défini comme extrême.

On observe alors comment la charge extrême Incendie se développe. On utilise d'une part la méthode Chain Ladder pour liquider les sinistres à la garantie, au produit mais aussi au critère de risque. On observe alors les limites de cette approche en fonction des garanties.

La modélisation de la fréquence grave et de la prime pure grave Incendie (ou extrême) se fait en partant de l'hypothèse d'indépendance entre fréquence et coût moyen. Cela permet de décomposer l'étude en 4 sous-modèles : fréquence, coût moyen, probabilité d'un sinistre d'être grave puis coût moyen grave. Cette partie de l'étude est l'opportunité de challenger différentes méthodes. On utilise des méthodes de régression linéaire généralisée ou logistique ainsi qu'une méthode Gradient Boosting Machine.

Une fois les modèles validés, on les implémente dans le processus de pilotage de la garantie. Cela nous permet d'avoir une vision plus juste des indicateurs de rentabilité de l'Incendie par profil de risque des assurés. Les mesures tarifaires de l'année suivante se baseront en partie sur ces nouveaux indicateurs techniques.

Mots Clés : Assurance Habitation, Incendie, IARD, Chain-Ladder, Théorie des Valeurs Extrêmes, Modèles Linéaires Généralisés, Gradient Boosting Machine, Pilotage, Rentabilité technique.

Abstract

For Home Insurance for individuals, the fire pure premium represents a large part of the total pure premium. The profitability of this guarantee is greatly degraded by claims of high intensity and low frequency. It is essential to understand and manage these losses according to the risk profiles. To do this, it is necessary to characterize these claims by a threshold. I rely on the theory of extreme values to determine this threshold beyond which, the loss is defined as extreme.

We observe how the extreme fire claims are developing. We use the Chain Ladder method to settle claims against the guarantee, the product but also the risk criterion. We then observe the limits of this approach.

The modelization of the extreme frequency and the extreme pure premium is done with the hypothesis of independence between frequency and average cost. This allows the study to be decomposed into four sub-models: frequency, average cost, extreme frequency and extreme average cost. This part is the opportunity to challenge different methods. We use generalized linear model as well as Gradient Boosting Machine.

Once the models have been validated, they are implemented in the guarantee management process. This allows use to have a better vision of the indicators of profitability of the Fire guarantee by risk profile of the insured. Tariff measures for the following years will be based in part on these new technical indicators.

Keywords: *Home Insurance, Fire, P&C insurance, Chain-Ladder, Theory of extreme value,* generalized linear model*, Gradient Boosting Machine, Pilotage, Profitability.*

Remerciements

Je remercie mon tuteur d'alternance Mathieu RIOULT ainsi que mon manager Ewen WYLCZIK qui m'ont accompagné et guidé tout au long de cette année. Je remercie également le reste de l'équipe Tarification Statistique et Pilotage MMA pour leurs conseils et leur soutien lors de ces deux années.

Enfin, je remercie le corps enseignant du Master Actuariat de l'ISFA pour la formation apportée qui m'a permis d'avoir toutes les connaissances nécessaires ainsi que l'autonomie pour mon entrée sur le marché du travail.

Introduction

Le marché de l'assurance IARD particuliers est fortement concurrentiel. Une revalorisation du tarif peut alors rapidement mener à une chute conséquente de portefeuille lié à un mauvais positionnement tarifaire par rapport à la concurrence. Les produits sur ce marché sont alors tout justes à l'équilibre en termes de rentabilité. A cela s'ajoute un contexte financier avec des taux nuls, voir négatifs qui ne permettent plus de dégager de rendement ce qui ne compense plus l'absence de rentabilité du marché IARD particuliers. Il apparait alors nécessaire de re-dégager de la marge par le tarif. Un pilotage efficace et le plus juste possible des résultats des produits d'assurance est donc indispensable afin de justifier les revalorisations tarifaires.

Au sein de la marque MMA, ce changement d'environnement de taux s'accompagne d'une nouvelle offre MRH destinée aux particuliers en production depuis peu. Cela a permis à la marque de se repositionner sur le marché de l'assurance Habitation. Cela a été l'occasion de revoir le pilotage de la sinistralité technique du produit. La garantie Incendie représentant une forte part de la sinistralité MRH, c'est donc le pilotage de cette garantie qui a été travaillé dans le cadre de ce mémoire. La charge globale de cette garantie est fortement portée par des sinistres de faible fréquence mais avec une sévérité importante. C'est donc par cet aspect d'extrême qu'est retraitée la charge sinistre Incendie.

Pour travailler ces sinistres extrêmes, il est nécessaire de les définir. Le choix qui est fait est de définir un sinistre extrême comme un sinistre ayant une charge à date dépassant un seuil. Ce seuil est déterminé à partir de résultats de la Théorie des Valeurs Extrêmes. Cela nous permet alors de distinguer la charge attritionnelle de la charge grave.

Chez MMA, la sinistralité est travaillée selon deux aspects. Un premier aspect dit « économique » qui représente la charge à l'ultime de l'année, puis un second dit « technique » qui représente le niveau de sinistralité attendue. Les deux aspects sont retravaillés dans ce mémoire. La charge « économique » est liquidée par garantie, par produit et critère de risque. La charge « technique » est retravaillée grâce aux modèles de grave construits par la suite.

L'objectif est d'obtenir des indicateurs de rentabilité « techniques » qui ne seront pas perturbés par la survenance d'un sinistre de forte intensité sur une population bien précise. Les modèles permettront de décrire le risque Incendie selon le profil de risque. Le nouvel indicateur se doit alors d'être stable dans le temps et de refléter le risque Incendie couvert par le contrat. La rentabilité observée de cette population peut se dégrader par un seul sinistre. L'idée est de répartir ce sinistre extrême non pas uniquement sur la population à laquelle il appartient mais bien à l'ensemble du produit selon les profils de risque.

C'est pour cela que nous souhaitons comprendre le risque Incendie et en particulier sa partie extrême. La modélisation de prime pure dans cette étude n'a pas pour but de valoriser un tarif mais à piloter la charge Incendie. Elle a vocation de décrire le risque grave Incendie. Pour ce faire, la prime pure grave a été décomposée en deux parties.

La première partie consiste à modéliser la fréquence de survenance d'un grave Incendie. Plutôt que de modéliser de façon directe la fréquence, il été choisi de procéder en deux étapes : une fréquence Incendie tous sinistres confondus, puis la probabilité d'un sinistre d'être grave.

La fréquence Incendie est modélisée par une GLM. On challenge la méthode classique de régression logistique sur la probabilité d'être grave. En effet, l'utilisation d'une régression logistique est comparée à l'utilisation du Gradient Boosting Machine, méthode basée sur les arbres de régression dans le cadre de la modélisation d'un évènement rare tel que la probabilité qu'un sinistre soit grave.

La seconde étape est de travailler le coût moyen d'un sinistre Incendie. Une première méthode basée sur la Théorie des Valeurs Extrêmes donne un coût moyen non segmenté. Cependant, cette approche ne semble pas satisfaisante à la vue de l'utilisation du modèle que nous souhaitons faire. C'est pourquoi, nous modélisons le coût moyen grave par une GLM par la suite afin d'obtenir un coût moyen par profil de risque. Une prime pure grave est alors obtenue à partir de ces modèles de fréquence et de coût moyen grave.

Cette fréquence et cette prime pure grave servent de clé de répartition pour lisser la sinistralité extrême sur l'ensemble du portefeuille du produit. La clé de répartition choisie entraine des changements conséquents sur nos indicateurs de rentabilité. Nous sommes alors amenés à comparer les impacts de ces modèles et à trancher sur la clé de répartition à utiliser.

I. Présentation de l'entreprise

J'effectue cette alternance au sein de la marque MMA. Celle-ci fait partie du groupe Covéa, regroupant 3 marques : MMA, MAAF et GMF. Je présente succinctement le groupe Covéa dans la section qui suit. Je développe dans une deuxième section les spécificités de la société MMA.

A. Covéa

1. Présentation globale du groupe

Représenté par près de 23 000 collaborateurs sous les trois marques à travers le monde dont 21 000 en France, le portefeuille est composé de plus de 10,7 millions de véhicules et de 8 millions d'habitations assurés permettant au groupe d'être leader sur le secteur de l'assurance aux biens pour un total de près de 11,5 millions de sociétaires. Le groupe est également deuxième sur les assurances pour les entreprises avec plus d'1 million de contrats.

Figure 1 : Primes acquises par marché du Groupe Covéa en 2018

Très implanté en France par son histoire, le groupe Covéa y fait la plus grande partie de son chiffre d'affaires : 16,9 Milliards en France et 2,0 Milliards à l'internationale, principalement au Royaume-Uni et en Italie.

2. Historique

2003	•MAAF Assurance et MMA se rapprochent pour former une SGAM (Société de groupe d'Assurance Mutuelle), nouvel sturcture créée l'an passé dans le droit français créant ainsi le groupe Covéa, sous la présidence de Jean-Claude Seys.
2004	•Covéa Fleet et Covéa Risks sont créés ainsi que Covéa AIS, en charge de la gestion des sinsitres pour toutes les marques et Covéa, en charge de la gestion des risques automobiles aggravés.
2005	•Le groupe AZUR-GMF rejoint Covéa, devenant alors le leader de l'assurance aux biens.
2006	•Des fonctions transverses dans le groupe sont adoptées pour le secretariat général, l'informatique,la R&D et les finances. •MMA et AZUR fussionnent sous le nom de MMA.
2008	•Thierru Derez succède à Jean-Claude Seys à la présidence du groupe.
2009	•Fidélia devient l'assisteur de toutes les marques du groupe.
2010	•Covéa Immobiler, pôle de compétences immobilières du groupe est créé
2011	•APGIS , institut de prévoyance, rejoint le groupe. •Une direction transversale Covéa Santé et Prévoyance est mise en place.
2012	•Deux nouvelles entités sont créées : Covéa Insurance et Covéa Coopérations. Elles ont pour but d'optimiser et de simplifier l'organisation juridique et financière du groupe.
2013	•Une direction commune aux marques du groupe Assurance Covéa est ouverte. •La mutuelle SMI vient renforcer le groupe dans son offre en santé et prévoyance collective en entreprise.
2014	•Le Groupe annonce sa transformation en une seule entreprise unique avec trois marques.
2015	•Des directions communes aux marques sont mises en place. La stratégie est définie au niveau du groupe.
2017	•Un statut commun pour tous les salariés de Covéa quelque soit la marque est signé.

Figure 2 : Historique du Groupe Covéa de 2003 à 2018

3. Les trois marques

Covéa se construit principalement autour de ses trois marques historiques que sont MAAF, MMA et GMF avec chacune une stratégie définie au niveau du groupe.

a) MAAF « La référence Qualité Pro »:

Apparue en 1950 à Niort sous le nom de Mutuelle Assurance Automobile Artisanale de France (MAAAF), elle avait pour ambition de proposer aux artisans une mutualisation des risques sur leurs véhicules professionnels. La mutuelle s'est diversifiée couvrant désormais dans le domaine de la vie, de l'épargne, de la santé et de la prévoyance ainsi que des services financiers. Dans le groupe Covéa depuis sa création en 2003, elle continue de gérer un portefeuille de près de 3,7 millions de sociétaires et clients tous domaines confondus, de plus de 4 millions de véhicules, 2,6 millions d'habitations, le tout articulé autour des trois sociétés MAAF Assurances pour les dommages aux biens, MAAF Vie et MAAF Santé. Dans la stratégie du groupe, MAAF s'oriente principalement vers les particuliers et les professionnels.

b) MMA «Entrepreneurs d'assurances » :

Les Mutuelles du Mans Assurances sont à l'origine un regroupement de plusieurs petites mutuelles de la région du Mans dont la plus ancienne remonte à 1828. Elles se développent jusqu'en 2003 où elles fondent avec MAAF le groupe Covéa. MMA distribue des offres en dommages aux biens, en risques professionnels, en risques d'entreprise, en vie, en épargne en santé prévoyance mais aussi des services financiers. Elle gère alors un portefeuille de près de 3.3 millions de clients, entreprises, professionnels ou bien particuliers. Dans la stratégie du groupe Covéa, MMA est le fer de lance en assurances pour les entreprises et les professionnels.

c) GMF « Assurément Humain » :

Proposant dès 1934 des contrats moins chers pour les agents de l'Etat, la Garantie Mutuelle des Fonctionnaires est représentée actuellement par plus de 3,5 million de sociétaires. Elle propose une gamme de produits large sur les domaines des dommages aux biens, de l'épargne et de la vie, de la Santé et de la prévoyance et aussi des solutions d'assistance et de protection juridique. Elle propose également des offres spécifiques aux agents du Service Public selon la stratégie du Groupe.

Autour de ses trois marques, ils coexistent de multiples entités complétant les marques principales du groupe : AGPIS, institut de prévoyance, SMI proposant des assurances collectives de personnes, Covéa Immobilier dédié à l'optimisation du patrimoine immobilier des trois marques, Covéa Finance en gestion de portefeuille, Fidélia assisteur unique du groupe, ou bien APJ la protection juridique pour MAAF et GMF ainsi que DAS pour la marque MMA.

B. MMA : Entrepreneurs d'assurances

MMA est une assurance multi spécialiste proposant des solutions de risques de dommages aux biens, de prévoyance, de retraite, d'épargne ou d'assurance vie à travers ses trois sociétés MMA IARD, MMA VIE et DAS protection juridique. L'assurance offre également une couverture pour les risques des pros et entreprises, se positionnant principalement sur ce segment de marché.

Figure 3 : Chiffres clés MMA sur l'année 2018

1. Historique

1828	•Le futur Maire du Mans et député de la Sarthe, Louis Basse, fonde la Mutuelle Immobilière du Mans visant à protéger les logements contre les Incendies.
1842	•Louis Basse fonde une sconde société, la Mutuelle Mobilière du Mans.
1883	•Jean-Maris Lelièvre , un avaocat de Mamers cherche à assurer ses dossiers et archives contre l'Incendie. Ne trouvant pas d'assureur, il fonde la Mutuelle générale Française Accident (MGFA).
1884	•Les trois Mutuelles du Mans se rapprochent en faisant administrateur commun.
1917	•La Défense Automobile et Sportive (DAS) est créée par Georges Durand, cofondateur des 24 Heures du Mans, proposant une solution contre les dangers de la circulation routière, et rejoint l'administration commune aux Mutuelles du Mans.
1971	•La DAS rejoint le groupe des Mutuelles du Mans.
1987	•C'est la naissance du Groupe des Mutuelles du Mans Assurances avec les trois entités de la marque: Mutuelle du Mans IARD, Mutuelle du Mans Vie et DAS.
1989	•Les Mutuelles du Mans innovent alors. Elles proposent la première assurance automobile au kilomètre, font des partenariats avec des acteurs de la distribution, de la banque et des caisses de retraite.
1999	•La marque change de nom en 1999 pour devenir MMA.
2003	•MMA fonde avec MAAF le groupe Covéa.
2005	•MMA et AZUR-Assurance fusionnent
2008	•MMA participe à la première opération de naming en France avec le MMArena, le stade du Mans
2015	•MMA change de signature pour "Entrepreneurs d'assurances" accompagnant le virage stratégique de la marque vers le marché des Professionels et Entreprises

Figure 4 : Historique de la marque MMA

2. Distribution

MMA distribue ces produits par deux canaux principaux :

I. Les Agents Généraux :

Plus de 1700 points de ventes tenus par des agents généraux à travers tout le territoire, ils représentent le principal canal de distribution MMA pour tous les types de produits et tous types de clientèle, aussi bien particuliers qu'entreprises. Ils sont répartis sur 10 directions régionales dans toute la métropole.

II. Les Courtiers :

Ils distribuent uniquement les produits IARD de la marque avec près de 400 courtiers partenaires et se concentrent principalement sur le marché des Pros et Entreprises.

C. L'équipe Tarification, Statistiques et Pilotage Habitation, Vie Privée et Prévoyance

L'équipe au sein de laquelle je me trouve est constituée de 5 collaborateurs. Elle opère sur 3 marchés des particuliers distincts :

- L'Habitation avec deux produits majeurs à la vente : une offre habitation pour les logements des particuliers ainsi qu'une offre pour les propriétaires non occupants
- La Vie Privée avec des produits plus atypiques allant de l'assurance cheval à l'assurance bateau
- La prévoyance du particulier avec les produits Garantie Accident de la Vie (GAV), Protection Accident de la Vie (PAV) et Forfait Accident.

Elle a pour mission récurrente de mettre en place les mesures tarifaires à venir ainsi que de contrôler la bonne implémentation des tarifs. C'est dans ce contexte que s'inscrit l'étude présentée ici. En effet, pour prendre les mesures tarifaires adéquates, l'équipe se base sur des indicateurs de rentabilité « économique » et « technique ». L'indicateur « économique » représente la rentabilité de l'année à l'ultime lorsque les sinistres seront clos dans leur totalité. Les indicateurs « techniques » représentent la rentabilité attendue à priori.

Il apparait alors que ces indicateurs se doivent d'être le plus performant possible, notamment sur le produit principal Habitation et sur une de ses garanties majeures : l'Incendie. La charge de l'Incendie étant composée avec une forte proportion de sinistres de forte intensité, on accorde une attention particulière à ces extrêmes. C'est dans ce cadre que se situe la suite de l'étude.

II. Définition du sinistre grave

Afin de pouvoir traiter les sinistres extrêmes, il convient de les définir de façon claire et précise.

Il est choisi de caractériser un grave par un seuil de coût au-delà duquel il devient grave. On prend comme vision du coût, non pas sa première évaluation ou son estimation à l'ultime mais sa dernière évaluation connue. De cette manière, un sinistre en deçà du seuil peut voir son coût réévalué et devenir grave et inversement, un sinistre grave peut se voir réévalué à la baisse et passer en dessous du seuil. Le sinistre est extrême uniquement en fonction de son coût à la date *t*.

La théorie des valeurs extrêmes nous permet d'estimer des variables aléatoires avec une fréquence faible mais une sévérité forte. En plus d'être extrême, cette sévérité est également très hétérogène et peut alors atteindre des valeurs très distantes les unes des autres. J'introduis dans cette partie les résultats fondamentaux issus de cette théorie afin de déterminer le seuil de grave.

Mise en « as if »

Avant de commencer, les coûts sont revalorisés. Cela permet de mieux prendre en compte les coûts de sinistres passés en les réactualisant avec un indice permettant de prendre en compte l'évolution des coûts d'indemnisation. On utilise trois indices fournis trimestriellement qui sont couramment utilisés, notamment à des fins contractuelles :

- L'indice du coût de la Construction FFB (Fédération Française du Bâtiment). Cet indice reflète le coût de la construction d'un immeuble à Paris.
- L'Indice des Prix à la Consommation (IPC)
- L'Indice des Prix à la Consommation hors Tabac

Je construits un indice composé avec pour base 1, la moyenne de l'indice en 2018 avec la pondération suivante :

- 50% ICC FFB
- 25% IPC
- 25% IPC hors Tabac

Cette pondération a été déterminée de façon à avoir un coût moyen qui paraisse le plus stable possible, en relevant à la hausse les sinistres les plus anciens

On obtient ainsi l'évolution d'indice suivante :

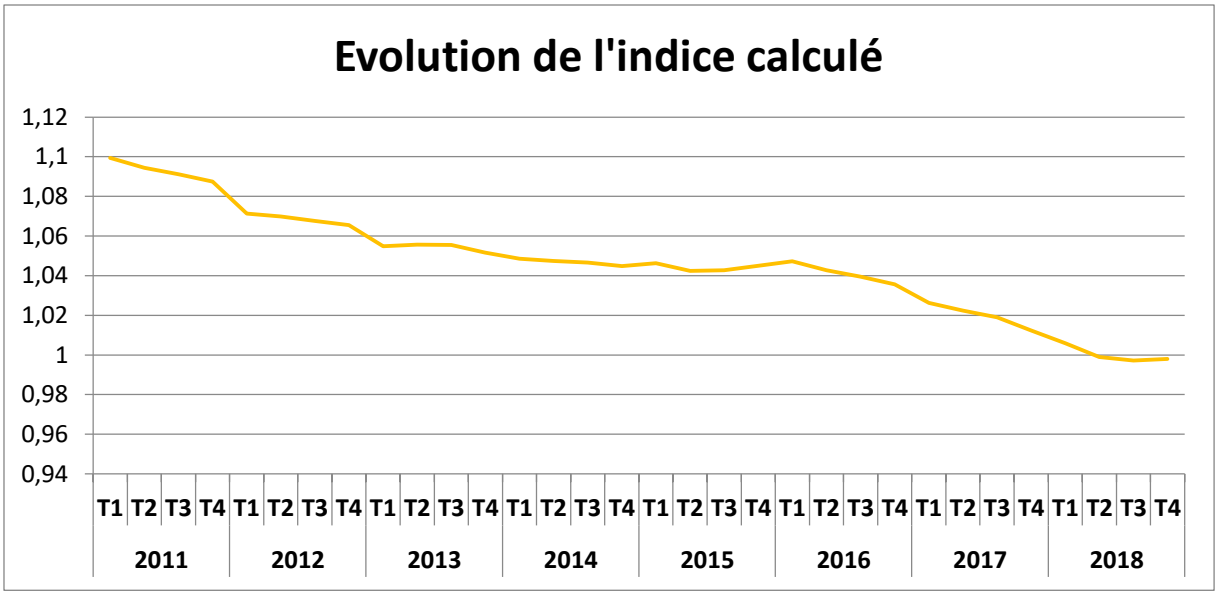


Figure 5 : Evolution de l'évolution de l'indice calculé

On revalorise le coût du sinistre en fonction de sa date de survenance. On peut revaloriser à partir de la dernière date d'évaluation du sinistre mais cela n'a pas été retenu. A titre d'exemple, un sinistre Incendie survenu en mars 2011 est revalorisé à +10% par rapport à son coût observé.

Dans la suite de ce mémoire, le coût d'un sinistre correspond au coût revalorisé de celui-ci.

A. Méthode descriptive

Le choix du seuil peut se faire de manière arbitraire. On peut choisir arbitrairement une définition comme la suivante : « Un sinistre est extrême si son coût est supérieur au 1% des sinistres les plus graves ». Le choix du seuil par cette méthode permet alors de définir un sinistre extrême par sa rareté.

On souhaite cependant qu'un sinistre extrême soit également synonyme d'intensité. Pour cela, j'observe pour différents seuils fixés les quantiles en nombre de sinistres mais également le pourcentage de la charge Incendie qui serait alors considéré comme grave. La première mesure permet de distinguer la rareté de l'évènement alors que la seconde caractérise son intensité.

Le choix du seuil est ensuite déterminé par avis d'expert à partir des deux critères de rareté et d'intensité.

Pour des seuils allant de 10 000 € à 500 000 €, on obtient les résultats suivants :

Seuil	% des sinistres graves	% de la charge >= Seuil
10 000€	90.6%	76.1%
25 000€	94.8%	65.1%
50 000€	96.8%	54.1%
75 000€	97.6%	46.6%
100 000€	97.9%	40.6%
150 000€	98.5%	31.0%
200 000€	98.8%	23.3%
250 000€	99.2%	18.3%
350 000€	99.5%	11.3%
500 000€	99.8%	6.0%

Figure 6 : Décomposition en nombre et en charge de la sinistralité grave pour différents seuils

Pour un seuil inférieur à 50 000€, le nombre de sinistres définis comme graves parait encore trop important. Plus de 3% des sinistres apparaissent comme extrêmes. Cela nous apparait comme une proportion trop importante pour qualifier l'évènement de rare. Ce seuil semble alors trop faible.

Ensuite, pour un seuil supérieur à 250 000€, le caractère « rare » de l'évènement semble respecté. En effet, moins de 1% des sinistres sont alors définis comme grave. Cependant, la notion d'intensité ne semble pas être respectée, une part de surcrête grave inférieure à 20% nous parait trop peu importante.

Les seuils compris entre 75 000€ et 200 000€ semblent répondre aux critères de rareté et d'intensité souhaités. Moins de 2,5% des sinistres sont définis alors comme graves, le caractère rare semble alors respecté. De plus, la surcrête grave représente entre 23.3% et 46.6% de la charge totale Incendie ce qui valide le critère d'intensité recherché.

Nous nous contentons pour le moment de cet intervalle de seuil. Le seuil définitif sera choisi selon les résultats des méthodes graphiques du *mean excess plot* (moyenne des excès) et du *Hill plot* (représentation de l'estimateur de Hill) qui sont abordées dans la prochaine partie.

B. Fonction mean excess

On définit la fonction de *mean excess* du coût tel que :

$$e(x) = E[C - x | C > x]$$

Cela représente l'espérance de la surcrête d'un sinistre pour un seuil de *x* si ce sinistre dépasse le seuil *x.*

La théorie des valeurs extrêmes nous indique que si C suit une loi de Pareto Généralisé $GPD(\beta, \gamma)$, alors la fonction de *mean excess* est de la forme suivante :

$$e(x) = \frac{\beta - \gamma x}{1 - \gamma}$$

On note qu'une contrainte existe : si $\gamma \geq 1$ alors l'espérance n'existe pas.
Pour appliquer utiliser ce résultat, il faut donc impérativement que $\gamma < 1$. Nous verrons à l'aide du Hill Plot que cette hypothèse semble respectée.

On remarque que la fonction de *mean excess* est linéaire en *x*, le seuil. C'est ce comportement linéaire qui nous permettra de détecter un seuil pertinent.

De plus, si $C - x | C > x \sim GPD(\gamma, \beta)$ alors :

$$Pour\ u > x, \qquad C - u | C > u \sim GPD(\beta + \frac{\gamma}{1 - \gamma}(u - x), \gamma)$$

Le comportement linéaire se conserve par translation positive du seuil et ne remet pas en cause l'existence de l'espérance. En effet, la translation n'influe que sur le facteur β.

On cherchera donc à représenter $(x, e(x))$. La vraie valeur de $e(x)$ est cependant inconnue dans la majorité des cas. Cette espérance de l'excès est donc estimée de façon empirique à partir d'un échantillon de sinistre $(C_1, \dots, C_n)$.

$$e_n(x) = \frac{1}{\sum_{i=1}^{n} 1(C_i > x)} \sum_{i=1}^{n} 1(C_i > x) * (C_i - x)$$

On trace alors les points $(x, e_n(x))_{i=1,\dots,n}$ pour remarquer le *x* le plus petit où la courbe devient linéaire.

Application

On trace le *mean excess plot* comme énoncé dans le chapitre précèdent pour des seuils compris en 1 000 € et 500 000€.

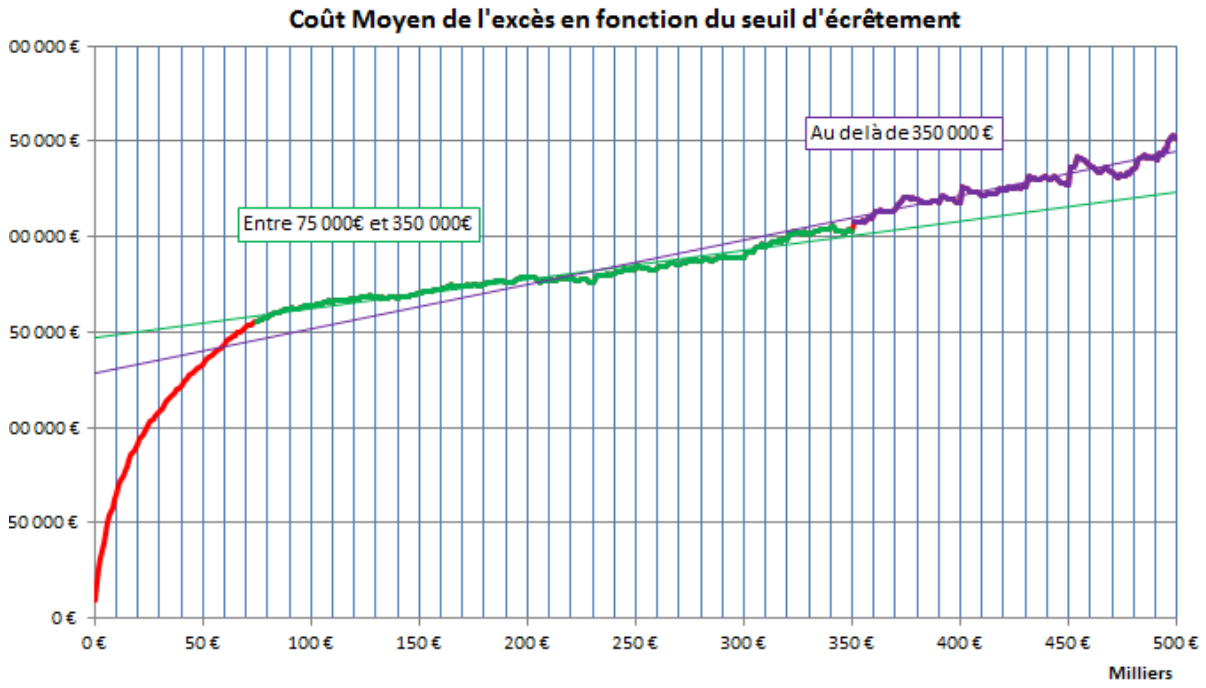


Figure 7 : MeanExcess Plot pour les sinistres incendie

On cherche donc un seuil à partir duquel la courbe semble devenir linéaire. A partir de 75 000€, la courbe est bien linéaire. Un seuil supérieur ou égal est adapté pour qualifier un sinistre Incendie de grave.

Au-delà de 350 000€, la moyenne de l'excès commence à être quelque peu erratique malgré que l'on constate une tendance linéaire. La tendance linéaire semble cependant plus forte que sur les seuils compris entre 75 000€ et 350 000€.

C. Estimateur de Hill

On a remarqué précédemment que le paramètre de queue γ était invariant par translation de seuil. On en déduit une seconde méthode pour choisir un seuil de grave. On représente la courbe (x, γ). Cependant, γ est souvent inconnu. C'est pourquoi on passe par l'estimateur de Hill du paramètre de queue.

$$\gamma'_{n,k} = \frac{1}{k}\sum_{i=1}^{k} \log\left(\frac{C_{(i)}}{C_{(k+1)}}\right) si\ \gamma > 0$$

Ce dernier nous permet de représenter les $(C_{(i)}, \gamma'_{n,k})$ qui approximent le paramètre de queue qui doit donc se stabiliser à partir d'un certains k. Le plus petit seuil à partir duquel le paramètre de queue semble invariant apparait alors comme un bon choix.

Application

On construit donc le *Hill Plot* à partir de l'estimateur décrit précédemment.

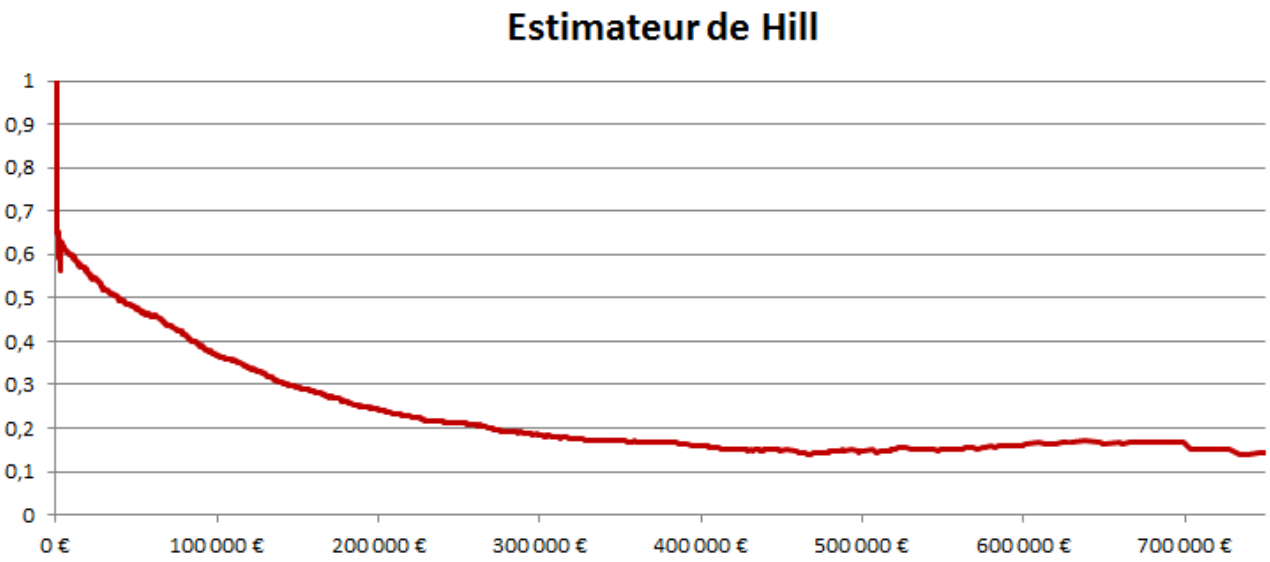


Figure 8 : *Hill plot* pour les sinistres incendie

On constate que γ est estimé <1 ce qui nous permet d'utiliser le *Mean Excess Plot* qui nous imposait cette contrainte.

Contrairement au *mean excess plot*, un coût inférieur à 350 000€ ne semble pas acceptable. En effet, l'estimateur est fortement décroissant jusqu'à 350 000€ et ne se stabilise qu'à partir de ce seuil.

D. Choix du seuil

Anciennement, le seuil de grave Incendie utilisé au sein de l'équipe était de 200 000€. Ce dernier cohabitait avec un seuil de grave utilisé par d'autres équipes qui était de 150 000€. La révision du seuil est une occasion d'harmoniser le traitement de l'Incendie au sein du Groupe Covéa. Ces deux seuils ne semblent pas remis en question au vu de l'approche descriptive et par le *mean excess plot* contrairement au *Hill Plot*.

On constate que si on se base sur le Hill Plot pour faire notre choix, on est alors imposé de choisir un seuil minimum de 350 000€. Cela représente alors seulement 240 sinistres depuis 2011 à 2018. On rappelle que la finalité de l'étude est de modéliser la sinistralité extrême Incendie. Il est donc souhaitable de ne pas avoir un nombre de sinistres trop faible afin d'apporter une robustesse aux modèles construits. On rejette donc ce seuil de 350 000€ qui parait pourtant cohérent d'après le Mean Excess Plot et le Hill Plot mais qui nous compliquerait les modélisations et qui serait contradictoire au seuil de 150 000€ utilisé d'autre part au sein du groupe.

Etant donné que le seuil de 150 000€ est acceptable à partir du Mean Excess Plot, qu'il qualifie correctement la rareté et l'intensité d'un grave, on choisit donc la définition suivante pour un sinistre extrême : **« Un sinistre est qualifié de grave si son coût à la dernière vision est supérieur ou égal à 150 000€ ».**

III. Liquidation au critère

En janvier 2019, une nouvelle version de notre produit MRH a été mise en production. Ce changement de version nous encourage à revoir notre processus de pilotage de la sinistralité et notamment de la charge ultime qui avait été fait jusqu'à présent à une maille grossière. Historiquement, on nous fournit une charge ultime par marché (Habitation, Vie Privée, Prévoyance) et par garantie (Incendie, Responsabilité Civile, Dégâts des eaux, ..). Ces ultimes au marché et à la garantie sont ensuite redistribués par produit à l'aide de clé de répartition par nos soins. L'idée est de revoir ces clés de répartition et de les affiner afin de descendre à une maille plus fine : au produit, à la garantie, au critère de risque. Cette amélioration de notre pilotage de la sinistralité ultime apporte une robustesse dans nos indicateurs de rentabilités. Cette précision nouvelle est donc une aide précieuse dans l'aide à la prise de décision des mesures tarifaires.

Dans cette partie, nous cherchons à vérifier si, en appliquant Chain Ladder pour chaque garantie sur notre produit Habitation principal et notre produit Propriétaire Non Occupant, on observe bien des cadences de règlement différentes. On vérifie également si celles-ci sont différentes d'une population à l'autre. Cela justifie alors notre nouvelle approche de répartition de la charge ultime.

A. Chain Ladder

Pour utiliser la méthode de Chain Ladder, les données doivent être sous la forme de triangle, c'est-à-dire qu'elles doivent être présentées en vision cumulée selon l'année de survenance et également selon l'année de développement, comme le montre le schéma ci-dessous.

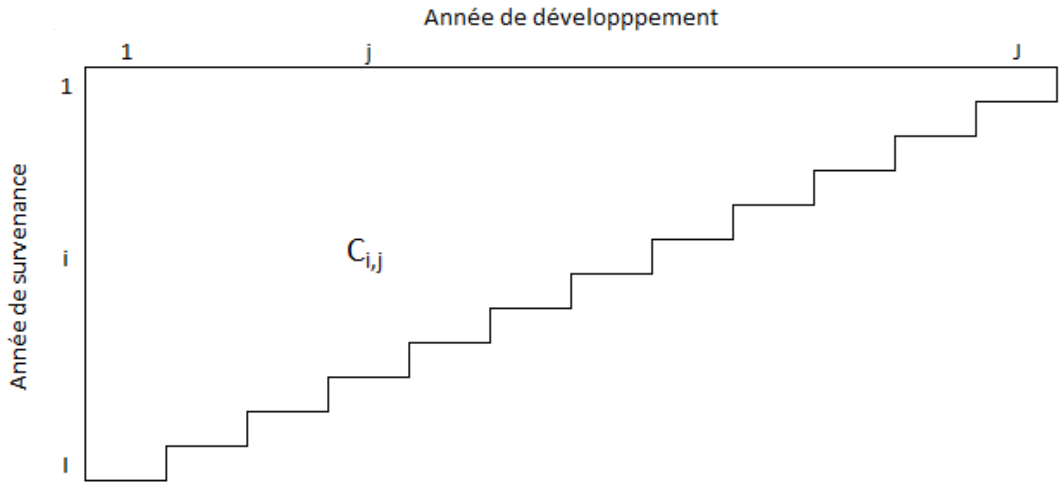


Figure 9 : Triangle de liquidation utiliser pour Chain Ladder

Nous noterons :
- i ε {1, .. , n} les années de survenance (en ligne),
- j ε {1, .. , n} les trimestres de développement (en colonne),
- $C_{i,j}$ les charges cumulée connues, pour 1≤ i ≤ n et j≤ n-i+1.

La méthode de Chain-Ladder repose sur l'existence d'une relation de récurrence :
H_0 : $C_{i,j} = \lambda_j C_{i,j+1}$ pour tous i, j = 1, .. , n.
On définit alors un estimateur de ce coefficient de passage λ_j pour tout j ε {1, .. , n} :

$$\lambda'_j = \frac{\sum_{i=1}^{n-j} C_{i,j+1}}{\sum_{i=1}^{n-j} C_{i,j}}$$

A partir de cet estimateur, il est maintenant possible de compléter le triangle de charges cumulées par des projections de valeurs futures :

$$C'_{i,j} = \lambda'_{n+1-j} .. \lambda'_{j+1} C_{i,n+1-j}$$

B. Application et Interprétation

1. Base d'étude

Il a été nécessaire de récupérer un historique de sinistres le plus conséquent possible. Les sinistres de 2013 à 2018 ont été récupérés avec l'évolution de leur coût trimestriellement. En effet, certaines garanties se liquidant assez vite, il est intéressant de voir une évolution à une maille plus fine qu'une maille à l'année.

Egalement, pour des questions de cohérence avec certaines de nos autres études, il a été choisi de prendre la vision fin mars/ N+1 de la charge de l'année N comme première vision.

Exemple : on prend comme première vision d'un sinistre survenu en 2015 son coût vu à fin mars 2016 puis pour les visions suivantes, son coût vu à fin juin 2016, fin septembre 2016, fin décembre 2016, …, fin mars 2019.

Pour chaque sinistre, on a donc :

- La garantie sinistrée
- Son année de survenance
- Son coût à 03/N+1, 06/N+1,…, 09/N+6 selon l'année de survenance des sinistres
- La typologie (grave/attritionnelle)

Il n'y a pas de segmentation supplémentaire sur le produit Propriétaire Non Occupant. Pour notre produit Habitation, on segmente également sur les deux critères caractérisant le risque sinistré :

- La qualité juridique (Propriétaire/Locataire)
- Le type de bien assuré (Maison/Appartement)

Cela nous permet de travailler notre sinistralité à l'ultime sur nos deux critères principaux. Afin d'avoir des populations suffisamment grandes et relativement homogènes, on décide de ne pas sélectionner plus de critères pour ne pas pénaliser les résultats obtenus par Chain Ladder.

On obtient les coefficients de passage d'une période à l'autre en appliquant la méthode Chain Ladder sur chaque population.

2. Application

a) Par garantie

On applique Chain Ladder dans un premier temps par produit et par garantie sans descendre à une maille plus fine. Cela nous permet de constater des comportements différents de la liquidation au sein d'une même garantie pour un même marché. On prend l'exemple de 2 produits (Produit Habitation et Produit Propriétaire Non Occupant) sur la garantie Incendie.

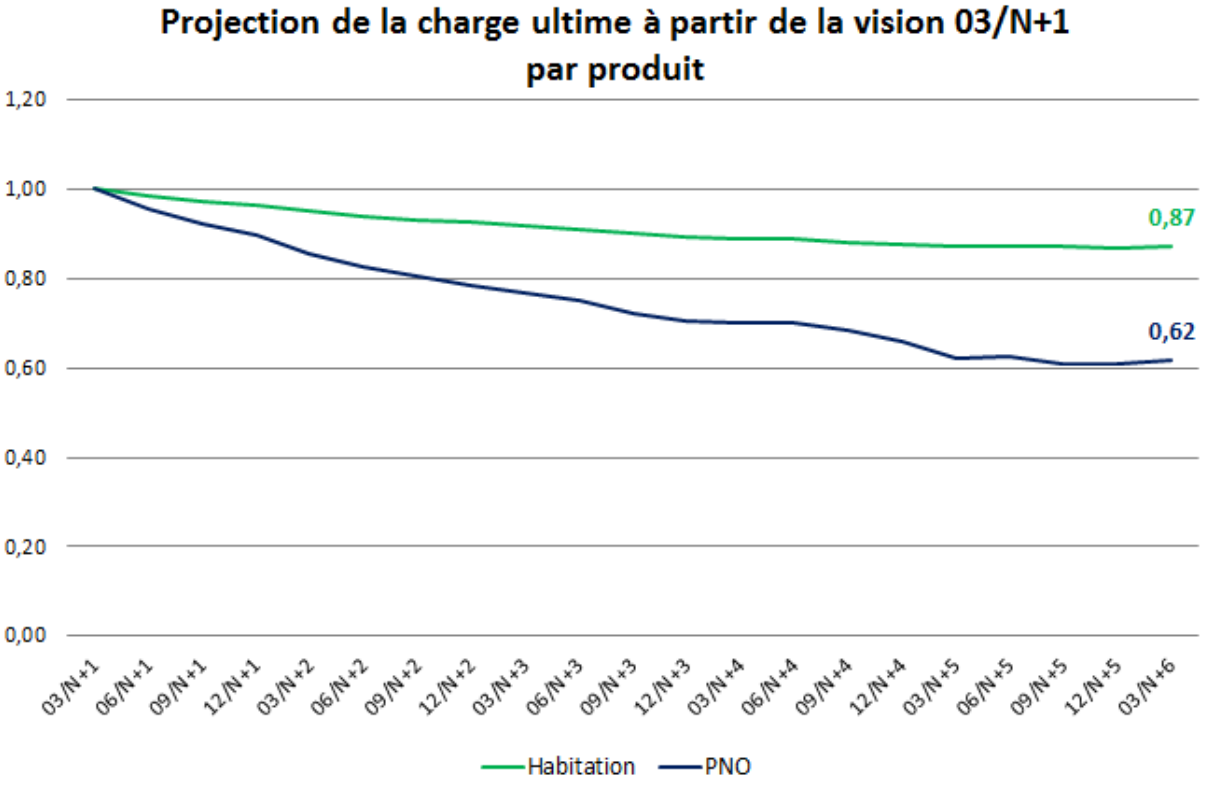


Figure 10 : Evolution des coefficients de liquidation en fonction de l'horizon de projection

A partir de la charge Incendie du produit Habitation observée à fin mars de l'année suivante, il est estimé qu'il faut appliquer un coefficient de 0.87. Autrement dit, il y a un dégonflement de 13% de la charge sinistre soit des bonis de liquidation de 13% entre la charge constatée en mars N+1 et la charge ultime. Sur le produit Propriétaire Non Occupant, on constate également un dégonflement de la charge Incendie mais bien plus fort :-38%. Cet écart de liquidation entre les deux produits est loin d'être négligeable, et caractérise une hétérogénéité au sein d'un même marché pour une garantie donnée. Il peut s'expliquer en partie par des risques couverts différents d'un produit à l'autre mais aussi des règles de gestion sinistres distinctes en fonction du produit.

b) Par risque

De la même manière, on peut observer la charge Incendie selon le type de bien et la qualité juridique. J'applique de nouveau la méthode Chain Ladder pour chaque garantie sur le produit Habitation et sur ces 4 populations :

- Propriétaire de Maison
- Locataire de Maison
- Propriétaire d'Appartement
- Locataire d'Appartement

Pour constater les dégonflements de la charge sur ces 4 populations, on ne représente que le coefficient de passage à l'ultime :

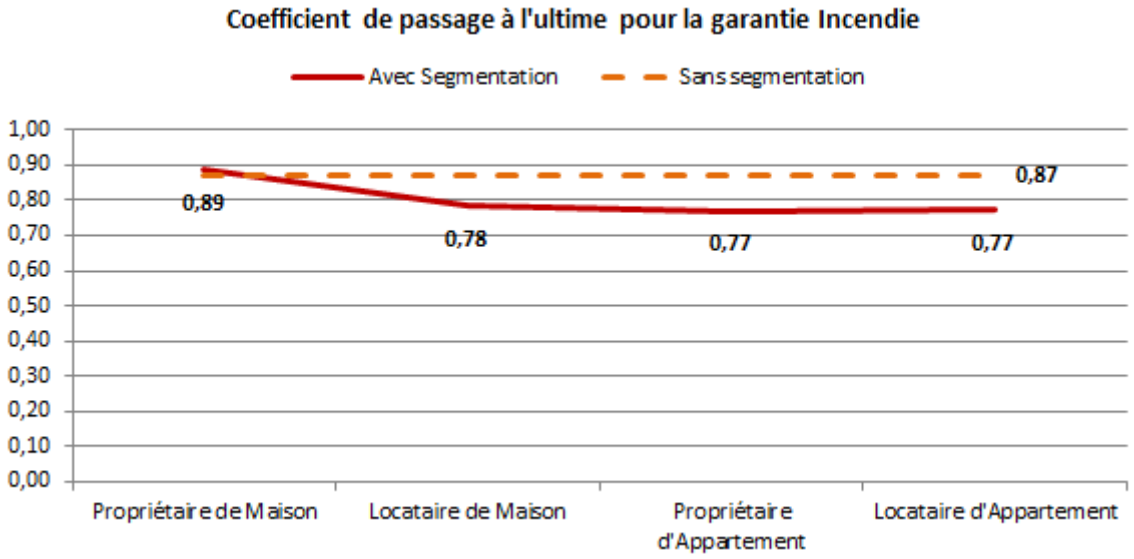


Figure 11 : Coefficient d'ultimisation par population

En orange, on représente le coefficient obtenu précédemment en appliquant la méthode Chain Ladder sur le produit Habitation et la garantie Incendie.

En rouge, on représente le coefficient obtenu en appliquant Chain Ladder par critères de risque.

On en déduit que le coefficient de 0.87 présenté précédemment était fortement porté par le coefficient des Propriétaires de Maison, population qui représentent la majorité de la sinistralité Incendie. Sans ce détail au critère de risque, on aurait surestimé la charge à l'ultime sur les trois autres populations.

On s'assure avec cette nouvelle approche de prendre des mesures tarifaires plus justes en obtenant une charge ultime plus fiable au critère de risque.

c) Limite

On peut reprocher à cette approche à une maille plus fine de mal prendre en compte la sinistralité extrême sur certaines populations. C'est notamment le cas sur la sinistralité Responsabilité Civile : un coût d'un sinistre RC pouvant évoluer brusquement, l'estimation du coefficient à l'ultime sur une relativement petite population peut se voir dégradée par l'évolution forte d'un unique sinistre. La sinistralité CATNAT et plus précisément Sécheresse pose également problème. La déclaration de sinistre étant tardive, la vision de la charge à fin mars de l'année suivant la survenance ne reflète pas correctement la sinistralité potentielle.

Cette approche au critère de risque est retenue pour calculer notre charge ultime par population. Cette charge nous permettra de calculer des indicateurs dits « économiques » qui rentrent dans la prise de décision des mesures tarifaires. Cet aspect « économique » sera détaillé plus tard dans ce mémoire.

IV. Modélisation de la prime pure grave

Dans cette partie, je détaille le modèle de prime pure grave développé. Dans une première partie, on s'attarde sur la méthodologie employée et sur les modèles utilisés. Ensuite, on développe la construction du modèle de fréquence grave puis le coût moyen grave. On peut alors étudier la prime pure grave modélisée.

Cette prime pure nous permet par la suite de redistribuer la sinistralité grave sur l'ensemble de notre produit Habitation. Ce lissage de la sinistralité extrême est une partie de notre processus de calcul de nos indicateurs de rentabilité techniques. Ces indicateurs sont des aides lors de la décision des mesures tarifaires. La méthode employée pour redistribuer la surcrête grave peut donc avoir un impact important. La méthode précédemment employée utilisait une fréquence grave. Cette fréquence ne permettait pas de piloter de manière satisfaisante la sinistralité sur les grands risques. Le produit Habitation est une offre destinée au grand public. Cependant, il est possible de souscrire des risques pouvant aller jusqu'à 30 pièces principales, des biens classés ou inscrits à l'inventaire des monuments historiques ou même encore des châteaux. La présence de ces grands risques apporte une atypie au portefeuille MMA. Une attention particulière est donc portée pour ces biens assurés. L'utilisation d'une prime pure grave permet de mieux les piloter et d'aider la prise de décision lors des choix de mesures tarifaires.

A. Méthodologie

1. Décomposition en sous-modèles

Anciennement, on ne traitait les sinistres extrêmes au sein de MMA qu'à partir de la fréquence. Lors de cette étude, nous avons donc dans un premier temps challenger le modèle de fréquence grave préexistant. Le modèle existant modélisait la fréquence grave Incendie comme une fréquence classique : avec 1 GLM log-Poisson. On décide donc de décomposer la fréquence grave en 2 blocs. Cela nous permet d'introduire plus de segmentation dans le modèle. En effet, le modèle de fréquence fait ressortir 32 variables explicatives. La fréquence grave est donc a minima segmentée sur ces 32 critères significatifs (contre seulement 13 dans la précédente version du modèle de fréquence grave).

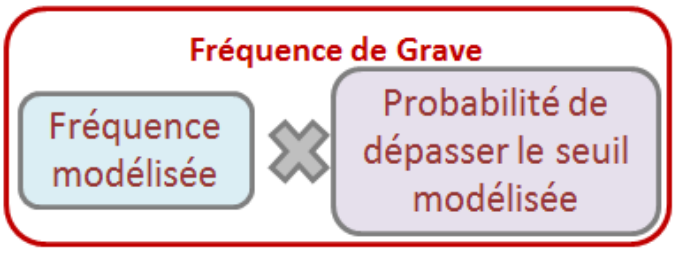


Figure 12 : Décomposition de la fréquence de grave

Cette décomposition est valable sous l'hypothèse classique de l'indépendance de la fréquence et du coût moyen.

On constate cependant que la fréquence ne suffit pas pour expliquer la sinistralité extrême Incendie (cf Partie V). En effet, on observe des disparités fortes sur le coût des graves en fonction du type de bien. Il a donc été décidé d'approfondir l'étude de grave avec l'enrichissement d'un modèle de coût moyen nous permettant d'obtenir ainsi une prime pure.

On se retrouve alors avec trois sous-modèles pour obtenir notre prime pure modélisée :

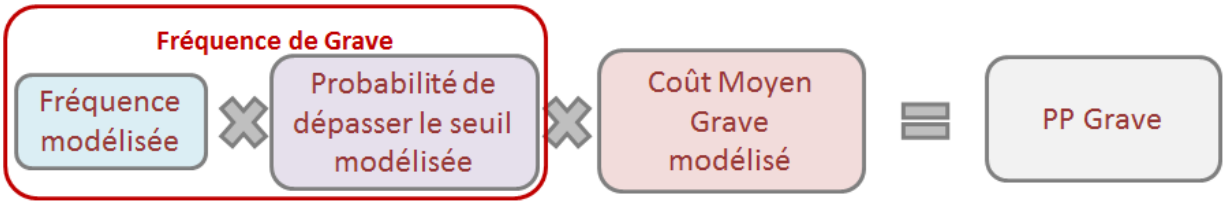


Figure 13 : Décomposition de la prime pure grave

En plus de traiter la partie grave, je traite également la prime pure Incendie globale. La fréquence étant déjà construite, on capitalise sur ce modèle que l'on complète d'un coût moyen global. Cela permettra ensuite d'observer la part de prime pure grave au sein de la prime pure globale Incendie modélisée.

On se retrouve avec le schéma de modélisation suivant :

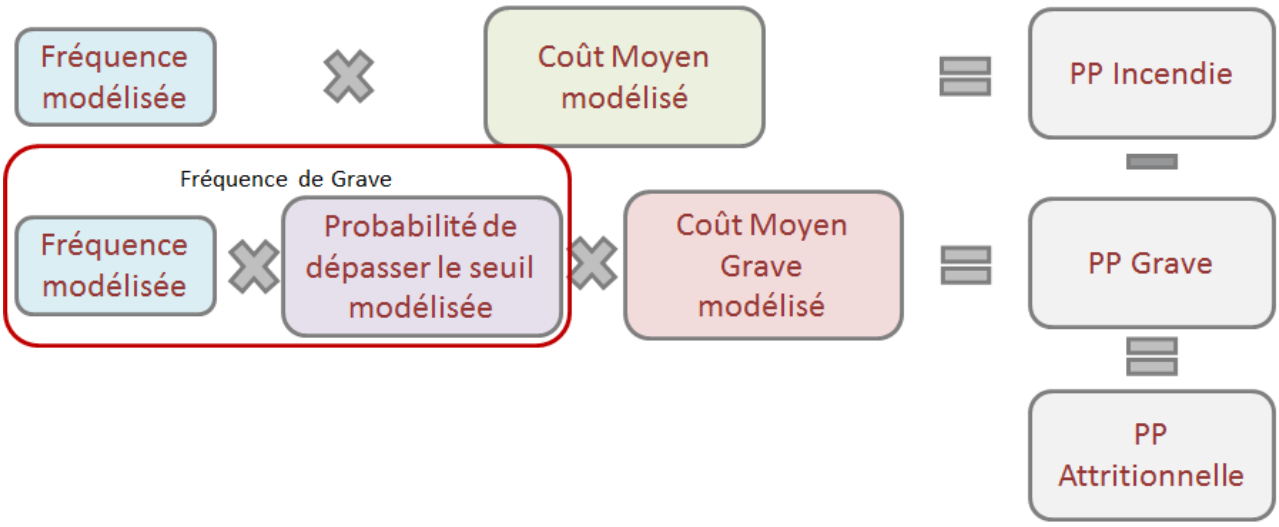


Figure 14 : Décomposition de la prime pure incendie

À partir de la prime pure grave et globale incendie, on peut déduire une prime pure attritionnelle par soustraction. Le modèle de grave n'étant conditionné par le niveau de prime pure globale, on peut se retrouver avec prime pure grave modélisée supérieure à la prime pure globale. La partie attritionnelle serait alors négative. Étant donné que seul le modèle de prime pure grave a un impact concret dans notre processus de revalorisation tarifaire, nous ne considérons pas ce problème comme bloquant et on conserve donc l'approche choisie.

Pour obtenir ces primes pures, je construis donc ces quatre modèles expliqués par plusieurs critères endogènes et exogènes. Pour ce faire, je détaille les deux pratiques utilisées : les GLM et le Gradient Boosting Machine (appliqué uniquement sur le dépassement de seuil).

2. Rappel sur les modèles linéaires généralisés

a) Étude de Loi

On étudie dans un premier temps les lois de distribution des variables modélisées. Les paramètres des lois sont estimés. Pour un modèle de fréquence, les lois testées sont généralement les lois de Poisson ou bien Binomiale Négative. Pour un modèle de coût moyen, on s'attarde sur les lois normales, log-normales, exponentielles, ou gammas. La seule restriction est qu'il s'agisse d'une loi appartenant à la famille exponentielle, c'est-à-dire avec une densité de la forme :

$$f(x,\theta,\varphi) = \exp\left(\frac{x\theta - b(\theta)}{a(\varphi)} + c(y,\varphi)\right)$$

- a(.), b(.) et c(.) sont des fonctions
- θ le paramètre naturel
- φ *un paramètre de nuisance*

b) Sélection des variables explicatives

Nous avons à notre disposition un grand nombre de données que nous pouvons utiliser pour modéliser notre prime pure. Cependant, certaines de ces variables sont fortement corrélées entre elles, il convient de les retraiter. L'utilisation d'un modèle linéaire implique que les variables explicatives ne soient pas corrélées linéairement entre elles. Pour déterminer si des variables sont corrélées entre elles, on se base sur l'utilisation du V de Kramer comme mesure de corrélation.

Si le V de Kramer est trop élevé entre deux variables var1 et var2, on dispose de deux choix :

- Supprimer une des deux variables candidates. Il peut être intéressant alors de conserver la variable la plus corrélée à la variable cible.
- Croiser les deux variables candidates

Le modèle est ainsi décorrélé des effets, une variable n'en influençant pas une autre.

c) Modélisation

On chercher à modéliser la fréquence ou le coût moyen par un Modèle Linéaire Généralisé (GLM).

$$g\big(E(Y|X)\big) = \sum_k \beta_k x_k$$

- Y la variable cible à modéliser
- X les variables explicatives, x_k les modalités des variables
- β_k les estimations des paramètres pour chaque modalité x_k
- g la fonction de lien du modèle

Il convient alors de choisir la fonction de lien adéquate. En fonction de la loi déterminée en amont, il est judicieux de choisir sa fonction de lien canonique, c'est-à-dire une fonction reliant le paramètre naturel θ à l'espérance de la loi μ.

Exemple :

Loi	**Θ(μ)**	**Lien**
Normale (μ, σ²)	μ	Identité
Bernoulli (1, μ)	$log(\frac{\mu}{1-\mu})$	LOGIT
Poisson (μ)	$log(\mu)$	LOG
Gamma (μ, ν)	$-\frac{1}{\mu}$	Inverse

Figure 15 : Loi et fonction de lien canonique

Dans la majorité des cas, il est approprié de choisir la fonction de lien canonique.

Cependant, pour le cas de la modélisation du coût moyen, on ne peut pas choisir la fonction de lien canonique, les valeurs prédites étant strictement positives et le lien inverse ne le garantissant pas. On choisit donc la fonction de lien Log par défaut.

On utilise alors les procédures disponibles sous SAS permettant d'effectuer des GLM. Cette procédure va maximiser la vraisemblance par la méthode de Newton-Raphson. On profite des fonctionnalités du processus de sélection *« stepwise »* proposées par la procédure. Celle-ci nous propose des statistiques type III. Il s'agit d'un test de rapport de

vraisemblance. On considère un modèle avec toutes les variables et on teste le retrait de chacune. On peut ainsi enlever des effets indésirables qu'une variable non significative apporterait.

Exemple :

Résumé des sélections			
Etape	Effet saisi	Nombre d'effets dans	p-value
0	Intercept	1	.
1	ANNEE	2	<.0001
2	RISQUE	3	<.0001
3	CP_MAISONHOTE	4	<.0001
4	NB_PIECE_SUP40	5	0.0001
5	ARTISAN	6	0.0003
6	FORMULE	7	0.0038
7	QUALITE	8	0.0014
8	CP_POURCOMPTE	9	<.0001
9	AGE	10	0.0011

Figure 16 : Sélection stepwise des variables

L'ensemble des p-value est inférieure à 0.05. Toutes les variables peuvent donc être considérées comme significatives. Dans le modèle, la sélection *stepwise* a considéré l'ensemble des variables comme significatives.

Ensuite, pour maximiser la vraisemblance, l'algorithme de Newton-Raphson est utilisé pour estimer chaque coefficient.

SAS accompagne chaque estimation de sa p-value du test de nullité du coefficient. Ainsi, si la p-value est supérieure au seuil 5% fixé, il faut regrouper la modalité avec une autre, de préférence la modalité de référence sauf si cela n'a pas de sens.

Exemple :

Paramètres estimés				
Paramètre	Valeur estimée	Erreur type	Valeur du test t	Pr > |t|
Intercept	-1.4684	0.02524	-58.17	<.0001
Q_PIEC_SUPE d:>3	-0.04685	0.2256	-0.21	0.8355
Q_PIEC_SUPE c:2-3	0.1878	0.05397	3.48	0.0005
Q_PIEC_SUPE b:1	0.04546	0.02261	2.01	0.0444
Q_PIEC_SUPE a:0	0	.	.	.
form 4	0.1433	0.07377	1.94	0.0520
form 1	0.1013	0.07338	1.38	0.1672
form 2	0.07653	0.02936	2.61	0.0092
form 3	0	.	.	.
Q_nb_enfa_2 b.>=1 enf	-0.06555	0.02426	-2.70	0.0069
Q_nb_enfa_2 a.0 enf	0	.	.	.
CD_TYPE_HABI_2 A	-0.1668	0.04228	-3.95	<.0001
CD_TYPE_HABI_2 M-H	0	.	.	.

Paramètres estimés				
Paramètre	Valeur estimée	Erreur type	Valeur du test t	Pr > |t|
Q_Age_2 b.26-40 ans	0.1009	0.03112	3.24	0.0012
Q_Age_2 c.41-50 ans	0.09075	0.02896	3.13	0.0017
Q_Age_2 d.51-60 ans	0.08848	0.02638	3.35	0.0008
Q_Age_2 a.<=25 et >=61	0	.	.	.
BC_C018 1	0.2745	0.1002	2.74	0.0062
BC_C018 0	0	.	.	.
BC_C106 1	0.3344	0.1032	3.24	0.0012
BC_C106 0	0	.	.	.
Q_CS2_2 a:0-3	0.09497	0.03120	3.04	0.0023
Q_CS2_2 c:>=10	0.08217	0.02170	3.79	0.0002
Q_CS2_2 b:03-09	0	.	.	.
CD_QLTE_ASSU_HABI L	-0.1843	0.04307	-4.28	<.0001
CD_QLTE_ASSU_HABI P	0	.	.	.
Anc_cont	-0.00432	0.001622	-2.66	0.0078

Figure 17 : Estimation des paramètres du modèle

On constate que la p-value pour la modalité **1** et **4** de la variable **form** (représente les formules de montée en gamme) est >0.05. Il faut donc les regrouper avec d'autres modalités. La modalité référence étant sur la modalité **3**, il faudrait, si l'on suit la logique du test de nullité du coefficient, accepter l'hypothèse de coefficient nul et donc regrouper avec la modalité **3**. Cependant, dans un souci de cohérence, il est préférable de les regrouper de façon logique, c'est-à-dire non pas regrouper **1, 3 et 4** mais le regroupement **1-2** et **3-4**. En effet, le regroupement 1-3 est incohérent.

De la même manière, la modalité **« d :>3 »** de la variable **Q_PIEC_SUPE** (représente le nombre de pièces supérieur à 40m²) ne sera pas regroupée avec la modalité de référence **« a : 0 »** mais avec la modalité **« c :2-3 ».** En effet, cela semble plus logique de regrouper les habitations ayant plus de 3 pièces supérieures à 40m² avec celles en ayant 2 plutôt qu'avec celles n'en ayant aucune.

Il convient de répéter ce processus jusqu'à ce que toutes les variables aient un effet significatif et également pour toutes les modalités.

3. Gradient Boosting Machine

On cherche à construire un modèle M tel que :

$$Y = M(X) + e$$

- Y, la variable à expliquer
- X, les variables explicatives
- M, le modèle
- e, l'erreur du modèle

Il s'agit d'une méthode de Machine Learning tout comme les Random Forest qui sont plus populaires. Elles permettent de prédire dans des cas de régression ou bien de classification.

Le Gradient Boosting conjugue deux méthodes différentes :

- La Descente de Gradient
- Le Boosting

a) Descente de Gradient

C'est une méthode itérative pour approcher la solution d'un problème d'optimisation.

-Problème d'optimisation : $\min_x f(x)$

-x0, point de départ

-e, le seuil de précision

Etape i :

-Calcul du gradient de f en x_i, $\nabla f(x_i)$

-Si $\nabla f(x_i) < e$, alors on arrête

-Sinon, $x_{i+1} = x_i - \propto_i * \nabla f(x_i)$, il existe différentes méthodes pour choisir le $\propto_i$, on le choisit constant

b) Agrégation par Boosting

C'est une méthode d'agrégation de modèle, que l'on appelle alors *Weak Learner* ou *Apprenant faible*. C'est une des différences avec les Random Forest qui utilisent une méthode d'agrégation appelé Bagging.

	Modèle Construit à l'Etape i	
Etape i	**Boosting**	**Bagging**
1	$Y = M_1(X) + e_1$	$Y = M_1(X) + e_1$
2	$e_1 = M_2(X) + e_2$	$Y = M_2(X) + e_2$
…	…	…
I	$e_{i-1} = M_i(X) + e_i$	$Y = M_i(X) + e_i$
…	…	…
N	$e_{N-1} = M_N(X) + e_N$	$Y = M_N(X) + e_N$
Modèle finale :	$\sum_{i=1}^{N} w_i * M_i$ Les w_i correspondent à une pondération des modèles selon leur qualité prédictive	$\frac{1}{N}\sum_{i=1}^{N} M_i$ Il s'agit de la moyenne des modèles construits

Le Bagging consiste à construire des modèles en parallèle, contrairement au Boosting qui en construit en série.

c) Gradient Boosting Machine

Dans le cas du Gradient Boosting, les *Weak Learner* sont des arbres de décision. On cherche alors à optimiser une fonction de perte évaluant la pertinence du modèle.

- Soit M, un modèle
- Soit j, la fonction de perte pour une observation, $j(y_i, M(x_i))$
- Soit J, la fonction de perte globale, $J(Y, M) = \sum_i^N j(y_i, M(x_i))$
- Problème : Minimiser $J(Y, M)$ par rapport à M
- M_0, le modèle initial
- e, seuil de tolérance

<u>Étape k intermédiaire dans le cas général :</u>

- Calcul du gradient de J, $\nabla J(Y, M_k) = (\frac{\partial J(Y, M_k)}{\partial M_k(x_i)})_i$
- Si $|\nabla \mathrm{J}(\mathrm{Y}, M_k)| < e$, on arrête
- Sinon, $M_{k+1}(x_i) = M_k(x_i) - \propto_i * \nabla j(y_i, M_k(x_i))$

En prenant comme fonction de perte j définie comme suit :

$$j(y_i, M(x_i)) = \frac{1}{2}(y_i - M(x_i))^2$$

On a donc comme fonction de perte globale :

$$J(Y, M) = \frac{1}{2}\sum_{i=1}^{N}(y_i - M(x_i))^2$$

<u>Étape k intermédiaire avec la fonction de perte quadratique :</u>

- $\nabla J(Y, M_k) = (\frac{\partial J(Y, M_k)}{\partial M_k(x_i)})_i = (-(y_i - M_k(x_i)))_i = (e_{k,i})_i$, le gradient correspond donc à l'erreur du modèle M_k
- On aimerait alors $M_{k+1} = M_k + \propto_k * e_k$
- On va alors construire un nouveau W*eak Learner* m_k tel que $e_k = m_k(X) + e$ pour estimer l'erreur du modèle précédent.

- $M_{k+1} = M_k + \propto * m_k$, le nouveau modèle à l'étape k, alpha est appelé vitesse d'apprentissage

Chaque nouveau modèle intermédiaire vient palier les lacunes du modèle précédent.

On utilise alors le logiciel R afin de modéliser la probabilité de dépasser le seuil de 150 000€. Il faut alors considérer plusieurs paramètres :

- Nombre d'arbres *Weak Learner*
- Profondeur des arbres
- Vitesse d'apprentissage
- % de la base d'apprentissage pour chaque arbre

Il est important de bien les choisir, un mauvais paramétrage pouvant amener à un sur apprentissage.

B. Fréquence grave Incendie

La prochaine étape dans la construction de notre modèle est le modèle de fréquence grave Incendie. Celui-ci vient se multiplier dans un second temps avec le modèle de dépassement de seuil pour obtenir notre modèle complet.

Dans un premier temps, je détaille le modèle de fréquence Incendie

Pour modéliser le taux de dépassement de seuil, je propose deux méthodes différentes. Une première classique en statistique avec une régression logistique puis une seconde plus innovante avec un Gradient Boosting Machine. Je compare les résultats obtenus pour déterminer laquelle est la plus adaptée.

1. Modèle de Fréquence Incendie

a) Base d'étude

L'étude se fait à partir d'un historique 2011-2018 sur l'ensemble des observations du produit 410 souscrit en agence. Pour chaque observation de la base d'étude, on distingue les données suivantes :

- <u>Données Clients :</u>
  - Age
  - Nombre d'enfants
  - Nombre de personnes dans le foyer
  - Nombre de contrats MMA
  - Client MRH, Client Epargne, Client Auto, …
  - Montant total des cotisations
- <u>Données Habitation :</u>
  - Type de bien
  - Qualité juridique
  - Résidence principale ou secondaire
  - …
- <u>Données Géographique :</u>
  - Données INSEE à l'iris
- <u>Données Contrats :</u>
  - Formule souscrite
  - Renfort
  - Ancienneté du contrat
  - Conditions Particulières
  - …

Ces données explicatives viennent compléter le nombre de sinistres survenus sur l'observation et l'exposition en risque-année (1 année d'exposition = 1 RA, 6 mois d'exposition = 0.5 RA) de cette exposition.

Ceci représente une exposition totale de 9 626 221 risques-année pour un total de 49 511 sinistres Incendie survenus, soit une fréquence Incendie de 0.51%.

Périmètre de la garantie Incendie

Il est nécessaire de préciser que le périmètre des sinistres Incendie a été modifié à deux reprises sur cette période :

- Une période avec uniquement de l'Incendie
- Une période avec l'Incendie et une partie du dommage électrique immobilier
- Une dernière période avec l'inclusion complète du dommage électrique

Pour traiter ce changement de périmètre, il a été ajouté comme donnée le prorata de l'observation passée dans chaque situation. Cela repose sur une hypothèse forte : l'ajout du dommage électrique immobilier à la même influence proportionnelle quel que soit le profil de risque. Cette solution a tout de même été retenue afin de conserver un historique le plus profond possible. De plus, le périmètre de l'Incendie de l'ancienne offre n'étant pas identique à celui de la nouvelle (pour laquelle nous n'avons pas suffisamment de données, seulement 6 mois depuis la mise en production), cette méthode permettra de valoriser l'absence de dommage électrique immobilier de la nouvelle offre.

Données INSEE

On déplore l'absence de zonier Incendie, le zonier actuellement en place a été réalisé il y a maintenant plusieurs années et ne segmente plus suffisamment le risque Incendie. Il a donc été décidé de ne pas s'en servir et de s'appuyer sur des données INSEE à l'iris pour expliquer le risque géographique.

Ces variables à l'iris, maille la plus fine à laquelle MMA travaille, peuvent être classées en plusieurs thèmes :

- L'état du logement (nombre de logements, de résidences principales, de maisons, de pièces,…)
- L'ancienneté de l'habitat
- La situation socio-économique de l'iris
- La situation démographique de l'iris

Ces variables sont alors traitées afin d'obtenir des proportions plutôt que des nombres bruts, on se sépare ainsi de quelques corrélations évidentes. Un iris avec un grand nombre de logements aura très probablement un grand nombre de résidences principales par exemple. Ces parts ont ensuite été qualifiées de manière à obtenir la distribution suivante :

- Modalité a : Valeur inférieure au quantile 5%
- Modalité b : Valeur comprise entre le quantile 5% et le quantile 10%
- Modalité c : Valeur comprise entre le quantile 10% et le quantile 25%
- Modalité d : Valeur comprise entre le quantile 25% et le quantile 50%
- Modalité e : Valeur comprise entre le quantile 50% et le quantile 75%
- Modalité f : Valeur comprise entre le quantile 75% et le quantile 90%
- Modalité g : Valeur comprise entre le quantile 90% et le quantile 95%
- Modalité h : Valeur supérieure au quantile 95%

On analyse ensuite le V de Kramer pour chaque couple de variables quantitatives. On traite les couples avec un coefficient supérieur à 0,6, caractéristique d'une forte corrélation.

Suite à ce traitement et après sélection dans les différents modèles présentés plus tard, on retient les variables suivantes à l'iris :

Nombre moyenne de pièces sur les maisons	**% de cadre**
% de logements de plus de 5pièces	**% de professions intermédiaires**
% de logements avec chauffage individuel	**% de retraités**
% de maisons	**% de logements construits entre 1920 et 1945**
% de couples avec enfant	**% de logements construits avant 1919**
% de couples sans enfant	**% de logements de moins de 30m2**
% de personnes âgées de 30 à 44 ans	**% de logements en location**
% de personnes âgées de 75 ans ou plus	**% de Maisons construites après 1991**
% d'agriculteurs	**% de Maisons construites avant 1919**
% d'artisans	**% de logements supérieurs à 100m2**

Figure 18 : Variables INSEE explicatives retenues pour expliquer la prime pure grave

Des cartes représentants certaines de ces données à l'échelle de la France sont disponibles en annexe de ce mémoire.

b) Conclusion

Pour modéliser la fréquence Incendie, j'ai choisi une approche dite « stepwise » pour sélectionner mes variables explicatives.

Cela a permis de présélectionner un total de 42 variables avec un effet significatif. Cela parait trop important pour être réaliste, j'ai donc choisi de stopper la sélection à l'étape présentant un BIC minimum sur l'échantillon de validation. L'intérêt de ce critère vis-à-vis de l'AIC est de prendre en compte la complexité du modèle.

On retient les variables suivantes pour expliquer la fréquence incendie :

Variables explicatives de la fréquence incendie	
Endogènes	***Exogènes à l'iris (Source : INSEE)***
Age de l'assuré (~)	*Moyenne de pièces principales sur les maisons(+)*
Ancienneté du contrat (-)	*% de maisons(+)*
Année (-)	*% de couples sans enfant (~)*
Maison en construction (-)	*% de personnes âgées de 30 à 44 ans(-)*
Assurance pour compte du propriétaire (+)	*% de cadres(-)*
Isolement du bien(+)	*% de professions intermédiaires(-)*
Capital mobilier assuré (~)	*% de logement de moins de 30m²(-)*
Résidence principale ou secondaire	*% de maisons construites après 1991*
Catégorie socio-professionnelle de l'assuré	*% de maisons construites avant 1919(-)*
Formule souscrite	*% de logement de plus de 100m²(+)*
Nombre d'enfants dans le foyer de l'assuré (+)	*% de logements de 5 pièces ou plus(-)*
Nombre de pièces principales (+)	*% de logement avec chauffage individuel(~)*
Nombre de pièces supérieures à 40m² (+)	
Dépendance à une autre adresse (+)	
Périmètre Garantie	
Qualité juridique	
Superficie de la dépendance (+)	
Type de bien	
+ : effet à la hausse	*- : effet à la baisse*

Figure 19 : Tableau récapitulatif des variables explicatives de la fréquence

Pour analyser les estimations du modèle de fréquence, on représente trois quantités :

- L'exposition de 2011 à 2018 par critère
- L'effet décorrélé « toutes choses égales par ailleurs » qui correspond aux rapports de côte, c'est-à-dire l'exponentielle du paramètre estimé.
- L'effet observé, c'est-à-dire, le rapport de la fréquence observée de la modalité sur la fréquence observée de la modalité de référence

(1) Par risque

On commence dans un premier temps par analyser les résultats en fonction de la qualité juridique de l'assuré (Locataire/ Propriétaire).

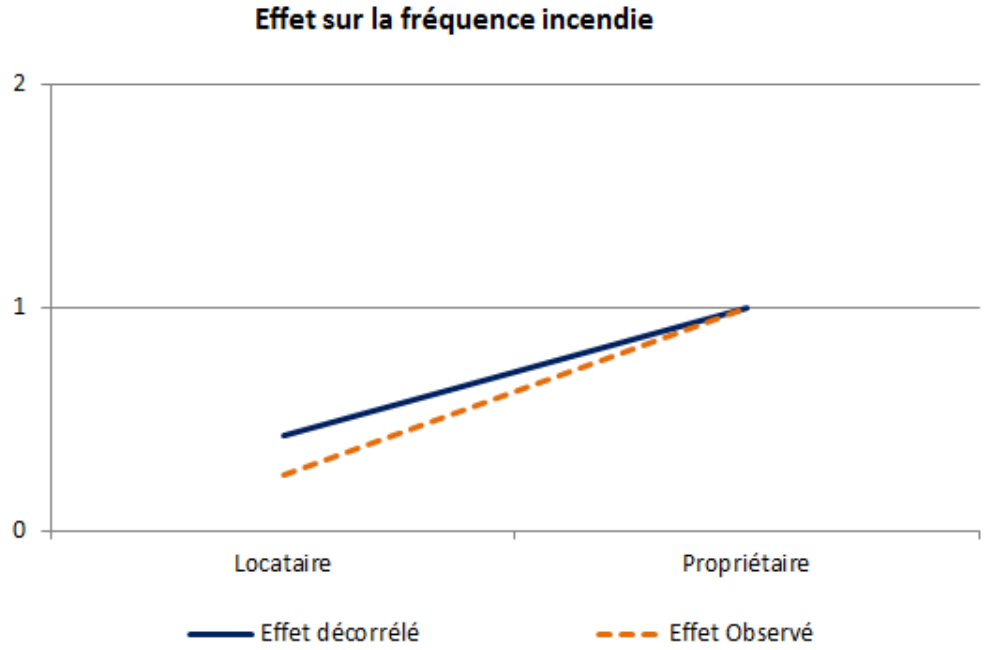


Figure 20 : *Odds Ratio* en fonction de la qualité juridique sur la fréquence

D'après le modèle, le seul fait d'être locataire de son habitation, à risque équivalent, décroit le risque de survenance d'un sinistre incendie de presque 50%. Cela correspond à ce que l'on observe sur la courbe bleue. Cependant, on constate grâce à la courbe orange que la fréquence incendie réelle sur les locataires est environ quatre fois inférieure à celle des propriétaires.

L'écart entre cet effet modélisé et cet effet observé correspond à l'information apportée par d'autres critères dans le modèle, ou bien non expliquée. Dans ce cas, cet écart s'explique en grande partie par le nombre de pièces et le type de bien.

Pour le visualiser, on représente les effets en fonction de ces 2 critères. La référence se situe sur les maisons de 4 pièces.

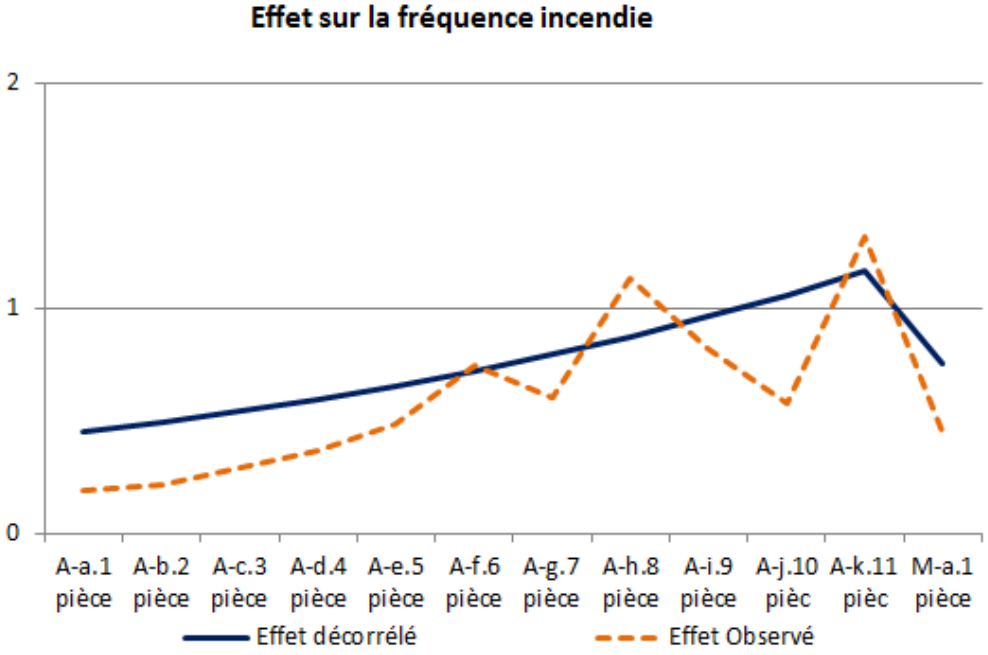


Figure 21 : *Odds Ratio* en fonction du nombre de pièces pour un appartement sur la fréquence

La fréquence croit en fonction du nombre de pièces, ce qui est conforme à ce qui est pressenti. 95% des appartements ont moins de 4 pièces. Pour la quasi-totalité des appartements, on a donc un risque diminué, toute chose égale par ailleurs, compris entre -40% pour les 4 pièces et -55% pour les 1 pièces par rapport à une maison de 4 pièces. En multipliant les effets locataire (~-50%) et appartement (entre -55% et -40%), on se retrouve proche de la fréquence observée avec la courbe orange (fig21).

(2) Aspect géographique

Pour expliquer le risque géographique, il a été décidé d'utiliser des données exogènes sans tenir compte d'un zonier Incendie. L'INSEE met à disposition des données à l'iris. On a donc utilisé ces dernières pour expliquer l'aspect géographique. Pour visualiser l'effet sur la fréquence Incendie de ces données spatial, on applique pour chaque iris uniquement les effets décorrélés des variables INSEE.

Figure 22 : Représentation des effets des variables INSEE sur la fréquence

Les côtes méditerranéennes et atlantiques, l'Aquitaine ainsi que l'Alsace apparaissent alors comme des zones avec risque Incendie plus élevé. Inversement, le Nord-Pas de Calais semble moins à risque.

L'utilisation de ces données exogènes semble prometteuse. Lors de la construction du futur zonier Incendie, il semble opportun de se pencher sur cette approche. Ici, la carte représente le risque spatial pour les maisons et appartements confondus. Il est sûrement nécessaire de faire un distinguo pour un zonier.

(3) Aspect temporel

On suppose que ces données INSEE et les données endogènes permettent d'expliquer la quasi-totalité de la fréquence Incendie. On souhaite tout de même connaitre une information supplémentaire : une tendance temporelle. Pour traiter cela, une variable continue traduisant l'année de survenance est introduit dans le modèle. On suppose alors que le paramètre estimé sur cette variable traduit une tendance de fond de la fréquence modélisée. En effet, une grande partie des effets est sensée avoir été neutralisée. La variable temporelle doit traduire cette évolution dans le temps.

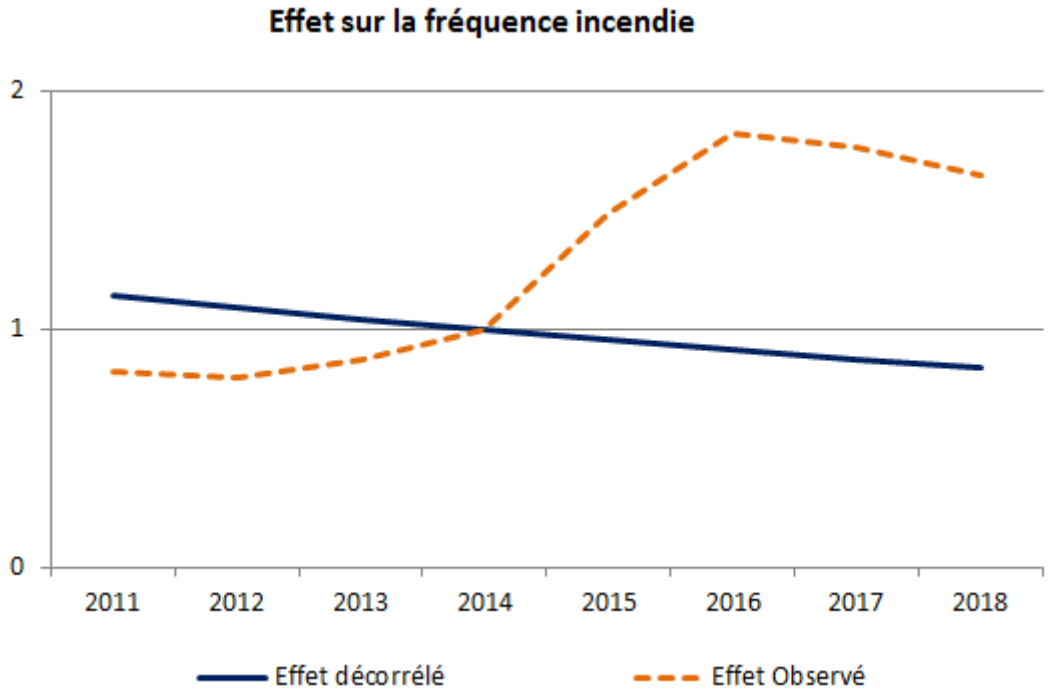


Figure 23 : *Odds Ratio* en fonction de l'année d'exercice sur la fréquence

On note tout d'abord que l'effet observé ici prend en compte le changement de périmètre de garantie Incendie avec l'ajout progressif du dommage électrique immobilier. Le périmètre étant plus étendu, il est normal de constater une hausse de la fréquence. Ce changement de périmètre est neutralisé par l'utilisation d'une donnée indiquant le périmètre de la garantie pour chaque observation. L'effet décorrélé représenté sur le graphique ci-dessus ne prend ainsi en compte que l'effet tendanciel. Pour ce faire, je réalise une sélection « stepwise » avec uniquement les variables temporelles brutes ou

transformées. La quantité temporelle qui apparait comme la plus adaptée au modèle de fréquence est l'année de survenance brute.

Cela a permis de modéliser une tendance à la baisse pour la fréquence Incendie avec une forte diminution annuelle d'environ -4%.

On représente ci-dessous la fréquence observée et la fréquence modélisée sur l'ensemble des observations pour s'assurer que le modèle est robuste.

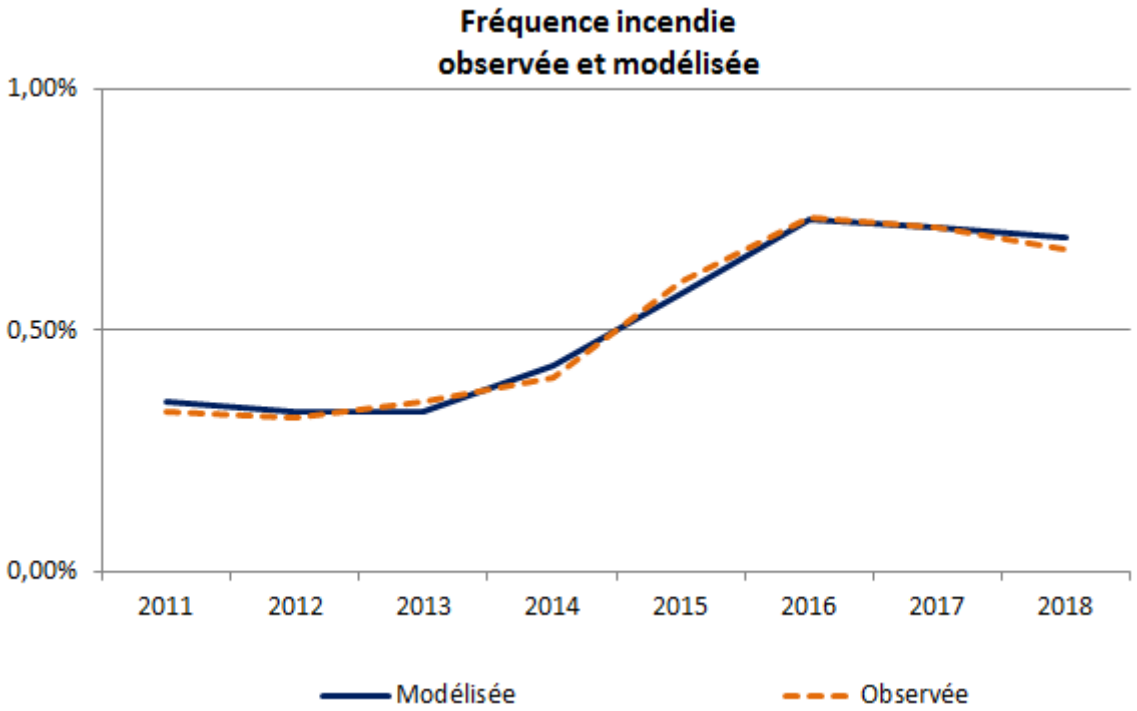


Figure 24 : Fréquence Incendie Observée et Modélisée par année d'exercice

Afin de finaliser la fréquence grave Incendie, on s'attarde sur la construction du modèle de dépassement de seuil qui complète le modèle de fréquence qui vient d'être présenté.

2. Modèle de dépassement de seuil

Dans cette partie, on modélise la probabilité que le coût d'un sinistre dépasse le seuil de 150 000€. La construction de ce modèle se caractérise par le faible nombre de sinistres graves observés et cela malgré une période d'observation plus importante. On en profite alors pour challenger les GLM utilisés habituellement avec un Gradient Boosting Machine. Pour comparer les deux modèles, on prend en compte leur capacité de prédiction avec l'AUC et la complexité de mise en œuvre.

On détaille d'abord la base d'étude. Ensuite, on rappelle succinctement les principes de la régression logistique ainsi que du Gradient Boosting Machine. Enfin, on confronte les deux modélisations pour choisir la plus efficace avant d'analyser le modèle sélectionné.

a) Base d'étude

Notre base d'étude se constitue de l'ensemble des sinistres Incendie de 2011 à 2018. Les données explicatives à disposition sont identiques à celles de la partie précédente. On a donc à disposition 42 087 sinistres Incendie dans la base d'apprentissage et 7 425 dans la base de test. Parmi ces 49 512 sinistres, seulement 757 sont supérieurs à 150 000€ soit 1,5%.

La base est décomposée en 3 échantillons :

- 30% pour l'échantillon de test
- 70% pour les échantillons d'apprentissage et de validation
    - 70% de ces 70% soit 49% de la base totale pour l'apprentissage
    - 30% de ces 70% soit 21% de la base totale pour la validation

b) Modélisation par Régression logistique

Comme pour la modélisation de la fréquence, on réutilise une régression linéaire généralisée et plus particulièrement une régression logistique.

La variable cible étant binaire, 1 pour un sinistre grave et 0 sinon, on utilisera une fonction de lien de type logit, fonction de lien canonique d'une loi Bernoulli.

$$g(p) = \sum_{k}^{M} \beta_k x_k$$

- p, la probabilité prédite
- X les variables explicatives, x_k les modalités des variables
- β_k les estimations des paramètres pour chaque modalité x_k
- g la fonction de lien LOGIT

Ainsi $g(p) = \log(\frac{p}{1-p})$ et donc $p = \frac{e^{\sum_k^M \beta_k x_k}}{1+e^{\sum_k^M \beta_k x_k}}$

Le modèle impose alors à p d'être compris en 0 et 1.

Le retraitement des variables et des modalités se fait de la même manière.

c) Modélisation par Gradient Boosting Machine

On rappelle que l'échantillon d'apprentissage est coupé en 2 échantillons distincts pour le GBM : 70% pour l'apprentissage et 30% pour la validation. On choisit l'AUC calculé sur l'échantillon de validation comme critère de choix.

On compare la précision du Gradient Boosting et de la régression logistique grâce à la courbe ROC. Elle représente la spécificité en fonction de la sensibilité.

(1) Courbe ROC et AUC

Pour $\propto$ entre 0 et 1.

Si $p > \propto$ alors on prédit un sinistre grave. On construit la matrice de confusion suivante :

	Grave	**Non grave**	VP=Vrai Positif
Grave prédit	**VP**	**FP**	FP=Faux Positif
Grave non prédit	**FN**	**VN**	FN=Faux Positif VN=Vrai Négatif

Figure 25 : Matrice de concordance

On définit la sensibilité et la spécificité de la manière suivante :

$$Sensibilité = \frac{VP}{VP + FN}$$

$$Spécificité = \frac{VN}{VN + FP}$$

La sensibilité traduit la probabilité de prédire un grave parmi les graves réels. Inversement, la spécificité correspond à la probabilité de ne pas prédire de grave parmi les sinistres attritionnels.

Il est alors possible de tracer la courbe ROC représentant la sensibilité en fonction de la spécificité pour différents seuils α. On superpose alors la première bissectrice correspondant à la courbe ROC d'un modèle qui prédit de manière aléatoire et uniforme un sinistre grave. On évalue alors la qualité du modèle par son aire sous la courbe ROC (*Area Under the Curve*) AUC.

Plus l'AUC est proche de 1, plus le modèle catégorise correctement. Un AUC de 1 correspond à une classification sans erreur.

Un AUC de 0.5 correspond à une absence de modèle. Une valeur proche de 0,5 n'est donc pas intéressante.

Dans la suite, on détaille le choix des paramètres du GBM à savoir, nombre d'arbres construits, vitesse d'apprentissage et profondeur d'interaction (nombre de fois qu'une variable peut apparaitre dans l'arbre).

(2) Choix des paramètres

On commence par choisir un nombre d'arbres pour les choix du alpha (vitesse d'apprentissage) et n (profondeur d'interaction). Par défaut, la vitesse d'apprentissage est à 0,01 et la profondeur d'interaction à 1.

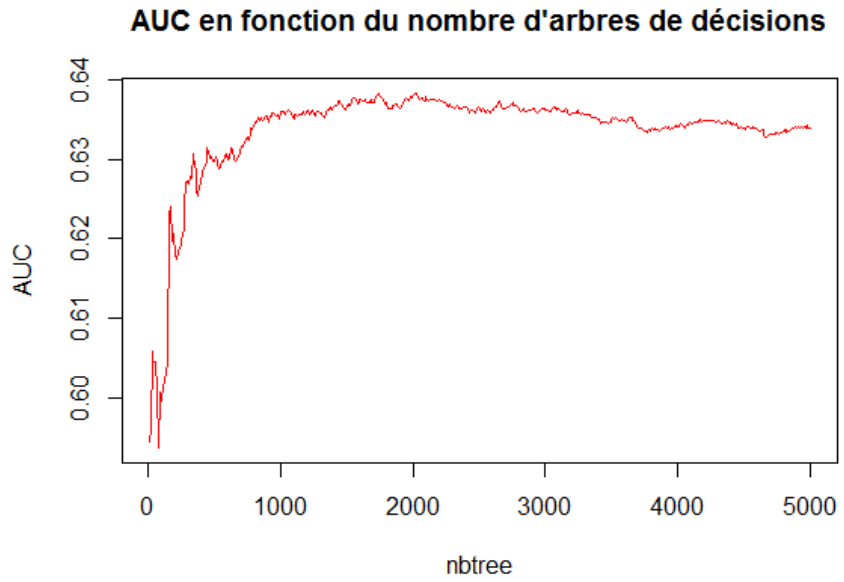


Figure 26 : AUC en fonction du nombre d'arbre sur l'échantillon de validation

A partir de **1 500 arbres**, on remarque que l'AUC n'augmente plus. On se fixe donc ce nombre d'arbres pour la suite du paramétrage.

On teste ensuite 5 vitesses d'apprentissage distinctes:

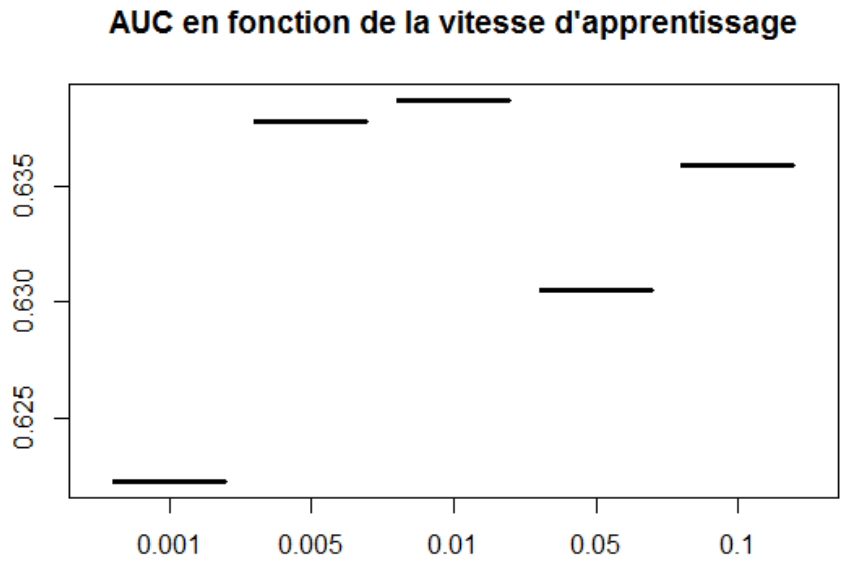


Figure 27 : AUC en fonction de la vitesse d'apprentissage sur l'échantillon de validation

La vitesse présentant la meilleure AUC est également le paramètre le plus parcimonieux. Il représente un juste milieu entre faible vitesse mais nécessitant un très grand nombre d'arbres et une vitesse grande mais peu précis. **On fixe la vitesse d'apprentissage à 0,01.**

Enfin, on teste les profondeurs d'interaction de 1 à 10, du modèle le plus simple au plus complexe.

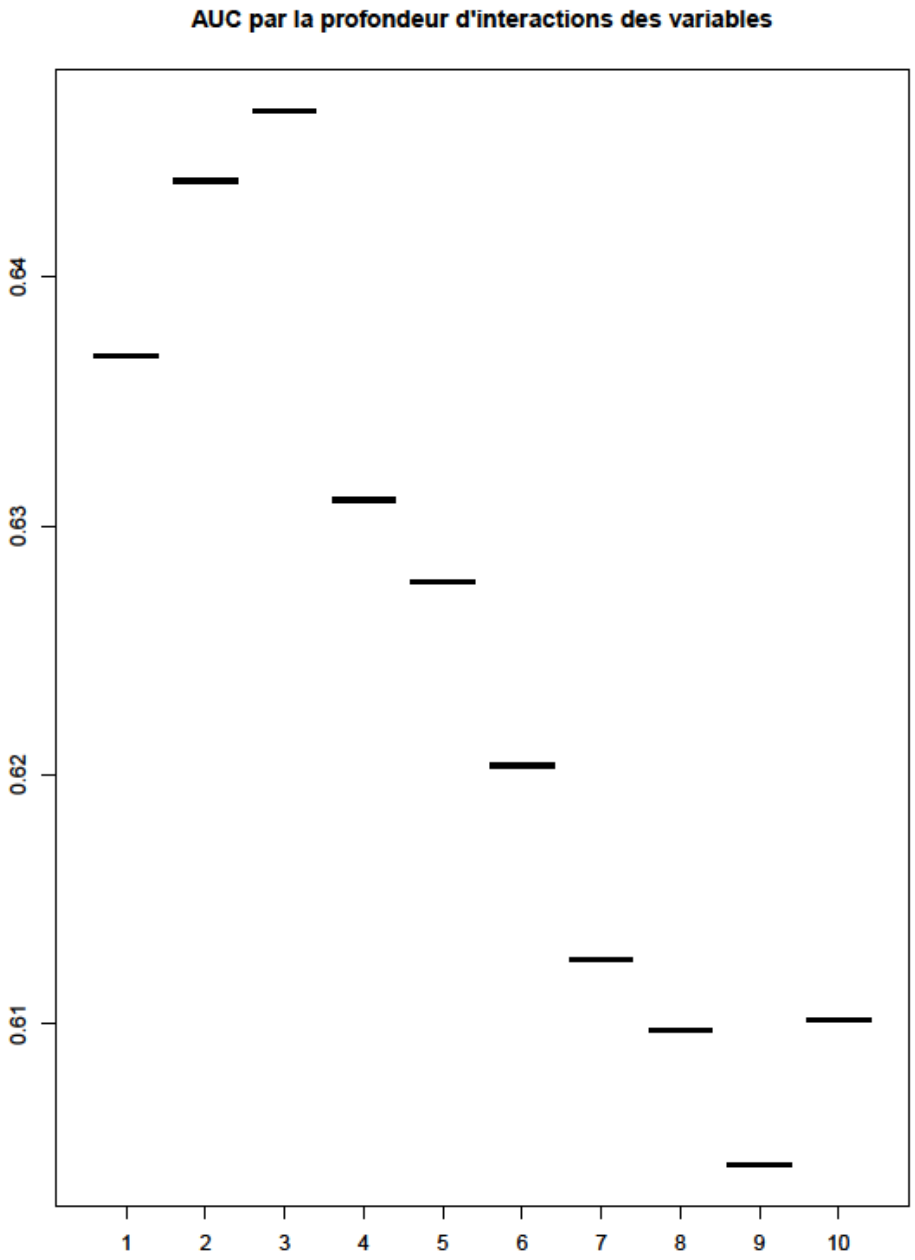


Figure 28 : AUC en fonction de la profondeur d'interaction sur l'échantillon de validation

L'AUC la plus élevée est pour une profondeur à 3.

On a donc nos 3 principaux paramètres : 1500 arbres, une vitesse d'apprentissage de 0,01 et une profondeur de 3. La paramétrisation du GBM permet de construire un modèle parcimonieux. Il se doit d'être performant mais également peu complexe. Il permet ainsi d'éviter à minima les cas de sur-apprentissage.

d) Comparaison des deux méthodes

Par la sélection *stepwise*, on obtient 8 variables explicatives pour la régression logistique contre 56 qui apparaissent comme explicative pour le Gradient Boosting Machine.

On compare la précision du Gradient Boosting et de la régression logistique grâce à l'AUC sur les différents échantillons construits.

	AUC d'apprentissage	AUC de test
GBM	**0.767034**	**0.609844**
Régression logistique	**0.656977**	**0.661266**

Figure 29 : AUC par échantillon et modèle

Pour l'ensemble des échantillons et des méthodes de modélisations, les AUC ne sont pas très élevées et comprises entre 0,6 et 0,8. On remarque cependant que l'AUC d'apprentissage du GBM semble acceptable. Cependant, elle est très supérieure à l'AUC de test. On détecte alors un possible phénomène de sur-apprentissage du GBM. Le GBM modélise bien sur l'échantillon d'apprentissage mais seulement sur celui-ci. Il reproduit la même structure sur les autres échantillons.

L'AUC d'apprentissage de la régression logistique est 10 points en dessous de celle du GBM mais est très proche son l'AUC de test.

Le choix du modèle se fait à partir de l'AUC calculée sur l'échantillon de test. Celui de la régression logistique étant 6 points supérieurs à celui du GBM, on écarte le gradient Boosting Machine comme modèle. De plus, l'utilisation du GBM nous impose d'utiliser un logiciel différent de celui utilisé habituellement : SAS. Cela peut être problématique à terme dans la maintenance du modèle par la méconnaissance de R ou Python et du machine learning.

	Modèle linéaire généralisé	Gradient Boosting
Avantage	• Facile à interpréter • Méthode classique	• Adapté à de nombreux problèmes de modélisation que ce soit en régression ou en classification
Inconvénient	• Beaucoup de retraitement • Problème de convergence • Nombre de variables explicatives moins important • Connaitre la variable modélisée (loi)	• Effet Boite Noire -> problème d'interprétation • Méthode encore peu répandue • Attention au sur-apprentissage : le paramétrage (échantillonnage, vitesse d'apprentissage,…) peut être compliqué

Figure 30 : Avantage/Inconvénient des méthodes employées

e) Conclusion

La sélection des variables se fait par « stepwise » ce qui nous permet de retenir les effets significatifs suivants :

Variables explicatives de la probabilité de dépasser le seuil	
Endogènes	***Exogènes à l'iris (Source : INSEE)***
Ancienneté du contrat(-)	*% de logements de 5 pièces ou plus(+)*
Assurance pour compte du propriétaire(+)	*% de logement avec chauffage individuel(~)*
Chambre d'hôte(+)	
Nombre de pièces supérieures à 40 m²(+)	
Nombre de pièces principales(+)	
Qualité juridique	
Périmètre Garantie	
Type de bien	
+ : effet à la hausse	*- : effet à la baisse*

Figure 31 : Tableau récapitulatif des variables explicatives du dépassement de seuil

Comme pour la fréquence, on peut analyser le modèle de dépassement de seuil avec les mêmes quantités.

L'AUC de la régression logistique apparait faible avec seulement 0,65.

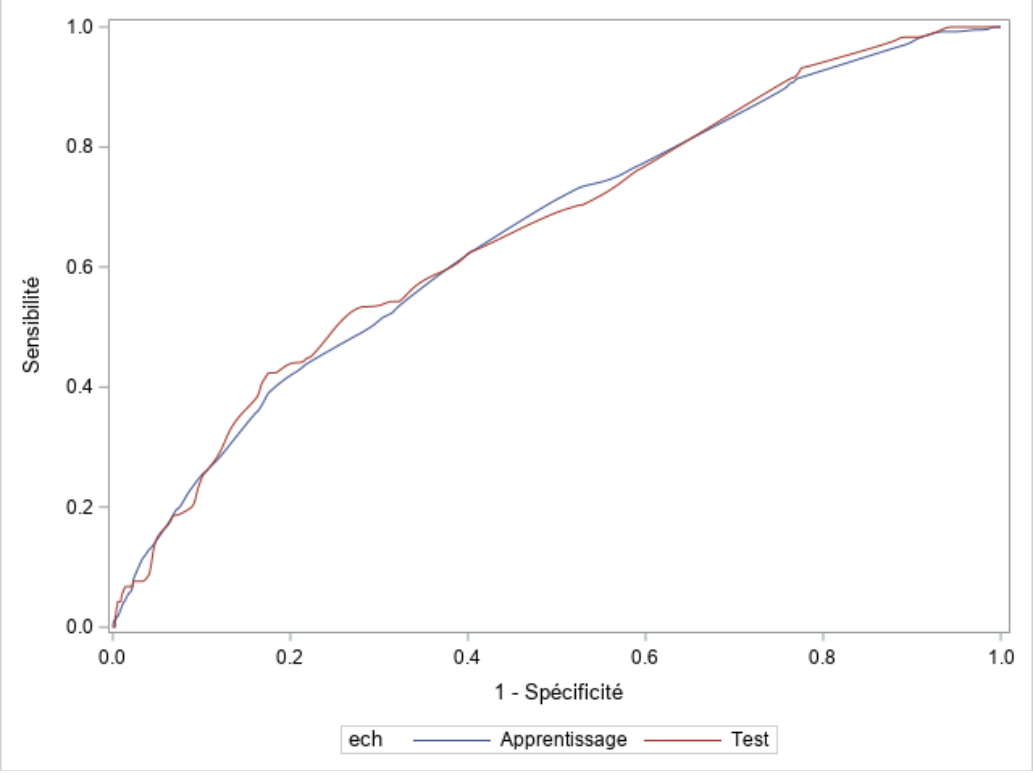


Figure 32 : Courbe ROC par échantillon sur le modèle GLM retenu

L'évènement modélisé étant particulièrement rare et atypique, on se satisfaisant d'une AUC faible. On propose cependant une piste d'amélioration future :

- Le déséquilibre entre le nombre de sinistres graves et non graves n'a pas été pris en compte dans cette étude. En effet, seulement 1.5% des sinistres Incendie sont identifiés comme graves. Cela implique un biais dans l'estimation des paramètres et notamment de l'*intercept*. Il est possible de le corriger en échantillonnant, dans un premier temps, de façon équilibrée (50% de grave/50% d'attritionnel). Ensuite, on dispose de 2 méthodes : une par pondération des observations et une par correction sur l'estimation de l'*intercept*. Cela n'ayant pas était fait dans cette étude, on ne développe pas ces solutions.

3. Fréquence grave Incendie

Pour obtenir le modèle de fréquence extrême Incendie, je multiplie la fréquence modélisée par la probabilité de dépassement de seuil sur l'ensemble des observations. Pour chaque observation, on dispose ainsi d'une fréquence grave modélisée.

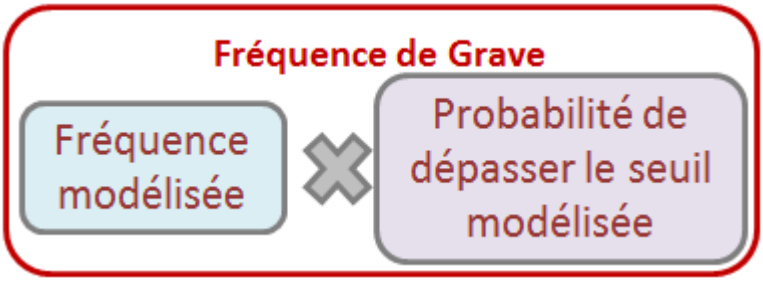


Figure 12 : Décomposition de la fréquence de grave

On obtient les effets décorrélés par produit des deux sous modèles. Je ne m'attarde pas sur cela, l'analyse est présentée au niveau de la prime pure en aval.

On observe initialement 757 sinistres graves Incendie. On en constate 776 modélisés. Les nombres étant proches, on se satisfait d'un tel résultat. On constate tout de même une surestimation de 2.5%, liée en partie au biais de surestimation sur l'*intercept* dû à la faible proportion de graves parmi les sinistres. Comme évoqué en amont, ce point sera étudié par la suite en dehors de ce mémoire.

Pour arriver à la prime pure grave Incendie, on complète par les modèles de coût moyen que je développe par la suite.

C. Coût moyen Incendie

1. Coût moyen grave

La modélisation du coût moyen grave est faite à partir de l'ensemble des sinistres Incendie survenus entre 2011 et 2018 avec un coût supérieur à 150 000 € en mars 2019 ce qui les qualifie comme grave. Les variables explicatives sont identiques à celles présentées dans le cadre de l'étude de la fréquence et du dépassement de seuil.

Au total, on travaille sur 757 sinistres graves Incendie.

Pour obtenir une prime pure, il est nécessaire de développer un modèle de coût moyen. La finalité étant d'avoir une clé de répartition qui puisse nous permettre de répartir la charge grave annuelle, il nous est indispensable d'avoir un coût moyen segmenté. Sans segmentation, cela revient à répartir selon la fréquence uniquement. Pour obtenir cette segmentation, on utilise un GLM Gamma avec une fonction de lien logarithme pour modéliser le coût moyen grave.

Goodness-of-Fit Tests for Gamma Distribution				
Test	Statistique		p-value	
Kolmogorov-Smirnov	D	0.03311978	Pr > D	>0.250
Cramer-von Mises	W-Sq	0.17831596	Pr > W-Sq	>0.250
Anderson-Darling	A-Sq	1.27141001	Pr > A-Sq	0.244

Goodness-of-Fit Tests for Lognormal Distribution				
Test	Statistique		p-value	
Kolmogorov-Smirnov	D	0.0843289	Pr > D	<0.010
Cramer-von Mises	W-Sq	1.7143016	Pr > W-Sq	<0.005
Anderson-Darling	A-Sq	10.3678926	Pr > A-Sq	<0.005

Goodness-of-Fit Tests for Pareto Distribution				
Test	Statistique		p-value	
Kolmogorov-Smirnov	D	0.03228444	Pr > D	0.094
Cramer-von Mises	W-Sq	0.18083851	Pr > W-Sq	0.038
Anderson-Darling	A-Sq	1.07382192	Pr > A-Sq	0.054

Figure 33 : Test de lois pour le coût moyen grave

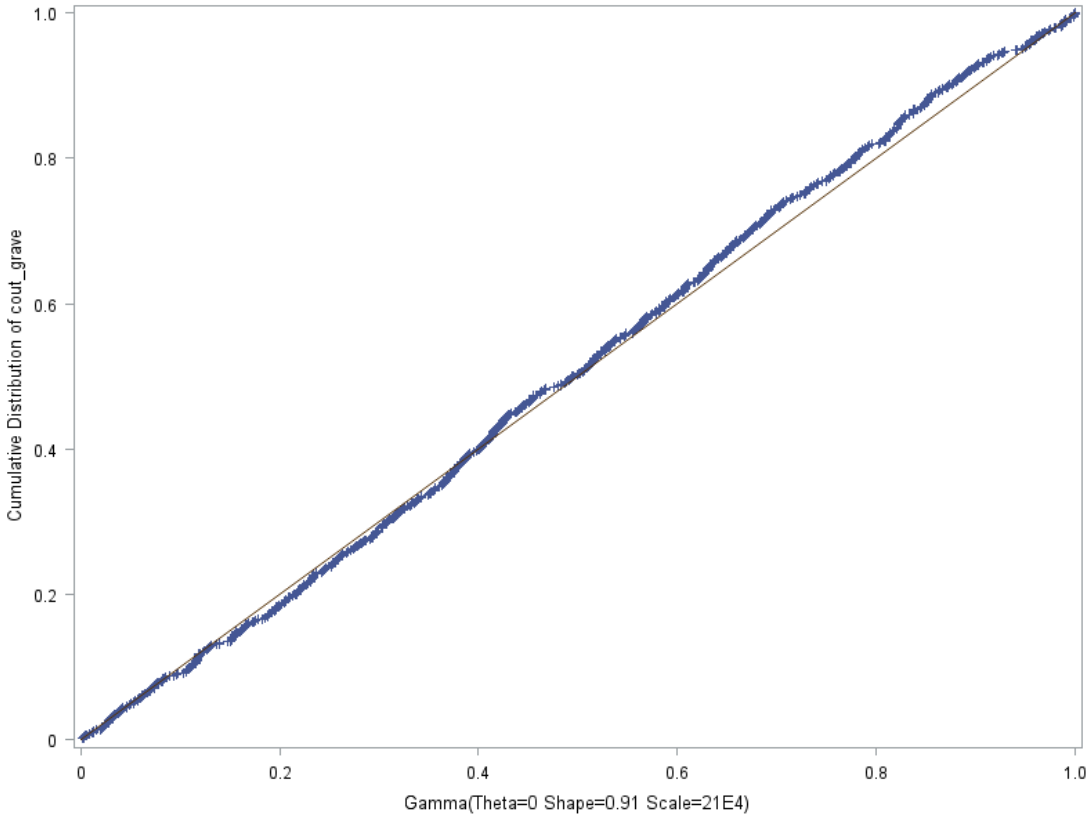


Figure 34 : PP-plot du coût moyen grave avec une loi Gamma

Parmi les lois testées, la loi Gamma ressort se distingue fortement. En effet, avec des pvalues>0.25, on peut considérer que le coût moyen grave suit une loi Gamma. De plus, il est intéressant de noter que le paramètre de forme estimé est relativement proche de 1 ce qui aurait été équivalent à une loi exponentielle.

La sélection de variables se fait par une méthode « stepwise » qui nous permet de retenir les effets significatifs suivants :

Variables explicatives du coût moyen grave	
Endogènes	***Exogènes à l'iris (Source : INSEE)***
Age de l'assuré(~)	*% de logements construits avant 1919(+)*
Catégorie socio-professionnelle	*% de couples avec enfants(+)*
Formule souscrite	
Nombre de pièces supérieures à 40m²(+)	
Nombre de pièces principales(+)	
Type de bien	
+ : effet à la hausse	*- : effet à la baisse*

Figure 35 : Tableau récapitulatif des variables explicatives du coût moyen grave

Le risque lié à l'assuré lui-même est expliqué par son âge, sa CSP ou bien la formule qu'il a souscrite qui peut traduire son niveau d'aversion au risque.

Les variables retenues ne sont pas surprenantes : le type de bien, le nombre de pièces principales ainsi que le nombre de pièces supérieures à 40m² caractérisent la matérialité du bien assuré.

Cela est complété par les données géographiques qui reprennent les informations sur la typicité de l'iris où se situe le bien. On note tout de même que la part de logements construits antérieurement à 1919 peut contenir également des informations sur la matérialité du bien, ne disposant de données sur l'ancienneté de l'habitat sur le risque assuré.

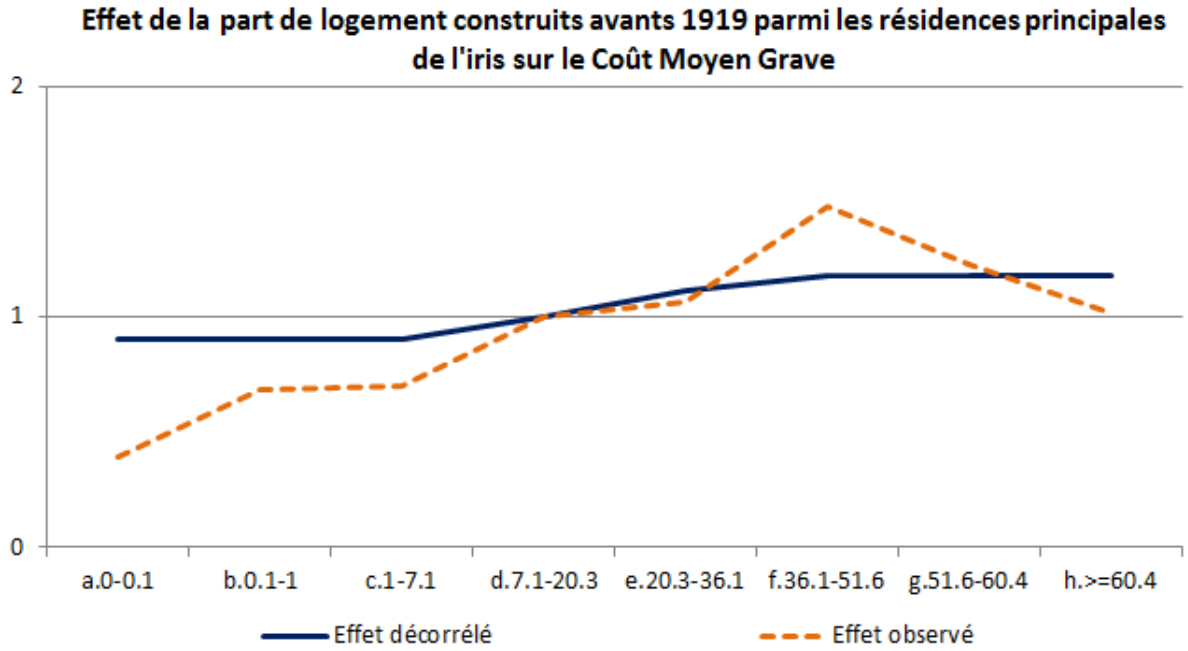


Figure 36 : *Odds Ratio* en fonction de la part de logements construits avant 1919 sur le coût moyen

On observe que l'ancienneté de l'habitat à l'iris semble aggraver le coût moyen grave. En effet, si la concentration de logement construits avant 1919 est supérieure à 36.1% ce qui correspond aux modalités f, g et h, le coût moyen est dégradé de +18% du seul effet de cette ancienneté de l'habitat environnant le bien assuré.

A ce stade de l'étude, nous avons le modèle de prime pure grave. On utilise cette dernière afin de retraiter le coût dans la partie suivante.

2. Coût moyen

La base d'étude du coût moyen Incendie est identique à celle du dépassement de seuil.

Afin de traiter l'hétérogénéité des sinistres graves qui posent problème, le coût est retravaillé.

Soit $(C_i)_{i=1..n}$ les coûts des n sinistres.

Soit $(PPG_i)_{i=..n}$ les primes pures grave Incendie modélisées pour les n sinistres

On pose :

$$E_i = \min(C_i, 150000)$$ le coût écrêté du sinistre *i*

$$S = \sum_{i=1}^{n} \max(C_i - 150\,000{,}0)$$ la surcrête totale

De là, on définit le coût à modéliser comme :

$$C_i' = E_i + \frac{PPG_i}{\sum_{i=1}^{n} PPG_i} * S$$

La prime pure grave modélisée nous sert de clé de répartition de la surcrête grave. On lisse les sinistres extrêmes en diminuant l'écart type du coût moyen et en affinant la queue de distribution tout en conservant un coût moyen identique.

On utilise un GLM Log-Gamma pour traiter ce coût recalculé. La fonction Log n'est pas la fonction canonique de la loi Gamma. Cependant, l'utilisation de celle-ci impose une positivité à la variable modélisé, chose indispensable pour un coût moyen.

Comme pour les précédents modèles, les effets sont sélectionnés par une méthode « stepwise ». Les effets du modèle de coût moyen incendie sont alors les suivants :

Variables explicatives du coût moyen	
Endogènes	***Exogènes à l'iris (Source : INSEE)***
Logarithme de l'année(+)	*% de logements de 5 pièces ou plus(-)*
Capital mobilier assuré(~)	*% de logements avec chauffage individuel(~)*
Chambre d'hôtes(+)	*% de personnes de plus de 75 ans(~)*
Catégorie socio-professionnelle	*% d'artisans ou commerçants(~)*
Formule souscrite	*% d'agriculteurs ou salariés agricoles(-)*
Nombre d'enfants dans le foyer de l'assuré [~)	*% de logements construits en 1920 et 1945(+)*
Nombre de pièces supérieures à 40m²(+)	*% de logements construits avant 1919(+)*
Nombre de pièces principales(+)	*% de logements de moins de 30m²(~)*
Périmètre Garantie	*% de logements en location(-)*
Superficie de la dépendance(+)	
Type de bien	
+ : effet à la hausse	*- : effet à la baisse*

Figure 37 : Tableau récapitulatif des variables explicatives du coût moyen

a) Aspect Temporel

Comme pour la fréquence, une tendance temporelle se distingue. La variable temporelle qui apparait comme la plus explicative est le logarithme de l'année de survenance.

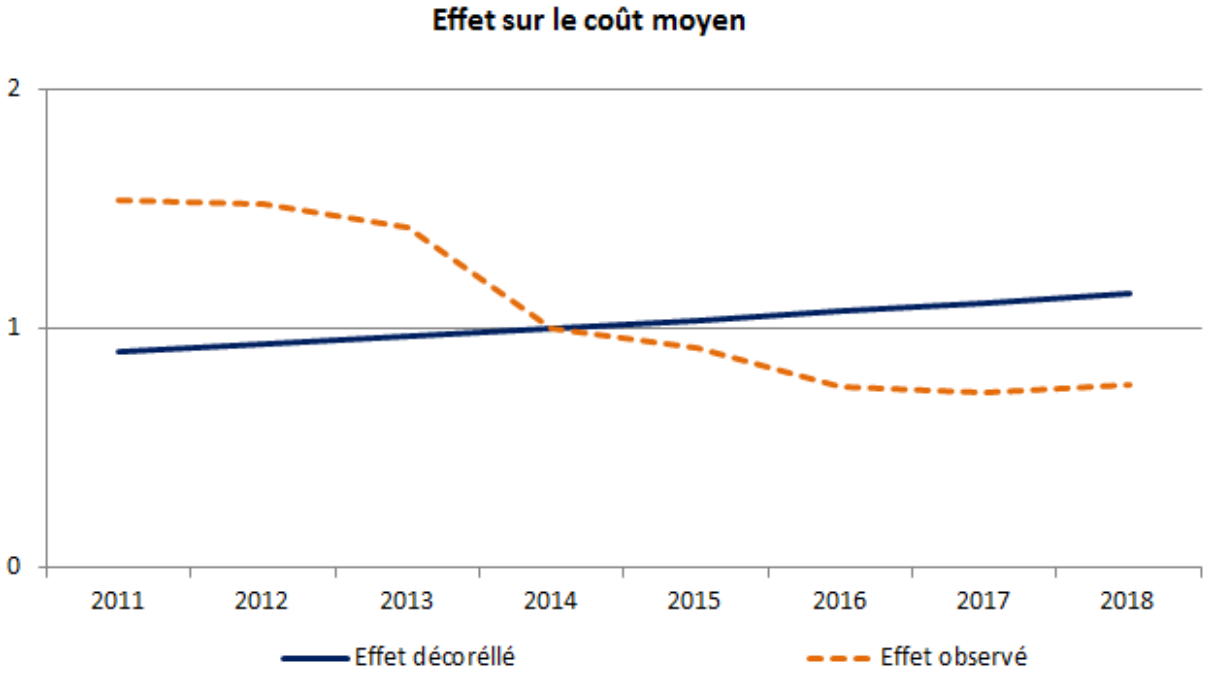


Figure 38 : *Odds Ratio* en fonction de l'année d'exercice sur le coût moyen

On observe un coût moyen observé en baisse. Le dommage électrique immobilier a un coût moyen plus faible que l'incendie. L'extension de la garantie incendie avec l'inclusion de ces sinistres dommages électriques induit donc une baisse de son coût moyen. Je rappelle que pour tenir compte de ce changement de garantie, le périmètre de la garantie à la survenance du sinistre est introduit comme variable explicative du modèle. L'effet décorrélé ne tient alors compte que de l'évolution tendancielle du coût moyen.

Contrairement à la fréquence, le coût moyen semble suivre une tendance haussière. En effet, celui-ci semble augmenter d'environ +3 à +4% par an.

b) Aspect géographique

On isole l'effet géographique lié aux variables INSEE de la même manière que pour la fréquence puis on le cartographie à l'échelle de la France.

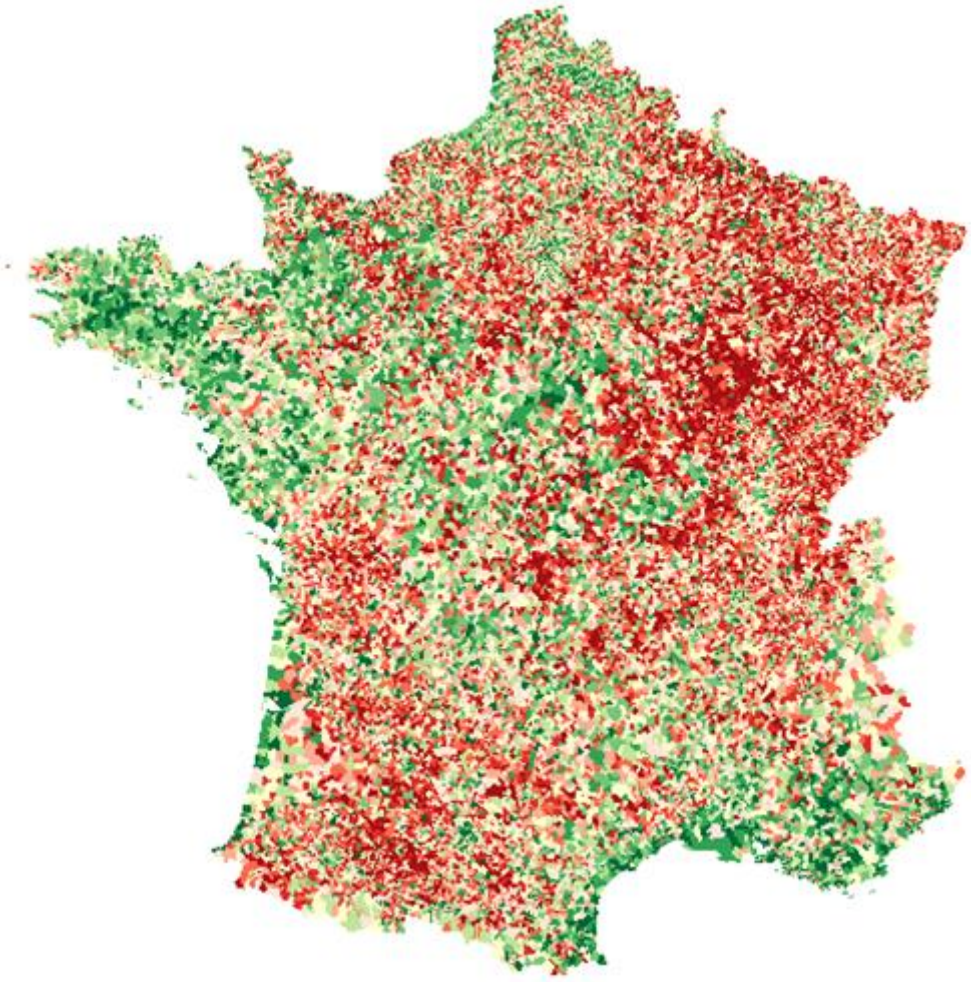


Figure 39 : Représentation des effets des variables INSEE sur le coût moyen

Le risque qui semble accru en fréquence sur les côtes atlantiques et méditerranéennes n'apparait plus ainsi. Cependant l'Aquitaine semble toujours fortement touchée mais également l'est de la France ce qui semble correspondre à des régions boisées et donc possiblement à des moyens de chauffage au bois plus répandus.

A partir de cette carte et de celle obtenue pour la fréquence, on peut visualiser l'effet spatial sur la prime pure Incendie que nous verrons dans la suite de l'étude.

D. Prime pure et prime pure grave

L'ensemble des sous-modèles nécessaires au calcul d'une prime pure grave Incendie vient d'être détaillé.

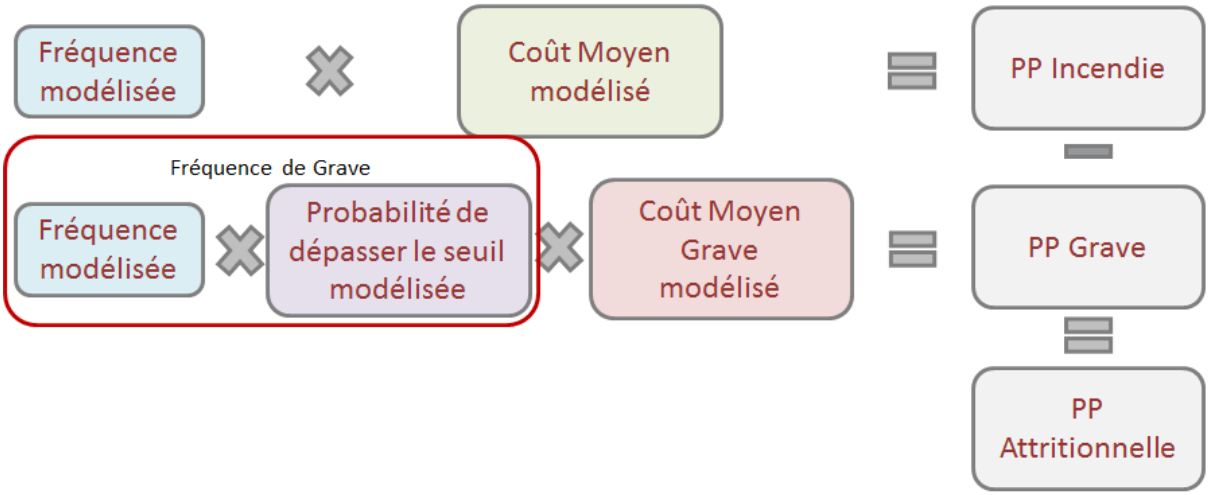


Figure 14 : Décomposition de la prime pure incendie

On rappelle qu'avec cette approche, il y a la possibilité d'obtenir une prime pure attritionnelle aberrante qui est négative du fait l'absence de contrainte entre la pure globale et la prime pure grave. Cependant, ce problème est accepté car seule la prime pure grave nous est nécessaire pour répartir la surcrête grave pour l'obtention de nos indicateurs de rentabilité « technique ». Le calcul des primes pures globales et attritionnelles n'est utilisé qu'à titre d'information. En effet, la construction du modèle de coût moyen est un moindre effort pour connaitre la part de la sinistralité grave aux critères de risque.

1. Variables explicatives

Je commence par présenter un résumé des effets introduits dans les différents modèles.

Variable	Fréquence	coût moyen	Dépassement de seuil	coût moyen grave
Age de l'assuré	-			x
Ancienneté contrat	-	-	-	
Année	-	+		
4 CP*	x	x	x	
Isolement	+			
Capital Mobilier	+	x		
RP/RS	x			
CSP	x	x	x	x
Formule	x	x	x	x
Nombre d'enfants	+	+		
Nombre de pièces > 40m2	+	+	+	+
Maison / Appartement	x	x	x	x
Nombre de pièces	+	+	+	+
20 Données Insee à l'iris	x	x	x	x
Dépendance à une autre adresse	+			
Qualité juridique	x		x	
Superficie dépendance	+	+		
Périmètre garantie (3 variables)	x	x	x	
	+ : effet à la hausse	- : effet à la baisse	x : pas de tendance particulière	

Figure 40 : Tableau récapitulatif des variables explicatives de la prime pure

L'ensemble de ces données nous permettent de comprendre le risque Incendie dans sa globalité ainsi que dans ses extrêmes. On remarque un nombre important de variables exogènes INSEE qui nous permettent d'appréhender la part du risque lié à l'aspect spatial. Cela apporte une complexité au modèle.

2. Analyse par critères de risque

Sans surprise, les effets sont à la hausse sur les critères physiques du risque : maison, propriétaire, nombre de pièces, superficie de la dépendance et nombre de pièces supérieures à 40m². Par produit des effets décorrélés de chaque sous modèle, on obtient l'effet décorrélé pour les primes pures. On isole ainsi les effets que l'on souhaite analyser.

On s'attarde désormais sur les résultats des principaux critères de risque.
Le graphique se compose des éléments suivants :

- La courbe bleue correspond à l'effet décorrélé pour la prime pure grave
- La courbe verte correspond à l'effet décorrélé pour la prime pure globale

Sur le critère de la qualité juridique, on obtient ainsi le graphique suivant :

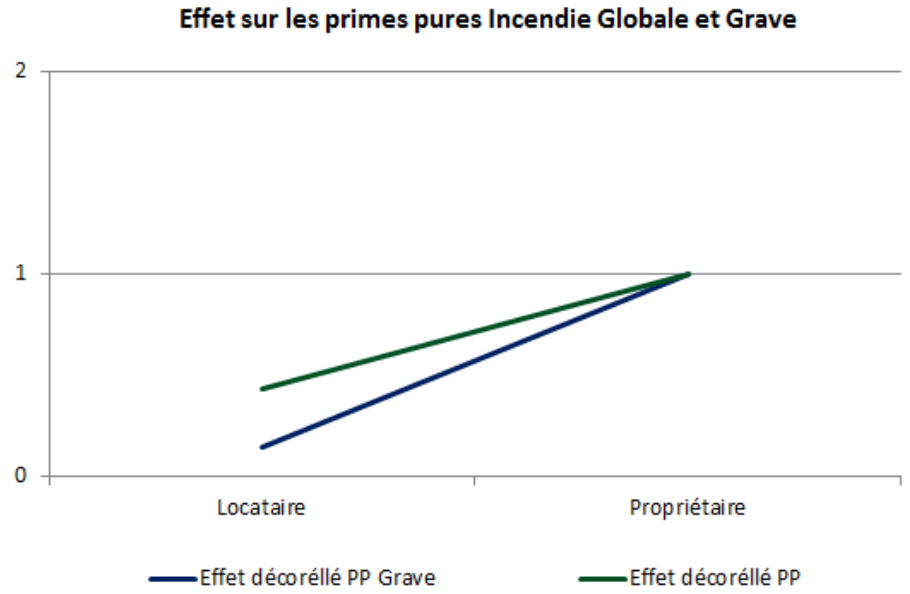


Figure 41 : *Odds Ratio* en fonction de la qualité juridique sur la prime pure grave et incendie

Les locataires semblent très fortement avantagés par rapport aux propriétaires. Du seul fait d'être locataire, la prime pure modélisée est réduite de 57% par rapport à un propriétaire et de 86% sur la partie grave de la prime pure.

L'effet est accentué pour les locataires d'appartement, ceux-ci bénéficiant d'un effet à la baisse pour ces 2 critères.

Afin de s'assurer de la fiabilité des modèles, on représente la prime pure globale incendie, la prime pure grave incendie ainsi que la fréquence observée et modélisée.

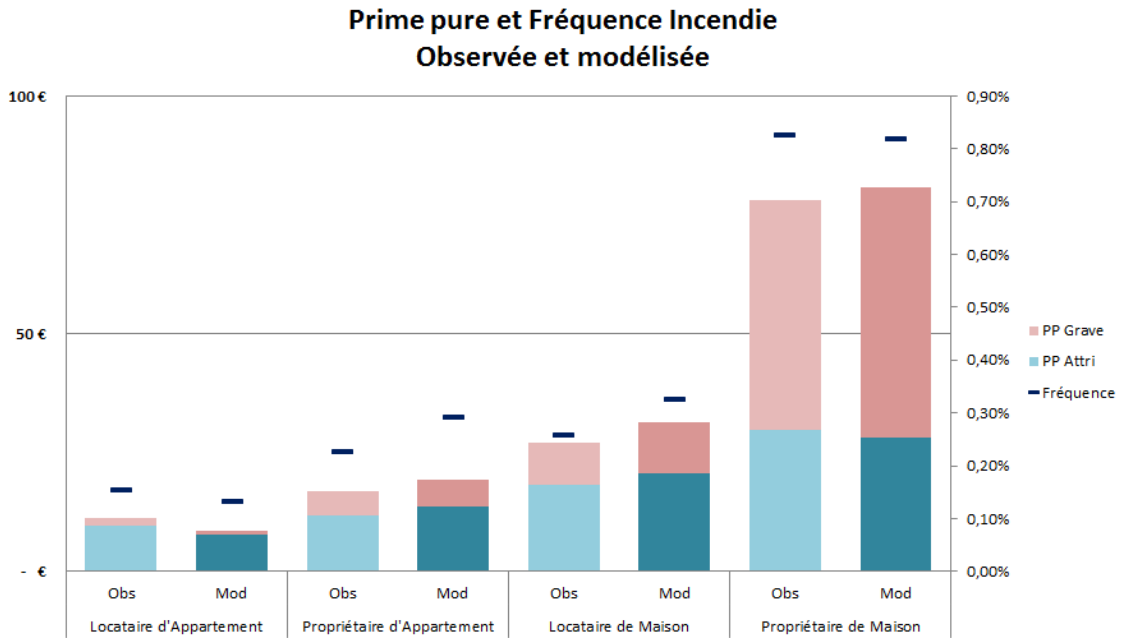


Figure 42 : Prime pure décomposée et fréquence en observée et modélisée en fonction de la qualité juridique et le type de maison

On remarque que les propriétaires de maison sont plus sujets au risque Incendie que les autres populations. La partie grave représente plus de 60% de la prime pure sur ce profil de risque. On zoome sur cette population de propriétaires de maison.

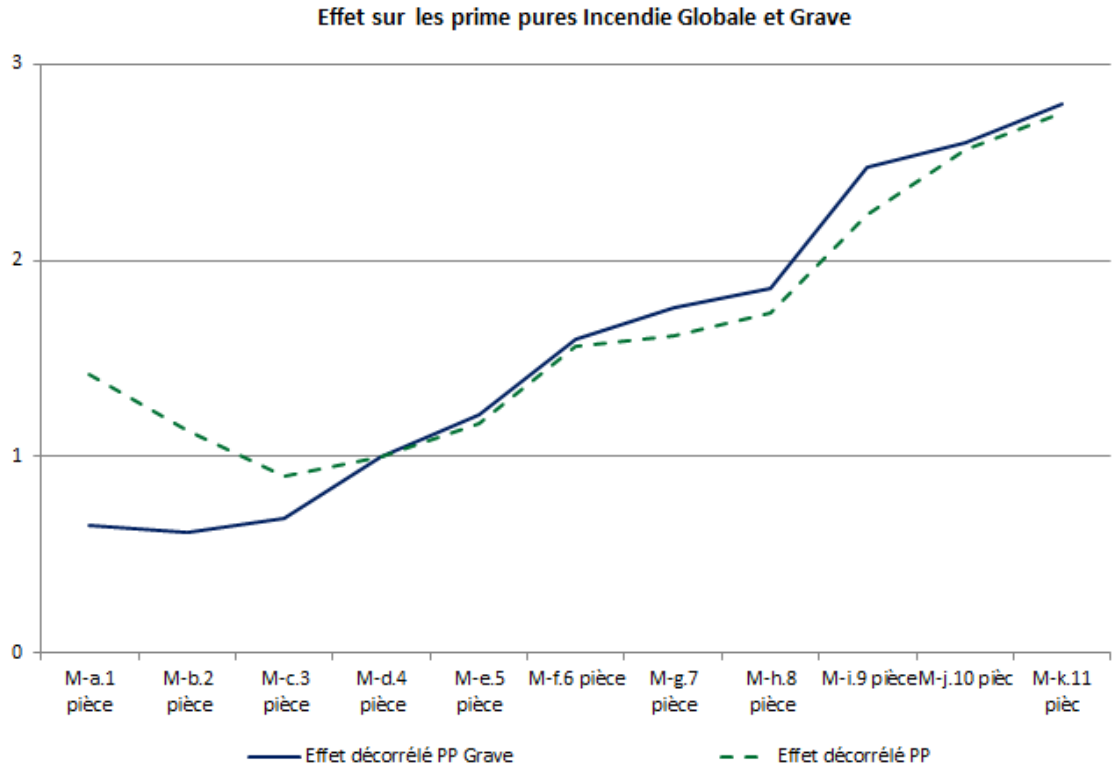


Figure 43 : *Odds Ratio* en fonction du nombre de pièces sur les propriétaires de maison sur la prime pure grave et incendie

L'effet du nombre de pièces principales est intéressant ici. En effet, on constate un effet décorrélé de prime pure plus fort sur les 1-2 pièces que sur les 3-4 pièces. Cette observation se vérifie également avec les valeurs observées et modélisées :

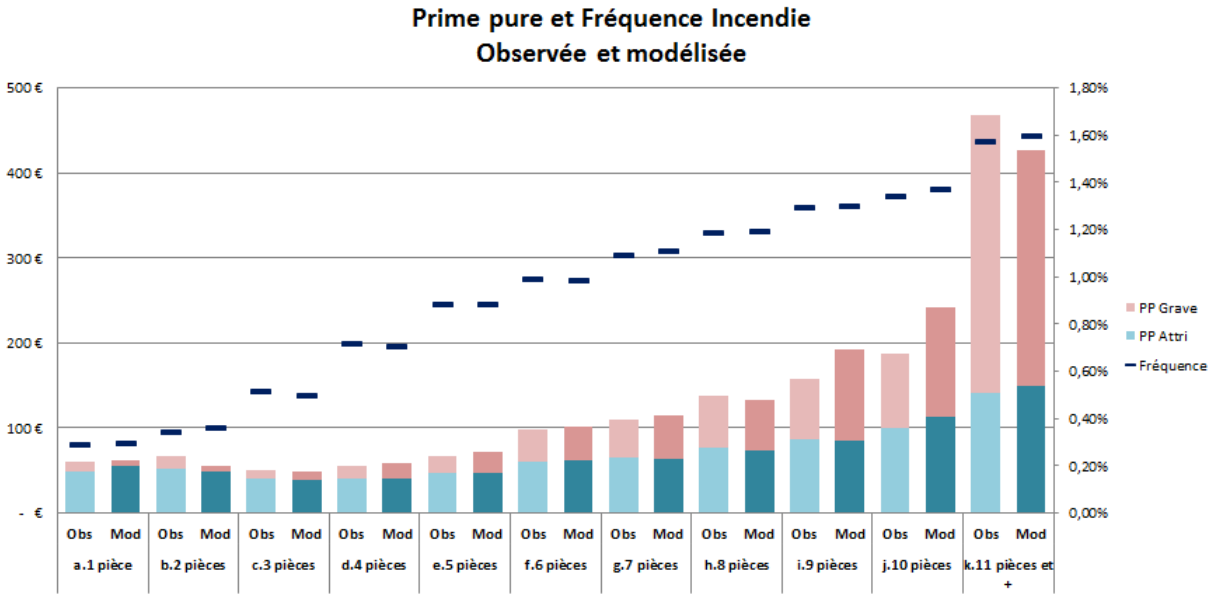


Figure 44 : Prime pure décomposée et fréquence en observée et modélisée en fonction du nombre de pièces sur les propriétaires de maison

Ce prime pure plus forte sur les petites maisons peut être expliqué en partie par des biens assurés de type Loft souscrits comme des maisons. Ces derniers sont plus risqués que des maisons plus standard avec 3 ou 4 pièces principales.

Au-delà de 3 pièces principales, on trouve une tendance à la hausse que cela soit pour la prime pure grave ou pour la prime pure globale. Les biens de 11 pièces et plus se démarquent tout de même des biens avec un nombre de pièces plus modeste. La prime pure grave modélisée est de 129€ pour une maison de 10 pièces puis passe à 278€ pour une maison de 11 pièces et plus, soit +115%. L'effet décorrélé nous indique cependant un risque accru de seulement 19%. Ce type de biens avec un grand nombre de pièces ont souvent une accumulation de critères aggravants : pièces de plus de 40m², grandes dépendances, château, manoir,…

La seule présence d'une pièce de plus de 40m² augmente la prime pure de +47% et de +79% pour la prime pure grave, cela quel que soit le nombre de pièces principales.

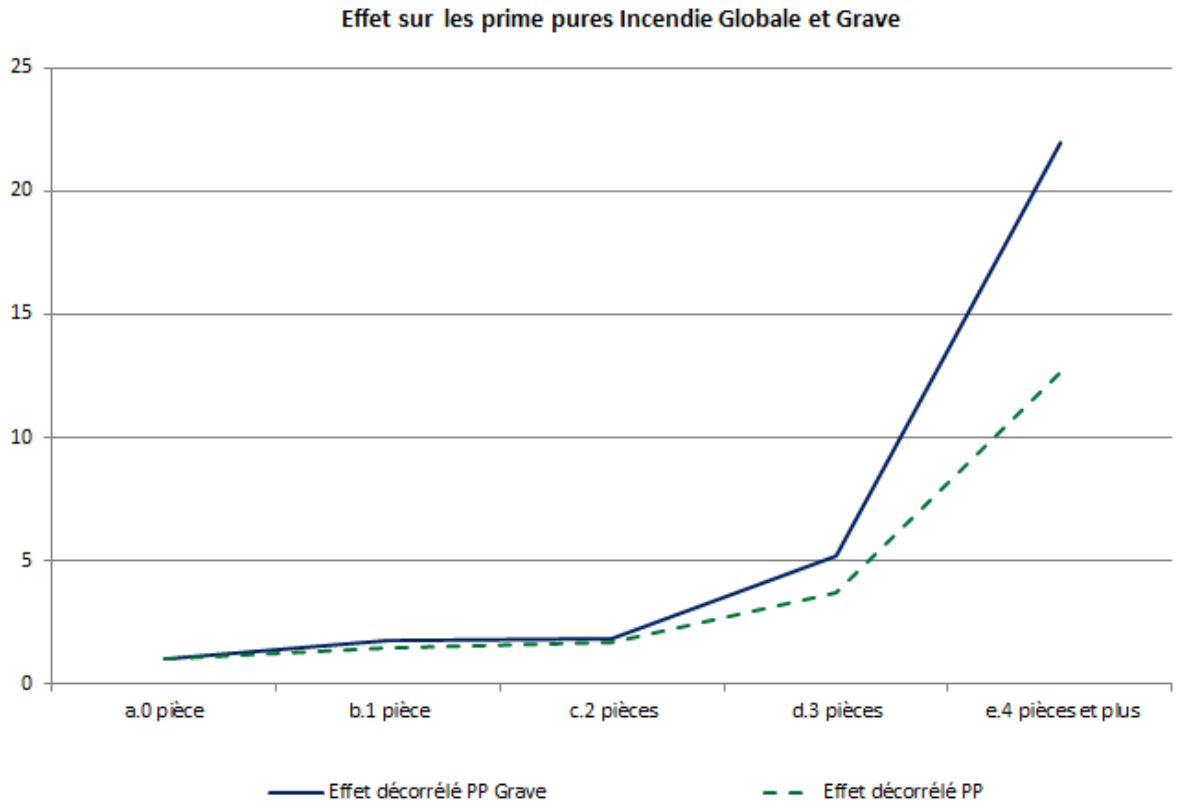


Figure 45 : *Odds Ratio* en fonction du nombre de pièces de plus de 40m² sur la prime pure grave et incendie

Avec cette approche, on ne dispose pas d'un équivalent en nombre de pièces principales de la pièce de plus de 40m². On a tendance à considérer instinctivement une pièce de 40m² comme équivalent à 1,5 ou 2 pièces principales. Dans notre modèle, les effets des deux variables Pièces>40m² et Pièces Principales se multiplient. Ainsi, pour le cas d'une maison de 4 pièces, sans la présence de pièces>40m², l'effet décorrélé est neutre, 4 pièces étant la référence. Si une des 4 pièces est supérieure à 40m², alors l'effet combiné est de +79% soit environ l'effet décorrélé d'une maison de 7 pièces. Pour une maison de 3 pièces dont une de 40 m ², l'effet combiné est de +23%, soit environ l'équivalent d'une maison de 5 pièces sans pièce supérieure à 40m². Pour une maison de 6 pièces dont 1 de plus de 40m², l'effet combiné est de +186%. L'équivalent est alors une maison de plus de 11 pièces de moins de 40m².

L'effet de la présence d'au moins 4 pièces de plus de 40m² est de +2 100% pour la prime pure grave et +1 165% pour la prime pure Incendie. Un seul sinistre de 4.7M€ est survenu sur cette population entre 2011 et 2018. Appliquer tel quel, cet effet entraine une dérive de la prime pure modélisée.

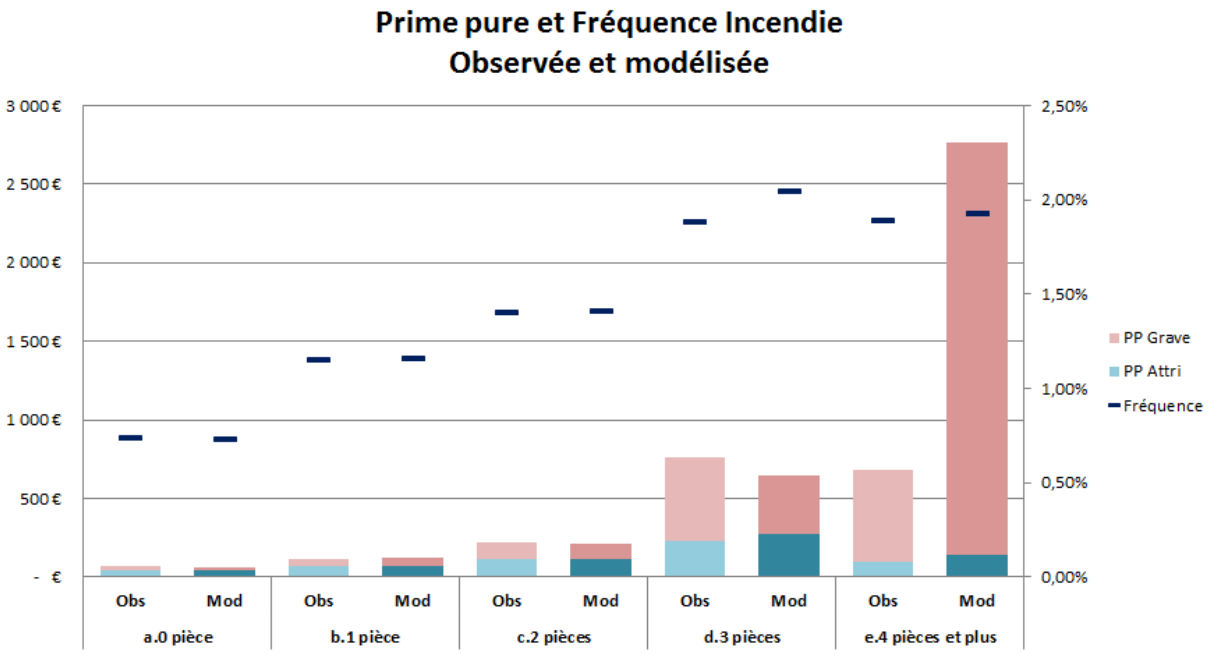


Figure 46 : Prime pure décomposée et fréquence en observée et modélisée en fonction du nombre de pièces de plus de 40m²

Ce seul sinistre de 4.7M€ sur un bien de type Château et au vu du faible volume sur ce profil de risque (moins de 4000 années d'exposition en 8 ans), on se permet de corriger manuellement cet effet en mutualisant l'effet sur les 3 pièces ou plus supérieures à 40m². Cela ramène l'effet de +2 100% à un niveau de de +651%. La prime pure grave initialement estimée à 2 628€ passe alors à 897 €, plus proche de l'observé.

3. Aspect géographique

Après cette analyse sur ces quelques critères, je propose de s'intéresser aux effets géographiques déjà présentés pour la fréquence et pour le coût moyen.

On peut se poser l'intérêt de variables comme la part de logements en location à l'iris étant donné que l'on dispose de la qualité juridique. Cela peut sembler répétitif. Ce que nous montre le modèle, c'est que, malgré ce qu'on aurait tendance à penser, cette part de logements en location apporte bien une information complémentaire.

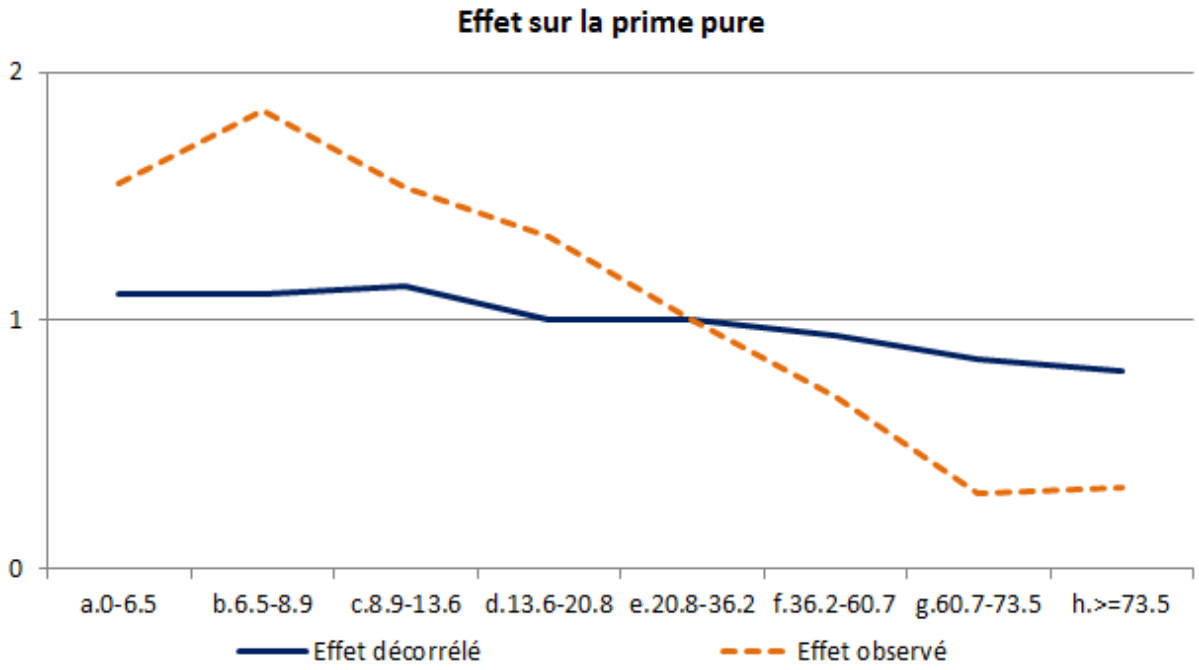


Figure 47 : *Odds Ratio* en fonction de la part de logement en location sur la prime pure incendie

L'effet décorrélé est loin d'être négligeable et celui-ci s'applique, que le bien assuré soit en location ou non. L'effet est à la baisse allant jusqu'à -20% pour les iris avec une concentration de logements en location au-delà de 73.5%.

En observant les primes pures sur les propriétaires de maison, on constate également une baisse avec cette concentration aussi bien en observé qu'en modélisé.

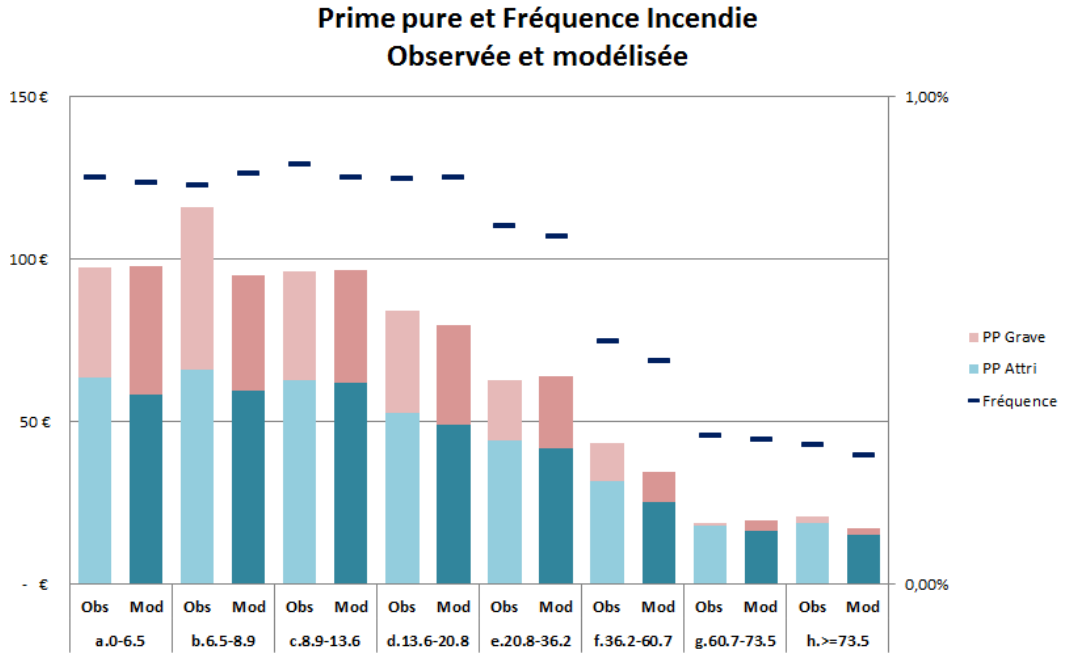


Figure 48 : Prime pure décomposée et fréquence en observée et modélisée en fonction de la part de logement en location à l'iris

On en conclut que cette concentration traduit bien une typicité de l'iris et ne prend pas en compte la qualité juridique.

Les estimations de paramètres sur les variables INSEE nous fournissent des informations sur les risques spatiaux. En agrégeant les effets déjà présentés sur la fréquence et le coût moyen, on obtient une carte de France du risque spatial à l'iris pour la prime pure Incendie.

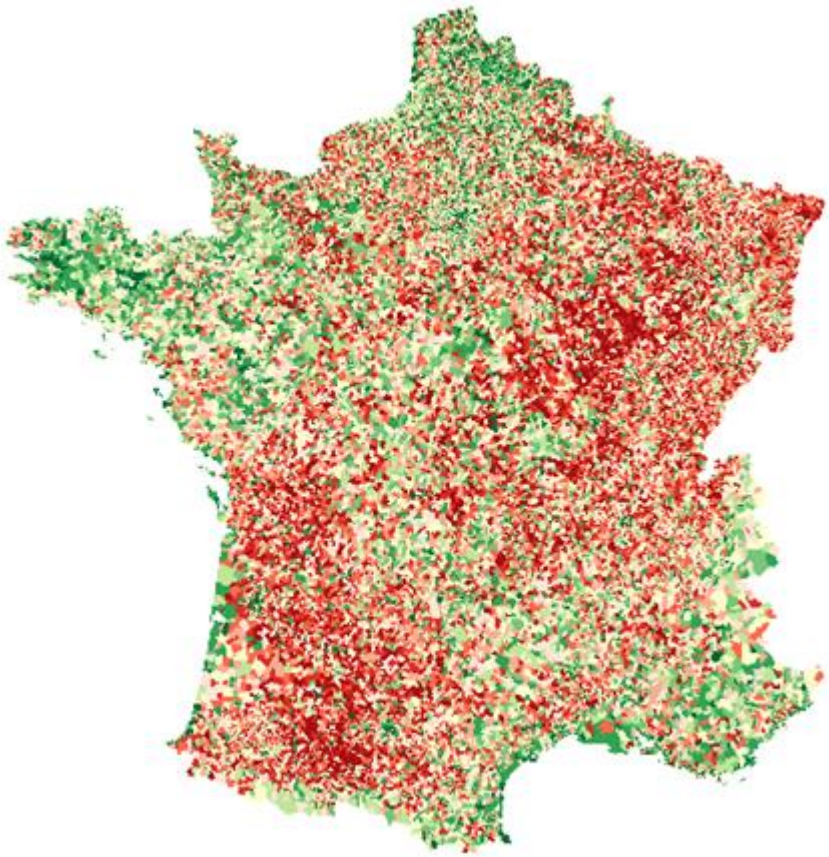


Figure 49 : Représentation des effets des variables INSEE sur la prime pure

La carte des effets prime pure semble fortement corrélée à la carte de la concentration en logements construits avant 1919 ci-dessous.

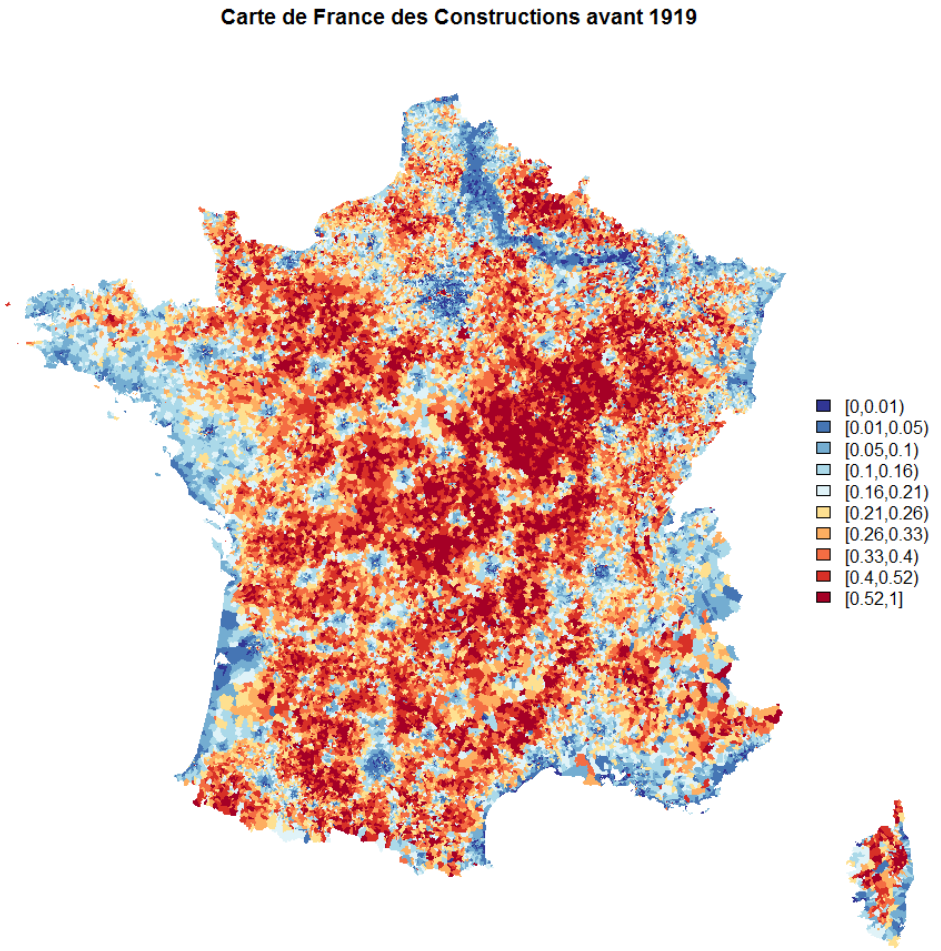


Figure 50 : Représentation géographique de la concentration en logement construit avant 1919

En effet, cette concentration à un effet à la hausse sur la prime pure Incendie.

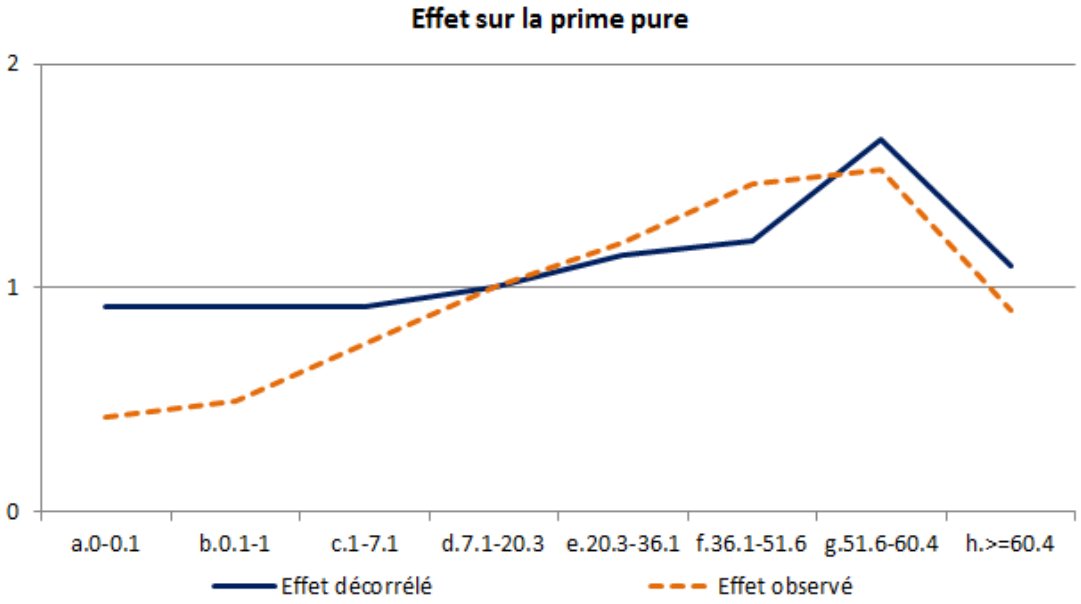


Figure 51 : *Odds Ratio* en fonction de la part de logement construit avant 1919 sur la prime pure incendie

L'effet peut aller jusqu'à une hausse de +66% due à cette ancienneté de l'habitat pour les concentrations comprises entre 51.6% et 60.4%.

Le traitement de ces données INSEE semble être une piste intéressante pour la constitution d'un zonier Incendie. On remarque cependant que contrairement au modèle construit dans ce mémoire, il peut être plus judicieux de distinguer les maisons des appartements.

La construction de nouveau zonier appartement et maison a donc été entrepris suite à cela courant 2021. En effet, les zoniers ne semblent pas expliquer correctement la prime pure incendie :

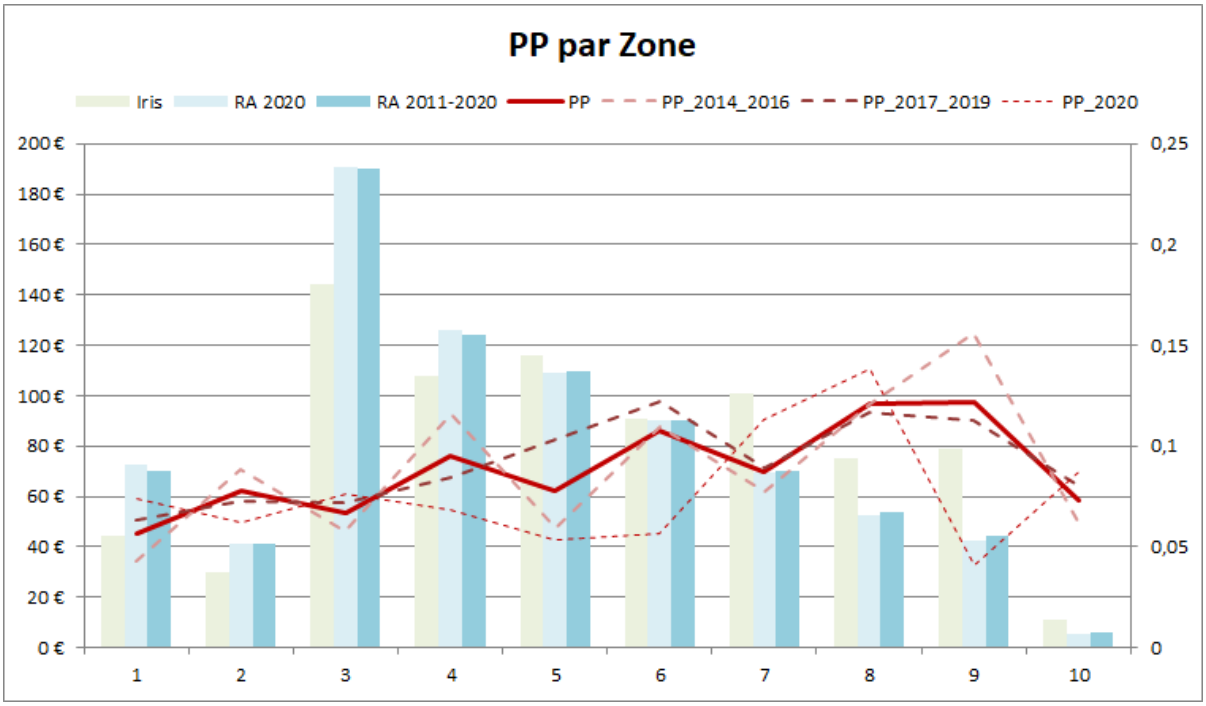


Figure 52 : Prime Pure Maison en fonction de l'ancien zonier Incendie maison

N'étant pas l'objet de ce mémoire, je développe succinctement la méthode employée pour la construction de ces deux nouveaux zoniers.

En capitalisant sur les données INSEE et la méthode employée pour réaliser ***la figure 46***, des modèles de prime pure Maison et Appartement incendie ont été reconstruits. À partir de plusieurs modèles GLM agrégés par Bagging. On identifie l'effet géographique uniquement par les données INSEE.

Les zoniers obtenus en sortie de modèle ont ensuite été lissés par cercles concentriques afin d'apporter une cohérence locale dans notre tarif. Le but est alors d'éviter de passer d'un IRIS classé en zone 10 à un IRIS classé en 1 juste en déplaçant le risque dans un IRIS voisin. Cette étape est primordiale étant donné que ces zoniers ont pour vocation à

être tarifaires par la suite. On obtient alors le nouveau zonier technique suivant en maison :

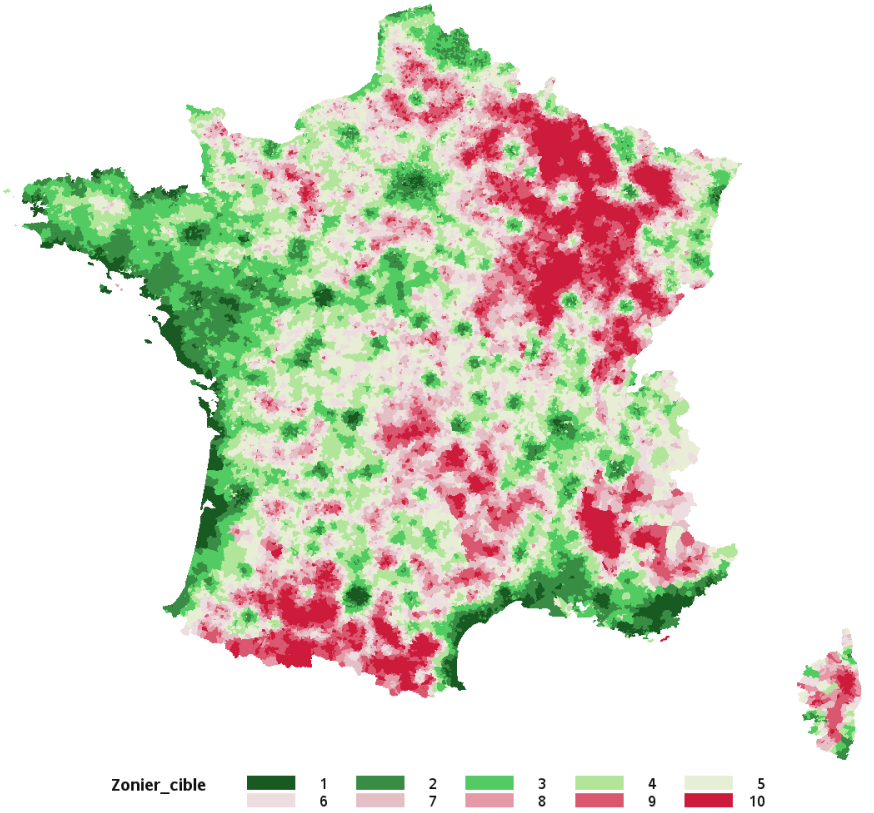


Figure 53 : Nouveau Zonier technique Incendie Maison

Ce nouveau zonier maison nous satisfait alors bien plus que le précédent. On trouve une cohérence globale avec des zones montagneuses ainsi que l'est de la France plus risqués à l'inverse des côtes méditerranéennes et atlantiques qui ont un risque plus faible. Enfin, on se satisfait de la prime pure que l'on observe sur ce zonier :

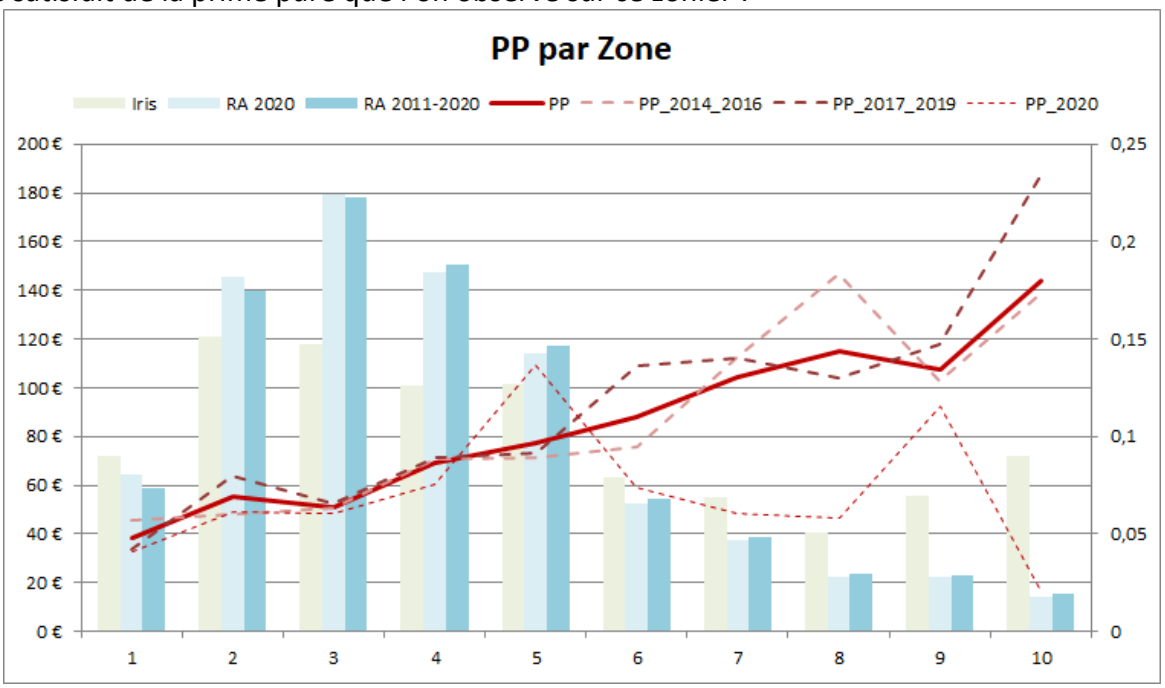


Figure 54 : Prime Pure Maison en fonction du nouveau zonier Incendie maison

En effet, cette dernière croit avec le niveau de la zone. Un travail identique a également été effectué sur les risques appartements.

Ces nouveaux zoniers et les nouvelles pentes tarifaires associées seront mis production dès Janvier 2022.

4. Aspect temporel

Les modèles de fréquence et de coût moyen ont des composantes temporelles. Je les associe afin d'obtenir l'effet temporel isolé de la prime pure.

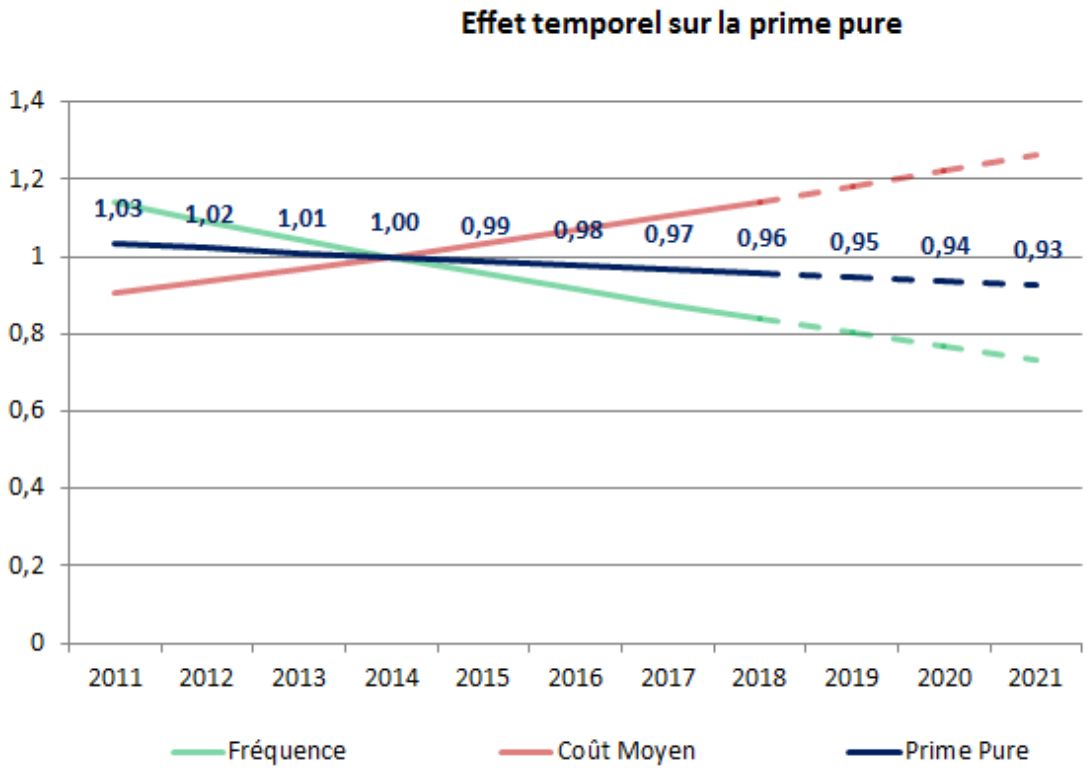


Figure 55 : *Odds Ratio* en fonction de l'année d'exercice sur la prime pure incendie

On rappelle que la tendance estimée pour la fréquence est d'environ -4% par an et que le coût moyen est en hausse de 3 à 4 % annuellement. Les variables temporelles étant quantitatives, on est en mesure de projeter l'effet du temps à court terme. On est alors en mesure d'envisager une évolution de prime pure Incendie pour les années à venir. On modélise une tendance à la baisse d'environ 1% par an.

Une des applications de cela est de projeter la prime pure à court terme. Pour un portefeuille 2019 identique à celui de 2018, on peut alors estimer une prime pure Incendie attendue.

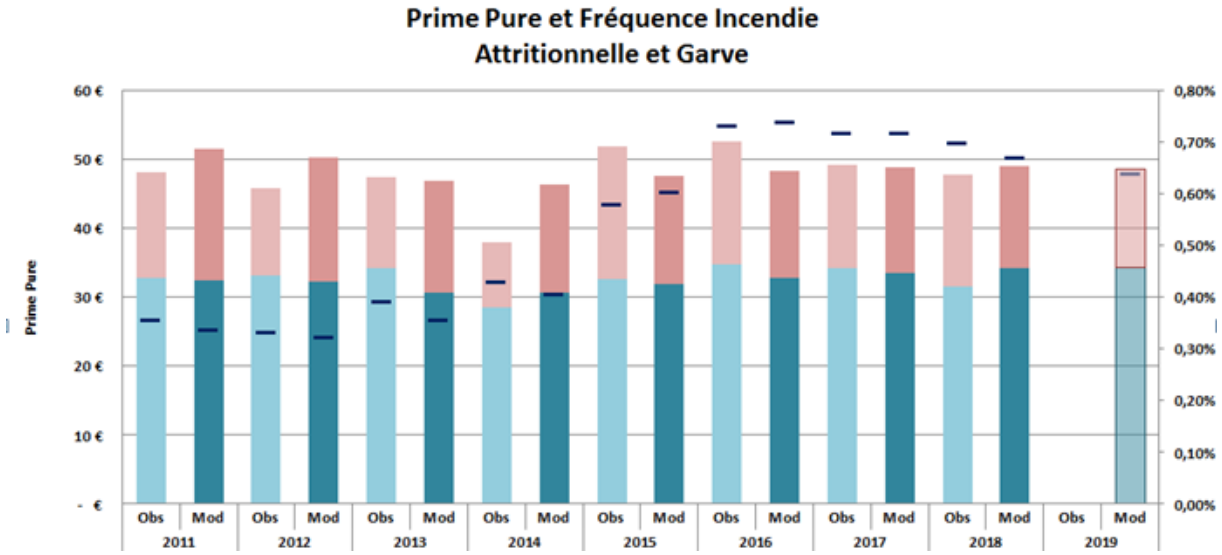


Figure 56 : Prime pure décomposée et fréquence en observée et modélisée en fonction de l'année d'exercice

Je souligne que le modèle de dépassement de seuil ainsi que le modèle de coût moyen grave n'ont pas intégré de composante temporelle comme variables explicatives. La tendance estimée de la prime pure grave est donc identique à celle de fréquence déjà montrée.

Ce traitement temporel nous interroge sur notre vision de la sinistralité attendue. Cette approche nous permet d'isoler l'effet temporel et ainsi d'obtenir une prime pure qui prend en compte le risque lié au risque et le temps de manière distingue. Sans cette donnée temporelle, la prime pure 2019 projetée serait identique à celle de 2018. Cette notion de temporalité nourrit actuellement des réflexions sur nos indicateurs de rentabilité, que nous détaillons dans la partie suivante.

5. Amélioration

Afin d'améliorer les modèles déjà construits, plusieurs voies semblent se dessiner :

- Le biais dans la modélisation du dépassement de seuil déjà évoqué plus tôt
- La construction d'un zonier Incendie Appartement et Maison qui simplifierait les modèles en réduisant le nombre de paramètres à estimer dans chaque modèle
- Le traitement de l'aspect temporel qui peut être amélioré
- Un modèle de prime pure grave distinct pour les appartements et les maisons. Cela entrainera surement une nouvelle révision des seuils de grave avec un seuil différent pour les appartements et les maisons

V. Impact sur les indicateurs de rentabilité

Lors de la décision des mesures tarifaires, on dispose de deux catégories d'indicateur afin d'analyser la rentabilité : les indicateurs économiques et les indicateurs techniques. Je définis ces indicateurs, puis j'expose les impacts sur nos indicateurs de rentabilité de l'implémentation de la répartition de la surcrête grave Incendie grâce au modèle de prime pure.

A. Visions techniques et économiques

1. Indicateurs économiques

a) Définition

Le S/C économique est le rapport entre les charges sinistres ultimes de l'année et les primes de l'année encaissées. Ce S/C prend donc en compte tous les évènements survenus dans l'année pouvant impacter les charges sinistres (évènements climatiques, sinistres graves,…). Les charges sinistres ultimes couvrent les indemnités versées aux assurés mais également les charges non encore versées, les sinistres survenus mais dont l'assureur ignore encore l'existence, les sinistres connus mais non encore indemnisés et les sinistres dont le montant va encore évoluer.

Le ratio combiné économique est un indicateur de rentabilité de l'exercice pour l'entreprise. Le ratio combiné prend en compte les primes encaissées pendant l'exercice, les charges ultimes estimées pour les sinistres survenus au cours de l'année, le poids de la réassurance et également les frais liés à l'exploitation de la société sur l'exercice.

b) Répartition de la charge économique

Les charges ultimes et les primes pures sont fournies par une équipe externe à la maille segments de marché et garanties. Il incombe alors à l'équipe tarification de réallouer les charges à la maille produit, garantie et même critère de risque.

L'étude sur la liquidation par critère de risque présenté en II permet d'affiner notre clé de répartition. On avait précédemment une clé de répartition par marché, on dispose désormais d'une clé par produit, garantie et par critère de risque pour le produit principal.

L'ancienne méthode utilisée consistait à appliquer un unique coefficient d'ultimisation sur l'ensemble des sinistres ouverts et clos.

Figure 57 : Ancienne méthode d'ultimisation

Dorénavant, on est capable de répartir les charges sur les populations suivantes et garanties suivantes:

8 Populations	
Produit 410 Prop. De Maison	Produit 410 Prop. d'Appart.
Produit 410 Loc. de Maison	Produit 410 Loc. d'Appart.
Produits 410 Autres	
Produit PNO	
Autres produits	

Figure 58 : Segmentation en 8 populations

Pour chaque garantie, on applique aux sinistres ouverts un coefficient de passage permettant de retrouver la charge économique à la garantie fournie à laquelle on doit se raccorder à la maille marché.

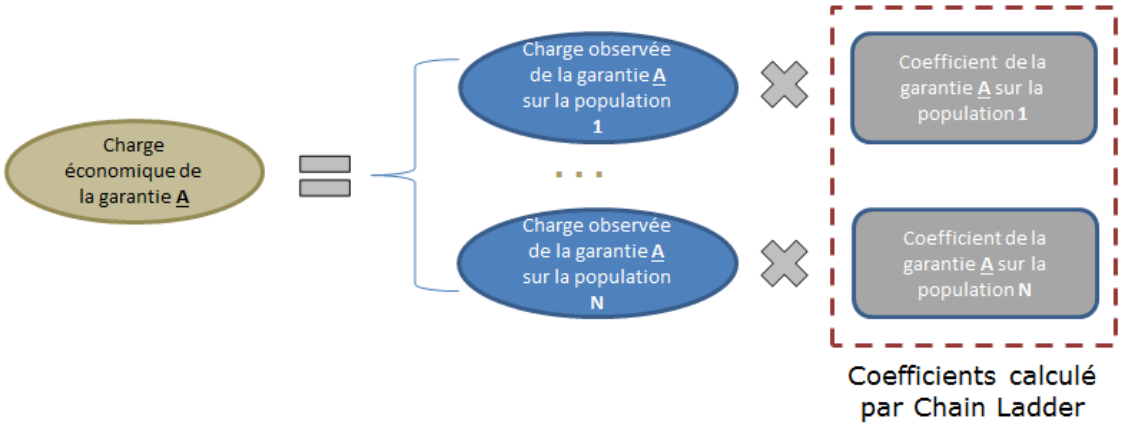


Figure 59 : Nouvelle méthode d'ultimisation

2. Indicateurs techniques

a) Définition

Le S/C technique est le rapport entre les charges techniques de l'année et les primes encaissées de l'année. La charge technique correspond à la charge de l'année attendue a priori. C'est-à-dire que la charge technique ne prend pas en compte tous les évènements particuliers de l'exercice pouvant impacter la charge sinistre. C'est la charge estimée correspondant à une année standard en termes de sinistralité. La charge technique peut être obtenue à partir de la prime pure technique.

Le ratio combiné technique est calculé à partir des primes encaissées, de la charge technique, de la réassurance et des frais d'exploitation de l'entreprise.

b) Répartition de la charge technique

A disposition de l'équipe tarification MRH, il y a actuellement trois charges techniques :

- Une charge technique climatique
- Une charge technique CATNAT
- Une charge technique autres garanties

Pour obtenir la charge technique par observation, on applique un coefficient de passage sur ces trois regroupements de garanties au coût observé. De cette façon, on retrouve la charge technique fournie en amont.

On considère la sinistralité attritionnelle observée comme suffisamment robuste et proche de la sinistralité attritionnelle attendue pour ne pas la retraiter. Seule la partie grave est donc redistribuée pour obtenir la charge technique par observation. La répartition de la surcrête se schématise de la façon suivante :

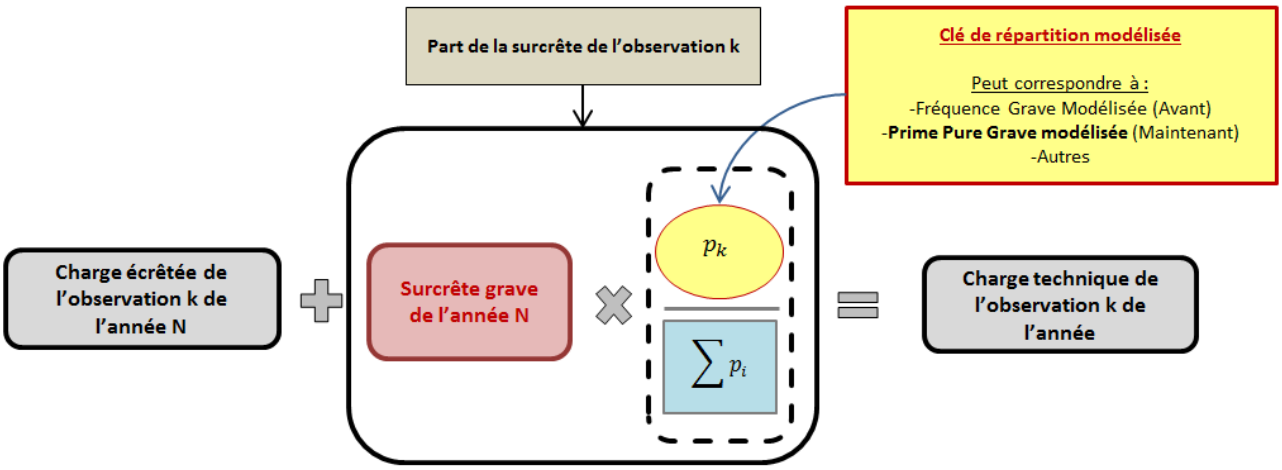


Figure 60 : Schématisation du calcul de la charge technique

De cette manière, la charge technique d'une observation est donc strictement positive. En l'absence de sinistre, l'observation récupère à minima sa part de surcrête grave de l'année en cours. Une observation avec un sinistre grave Incendie a une charge technique de 150 000€ à laquelle s'ajoute sa part de surcrête.

Ce processus de redistribution de la surcrête est réalisé pour les garanties principales : Responsabilité Civile, Dégâts des Eaux, Dommages Electriques, Vol,... Pour l'ensemble de ces garanties hors Incendie, la répartition se fait via un modèle de fréquence grave.

La prime pure grave modélisée a pour objectif de servir également de clé de répartition. Cette réallocation de la surcrête se faisait anciennement par un modèle de fréquence grave uniquement. Cela semblait insuffisant notamment sur les grands risques (plus de 10 pièces principales, grandes dépendances,…). La composante de coût moyen grave ajoutée permet de réallouer de manière plus juste la part de charge grave attendue.

Ce sont ces charges techniques alors obtenues qui nous permettent de calculer nos indicateurs de rentabilité par critère de risque. On propose par la suite de visualiser les impacts liés au choix de la clé de répartition.

B. Indicateurs à la Garantie Incendie

On cherche à visualiser les indicateurs de rentabilité par critère sur la garantie Incendie. On représente 5 quantités :

- Le **S/C économique moyen 4 ans 2015-2018**
- Le **S/C technique 2018** obtenu par répartition de la surcrête grave avec le **modèle de fréquence grave**
- Le S/C technique moyen 4 ans 2015-2018 obtenu par répartition de la surcrête grave avec le modèle de fréquence grave
- Le **S/C technique 2018** obtenu par répartition de la surcrête grave avec le **modèle de prime pure grave**
- Le S/C technique moyen 4 ans 2015-2018 obtenu par répartition de la surcrête grave avec le modèle de prime pure grave

Si le modèle de répartition est efficace, le S/C technique est sensé avoir une tendance identique au S/C économique moyen. L'économique 4 ans considère la sinistralité sur 4 ans et est donc relativement proche de ce que l'on pourrait attendre a priori comme rentabilité.

On s'intéresse à la rentabilité de l'Incendie en fonction de la qualité juridique et du type de bien.

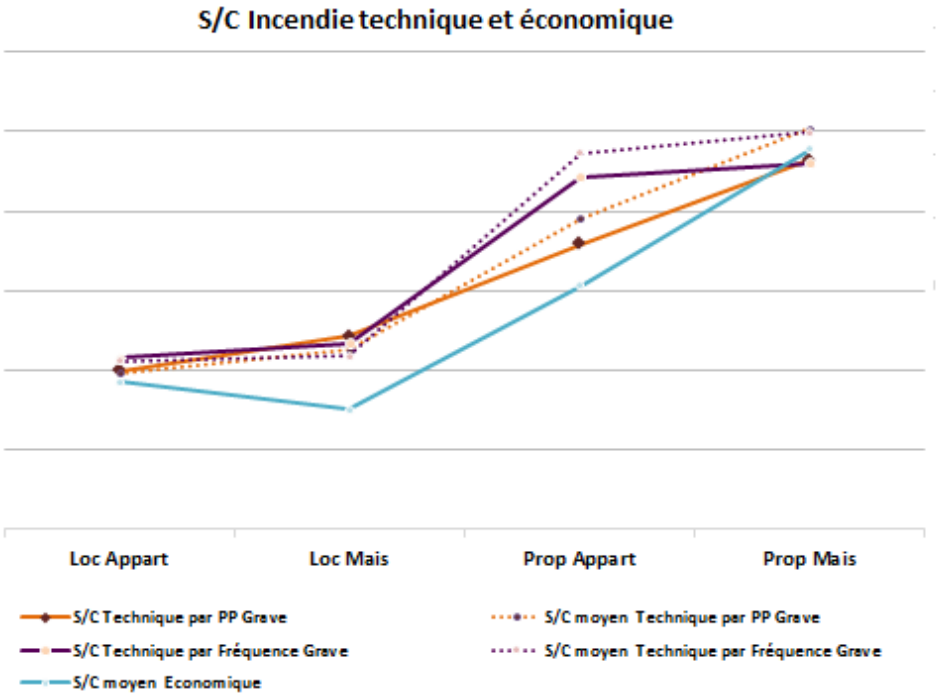


Figure 61 : S/C en fonction de la qualité juridique et du type de logement

L'utilisation de la prime pure grave favorise les appartements. L'effet de ce critère est à la baisse dans le modèle de fréquence mais également dans le modèle de coût moyen grave. Il est donc naturel de constater une amélioration sur ce type de risque. La prime pure améliore grandement les résultats sur les propriétaires d'appartements avec une baisse de -17pts et une baisse moindre de -3.3pts sur les locataires d'appartements. Par effet de bascule, les résultats des maisons sont détériorés : +0.9pt pour les propriétaires et +1.9pts pour les locataires.

A première vue, la répartition par prime pure grave semble plus efficace. Le S/C technique 2018 et moyen obtenus ainsi semblent plus proches du S/C économique moyen.

Je parlais en amont que la répartition par fréquence grave avait quelques difficultés à lisser la charge extrême sur les grands risques. Pour vérifier cela, on observe les S/C Incendie de la maison en fonction du nombre de pièces.

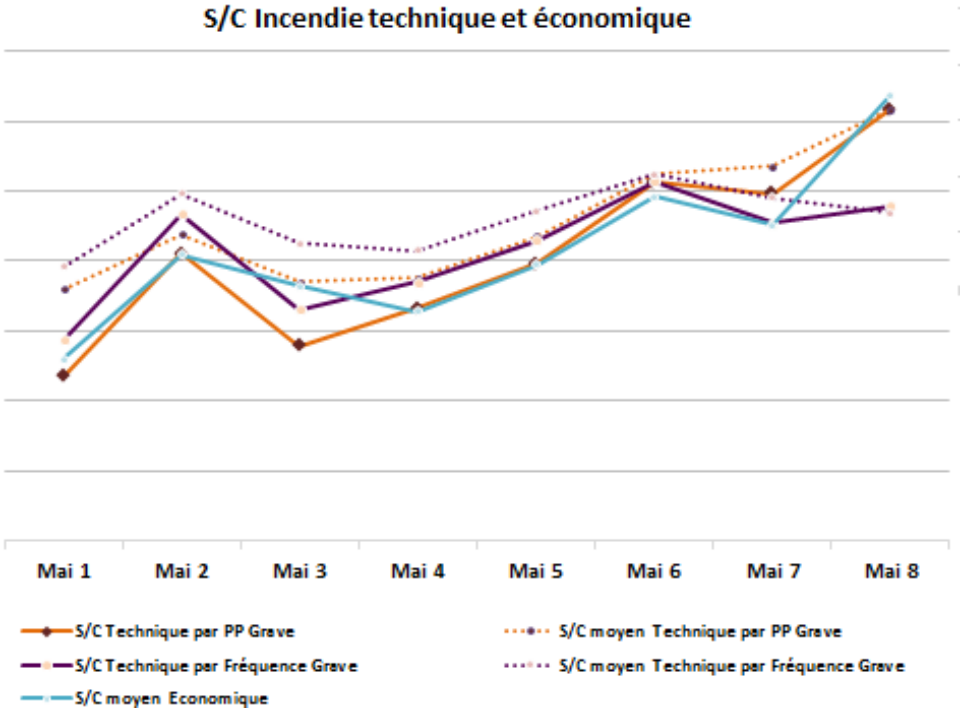


Figure 62 : S/C en fonction du nombre de pièces sur les maisons

Pour les maisons de 8 pièces et plus, le S/C technique obtenu par fréquence grave est en dessous de quasiment 30pts des 2 autres S/C. Sur ce critère de risque aussi, le S/C technique issu de la prime pure grave semble plus robuste. Sa tendance semble quasi identique à l'économique. La répartition par prime pure transfert une partie de la charge qui est allouée avec la fréquence grave aux maisons de 5 pièces et aux maisons de plus de 7 pièces. Cette approche semble corriger en partie les défauts liés à la fréquence grave observés sur les grands risques.

La présence de pièce supérieure à 40m² est également caractéristique des grands risques.

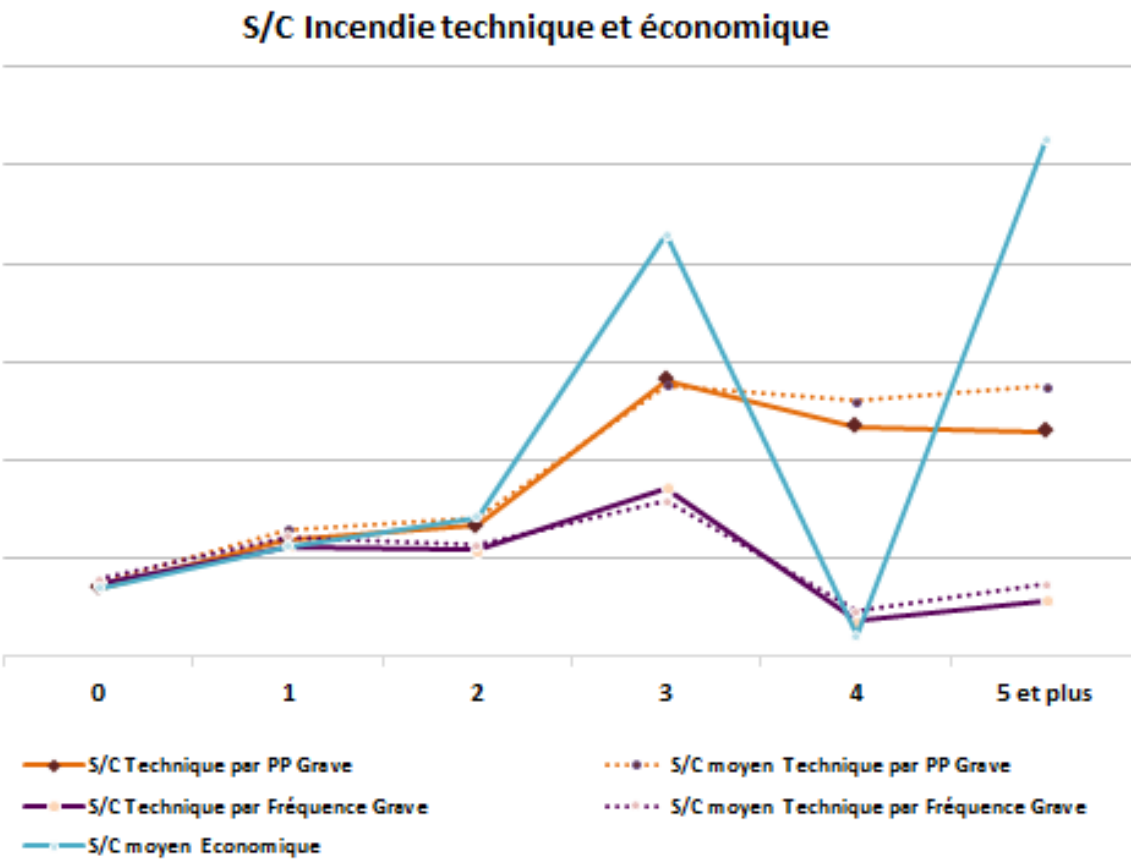


Figure 63 : S/C en fonction du nombre de pièces de plus de 40m² sur les maisons

Le S/C technique par prime pure grave semble ici bien plus effcicace que celui par fréquence grave. Au-delà de 2 pièces supérieures à 40m², le technique lié à la fréquence donne un S/C de 107% alors que l'économique moyen et le technique sont aux alentours des 180%. Le technique par prime pure grave apparait comme plus cohérent vis-à-vis de l'économique.

L'ensemble des constats semble favoriser l'utilisation de la prime pure grave afin de piloter la sinistralité technique. Cette approche semble bien mieux appréhendée la sinistralité extrême, notamment sur les grands risques. A la maille garantie, l'impact lié au changement de clé de répartition semble important. Le S/C technique au critère est dorénavant plus juste. L'effet est plus mesuré sur les résultats « Toutes Garanties ».

C. Indicateurs Toutes Garanties

On représente maintenant l'impact du changement de répartition de surcrête Incendie grave avec l'ensemble des garanties. La sinistralité technique des autres garanties reste inchangée. On trace les 5 quantités suivantes:

- Le **Ratio Combiné économique moyen 4 ans 2015-2018**
- Le **Ratio Combiné technique 2018** obtenu par répartition de la surcrête grave avec le **modèle de fréquence grave**
- Le Ratio Combiné technique moyen 4 ans 2015-2018 obtenu par répartition de la surcrête grave avec le modèle de fréquence grave
- Le **Ratio Combiné technique 2018** obtenu par répartition de la surcrête grave avec le **modèle de prime pure grave**
- Le Ratio Combiné technique moyen 4 ans 2015-2018 obtenu par répartition de la surcrête grave avec le modèle de prime pure grave

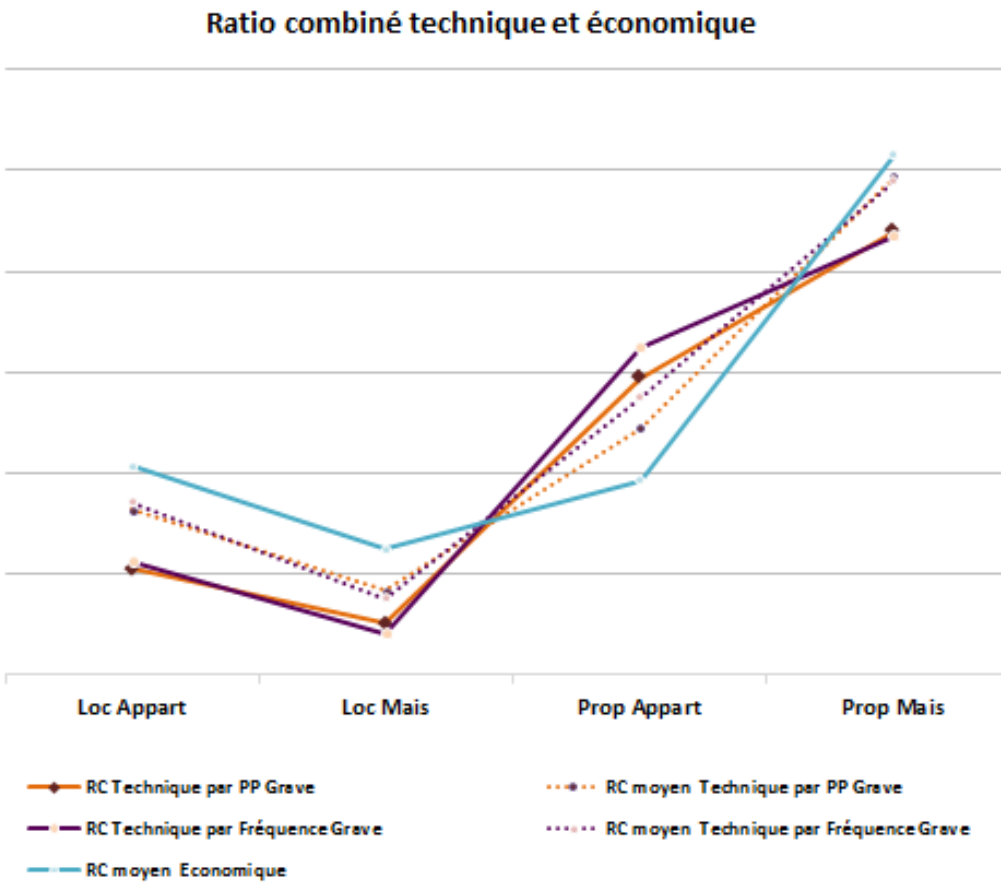


Figure 64 : Ratio Combiné en fonction de la qualité juridique et du type de logement

Sur la qualité juridique croisée avec le type de bien, on constate des écarts bien plus faibles en Toutes Garanties qu'à la garantie Incendie. Le Ratio Combiné technique 2018 des appartements est amélioré d'0.8pt avec pour contrepartie une dégradation de 0.3pt sur les maisons. A cette maille, l'utilisation d'une approche ou d'une autre ne semble pas impacter grandement sur les résultats techniques. Dans ce cas, seul le Ratio Combiné technique 2018 des propriétaires d'appartements semble avoir un effet remarquable : -1.4pt.

D. Conclusion

L'usage de la fréquence grave pour redistribuer la sinistralité extrême était initialement dû à la complexité de modéliser un coût moyen grave. Leur faible fréquence rend leur modélisation complexe. Le choix de traiter le coût moyen grave de façon classique par un GLM est certes discutable mais il apporte une information complémentaire qui précise notre clé de répartition de la surcrête Incendie.

Comme nous avons pu l'observer en comparant les résultats économiques et techniques, l'utilisation de la prime pure grave semble plus efficace pour lisser la sinistralité extrême que l'utilisation de la fréquence grave. La justesse des indicateurs obtenus par ce lissage est primordiale afin de prendre les mesures tarifaires les plus adéquates.

Le calcul de la charge technique au critère est perfectible. Deux axes d'amélioration semblent se dégager :

- Améliorer le modèle de prime pure grave
    - Correction du biais lié au déséquilibre du nombre de sinistres graves dans le modèle de dépassement de seuil
    - Utilisation des zoniers Incendie en remplacement des données exogènes
    - Modèles de prime pure grave distincts pour appartement et maison

- Revoir notre approche de la charge technique.
    - Le coefficient de raccordement appliqué à l'Incendie est identique à celui des autres garanties, excepté des deux déjà citées (climatique et CATNAT). L'inconvénient de ce raccordement sur le regroupement « autres garanties » est que l'on ne peut pas retrouver un niveau de charge attendue à priori, tel que la définition de technique l'implique. En effet, pour une année avec une forte sinistralité Incendie observée, la charge technique Incendie se verra dégradée car celle-ci occupera un poids plus important qu'à l'accoutumée au sein du regroupement « autres garanties ».
    - L'obtention d'une charge technique à la garantie et non plus seulement pour la CATNAT, le climatique et les autres garanties, permettrait de régler ce problème.

Conclusion

Les sinistres extrêmes, de par leur rareté et leur sévérité, sont sources d'une grande incertitude. Cela est d'autant plus vrai sur l'Incendie en Habitation où la prime pure de cette garantie représente une grande part de la prime pure globale du segment de marché.

Il a d'abord été question de redéfinir ce qu'était un sinistre grave en Incendie. Je me suis basé sur trois approches pour faire mon choix. La première consistait à constater le nombre de sinistres et la charge qui seraient alors qualifiés de graves pour certains seuils. Je recherchais un seuil qui me permette de traduire l'aspect rare et intense d'un évènement grave. Les deux autres méthodes sont visuelles et s'appuient sur des résultats de la théorie des Valeurs extrêmes. Le *Hill Plot* m'encourageait à choisir un seuil de 350 000€ ce que déconseillait mon approche descriptive. Cela représentait trop peu de sinistres pour modéliser leur fréquence et leur coût moyen de manière fiable. Le *Mean excess Plot* quant à lui semblait concorder avec l'analyse descriptive ce qui a permis de fixer le seuil de grave Incendie.

Le pilotage de la sinistralité en Habitation a été revu avec, au début, le traitement des charges ultimes à une maille plus fine que ce qui été fait précédemment. La méthode de Chain Ladder est suffisamment robuste pour effectuer les calculs d'ultimes par population pour la majorité des cas. Cependant, il ne prend pas en compte correctement les sinistres qui se développent de façon brusque comme cela peut être le cas sur des sinistres extrêmes, notamment en Responsabilité Civile ou en Incendie. La méthode de Chain Ladder appliquée à la garantie CATNAT peut également poser problème en raison de la déclaration tardive des sinistres sécheresse. J'ai pu tout de même démontrer que le développement de la charge diffère selon le produit mais aussi selon le type de bien et la qualité juridique au sein d'une même garantie. Les charges ultimes obtenues via cette étude rentrent désormais dans le pilotage de la charge économique qui enrichit les réflexions autour des mesures tarifaires à venir.

Après avoir retraité la vision économique de la sinistralité, on s'est intéressé à la vision technique. Cette sinistralité représente la charge attendue a priori. Pour piloter cette charge technique, la surcrête grave est redistribuée selon le profil de risque. Précédemment, la surcrête était redistribuée avec un modèle de fréquence grave. Ce modèle de fréquence grave a tout d'abord été remis à jour en le décomposant en deux sous-modèles : un modèle de fréquence Incendie et un modèle de dépassement de seuil. Afin de challenger les méthodes statistiques classiques, j'ai entrepris de modéliser le dépassement de seuil à l'aide d'un GBM. La régression logistique a cependant donné des résultats plus satisfaisants.

Afin d'obtenir un modèle de prime pure grave, nous avons construit deux modèles de coût moyen : pour les sinistres graves et pour l'ensemble des sinistres Incendie. L'intérêt était ici d'avoir un coût moyen grave segmenté. Pour cela, on a choisi une approche par modèle linéaire généralisé.

L'utilisation de données exogènes INSEE nous a permis d'expliquer le risque spatial et de construire une ébauche de zonier Incendie. Cela a permis d'aboutir à la construction de

nouveaux zoniers à partir de ces données exogènes. Par la suite, on pourra utiliser ces zoniers à la place de données INSEE afin de simplifier nos futures études.

Le modèle de prime pure grave étant maintenant complet, on a traité la surcrête grave Incendie avec cette prime pure mais aussi avec le modèle de fréquence grave. On a alors été en mesure de calculer les indicateurs de rentabilité technique à la garantie avec ces deux approches. Le S/C technique issu de la répartition par la prime pure grave se comporte de manière similaire au S/C économique moyen sur 4 ans. Cette approche est désormais celle que l'équipe tarification utilise pour piloter la charge technique Incendie et celle qui nous permet de faire nos recommandations de revalorisation.

Ces évolutions dans le pilotage de la sinistralité permettent de calculer des indicateurs de rentabilité plus robustes et qui représentent plus fidèlement le risque réel. Ces derniers sont primordiaux lors de la décision des mesures tarifaires à apporter. Les travaux présentés dans ce mémoire aident ainsi à la prise de décision lors des revalorisations tarifaires annuelles.

Bibliographie

Le zonier en tarification IARD : approche comparative de deux techniques de construction d'un critère de segmentation géographique en assurance habitation, Claudine Ferrier, Mémoire d'actuariat, 2016
Disponible sur : http://www.ressources-actuarielles.net

Modèle de provisionnement des sinistres graves et son allocation économique aux différentes succursales d'AXA Corporate Solutions, Duc Hien VU, Mémoire d'actuariat, 2015
Disponible sur : http://www.ressources-actuarielles.net

Nouvelle modélisation du risque extrême dans la tarification de la garantie incendie en assurance multirisques habitation, Fatima-Zahra NAJI, Mémoire d'actuariat, 2016
Disponible sur : http://www.ressources-actuarielles.net

De la qualité d'un score de classification, Arthur CHARPENTIER, 2010
http://freakonometrics.hypotheses.org/2078

Utilisation de la théorie des valeurs extrêmes dans le contexte solvabilité 2, Frédéric PLANCHET, 2016
http://www.ressources-actuarielles.net/EXT/ISFA/fp-isfa.nsf/0/0b9df464e9543283c1256f130067b2f9/$FILE/Seance1.pdf

Package 'gbm' :
https://cran.r-project.org/web/packages/gbm/gbm.pdf

Page Wikipédia du Gradient Boosting :
https://en.wikipedia.org/wiki/Gradient_boosting

A Gentle Introduction to the Gradient Boosting Algorithm for Machine Learning :
https://machinelearningmastery.com/gentle-introduction-gradient-boosting-algorithm-machine-learning/

Documentation sur les procédures SAS :
https://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#genmod_toc.htm
https://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#corresp_toc.htm
https://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#fastclus_toc.htm
https://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#logistic_toc.htm

Site de l'Institut Mathématiques de Toulouse, « *Agrégation de modèles* » :
https://www.math.univ-toulouse.fr/~besse/Wikistat/pdf/st-m-app-agreg.pdf

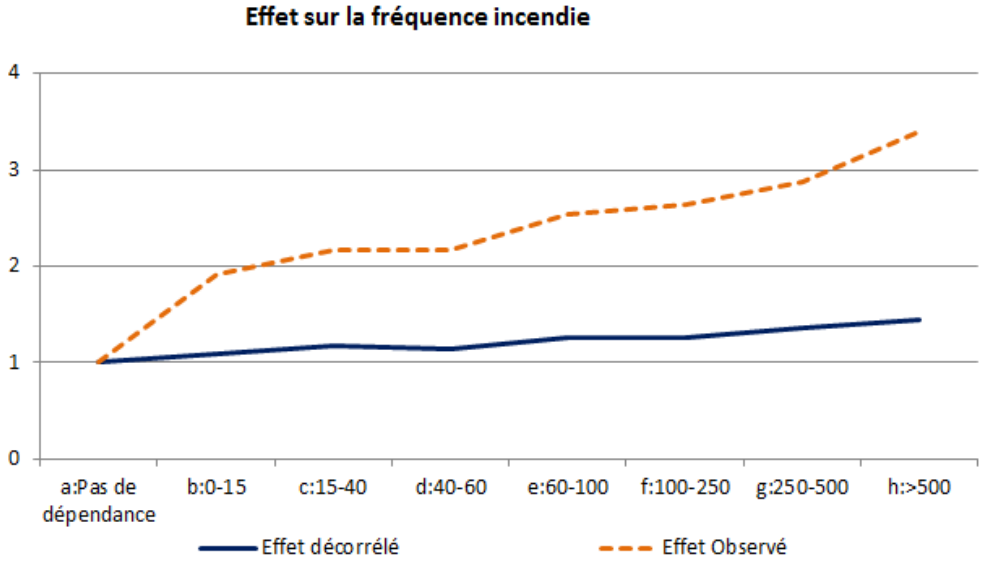
Annexe

Corrélations des variables INSEE catégorisés

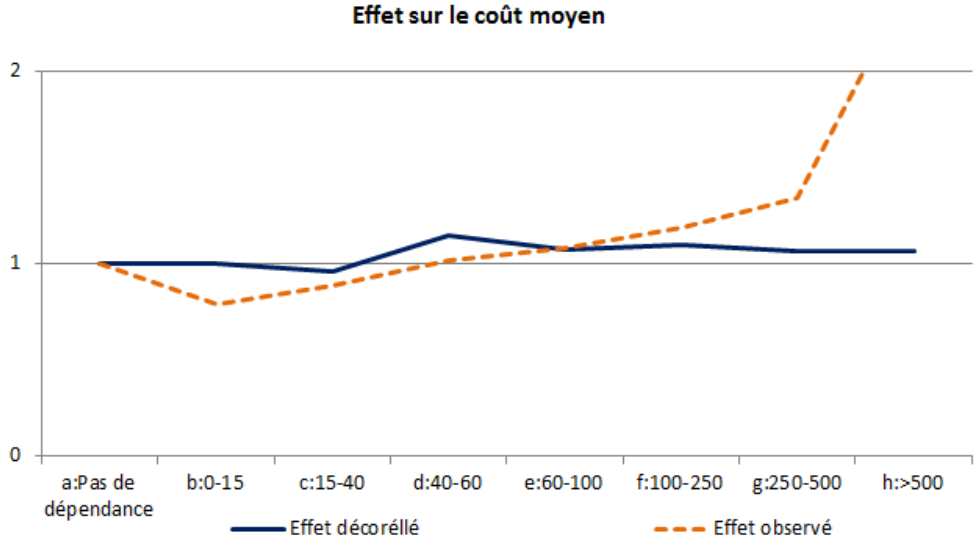
V_Cramer	Variable 1	Variable 2
1,00	PART_POP_ETR	PART_POP_FR
0,93	PART_POP_MARIEE	PART_POP_NONMARIEE
0,93	PART_1564_ACTIF	PART_1564_INACTIF
0,89	PART_APPART	PART_APPART_RP
0,87	POP	POP_15ansetplus
0,86	PART_MAISON	PART_MAISON_RP
0,85	POP	POP_1564
0,83	PART_APPART_RP	PART_MAISON_RP
0,81	POP_1564	POP_15ansetplus
0,80	NB_MEN	POP_15ansetplus
0,80	PART_APPART	PART_MAISON
0,79	PART_APPART	PART_MAISON_RP
0,77	PART_APPART_RP	PART_MAISON
0,76	NB_MEN	POP
0,76	PART_RP_LOC	PART_RP_PROP
0,75	NB_Logement	NB_MEN
0,74	PART_RPMAISON_AVT19	PART_RP_AVT1919
0,74	NB_MEN	POP_1564
0,72	PART_MEN_FAMILLE	PART_MEN_SEUL
0,70	NBP_MOY_RP	PART_5PP_RP
0,66	NB_Logement	POP_15ansetplus
0,63	PART_RPMAISON_0612	PART_RP_0612

0,63	NB_Logement	POP
0,62	PART_RPMAISON_9105	PART_RP_9105
0,61	PART_1P_RP	PART_RP_INF30m2
0,61	NBP_MOY_MAISON_RP	NBP_MOY_RP
0,61	NB_Logement	POP_1564
0,61	NBP_MOY_MAISON_RP	PART_5PP_RP
0,60	PART_RPMAISON_2045	PART_RP_2045

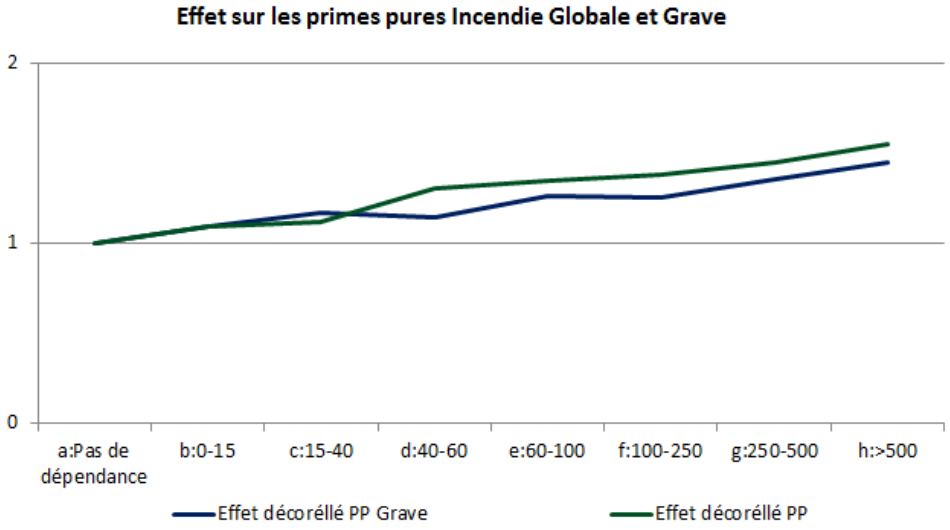
Effet de la superficie de la dépendance sur la fréquence incendie



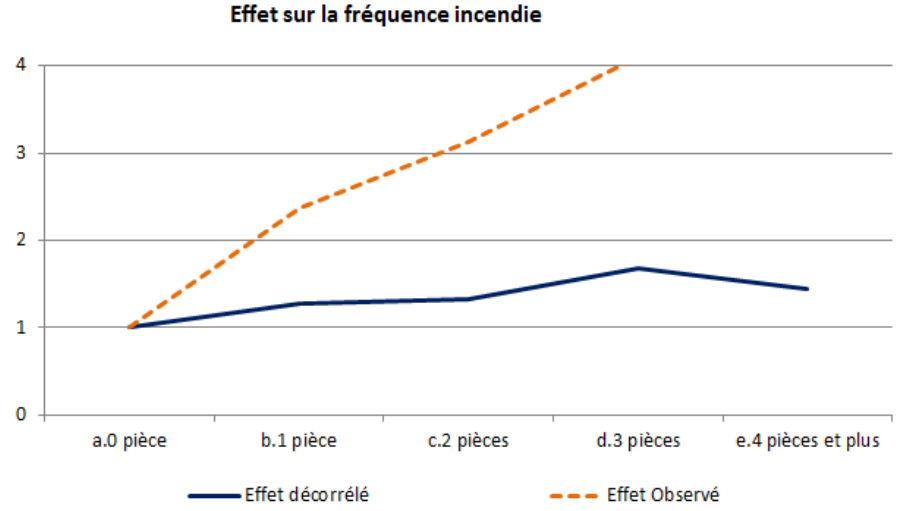
Effet de la superficie de la dépendance sur le coût moyen incendie



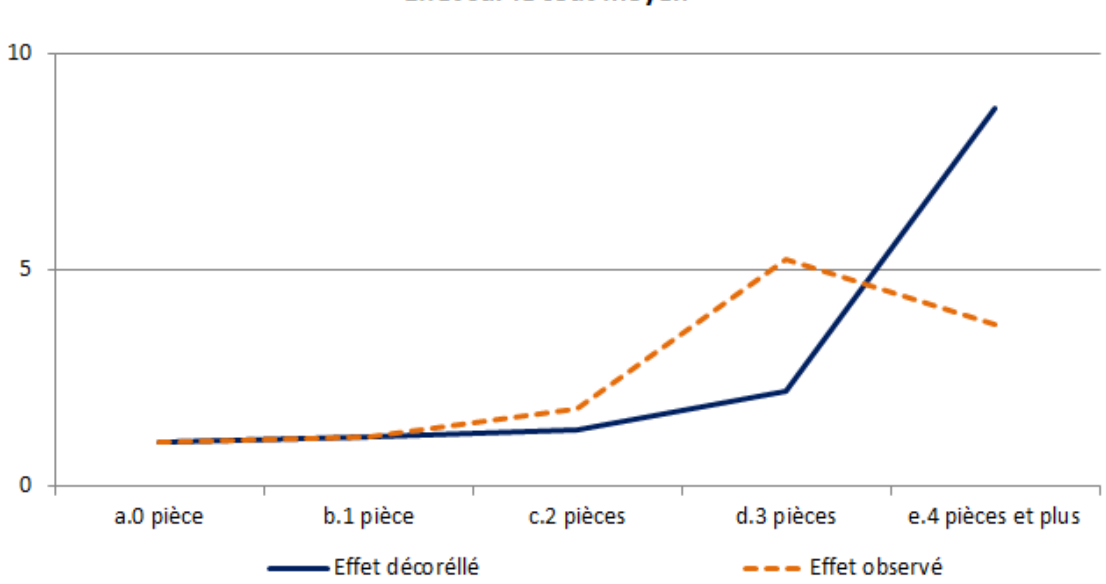
Effet de la superficie de la dépendance sur la prime pure incendie et la prime pure grave incendie



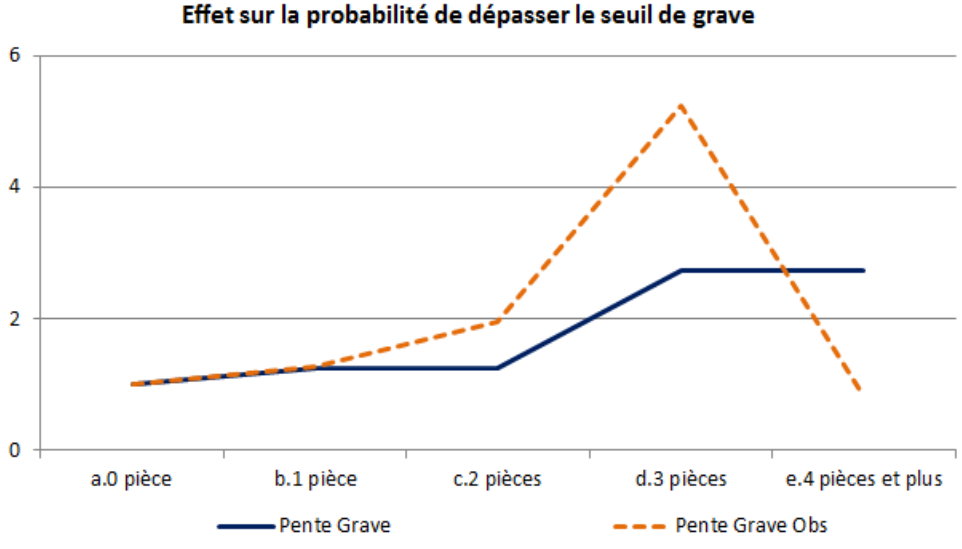
Effet du nombre de pièces de plus de 40m² sur la fréquence incendie



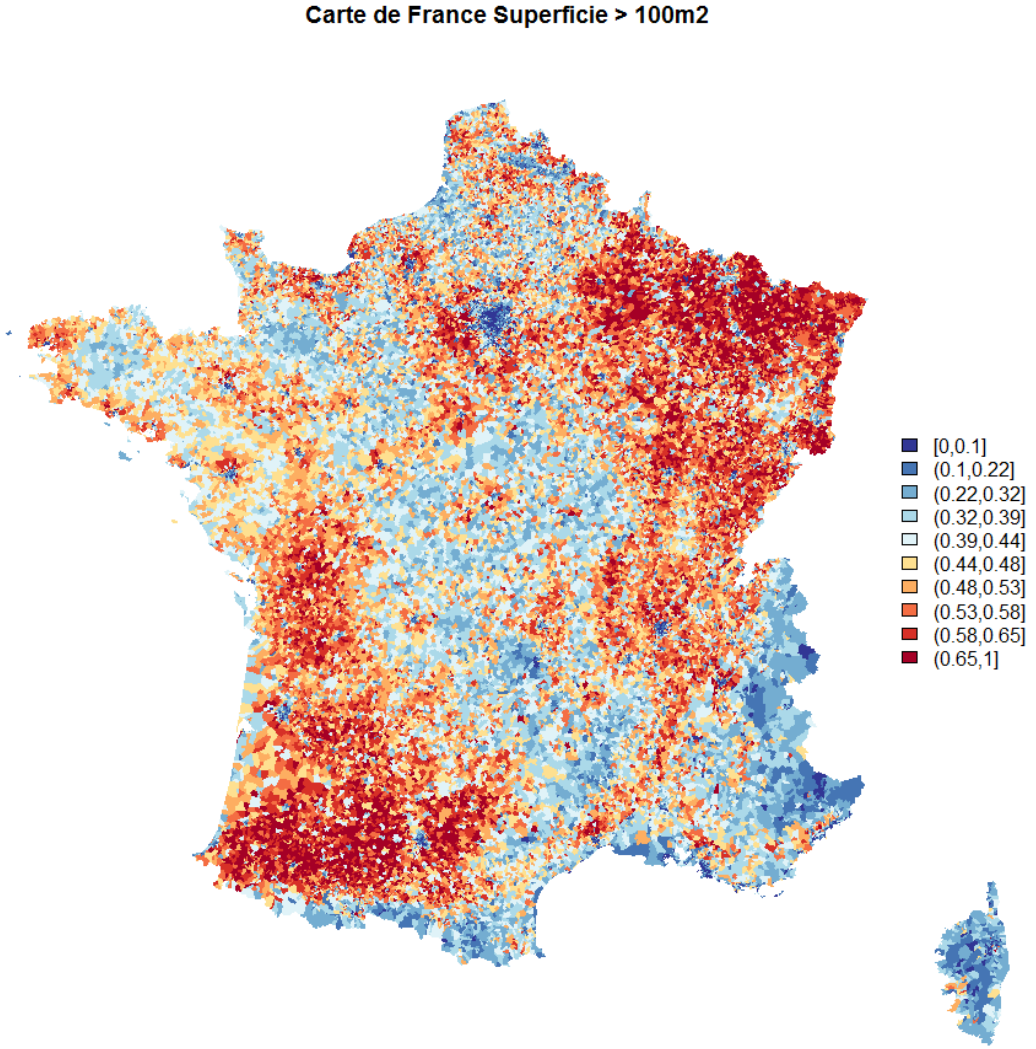
Effet du nombre de pièces de plus de 40m² sur le coût moyen incendie



Effet du nombre de pièces de plus de 40m² sur la probabilité de dépasser le seuil



Carte de la concentration en logements de plus de 100m²



Effet de la concentration en logements de plus de 100m² sur la prime pure