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

le 8 septembre 2023

Par : Krispanga NIKIEMA
Titre : Impact de l'ajout d'*Open Data* dans la modélisation de la garantie climatique en MRH.

Confidentialité : Oui (Durée : 2 ans)

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

Membre présent du jury de l'Institut des Actuaires :
David DUBOIS
Florence PICARD
Sonia GUELOU
Signature :

Entreprise :
AXA FRANCE IARD
Signature :

Membres présents du jury de l'EURIA :
Pierre AILLIOT

Directeur de mémoire en entreprise :
Brice DECAUX
Signature :

Invité :

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

Signature du responsable entreprise :

Signature du candidat :

EURIA
EURo Institut d'Actuariat

6, avenue le Gorgeu
CS 93837
29238 Brest Cedex 3

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

Résumé

Face aux défis posés par le changement climatique et à l'essor des *Open Data*, ce mémoire se concentre sur l'impact de l'intégration de ces données dans la modélisation de la garantie climatique des propriétaires de maison en MultiRisques Habitation (MRH). Cette garantie prend en charge les sinistres issus d'événements climatiques tels que les tempêtes, inondations, grêles, et autres événements ne faisant pas l'objet d'un arrêté de catastrophe naturelle. Les tempêtes sont à l'origine de la majorité des sinistres, dépassant les 60 % en nombre et en coût. Par ailleurs, le ratio Sinistres sur Cotisations (S/C) de cette garantie excède 180 % pour les propriétaires de maison, qui forment plus de 90 % du portefeuille. Il convient de noter que les catastrophes naturelles et la sécheresse sont couvertes par des garanties distinctes.

Compte tenu de l'impact du changement climatique sur la sinistralité des risques climatiques, cette étude vise à améliorer les performances du modèle, et ainsi obtenir de meilleurs ratios S/C. C'est dans cette optique que la base de Demandes de Valeurs Foncières (DVF), qui fournit les valeurs foncières des transactions immobilières en France entre 2017 et 2022, sera explorée. Cette étude de la base DVF vise à optimiser particulièrement le modèle de coût moyen. En effet, les bases de données internes de la MRH contiennent des informations sur les biens mobiliers des assurés, mais aucune donnée sur la valeur immobilière des habitations, que nous supposons plus pertinente, notamment dans le cas des sinistres liés aux tempêtes.

Dans le cadre de la modélisation, les Modèles Linéaires Généralisés pénalisés (GLM) de type LASSO, reconnus pour leur efficacité, seront privilégiés. De plus, une approche zonière, particulièrement pertinente en MRH sera adoptée. En effet, cette méthode permettra d'ajouter des informations géographiques essentielles pour une meilleure compréhension des risques associés à chaque zone. C'est à la suite de ce zonage que les données ouvertes seront ajoutées, ce qui permettra l'évaluation de leur impact sur les performances des modèles.

Par ailleurs, dans le but d'optimiser davantage la modélisation et d'apporter une dimension supplémentaire à l'analyse, l'étude envisage d'explorer le potentiel des forêts aléatoires, une méthode d'apprentissage statistique avancée. En effet, les forêts aléatoires présentent plusieurs avantages, notamment la capacité de traiter de nombreuses variables tout en garantissant une robustesse dans la modélisation.

Finalement, ce mémoire cherche à apporter une nouvelle perspective sur la modélisation de la garantie climatique en intégrant des données ouvertes, tout en combinant différents types de méthodologies.

Mots clefs: *MultiRisques Habitation, climatique, propriétaires de maison, données ouvertes, modélisation, zonier, Demandes de Valeurs Foncières, Modèles Linéaires Généralisés, Forêts aléatoires*

Abstract

Faced with the challenges of climate change and the rise of Open Data, this thesis examines the impact of integrating these data into the modelling of homeowners' climate cover in Home Multi-Risk Insurance (MRH). This guarantee covers claims arising from climatic events such as storms, floods, hails and other events not subject to a natural disaster decree. Storms account for the majority of claims, exceeding 60 % in number and cost. Furthermore, the claims to contributions ratio (S/C) for this coverage exceeds 180 % for homeowners, who account for more than 90 % of the portfolio. It should be noted that natural disasters and drought are covered by separate guarantees.

Given the impact of climate change on the claims frequency of climatic risks, this study aims at significantly improving the model's performance, and thus obtain better S/C ratios. In this context, the Real Estate Value Requests (DVF) database, which provides the real estate transactions in France between 2017 and 2022, will be explored. This study of the DVF database is aimed specifically at optimising the severity model. Indeed, the MRH's internal databases contain information on policyholders' personal property, but no data on the real estate value of policyholders' homes, which we assume to be more relevant, particularly in the case of storm-related claims.

In terms of modelling, we will prioritize Penalized Generalized Linear Models (GLM) of the LASSO type, known for their effectiveness. In addition, a zoning approach, particularly relevant in MRH will be adopted. Indeed, this method will make it possible to add geographical information, which is essential for a better understanding of the risks associated with each zone. It is following this zoning that Open Data will be added, allowing for the evaluation of their impact on the modelling.

Furthermore, to further optimise the modelling and add an extra dimension to the analysis, the study plans to explore the potential of random forests, an advanced statistical learning method. Indeed, this method has several advantages, including the ability to handle many variables while ensuring robustness in modelling.

In conclusion, this thesis seeks to provide a new perspective on modelling the climatic guarantee by integrating Open Data, while combining different types of methodologies.

Keywords: *Home Multi-Risk Insurance, climatic, homeowners, Open Data, modelling, zoning, Real Estate Value Requests, Generalized Linear Models, Random Forests*

Remerciements

Je tiens à exprimer ma profonde gratitude envers les personnes qui ont contribué de manière significative à la réalisation de ce mémoire. Leur soutien, leurs suggestions et leur disponibilité ont été d'une importance capitale pour mener à bien ce travail de recherche.

Tout d'abord, je voudrais remercier chaleureusement Brice DECAUX, mon tuteur professionnel, pour son encadrement régulier, sa disponibilité constante et sa précieuse contribution à travers ses nombreux conseils avisés. Sa présence tout au long de ce projet a été une source d'inspiration et d'apprentissage. Je lui suis reconnaissante pour sa patience et sa capacité à répondre à mes interrogations, ce qui a grandement enrichi ma réflexion.

Je souhaite également exprimer ma gratitude envers Isabelle RIVOALEN, ma tutrice mémoire, pour sa patience et sa disponibilité. Ses conseils éclairés ont été d'une grande valeur pour moi.

J'adresse mes sincères remerciements à Pierre AILLIOT, mon tuteur académique, pour sa bienveillance et son accompagnement. Son expertise a été d'une grande aide dans l'orientation de ma réflexion.

Un grand merci à Christophe GALLOIS, mon manager, et à toute l'équipe Affaires Nouvelles de la MRH pour leur bienveillance et leur soutien tout au long de cette période de recherche.

Ensuite, j'exprime ma reconnaissance envers l'équipe pédagogique de l'EURIA pour leur encadrement attentif et leur soutien tout au long de ma formation. Leur expertise et leurs conseils ont joué un rôle déterminant dans mon parcours académique.

Enfin, je tiens à remercier ma famille pour leur soutien indéfectible et leur encouragement tout au long de ce projet. Leur présence et leur soutien moral ont été une source de motivation essentielle.

Table des matières

Table des figures

Liste des tableaux

Introduction

Ce mémoire a été réalisé au sein de l'équipe Actuariat tarification Affaires Nouvelles et Remplacements (AN) de la direction MultiRisques Habitation (MRH) d'AXA France. Ce type d'assurance a pour but de protéger l'habitation, les biens, ainsi que la responsabilité civile des clients vis-à-vis des tiers. Il couvre des sinistres variés tels que le dégât des eaux, les évènements climatiques, les catastrophes naturelles, et bien d'autres. Le marché de l'habitation est hautement compétitif. Cette concurrence s'est intensifiée avec l'arrivée de nouveaux acteurs, tels que les bancassureurs, et avec l'entrée en vigueur de la loi Hamon en janvier 2015. En effet, cette loi permet aux assurés de résilier un contrat d'assurance après la première année d'échéance, leur donnant l'opportunité de profiter d'un tarif plus compétitif auprès d'un concurrent. Cette évolution législative a modifié la dynamique de production du produit MRH. Par conséquent, pour préserver les résultats sans compromettre la compétitivité de l'entreprise, il est nécessaire de revoir régulièrement le tarif du produit en mettant à jour les données de modélisation.

En France, les charges des sinistres associés aux évènements climatiques ont considérablement augmenté au cours de ces quarante dernières années. Cette augmentation est illustrée par le fait que les indemnisations sont passées d'un coût moyen annuel d'un peu plus d'1 milliard d'euros dans les années 1980 à environ 4 milliards d'euros aujourd'hui. Selon une étude exclusive publiée en octobre 2021 par les experts de France Assureurs (ou ex Fédération Française de l'Assurance), il est estimé que les coûts liés aux aléas climatiques pourraient atteindre 143 milliards d'euros au cours des 30 prochaines années en France. Cela représente une augmentation de 93 % par rapport aux coûts liés à la période entre 1989 et 2019. D'après cette même étude, le changement climatique contribuerait à un peu plus d'un tiers de cette hausse. Ces projections mettent en évidence l'impact croissant du changement climatique sur les coûts des sinistres liés aux événements climatiques en France et mettent en évidence les défis économiques auquels les assureurs devront faire face dans les décennies à venir. Cela explique l'urgence d'une action de la part des assureurs afin de limiter les pertes pouvant être dues à ce type de sinistres. Néanmoins, il convient de souligner que dans le contexte des assurances MRH, la modélisation de la garantie climatique présente une complexité particulière en raison de la nature aléatoire et difficilement prévisible des événements climatiques.

La garantie climatique du produit MRH « Ma Maison », lancé sur le marché en 2017, couvre les sinistres liés aux événements climatiques tels que les tempêtes, les inondations, la grêle, et tout autre événement climatique n'ayant pas fait l'objet d'un arrêté de catastrophe naturelle. Il est essentiel de noter que lorsqu'un sinistre lié à un évènement climatique est déclaré, une garantie climatique est automatiquement ouverte. Si plus tard l'évènement climatique concerné fait l'objet d'un arrêté de catastrophe naturel, une garantie catastrophes naturelles CATNAT est ouverte pour l'indemnisation des dommages qui y sont liés. Les tempêtes représentent la majorité des sinistres climatiques, totalisant plus de 60 % des incidents en nombre et en coût. Il est important de souligner que la sécheresse fait l'objet d'une garantie à part. Cela montre la complexité et la nuance de la couverture offerte par la MRH.

Il a été observé que le ratio Sinistres sur Cotisations (S/C) de cette garantie dépasse les 180 % pour les propriétaires de maisons, qui constituent plus de 90 % du portefeuille. Compte tenu de ces observations et de l'impact du changement climatique sur la fréquence et le coût des sinistres, il devient impératif de raffiner le modèle existant afin d'améliorer le ratio S/C des propriétaires de maisons.

Dans ce contexte, une modélisation de la prime pure sera effectuée, axée spécifiquement sur les propriétaires de maisons. Cette étude a pour objectif d'explorer l'impact de l'utilisation de données ouvertes sur l'optimisation des modèles de coût moyen et de fréquence. Si cette étude s'avère être fructueuse,

ces données ouvertes seront intégrées aux bases de données de l'équipe afin d'être testées sur d'autres garanties. Les données ouvertes qui seront étudiées sont la base de Demandes de Valeurs Foncières (DVF) et de la base des arrêtés Catastrophes Naturelles (CATNAT). Selon les sources gouvernementales, la base des arrêtés Catastrophes Naturelles est la base qui recense « les arrêtés interministériels de reconnaissances de l'état de catastrophe naturelle délivrés pour un ensemble de communes, un aléa et une période donnée, après examen des demandes de reconnaissance déposées par les maires des communes concernées». L'intérêt principal de l'exploitation de cette base de données est de déterminer si la publication d'un arrêté de catastrophe naturelle pour une commune spécifique peut améliorer la prédiction de la fréquence des sinistres climatiques. De plus, une exploration supplémentaire sera menée sur la base Météonet, qui compile les données météorologiques pour le Nord-Ouest et le Sud-Est de la France. La variable d'intérêt sera la vitesse des vents, avec un objectif d'amélioration similaire à celui de la base des arrêtés CATNAT.
La base DVF, qui recense les transactions immobilières en France entre 2017 et 2022, représente une ressource précieuse pour évaluer la valeur des maisons assurées. Notons que l'analyse de cette base vise à combler le manque de données internes de la MRH concernant la valeur immobilière des biens, qui est particulièrement pertinente dans le cas des sinistres liés aux tempêtes. En exploitant ces données conjointement avec l'expertise interne de la MRH, nous visons à améliorer la précision des résultats et à mieux gérer les risques associés aux sinistres climatiques pour les propriétaires de maisons.

Afin d'évaluer l'impact de l'utilisation des données ouvertes dans la modélisation de la garantie climatique du produit Ma Maison, la première partie présentera en détails les caractéristiques actuelles du marché de l'habitation, la méthode de modélisation (fréquence-coût), la présentation et le traitement des bases de données internes à l'équipe. La deuxième partie sera consacrée à la présentation et au traitement détaillé des bases de données ouvertes.
Dans la troisième partie, la garantie climatique sera modélisée pour les propriétaires de maison en utilisant les Modèles Linéaires Généralisés (GLM) pénalisés étant donné leur pertinence et leur utilisation courante dans l'équipe. Le modèle climatique actuel a d'ailleurs été élaboré à partir d'un GLM. Les performances de ces modèles seront comparées à celles d'une forêt aléatoire, une méthode d'apprentissage statistique, afin d'enrichir l'analyse. Finalement, dans une dernière partie, dans le cadre de la modélisation, il sera abordé la thématique du zonier. Cette notion doit être prise en compte en assurance Habitation. En effet, il existe des variables qui traduisent l'environnement géographique où évolue le contrat. C'est après la prise en compte de cette information que seront rajoutées les variables liées aux données ouvertes, ce qui nous permettra de clore l'étude en évaluant l'impact de ces données sur la modélisation.

Chapitre 1

Contexte de l'étude

1.1 L'assurance MultiRisques Habitation (MRH)

À noter : les chiffres et les graphiques qui sont présentés dans les sections 1.1.1 et 1.1.2 proviennent de données de France Assureurs (ex FFA).

1.1.1 Présentation du marché IARD en France

Le marché des **I**ncendies, **A**ccidents et **R**isques **D**ivers (**IARD**) regroupe les différentes couvertures mises en place pour protéger les clients [1] contre les dommages liés à leurs biens matériels et leur responsabilité envers des tiers. À l'inverse, les mutuelles et les assureurs vie se concentrent sur la protection des personnes.

En 2021, on compte 171 compagnies d'assurances non-vie [2] en France. Les assurances de biens et de responsabilité comptabilisent 63.2 milliards d'euros de cotisations, sur un total de 238 millards d'euros de cotisations pour l'ensemble des assurances françaises. Cela représente une hausse de 4.9 % par rapport à 2020. Sur ces 63.2 milliards d'euros de cotisations, 39.3 milliards concernent les assurances destinées aux particuliers.([France Assureurs, 2021b]
Cependant, il convient de souligner que cette évolution masque des croissances contrastées. En effet, si certaines catégories d'assurances telles que les assurances professionnelles ou agricoles ont connu une croissance dynamique, la croissance a été plus modérée pour les assurances destinées aux particuliers [3].
Ce phénomène s'explique par le fait que les assurances de particuliers ont été freinées par des engagements pris par ces dernières pour soutenir leurs clients lors de la crise de la Covid-19, notamment par le biais de réductions de primes. En opposition, les assurances professionnelles, agricoles, de la construction et des transports ont bénéficié d'une reprise économique.
Les deux plus grands types de couvertures présents sur le marché IARD sont l'assurance Automobile et l'assurance Habitation. Le graphique suivant illustre la répartition des cotisations dans le domaine des assurances de biens et de responsabilité en 2021.

1. Particuliers et professionnels
2. IARD et santé
3. Automobile, habitation, etc.

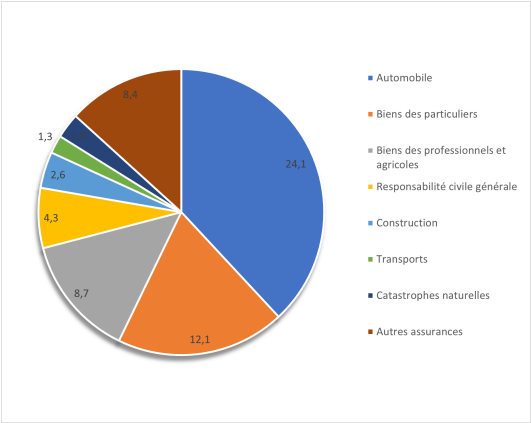


FIGURE 1.1 – Répartition des cotisations IARD en 2021 en Mds d'€. *Source des données : FFA*

En ce qui concerne la sinistralité globale en assurance de biens et responsabilité, la charge de prestations connaît une hausse globale de 1.3 % par rapport à 2020. Cette évolution masque encore des croissances contrastées en fonction des domaines : la hausse est marquée en automobile et en dommages aux biens des particuliers, tandis que les assurances de dommages aux biens des professionnels et agricoles connaissent un net recul.
Nous nous intéressons à l'évolution du ratio combiné au fil des années. Ce ratio est utilisé pour évaluer la performance financière d'une compagnie d'assurance en mesurant le rapport entre les coûts des sinistres et les primes encaissées :

$$\text{ratio combiné}_n = \frac{\text{prestations versées}_n + \text{frais généraux}_n + \text{commissions}_n + \text{réassurance}_n}{\text{cotisations encaissées}_n}$$

Où n correspond à l'année n.

À priori, un ratio combiné inférieur à 100 % indique une rentabilité potentielle, tandis qu'un ratio supérieur à 100 % indique une perte potentielle. Un ratio de 100 % représente un point d'équilibre. Néanmoins, d'autres facteurs pourraient influencer la santé financière globale d'une compagnie d'assurance. Nonobstant ceci, il est important pour les compagnies d'assurance de surveiller et de maintenir un ratio combiné optimal afin de garantir leur viabilité financière sur le long terme. De plus, le ratio combiné offre un avantage supplémentaire : il permet de comparer la performance d'une année à l'autre, offrant ainsi une perspective sur l'évolution de la rentabilité [1]. Le graphique ci-dessous illustre l'évolution du ratio combiné net de réassurance du marché IARD depuis 2016. En 2021, pour l'ensemble des assurances de biens et de responsabilité, le ratio s'établit à 96.9 %, soit une hausse de 0.2 % par rapport à 2020.

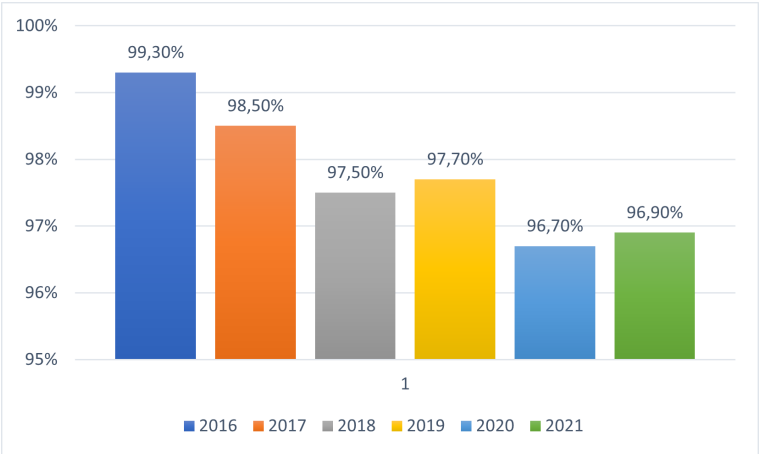


FIGURE 1.2 – Évolution du ratio combiné net de réassurance du marché IARD

Le ratio combiné de l'année 2020 est inférieur à celui des autres années. Cela signifie que, malgré les défis sans précédents posés par la pandémie de COVID-19, les assureurs IARD ont enregistré une

1. L'évolution de la charge peut être par exemple liée à l'évolution du ptf

meilleure performance cette année-là par rapport aux autres années de la période étudiée. Plusieurs raisons peuvent être avancées pour expliquer cette tendance. Notamment, la réduction des sinistres est un facteur clé. En effet, avec les confinements et les restrictions de déplacement imposées pour lutter contre la propagation du virus, certains types de sinistres, en particulier ceux liés à l'automobile, ont connu une baisse significative.

En 2022, en raison du contexte de guerre en Ukraine, l'inflation a augmenté, ce qui a eu un impact significatif sur les activités d'assurance non-vie. La conséquence principale de cette situation a été une augmentation des prestations à payer et des frais de gestion pour les assureurs, entraînant ainsi une dégradation du ratio combiné. Toutefois, il convient de noter que ce segment d'activité est principalement lié à des contrats d'une durée d'un an, ce qui permet aux assureurs de revaloriser leurs primes à l'échéance. Par conséquent, cette détérioration pourrait être de nature temporaire. [1]
Cependant, malgré la possibilité d'ajuster les primes d'assurance à la hausse, deux éléments limitent cette démarche. D'une part, la pression concurrentielle entre les assureurs exerce une contrainte sur l'augmentation des primes, car ils doivent rester compétitifs sur le marché. D'autre part, la pression politique peut également limiter les ajustements tarifaires, car les décideurs politiques cherchent à protéger les consommateurs et à éviter des hausses excessives qui pourraient impacter leur pouvoir d'achat.

Le sujet de ce mémoire étant centré sur une garantie de l'assurance **M**ulti**R**isques **H**abitation (**MRH**), nous allons à présent passer à l'examen de ce marché, qui constitue un segment spécifique du marché des assurances de biens et de responsabilité (IARD) en France.

1.1.2 Marché de l'assurance MultiRisques Habitation

Définition

Un contrat d'assurance Multirisques habitation (MRH) est un contrat comprenant plusieurs garanties conçues pour protéger le patrimoine familial [2] en cas de sinistre pour lesquels un foyer est victime ou responsable. Cette assurance est obligatoire pour les locataires et les copropriétaires. En revanche, elle n'est pas obligatoire pour les propriétaires occupant leur logement, ni pour les propriétaires non occupants dont le logement ne fait pas partie d'une copropriété.([EMH, 2018])

Les biens couverts par ce type de contrat d'assurance incluent :

- Les bâtiments (Locaux à usage d'habitation, dépendances, garages, caves...) ;
- Le mobilier personnel ;

En revanche, les bâtiments en cours de construction ne sont pas assurables par ce type de contrat.

Principales garanties proposées en MRH

Les principales garanties proposées par un contrat d'habitation sont :

- La Responsabilite Civile (RC) [3] : la garantie RC couvre les dommages causés par l'assuré, ses proches, ses biens ou ses animaux dans le cadre de sa vie privée ;
- La garantie incendie-explosion ;
- La garantie dégâts des eaux ;
- La garantie vol et la garantie vandalisme ;
- La garantie bris de glace ;
- La garantie catastrophes naturelles et technologiques ;
- La garantie tempête et autres événements climatiques : elle couvre obligatoirement les effets du vent dû aux tempêtes, ouragans et cyclones. Cette garantie est également accompagnée d'une protection contre la grêle, qui couvre les dommages causés par l'impact de la grêle sur les toitures, et enfin les dommages causés par le poids de la glace ou de la neige accumulée sur les toitures ([Ministère de l'Économie, des Finances et de la Relance, 2022]).

1. Voir évolution de l'indice FFB en annexes
2. Mobilier, habitation
3. Pour les co-propriétaires, l'obligation d'assurance se limite à la garantie responsabilité civile envers la copropriété, les voisins, les locataires, etc.

En complément de ces garanties, il est possible d'ajouter plusieurs options supplémentaires qui dépendront des offres proposées par chaque assureur. Ces options peuvent inclure la couverture des dommages électriques, une protection juridique ou même la prise en charge des frais liés à la casse des appareils nomades, etc.

Maintenant que l'assurance multirisques habitation (MRH) est définie, il convient d'expliciter l'état actuel de son marché.

Le marché de l'habitation

Après une période de ralentissement en 2020 en raison de la Covid-19 (baisse de souscription dû au confinement), le marché de l'assurance habitation se redresse en 2021. En effet, cette année-là, le nombre de contrats MRH atteint 44.4 millions, soit une progression de 2.2 % par rapport à l'année précédente. En parallèle, le nombre de logement n'a augmenté que de 1.1 %.
De plus, le montant total des cotisations, incluant toutes les garanties, enregistre une progression de 3.6 %, atteignant ainsi un total de 11.7 milliards d'euros. La prime moyenne atteint 263 € HT, contre 260 € en 2020.([France Assureurs, 2021b])

Parallèlement, il est important de noter que la fréquence et le coût des sinistres sur ce marché ont connu une augmentation significative en 2021. En effet, on a enregistré 3.6 millions de sinistres indemnisés, représentant une charge totale de 6 milliards d'euros, soit une hausse de 13 % par rapport à l'année précédente.

Ces données montrent une hausse des risques associés aux sinistres dans le secteur de l'assurance habitation, se rapprochant des niveaux observés avant la crise de la Covid-19. Bien que la fréquence des sinistres ait légèrement diminué de 1,5 % par rapport à 2019, le coût moyen a augmenté de 6,3 % depuis 2020. Cependant, il reste presque stable si l'on compare à 2019, avec une hausse minime de 0,8 %. Ces tendances devront être prises en compte par les assureurs lors de la tarification de leurs contrats MRH.

En dernier lieu, il est important de spécifier que le caractère concurrentiel du marché de l'habitation peut affecter son essor. Si l'entrée des bancassureurs sur le marché est une cause de croissance de ce phénomène, un deuxième exemple est la mise en place de la loi Hamon en 2015, permettant aux assurés de résilier leurs contrats d'assurance après une année de souscription. La possibilité de résilier facilement un contrat MRH après une année peut conduire à une plus grande volatilité des clients et à une augmentation de la concurrence entre les assureurs. Les assurés sont plus enclins à rechercher de meilleures offres et à changer de compagnie d'assurance pour obtenir des primes plus avantageuses ou des garanties plus adaptées à leurs besoins.

Le contexte du marché de l'assurance MRH ayant été présenté, l'attention peut désormais être portée sur l'offre d'AXA dans ce marché.

AXA France sur le marché de la MRH

Dans ce mémoire, l'orientation de la gestion des risques MRH se concentre exclusivement sur les **particuliers**. Les garanties principales intégrées dans le socle de garanties des produits de ce portefeuille correspondent à celles mentionnées précédemment, accompagnées d'options supplémentaires telles que la couverture de la casse des appareils nomades, des dommages aux appareils électriques et une protection juridique.
En 2021, avec un chiffre d'affaires de 986 millions d'euros, AXA se positionne comme le 4ème assureur sur le marché de l'assurance habitation. Si depuis 2014 ce chiffre d'affaires tourne autour du milliard d'euros, le nombre de contrats a baissé de 11 % (hausse de la concurrence). Cette diminution s'est principalement manifestée à partir de 2016 et ce, même si depuis 2017 AXA a adopté le Web comme nouveau canal de distribution en plus des canaux de distribution traditionnels (courtage, agences...).

L'étude présentée dans ce mémoire sera centrée sur les deux principaux produits de la MRH : «***Confort***» et «***Ma Maison***» :

— Le produit «*Confort*», qui était auparavant le produit MRH d'AXA, va progressivement être supprimé. Cette offre est désormais limitée aux souscriptions en Outre-mer (DOM) et à Monaco. Cependant, les assurés qui ont déjà souscrit ce produit ne sont pas obligés de le modifier. Il offre une gamme de garanties complètes, telles que celles mentionnées dans la partie 1.1.2. ;

— Le produit «*Ma Maison*», lancé en 2017, qui a été conçu pour offrir une plus grande flexibilité par rapport à *Confort*. Ce nouveau produit permet aux assurés de personnaliser leur contrat en fonction de leurs besoins spécifiques. En plus d'un ensemble de garanties de base, « Ma Maison » propose une variété d'options parmi lesquelles les assurés peuvent choisir pour adapter leur contrat selon leurs préférences : par exemple, les couvertures pour les dommages électriques, le bris de glace, le vol et le vandalisme deviennent des options pouvant être souscrites séparément. De plus, la couverture contre le gel est automatiquement incluse dans la couverture des dommages causés par les eaux (ou garantie dégats des eaux). Cette approche permet à AXA de proposer des primes plus adaptées aux risques spécifiques auxquels sont exposés les assurés.

Aujourd'hui, plus d'un quart des contrats du portefeuille correspondent au produit «*Ma Maison*». L'évolution de cette part au cours du temps est donné par le graphique suivant :

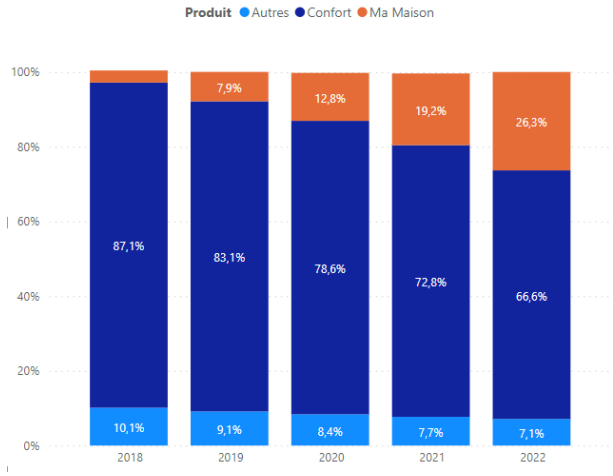


FIGURE 1.3 – Répartition des différents types de contrats dans le portefeuille

Les contrats « Autres » incluent les contrats spécifiques tels que les Grandes-Demeures ou les Mobil-Homes.

Ayant passé en revue l'offre d'AXA en matière d'assurance multirisques habitation, l'attention sera à présent portée sur la garantie climatique, objet d'étude de ce mémoire.

1.2 La garantie climatique

Dans cette section, il sera abordé un aspect essentiel de l'offre d'AXA en matière d'assurance multirisques habitation : la garantie climatique. Cette garantie faisant l'objet de ce mémoire, nous allons maintenant nous concentrer sur son analyse.

1.2.1 Problématique des évènements climatiques en France

Lorsqu'un événement météorologique met en péril les constructions, les habitants ou les activités humaines, il devient un risque climatique. En France, les risques climatiques varient en fonction des saisons et des régions. Les principales menaces climatiques sont les suivantes :

— Les inondations ;

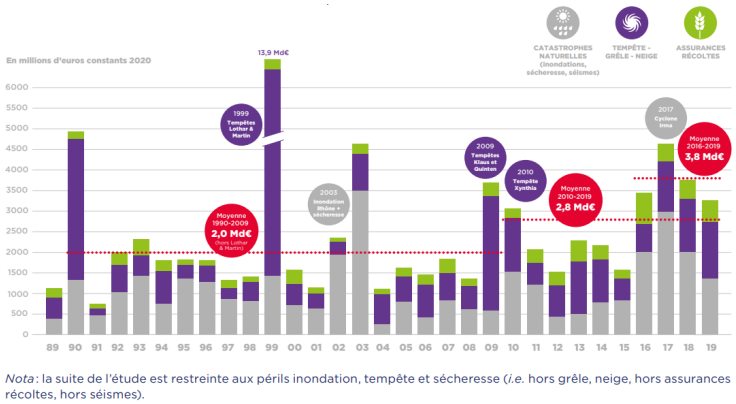
— Les tempêtes sur l'ensemble du territoire, et les cyclones en plus en outre-mer ;

— Les canicules en été ;

— Les vagues de froid en hiver. [1]

1. D'après le site d'AXA

En France, les charges de sinistres associées aux évènements climatiques ont considérablement augmenté au cours de ces quarante dernières années. En effet, les indemnisations sont passées d'un coût moyen annuel d'un peu plus d'1 milliard d'euros dans les années 1980 à environ 4 milliards d'euros aujourd'hui. [1]
Le graphique ci-dessous met en évidence le coût des évènements climatiques en France entre 1989 et 2019 ([France Assureurs, 2022]).



Nota : la suite de l'étude est restreinte aux périls inondation, tempête et sécheresse (*i.e.* hors grêle, neige, hors assurances récoltes, hors séismes).

FIGURE 1.4 – Coût des événements naturels en France sur la période 1989-2019 en Mds d'euros, *Source : FFA*

Entre 1989 et 2019, la plus grande part d'indemnisations versées par les assurances dans le cadre du climatique est causé par les tempêtes :

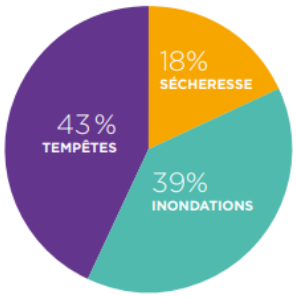


FIGURE 1.5 – Répartition du cumul des indemnisations versées par les assureurs entre 1989 et 2019 par type de péril, *Source : FFA*

Selon une étude exclusive publiée en octobre 2021 par les experts de France Assureurs (FFA), il est estimé que les coûts liés aux aléas climatiques pourraient atteindre 143 milliards d'euros au cours des 30 prochaines années en France. Cela représente une augmentation de 93 % par rapport aux coûts liés à la période entre 1989 et 2019. D'après cette même étude, le changement climatique contribuerait à un peu plus d'un tiers de cette hausse ([France Assureurs, 2021a]). Ces projections mettent en évidence l'impact croissant du changement climatique sur les coûts des sinistres liés aux événements climatiques en France et soulignent les défis économiques auxquels le pays et les assureurs devront faire face dans les décennies à venir. Cela explique l'urgence d'une action de la part des assureurs afin de limiter les pertes pouvant être dues à ce type de sinistres.
Le paragraphe suivant analysera la sinistralité de la garantie climatique dans le cadre de l'offre MRH d'AXA.

1.2.2 Etude de la sinistralité de la garantie climatique

Cadre de l'étude

Dans le cadre des produits d'AXA en MRH, la garantie climatique englobe la couverture des sinistres liés aux tempêtes, aux inondations, à la grêle et aux autres événements climatiques **qui n'ont**

1. Selon la FFA, [France Assureurs, 2021a]

pas été déclarés comme des catastrophes naturelles. Chez AXA, **chaque péril**[1] **lié à une garantie est appelé Unité de Prestations (UP)**. Ainsi, la garantie climatique comporte 4 UP différents :

— l'UP TEMP pour les tempêtes ;

— l'UP INOND pour tous les types d'inondations (pluviale ou submersive) ;

— l'UP GRELE pour les grêles ;

— l'UP NATUR pour tous les autres évènements climatiques n'ayant pas fait l'objet d'un arrêté catastrophe naturelle : fortes pluies, vents, etc.

Ainsi, il est important de souligner que **la sécheresse et les catastrophes naturelles proprement dites sont exclues de la garantie climatique et font l'objet de garanties distinctes**.

L'objet de ce paragraphe est d'analyser la sinistralité de la garantie climatique. La période d'étude s'étend du **1er janvier 2016 au 30 septembre 2022** et se restreint à la **France métropolitaine**[2]. Cette étude a été réalisée à l'aide de SAS et Python. Trois variables segmentantes seront considérées : «NewSegm», «Qualhab», «CDREGION». Elles correspondent respectivement au segment MRH, à la qualité et au type de logement de l'assuré, et à la région AXA. Les modalités de chacune de ces variables sont explicitées dans les deux tableaux suivants.

New_Segm	Signification New_Segm	Qualhab	Signification Qualhab
LA1p	Locataire Appartement 1 pièce	LocA	Locataire Appartement
LA2p	Locataire Appartement 2 pièces		
LA3p	Locataire Appartement 3 pièces		
LA4+	Locataire Appartement 4 pièces et plus		
LocM	Locataire de Maison	LocM	Locataire Maison
PM4-	Propriétaire de Maison 4 pièces ou moins	PM	Propriétaire Maison
PM56	Propriétaire de Maison 5 ou 6 pièces		
PM7+	Propriétaire de Maison 7 pièces et plus		
PNOM	Propriétaire Non Occupant Maison		
PNOA	Propriétaire Non Occupant Appartement	ProA	Propriétaire Appartement
ProA	Propriétaire Appartement		

FIGURE 1.6 – Premier type de segmentation par profil : NewSegm et Qualhab

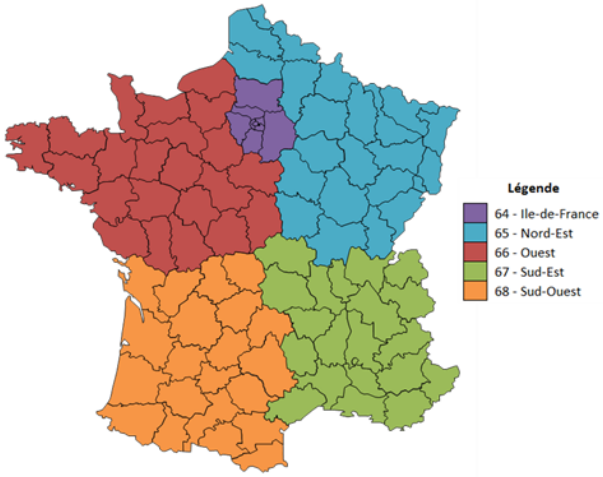


FIGURE 1.7 – Second type de segmentation : Régions AXA

NB : Par soucis de simplification d'écriture, les abréviations de chacune des modalités des variables segmentantes seront utilisées dans la suite du mémoire.

1. Un péril désigne un événement soudain et imprévu pouvant causer des dommages à des biens ou des personnes.
2. Dans tout le mémoire

Répartition des contrats

La répartition des contrats en fonction du profil de l'assuré et des régions AXA est donnée par les graphiques suivants.

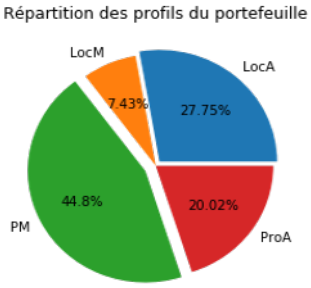


FIGURE 1.8 – Répartition des profils dans le portefeuille

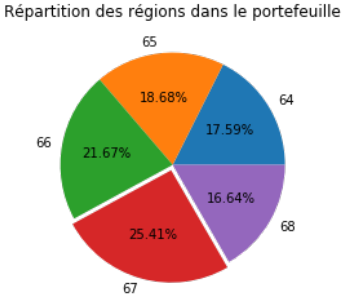


FIGURE 1.9 – Répartition des régions dans le portefeuille

La région **S**ud-**E**st (**SE**) comptabilise le plus de contrats, suivie par l'**O**uest (**OU**), le **N**ord-**E**st (**NE**), l'**Î**le-de-France (**IDF**) et enfin le **S**ud-**O**uest (**SO**). Les propriétaires représentent 64.82 % du portefeuille. En particuliers, les **P**ropriétaires de **M**aison (**PM**) décrivent 44.8 % du portefeuille. [1]

Analyse du S/C climatique

L'analyse du S/C climatique fera l'objet de ce paragraphe. Pour ce faire, dans un premier temps, la répartition de la prime acquise sera étudiée. La période d'étude s'étend toujours du **1er janvier 2016 au 30 septembre 2022**.

Primes acquises climatiques

Vision univariée

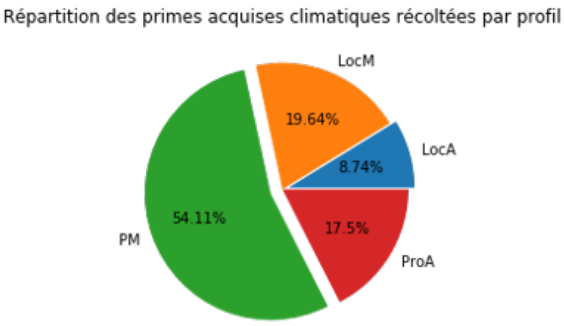


FIGURE 1.10 – Répartition des primes acquises climatiques récoltées par profil

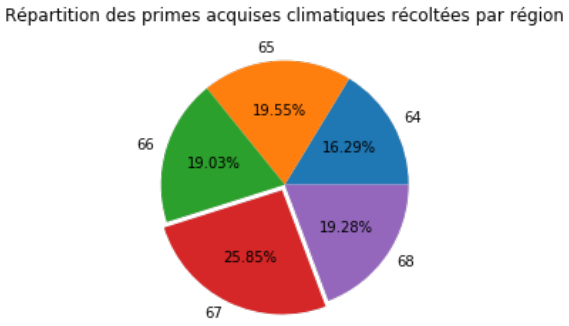


FIGURE 1.11 – Répartition des primes acquises climatiques récoltées par région

Le Sud-Est (SE) se distingue par le montant le plus élevé de primes acquises climatiques, suivi du Nord-Est (NE) et du Sud-Ouest (SO). Concernant le profil des assurés, bien que les Propriétaires de Maison (PM) représentent 44,8 % du portefeuille, ils génèrent 54,11 % des primes acquises climatiques. Cela met en évidence une prime moyenne plus élevée pour ce segment. À l'inverse, les Propriétaires d'Appartement (ProA) représentent 20 % du portefeuille, mais ne contribuent qu'à 17,5 % des primes acquises climatiques.

1. Rappel : 64 - IDF, 65 - NE, 66 - OU, 67 - SE, 68 - SO

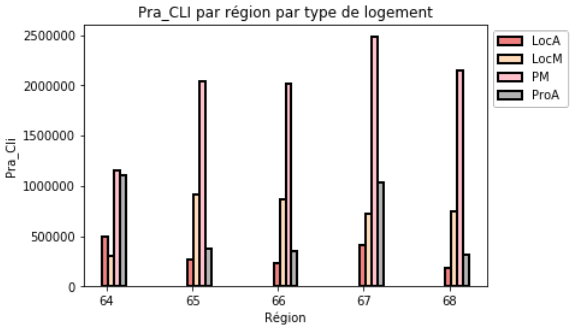


FIGURE 1.12 – Primes acquises climatiques par région et par profil

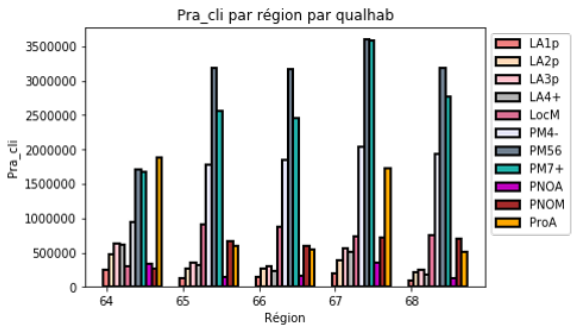


FIGURE 1.13 – Primes acquises climatiques par région et par NewSegm

D'un point de vue bivarié, c'est **la combinaison (PM, SE) qui représente le plus grand montant des primes acquises récoltées**. Au sein des PM, ce sont les PM56 et les PM7+ qui représentent les plus grands montants de primes acquises récoltées dans toutes les régions sauf en Île-de-France (IDF) où ce sont les ProA.

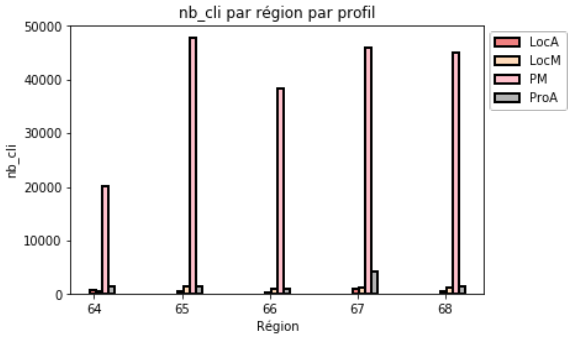


FIGURE 1.14 – Nombre de sinistres climatiques par région et par profil

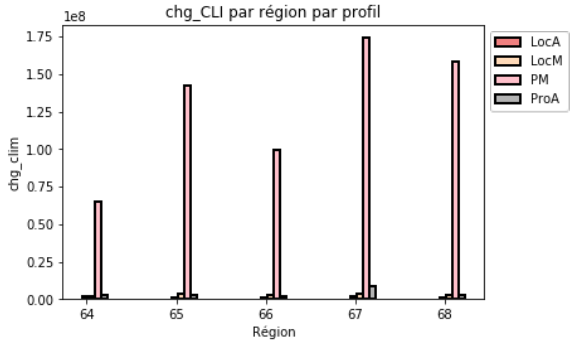


FIGURE 1.15 – Charges des sinistres climatiques par région et par profil

Concernant la sinistralité à proprement parler, c'est dans le NE qu'on observe le plus de sinistres climatiques pour les PM. Il est suivi de près par le Sud. **Les PM se démarquent des autres profils dans toutes les régions**. C'est dans le SE que les charges climatiques sont les plus élevées. Viennent ensuite le SO, le NE, l'OU et enfin l'IDF [1]. Notons que les charges de sinistres comprennent les paiements, les provisions et les recours.

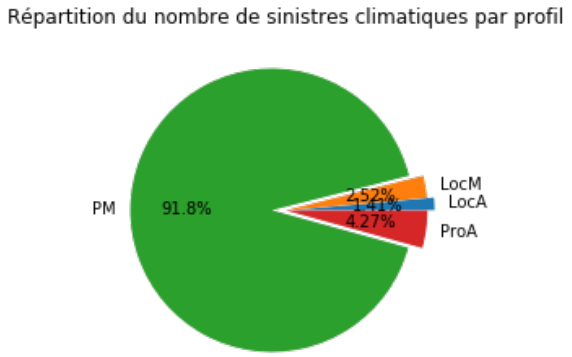


FIGURE 1.16 – Répartition du nombre de sinistres climatiques par profil

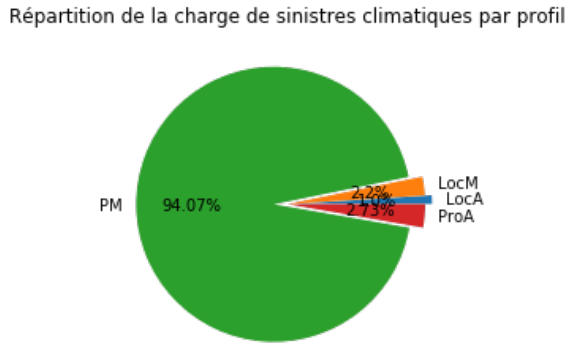


FIGURE 1.17 – Répartition de la charge des sinistres climatiques par profil

Entre 2016 et 2022, les Propriétaires de Maison (PM) ont déclaré 91,8 % des sinistres climatiques, bien qu'ils ne représentent que 44 % du portefeuille. De plus, ils ont généré 94,07 % des charges liées à ces sinistres durant cette période. Ces chiffres soulignent la sensibilité accrue des PM aux événements

1. Rappel : 64 - IDF, 65 - NE, 66 - OU, 67 - SE, 68 - SO

climatiques par rapport aux autres profils. À l'inverse, les Propriétaires d'Appartement (ProA) ont déclaré 4,27 % des sinistres climatiques, mais ces sinistres ne représentent que 2,72 % des charges. Cela indique que, bien que les ProA déclarent mo ins de sinistres, l'intensité de leur sinistralité est nettement inférieure à celle des autres segments.

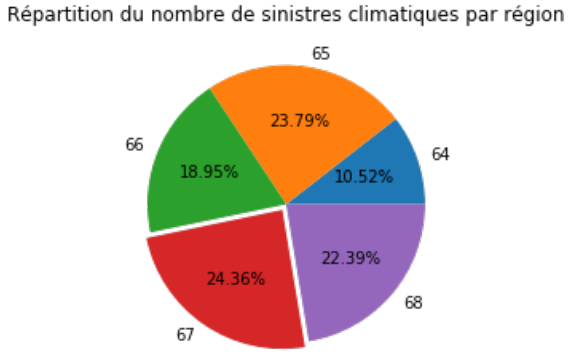


FIGURE 1.18 – Répartition du nombre de sinistres climatiques par région

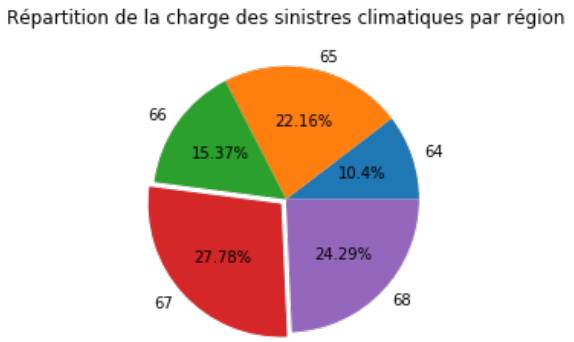


FIGURE 1.19 – Répartition de la charge des sinistres climatiques par région

Finalement, ces graphiques soulignent que c'est dans le SE que le nombre et les charges liés aux climatiques sont les plus importants. Viennent ensuite le SO et le NE, l'OU et enfin l'IDF. Le graphique suivant met en avant que le nombre de sinistres (gauche) et les montants des charges dépensées liées aux climatiques (droite) sont plus élevés dans toutes les régions pour les PM56, PM7+, PM4- et PNOM (dans cet ordre).

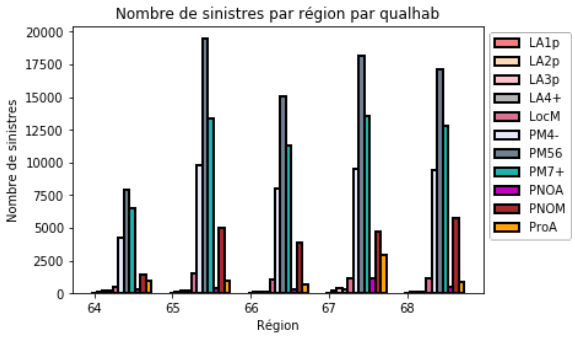


FIGURE 1.20 – Répartition du nombre de sinistres par région et NewSegm

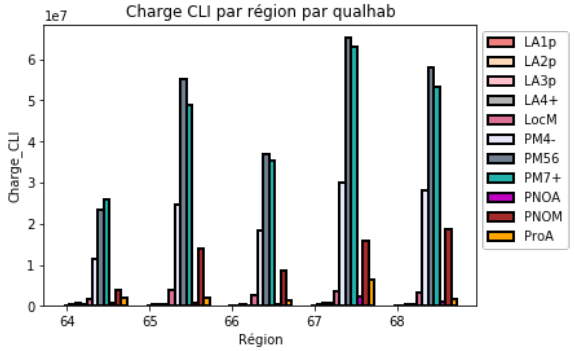


FIGURE 1.21 – Répartition de la charge des sinistres par région et NewSegm

Pour ce paragraphe uniquement, on définit la fréquence comme étant le rapport entre le nombre de sinistres et le nombre de contrats :

$$\boxed{\text{fréquence} = \frac{\text{nombre de sinistres}}{\text{nombre de contrats}}}$$

On peut ainsi observer la fréquence des sinistres climatiques d'un point de vue bivarié :

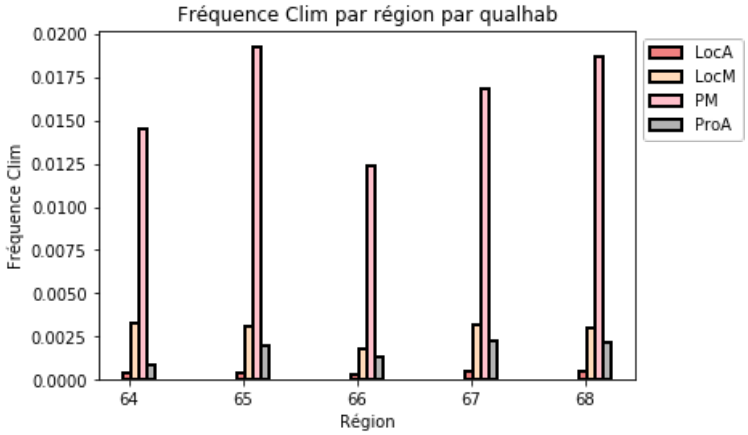


FIGURE 1.22 – Fréquence des sinistres climatiques par région et par profil

La fréquence globale des climatiques s'élève à 0.79 %. Le Nord-Est (NE) est la région où la sinistralité est la plus fréquente, suivie de près par le Sud-Ouest (SO). Ensuite, viennent le Sud-Est (SE), l'Île-de-France (IDF), et enfin l'Ouest (OU) [1]. Cela peut s'expliquer par le fait que le NE et le Sud sont les régions les plus sujettes aux tempêtes, le type de sinistre climatique le plus fréquent en France (cf 1.2.1). En ce qui concerne la fréquence des sinistres, les propriétaires occupants (PM hors PNOM) se distinguent toujours des autres profils, en enregistrant un nombre de sinistres plus élevé.

De même, on définit le coût moyen comme étant le rapport entre le montant des sinistres et le nombre de sinistres :

$$\boxed{\text{coût moyen} = \frac{\text{montant des sinistres}}{\text{nombre de sinistres}}}$$

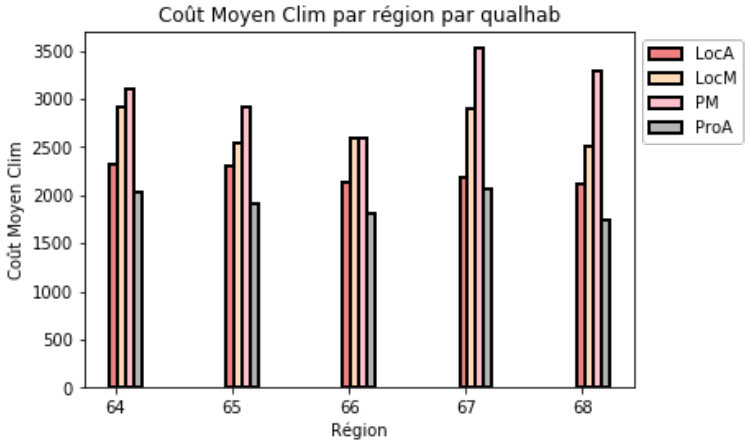


FIGURE 1.23 – Coût moyen des sinistres climatiques par région et par profil

Le coût moyen global des climatiques s'élève à 3161.75 €. En ce qui concerne les coûts moyens, la combinaison (SE, PM) se distingue une fois de plus en termes de coût élevé. Ensuite, viennent les combinaisons (SO, PM), (IDF, PM) et (NE, PM), ainsi que l'(OU, PM/LocM). Il est intéressant de notifier que dans la région de l'OU, les coûts moyens liés aux événements climatiques sont similaires pour les sinistres couverts par les polices PM et LocM.

Finalement, on analyse le ratio « sinistres sur cotisations » (S/C) des climatiques. Le ratio S/C est un indicateur clé utilisé dans le domaine de l'assurance non-vie pour évaluer la rentabilité d'un portefeuille d'assurance. Il permet de mesurer la proportion des sinistres par rapport aux cotisations collectées sur une période donnée. Ce ratio se calcule par la formule suivante :

1. Rappel : 64 - IDF, 65 - NE, 66 - OU, 67 - SE, 68 - SO

$$\boxed{\text{S/C} = \frac{\text{somme des charges}}{\text{somme des cotisations}}} \tag{1.1}$$

Lorsque le ratio S/C atteint 100 %, cela indique que, sur une période donnée, l'assureur a équilibré ses sinistres avec les cotisations perçues, ne réalisant ni bénéfice ni perte à ce titre. Toutefois, il est essentiel de noter que ce ratio ne tient pas compte des frais généraux de l'assureur, contrairement au ratio combiné (voir section 1.1). Par conséquent, un ratio de 100 % ne laisse aucune marge pour couvrir ces frais supplémentaires, ce qui pourrait compromettre la rentabilité globale de l'assureur. De la même manière, un S/C supérieur à 100 % traduit une perte opérationnelle. Cela soulève des préoccupations quant à la capacité de l'assureur à couvrir ses sinistres avec les primes collectées, sans même considérer d'autres frais. Finalement, c'est un ratio S/C strictement inférieur à 100 % qui est le plus favorable pour l'assureur, suggérant que les primes couvrent non seulement les sinistres, mais offrent aussi une marge pour d'autres charges et, potentiellement, un bénéfice. Les graphiques qui suivent illustrent le S/C climatique par région et par profil.

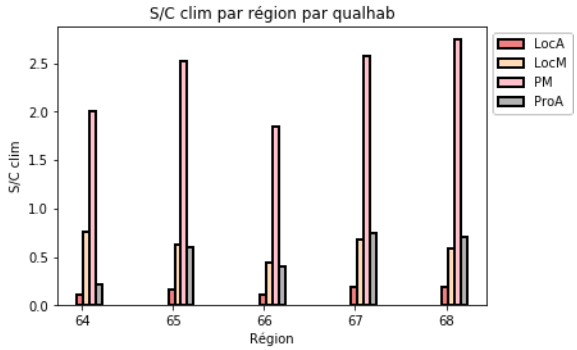


FIGURE 1.24 – S/C climatiques par région et par profil

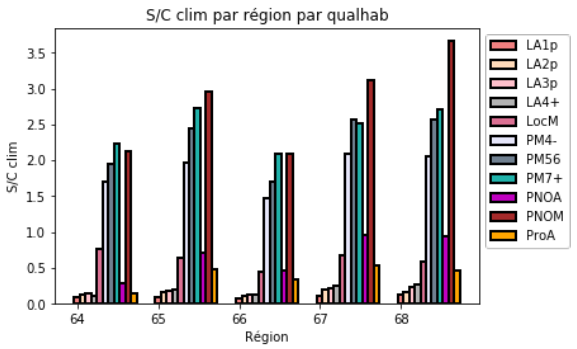


FIGURE 1.25 – S/C climatiques par région et par NewSegm

Le S/C climatique global vaut 174.4 % (Valeur moyenne entre le 1^{er} janvier 2016 et le 30 septembre 2022). Il est nettement plus élevé pour les PM dans toutes les régions, avec des valeurs comprises entre 180 % et 280 %. Plus précisément, les PNOM ont les S/C Climatique les plus élevés dans toutes les régions à part l'IDF. Pourtant, il a été vu précédemment que ce n'est pas le profil qui a les charges de sinistres climatiques les plus élevées (cf figure 1.21). Cette observation suggère que les cotisations des PNOM ne sont pas adéquates pour couvrir leurs sinistres.
Par conséquent, en cas de révision de la prime pour les PM, il faudra inclure les PNOM dans le calcul. En effet, en raison de la nature spécifique des propriétaires non occupants, leur prime pure était initialement calculée séparément (avec les PNOA). Dans le cadre de l'étude de ce mémoire, ils seront pris en compte lors d'un éventuel ajustement de la prime pour les PM.

Pour conclure la première étape de cette analyse, il y a plus de contrats dans le Sud-Est. Dans cette région, les primes acquises climatiques récoltées sont plus élevées que dans les autres régions. En termes de charges, c'est également la région présentant un montant de charges climatiques plus élevé. Viennent ensuite le Sud-Ouest, le Nord-Est, l'Ouest et enfin l'Île-de-France. En termes de segmentation, **ce sont les Propriétaires de Maisons qui se démarquent des autres profils**. En effet, ils représentent environ 92 % du nombre de sinistres climatiques et plus de 94 % des charges de sinistres climatiques du portefeuille. Les événements climatiques ont un impact majeur sur les propriétés immobilières, et les charges qui en résultent sont généralement supportées par les propriétaires, sauf si la responsabilité du locataire est établie. Cette tendance souligne pourquoi les propriétaires sont particulièrement exposés aux risques climatiques.

Finalement, c'est la combinaison **(PM ; SE)** et plus précisément **(PM56 et PM7+ ; SE)** qui est **la plus sinistrée**. **Les Propriétaires Non Occupants Maison ont le S/C le moins performant**. Cela explique la nécessité de refaire une modélisation de la prime pure climatique, en les incluant éventuellement.

Analyse des UP climatiques

Ce paragraphe se concentre sur l'analyse de la sinistralité par UP [1] climatique. L'objectif est de voir s'il existe un UP particulier qui se démarque pour chacune des régions.

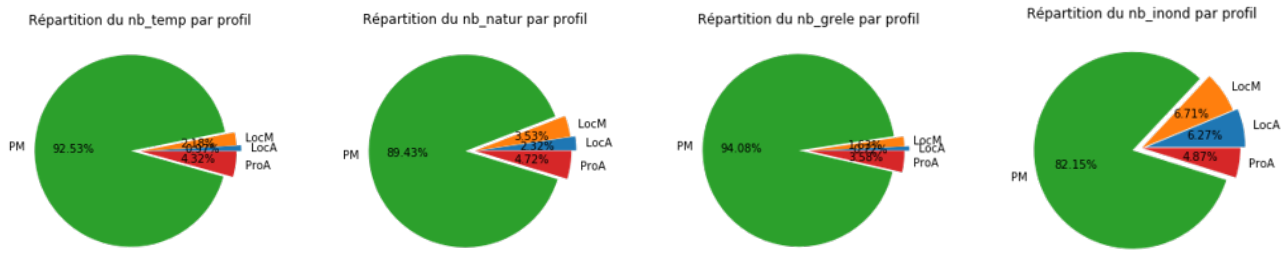


FIGURE 1.26 – Répartitions par profil du nombre de sinistres par UP

En termes de nombre de sinistres, les PM représentent la part la plus élevée pour tous les UP : leur part se situe entre 82 et 94% en fonction de l'UP. L'observation est similaire en termes de charges (entre 86 et 96 % en fonction de l'UP). Cela est cohérent compte tenu des statistiques précédentes. Les graphiques suivants montrent la répartition du nombre de sinistres par UP, en fonction des régions AXA.

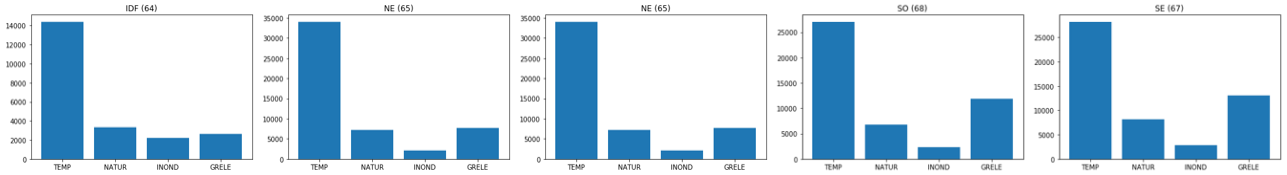


FIGURE 1.27 – Nombre de sinistres climatiques par péril et par région

Il ressort que les tempêtes sont les UP climatiques les plus fréquents dans toutes les régions. Finalement les tempêtes représentent plus de 60 % des sinistres climatiques. En termes de charges on constate sur les graphiques suivants un phénomène surprenant :

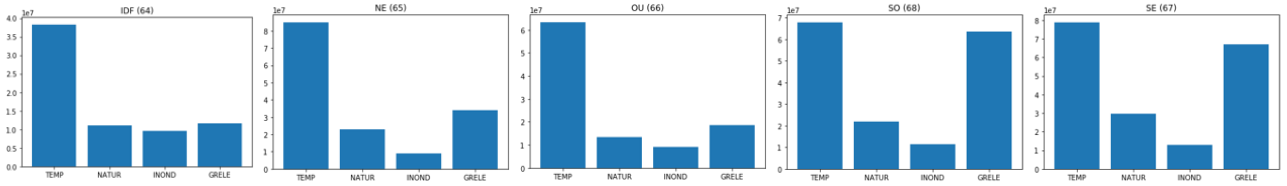


FIGURE 1.28 – Charge des sinistres climatiques par péril et par région

Si en termes de charges l'UP tempête se démarque en IDF, dans le NE et l'OU, on remarque que dans le Sud la charge totale des sinistres grêle est quasi équivalente à celle des sinistres tempête. La combinaison de deux phénomènes explique ce constat :

— En 2022, on assiste à une augmentation considérables des sinistres graves grêle [2]. Le graphique suivant le met en évidence :

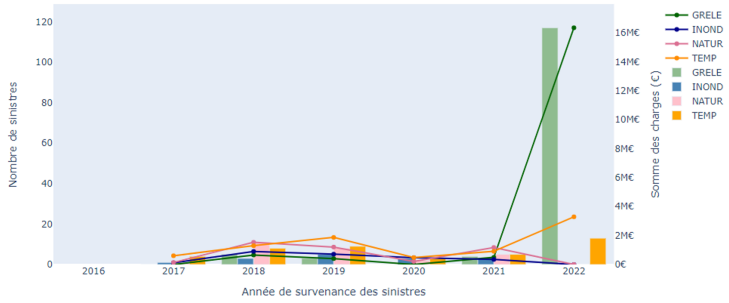


FIGURE 1.29 – Nombre de sinistres graves climatiques et charges associées

1. Unités de Prestation
2. Dont la charge est supérieure à 100 000 €.

— Lorsqu'un sinistre lié à un évènement climatique est déclaré, il est automatiquement classé en tant que sinistre climatique jusqu'à preuve du contraire. Ainsi, si plus tard, un évènement climatique est déclaré comme catastrophe naturelle et qu'il s'avère que les dommages réclamés y sont en fait liés, les dommages en question seront alors transférés à la garantie des catastrophes naturelles (CATNAT). Par conséquent, si un décrèt est adopté, la charge des sinistres climatiques (ici les grêles) pourraient être réduite.

Finalement, c'est l'UP inondation qui a le coût moyen le plus élevé partout.

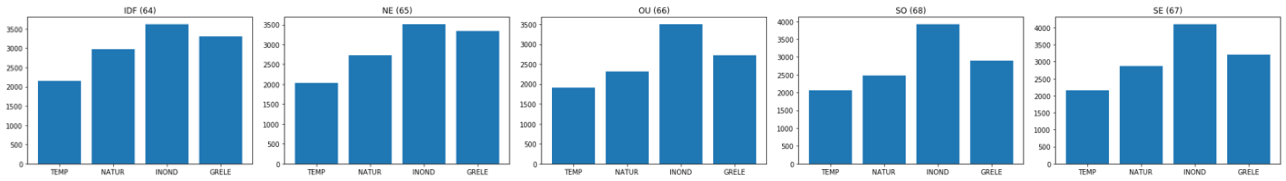


FIGURE 1.30 – Coût moyen des sinistres climatiques par péril et par région

Pour conclure l'analyse de la sinistralité de la garantie climatique, on retient que le profil majoritaire est le profil des Propriétaires de Maison. Le Nord-Est et le Sud présentent le plus de sinistres climatiques et la plus grande part de charges. L'UP tempête est le plus récurrent et le plus coûteux. Néanmoins, en 2022, dans le Sud, on a observé une série considérable de sinistres graves grêle. Cela a augmenté considérablement la charges de sinistres associée. Dernièrement, l'analyse des S/C climatique nous a permis d'émettre l'hypothèse qu'il faudrait mieux segmenter la modélisation de la prime pure des PM, en rajoutant éventuellement les PNOM.

Le paragraphe suivant détaillera les objectifs de l'étude, ainsi que les enjeux.

1.3 Objectifs de l'étude

Les propriétaires de maison occupants et les propriétaires non occupants de maison sont exposés à un risque accru de sinistres climatiques, ce qui soulève la nécessité d'améliorer la tarification de leur prime pure. Dans cette perspective, l'objectif de cette étude consistera à évaluer si l'utilisation de données ouvertes externes permet d'améliorer la modélisation de leur prime pure. Nous envisageons d'exploiter trois sources de données ouvertes pour enrichir notre analyse. Tout d'abord, la base des arrêtés relatifs aux catastrophes naturelles et la base MétéoNet, afin d'optimiser notre modèle de fréquence des sinistres. Ensuite, nous considérons la base de Demandes de Valeurs Foncières pour améliorer notre modèle de coût moyen des sinistres. Il est essentiel de noter que notre équipe n'a pas pour habitude de mener la tarification de la garantie climatique en séparant les profils. Comme évoqué précédemment, si cette étude s'avère fructueuse, les bases de données utilisées seront testées pour d'autres garanties et d'autres profils.

Dans le cadre de la modélisation, la méthode fréquence-coût moyen sera utilisée afin de modéliser la prime pure. Le prochain paragraphe justifiera l'utilisation de cette méthode. Celui d'après traitera de l'intérêt des données ouvertes.

1.3.1 Méthodologie de modélisation de la prime pure

Dans cette section, on définit l'espace probabilisé (Ω, A, P)*. Soit* X *est une variable aléatoire définie dans cet univers. On note* $\mathbb{E}[X]$ *l'espérance de cette variable aléatoire et* $\mathrm{Var}[X]$ *sa variance, que nous supposons bien définies. Soit* A *et* B *deux variables aléatoires définies dans l'espace probabilisé.*

La prime pure

Soit X_i le coût de sinistre d'un assuré i.

Afin de déterminer la prime réellement payée par le client, plusieurs étapes sont nécessaires. La première consiste à calculer l'**espérance du risque**, également appelée **prime pure**. En effet, dans le domaine de l'assurance, l'inversion du cycle de production oblige l'assureur à déterminer préalablement

cette prime pure, qui doit couvrir intégralement les charges de sinistres pour une police d'assurance donnée sur une période donnée. Elle doit être calculée de sorte à éviter le phénomène d'antisélection. En effet, le principe de la tarification repose sur celui de la segmentation. Ce dernier vise à regrouper les assurés en classes homogènes de risque de la façon la plus précise possible. Le phénomène de l'antisélection aurait pour conséquences que les risques les plus élevés [1] choisissent de s'assurer dans une compagnie tandis que les risques plus faibles se tournent vers la concurrence en raison d'une prime jugée trop élevée. Par conséquent, la segmentation des assurés permet de proposer des primes adaptées à chaque profil de risque, garantissant ainsi l'équilibre financier de l'entreprise (cf [PERRIN, 2021]). Néanmoins, une segmentation trop fine, frôlant l'individualisation du tarif, peut biaiser l'estimation de la charge moyenne de sinistre, compromettant les résultats de l'assureur. Il est donc essentiel de trouver un équilibre dans la segmentation.

Dans le cadre de ce mémoire, la modélisation de la sinistralité reposera sur le principe du modèle collectif qui différencie le modèle fréquence d'une part et le modèle coût d'une autre part. En utilisant cette approche de modélisation, déterminer la prime pure pour un assuré i consiste à évaluer l'espérance de la valeur totale des sinistres survenus au cours de la période considérée de ce même assuré i. On le note $\mathbb{E}[X_i]$. Ainsi, la variable aléatoire X_i se décompose de la manière suivante :

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

Avec :

— N_i : Nombre de sinistres survenus pour l'assuré i ;

— $C_{i,j}$: Le coût du sinistre n°j de l'assuré i.

Par conséquent, afin d'estimer $\mathbb{E}[X_i]$, il faut définir la fréquence des sinistres et le coût moyen.

Définition 1.3.1. *La fréquence de sinistres de l'assuré i correspond au ratio suivant :*

$$\boxed{freq_i = \frac{Nombre\ de\ sinistres_i}{Exposition_i}} \tag{1.2}$$

L'exposition correspond à la durée pendant laquelle le risque est couvert par le contrat d'assurance. Par exemple, si un contrat est souscrit le 1er février d'une année N, et reste en vigueur jusqu'au 31 décembre de cette même année, son exposition annuelle est de $\frac{11}{12}$. Si un sinistre survient pendant cette période de couverture, la fréquence observée serait de $1 \times \frac{12}{11} = 109.09\%$: un sinistre sur une période de 11 mois équivaut en moyenne à 1.09 sinistres par an en termes de risque.

Définition 1.3.2. *Le coût moyen des sinistres de l'assuré i correspond quant à lui au ratio suivant :*

$$\boxed{cm_i = \frac{Charge\ totale\ des\ sinistres_i}{Nombre\ de\ sinistres_i}} \tag{1.3}$$

La méthodologie du modèle collectif repose sur deux hypothèses :

1. Les charges des sinistres individuels sont des variables aléatoires **i**ndépendantes et **i**dentiquement **d**istribuées (**iid**). Cette hypothèse est admise ;
2. Le nombre total des sinistres est indépendant du coût de chaque sinistre. Pour confirmer la validité de cette hypothèse, une étude de la relation entre la fréquence des sinistres et leur coût sera effectuée afin de détecter une éventuelle corrélation entre les deux. Trois mesures de corrélation seront étudiées.

1. Ou assurés les plus risqués.

Définition 1.3.3. *Le coefficient de corrélation de Pearson :*

$$\boxed{\rho(A,B) = \frac{\mathbb{E}[A,B] - \mathbb{E}[A]\mathbb{E}[B]}{\sqrt{\mathrm{Var}[A]\mathrm{Var}[B]}}}$$

Le coefficient de corrélation de Pearson mesure la corrélation linéaire entre deux variables continues. Le coefficient prend des valeurs entre -1 et 1. Une valeur de 1 indique une corrélation positive parfaite[1]*, et une valeur de -1 indique une corrélation négative parfaite*[2]*. Une valeur de 0 indique une absence de corrélation linéaire entre les variables*[3]

Définition 1.3.4. *Le coefficient de corrélation de Spearman.*

Pour un échantillon de taille $n \in \mathbb{N}$, les variables de rang $(rg(A_i), rg(B_i))_{i\in[1,n]}$ sont calculées à partir des données (A_i, B_i). Ainsi, la corrélation de Spearman est définie par :

$$\boxed{\rho_s(A,B) = \frac{cov(rg(A_i), rg(B_i))}{\sigma_{rg}(A_i)\sigma_{rg}(B_i)}}$$

Où $cov(rg(A_i), rg(B_i))$ désigne la covariance des variables de rang et $\sigma_{rg}(A_i)$ et $\sigma_{rg}(B_i)$ sont les écarts-types.[4] *La corrélation de Spearman est utilisée lorsque deux variables semblent être corrélées, mais pas de manière linéaire. Elle cherche à établir un coefficient de corrélation entre les rangs des valeurs des deux variables, plutôt que leurs valeurs réelles. Elle mesure dans quelle mesure la relation entre deux variables peut être décrite par une fonction monotone. Une valeur de 0 indique une absence de corrélation selon l'ordre des rangs. Contrairement au coefficient de Pearson, le coefficient de Spearman est robuste : l'impact d'éventuelles données aberrantes sur sa valeur est limitée.*

Définition 1.3.5. *Le tau de Kendall.*

Le coefficient de Kendall, également appelé τ de Kendall, repose également sur l'ordre des observations, mais il diffère du coefficient ρ_s de Spearman par son principe de calcul. En effet, il repose sur la notion de paires discordantes et concordantes :

Deux paires d'observations (x_i ; y_i) et (x_j ; y_j) sont concordantes si (x_i - x_j)(y_i - y_j) > 0, discordantes si (x_i - x_j)(y_i - y_j) < 0. Cette mesure est moins sensible aux liens égaux[5] *que le ρ_s de Spearman, et elle convient mieux lorsque les données contiennent des valeurs répétées.*
Soit (A',B') un couple indépendant et de même distribution que (A,B). Le tau de Kendall se définit de la façon suivante :

$$\boxed{\tau(A,B) = P[(A-A')(B-B') > 0] - P[(A-A')(B-B') < 0]}$$

De même, si la valeur du tau est proche de 0, alors les variables A et B sont indépendantes.

Les résultats du calcul de chacun des indicateurs, calculés à l'aide du logiciel SAS, sont recensés dans le tableau suivant :

TABLE 1.1 – Mesures de dépendance

Mesure de dépendance	Valeur
ρ	0.01132
ρ_s	0.07069
τ	0.05772

1. Les variables varient dans le même sens.
2. Les variables varient ensemble dans des sens opposés.
3. Il est important de noter que le coefficient de corrélation de Pearson ne capture pas les relations non linéaires entre les variables.
4. Notons que cette définition correspond à la corrélation de Pearson des variables de rang.
5. Référence aux valeurs ayant le même rang.

Ainsi, la valeur de chacun des indicateurs étant faible (et proche de 0), on peut considérer qu'il y a bien indépendance entre la fréquence et le coût. Par conséquent, l'utiliser un modèle fréquence-coût est mathématiquement justifiée. $\mathbb{E}[X_i]$ peut s'écrire de la façon suivante :

$$\mathbb{E}(X_i) = \mathbb{E}[\mathbb{E}(X_i|N_i)] = \mathbb{E}[\mathbb{E}[\sum_{j=1}^{N_i} C_{i,j}|N_i]]$$

$$= \mathbb{E}[\mathbb{E}(N_i C_i|N_i)]$$

(car les $C_{i,j}$ sont iid)

$$= \mathbb{E}[N_i \mathbb{E}(C_i|N_i)]$$

$$= \mathbb{E}[N_i \mathbb{E}(Ci)]$$

Car C_i et N_i sont indépendants

$$= \mathbb{E}(N_i)\mathbb{E}(C_i)$$

$$= \text{Espérance du nombre de sinistres} \times \text{Espérance du coût des sinistres.}$$

On en déduit la formule de la prime pure d'un assuré i :

$$\boxed{PP_i = \frac{\text{Charge totale des } sinistres_i}{Exposition_i} = \text{fréquence des } sinistres_i \times \text{coût } moyen_i = freq_i \times cm_i} \tag{1.4}$$

La fréquence des sinistres est principalement influencée par le comportement de l'assuré, tandis que le coût est davantage déterminé par les caractéristiques du bien assuré. Ainsi, il est cohérent de s'attendre à ce que les variables prises en compte dans chaque modèle diffèrent. Ensuite, les modèles de fréquence sont généralement plus robustes et stables que les modèles de coût. En effet, les modèles de coût peuvent être moins fiables en raison d'un nombre d'observations réduit, puisque seuls les contrats avec des sinistres sont utilisés pour étudier la distribution des montants de sinistres. De plus, il existe un décalage temporel entre ces deux concepts : la survenance du sinistre est connue dès sa déclaration à l'assureur, tandis que le coût réel n'est établi qu'à la fin du processus de règlement éventuel. C'est en prenant en compte ces deux aspects dans notre modèle que nous serons en mesure de mieux appréhender les risques spécifiques aux Propriétaires de Maison, et de proposer une prime pure plus précise et adaptée à leur profil. Plus la détermination de la prime pure sera précise, plus l'étude de l'impact des données ouvertes le sera.

Le principe de tarification présenté, examinons l'intérêt d'utilisation des données ouvertes.

1.3.2 Intérêt des *Open data*

Les *Open Data* ou données ouvertes en français, sont des données accessibles gratuitement et sans restrictions, ce qui permet leur exploitation et leur réutilisation. Ces données peuvent aussi bien provenir de sources publiques, telles que les services gouvernementaux, les collectivités locales ou les communes, que de sources privées, comme les entreprises, les organisations non gouvernementales (ONG), les startups ou les fondations caritatives.
Les données ouvertes sont régies par quelques grands principes ([BOUCHER, 2016]) :

1. Disponibilité et accès : Les données doivent être facilement accessibles, de préférence via Internet ;
2. Réutilisabilité et redistributivité : Les données doivent être fournies dans des conditions qui permettent leur réutilisation, leur transformation, leur enrichissement et leur redistribution ;
3. Participation universelle : Aucune discrimination ne doit être faite quant aux finalités d'utilisation.

Depuis 2011, avec l'ouverture de data.gouv.fr, portail dédié à l'*Open Data*, la France se positionne dans le top 10 des pays qui mettent à disposition leurs données à l'échelle mondiale selon l'INSEE.

Dans notre cas, l'*Open Data* nous permet d'accéder à des ensembles de données climatiques précieuses provenant de diverses sources, telles que les agences météorologiques, les satellites, et d'autres. Ces données peuvent inclure des informations sur les températures, les précipitations, la vitesse des vents, les risques d'inondations, etc. Par conséquent, l'utilisation des *Open Data* peut contribuer à améliorer la précision des modèles, en fournissant des informations plus détaillées, récentes et fiables. Il est important de noter que la fiabilité des données est importante car la modélisation de la garantie climatique est un domaine complexe et en évolution constante.

***Open data* étudiées**

Afin d'illustrer le potentiel de l'*Open Data* dans la modélisation de la prime pure climatique en MRH, ce mémoire va se concentrer sur l'exploitation de plusieurs bases :

— La base des arrêtés catastrophes naturelles, disponible sur le site du gouvernement français « data.gouv.fr ». L'exploitation de cette base vise à améliorer le modèle fréquence ;

— La base de Demandes de valeurs foncières, également disponible sur le site du gouvernement. Dans la base de modélisation interne, nous disposons d'informations sur les caractéristiques des assurés et leur mobilier, mais pas sur les valeurs immobilières des biens assurés. Ainsi, l'exploitation de cette base vise à améliorer le modèle coût moyen.

— La base météorologique de Météonet, disponible sur le site de Météo France. L'exploitation de cette base vise à améliorer le modèle fréquence.

Avant de nous plonger dans l'analyse de ces données ouvertes, il est nécessaire d'introduire les bases de données internes et de décrire leur processus de traitement. En effet, c'est à partir des bases fréquence et coût moyen, générées par ces bases internes, que nous fusionnerons les informations issues des bases de données ouvertes.

1.4 Bases d'étude internes

Il a déjà été mentionné que pour réaliser une modélisation pertinente, il est primordial de créer une base de données robuste et aussi complète que possible. Elle inclura les informations des contrats d'assurance habitation entre le 01/01/2016 et le 31/12/2022, afin de recueillir un maximum d'informations concernant les contrats, les clients associés et les sinistres éventuels liés à la garantie climatique. Cela nous permettra d'avoir une période d'historique de 7 ans et de rester cohérent avec l'analyse de la sinistralité effectuée précédemment. Nous n'avons gardé que les contrats Confort et Ma Maison, afin de mieux cibler notre étude. Le traitement des bases a été effectué à l'aide de SAS.

1.4.1 Présentation des bases

La base image

La première base qui sera présentée est la base Contrats par image de risque. Cette base, appelée base image, constitue la base de données centrale. En effet, elle fournit des informations relatives à toutes les caractéristiques de l'habitation disponibles telles que le nombre de pièces du logement, l'ancienneté du logement, le type d'habitation [1], le type d'occupation [2], etc. Elle donne également les caractéristiques de souscription telles que l'exposition [3], la cotisation [4], les garanties et les options souscrites. Ces caractéristiques des contrats seront utilisées comme variables dans notre modélisation afin de mieux segmenter le risque climatique. Cette base est mise à jour mensuellement, offrant ainsi une image mensuelle des contrats.

1. Appartement, Maison
2. Occupant, non occupant, logement principal ou secondaire...
3. Défini dans la section précédente
4. Ce que paye effectivement l'assuré

Afin d'utiliser cette base il y a un traitement à effectuer. En effet, la base associe les variables de risque à un numéro de contrat : elle est structurée de manière à représenter chaque image de risque tout au long de la durée de vie du contrat. Plus précisemment, lorsqu'il y a des modifications d'informations, telles qu'un déménagement de l'assuré ou un changement de garantie, une nouvelle ligne est ajoutée à la base image pour refléter cette modification, tout en conservant le même numéro de contrat. On parle de nouveau risque. De plus, à chaque nouvel exercice, une nouvelle ligne est créée pour représenter encore une fois un nouveau risque.
Pour cette étude, il a été extrait à l'aide de SAS la dernière image du risque de chaque contrat pour chaque exercice, afin d'obtenir une base annuelle cohérente : ainsi, on construit une base de risque comportant une ligne par risque en portefeuille pendant la période d'étude, plutôt qu'une ligne par numéro de contrat. Ainsi, on garantit la stabilité et la fiabilité de la modélisation.

La base sinistres

Les bases relatives à la sinistralité appelées base Sinistres sont organisées par année de survenance. Ces bases permettent d'obtenir des informations détaillées sur les sinistres tels que les garanties impactées ou les flux comptables associés [1] : la charge totale par sinistre est calculée en prenant en compte le règlement principal, les réserves et les recours. Ces bases contiennent également des informations relatives à l'état du sinistre (en cours ou clos), etc. Les sinistres avec une charge nulle ou négative ont été écartés afin de garantir la qualité des données.
La vision des sinistres est fixée au 31/12/2022. Cela permet ainsi d'avoir au moins une année de développement pour les sinistres survenus en 2021 et avant. Etant donné le temps d'écoulement de charges des sinistres climatique, cela est suffisant.

Base Clients et Base Adresses

Deux autres bases de données vont être utilisées dans le cadre de cette étude, l'une est interne et l'autre est d'origine externe.

— La première base de données, interne, est la Base Clients, qui contient les caractéristiques propres aux assurés telles que leur âge, l'ancienneté de leur contrat MRH, leur ancienneté dans le portefeuille IARD AXA, leur statut marital, leur nombre d'enfants, et leur catégorie socio-professionnelle ;

— La seconde base est la Base Adresses. Cette source de données sera particulièrement utile pour la thématique de la zonification qui sera abordée plus tard dans l'étude. En effet, grâce à cette base on peut associer les attributs géographiques de chaque adresse liée à chaque contrat présent dans la base de modélisation :

1. L'adresse, le code postal, la commune et son code commune appelé aussi code INSEE associés au risque assuré. Le code **INSEE** est un identifiant numérique de longueur 5 attribué par l'**I**nstitut **N**ational de la **S**tatistique et des **E**tudes **E**conomiques à chaque commune en France. Il permet d'identifier de manière unique une commune donnée. Au 1er janvier 2022, la France compte 34 955 communes [2] ;
2. Le code **IRIS** (**I**lots **R**egroupés pour l'**I**nformation **S**tatistique) associé au risque assuré. Le code IRIS est un système de découpage territorial utilisé en France. Il permet de diviser le territoire en mailles de taille homogène afin de faciliter l'analyse et la comparaison des données statistiques. Chaque IRIS correspond à une zone géographique définie, généralement de petite taille, et ainsi identifiée par un code unique. Il en existe 48 606 en France métropolitaine [3] ;
3. Les coordonnées géographiques exactes de type projection conique Lambert-93, qui correspond au type de projection le plus utilisé en France ;

1. Paiements, provisions, recours
2. D'après Bis 163 le nombre de communes et depci fiscalite propre de collectivites-locales.gouv.fr.
3. D'après : INSEE, Découpage infra-communal paru le 17/09/2020.

4. La qualité géocodage de l'adresse : « 4 » si l'adresse est exacte, « 3 » si le numero de rue est approche, « 2 » pour le centroïde de la voie, « 1 » pour le centroïde de la ville et « 0 » en cas d'erreur.

Les deux prochains paragraphes se concentreront sur le traitement des bases relatives à la sinistralité.

1.4.2 Sélection des sinistres attritionnels

Lorsque la sinistralité est modélisée, une hypothèse principale est d'assumer que le portefeuille d'assurance est composé de risques homogènes. Cependant, cette hypothèse est compromise lorsque le portefeuille contient des sinistres dits « graves », caractérisés par un montant élevé mais une occurrence rare. Ces sinistres, de par leur nature exceptionnelle, ne peuvent pas être mutualisés parmi l'ensemble des assurés. Ils nécessitent donc une gestion spécifique, distincte de celle des autres sinistres, et ne peuvent être pleinement intégrés au mécanisme classique de mutualisation des risques.

Afin de faire face à cette situation, il est nécessaire de déterminer un seuil d'écrêtement (ou seuil grave de sinistre). Ce seuil correspond au montant au-delà duquel les sinistres seront considérés comme graves et seront traités séparément du reste du portefeuille. Les sinistres dont les montants sont inférieurs à ce seuil, dits attritionnels, seront quant à eux inclus dans le système de mutualisation des risques avec les autres assurés. Ainsi, l'écrêtement consiste à identifier un changement dans la partie extrême de la distribution des sinistres. On a :

$$\text{charge attritionnelle} = \min(\text{charge réelle}, \text{seuil grave}) \tag{1.5}$$

Et

$$\text{charge grave écrêtée} = (\text{charge réelle} - \text{seuil grave}) \cdot \mathbb{1}_{[\text{charge réelle}>\text{seuil grave}]} \tag{1.6}$$

Dans le cadre de ce mémoire, **l'analyse se concentrera uniquement sur l'impact de l'utilisation des données ouvertes sur les sinistres attritionnels**. La détermination du seuil sera effectuée à l'aide de la théorie des valeurs extrêmes.

1.4.3 Calcul du seuil

Afin de trouver un seuil, nous ferons appel à la théorie des valeurs extrêmes. Le théorème de Pickands (1975) est particulièrement intéressant, car il établit que sous certaines conditions favorables et pour un seuil élevé suffisamment choisi, la loi de Pareto généralisée (ou GPD pour *Generalized Pareto Distribution*) constitue une excellente approximation de la loi des excès F_u d'une variable aléatoire Y. En exploitant cette propriété, nous serons en mesure de mieux caractériser les sinistres graves et d'estimer de manière précise un seuil. F_u est défini de la façon suivante :

$$F_u(y) = P[Y - u \leq y | Y > u], \quad y \geq 0$$

Pour une variable aléatoire Y, les dépassements $(Y - u)$ au-delà d'un certain seuil u suivent une loi de Pareto généralisée (GPD), notée $GPD_{(\sigma,\gamma)}(y)$. Le paramètre σ est positif et appelé paramètre d'échelle, tandis que γ représente le paramètre de forme. Une valeur élevée du paramètre de forme γ indique une présence accrue de valeurs extrêmes dans la distribution. La fonction de répartition d'une GPD se définit ([Raillard, 2021]) :

$$\forall y \geq 0 \; P[Y - u > y | Y > u] = \begin{cases} 1 - \left(1 + \gamma \frac{y}{\sigma}\right)^{-\frac{1}{\gamma}} & \text{si } \gamma \neq 0 \\ 1 - \exp\left(-\frac{y}{\sigma}\right) & \text{si } \gamma = 0 \end{cases} \tag{1.7}$$

Afin de déterminer la valeur du seuil approprié, la fonction de dépassement moyen des excès sera examinée, ainsi que la stabilité des paramètres associés aux propriétés de la loi GPD. Le choix du seuil représente un défi, car il doit être suffisamment élevé pour que l'approximation GPD soit valide, tout en évitant d'avoir un faible nombre de données pour estimer les paramètres du modèle. Cette étude sera réalisée sur R (librairies `ismev` , `extRemes`, `MASS` et `evd`).

Méthode de la stabilité du paramètre d'échelle

La première méthode étudiée sera l'analyse des paramètres de la GPD. Cette loi présente une propriété de stabilité par seuil, ce qui signifie que si les dépassements $(Y-u)$ suivent une $GPD_{(\sigma_u,\gamma_u)}$, alors pour tout seuil $v > u$, les dépassements $(Y-v)$ suivent également une GPD de paramètres σ_v et γ_v. En fait, seul le paramètre d'échelle σ diffère. Il est une fonction linéaire du seuil : $\sigma_v = \sigma_u + \gamma_u(v-u)$. Le paramètre de forme est en réalité identique pour tout u ([DERKAOUI, 2021]) : $\gamma_u = \gamma_v = \gamma$. L'analyse de la stabilité du paramètre d'échelle est effectuée graphiquement : son estimation est représentée pour plusieurs seuils, incluant les intervalles de confiance à 95 % (cf les traits verticaux). Le seuil retenu est celui correspondant à la plus petite valeur de u pour laquelle le paramètre d'échelle est stable.

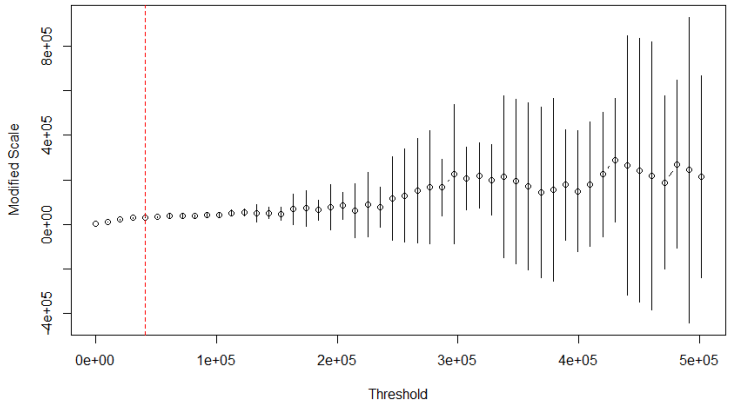


FIGURE 1.31 – Analyse de la stabilité du paramètre d'échelle σ

Le seuil choisi est donné par la droite verticale rouge. On obtient $u_1 = 40\ 936$.

Méthode de l'analyse de la fonction moyenne des excès

La fonction moyenne des excès se définit de la façon suivante :

$$e(v) = E[Y - v | Y > v] = \frac{\sigma + \gamma(v-u)}{1-\gamma}; \quad \text{pour } v > u \text{ et } \gamma < 1 \tag{1.8}$$

Cette fonction est linéaire par rapport à v si l'approche GPD est valide. Son estimateur est défini par :

$$e_n(v) = \frac{1}{N_v} \sum_{i=1}^{N_v} (X_i - v)_+$$

où N_v est le nombre de données supérieures à v.

Si la variable suit une loi de Pareto généralisée pour un seuil donné, alors le graphique de l'estimateur de la fonction de dépassement moyen devra être approximativement linéaire au-delà de ce seuil ([DERKAOUI, 2021]).

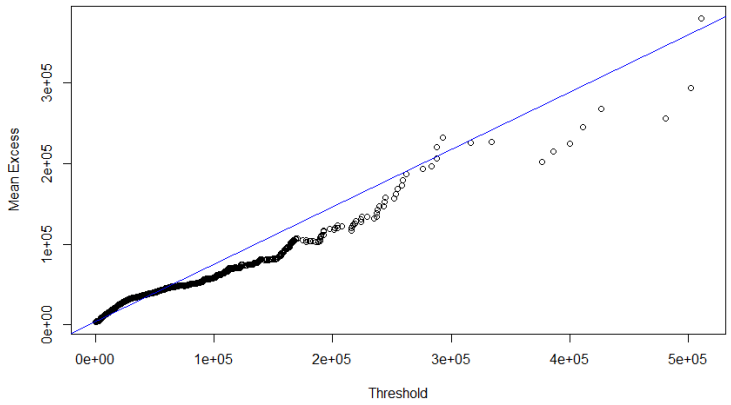


FIGURE 1.32 – Fonction moyenne des excès. *La droite bleue correspond à la droite de régression du nuage de points.*

On lit ainsi u_2= 40 408.

Méthode de l'estimateur de Hill

La dernière méthode utilisée pour définir le seuil : il s'agit de l'estimateur de Hill qui n'est utilisable que pour les distributions à queue lourde [1], soit des distributions où le paramètre de forme γ est positif. L'estimateur s'écrit :

$$\hat{\gamma}_n(k) = \frac{1}{k}\sum_{j=1}^{k} \ln\left(\frac{X_{(i)}}{X_{(k+1)}}\right) \tag{1.9}$$

où :

— $X_(i)$ est la statistique d'ordre associée à l'échantillon $X_1, \ldots, X_n$;

— k représente un nombre d'excès inférieur ou égal à n ;

— n est le nombre d'observations.

Lorsque le graphique de l'estimateur de Hill est représenté, ce qui correspond à la valeur de l'estimateur en fonction de l'indice k de la statistique d'ordre, la première étape consiste à trouver une zone où l'estimateur semble stable et robuste. Le seuil optimal correspondra au plus petit seuil u appartenant à cette zone ([DERKAOUI, 2021]) :

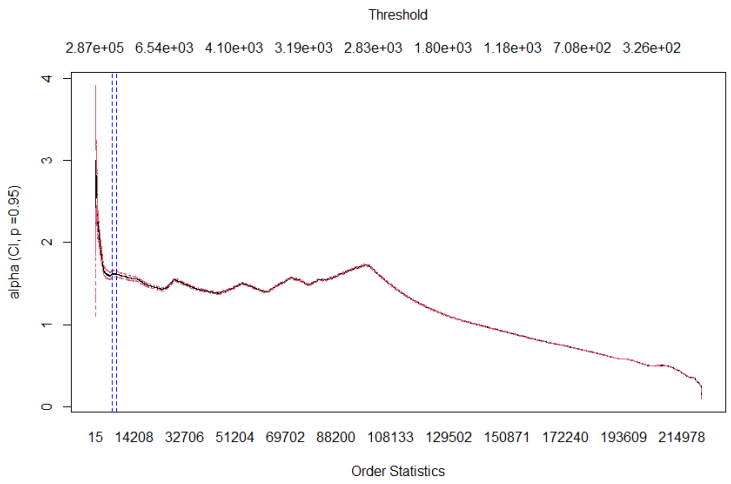


FIGURE 1.33 – Analyse de l'estimateur de Hill

La zone retenue est celle où la statistique d'ordre est comprise entre 6100 et 7500 (cf traits verticaux bleus). Cette zone correspond à une zone de seuil comprise entre 39315.1 et 42565.51. On garde ainsi

1. Si la fonction moyenne des excès est croissante, alors la queue est lourde, comme c'est le cas ici. cf [SCHRYVE, 2018]

$u_3 = 39315.1$.

Les trois seuils retenus sont proches. Le choix final du seuil sera déterminé à l'aide de graphiques quantile-quantile pour tous les seuils compris entre 39 000 (entier inférieur de u_3) et 41 000 (entier supérieur à u_1). Ce type de graphique permet de tester visuellement l'adéquation d'une famille de lois à des données. Dans notre cas, il affichera un nuage de points ayant pour abscisse les quantiles théoriques d'une GPD dont les paramètres correspondent aux paramètres estimés, et pour ordonnée les quantiles empiriques de la distribution observée. S'il y a adéquation, les quantiles de la famille de lois testées et les quantiles de notre échantillon seront linéairement liés.

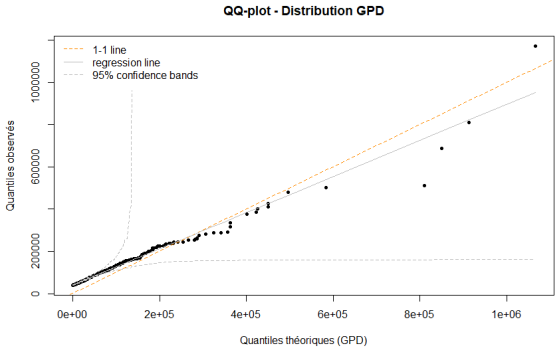


FIGURE 1.34 – Graphique Quantile-Quantile avec u = 40 000

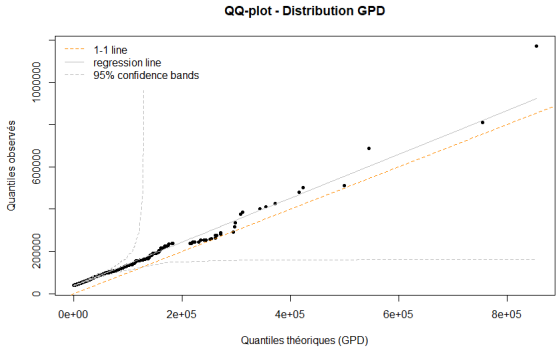


FIGURE 1.35 – Graphique Quantile-Quantile avec u = 41 000

La queue de distribution semble bien correspondre à une GPD. Le graphique associé à **u = 41 000** est particulièrement convaincant de par son alignement avec le nuage de points. En utilisant ce seuil, 0.7 % des contrats présentent une charge de sinistre grave, avec un coût moyen de 3510 €. Ce seuil sera donc retenu.

1.4.4 Vieillissement des sinistres

Problématique

Le coût d'un sinistre pour l'assureur évolue dans le temps. Lorsqu'un sinistre survient, une évaluation dite forfaitaire lui est attribuée en fonction de sa nature à l'ouverture du dossier. Dans le cas des sinistres à montant important, ce montant sera réévalué à la hausse ou à la baisse par un expert. Si ce n'est pas le cas, le montant évoluera selon les règlements effectués par l'assureur, c'est-à-dire en fonction de l'indemnisation réelle (sous facture). Par conséquent, le coût final du sinistre peut être sujet à des changements au fil du temps, pouvant s'étendre sur plusieurs années selon les garanties associées. Dans le cadre de notre modélisation, les charges de sinistres (notamment celles de 2022) n'ont pas nécessairement atteint leur coût final, ce qui peut biaiser notre étude de sinistralité et donc la modélisation. Pour obtenir une vision plus juste de la sinistralité finale attendue, nous devons donc « vieillir » nos sinistres : les évolutions du coût des sinistres au fil du temps seront estimées.

Pour effectuer ce vieillissement, nous appliquerons la méthode de Chain Ladder, qui s'appuie sur des triangles de sinistres. Ces triangles illustrent la progression du cumul des sinistres d'une année comptable à l'autre, pour chaque année de survenance. Grâce à la méthode de Chain Ladder, nous serons en mesure d'extrapoler les valeurs observées jusqu'à l'évaluation définitive du coût total ([ANGOUA, 2023]).

Application

Afin d'expliquer de façon plus visuelle cette méthode, les notations adoptées sont les suivantes :

— i : l'année d'origine, avec $i = 2008, \ldots, 2022$: on choisit une plage d'année large pour s'assurer de la robustesse de la méthode. Les sinistres compris entre 2008 et 2022 sont donc choisis ;

— j : l'année de développement, avec $j = 1, \ldots, 14$;

— $X_{i,j}$: la mesure de sinistralité correspondant à l'année d'origine i et à l'année de développement j ;

— $C_{i,j}$: coefficient du triangle de montants cumulés tel qu'il est défini par $C_{ij} = \sum_{h=1}^{j} X_{ih}$.

Ainsi $X_{i,j} = C_{i,j} - C_{i,j-1}$. Le triangle de charges cumulées présente les montants cumulés des sinistres pour chaque année d'origine i et pour chaque année de développement j. Chaque cellule du tableau contient le coefficient $C_{i,j}$ correspondant. Un aperçu est donné par le tableau suivant :

Année de survenance i / Année de développement j	0	1	...	j	...	n-1	n
1	$C_{1,0}$	$C_{1,1}$	...	$C_{1,j}$	...	$C_{1,n-1}$	$C_{1,n}$
2	$C_{2,0}$	$C_{2,1}$	...	$C_{2,j}$	...	$C_{2,n-1}$	
⋮	⋮	⋮	⋱	⋮			
i	$C_{i,0}$	$C_{i,1}$	...	$C_{i,j}$			
⋮	⋮	⋮					
n-1	$C_{n-1,0}$	$C_{n-1,1}$					
n	$C_{n,0}$						

TABLE 1.2 – Triangle des charges cumulés

La méthode de Chain Ladder consiste à déterminer le triangle inférieur du tableau en estimant les montants non connus. Cette méthode repose sur les hypothèses suivantes :

1. Aucun sinistre ne peut être encore ouvert après l'année de développement n. Ici n=14 ;
2. Les versements cumulés sont indépendants entre les années de survenance ;
3. La répartition des paiements est supposée constante dans le temps. Ce dernier point peut se réécrire comme suit :

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

Le triangle de sinistres se remplit en employant des proportions, désignées sous le terme de facteurs de développement ($\hat{f}_j$), pour chaque année de développement j :

$$\hat{f}_j = \frac{\sum_{i=1}^{n-j-1} C_{i,j+1}}{\sum_{i=1}^{n-j-1} C_{i,j}} \quad (2.7) \tag{1.10}$$

Une fois ces facteurs calculés, on peut obtenir une estimation de la charge cumulée finale de l'année i ($\hat{C}_{i,n}$) :

$$\hat{C}_{i,n} = \hat{C}_{i,n-1} \times \hat{f}_{n-1} = \hat{C}_{i,n-2} \times \hat{f}_{n-1} \times \hat{f}_{n-2} = \ldots = C_{i,n-i+1} \prod_{j=1}^{n-i} \hat{f}_{n-j} \tag{1.11}$$

Dans le cadre de l'étude, les données historiques des sinistres de 2008 à 2022 seront utilisées afin d'évaluer la charge finale des sinistres survenus entre 2016 et 2022. L'historique de développement des sinistres est à la fois stable et conséquent, garantissant ainsi une évaluation aussi précise que possible de la charge ultime. Cette méthode a été exclusivement mise en œuvre pour les sinistres attritionnels (voir section 1.4.2).

TABLE 1.3 – Triangle de charges cumulées initial

	Triangle appliqué à nos données													
Année de survenance	0/1	1/2	2/3	3/4	4/5	5/6	6/7	7/8	8/9	9/10	10/11	11/12	12/13	13/14
2008	20 692 066	22 663 586	22 953 622	23 004 193	23 062 729	23 061 808	23 059 554	23 087 164	23 051 802	23 036 161	23 043 281	23 043 485	23 036 918	23 044 587
2009	139 091 099	136 021 035	137 434 499	137 811 910	137 835 958	137 852 558	137 861 309	137 904 409	137 882 340	137 885 431	137 885 431	137 886 991	137 890 130	137 890 130
2010	57 051 743	55 900 652	55 982 241	55 992 042	56 097 770	55 922 240	55 854 179	55 884 157	55 907 706	55 911 936	55 887 057	55 850 026	55 850 026	
2011	25 015 745	32 653 253	32 213 768	32 330 085	32 289 910	32 260 400	32 264 163	32 250 392	32 247 922	32 285 460	32 285 460			
2012	38 544 509	39 599 262	39 795 798	39 900 021	39 881 566	39 851 354	39 735 816	39 747 253	39 699 559	39 701 843				
2013	68 684 780	81 556 488	82 582 567	82 525 939	82 666 609	82 512 230	82 454 415	82 462 472	82 461 707					
2014	69 781 732	72 957 937	72 716 695	72 958 524	72 831 905	72 793 650	72 775 531	72 754 989						
2015	34 504 295	36 288 197	36 539 309	36 531 702	36 523 238	36 485 792	36 489 102							
2016	46 459 630	46 572 352	46 763 050	46 583 823	46 515 719	46 507 012								
2017	67 042 085	69 644 319	69 297 873	69 013 605	69 028 391	68 974 107								
2018	88 073 189	82 416 422	82 809 641	82 923 469	82 719 926									
2019	69 240 122	82 995 915	84 733 355	85 059 456										
2020	54 269 503	57 658 510	58 381 725											
2021	58 019 854	57 621 876												
2022	146 224 323													

Les coefficients de développement sont résumés dans le tableau suivant :

Table 1.4 – Coefficients de développement calculés

j	0/1	1/2	2/3	3/4	4/5	5/6	6/7	7/8	8/9	9/10	10/11	11/12	12/13	13/14
f	1,05	1,01	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00

Les sinistres attritionnels présentent une stabilisation de charge après trois années de développement (les facteurs sont égaux à 1 après 3 ans). Ainsi, la charge finale des sinistres attritionnels donnés par année de survenance sont :

Table 1.5 – Charges ultimes par année de survenance

2016	46 508 138
2018	68 974 107
2019	82 719 926
2020	85 059 456
2021	58 381 725
2022	57 621 876
2022	146 224 323

Dans le cadre du vieillissement des sinistres, il y a un aspect qui n'a pas encore été pris en compte : l'inflation. En réalité, il est raisonnable de postuler qu'un sinistre qui s'est produit en 2016 n'aurait pas le même coût financier en 2018, en raison de l'inflation. Pour prendre en compte ce facteur, l'ajustement du coût du sinistre aurait pu être effectué en se basant sur l'indice fourni par la Fédération Française du Bâtiment (FFB). Hors cet indice, conçu pour refléter les coûts de construction de logements neufs, pourrait ne pas capturer précisément l'évolution des coûts de réparation, en particulier pour les sinistres climatiques. Les spécificités des réparations liées aux événements climatiques peuvent différer de celles des constructions neuves, rendant l'indice FFB moins pertinent pour estimer ces coûts. Par conséquent, l'année de survenance du sinistre a été incluse en tant que variable explicative dans la base de données utilisée pour la modélisation, permettant ainsi de prendre en compte l'impact de l'inflation.

1.4.5 Construction de la base de modélisation à partir des données internes

Une jointure est effectuée entre les différentes tables mentionnées précédemment. Des retraitements (nettoyages et préparations) vont ensuite être nécessaires afin d'assurer la fiabilité des bases de modélisation : la base fréquence et la base coût moyen.

Traitements massifs

Pour commencer, deux nouvelles variables ont été générées à partir de la variable concernant les montants des biens mobiliers. La première, MTOBJVAL (Montant objet de valeur), représente la valeur des biens mobiliers de valeur, qui est calculée en multipliant la valeur des biens mobiliers par le taux des objets de valeur. La deuxième, MTCAPAUT (Montant capitaux assurés), représente la valeur des biens mobiliers communs, calculée en multipliant la valeur des biens mobiliers par l'expression (1 - taux_des_objets_de_valeur).

Deuxièmement, afin d'augmenter la robustesse et la pertinence de la modélisation, plusieurs ajustements ont été apportés aux variables. Pour éviter des modalités avec des effectifs considérés comme « négligeables » pour certaines variables, une fusion de certaines modalités a été réalisée. Cela a été particulièrement pertinent pour les variables ayant de multiples modalités, comme celle concernant l'ancienneté du client. De plus, nous avons discrétisé quelques variables continues ce qui permettra de détecter les effets non linéaires potentiels dans la répartition des données. Les variables ainsi transformées comprennent :

- **Âge du client** :
  - Moins de 18 ans ;
  - 18 - 20 ans ;
  - Plus de 100 ans ;

  - Tranches de 5 ans pour les autres âges.
- **Dimension des dépendances** :
  - Plus de 500 m^2 ;
  - Tranches de 25 m^2 pour les autres dimensions.
- **Montant des objets de valeur** :
  - Moins de 500 € ;
  - Plus de 150 500 € ;
  - Tranches de 1 000 € pour les autres montants.
- **Montant des capitaux assurés** :
  - Moins de 1 500 € ;
  - Plus de 350 500 € ;
  - Tranches de 1 000 € pour les autres montants.
- **Âge du bâtiment** :
  - Moins de 5 ans ;
  - Entre 5 et 10 ans ;
  - Plus de 10 ans.

Génération des bases

Pour rappel, l'étude concerne uniquement les propriétaires de maisons occupants et non occupants. Ils sont alors sélectionnés dans un premier temps.

Afin de construire la base de fréquence, une première étape consiste à créer la variable fondamentale, intitulée « nb_attri_cli ». Cette dernière représente le nombre de sinistres climatiques attritionnels rattachés à chaque contrat, le seuil ayant déjà été fixé (cf 1.4.2). Ensuite, il est essentiel d'établir une variable essentielle : l'exposition du contrat, telle que définie précédemment (cf 1.3.1). Pour déduire cette variable, les dates d'affaire nouvelle (AN) et de résiliation (RS) pour chaque année d'occurrence, provenant de la base image, sont utilisées :

$$Exposition_{annee} = \frac{\min(\text{Date de fin d'année}, \text{Date RS}) - \max(\text{Date de début d'année}, \text{Date AN})}{\text{Nombre de jours dans l'année}}$$

Cette base est constituée de 9 379 782 entrées et 168 colonnes. Le nombre le plus élevé de sinistres climatiques enregistrés pour un seul contrat est de 4.

En ce qui concerne la base de coût moyen, les charges des sinistres non clos sont mises à jour dans les bases jointes : les facteurs de développement, déterminés précédemment, sont appliqués pour estimer les charges ultimes (cf 1.4.4). Avec ces charges révisées, une nouvelle variable est introduite : la charge vieillie attritionnelle (notée « chg_vieillie »). En dernier lieu, une sélection est effectuée pour ne retenir que les contrats ayant enregistré un sinistre avec une charge vieillie attritionnelle strictement positive. Finalement, cette base contient 168 446 entrées et 120 colonnes.

Par la suite, les variables temporelles associées aux événements de souscription, de renouvellement et de résiliation ont été éliminées, car elles étaient jugées non pertinentes pour la modélisation. Après cette suppression, un travail important a été réalisé pour assurer l'uniformité de la base de données. Cela comprenait une vérification approfondie de la qualité des données, avec une attention particulière pour détecter d'éventuelles valeurs aberrantes ou valeurs manquantes dans les données. Voici le principe de traitements des données manquantes :

- Si une variable présente une modalité vide qui est véritablement indéterminée, elle est remplacée par « 00 - NR ». C'est le cas, par exemple, pour la variable relative à la tranche d'âge ;
- Pour les variables dichotomiques où la valeur 0 est équivalente à une absence de données, la valeur manquante est remplacée par 0. Cela concerne les variables relatives aux antécédents par exemple.

La modélisation des GLM sera effectuée à l'aide d'Akur8, logiciel de modélisation développé par un prestataire externe à l'entreprise. Par conséquent, l'une des préoccupations majeures est de respecter les normes de confidentialité et de sécurité des données [1]. Ainsi, toutes les données personnelles identifiables, comme les numéros de contrat, de client, les noms et prénoms des assurés, ainsi que les informations de contact, ont été minutieusement supprimées pour protéger la confidentialité des clients.

1.4.6 Statistiques descriptives

Corrélations

Une analyse de corrélation a été réalisée pour étudier les relations entre les différentes variables présentes dans la base de données. Cette étape essentielle vise à comprendre les liens entre les variables et à prévenir tout biais potentiel dans les résultats du modèle causé par des corrélations élevées. L'objectif est de préserver le caractère interprétable du modèle. Une limite a été définie pour identifier les corrélations fortes : toute valeur supérieure à 0.5 a été considérée comme indicative d'une forte corrélation.

Le coefficient de corrélation de Spearman a été utilisé pour évaluer la corrélation entre les variables numériques, à l'aide de Python. La matrice de corrélation de Spearman est présentée ci-dessous.

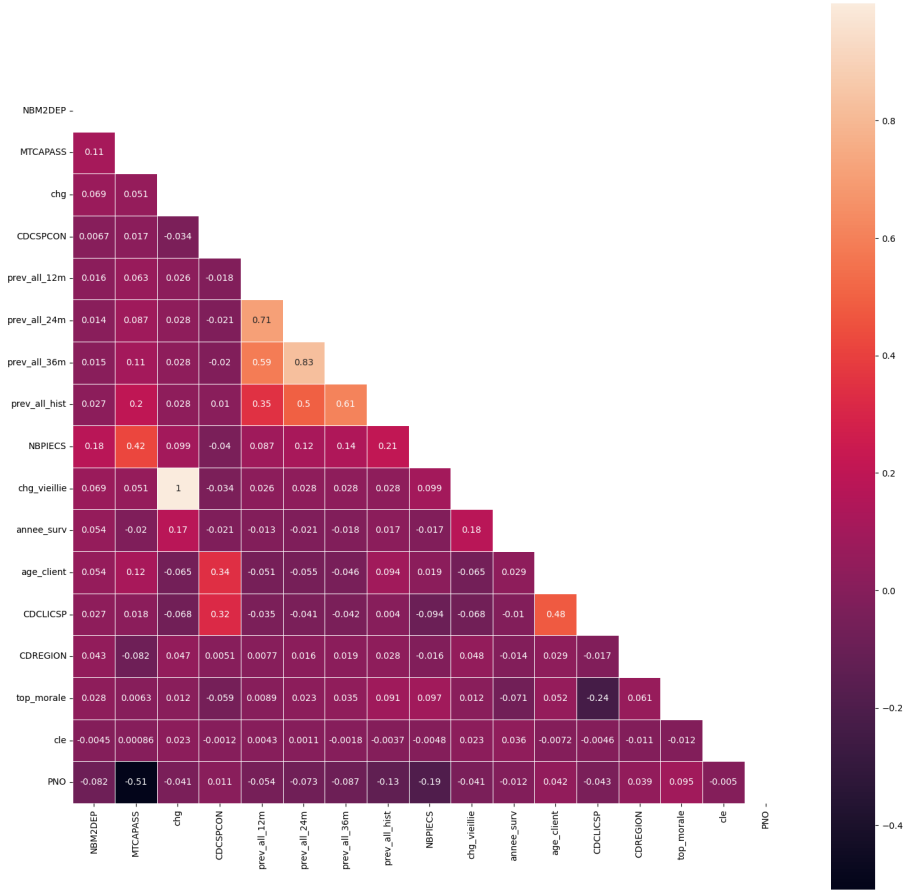


FIGURE 1.36 – Corrélations de Spearman pour des variables de la base coût moyen

À titre d'exemple, il existe une corrélation parfaite et positive entre la charge vieillie et la charge non vieillie. De plus, les antécédents de tous les sinistres sur un historique de 36 mois sont corrélés à 0.6 avec les antécédents sur 5 ans.

En ce qui concerne les corrélations entre les variables catégorielles, et celles entre les variables catégorielles et numériques, le score de corrélation ϕ_K a été utilisé. Le score de corrélation ϕ_K est une mesure qui évalue la force de la relation entre deux variables. Il varie entre 0 (absence totale de corrélation), et 1 (corrélation parfaite). Cette mesure est basée sur une version modifiée du test

1. En particulier, les dispositions du Règlement Général sur la Protection des Données (RGPD).

de contingence de Pearson χ^2 qui est un test d'hypothèse pour l'indépendance entre deux variables ([Baaka *et al.*, 2019]). Notons que lorsque les deux variables sont issues d'une distribution normale bivariée, le score coïncide avec la valeur absolue du coefficient de corrélation de Pearson [1]. Ainsi, ce score offre plusieurs avantages notables :

— Il est applicable de manière cohérente aux variables catégorielles, ordinales ou continues ;

— Il est en mesure de détecter des relations non-linéaires entre les variables ;

— Son calcul ne repose pas sur une formule fermée, offrant ainsi une grande flexibilité.

Ainsi, le score ϕ_K se présente comme une alternative robuste aux méthodes traditionnelles de corrélation, surtout dans les situations où les relations entre les variables ne sont pas purement linéaires. Ce score a été calculé à l'aide de l'outil Akur8.
À titre d'exemple, la tranche d'âge du client et le nombre d'enfants à charge affichent une corrélation positive de 0.591, tandis que l'âge du bâtiment présente une corrélation de +0.34 avec ces variables. Le réseau de distribution montre une corrélation de +0.664 avec la région AXA.

Variables sélectionnées selon leur importance

Après la sélection de la variable cible, un score d'importance est attribué à chaque variable par Akur8. Il s'agit du score d'« Importance Des Variables Univariées ». Grâce à ce score, les variables constituant une « fuite de cible » sont identifiées : il s'agit de variables à postériori qui « ne sont pas connues au moment de la prédiction » [2]. Pour faciliter la détection de telles variables, Akur8 évalue la puissance prédictive d'une unique variable. Les valeurs sont ensuite ajustées à la valeur médiane de la liste de toutes les variables. Bien que cette métrique de puissance prédictive ne possède pas une signification en magnitude absolue, l'ordre relatif des scores d'importance offre des indications sur les variables susceptibles de causer une fuite. Le concept mathématique sous-jacent repose principalement sur des méthodes de régularisation et des modèles statistiques bayésiens. En effet, l'importance des variables est mesurée en évaluant leur contribution dans le modèle à l'aide d'un cadre bayésien doté de priors [3] appropriés ([Yann TRAONMILIN, 2018]). Ce score est directement lié au degré de régularité nécessaire pour exclure la variable du modèle. Autrement dit, un score d'importance élevé par rapport aux autres indique une contribution significative de la variable à la performance du modèle, tandis qu'un faible score suggère une moindre pertinence de la variable.

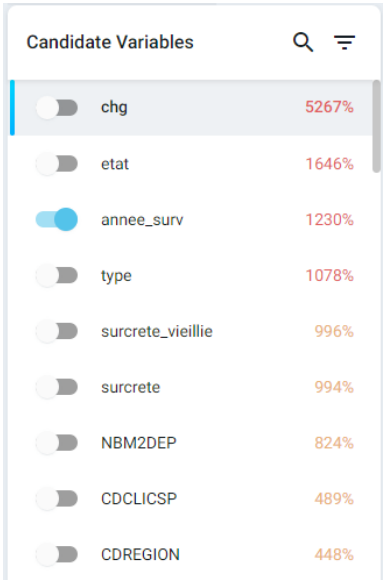


FIGURE 1.37 – Importance des variables pour le modèle Coût Moyen, *Akur8*

1. Une distribution normale bivariée est une extension de la distribution normale univariée à deux dimensions. Elle décrit la distribution conjointe de deux variables aléatoires continues.
2. Autrement dit, lorsqu'une variable explicative, utilisée pour prédire la cible, contient déjà des informations sur cette cible, rendant ainsi la prédiction biaisée ou trop optimiste.
3. Connaissances préalables.

Variable	Description
chg	Charge non vieillie des sinistres
etat	État des sinistres (ouvert ou clos)
annee_surv	Année de survenance du sinistre
CDCLISCSP	Catégorie socio-professionnelle

TABLE 1.6 – Description de quelques variables *(les autres seront explicitées plus tard)*

La variable de catégorie socio-professionnelle n'a pas été incluse en raison de son manque de fiabilité, principalement causé par de nombreuses valeurs manquantes et l'absence de mises à jour régulières.

Il est notable que l'importance accordée à la variable« chg » est exceptionnellement élevée par rapport à celle des autres. Étant donné que nous modélisons la charge vieillie, maintenir la variable « chg » constitue une fuite de cible. En effet, la variable « charge » est directement liée à la charge vieillie que nous cherchons à modéliser. Si nous utilisons la charge non vieillie comme variable, le modèle pourrait simplement « apprendre » cette relation directe, rendant la prédiction artificiellement précise. En d'autres termes, en utilisant « chg », le modèle pourrait simplement reproduire la charge vieillie sans vraiment comprendre les autres facteurs influençant cette charge. C'est pourquoi il est essentiel d'identifier et de gérer de telles variables pour éviter des prédictions biaisées.

Tenant compte de la corrélation des variables, de leur score d'importance, de leur pertinence [1] et des directives de l'entreprise, le tableau ci-après recense les variables retenues pour la modélisation de chacun des modèles (hors zonier) :

Modèle Coût Moyen			
Variable	Correspondance	Modalités	Importance de la variable
annee_surv	Année de survenance du sinistre	2016, 2017, 2018, 2019, 2020, 2021, 2022	1230%
OPT_AMG_PISC	Option Aménagement Piscine	0 (pas d'option), 1 (présence de l'option)	406%
NBPIECS	Nombre de pièces	1 à 23	359%
tr_NBM2DEP	Nombre de m2 de dépendance, par tranche	22 tranches	169%
CDRESID	Type de résidence	P (Principale - PO), S (Secondaire - PO), O (bien Occupé - PNOM), U(bien inoccupé - PNOM)	111%
Age_batiment	Age du batiment	3 modalités : "Moins de 5 ans", "Entre 5 et 10 ans", "Plus de 10 ans"	111%
DEDUCTIBLE_TYPE	Type de Franchise du contrat	3 modalités : Franchise 150 euros, Majoration, Rachat	108%
DISTRIB	Réseau de distribution	5 modalités : Agents, AXA Partenaires, Courtiers, Salariés, WEB	70%
tr_age_client	Âge du client, par tranche	19 intervalles d'âge de longueur 5 ans	67%

FIGURE 1.38 – Variables candidates pour le modèle coût moyen

Modèle Fréquence			
Variable	Correspondance	Modalités	Importance de la variable
NBPIECS	Nombre de pièces	1 à 23	1068%
OPT_AMG_PISC	Option Aménagement Piscine	0 (pas d'option), 1 (présence de l'option)	1024%
year_expo	Année d'exposition	2016, 2017, 2018, 2019, 2020, 2021, 2022	852%
IDINSERT	Présence d'un insert	1 (présence), 0 (absence)	727%
CDRESID	Type de résidence	P (Principale - PO), S (Secondaire - PO), O (bien Occupé - PNOM), U(bien inoccupé - PNOM)	723%
OPT_VAN	Option Valeur à Neuf	0 (pas d'option), 1 (présence de l'option)	626%
Age_batiment	Age du batiment	3 modalités : "Moins de 5 ans", "Entre 5 et 10 ans", "Plus de 10 ans"	534%
tr_MTOBJVAL	Montant des objets de valeur, par tranche	29 tranches	300%
tr_NBM2DEP	Nombre de m2 de dépendance, par tranche	22 tranches	229%
tr_age_client	Âge du client, par tranche	19 intervalles d'âge de longueur 5 ans	167%
DISTRIB	Réseau de distribution	5 modalités : Agents, AXA Partenaires, Courtiers, Salariés, WEB	115%
DEDUCTIBLE_TYPE	Type de Franchise du contrat	3 modalités : Franchise 150 euros, Majoration, Rachat	47%

FIGURE 1.39 – Variables candidates pour le modèle fréquence

Les variables liées aux zones géographiques, aux antécédents de sinistres climatiques et aux données ouvertes seront explicitées et intégrées lors de la phase spécifique de modélisation.
À présent que la base de données a été soigneusement nettoyée, restructurée et validée, il convient d'analyser les variables qui seront utilisées pour la modélisation.

Statistiques

Dans cette section, l'évolution de chaque variable des deux modèles (fréquence des sinistres climatiques et coût moyen des sinistres climatiques) en fonction des modalités choisies, notamment l'année, le nombre de pièces, l'âge du bâtiment et le type de résidence sera analysée.

1. Par exemple, dans le cadre de la modélisation climatique, la variable concernant les antécédents de vol sera écartée.

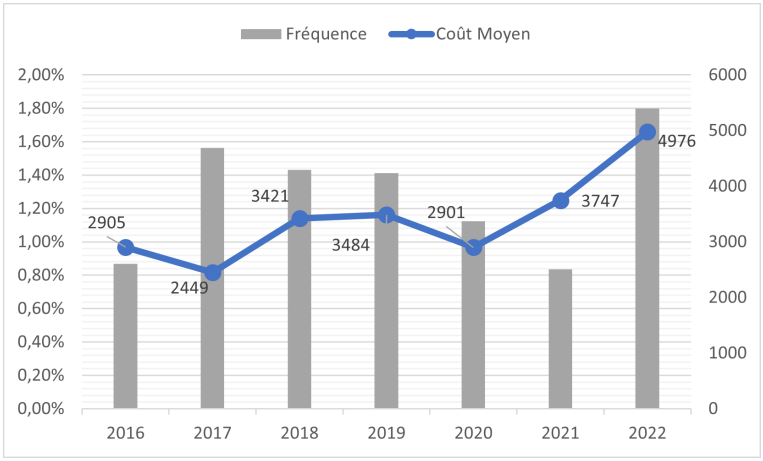


FIGURE 1.40 – Evolution de la fréquence et du coût moyen sur la période 2016 - 2022

L'analyse du graphique montre que le coût moyen des sinistres climatiques pour les propriétaires de maisons, qu'ils soient occupants ou non, a augmenté de 71 % depuis 2020, pour une augmentation du nombre de sinistres de 115 %. Cette augmentation est plus prononcée que celle observée entre 2017 et 2018, qui était de 40 %, malgré une baisse de la sinistralité de 8.46 %.

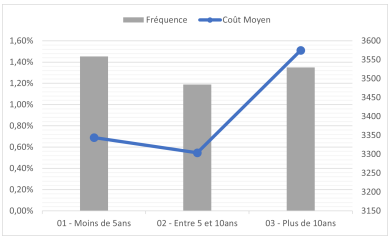


FIGURE 1.41 – Evolution de la fréquence et du coût moyen par âge du bâtiment

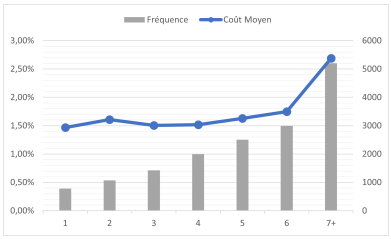


FIGURE 1.42 – Evolution de la fréquence et du coût moyen par nombre de pièces

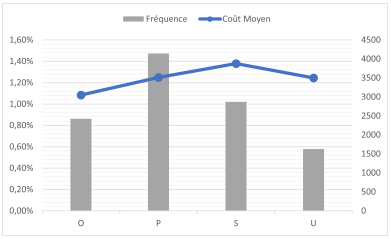


FIGURE 1.43 – Evolution de la fréquence et du coût moyen par type de résidence

Concernant l'âge du bâtiment (graphique de gauche), on note que 83 % des logements du portefeuille ont plus de 10 ans. Bien que les sinistres soient plus fréquents pour les logements de moins de 5 ans, le coût moyen est significativement plus élevé pour ceux de plus de 10 ans.

Le graphique central analyse l'évolution en fonction du nombre de pièces, un élément clé de la tarification en assurance habitation. Il apparaît que le nombre de pièces a une influence notable sur la fréquence des sinistres. Concrètement, une augmentation moyenne de 32 % est constatée à chaque ajout d'une pièce pour des logements allant de 1 à 6 pièces. Cette progression atteint 74 % lorsque le nombre de pièces dépasse le seuil des 7.
De plus, un phénomène intéressant est observé : le coût moyen des sinistres pour les logements de 2 pièces est plus élevé que pour ceux de 3 pièces. Cela s'explique en partie par la distribution géographique de ces logements par rapport aux zones où les sinistres coûtent le plus cher.

Enfin, le graphique de droite illustre l'évolution en fonction du type de résidence. Il est à noter que la fréquence des sinistres est plus élevée chez les Propriétaires Occupants (PO). Cependant, une différence notable est observée dans le coût moyen selon le statut d'occupation. Les PO possédant un logement secondaire ont un coût moyen supérieur à ceux possédant un logement principal. Par ailleurs, le coût moyen des Propriétaires Non Occupants (PNO) ayant un logement occupé est inférieur à ceux ayant un logement inoccupé.
L'explication de cette tendance réside en partie dans la répartition des logements : les résidences secondaires des PO ne représentent que 9,5 % du portefeuille, tandis que les résidences principales s'élèvent à 79 %. De plus, la zone Sud est celle où le coût moyen des sinistres est le plus prononcé. À titre d'exemple, 51 % des sinistres relatifs aux résidences secondaires proviennent du Sud, contre 45 % pour les résidences principales. Ces facteurs, combinés à l'effet volume, influencent considérablement le coût

moyen.

Le prochain chapitre sera dédié à la présentation des bases de données ouvertes et à leur préparation en vue de la modélisation.

Chapitre 2

Présentation et traitement des données ouvertes

L'importance des bases de données ouvertes ayant été précisée dans le chapitre précédent, ce chapitre sera consacré à leur présentation détaillée, ainsi qu'aux méthodes qui seront employées pour leur préparation et leur exploitation. Nous donnerons un aperçu des techniques et des défis associés à la manipulation de ces bases, garantissant ainsi qu'elles soient prêtes à être intégrées dans les bases de modélisation.

2.1 Base des arrêtés Catastrophes naturelles

2.1.1 Présentation

L'application GASPAR [1] de la Direction Générale de la Prévention des Risques (DGPR) occupe une position centrale dans le système d'information sur les risques naturels en France. Les bases de données GASPAR recensent les procédures administratives relatives aux risques. Les mises à jour de ces bases sont effectuées directement par les services départementaux ou régionaux ([Etalab, 2016]). Elles comprennent des informations sur divers documents, dont :

1. Les PPR (Plans de Prévention des Risques) naturels et assimilés, et les PPR technologiques ;
2. Les procédures de reconnaissance de l'état de catastrophes naturelles ;
3. D'autres documents d'information préventive tels que :
   - TIM : Dossier de Transmission d'Information au Maire ;
   - DICRIM : Document d'Information Communal des populations sur les Risques Majeurs ;
   - PCS : Plan Communal de Sauvegarde ;
   - AZI : Atlas des Zones Inondables ;
   - Arrêtés de catastrophes naturelles (arrêtés CATNAT [2]), actualisés dans les 30 jours après leur parution au Journal Officiel. Le Journal Officiel (souvent abrégé en « JO ») est la publication officielle de la République française où sont consignés tous les actes législatifs et réglementaires de la France. Il est édité par la Direction de l'information légale et administrative, une direction du Premier ministre.

Ainsi, grâce à GASPAR, des informations essentielles sur les risques naturels et technologiques sont centralisées et accessibles. Par conséquent, cela facilite la gestion des situations d'urgence, la prévention et la sensibilisation des populations locales face aux risques potentiels.

Il est ici question des arrêtés de catastrophe naturelle qui, par définition, sont « des arrêtés interministériels de reconnaissances de l'état de catastrophe naturelle délivrés pour un ensemble de communes, un aléa et une période donnée, après examen des demandes des maires concernés. » [3]

1. Gestion Assistée des Procédures Administratives relatives aux Risques naturels et technologiques
2. Cette abréviation sera souvent utilisée dans la suite du mémoire.
3. Définition selon le Dictionnaire des données GASPAR, georisques.gouv.fr

Au moment de l'extraction, ces données comprenaient une compilation complète des arrêtés de catastrophe naturelle émis entre juillet 1982 et avril 2022. Les données sont structurées sous la forme d'un tableau où chaque entrée correspond à un arrêté de catastrophe naturelle. Les entrées offrent des informations comme le code Insee de la commune, le département, le nom de la ville, la cause des dégâts, ainsi que les dates précises de début et de fin de la catastrophe. Il convient de souligner que si un événement affecte 500 communes, cela entraînera la publication de 500 arrêtés séparés.
Les détails des colonnes sont précisés dans le tableau ci-après [1] :

Colonne	*Description*
cod_nat_catnat	Code unique identifiant une reconnaissance (code national généré par Gaspar)
cod_commune	Code de la commune concernée
lib_commune	Nom de la commune concernée
num_risque_jo	Numéro du risque mentionné
lib_risque_jo	Libellé du risque mentionné dans le journal officiel
dat_deb	Date de début de l'événement
dat_fin	Date de fin de l'événement
dat_pub_arrete	Date de l'arrêté
dat_pub_jo	Date de publication au journal officiel
dat_maj	Date de mise à jour de la fiche GASPAR

FIGURE 2.1 – Dictionnaire de la base des arrêtés CATNAT, *georisques.gouv.fr*

Les périls climatiques recensés incluent, entre autres, inondations, sécheresse, tempête, secousses sismiques et poids de la neige. La carte suivante représente le nombre de catastrophes naturelles déclarées entre janvier 2011 et septembre 2021.

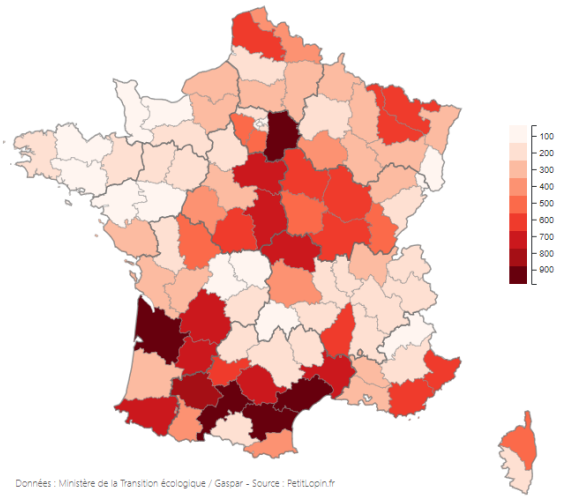


FIGURE 2.2 – Catastrophes naturelles entre janvier 2011 et septembre 2021. *Données : Ministère de la Transition écologique / Gaspar - Source : PetitLopin.fr*

D'après la source consultée pour obtenir la carte, le département ayant le plus d'arrêtés est l'Hérault (34) avec un total de 1082 arrêtés. La moyenne pour l'ensemble du pays est de 368 arrêtés, avec un écart-type de 256. À l'opposé, le département avec le moins d'arrêtés est Paris (75), avec 4 arrêtés au total.

L'enjeu est d'évaluer si un arrêté peut enrichir la prédiction des sinistres climatiques. Cette étude vise à déterminer leur pertinence pour anticiper ces événements. Le prochain paragraphe portera sur le traitement et l'exploitation de cette base.

1. Arrêtés de catastrophe naturelle en France métropolitaine, juin 2016, data.gouv.fr

2.1.2 Traitement

Comme mentionné dans la section 1.2.2, l'UP tempête représente la majorité des sinistres climatiques en nombre. Il semble donc pertinent d'analyser en premier lieu les arrêtés CATNAT liés aux tempêtes pour déceler une éventuelle corrélation avec les sinistres tempête entre 2016 et 2022 de la base de modélisation. Les corrélations de Pearson et de Spearman seront utilisées. La sélection des données pertinentes sera réalisée à l'aide de SAS, et les sinistres seront extraits de la base sinistre, comme référencé dans le paragraphe 1.4.1. Les arrêtés CATNAT relatifs aux tempêtes seront filtrés en utilisant le critère lib_risque_jo = « Tempête » (voir 2.1). Pour cette analyse, deux tableaux ont été élaborés, juxtaposant le nombre de sinistres tempête et d'arrêtés CATNAT tempête. Le premier tableau agrège les données par commune, tandis que le second le fait par département. Les cartes ci-dessous illustrent ces données départementales

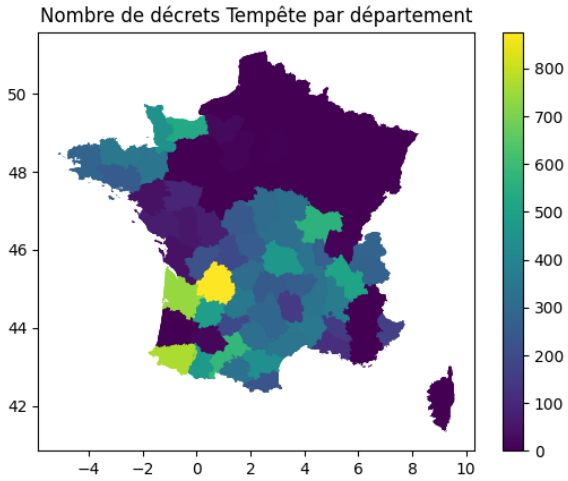


FIGURE 2.3 – Nombre de décrets tempête par département.

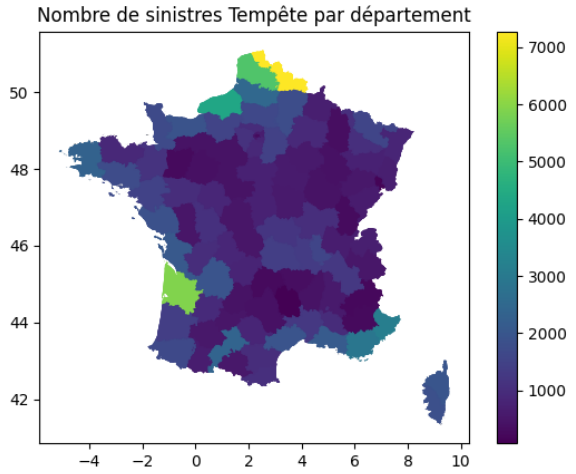


FIGURE 2.4 – Nombre de sinistres tempête entre 2016 et 2022 (Portefeuille MRH).

Visuellement, il est difficile de discerner une corrélation claire, bien que certains départements, comme la Gironde (33), affichent un nombre élevé de décrets et de sinistres tempête. En revanche, le département du Nord (59) se distingue avec un nombre élevé de sinistres mais sans aucun décret tempête associé. Au total, 15 195 décrets relatifs aux tempêtes sont recensés, principalement dans le Sud (notamment l'Ouest) et le Nord-Ouest.

À l'échelle communale, 14 114 zones du portefeuille MRH ont été touchées par un décret tempête. La commune du Bouscat (33069) détient le record avec 5 catastrophes naturelles liées aux tempêtes, mais avec 65 sinistres tempête enregistrés. À titre de comparaison, Bordeaux (33063), qui a le plus grand nombre de sinistres tempête du portefeuille, n'a déclaré qu'un seul décret de catastrophe naturelle pour une tempête. Il est intéressant de noter la proximité géographique entre Le Bouscat et Bordeaux, qui ne sont séparées que de 3 km.
Dans 4 278 communes (soit 12 %), aucun sinistre tempête n'est recensé, mais au moins un arrêté CATNAT est présent. À l'inverse, dans 13 372 communes, aucun arrêté n'est relevé, mais une fréquence non nulle de sinistres est observée. Cela met également en avant un manque de corrélation à cette échelle. Ces observations sont récapitulées dans la figure ci-dessous :

Nombre de décrets tempête	15 195		
Nombre de communes du portefeuille touchées par au moins un décret tempête	14 114	Nombre de départements du portefeuille touchées par au moins un décret tempête	62
Nombre de communes du portefeuilles n'ayant jamais eu de décrets	21 918	Nombre de départements du portefeuilles n'ayant jamais eu de décrets	38
Nombre de décrets maximal dans une commune	5	Nombre de décrets maximal dans un département	876
Commune correspondante	Le Bouscat (33069)	Département correspondant	Dordogne (24)
Nombre correspondant de sinistres tempêtes entre 2016 et 2022	65	Nombre correspondant de sinistres tempêtes entre 2016 et 2022	1 940
Nombre maximal de sinistres tempêtes dans une commune	372	Nombre maximal de sinistres tempêtes dans un département	7 275
Commune correspondante	Bordeaux (33063)	Département correspondant	Nord (59)
Nombre de décrets correspondant	1	Nombre de décrets correspondants	0
Distance entre Le Bouscat et Bordeaux (kilomètres)	3	Distance approximative Dordogne et Nord (kilomètres)	750
Nombre de communes avec 0 sinistres et au moins 1 arrêté	4 278	Nombre de départements avec 0 sinistres et au moins 1 arrêté	0
Nombre de communes avec 0 arrêté et au moins un sinistre	13 372	Nombre de départements avec 0 arrêté et au moins un sinistre	38

FIGURE 2.5 – Comparaison des sinistres tempête et des arrêtés CATNAT tempête par commune et par département

Il est à noter que, la corrélation de Pearson entre le nombre de sinistres tempête et le nombre de décrets tempête est de 0.1, et celle de Spearman est de 0.2. Ces résultats mathématiques s'expliquent par les analyses précédentes, mais aussi par les informations suivantes :

— En 1982, 75 % des décrets sont émis ;

— En 1983, 8 % des décrets sont émis ;

— En 1986, aucun décret n'est émis ;

— En 1987, 14 % des décrets sont émis ;

— En 1989, 3 % des décrets sont émis.

L'arrêté de catastrophe naturelle le plus récent lié à une tempête remonte à 1989. À ce jour, l'exploitation de cette base de données ne semble pas pertinente. Néanmoins, l'ambition d'évaluer l'impact de l'utilisation de données ouvertes sur le modèle fréquence demeure. Ainsi, la base de données Météonet, renseignant la vitesse des vents dans le Sud-Est et le Nord-Ouest de la France, sera examinée ultérieurement dans ce chapitre. La section suivante sera centrée sur l'analyse de la base de Demandes de Valeurs Foncières.

2.2 Base de Demandes de Valeurs Foncières (DVF)

2.2.1 Présentation

Lancé en avril 2019 par Etalab [1], le service Demande de Valeurs Foncières (DVF) a pour ambition de fournir en *Open Data* les informations relatives aux transactions immobilières réalisées à titre onéreux durant les cinq dernières années, couvrant le territoire métropolitain et les DOM-TOM. Seuls l'Alsace, la Moselle et Mayotte sont exclus de cette initiative. Ce service offre un accès instantané en ligne aux informations sur les mutations immobilières, optimisant la consultation et l'exploitation des données pour diverses études et analyses. Ces fichiers subissent des mises à jour deux fois par an, en avril et en octobre, et leur obtention ne nécessite aucune demande préalable. Il convient de mentionner que chaque mise à jour élimine les fichiers précédemment disponibles, ce qui garantit ainsi aux utilisateurs des données toujours à jour et pertinentes. Cette facilité est devenue possible grâce au décret n° 2018-1350 du 28 décembre 2018. Par conséquent, les parties concernées par les transactions ne peuvent pas s'opposer à la publication.

Pour les besoins de ce mémoire, les données brutes de la base DVF ne seront pas utilisées, mais plutôt aux données géolocalisées. Cette version améliorée est proposée par la DGFiP [2]. Elle offre un format standardisé et enrichi qui, contrairement à la base originale, donne accès aux coordonnées géographiques des propriétés vendues et standardise les données par année de transaction, facilitant ainsi leur réutilisation. Voici une liste des améliorations apportées par rapport aux données brutes [3] :

— « le format de téléchargement est en CSV avec séparateur virgule et encodage UTF-8 ;

— Renommage des colonnes pour un traitement informatique plus facile ;

— Normalisation des valeurs décimales (point comme séparateur décimal) ;

1. Administration publique Française qui a pour objectif d'améliorer le service public et l'action publique grâce aux données.

2. Direction générale des Finances publiques

3. Source : Demandes de valeurs foncières géolocalisées, data.gouv.fr, consulté le 14 juillet 2023

— Normalisation des codes INSEE (5 caractères) ;

— Normalisation des codes voie (FANTOIR) (4 caractères) ;

— Création d'un identifiant de parcelle compatible avec les fichiers cadastraux proposés par Etalab. Un fichier cadastral est un ensemble de documents cartographiques qui représentent la division d'une commune en unités de terrain, appelées parcelles. Ces parcelles sont identifiées par un numéro unique ;

— Date de mutation au format ISO-8601 ;

— Géocodage latitude/longitude à la parcelle en coordonnées WGS-84 : *World Geodetic System 1984*. Dans ce système, chaque point est défini par deux coordonnées (latitude et longitude). »

La description des colonnes de la base de données est illustrée dans la figure ci-dessous :

Nom de la colonne	Signification
id_mutation	Identifiant de mutation (non stable, sert à grouper les lignes)
date_mutation	Date de la mutation au format ISO-8601 (YYYY-MM-DD)
numero_disposition	Numéro de disposition
nature_mutation	Nature de la mutation
valeur_fonciere	Valeur foncière (séparateur décimal = point)
adresse_numero	Numéro de l'adresse
adresse_suffixe	Suffixe du numéro de l'adresse (B, T, Q)
adresse_code_voie	Code FANTOIR de la voie (4 caractères)
adresse_nom_voie	Nom de la voie de l'adresse
code_postal	Code postal (5 caractères)
code_commune	Code commune INSEE (5 caractères)
nom_commune	Nom de la commune (accentué)
ancien_code_commune	Ancien code commune INSEE (si différent lors de la mutation)
ancien_nom_commune	Ancien nom de la commune (si différent lors de la mutation)
code_departement	Code département INSEE (2 ou 3 caractères)
id_parcelle	Identifiant de parcelle (14 caractères)
ancien_id_parcelle	Ancien identifiant de parcelle (si différent lors de la mutation)
numero_volume	Numéro de volume
lot_1_numero	Numéro du lot 1
lot_1_surface_carrez	Surface Carrez du lot 1
lot_2_numero	Numéro du lot 2
lot_2_surface_carrez	Surface Carrez du lot 2
lot_3_numero	Numéro du lot 3
lot_3_surface_carrez	Surface Carrez du lot 3
lot_4_numero	Numéro du lot 4
lot_4_surface_carrez	Surface Carrez du lot 4
lot_5_numero	Numéro du lot 5
lot_5_surface_carrez	Surface Carrez du lot 5
nombre_lots	Nombre de lots
code_type_local	Code de type de local
type_local	Libellé du type de local
surface_reelle_bati	Surface réelle du bâti
nombre_pieces_principales	Nombre de pièces principales
code_nature_culture	Code de nature de culture
nature_culture	Libellé de nature de culture
code_nature_culture_speci	Code de nature de culture spéciale
nature_culture_speciale	Libellé de nature de culture spéciale
surface_terrain	Surface du terrain
longitude	Longitude du centre de la parcelle concernée (WGS-84)
latitude	Latitude du centre de la parcelle concernée (WGS-84)

FIGURE 2.6 – Colonnes des bases DVF géolocalisées. *Source : Demandes de valeurs foncières géolocalisées sur data.gouv.fr*

Il est à noter que la valeur foncière renseignée dans la base correspond à un prix « net vendeur », c'est-à-dire hors frais d'agence immobilière et frais de notaire. Si du mobilier est vendu avec le bien immobilier (par exemple, une cuisine équipée), sa valeur n'est pas incluse dans le prix affiché, bien que la TVA le soit. De plus, la variable « surface_reelle_bati » correspond à la surface réelle du bâtie et non pas à la surface Carrez [1], qui est présente dans le fichier mais rarement renseignée.

Les transactions renseignées dans cette base peuvent concerner des maisons, appartements, dépendances, locaux industriels, commerciaux, ou assimilés. Les dépendances associées à des résidences principales ou secondaires sont toujours listées séparément. Les données vont de 2017 à 2022. En effet, les données de 2017 ont été obtenues fin 2022, juste avant leur suppression. La première étape du travail a été de fusionner toutes ces données sur SAS, aboutissant à un total de 20 693 618 entrées. Le graphique ci-dessous illustre le nombre d'entrées disponibles chaque année :

1. La surface Carrez, nommée d'après la loi Carrez de 1996, correspond à la superficie des planchers des locaux clos et couverts après déduction des surfaces occupées par les murs, cloisons, marches et cages d'escaliers, gaines, embrasures de portes et de fenêtres.

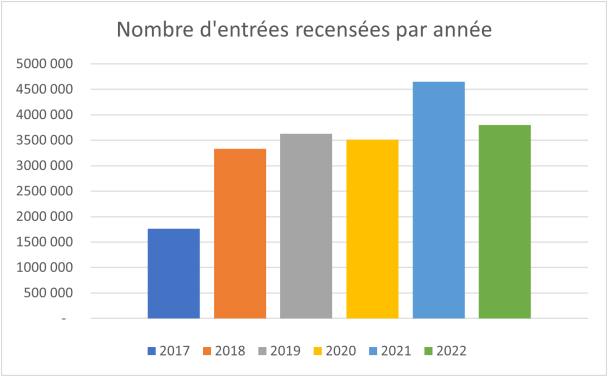


FIGURE 2.7 – Nombre d'entrées recensées dans la base par année

Il sera vu dans la sous-section suivante que le nombre d'entrées ne reflète pas nécessairement le nombre de transactions effectives.

2.2.2 Traitements

Afin de préparer les données pour la modélisation, une série d'étapes rigoureuses a été suivie afin d'assurer leur pertinence et leur fiabilité. Ces traitements ont été réalisés à l'aide de SAS.

Il est essentiel de rappeler que le cœur de cette étude est de déterminer dans quelle mesure l'intégration de cette base de données ouvertes peut optimiser la modélisation de la garantie climatique pour les propriétaires de maisons, qu'ils soient occupants ou non. L'hypothèse sous-jacente est que, dans une commune ou un IRIS où la valeur immobilière est supérieure à la moyenne, un sinistre pourrait engendrer des coûts plus importants. Pour étayer cette hypothèse, nous nous appuierons principalement sur la valeur foncière. De plus, un indicateur du prix au mètre carré sera élaboré en mettant en relation cette valeur foncière avec la surface réelle bâtie :

$$\text{prix\_m}^2 = \frac{\text{valeur\_fonciere}}{\text{surface\_reelle\_bati}}$$

Étape 1 : Sélection du type de local

Les valeurs foncières renseignées dans la colonne dédiée sont données par transaction. Ainsi, si une maison et une dépendance ont été vendues dans le cadre d'une même transaction, une unique valeur foncière sera attribuée à ces deux éléments [1]. Une entrée fera référence à la maison et une entrée à la dépendance. Il a été décidé de ne pas retenir les dépendances, car dans l'entreprise, elles relèvent d'une tarification spécifique. Par souci de cohérence avec le produit « Ma Maison » et pour avoir une valeur foncière spécifique à chaque maison proprement dite, le choix a été fait de maintenir les maisons (déterminées grâce à la variable type_local) qui sont associées à des transactions impliquant exclusivement des maisons. Après cette sélection, 3 876 153 entrées demeurent. Si les maisons incluses dans des transactions multiples avaient été conservées, 165 655 lignes supplémentaires auraient été ajoutées. Cette démarche est illustrée dans le schéma suivant :

id_mutation	date_mutation	nature_mutation	valeur_fonciere	type_local	nombre_pieces_principales	surface_reelle_bati	surface_terrain	longitude	latitude
2019-y	07/01/2021	Vente	890 000	Maison	10	350	400	5.209389	46.190084
2019-y	07/01/2021	Vente	890 000	Dépendance	0	-	400	5.209389	46.190084
2022-z	16/09/2022	Vente	200 000	Maison	4	110	800	4.941728	46.263782
2022-b	16/09/2022	Vente	500 000	Appartement	9	200	-	4.941799	46.263679

Filtration

id_mutation	date_mutation	nature_mutation	valeur_fonciere	type_local	nombre_pieces_principales	surface_reelle_bati	surface_terrain	longitude	latitude
2022-z	16/09/2022	Vente	200 000	Maison	4	110	800	4.941728	46.263782

FIGURE 2.8 – Sélection des Maisons

Le type de local à présent ciblé, le type de transaction sera choisi.

1. De même qu'une transaction concernant une maison et un appartement.

Étape 2 : Sélection du type de transaction

La base de données comprend divers types de mutations. Pour assurer la cohérence avec la valeur réelle de la propriété, seules les transactions renseignées comme étant des ventes ont été conservées, les autres types tels que les échanges ou les adjudications (Vente aux enchères) ont été exclus. Une fois les données filtrées en fonction du type de local et de transaction, nous nous sommes concentrés sur les valeurs manquantes.

Étape 3 : Valeurs manquantes

Les entrées avec des valeurs manquantes pour la « valeur foncière », principale variable d'intérêt, ont été éliminées, soit 8207 lignes. De la même manière, les colonnes affichant un nombre considérable de valeurs manquantes et sans pertinence pour l'étude ont été retirées de la base de données. Parmi ces colonnes, on retrouve :

— ancien code commune ;
— ancien nom commune ;
— lot1 numero ;
— lot1 surface carrez ;
— lot2 numero ;
— lot2 surface carrez ;
— lot3 numero ;
— lot3 surface carrez ;
— lot4 numero ;
— lot4 surface carrez ;
— lot5 surface carrez ;
— lot5 numero.

Par ailleurs, les lignes dépourvues de valeurs relatives à la surface du logement (indiquée dans la colonne « surface_reelle_bati »), une variable essentielle pour le calcul du prix au mètre carré, ont également été supprimées. Cette opération a entraîné la suppression de 150 602 lignes supplémentaires.

Étape 4 : Gestion des doublons

Il est à noter que la base de données contenait initialement 25 032 doublons parfaitement identiques. Ils ont été supprimés. Néanmoins, des doublons plus subtils se sont également manifestés en raison du fait que les valeurs foncières soient données par transaction. Ces doublons plus délicats à identifier sont illustrés et détaillés dans les schémas et exemples suivants :

— Cas 1 :

id_mutation	date_mutation	nature_mutation	valeur_fonciere	type_local	nombre_pieces_principales	surface_reelle_bati	surface_terrain	longitude	latitude
2018-x	16/09/2018	Vente	550 000	Maison	4	75	250	5.22044	46.200062
2018-x	16/09/2018	Vente	550 000	Maison	3	52	250	5.22044	46.200062

⇓

id_mutation	date_mutation	nature_mutation	valeur_fonciere	type_local	nombre_pieces_principales	surface_reelle_bati	surface_terrain	longitude	latitude
2018-x	16/09/2018	Vente	550 000	Maison	7	127	250	5.22044	46.200062

FIGURE 2.9 – Doublons de bâti

Il s'agit içi des transactions comprenant 2 maisons distinctes mais dans la même propriété (ou parcelle). Il a été choisi de les regrouper en une seule, en sommant la surface bâtie ;

— Cas 2 :

id_mutation	date_mutation	nature_mutation	valeur_fonciere	type_local	nombre_pieces_principales	surface_reelle_bati	nature culture	surface_terrain	longitude	latitude
2021-y	07/01/2021	Vente	258 000	Maison	4	117	sols	840	5.209389	46.190084
2021-y	07/01/2021	Vente	258 000	Maison	4	117	terrains d'agrément	551	5.209389	46.190084

⇓

id_mutation	date_mutation	nature_mutation	valeur_fonciere	type_local	nombre_pieces_principales	surface_reelle_bati	surface_terrain	longitude	latitude
2021-y	07/01/2021	Vente	258 000	Maison	4	117	1391	5.209389	46.190084

FIGURE 2.10 – Doublons de sols

Lorsque plusieurs types de sols impliquent l'achat d'une maison, ils apparaissent sous plusieurs lignes différentes. Le choix a été fait encore une fois de les regrouper.

— Cas 3 : Certaines transactions peuvent inclure plusieurs habitations. Bien que ces maisons puissent être situées dans des communes différentes, elles peuvent aussi être adjacentes. Afin de préserver la cohérence avec les objectifs de l'étude tout en conservant autant de lignes de données que possible, une décision a été prise de ne conserver que les transactions concernant l'achat de maisons situées sur un même cadastre. Ces maisons sont alors regroupées en une seule entrée. Deux maisons sont considérées comme appartenant à un même cadastre si les dix premiers caractères de leur identifiant de parcelle sont identiques. Un exemple concret de ce processus est exposé dans le schéma ci-dessous :

id_mutation	date_mutation	nature_mutation	valeur_fonciere	type_local	nombre_pieces_principales	surface_reelle_bati	code_commune	id_parcelle	surface_terrain	longitude	latitude
2020-a	16/09/2020	Vente	700 000	Maison	5	60	11111	11111000AZ0120	150	5.220604	46.193389
2020-a	16/09/2020	Vente	700 000	Maison	7	62	11111	11111000AZ0121	170	5.220605	46.193390
2020-b	06/12/2020	Vente	650 000	Maison	6	72	22222	22222000XY0150	180	5.209132	46.123578
2020-b	06/12/2020	Vente	650 000	Maison	6	53	22222	22222000AB0151	120	5.208416	46.125772
2020-c	18/12/2020	Vente	900 000	Maison	10	200	33333	33333000AC0500	300	5.20792	46.125619
2020-c	18/12/2020	Vente	900 000	Maison	7	180	44444	44444000AC0501	250	4.897598	46.341071

Filtration

id_mutation	date_mutation	nature_mutation	valeur_fonciere	type_local	nombre_pieces_principales	surface_reelle_bati	code_commune	surface_terrain	longitude	latitude
2020-a	16/09/2020	Vente	700 000	Maison	12	122	11111	320	5.220604	46.193389

FIGURE 2.11 – Doublons de parcelle

Étape 5 : Traitement des valeurs aberrantes

La gestion des valeurs aberrantes a été menée en tenant compte des considérations métiers. Les maisons dont la valeur foncière est inférieure à 65 000 euros ont été exclues de l'étude, cette fourchette de prix correspondant généralement aux mobil-homes ou à d'autres types d'habitations qui ne sont pas couverts par le produit « Ma Maison ». De même, les maisons dont la valeur foncière excède 2.5 millions d'euros ont également été écartées, cette limite correspondant au seuil des produits « Grandes Demeures » chez AXA. Après ces ajustements, l'analyse des boîtes à moustache (voir graphiques ci-dessous) n'a révélée aucune autre valeur aberrante.

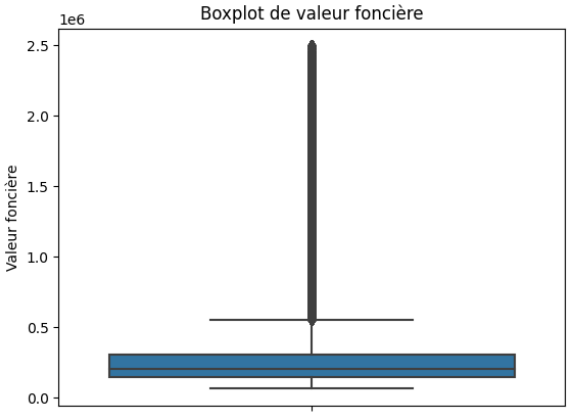


FIGURE 2.12 – Boxplot des valeurs foncières

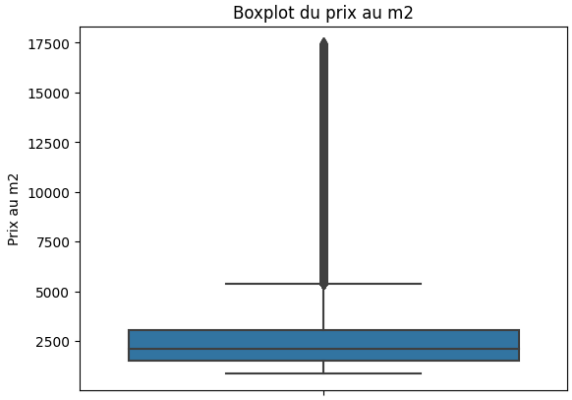


FIGURE 2.13 – Boxplot du prix au mètre carré

Finalement, la base d'étude recense 2 505 099 entrées, dont la répartition par année est données sur le graphique suivant :

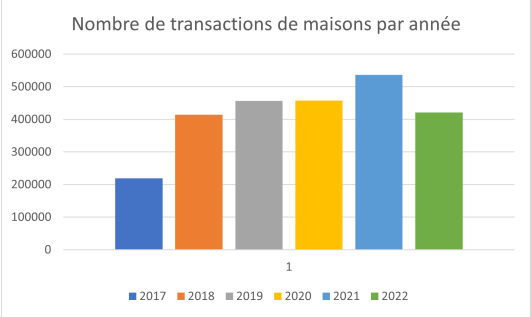


FIGURE 2.14 – Nombre de transactions de maisons par année

Une corrélation positive significative entre le nombre de transactions par ville et le nombre de sinistres climatiques a été observée. Cette corrélation a été mesurée en utilisant le coefficient de corrélation de Spearman. La valeur est de 0.568. Ce constat renforce l'intérêt de l'exploitation de la base DVF, et l'idée d'intégrer cette base dans la modélisation de la fréquence émerge. Le graphique suivant montre le nombre de transactions de maisons par département.

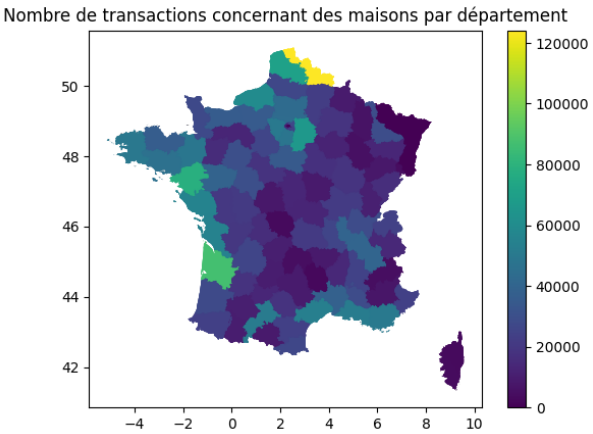


FIGURE 2.15 – Répartition du nombre de transactions de maisons par département en France

Globalement, c'est dans le département Nord (59) que le plus grand nombre de transactions de maisons est recensé. Un montant considérable est également observé en Gironde (33), Loire-Atlantique (44), Haute-Garonne (31), l'Hérault (34) et le Var (83) et en Seine-et-Marne (77).

Dans le paragraphe suivant, une série de graphiques sera présentée pour visualiser les transactions, les prix au mètre carré moyens, ainsi que les valeurs foncières moyennes par commune et par IRIS. Ces graphiques aideront à comprendre les tendances générales de ces variables.

2.2.3 Exploitation

Échelle départementale

À l'aide de SAS, les valeurs foncières et les prix au m^2 ont été agrégés par département. Les résultats sont illustrés sur le graphique ci-dessous.

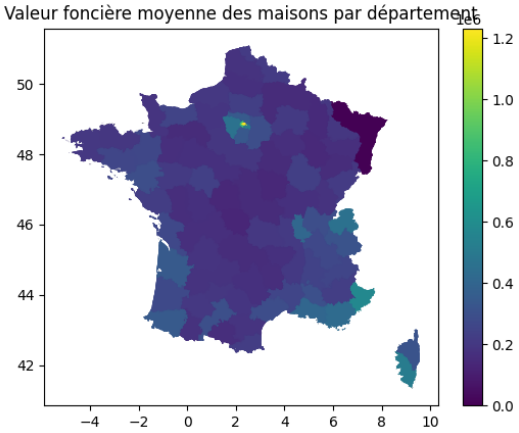


FIGURE 2.16 – Valeur foncière moyenne des maisons par département

Il est observé que les départements les plus « dynamiques », les plus denses, ou les plus touristiques affichent une valeur foncière moyenne plus élevée, ce qui est cohérent avec la réalité. Parmi eux figurent notamment Paris, les autres départements d'Île-de-France (surtout les Yvelines), le Var (83), les Alpes-Maritimes (06), les Bouches-du-Rhône (13), la Corse (2A et 2B), la Gironde (33) et le Rhône (69). Il est nécessaire de rappeler que la région Alsace n'est pas renseignée par la base DVF.

Échelle communale

De même, les valeurs foncières et les prix au m^2 ont été agrégés par commune. Il convient de rappeler que l'étude se concentre sur la France métropolitaine. Notons que 33 046 communes différentes sont renseignées dans la base DVF. [1]

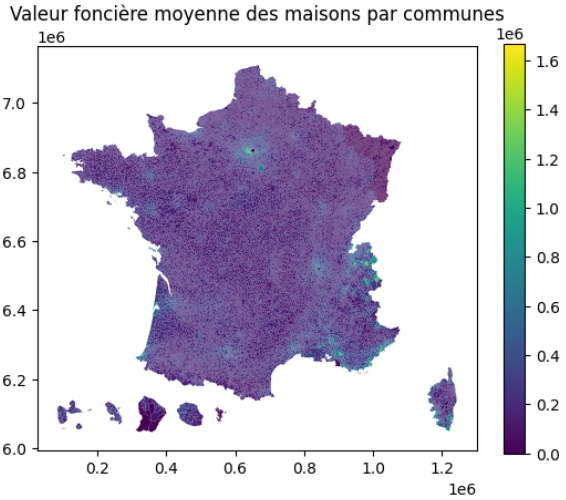


FIGURE 2.17 – Valeur foncière des maisons par communes

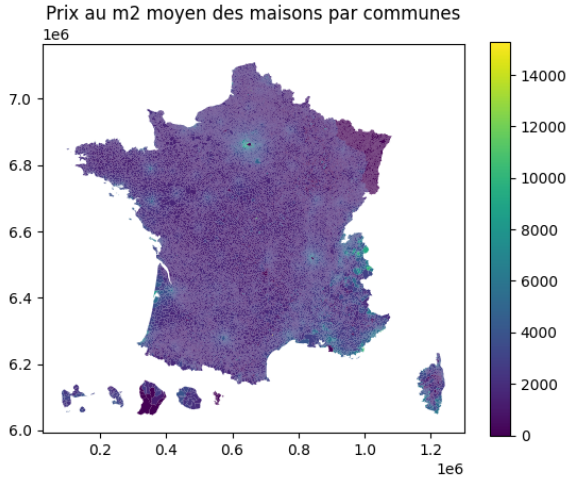


FIGURE 2.18 – Prix au m^2 des maisons par communes

Le graphique montre une tendance similaire à celle observée précédemment à l'échelle départementale. En effet, les communes les plus denses et touristiques affichent des valeurs foncières moyennes et des prix au mètre carré plus élevés, ce qui est cohérent avec la réalité. La distribution du nombre de transactions par commune est illustrée par le diagramme à moustache ci-dessous, mettant en avant la répartition inégale du nombre de transactions par communes au sein du territoire :

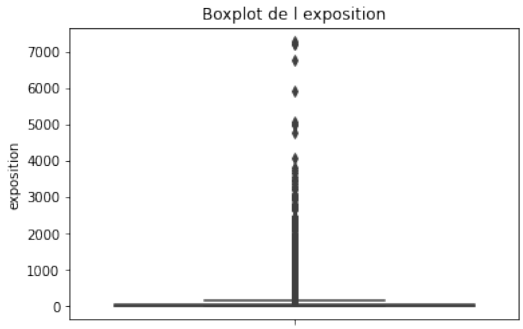


FIGURE 2.19 – Diagramme à moustache du nombre de transactions par INSEE

Bordeaux détient le record avec plus de 7000 transactions entre 2017 et 2022.

Échelle de l'IRIS

Un agrégation par IRIS a également été effectuée. Avant de mettre en avant les résultats, il est nécessaire de rappeler la définition d'un IRIS.

Définition 2.2.1. *Selon l'INSEE, les IRIS, ou Ilots Regroupés pour l'Information Statistique, sont définis comme des unités territoriales décrivant des « micro-quartiers ». Ces entités, composées de blocs contigus et homogènes, regroupent au moins 2 000 habitants. Ces « unités élémentaires » sont utilisées pour la collecte et l'analyse des données statistiques et démographiques. Il est recensé 48 606 IRIS en France métropolitaine.* [2]

Afin de déterminer de façon précise le code IRIS associé à chaque maison, l'outil interne GeoPoint a été utilisé. GeoPoint est un outil Python qui permet de trouver le niveau de risque d'une localisation

1. Les communes d'Outre-Mer sont renseignées dans les graphiques, mais seulement à titre informatif.
2. D'après : INSEE, Découpage infra-communal paru le 17/09/2020.

à partir d'un shapefile, où ici, le niveau de risque est l'IRIS. L'objectif de GeoPoint est expliqué dans le schéma ci-dessous :

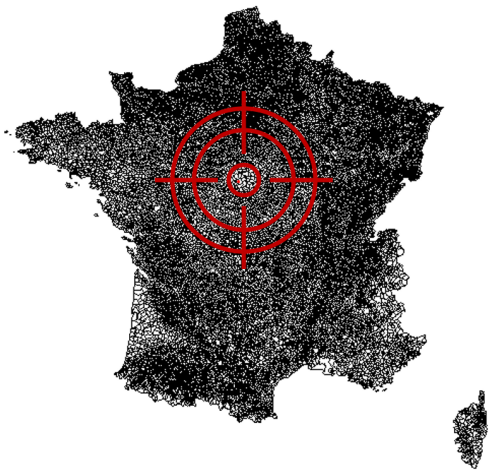


FIGURE 2.20 – Objectif de GeoPoint : Identifier le polygone contenant un point donné. *Source : document interne*

Un shapefile est un fichier composé d'informations sur des polygones et des niveaux de risque associés. Le shapefile des contours IRIS mis à disposition par le gouvernement, a été utilisé [1]. Une conversion de coordonnées a été nécessaire, réalisée grâce à la bibliothèque `pyproj` de Python. En effet, les coordonnées des sommets de chaque zone IRIS sont données en coordonnées côniques Lambert-93. Hors, les coordonnées disponibles dans la base DVF sont en coordonnées projetées WGS-84. Il a ainsi été choisi pour des raisons pratiques de convertir les coordonnées des sommets Lambert-93 en coordonnées WGS-84. Après cette étape, GeoPoint a été exécuté, prenant plus de 25 minutes. 43 890 IRIS sont renseignés dans la base DVF. Des exemples de résultats pour Paris sont présentés ci-dessous :

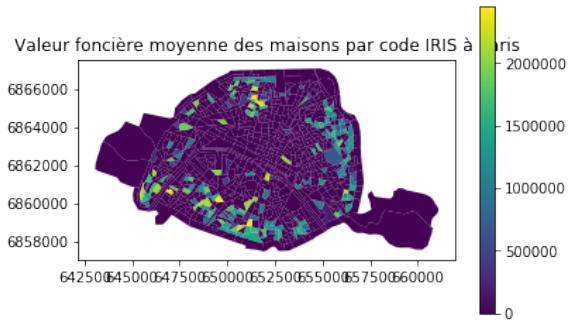


FIGURE 2.21 – Valeur foncière moyenne des maisons par IRIS sur Paris

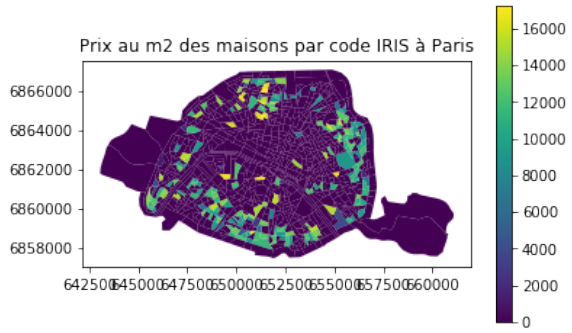


FIGURE 2.22 – Prix au m^2 moyen des maisons par IRIS sur Paris

Il y a un nombre relativement élevé d'IRIS manquants à Paris. Cette présence de données manquantes peut s'expliquer en partie par le fait que la majorité des transactions immobilières de cette ville concerne des appartements plutôt que des maisons à proprement parler. Cela met en avant le fait qu'il existe une limitation dans l'agrégation par cette maille, en raison d'un nombre limité voir insuffisant de données pouvant fausser les valeurs agrégrées :

1. Disponible sur le site de geoservice : https ://geoservices.ign.fr/contoursiris

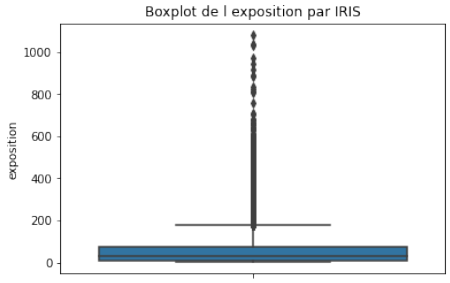


FIGURE 2.23 – Diagramme à moustache du nombre de transactions par IRIS, base DVF

Ainsi, une interrogation majeure émerge : est-il pertinent d'intégrer les données de la base DVF à la maille IRIS ? En effet, bien que l'agrégation par IRIS permette d'obtenir des informations sur les variables d'intérêt à une échelle micro-territoriale, elle pose également des questions sur la représentativité et la précision des données. Toutefois, notons que la répartition des données par IRIS semble plus équilibrée qu'avec l'INSEE.

2.2.4 Lissage des données

Un lissage des données agrégées sera effectué.
En modélisation, le lissage est un concept essentiel. Premièrement, le lissage permet de réduire le bruit inhérent aux données. Ce bruit peut être la conséquence d'erreurs de mesure, d'anomalies ou d'autres sources de variations aléatoires. Par conséquent, le lissage aide à rendre les modèles plus robustes, en les rendant moins sensibles à ces fluctuations aléatoires. Il convient toutefois de noter que l'utilisation du lissage doit être prudente et réfléchie. En effet, un lissage excessif peut éliminer des informations importantes. Il est donc essentiel de trouver un équilibre approprié lors de l'application de techniques de lissage.
Deux méthodes seront étudiées ici : le lissage par proches-voisins (*knn*) et celui de l'outil GeoRev. Ces méthodes permettront de lisser les données existantes et d'attribuer des valeurs aux zones INSEE et IRIS dépourvues de données.

Lissage k-proches voisins

L'algorithme des k plus proches voisins ou *k-nearest neighbors* (*k-nn*), est une méthode d'apprentissage supervisée non paramétrique basée sur la notion de proximité pour réaliser des prédictions. Cette technique, ayant trouvé ses origines dans un article de 1951 écrit par Evelyn Fix et Joseph Hodges, a été développée par Thomas Cover dans ses recherches sur les modèles de classification des voisins les plus proches. Dans le contexte de la régression, le *k-nn* prédit la valeur d'une observation en se basant sur la moyenne des valeurs de ses k $\in \mathbb{N}$ voisins les plus proches. L'efficacité de cette méthode dépend essentiellement de la manière dont la distance entre les observations est définie ([ibm, 2023]).

Il convient de noter que le *k-nn* est parfois qualifié d'« algorithme d'apprentissage paresseux », car il n'exécute pas de phase d'entraînement explicite. Au lieu de cela, il stocke l'ensemble des données d'apprentissage et effectue tous les calculs lors de la prédiction. Bien que cela rend l'algorithme simple à comprendre et précis (dans certains cas), il peut devenir inefficace lorsque la taille des données d'apprentissage augmente, ce qui peut compromettre les performances globales du modèle, et par conséquent la modélisation finale qui sera effectuée dans le cadre de ce mémoire.
Néanmoins, l'un des principaux avantages du *k-nn* est sa simplicité en termes de paramétrage. Il nécessite principalement la définition de k (le nombre de voisins), et d'une mesure de distance.
Pour cette étude, la distance euclidienne a été choisie, adaptée aux données en coordonnées projetées WGS-84. On note x le point d'étude et y un autre point du plan quelconque. La formule de la distance euclidienne est la suivante :

$$D(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}, \quad Ici\ n = 2 \tag{2.1}$$

Le choix de la valeur de k dans l'algorithme a un impact direct sur la performance du modèle. Des valeurs relativement faibles de k peuvent conduire à un surajustement (ou surapprentissage) des données : le modèle s'adaptera excessivement « bien » aux données d'apprentissage au détriment de sa capacité à généraliser à partir de nouvelles données. Inversement, des valeurs trop élevées de k peuvent conduire à un lissage excessif des valeurs prédites. Il est donc essentiel de trouver un juste milieu pour la valeur de k afin d'optimiser la performance du modèle en minimisant son erreur de prédiction sur de nouvelles données.

Le choix du nombre de voisins a été déterminé par validation croisée (*cross validation* en anglais) sur les données existantes d'INSEE/IRIS. Avant de fournir les résultats, il est essentiel de définir cette méthode.

Définition 2.2.2. *La validation croisée est une technique utilisée pour évaluer les performances d'un modèle en utilisant différentes partitions des données d'entraînement et de test. Elle permet de s'assurer que le modèle est robuste et peut généraliser à de nouvelles données. Le processus général de la validation croisée consiste à :*

1. *Diviser l'ensemble de données en p groupes différents (folds en anglais). Il est arbitrairement choisi p = 4, nombre couramment utilisé dans l'équipe. Cette décision est guidée par la dimension des bases de données agrégées, permettant d'équilibrer entre une évaluation précise du modèle et une optimisation du temps de calcul ;*

2. *Pour chaque fold :*

   — *Utiliser ce fold comme jeu de données de test.*

   — *Utiliser les 3 (p - 1) autres folds comme jeu de données d'entraînement.*

   — *Entraîner un modèle sur l'ensemble d'entraînement et évaluer ses performances sur l'ensemble de test. La métrique de performance utilisée ici sera la RMSE. La **RMSE** (**R**oot **M**ean **S**quared **E**rror), ou erreur quadratique moyenne, mesure l'écart moyen entre les valeurs réelles et les valeurs prédites par le modèle :*

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2} \tag{2.2}$$

   où :

   — *m représente le nombre d'observations dans le jeu de données ;*

   — *y_i est la valeur réelle de l'observation i ;*

   — *$\hat{y}_i$ est la valeur prédite par le modèle pour l'observation i.*

   Une RMSE plus faible indique un modèle plus précis, car cela signifie que les prédictions du modèle sont plus proches des valeurs observées.

3. *Calculer la moyenne du RMSE sur tous les folds afin d'obtenir une estimation globale des performances du modèle.*

Le principe est résumé dans le graphique suivant (cas de 5 folds) :

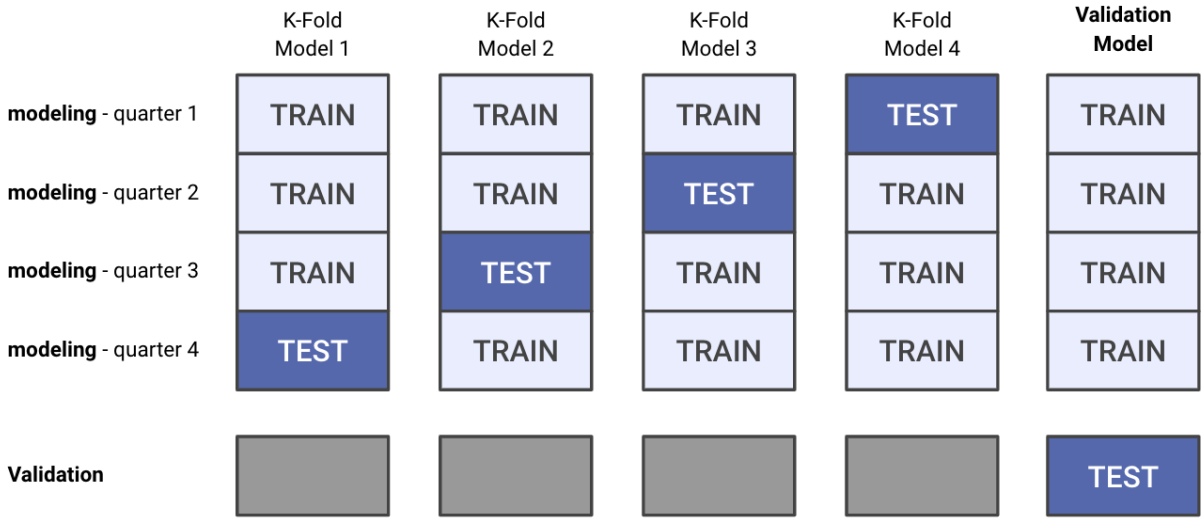


FIGURE 2.24 – *Cross validation*, source : Akur8

Ainsi, en divisant les données en plusieurs sous-ensembles et en entraînant et testant le modèle sur différents arrangements de ces sous-ensembles, la cross-validation permettra d'estimer la performance du modèle pour différentes valeurs de k.
Le nombre de voisins k qui offrira la RMSE la plus basse sans induire de surapprentissage sera retenu, tout en veillant à ce que la représentation visuelle sur la carte IRIS et INSEE soit jugée satisfaisante. En effet, afin de rester cohérent avec la réalité, les zones les plus denses doivent présenter des prix au m^2 et des valeurs foncières moyennes plus élevées que les autres.
Des valeurs de k comprises entre 1 et 100 seront testées. Le choix d'un maximum de 100 a été jugé adapté compte tenu des niveaux de maille d'agrégation INSEE et IRIS. Ce choix a également été influencé par des expériences antérieures avec des données analogues.
La mise en place du lissage en Python a été effectuée à l'aide de la bibliothèque `scikit-learn`. Les coordonnées des INSEE ont été récupérées sur le même site que celui qui fournit les contours IRIS, et les coordonnées IRIS correspondent aux centroïdes des contours des polygones IRIS (cf section 2.2.3).
Le récapitulatif des paramètres obtenus est donné dans le tableau suivant :

KNN				
Maille	Variable	Plage de k	k retenu	RMSE correspondant
INSEE	Valeur Foncière	(1; 100)	7	48104,13
	Prix au m2	(1; 100)	7	357,41
IRIS	Valeur Foncière	(1; 100)	9	79171,29
	Prix au m2	(1; 100)	6	536,29

FIGURE 2.25 – Choix des paramètres du knn

Les résultats cartographiques sont les suivants :

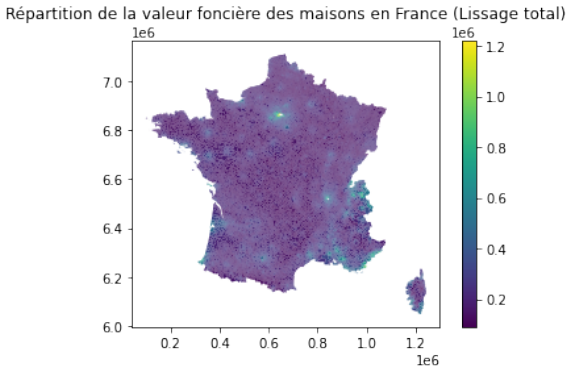


FIGURE 2.26 – Valeur foncière lissée par INSEE (*knn*)

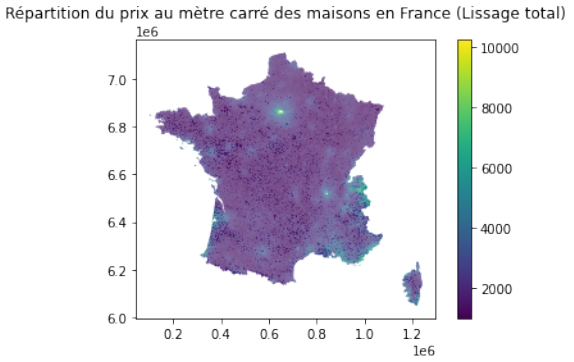


FIGURE 2.27 – Prix au m^2 lissé par INSEE (*knn*)

Sur les résultats du lissage effectué par zones INSEE, il est remarqué que les prix au m^2 supérieurs à 11 000 sont « perdus », caractéristiques des biens situés à Paris.

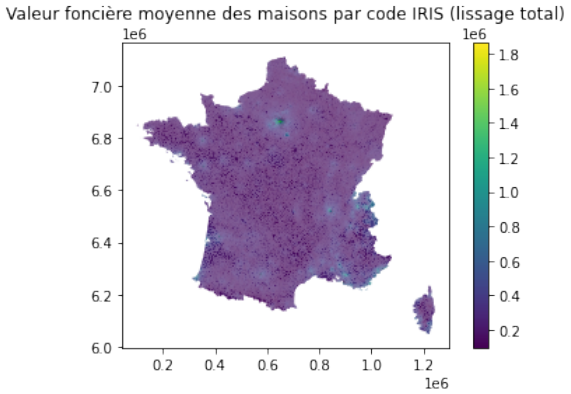


FIGURE 2.28 – Valeur foncière lissée par IRIS (*knn*)

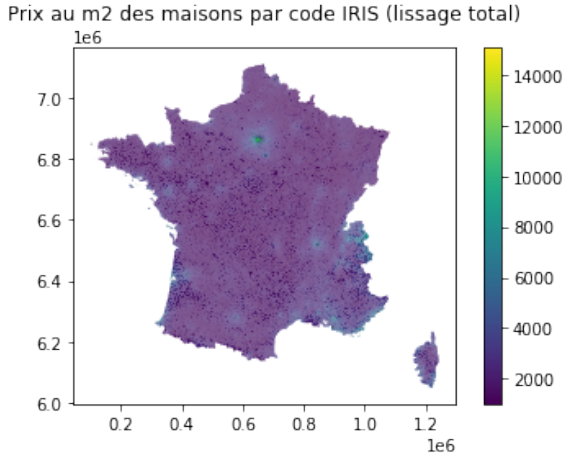


FIGURE 2.29 – Prix au m^2 lissé par IRIS (*knn*)

Le lissage par IRIS permet de garder une plus grande plage de valeur que le lissage par INSEE, ce qui est normal étant donné la maille plus fine. L'Alsace a des valeurs attribuées, mais elles semblent plus basses que la réalité : le prix au m^2 attribué vaut en moyenne 1900 contre 3957 €[1] dans la réalité. Même si le calcul du prix au m^2 faisant l'objet de ce mémoire n'est pas calculé de la même façon que celui de la réalité, nous cherchons néanmoins à nous en rapprocher au le plus possible.

Nous choisirons ensuite de quantiliser les résultats du lissage en 20 zones. Ce seront ces données qui seront utilisées pour le GLM. La raison de ce choix est similaire à celle qui a justifié la quantilisation dans les traitements massifs de la partie 1.4.5. Les résultats sont donnés ci-après :

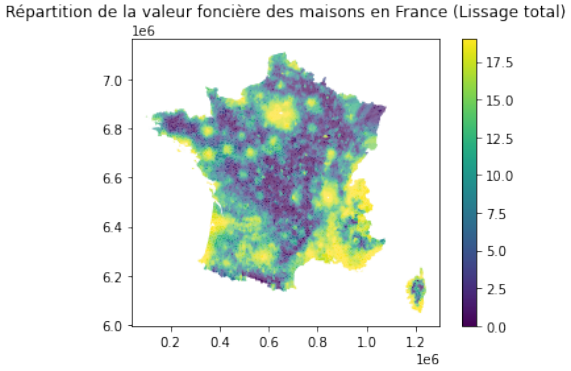


FIGURE 2.30 – Valeur foncière lissée et quantilisée par INSEE (*knn*)

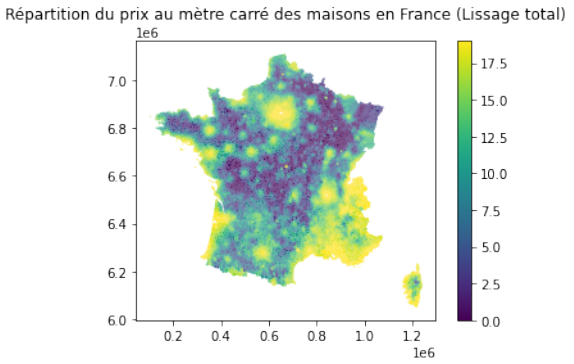


FIGURE 2.31 – Prix au m^2 lissé et quantilisé par INSEE (*knn*)

1. D'après SeLoger.com

L'Île-de-France et la Vendée sont regroupées dans la même zone. De plus, bien que le Drôme et le Var soient dans la même zone, le Var affiche des prix moyens nettement supérieurs, avec une différence notable de 2000 € pour le prix au m^2. Il serait envisageable d'utiliser davantage de quantiles pour une répartition plus précise. Cependant, cela pourrait complexifier l'analyse. Ainsi, nous choisirons de conserver ces résulats. Cette décision est également motivée par le fait que la prochaine méthode de lissage qui sera étudiée crée également des zones par quantilisation, permettant ainsi une comparaison sur une base identique.

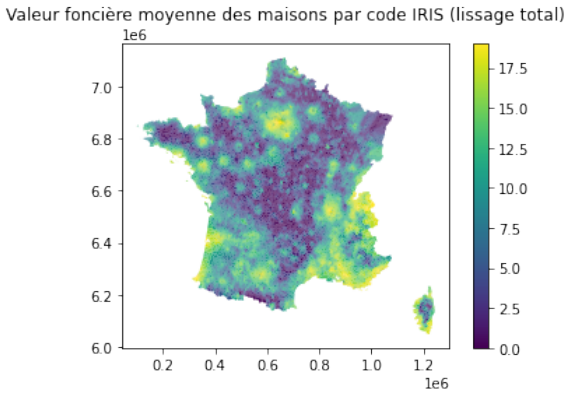


FIGURE 2.32 – Valeur foncière lissée et quantilisée par IRIS (*knn*)

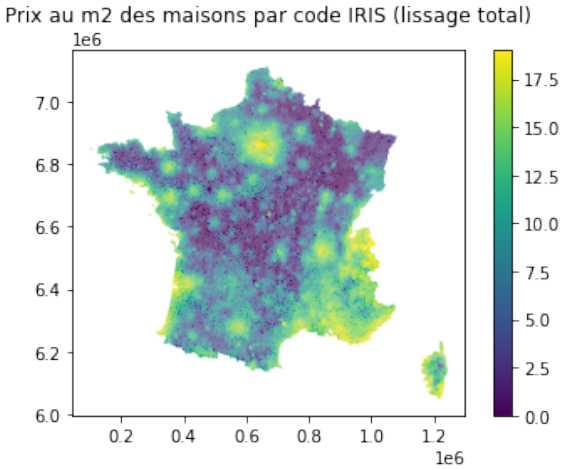


FIGURE 2.33 – Prix au m^2 lissé et quantilisé par IRIS (*knn*)

À l'échelle de l'IRIS, les valeurs attribuées au département Haut-Rhin (en Alsace) sont plus proches de la réalité.

Le prochain paragraphe aborde une autre technique de lissage, une méthode interne à l'entreprise appelée GeoRev. Il est espéré que cette méthode offre des plages de valeurs plus larges et des estimations plus réalistes pour certaines régions comme l'Alsace.

Lissage avec l'outil GeoRev

L'outil GeoRev est un outil de lissage qui allie à la fois la notion de distance, similaire aux *k-nn*, et la notion de crédibilité des données. La crédibilité, dans le contexte de l'outil GeoRev, fait référence à la fiabilité ou à la confiance accordée aux données d'une zone géographique spécifique. Elle est déterminée en fonction de l'exposition (ici le volume de transactions) dans cette zone. Ainsi, une zone avec un grand nombre de transactions est considérée comme ayant une haute crédibilité, car elle offre une représentation plus robuste et fiable. Inversement, une zone avec peu de transactions pourrait être considérée comme moins crédible, car elle est susceptible d'être influencée par des anomalies ou des variations temporaires. Comme pour le cas des *k-nn*, 20 zones de risque seront définies. Cela permettra de comparer les deux méthodes.
La distance utilisée par cet outil est la distance de Haversine. Elle correspond à une formule qui calcule la distance entre deux points sur la surface d'une sphère (la Terre ici), en utilisant leurs coordonnées de latitude et de longitude. Elle est donnée par :

$$d = 2r \times \arcsin\left(\sqrt{\sin^2\left(\frac{\Delta\text{lat}}{2}\right) + \cos(\text{lat}_1) \cdot \cos(\text{lat}_2) \cdot \sin^2\left(\frac{\Delta\text{long}}{2}\right)}\right)$$

où :

— Δlat est la différence de latitude entre les deux points (en radians) ;

— Δlong est la différence de longitude entre les deux points (en radians) ;

— r est le rayon de la sphère (ici la Terre).

La méthode de GeoRev est mise en œuvre en trois étapes. Elle sera détaillée pour la maille INSEE, l'approche étant la même à la maille IRIS ([TOESCA, 2020]).

1. Attribution d'un niveau de risque initial de 1 à 20 pour chaque INSEE ;
2. Complétion et lissage de la carte : Détermination d'une note lissée (1 à 20) pour chaque INSEE en France ;
3. Détermination finale du niveau de risques au sein de chaque INSEE : Attribution d'un niveau de risque final de 1 à 20 pour chaque INSEE.

Lors de l'étape 1, un coefficient spécifique est défini pour chaque polygone [1] :

$$\begin{aligned}\text{Coefficient du polygone} &= \text{Coefficient de crédibilité du polygone} \times \text{Note du polygone} \\ &+ (1 - \text{Coefficient de crédibilité du polygone}) \times \left(1 - \frac{\sum_{\text{voisin}=1}^{n} f(\text{Voisin}) \times \text{Note du voisin}}{\sum_{\text{voisin}=1}^{n} f(\text{Voisin})}\right)\end{aligned} \tag{2.3}$$

où :

— Le **coefficient de crédibilité du polygone** (cc) représente un facteur permettant de déterminer l'importance des zones avec le plus de contrats par rapport aux autres zones. Par exemple plus il y a d'exposition dans une commune, plus on peut « croire » directement le niveau de risque initial de cette commune ;

— La **note du polygone** est la note associée à la zone considérée (1 à 20) ;

— La **note du voisin** est la note associée à chaque zone voisine ;

— **f(Voisin)** est une fonction attribuant un poids à chaque voisin : on accorde plus d'importance aux zones voisines qui ont des nombres de contrats proches de la zone considérée ;

— **n** est le nombre total de zones voisines considérées.

Le coefficient de crédiblité (cc) se calcule à l'aide de la fonction suivante :

$$g(x) = \min\left(\frac{x^3}{\text{quantile}(99.5)^3}, 1\right) \quad \text{où x représente l'exposition}$$

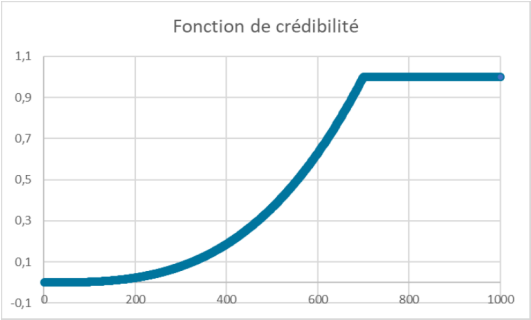


FIGURE 2.34 – Fonction de crédibilité de GeoRev

Ainsi, par cette fonction, une confiance totale (soit un coefficient de crédibilité de 1) est attribuée aux INSEE dont l'exposition est supérieure ou égale au quantile à 99,5 % des expositions observées. L'exposition est le nombre de transactions ici. Le fait que ce soit un polynôme du troisième degré permet une augmentation rapide du coefficient en fonction de l'exposition.

Concernant la fonction de pondération des voisins **f(Voisin)**, son principe est le suivant : deux voisins doivent avoir une note semblable s'ils sont similaires en termes de distance géographique et d'exposition :

$$f(Voisin) = f_{distgeo}(Voisin) \times f_{exposition}(Voisin) \tag{2.4}$$

Où

$$f_{distgeo}(Voisin) = \min\left(1, \frac{1}{\sqrt[3]{x/50}}\right) \quad \text{(Fonction positive et décroissante.)}$$

1. INSEE ou IRIS

Plus le polygone du voisin est proche du polygone étudié plus sa note doit être « impactante ». La racine cubique étant utilisée, la fonction tend rapidement vers 0 quand la distance augmente. Pour l'exposition :

$$f_{exposition}(Voisin) = \min\left(1, \frac{1}{\sqrt[3]{(|x-m|+1)/50}}\right)$$

Avec

— x : exposition du voisin

— m : exposition du polygone considéré

La fonction est symétrique par rapport à l'axe x, et donc f est tel que $f_{exposition}(x-m) = f_{exposition}(x+m)$. Cette fonction permet d'accorder plus d'importance aux villes voisines semblables à la ville de référence. Les courbes associées aux deux fonctions sont les suivantes :

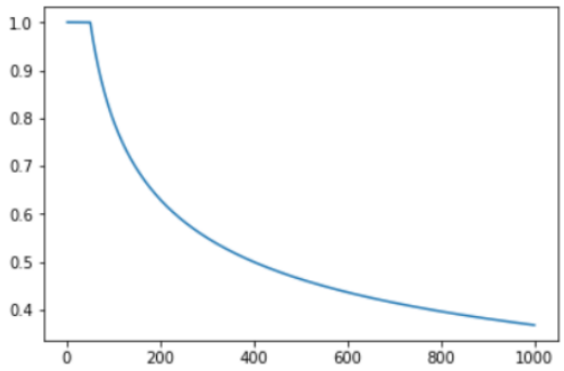


FIGURE 2.35 – Distance géographique. *La distance donnée en abscisse est en km. Source : [TOESCA, 2020]*

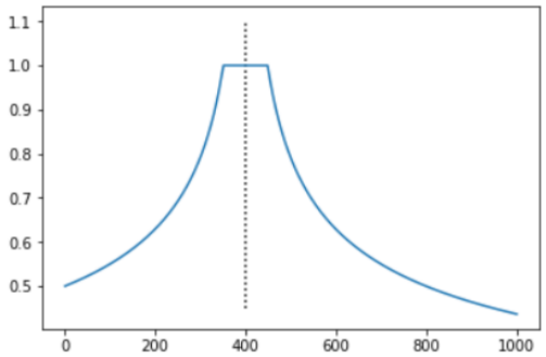


FIGURE 2.36 – Exposition. *La distance donnée en abscisse est en km. Source : [TOESCA, 2020]*

Ainsi, la fonction GeoRev disponible dans la bibliothèque `georev` de Python prend en entrée une table contenant :

— Les variables d'intérêt (valeurs foncières, prix au m^2 , code INSEE, code IRIS) ;

— L'exposition de chaque ville (volume de transactions).

Les paramètres d'entrée sont :

— La fonction de crédibilité de la note du polygone ;

— La fonction de pondération des voisins ;

— Le minimum d'exposition (ou de transactions) requis parmi les voisins : expo_min.

— Le minimum de voisins souhaité. Le nombre final de voisins par polygone sera le maximum entre nmin et le nombre minimal de voisins permettant de considérer au moins 1 % de l'exposition totale : nmin.

Le choix des paramètres nmin et expo_min a été fait de manière exhaustive. Pour chaque maille (INSEE ou IRIS) et chaque variable d'intérêt (valeur foncière ou prix au m^2), les plages d'entiers d'expo_min et nmin respectivement compris entre [quantile d'exposition 50 %,quantile d'exposition 65 %] et [1,100] ont été testées. Le choix de la plage de ces quantiles pour expo_min est expliqué par la volonté de trouver un équilibre entre un lissage excessif et un lissage insuffisant. Le critère de choix final des paramètres sera le même que celui des *k-nn* : ce sont les paramètres qui vont minimiser la RMSE mais qui permettront également d'avoir un résultat satisfaisant d'un point de vue métier (les grandes villes doivent rester les endroits les plus « chers »). Les choix finaux des paramètres sont recensés dans le tableau suivant :

GeoRev							
Maille	Variable	Plage de nmin	nmin retenu	Plage de expomin	expomin retenu	RMSE correspondant	Temps d'exécution par combinaison (en minutes)
INSEE	Valeur Foncière	(1; 100)	10	(19; 35)	19	48833,89	19,34
	Prix au m2	(1; 100)	10	(19; 35)	19	383,86	19,22
IRIS	Valeur Foncière	(1; 100)	10	(23; 45)	23	88850,06	40,10
	Prix au m2	(1; 100)	10	(23; 45)	23	656,67	40,06

FIGURE 2.37 – Paramètres de GeoRev

Les résultats cartographiques donnent :

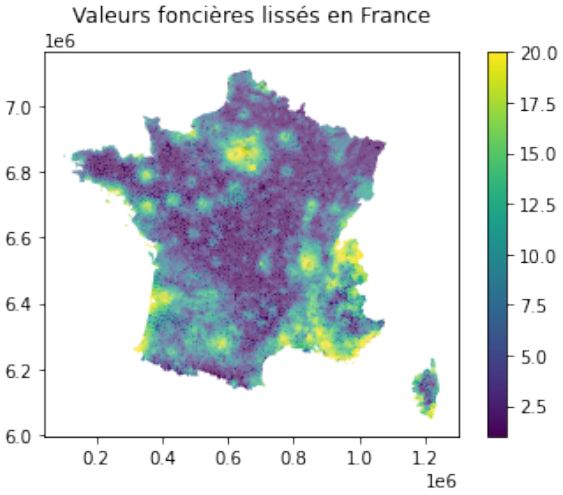


FIGURE 2.38 – Valeurs foncières lissé par IRIS (GeoRev)

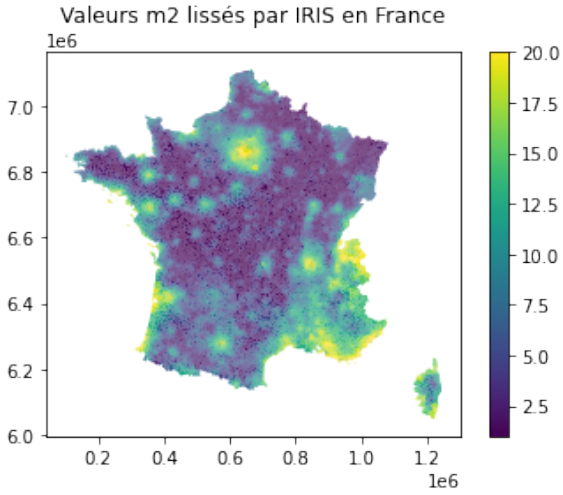


FIGURE 2.39 – Prix au m^2 lissé par IRIS (GeoRev)

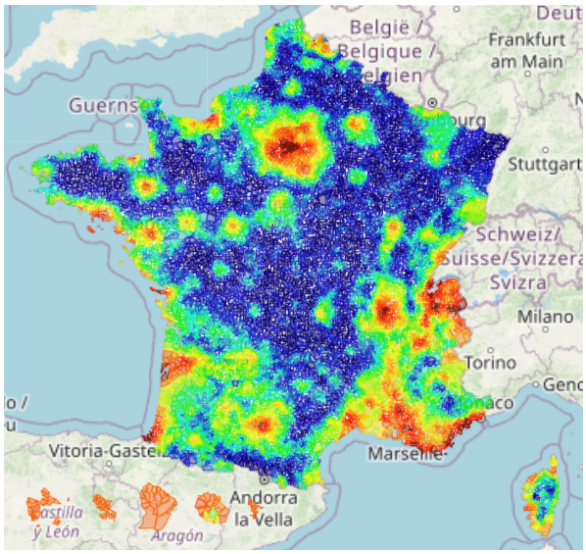


FIGURE 2.40 – Valeurs foncières lissé par INSEE (GeoRev), bibliothèque `maplot`

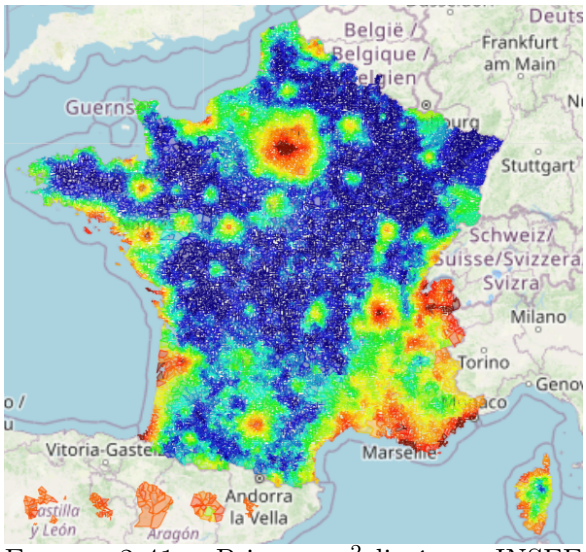


FIGURE 2.41 – Prix au m^2 lissé par INSEE (GeoRev), bibliothèque `maplot`

FIGURE 2.42 – Echelle

Les résultats obtenus par ce lissage sont cohérents avec la réalité, notamment dans le cas des grandes villes, et sont globalement en accord avec les observations réelles. Cette fois-ci, et contrairement aux *k-nn*, Paris et la Vendée ne sont pas dans la même zone. De plus, la représentation du Haut-Rhin est nettement plus fidèle pour les deux mailles concernées. Même si les performances sont relativement plus faibles que celles de la méthode *k-nn* (RMSE), l'intégration du facteur de crédibilité est un avantage non négligeable. Il a donc été décidé d'adopter les résultats fournis par la méthode GeoRev pour la suite du mémoire. Les résultats ont été fusionnés à l'aide de SAS à la base de modélisation avec pour

clé de jointure l'INSEE et l'IRIS.

La section suivante sera consacrée à la présentation de la base MétéoNet.

2.3 Base MétéoNet

2.3.1 Présentation

MétéoNet est une base de données météorologiques élaborée et proposée par MÉTÉO FRANCE, le service national de météorologie en France. L'ambition de cet organisme est d'offrir un ensemble de données accessible et immédiatement utilisable aux *Data Scientists* souhaitant explorer le domaine météorologique.

L'ensemble de données comprend des données d'observation au sol, d'observations des radars de pluie et de prévisions de modèles météorologiques. Chaque paramètre est mesuré toutes les 6 minutes. Les données couvrent deux zones géographiques spécifiques de 550 km x 550 km, situées dans le Nord-Ouest et le Sud-Est de la France, et s'étendent sur une période de trois ans, de 2016 à 2018 ([Météo-France, 2019]).

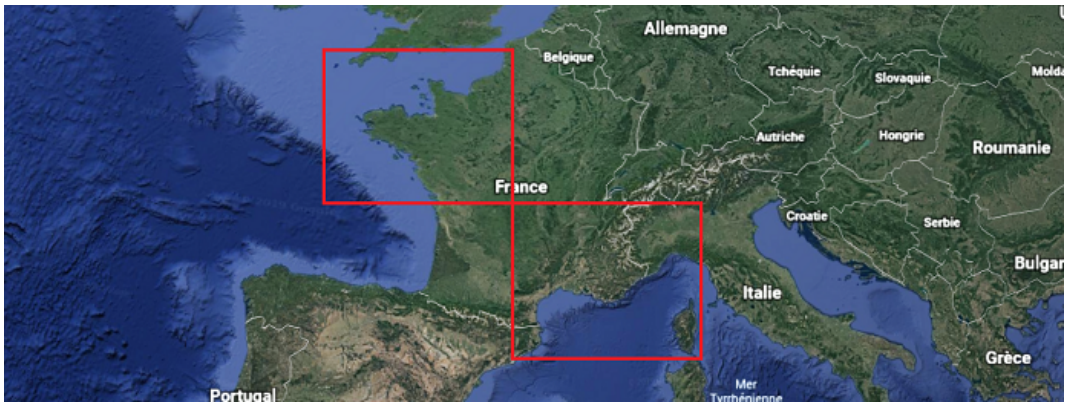


FIGURE 2.43 – Zones géographiques couvertes par la base MétéoNet. *Source : MétéoNet*

Il convient de définir les attributs présents dans la base de données, ce qui permettra d'avoir un aperçu plus précis des paramètres météorologiques donnés.

TABLE 2.1 – Informations sur les stations au sol et paramètres météorologiques

Attribut	**Description**	**Unité**
number_sta	Numéro de la station au sol	-
lat	Latitude WGS-84	Degrés
lon	Longitude WGS-84	Degrés
height_sta	Hauteur de la station	mètres
dd	Direction du vent	degrés
ff	Vitesse du vent	$m.s^{-1}$
precip	Précipitations pendant la période de rapport	$kg.m^2$
hu	Humidité	%
td	Température du point de rosée	Kelvin
t	Température	Kelvin
psl	Pression réduite au niveau de la mer	Pa

La variable d'intérêt dans le cadre de ce mémoire est « ff », représentant la vitesse du vent. En effet, les tempêtes étant les sinistres climatiques les plus fréquents du portefeuille, l'étude de cette variable est pertinente, afin d'examiner si son intégration peut affiner le modèle de fréquence, du moins pour les régions où des relevés sont disponibles. Des traitements préliminaires sont nécessaires pour rendre ces données exploitables.

2.3.2 Traitements

La base MétéoNet fournie est divisée en six sous-ensembles, chacun étant volumineux. Les données sur les dimensions sont recensées dans le tableau suivant :

TABLE 2.2 – Détails des sous-ensembles de la base MétéoNet

Nom de la base	Description	Nombre d'entrées	Taille (KB)
NW2016	Données Nord-Ouest 2016	21 921 197	1 738 008
NW2017	Données Nord-Ouest 2017	21 871 069	1 749 837
NW2018	Données Nord-Ouest 2018	22 034 571	1 758 821
SE2016	Données Sud-Est 2016	42 368 865	3 315 172
SE2017	Données Sud-Est 2017	41 838 616	3 295 627
SE2018	Données Sud-Est 2018	43 308 315	3 401 316

Les données ont été traitées à l'aide de SAS et Python.

Valeurs manquantes

Pour chaque sous-ensemble, seules les lignes contenant des données pour la variable « ff » ont été conservées, réduisant ainsi de près de moitié le nombre d'entrées de chaque sous-ensemble.

Doublons

461 stations différentes sont présentes dans la base. Bien que la base soit considérée comme « propre » d'après le site de MétéoNet, des doublons ont été identifiés. Ces doublons semblent avoir été causés par le codage des coordonnées. Un exemple de doublon est montré dans le tableau ci-dessous.

number_sta	lat	lon	date	height_sta	ff
1034004	45.769	5.688	01/01/2016 23:54	2	17
1034004	45.77	5.69	01/01/2016 23:54	2	17

FIGURE 2.44 – Exemple de doublons dans la base MétéoNet

Afin de garantir l'intégrité des données, ces doublons ont été fusionnés.

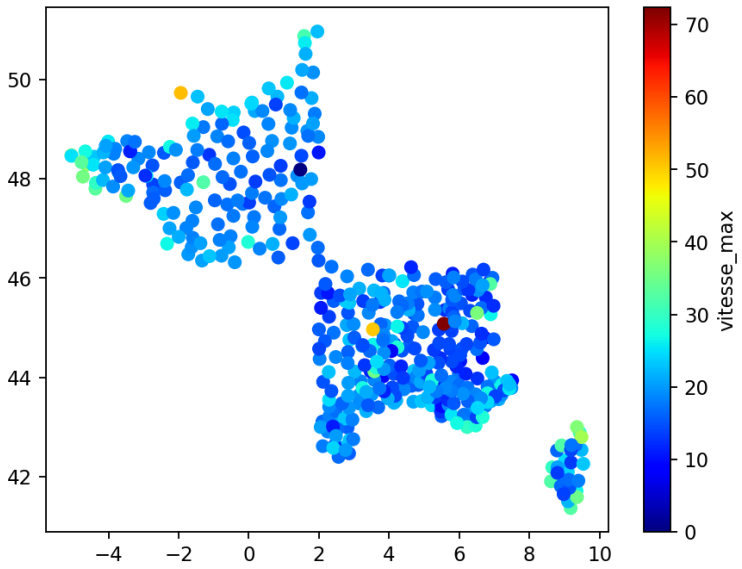


FIGURE 2.45 – Carte de France représentant la vitesse maximale du vent par station entre 2016 et 2018

2.3.3 Attribution du code INSEE

Les coordonnées des stations ont été récupérées et, à l'aide de GeoPoint sur Python, un code INSEE a été attribué à chaque station, en utilisant les mêmes fichiers shapefile que ceux de la partie 2.2.3.

2.3.4 Choix de l'étude

Une agrégation INSEE de la vitesse des vents a été choisie. À chaque INSEE a été attribué la vitesse maximale et la moyenne des 5 % de vitesses de vent les plus élevées. Ensuite, les valeurs ont été quantilisées en 10 zones. Bien que d'autres méthodes d'attribution de valeur auraient pu être envisagées, notamment en utilisant la théorie des valeurs extrêmes, cette approche a été jugée la plus pertinente pour une première analyse. Enfin, ces données ont été fusionnées avec les bases de fréquence et de coût moyen réduites, en ne conservant que les contrats situés dans les INSEE où des données de vitesse du vent étaient disponibles.

En conclusion, la base MétéoNet donne l'opportunité d'intégrer des données météorologiques dans la modélisation de la garantie climatique.

Le chapitre suivant sera centré sur la modélisation hors zonier, une étape essentielle pour assurer la fiabilité et la robustesse du modèle final, qui permettra d'évaluer l'impact des données ouvertes sur la modélisation de la garantie climatique des Propriétaires de Maison occupants et non occupants.

Chapitre 3

Modélisation de la prime pure climatique Hors Zonier

L'objectif de ce chapitre est de développer un modèle prédictif de la sinistralité future en s'appuyant sur les Modèles Linéaires Généralisés, plus couramment appelés ***GLM***. Nous explorerons également une méthode d'apprentissage statistique : les forêts aléatoires. La modélisation des GLM et des forêts aléatoires sera réalisée en suivant les méthodologies préconisées par l'équipe, anticipant l'intégration potentielle des bases de données ouvertes si les résultats de l'étude s'avèrent être satisfaisants.

Dans ce chapitre, on définit l'espace probabilisé (Ω, A, P)*. Soit* X *est une variable aléatoire définie dans cet univers. On note* $\mathbb{E}[X]$ *l'espérance de cette variable aléatoire et* $\mathrm{V}[X]$ *sa variance, que nous supposons bien définies.*

3.1 Méthodologie

3.1.1 Principe de la modélisation

Afin d'aborder la modélisation, les deux bases de modélisation fréquence et coût moyen seront séparées en trois :

- Les bases **HZ** (**H**ors **Z**onier) : Elles seront utilisées pour modéliser la sinistralité sans prendre en compte les informations géographiques, les antécédents de sinistres, et les données ouvertes. Ces bases représenteront 40 % du total des données ;
- Les bases **Z** (**Z**onier) : Elles serviront à modéliser la sinistralité en prenant en compte les variables géographiques, les variables antécédents et les données ouvertes. Les bases Z constitueront également 40 % des données.
- Les bases **V** (**V**alidation) : Elles seront utilisées exclusivement pour valider les modèles issus des bases Z. Aucun apprentissage ne sera réalisé sur les bases V. Les bases V représenteront 20 % des données. [1]

La méthodologie est résumée dans le schéma suivant :

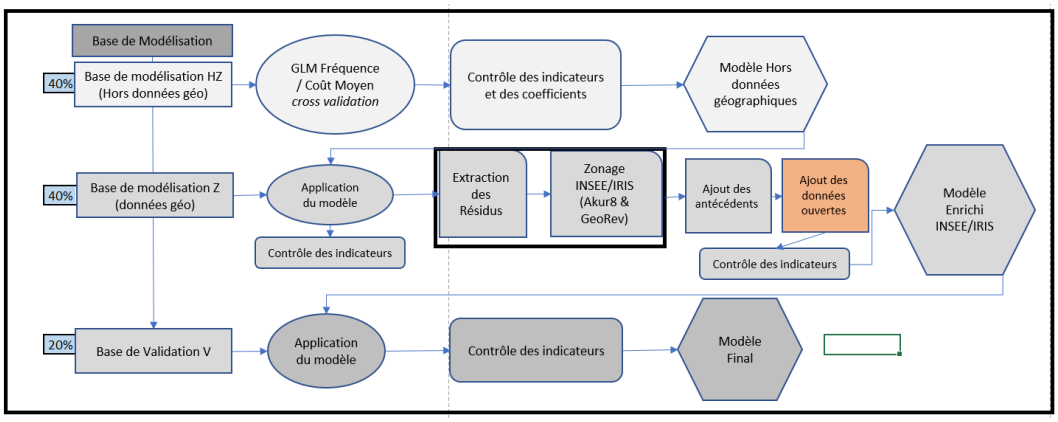


FIGURE 3.1 – Schéma de la méthodologie de modélisation

1. La répartition en % des données entre les trois bases suit les recommandations habituelles de l'équipe.

Dans ce chapitre, la modélisation HZ sera abordée. Plusieurs modèles seront entraînés et un seul sera choisi grâce aux indicateurs de performances, qui seront définis dans la prochaine section.
Les modèles seront évalués par validation croisée. Bien que cette méthode ait déjà été expliquée dans le cadre d'une recherche de paramètres (cf 2.2.2), voici l'explication de cette méthode dans le cadre de la modélisation : la base de données (en l'occurrence les bases HZ et Z) est encore une fois divisée en $k = 4$ *folds*, de manière aléatoire. Le choix de $k = 4$ est une pratique courante au sein de l'équipe. Ce choix offre encore une fois un équilibre entre la robustesse de la validation et des temps d'exécution raisonnables.

1. Sur k - 1 (= 3) *folds*, un modèle est formé, tandis que le *fold* restant sert de test pour évaluer les performances à l'aide des indicateurs ;
2. Cette procédure est répétée k (= 4) fois, garantissant ainsi k évaluations distinctes des indicateurs ;
3. La moyenne de ces indicateurs est calculée pour évaluer la performance globale du modèle ;
4. Enfin, les performances des modèles sur les k *folds* sont comparées. Le modèle final, choisi pour être déployé, est formé sur l'intégralité de la base d'entraînement.

Cette méthodologie permettra d'obtenir des modèles robustes et fiables pour prédire la sinistralité. Les métriques de performances qui seront utilisées pour évaluer la performance des modèles seront présentées dans le paragraphe suivant.

3.1.2 Indicateurs de performance

Afin d'évaluer la qualité et la précision des modèles, plusieurs indicateurs de performance seront pris en compte.

Indice de Gini

L'indice de Gini est couramment utilisé pour évaluer l'efficacité de la segmentation par un modèle. Il provient de la courbe de Lorenz, initialement conçue pour apprécier les inégalités de richesse d'une population. Cette courbe représente la distribution cumulée des valeurs d'une variable par rapport à la distribution cumulée des fréquences de cette même variable. Par exemple, l'axe horizontal peut représenter une partie croissante de la population, tandis que l'axe vertical indique combien cette partie possède ou reçoit. Cette représentation est pertinente pour différentes sortes de données statistiques. Voici à quoi ressemble cette courbe :

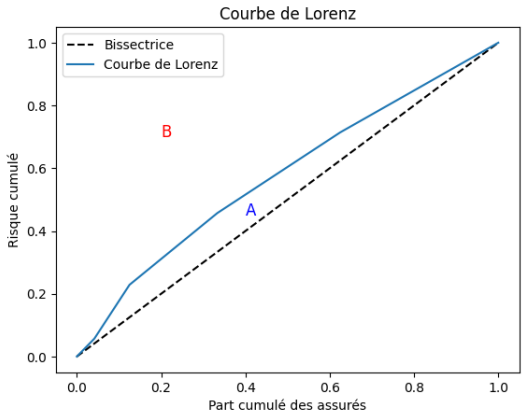


FIGURE 3.2 – Courbe de Lorenz

L'aire A est l'aire entre la courbe de Lorenz et la bissectrice, tandis que l'aire B est l'aire au-dessus de la courbe de Lorenz. Le gini est calculé comme suit :

$$\text{Gini} = \frac{\text{A}}{\text{A} + \text{B}} \tag{3.1}$$

Un indice de Gini de 0 signifie que la courbe de Lorenz coïncide avec la bissectrice, indiquant que les prédictions sont parfaitement alignées avec les valeurs observées. Dans ce contexte, le modèle prédit

le même niveau de risque pour chaque observation, indiquant une absence totale de segmentation. Autrement dit, toutes les prédictions sont identiques, quelles que soient les caractéristiques des assurés. Inversement, un indice de Gini de 1 démontre une segmentation maximale. Cela signifie que le modèle distingue efficacement les profils d'assurés en termes de sinistralité. Il est important de noter que cet indicateur ne renseigne pas sur l'exactitude absolue des prédictions.

Déviance

La déviance est une mesure statistique utilisée pour évaluer la qualité d'ajustement d'un modèle. Elle compare la vraisemblance du modèle calibré à celle d'un modèle saturé, qui est un modèle possédant autant de paramètres que d'observations et qui estime donc parfaitement les données. Mathématiquement, la déviance est définie comme :

$$\text{Déviance} = 2 \times (L_s - L_c)$$

où L_c est la log-vraisemblance du modèle étudié et L_s est la log-vraisemblance du modèle saturé.

L'intérêt de la déviance réside dans sa capacité à quantifier la qualité de la régression. Lorsque différents modèles sont comparés, celui avec la déviance la plus faible est généralement préféré car il offre un meilleur ajustement aux données. Cependant, il est essentiel de noter que la déviance ne prend pas en compte le risque de surapprentissage. De plus, sa sensibilité peut varier en fonction de l'ajout d'une variable au modèle, dépendant ainsi de la pertinence de cette variable.

RMSE

La **RMSE**[1] (***R**oot **M**ean **S**quared **E**rror*), ou erreur quadratique moyenne, mesure l'écart moyen entre les valeurs réelles (ou observées) et les valeurs prédites par le modèle :

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{3.2}$$

où :

— n représente le nombre d'observations dans le jeu de données ;

— y_i est la valeur réelle de l'observation i ;

— $\hat{y}_i$ est la valeur prédite par le modèle pour l'observation i.

Une RMSE plus faible suggère un modèle précis, car cela signifie que les prédictions du modèle sont alors proches des valeurs observées. Il est à noter que la RMSE est fortement influencée par les valeurs aberrantes (elle amplifie les erreurs). Cela est dû au fait qu'elle élève au carré les écarts entre les prédictions et les valeurs réelles.

MAE

La **MAE** (signifiant ***M**ean **A**bsolute **E**rror*) mesure l'erreur absolue moyenne entre les valeurs réelles et les valeurs prédites :

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{3.3}$$

Elle est exprimée dans l'unité de la variable cible, ce qui facilite son interprétation. Elle attribue un poids égal à toutes les erreurs, contrairement à la RMSE.

Résidus quantiles

Les résidus basés sur les quantiles sont des outils permettant d'évaluer l'adéquation entre les données et la loi théorique adoptée pour la modélisation. Ces résidus transforment la distribution du modèle en une distribution normale standard.

Ils sont exprimés par la formule :

$$r_{q,i} = \Phi^{-1}\{F(z_i; \tilde{\mu}_i, \tilde{\phi}\}$$

Où :

1. Rappel

— z est le vecteur des réponses suivant la loi $\mathcal{P}(\mu, \phi)$ avec $\mu_i = E[z_i]$ et ϕ est un paramètre commun à tous les z_i, supposés indépendants ;

— $F(z)$ est la fonction de distribution continue de la loi $\mathcal{P}(\mu, \phi)$ et $F(z_i; \mu_i, \phi)$ est uniformément distribuée sur l'intervalle [0,1] ;

— Φ représente la fonction de répartition de la distribution normale standard.

Si les résidus basés sur les quantiles suivent une distribution normale le long de l'axe des ordonnées, cela indique que la loi choisie pour la modélisation est bien adaptée aux données observées.

Erreur Totale

L'erreur totale évalue l'ajustement global du modèle en mesurant l'écart moyen entre les valeurs prédites et les valeurs réelles :

$$\text{Erreur totale} = \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)}{\sum_{i=1}^{n} y_i} \tag{3.4}$$

Elle offre une vue d'ensemble de la performance du modèle.

L'évaluation à l'aide de ces indicateurs est essentielle pour comparer la qualité des différents modèles. Dans la prochaine section, nous explorerons la modélisation avec les GLM en utilisant Akur8, outil privilégié par l'équipe pour les modélisations de GLM.

3.2 GLM

Avant d'aborder l'aspect théorique du GLM, le modèle linéaire doit être présenté.

3.2.1 Modèle linéaire gaussien

Un modèle linéaire vise à exprimer une variable aléatoire Y en fonction de plusieurs variables explicatives $X_1, X_2, ..., X_p$ (p $\in \mathbb{N}$). Cela peut se formuler comme suit :

$$Y_i = \beta_0 + \beta_1 x_{i,1} + \ldots + \beta_p x_{p,i} + W_i; \quad i = 1, \ldots, n; \quad \text{n est le nombre d'observations} \tag{3.5}$$

Où :

— Les variables $x_{i,j}$ sont les observations connues donc déterministes. C'est la réalisation de la variable explicative X_j pour l'observation i ($1 \leq i \leq n$, $1 \leq j \leq p$) ;

— Les W_i sont des variables aléatoires iid, suivant une loi normale centrée et de variance $\sigma^2 > 0$. Ils sont souvent appelés erreur de modélisation ;

— Le paramètre β_j ($1 \leq j \leq p$), associé de manière non aléatoire à la variable explicative X_j, demeure inconnu. Cela correspond à un coefficient fixe et constant qui mesure la relation entre la variable explicative X_j et la variable de réponse Y dans le modèle linéaire ([VERMET, 2020]).

Cependant, le modèle linéaire traditionnel présuppose une distribution normale des données, ce qui peut n'est pas toujours réaliste. Par exemple, cette hypothèse impliquerait que les sinistres pourraient avoir des coûts négatifs ou des fréquences négatives. C'est pour pallier à cette limitation que les modèles linéaires généralisés (*GLM*) sont utilisés. Ces modèles offrent une flexibilité en n'exigeant pas l'hypothèse de normalité des observations et en permettant une extension à la famille exponentielle. En d'autres termes, les *GLM* prennent en compte la distribution spécifique des données et permettent une meilleure adaptation aux caractéristiques des sinistres observés.

3.2.2 Modèles Linéaires Généralisés

Le modèle linéaire généralisé (GLM) est traditionnellement employé en tarification pour sa simplicité d'interprétation et sa solide performance. Bien que la distribution des Y_i (où les Y_i sont iid) ne soit pas forcément normale, elle doit tout de même appartenir à la famille exponentielle. Il est dit qu'une distribution appartient à la famille de dispersion exponentielle si sa fonction de densité prend la forme suivante [1] :

$$f_{(\theta;\phi)}(y) = \exp\left(\frac{y\theta - a(\theta)}{\phi} + c_\phi(y)\right) \tag{3.6}$$

où :

— θ désigne un paramètre réel, également appelé paramètre naturel ou canonique,

— ϕ symbolise le paramètre de dispersion (> 0),

— $a(\theta)$ est une fonction convexe de classe C^2,

— $c_\phi(y)$ est une fonction indépendante de θ.

Ainsi, si une variable aléatoire Y suit cette forme de distribution exponentielle, alors :

— $\mathbb{E}(Y) = a'(\theta)$

— $\mathbb{V}(Y) = a''(\theta)\phi$

Chaque loi de la famille exponentielle possède une fonction de lien spécifique, appelée fonction de lien canonique, qui fait le lien entre l'espérance mathématique et le paramètre naturel θ. Finalement, le GLM propose une relation plus générale entre les variables explicatives X et la variable réponse Y, qui peut être généralisée comme suit :

$$E(Y) = g^{-1}(X^T\beta); \quad (X^T \text{ est la transposée de X}),\ \beta \text{ est le vecteur des paramètres du modèle} \tag{3.7}$$

où g est une fonction de lien monotone et dérivable (bijective), appelée fonction de lien. Finalement, un modèle est qualifié de Modèle Linéaire Généralisé s'il remplit deux conditions principales :

1. La loi de $Y|X = x$ appartient à la famille exponentielle.
2. Il existe une fonction bijective g telle que $g(E(Y|X = x)) = X^T\beta$. C'est la fonction de lien.

Pour estimer les différents paramètres $\beta_0, \beta_1, ..., \beta_p$ du modèle, la méthode du maximum de vraisemblance (ou de la log-vraisemblance) peut être utilisée. Dans le cas des modèles de la famille exponentielle, la log-vraisemblance est définie comme :

$$l(\beta, \phi, y) = \sum_{i=1}^{n} \frac{Y_i\theta_i - a(\theta_i)}{\phi} + c_\phi(Y_i) \tag{3.8}$$

Pour maximiser la vraisemblance, la dérivée s'annule en résolvant l'équation :

$$\frac{\partial l(\beta, \phi, y)}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial}{\partial \beta_j}\left(\frac{Y_i\theta_i - a(\theta_i)}{\phi} + c_\phi(Y_i)\right) = 0; \quad 1 \leq j \leq p \tag{3.9}$$

L'intercept, noté β_0, symbolise la classe de référence. Concernant les autres coefficients de β, si $\beta_j > 0$ (ou inversement $\beta_j < 0$), cela signifie qu'une observation possédant la modalité X_j présentera un profil plus risqué (ou respectivement moins risqué) par rapport à la classe de référence.

1. cf [CHELBON, 2022]

3.2.3 Pénalisation

Les GLM utilisés dans l'équipe sont pénalisés. La pénalisation est un concept qui vise à sélectionner les variables explicatives ayant le plus d'influence significative sur la variable réponse Y. À la différence des modèles linéaires généralisés classiques qui tiennent compte de toutes les variables et de leurs modalités, la pénalisation permet d'établir des seuils de significativité pour écarter certaines variables. Cela est nécessaire, car un grand nombre de variables peut entraîner une variance élevée du modèle, conduisant à un surapprentissage des données [1]. En réalité, réduire le nombre de variables améliore l'interprétation des modèles, car en limitant le nombre de variables incluses, nous pouvons nous concentrer sur les facteurs les plus pertinents pour expliquer la variable cible.

La méthode de pénalisation qui sera utilisée dans ce mémoire est appelée ***LASSO*** (***L**east **A**bsolute **S**hrinkage and **S**election **O**perator*). Elle intègre un terme additionnel en norme $\mathbb{L}_1$. Cet élément contraint le processus d'optimisation à ajuster la valeur des coefficients. L'objectif est d'augment délibérément le biais du modèle pour en diminuer la variance, ce qui le rend plus robuste. L'estimateur associé à la régression LASSO est défini comme suit :

$$\beta_{\text{LASSO}} = \underset{\beta}{\operatorname{argmin}} \left(\sum_{i=1}^{n} (Y_i - \beta X_i)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right) \quad \text{avec } \lambda \in \mathbb{R}^+ \tag{3.10}$$

Le terme $\lambda \sum_{j=1}^{p} |\beta_j| = \lambda \|\beta\|_1$ représente la pénalité $\mathbb{L}_1$. Notons que si le coefficient de pénalisation λ est nul, l'estimateur se réduit à celui d'une régression non pénalisée. Ainsi, λ est un paramètre déterminant car il régule l'importance conférée à la pénalité. Pour être plus précis, le LASSO vise à simplifier le modèle en incitant les variables les moins pertinentes à adopter un coefficient nul (cas où λ est élevé). Cette méthode a l'avantage de produire un modèle épuré avec de nombreux coefficients fixés à zéro. Des études internes antérieures ont démontré l'adéquation de cette technique de pénalisation pour le portefeuille étudié. La pénalisation sera mise en œuvre à l'aide d'Akur8. La métrique d'évaluation utilisée pour choisir le meilleur λ par validation croisée est la déviance.

Les deux paragraphes suivants porteront sur la modélisation de la fréquence et du coût moyen.

3.2.4 Modèle fréquence

Dans cette section, la modélisation GLM hors zonier employée pour prédire le nombre annuel de sinistres associés à la garantie climatique des PM et PNOM, est expliquée. L'objectif est de projeter le nombre moyen de sinistres qu'un assuré pourrait rencontrer au cours d'une année. La variable clé, le nombre de sinistres climatiques attritionnels climatiques, est ajustée en fonction de l'exposition, telle que définie dans le paragraphe 1.3.1. La variable d'intérêt devient $\frac{\text{nombre de sinistres attritionnels}}{\text{exposition}}$.

Choix de la loi

Une loi de Poisson et une loi binomiale négative ont été ajustées aux données en utilisant la bibliothèque R `MASS`. Les distributions résultantes, dites théoriques, ont été comparées à celles des données, et des graphiques quantile-quantile ont été tracés pour visualiser la concordance entre les deux distributions. Suite à cette analyse, il est apparu que la loi binomiale négative était la plus adaptée pour représenter les données. Par conséquent, cette loi a été retenue pour modéliser la fréquence des sinistres attritionnels climatiques. La fonction de lien associée est la fonction logarithmique (*log*).

Détermination du modèle Hors Zonier fréquence

Lors de l'exécution d'un GLM sur Akur8, les résultats sont présentés sous forme de recherche sur grille (*grid search*). Sur cette grille, chaque point représente un modèle, l'axe des abscisses correspond au nombre de variables du modèle, tandis que l'axe des ordonnées illustre les performances du modèle (en fonction de l'indicateur choisi). Les résultats sont illustrés ci-dessous :

1. Rappel : il y a surapprentissage lorsque le modèle s'ajuste de façon correcte aux données sur lesquelles il a appris, mais performe mal sur de nouvelles données. La capacité de prédiction du modèle est ainsi réduite.

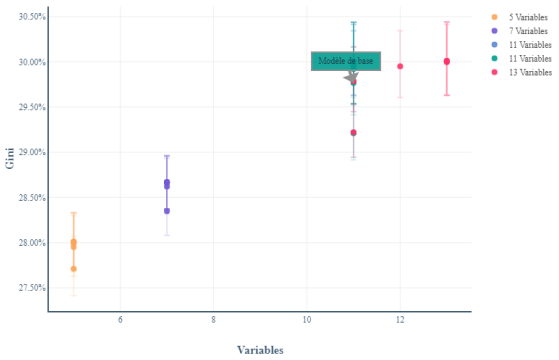


FIGURE 3.3 – Résultats du Grid Search Fréquence selon l'indice de Gini

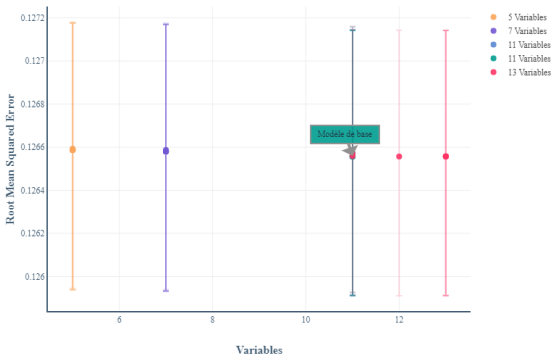


FIGURE 3.4 – Résultats du Grid Search Fréquence selon la RMSE

Le graphique de gauche illustre que les performances de segmentation, mesurées par l'indice de Gini, tendent à s'améliorer avec l'augmentation du nombre de variables. Il est à noter que pour un nombre de variables donné, plusieurs modèles peuvent être proposés. Le modèle le plus « bas » sur le graphique correspond à une version plus lissée du modèle, tandis que le modèle le plus « haut » est une version plus bruitée. Pour illustrer ce phénomène, considérons les deux modèles extrêmes à 7 variables :

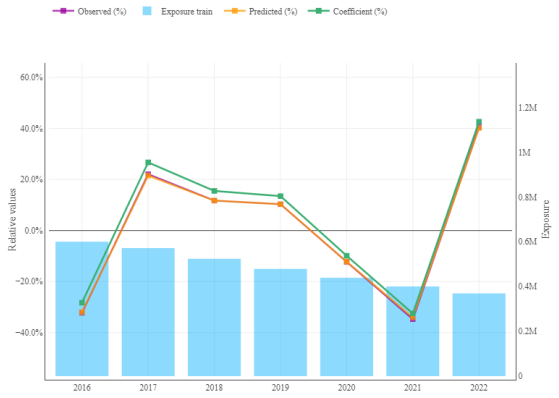


FIGURE 3.5 – Variable *year_expo* du modèle à 7 variables le plus bruité

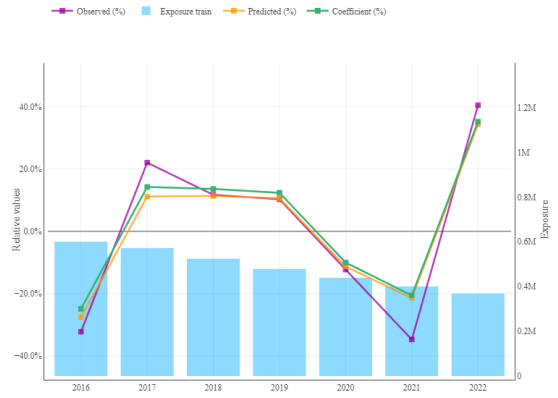


FIGURE 3.6 – Variable *year_expo* du modèle à 7 variables le plus lissé

Sur le graphique de gauche, le modèle le plus bruité montre que les prédictions correspondent étroitement aux valeurs observées, suggérant un potentiel risque de surapprentissage. À l'inverse, le modèle le plus lissé offre une généralisation plus large. Bien que le modèle bruité présente des performances d'apprentissage élevées, l'objectif de la sélection de modèle est de trouver un équilibre : un modèle qui offre des performances satisfaisantes, qui n'est ni trop bruité (pour permettre la généralisation) ni trop lissé (éviter la perte d'information), et qui est pertinent d'un point de vue professionnel [1]. Par exemple dans le cas de la variable *year_expo*, des prédictions lissées sont privilégiées afin de généraliser les tendances annuelles.
Le modèle qui répond le mieux à ces critères est désigné comme le « Modèle de base » sur la figure 3.5. Il s'agit d'un modèle composé de 11 variables. Avant d'analyser les coefficients de chacune des variables de la prédiction, une attention sera portée sur les indicateurs de performance de ce modèle (dont ceux de la *cross validation*), illustrées dans le tableau ci-dessous :

1. Par exemple, la variable CDRESID, correspondant au type d'occupation, est essentielle en tarification.

METRIC	FOLD	TRAIN FULL	TRAIN K-FOLD	TEST K-FOLD
GINI		29.93%	29.94%	29.77%
	1		29.9%	30%
	2		29.83%	30.22%
	3		30.06%	29.44%
	4		29.96%	29.41%
RMSE		0.1266	0.1266	0.1266
	1		0.1268	0.1258
	2		0.1264	0.1269
	3		0.1263	0.1272
	4		0.1266	0.1263
MAE		0.02522	0.02522	0.02522
	1		0.02523	0.02521
	2		0.02519	0.02526
	3		0.02518	0.02525
	4		0.02528	0.02517
OBSERVED TARGET AVERAGE		0.01281	0.01281	0.01281
	1		0.01282	0.0128
	2		0.01279	0.01286
	3		0.01279	0.01287
	4		0.01284	0.01272
PREDICTED TARGET AVERAGE		0.01281	0.01281	0.01281
	1		0.01282	0.01281
	2		0.01279	0.0128
	3		0.01279	0.01278
	4		0.01284	0.01285

TABLE 3.1 – Métriques de performance du modèle fréquence
Observed Target Average : Moyenne de la valeur cible observée.
Predicted Target Average : Moyenne de la valeur cible prédite.

Il est observé, à partir des résultats des *folds* par estimateur, que le modèle ne présente pas de surapprentissage. En effet, les performances *train/test* sont équivalentes. Une concordance est notée entre la valeur moyenne observée et celle prédite, indiquant une erreur totale relativement faible et donc une bonne adéquation du modèle aux données. De plus, l'indice de Gini est du même ordre que celui des standards habituels [1], ce qui suggère une capacité satisfaisante du modèle à segmenter les différents profils.
Enfin, le graphique des résidus quantiles montre que ces derniers suivent une distribution normale selon l'axe des abscisses, ce qui conforte le choix de la Binomiale Négative :

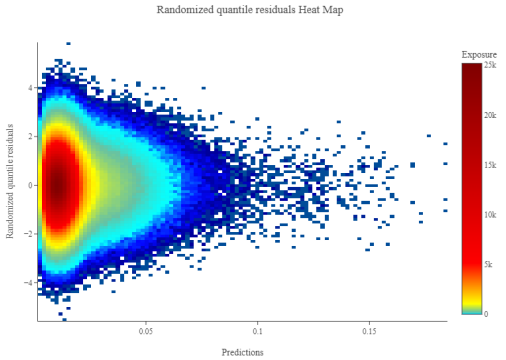


FIGURE 3.7 – Résidus Quantile modèle fréquence

Les performances étant satisfaisante, il convient à présent d'analyser le spread de coefficient de

1. Résulats de précédentes modélisations dans l'équipe.

chaque variable de ce modèle afin d'évaluer leur impact. Les coefficients de modèle sont donnés en pourcentage (normalisés). Deux méthodes distinctes de calcul du « spread » seront considérées ici :

— La première méthode, appelée « spread 100/0 », mesure la différence entre les coefficients maximum et minimum pour une variable donnée. En prenant en compte les coefficients normalisés, le spread est calculé par :

$$\text{Spread} = \frac{\max(\text{coefficient}) + 1}{\min(\text{coefficient}) + 1} - 1 \tag{3.11}$$

— La seconde méthode, le « spread 95/5 », est calculée de manière similaire au spread 100/0, mais avec une étape préliminaire. En effet, les 5% des expositions les plus risquées et les 5% des expositions les moins risquées sont éliminés. Le spread est ensuite déterminé à partir des 90% d'expositions restantes.

La figure suivante présente les variables ainsi que les spreads des coefficients à 100 % et à 95 % pour chacune d'entre elles :

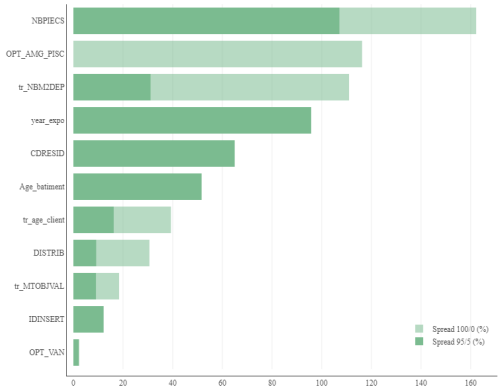


FIGURE 3.8 – Variables du modèle de base, avec les spreads de coefficients à 100 % et à 95 %

Il est observé une différence notable entre le spread à 95 % et le spread à 100 % de la variable relative à l'option piscine (OPT_AMG_PISC). Cela est due à plusieurs facteurs. En effet, cette variable a deux modalités et 96 % de l'exposition correspond à des contrats sans option, et donc de coefficient 0 %. Concernant la variable relative à l'option valeur à neuf (OPT_VAN), en raison de la faiblesse de ses deux mesures de spread faibles, soulevant la question de sa pertinence, elle a été retirée. Elle attribuait un coefficient de 2.27 % aux individus ayant opté pour cette option et 0 dans le cas contraire. La variable concernant l'âge du bâtiment a également été écartée, car en écartant la modalité faisant référence aux valeurs manquantes, le spread de coefficient est minime. Ce résultat a suscité une certaine surprise : intuitivement, on pourrait s'attendre à ce qu'un bâtiment plus ancien influence davantage la fréquence des sinistres. Une attention a été portée sur la variable de la dimension de dépendance (tr_NBM2DEP) étant donné la différence entre le spread à 100/0 et celui de 95/5. Le graphique suivant montre les valeurs observées, les valeurs prédites et les coefficients attribués à chaque tranche :

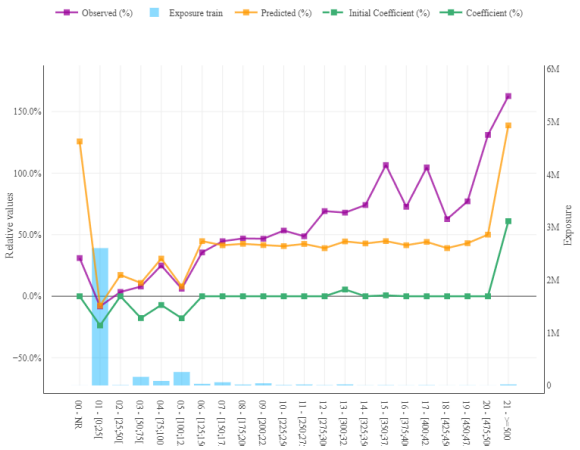


FIGURE 3.9 – Cas de la variable nombre de mètre carré des dépendances

Le coefficient de la 13^{me} tranche surpasse ceux des tranches supérieures, alors que d'un point de vue métier ce n'est pas cohérent. Cela pourrait être dû à la faible exposition de ces tranches, influençant ainsi les performances de prédiction et l'ajustement des coefficients. De plus, jusqu'à la 6^{me} tranche, les prédictions sont trouvées trop proches des valeurs réellement observées, suggérant un potentiel risque de surajustement. À partir de la 6^{me} tranche, une sous-prédiction notable a été constatée. Ainsi, des modifications ont été jugées nécessaires pour améliorer la précision du modèle : une linéarisation des coefficients pour les segments jugés trop bruités et une modification des coefficients pour atténuer la sous-prédiction. Le résultat final donne :

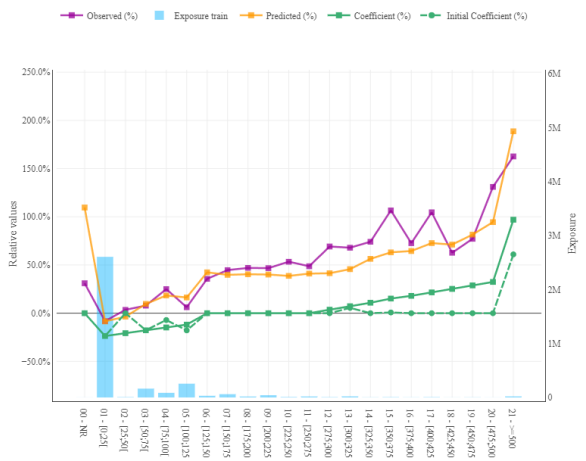


FIGURE 3.10 – Modification de la variable nombre de mètre carré des dépendances

La variable relative au nombre de pièces a également été traitée de façon similaire.

Validation du modèle

Enfin, le modèle de base modifié a été *refitté* afin d'évaluer ses performances :

TABLE 3.2 – Résultats des métriques de performance, modèle final Fréquence HZ

METRIC	TRAIN K-FOLD	TEST K-FOLD
GINI	29.23%	29.23%
RMSE	0.1266	0.1266
MAE	0.02522	0.02522
OBSERVED TARGET AVERAGE	0.01281	0.01281
PREDICTED TARGET AVERAGE	0.01281	0.01281

Observed Target Average : Moyenne de la valeur cible observée.
Predicted Target Average : Moyenne de la valeur cible prédite.

Les performances de ce modèle sont satisfaisantes : elles sont quasiment identiques à celles du modèle de base, et il n'y a pas de surapprentissage. La stabilité de l'indice de Gini est un autre critère essentiel pour valider ce modèle. Le graphique ci-après démontre cette stabilité, où l'on observe que les différentes courbes de Lorenz coïncident presque parfaitement pour chaque sous-échantillon de la base de validation.

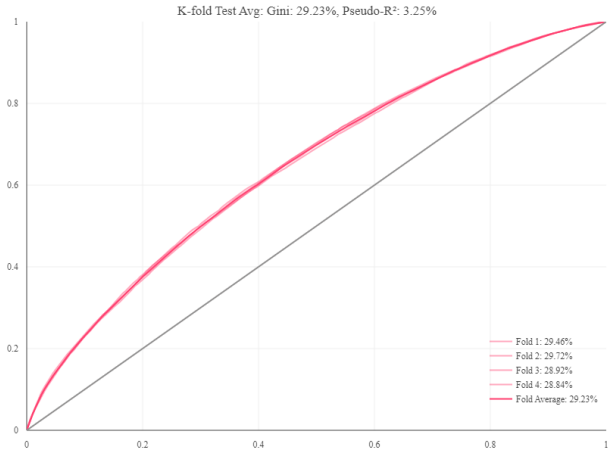


FIGURE 3.11 – Courbe de Lorenz du modèle final

Le gini moyen est inférieur à celui du modèle de base de 0.7 %, mais reste satisfaisant.
Il est désormais essentiel de comparer les fréquences observées avec celles prédites sur l'échantillon de validation. La validation par quantiles est utilisée pour évaluer la capacité prédictive du modèle. Cette méthode consiste à créer des quantiles basés sur la fréquence prédite par le modèle, puis à représenter graphiquement les fréquences observées et prédites. Si le modèle est bien calibré, la fréquence prédite devrait se rapprocher de la fréquence observée. Cette tendance est illustrée par la *lift-curve* :

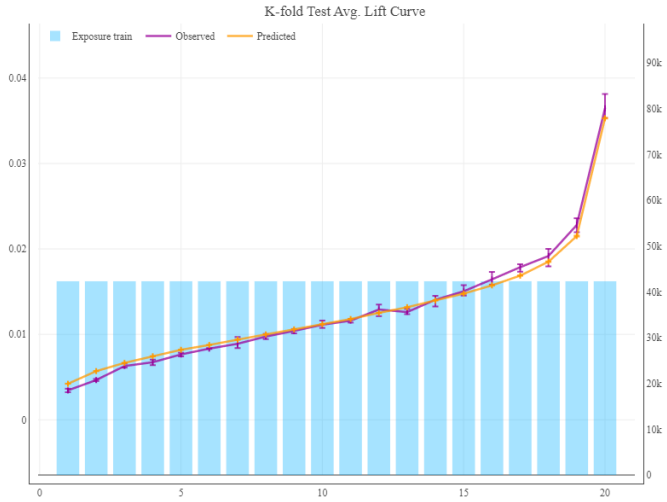


FIGURE 3.12 – *Lift-Curve* du modèle final Fréquence, 20 quantiles

La validation visuelle montre une concordance entre la fréquence prédite et la fréquence observée. Cette observation est confirmée numériquement, avec des valeurs de 1.28 % pour les deux.

En conclusion, ce modèle sera retenu pour la fréquence HZ. Tous les tests ont validé la fiabilité des résultats produits par ce modèle, montrant une constance dans les indicateurs et une capacité de prédiction robuste, tout en éliminant le risque de surapprentissage. L'étape suivante portera sur la modélisation de la seconde composante de la prime pure : le coût moyen.

3.2.5 Modélisation du Coût Moyen

Le modèle du coût moyen est conçu pour prédire la charge moyenne associée à un sinistre climatique sur une période d'un an. Par conséquent, la variable cible est la charge vieillie attritionnelle. Bien que la sélection du modèle final pour le coût moyen repose sur les mêmes critères que celle de la fréquence, nous aborderons ce sujet de manière plus synthétique. Deux lois sont possibles pour ce modèle : la loi gamma et la loi inverse gaussienne, toutes deux associées à une fonction de lien *log*. La première étape consistera à choisir l'une des deux.

Choix de la loi

En utilisant la démarche adoptée pour la fréquence dans le processus de sélection entre les deux lois, une loi inverse gaussienne et une loi gamma ont été ajustées à l'aide de la bibliothèque `scipy.stats` de Python. Afin de prendre une décision éclairée sur la loi la plus adaptée, nous avons confronté les distributions théoriques aux distributions observées à l'aide de graphiques. Les graphiques quantile-quantile ont également été employés pour évaluer la concordance entre les distributions théoriques et les données réelles. Les illustrations suivantes mettent en évidence ces comparaisons.

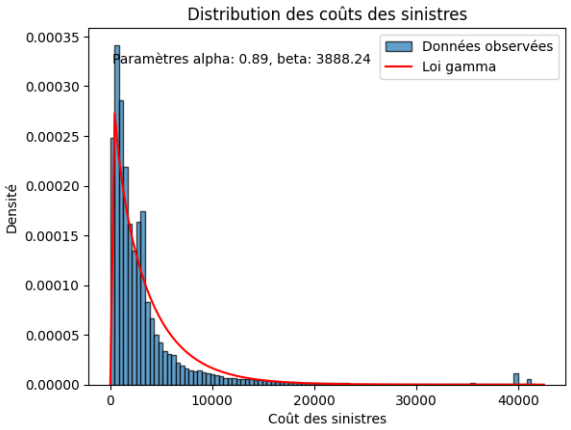


FIGURE 3.13 – Distribution des coût et distribution de la loi gamma ajustée

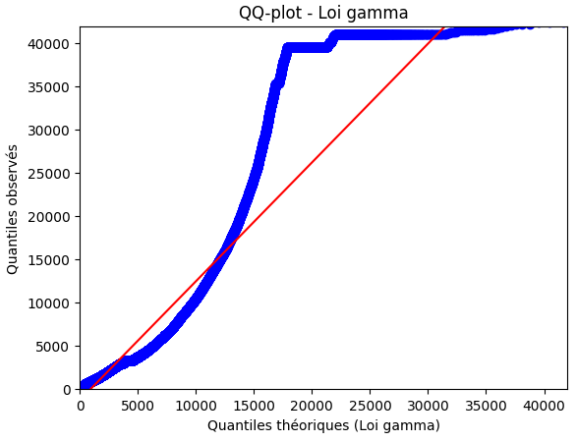


FIGURE 3.14 – Graphique Quantile-Quantile gamma

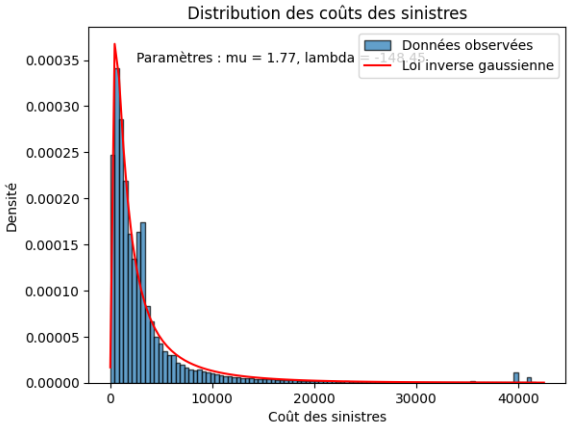


FIGURE 3.15 – Distribution des coût et distribution de la loi inverse gaussienne ajustée

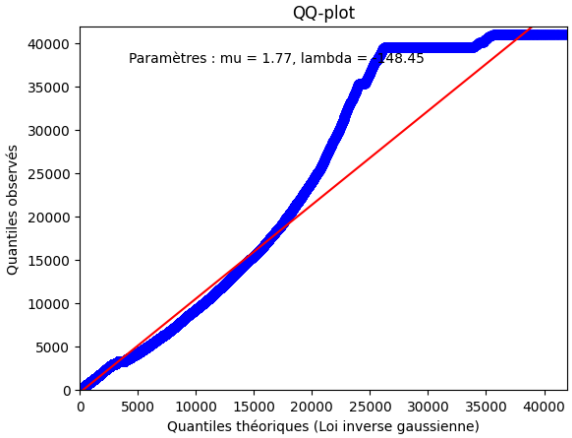


FIGURE 3.16 – Graphique Quantile-Quantile inervse gaussienne

La loi inverse gaussienne offre le meilleur ajustement à nos données, se révélant ainsi plus pertinente. C'est donc cette loi qui a été choisie pour la modélisation.

Choix et validation du modèle HZ Coût Moyen

Après avoir adopté une méthodologie similaire à celle du modèle de fréquence, nous avons abouti à un modèle de base ajusté pour le modèle HZ. Ce modèle intègre sept variables distinctes. Ci-dessous, nous détaillons chacune de ces variables ainsi que leurs spreads associés :

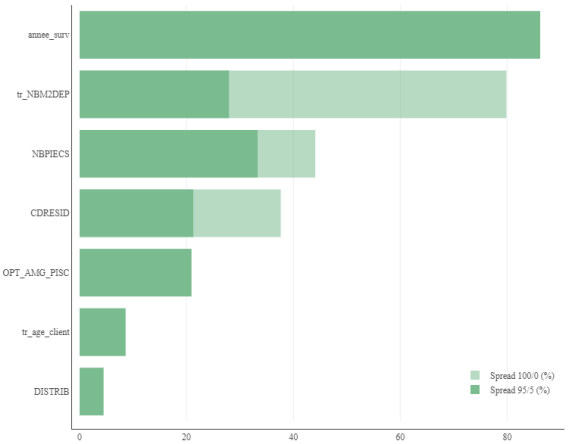


FIGURE 3.17 – Variables du modèle Coût Moyen sélectionné et leurs spreads associés

Il est à noter que ces variables sont également présentes dans le modèle de fréquence. De plus, ce modèle ne présente pas de surapprentissage. En effet, les performances entre les données d'entraînement et de test sont équivalentes. Elles sont résumées dans le tableau suivant :

TABLE 3.3 – Résultats des métriques de performance, modèle final Coût Moyen HZ

METRIC	TRAIN K-FOLD	TEST K-FOLD
GINI	17.41%	17.41%
RMSE	5282	5281
MAE	2883	2883
OBSERVED TARGET AVERAGE	3510	3510
PREDICTED TARGET AVERAGE	3465	3465

Observed Target Average : Moyenne de la valeur cible observée.
Predicted Target Average : Moyenne de la valeur cible prédite.

Il convient également d'analyser la courbe de Lorenz et la *lift-curve* :

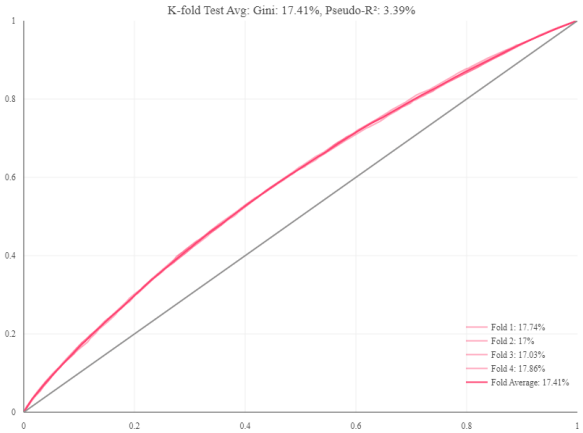


FIGURE 3.18 – Courbe de Lorenz du modèle final coût moyen

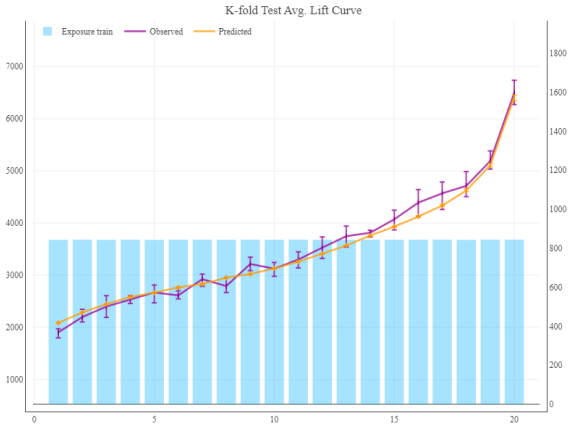


FIGURE 3.19 – *Lift-Curve* du modèle final coût moyen

L'indicateur de Gini demeure stable pour chaque sous-échantillon de la base de validation. De plus, le coût moyen prédit est sensiblement équivalent au coût moyen observé, indépendamment du quantile étudié. En fait, le coût moyen prédit, évalué à 3465, reste relativement proche du coût moyen observé qui est de 3510. En conclusion, ce modèle sera retenu pour le coût moyen HZ.

La partie suivante sera consacrée à la modélisation HZ à l'aide d'une méthode d'apprentissage statistique : les forêts aléatoires. L'objectif sera de comparer ses performances à celles du GLM.

3.3 Forêt aléatoire

Dans cette étude, une méthode d'apprentissage statistique sera explorée : les forêts aléatoires (ou *Random Forests* en anglais). L'apprentissage statistique suppose que les données à prédire sont générées de manière indépendante et identiquement distribuées. L'objectif est de construire un algorithme capable d'apprendre à prédire la variable dépendante Y en fonction des variables explicatives X : $Y = f(X) + \epsilon$, où ϵ est un bruit centré. Dans le cadre de la régression, $f(X) = E[Y|X = x]$ ([Paglia et Phélippé-Guinvarc'h, 2011]). La base de modélisation utilisée pour les forêts aléatoires sera légèrement différente de celle utilisée pour les *GLM* : la dimension des dépendances et l'âge des clients ne seront pas transformés en tranches.
Le choix d'explorer une autre technique de modélisation dans le cadre de ce mémoire est motivé par la volonté de déterminer si la combinaison de données ouvertes avec une technique d'apprentissage différente peut permettre d'améliorer au maximum les performances du modèle. Les forêts aléatoires offrent une flexibilité et une robustesse qui pourraient s'avérer bénéfiques dans cette optique.

3.3.1 Aspect théorique

Pour comprendre le fonctionnement des forêts aléatoires, il est essentiel de se familiariser d'abord avec les arbres de décision *CART* (*Classification And Regression Trees*).

Les arbres de décision, qui englobent à la fois les arbres de régression et de classification, sont des outils non paramétriques destinés à la segmentation. Leur objectif est d'identifier des critères qui permettent de diviser les observations en deux classes distinctes. Ainsi, l'algorithme en question commence par choisir la variable qui permet de créer deux sous-groupes les plus homogènes possible et les plus distincts. Ensuite, il détermine la meilleure manière de segmenter en fonction de cette variable. De ce point de départ, nommé nœud racine, émergent deux branches, correspondant à deux résultats possibles. Chaque branche peut soit mener à un autre nœud, introduisant un test supplémentaire, soit aboutir à un nœud terminal, ou feuille. À chaque étape de sa construction, l'arbre de décision évolue, choisissant une variable explicative et formant une nouvelle branche. Le choix de la variable et du point de division à chaque nœud est guidé par la volonté de minimiser ce qu'on appelle « l'impureté », évaluée par la variance pour une régression. L'avantage des arbres de décision est leur lisibilité et leur interprétabilité : cela se traduit par un ensemble de règles claires et simples. Cependant, ils présentent une certaine instabilité, ce qui peut entraîner des variations significatives dans les prédictions obtenues. C'est pourquoi les forêts aléatoires seront utilisées. En effet, elles sont une extension des arbres de décision. Au lieu de construire un seul arbre, une forêt aléatoire construit de nombreux arbres de décision et les combine pour obtenir une prédiction plus précise et stable.

Random Forest (RfR)

Dans le cadre des *Random Forests*, le processus de construction d'arbres est répété de nombreuses fois. La solution finale est alors déduite de la combinaison des réponses de chaque arbres. En d'autres termes, une forêt aléatoire est essentiellement une collection d'arbres de décision, où chaque arbre est formé à partir d'un échantillon aléatoire des données. Cette approche vise à éviter le piège du surapprentissage, principal inconvénient des arbres de décision. En construisant de nombreux arbres, la variance de l'estimation est grandement réduite, conduisant à des prédictions plus fiables et donc des estimations plus robustes. Pour construire une forêt aléatoire efficace, quelques paramètres (dits hyperparamètres) doivent être déterminés :

— Le ***max features***. Il correspond au nombre de variables sélectionnées de manière aléatoire pour la construction de chaque arbre. Ce paramètre est essentiel, car le fait de ne prendre qu'un sous-ensemble des variables explicatives pour chaque arbre permet de les décoreller les uns des autres. En effet, si une variable explicative est fortement liée à la variable cible Y, elle sera probablement utilisée dans la majorité des arbres. Par conséquent, la plupart de ces arbres auront une structure similaire, leurs prédictions seront étroitement liées. Ainsi, un nombre élevé pour ce paramètre augmente le risque de sur-apprentissage ;

— Le ***n estimators***. Il représente le nombre d'arbres dans la forêt ou l'ensemble. Il spécifie le nombre d'arbres à construire avant de faire une moyenne des prédictions. En général, un nombre plus élevé d'arbres augmente la performance et rend les prédictions plus stables, mais cela peut aussi ralentir considérablement le calcul, surtout si le nombre d'arbres choisi est très élevé.

— Le ***max depth***. Il définit la profondeur maximale de chaque arbre. Une profondeur trop grande peut conduire à un modèle trop complexe, qui risque de sur-apprendre sur les données d'entraînement. À l'inverse, une profondeur trop faible peut empêcher le modèle de capturer les nuances des données, conduisant à un sous-apprentissage ;

— Le ***min samples split***. Ce paramètre détermine le nombre minimum d'échantillons requis pour diviser un nœud interne. Une valeur élevée peut empêcher la division d'un nœud, même si celle-ci est justifiée, conduisant à un sous-apprentissage. À l'inverse, une valeur trop faible peut conduire à des divisions inutiles, et donc à un sur-apprentissage.

Notons que si le paramètre *max depth* n'est pas spécifié, alors les nœuds sont développés jusqu'à ce que toutes les feuilles soient pures ou jusqu'à ce que toutes les feuilles contiennent moins d'échantillons que la valeur spécifiée par *min samples split*[1].

Une recherche exhaustive (ou *grid search*) avec *cross validation* sera mise en œuvre sur toutes les combinaisons possibles des paramètres afin d'identifier la valeur optimale pour chacun d'entre eux. Cette approche est mise en œuvre grâce à la librairie `sklearn.modelselection` de Python. L'objectif est de trouver la combinaison qui minimise l'erreur de prédiction, basée sur l'indicateur RMSE. Par ailleurs, nous ferons appel à l'optimisation bayésienne, une technique efficace pour réduire les temps de calcul. En utilisant l'optimisation bayésienne, chaque ensemble de solutions possibles est exploré en prenant en compte les résultats passés. Dans cette approche, une fonction objectif est cherchée à être optimisée de manière progressive, sur la base d'un score préalablement défini. Dans notre contexte, cette fonction objectif s'appuie sur les résultats de la validation croisée réalisée à l'aide d'un régresseur de type forêt aléatoire. L'optimisation bayésienne est accessible par la librairie `bayesopt` de Python.

Les *Random Forest* nécessitent des entrées numériques. Par conséquent, les variables catégorielles ont été transformées en utilisant la méthode de *Ordinal Encoding* grâce à la librairie `sklearn.preprocessing` de Python. Cette technique attribue une valeur numérique unique à chaque catégorie d'une variable, en fonction de l'ordre des catégories. Par exemple, pour la variable « Age bâtiment », ayant les catégories « Moins de 5 ans », « Entre 5 et 10 ans », « Plus de 10 ans », l'encodage ordinal pourrait attribuer les valeurs 0, 1 et 2 respectivement. L'avantage de cette méthode est qu'elle conserve l'information sous une forme numérique sans augmenter la dimensionnalité des données, contrairement à d'autres méthodes comme le *one-hot encoding*.

3.3.2 Modèle Fréquence

Mise en oeuvre

La configuration de paramètres suivante a été explorée pour le Random Forest en utilisant une fonction de perte quadratique[2]. Voici les détails de cette configuration :

— Le nombre minimal d'observations nécessaires pour constituer une feuille a été défini dans une plage allant de 1 % à 5 % du nombre total d'observations (***min samples split***) ;

— Compte tenu du fait que la base de modélisation comprend 12 variables (voir chapitre 1), une décision a été prise de manière arbitraire pour en prendre au plus 10 lors de la création de chaque arbre (***max features***) ;

— Le nombre de forêts a été défini dans un intervalle allant de 100 à 500 (***n estimators***) ;

— La profondeur maximale des arbres a été définie dans un intervalle allant de 1 à 10 (***max depth***).

1. D'après [scikit-learn Developers, 2023]
2. Cela correspond au RMSE élevé au carré

L'optimisation de ces hyperparamètres a été réalisée à l'aide de l'optimisation bayésienne. Cette méthode a été choisie car le Grid Search ne s'était pas achevé, même après 6 jours de calculs. Afin de garder des temps de calcul raisonnables, le paramètre d'itération de l'optimisation bayésienne (n_iter) a été fixé à 100. Cela correspond au nombre de fois où l'algorithme va évaluer une configuration d'hyperparamètres. En effet, chaque itération consiste en la suggestion d'une nouvelle configuration d'hyperparamètres, suivie de l'évaluation du $RMSE^2$ de cette configuration.

Résultats

Les paramètres optimaux déterminés par l'algorithme sont :

Paramètre	**Valeur**
n_estimators	458
max_features	10
max_depth	8
min_samples_split	1% des lignes de la base fréquence

TABLE 3.4 – Paramètres optimaux déterminés par l'algorithme

Les métriques correspondantes sont :

Métrique	**Valeur**
MAE (*test*)	0.028167
MAE (*train*)	0.028160
RMSE (*test*)	0.214454
RMSE (*train*)	0.216907
Gini (*train*)	32.37 %
Gini (*test*)	30.24 %

TABLE 3.5 – Résultats de l'évaluation du modèle

Si on regarde le MAE et le RMSE, les performances du train et du test étant équivalentes. Néanmoins, on note une différence de 2 points entre le gini de *train* et *test*. À présent, il convient de regarder l'importance des variables. Dans le cadre du Random Forest de la librairie `scikit`, le score d'importance est calculé par le processus suivant :

1. **Diminution de l'Impureté** : Lorsqu'une variable est utilisée pour diviser un nœud dans un arbre, l'impureté du nœud parent est comparée à celle des nœuds enfants. La réduction de l'impureté est cumulée pour chaque variable à travers tous les arbres.

2. **Moyennisation sur l'Ensemble des Arbres** : L'importance de chaque variable est calculée en faisant la moyenne de ses diminutions d'impureté sur l'ensemble des arbres de la forêt.

3. **Normalisation** : Les scores d'importance obtenus pour chaque variable sont normalisés afin que leur somme totale soit égale à 1.

Ainsi, l'importance des variables est donnée par le tableau suivant :

Variable	Importance
OPT_AMG_PISC	26.31%
NBPIECS	21.72%
year_expo	20.41%
NBM2DEP	9.17%
tr_MTOBJVAL	6.49%
age_client	4.86%
CDRESID	5.09%
IDINSERT	2.38%
Age_batiment	2.19%
DISTRIB	0.85%
DEDUCTIBLE_TYPE	0.49%

TABLE 3.6 – Importance des variables classées par ordre décroissant, Fréquence

L'analyse de l'importance des variables dans le modèle Random Forest offre un aperçu précieux sur les facteurs qui influencent le plus la variable cible. Dans le cas de ce modèle :

1. **Variables Dominantes** : Les variables `OPT_AMG_PISC`, `NBPIECS`, et `year_expo` sont les plus influentes, avec des importances respectives de 26.04%, 21.45% et 20.14%. Cela suggère que ces variables jouent un rôle essentiel dans la prédiction de la variable cible. Par exemple, cela pourrait signifier que la présence d'une piscine (`OPT_AMG_PISC`), le nombre de pièces (`NBPIECS`) et l'année d'exposition (`year_expo`) sont des facteurs déterminants. Il est intéressant de noter que ces mêmes variables se distinguent également par leur importance dans le modèle GLM fréquence ;
2. **Variables Modérément Importantes** : Les variables comme `NBM2DEP`, `tr_MTOBJVAL`, `age_client`, `CDRESID`, et `IDINSERT` ont une importance modérée, suggérant qu'elles contribuent également à la prédiction, mais dans une moindre mesure par rapport aux variables dominantes ;
3. **Variables Moins Importantes** : Les variables `Age_batiment`, `DISTRIB`, et `DEDUCTIBLE_TYPE` ont une importance relativement faible. Cela ne signifie pas nécessairement qu'elles sont inutiles, mais dans le contexte du modèle Random Forest, leur contribution à la prédiction est moindre. Dans le cadre du GLM fréquence, Age_batiment a été supprimée.

Ces résultats seront comparés à ceux du GLM Fréquence par la suite.

3.3.3 Modèle Coût Moyen

Il convient de rappeler que la quantité de données de la base coût moyen est réduite par rapport à la base fréquence. Cela offre l'avantage de pouvoir effectuer et un *Grid Search* et une optimisation bayésienne tout en maintenant un temps d'exécution acceptable.La configuration de paramètres suivante a été explorée, avec un total de 400 arbres [1] et une fonction de perte de type quadratique :

— Le nombre minimal d'observations nécessaires pour former une feuille a été défini dans une plage allant de 1 % à 10 % [2] du nombre total d'observations (***min samples split***) : [1 % de la base, 2.5 % de la base, 5 % de la base, 10 % de la base] ;

— Étant donné que la base de modélisation contient 10 variables (voir chapitre 1), il a été décidé arbitrairement de considérer au maximum 6 lors de la création de chaque arbre (***max features***) : [3, 4, 5, 6] ;

— Le nombre d'arbres a été fixé à 400 (***n estimators***) ;

— La profondeur maximale des arbres a été définie comme [5, 10] (***max depth***).

Pour l'optimisation bayésienne, les mêmes plages de test ont été choisies, mais sous forme d'intervalles continus. Le nombre d'itérations a été fixé à 300.

Les résultats obtenus sont les suivants :

1. Ce nombre est considéré comme suffisant au vu des résultats du modèle de fréquence.
2. Étant donné la taille réduite de la base, il est possible d'aller jusqu'à 10 % tout en conservant des temps de calcul raisonnables.

Paramètre	**Valeurs**	
	Grid Search	**Optimisation Bayésienne**
n_estimators	400	400
max_features	6	9
max_depth	10	8
min_samples_split	1% des lignes de la base coût moyen	1% des lignes de la base coût moyen

TABLE 3.7 – Paramètres optimaux déterminés par l'algorithme, modèle coût moyen

Les métriques correspondantes sont :

Métrique	**Valeurs**	
	Grid Search	**Optimisation Bayésienne**
MAE (*test*)	2908.90	2909.15
MAE (*train*)	2900.29	2900.04
RMSE (*test*)	5294.32	5295.01
RMSE (*train*)	5277.97	5277.61
Gini (*train*)	19.30 %	19.25%
Gini (*test*)	17.94 %	17.86%

TABLE 3.8 – Résultats de l'évaluation du modèle

En analysant les métriques, on observe des différences significatives entre les valeurs d'entraînement et de test. Plusieurs raisons peuvent expliquer ces écarts :

1. **Taille de la base de données** : La base de données pourrait ne pas être suffisamment volumineuse pour permettre au modèle d'apprendre des caractéristiques générales, ce qui pourrait entraîner un léger surapprentissage sur les données d'entraînement et, par conséquent, des performances réduites sur les données de test ;
2. **Sélection des hyperparamètres** : Les plages d'hyperparamètres choisies pour la recherche pourraient ne pas être optimales. Cette hypothèse a été vérifiée en testant d'autres combinaisons d'hyperparamètres, mais les écarts étaient encore plus grands.

Variable	**Importance**	
	Grid Search	**Optimisation Bayésienne**
OPT_AMG_PISC	10.539%	10.273%
NBPIECS	15.859%	16.363%
NBM2DEP	16.789%	15.863%
age_client	3.669%	3.843%
CDRESID	1.849%	1.923%
Âge_batiment	0.679%	0.813%
DISTRIB	0.469%	0.703%
DEDUCTIBLE_TYPE	0.419%	0.643%
annee_surv	49.729%	49.773%

TABLE 3.9 – Importance des variables classées par ordre décroissant pour les méthodes Grid Search et Optimisation Bayésienne

L'analyse des importances des variables montre des tendances similaires entre les méthodes Grid Search et Optimisation Bayésienne, bien que de légères variations soient observées. La variable `annee_surv` est clairement dominante en termes d'importance, suivie de `NBM2DEP` et `NBPIECS`. D'autres variables, comme `age_client` et `CDRESID`, ont une importance moindre, mais contribuent néanmoins à la performance du modèle. Les variables `DISTRIB` et `DEDUCTIBLE_TYPE` ont la plus faible importance, suggérant un impact limité sur les prédictions.

Il est également intéressant de noter que les importances des variables sont relativement stables entre les deux méthodes. Cela renforce la confiance dans la robustesse des importances dérivées et suggère que les deux méthodes fournissent des informations cohérentes sur la pertinence des variables. Cependant, pour les analyses ultérieures, nous avons choisi de conserver les estimations obtenues à partir de la méthode *Grid Search*, car elle offre des métriques légèrement plus performantes que celles de l'optimisation bayésienne.

3.4 Synthèse de la modélisation hors zonier

Dans cette section, les performances des modèles GLM et RfR en termes de fréquence et de coût moyen seront comparées.

Modèle	**Fréquence**		**Coût moyen**	
	RMSE	**Gini**	**RMSE**	**Gini**
GLM pénalisé	0.1266	29.23 %	5281	17.41 %
RfR	0.2145	31.24 %	5294	17.94 %

TABLE 3.10 – Comparaison des métriques d'évaluation entre les GLM et RfR sur la base de validation

Il est également important de noter que, bien que le RfR ait montré une meilleure performance en termes de coefficient de Gini, le RMSE était plus élevé que celui du GLM. Cela suggère que, bien que le RfR puisse mieux segmenter les observations, il peut aussi faire des erreurs plus importantes en termes de magnitude. Cette observation est essentielle, car dans le contexte de tarification en assurance, il est essentiel de minimiser l'erreur absolue pour éviter des sous-estimations ou des surévaluations significatives des primes.

Les GLM, en revanche, présentent des performances prédictives solides. De plus, ils sont plus interprétables que les RfR, ce qui est un avantage considérable dans le secteur de l'assurance où la compréhension et l'explication des modèles sont essentielles pour la prise de décision et la communication avec les parties prenantes.

En conclusion, bien que les RfR aient leurs avantages et puissent être plus performants dans certaines situations, les GLM semblent être le choix optimal pour cette étude en particulier. Leur simplicité, leur interprétabilité et leur efficacité en font un outil précieux pour la modélisation de la prime pure climatique des Propriétaires de Maison en MRH. C'est donc à l'aide des GLM que nous évaluerons dans le chapitre suivant l'impact de l'intégration des données ouvertes sur la modélisation de la prime pure climatique pour les propriétaires de maisons en MRH.

Chapitre 4

Ajout du signal géographique et des données ouvertes

Cette section est centrée sur l'enrichissement des modèles GLM hors zonier par l'intégration d'informations géographiques (ou zonier), d'antécédents de sinistres climatiques et des données ouvertes. La méthodologie sera détaillée à chaque étape.
L'ajout d'une dimension géographique, sous forme de zonier [1], repose sur l'hypothèse que le risque associé à un bien assuré ne dépend pas uniquement des caractéristiques de ce bien ou de son propriétaire. En effet, la localisation géographique du bien peut influencer le risque, notamment en matière de sinistres climatiques. Le zonier, en établissant une relation entre une zone géographique et un coefficient de risque, permet d'intégrer cette dimension spatiale dans la tarification. L'objectif est d'affiner la tarification en fonction de la localisation, améliorant ainsi la capacité prédictive et la segmentation du modèle([2]).

4.1 Aspect théorique

La construction du zonier repose sur plusieurs étapes clés ([PERRIN, 2021]) :

1. Analyse du résidu (Écart entre la variable prédite [3] et la variable observée). Ce résidu est présumé être composé d'un signal géographique ainsi que d'un bruit :

$$\text{var\_prédite} = \text{var\_observée} + \text{bruit}$$

où

$$\text{var\_observée} = \text{var\_non\_géographique\_observée} + \text{var\_géographique\_observée}$$

2. Les résidus sont agrégés à une maille géographique (INSEE ou IRIS dans le cadre de ce mémoire) ;
3. Création de zones par découpage des résidus ;
4. Lissage géographique permettant d'homogénéiser les zones géographiques ;
5. Intégration de la variable « zonier » au modèle GLM hors zonier existant pour qu'elle serve de variable tarifaire basée sur la géographie. Les coefficients du modèle original auront été fixés, ce qui évite ainsi leur modification lors de l'ajout de cette nouvelle variable. Cette démarche vise à conserver l'information déjà expliquée par les variables internes, prévenant tout biais pouvant être introduit par l'intégration d'un élément supplémentaire au modèle ;
6. Attribution d'un coefficient à chaque zone définie.

Les 4 premières étapes seront effectuées à l'aide d'Akur8 mais aussi de Georev. Les résultats de la cinquième étape obtenus par ces deux méthodes seront confrontés afin de sélectionner la plus performante.

1. Association d'un coefficient de risque à une région géographique spécifique.
2. [PERRIN, 2021]
3. Fréquence ou coût moyen.

Par la suite, nous intégrerons la variable relative aux antécédents de sinistres climatiques des PM et PNOM en adoptant une approche similaire. L'inclusion des antécédents à ce stade est justifiée par leur définition au niveau contractuel (ou au niveau géographique de l'adresse). Par conséquent, ils contiennent un aspect géographique qui pourrait améliorer les performances des modèles suite à l'intégration du Zonier.

En dernier lieu, nous intégrerons les variables issues de la base DVF et MétéoNet. Le fait qu'elles contiennent des informations agrégées à la maille INSEE ou IRIS, les caractérise comme des données à dimension géographique. De la même manière que pour la variable des antécédents, nous nous attendons à ce que leur inclusion « affine » le modèle [1]. L'ajout de ces variables en fin de modélisation s'explique par notre volonté d'appréhender l'effet marginal qu'elles peuvent avoir, tout en maintenant un modèle climatique opérationnel sans ces données. Il convient de préciser que nous avons expérimenté l'ajout de ces variables avant celle du zonier. Toutefois, cette stratégie n'a pas produit de résultats significativement différents de la méthodologie que nous avons choisie d'adopter.

Ainsi, cette méthode nous permettra d'évaluer l'impact de l'intégration de données ouvertes dans la modélisation de la garantie climatique des propriétaires de maisons en MRH. Le prochain paragraphe sera centré sur les étapes 4 et 5.

4.2 Lissage des résidus et intégration (étapes 4 et 5)

Dans le cadre du lissage, 20 zones ont été choisies pour chaque modèle et chaque méthode. La paramétrisation de GeoRev s'est inspirée de celle adoptée pour le lissage de la base DVF. Le paramètre nmin a été fixé à 50, mais cette fois-ci le paramètre expo_min a été fixé au quantile d'exposition à 95 %. En effet, un quantile inférieur à celui-ci ne fournissait pas des résultats assez lissés. En intégrant les résultats de chaque méthode à la base de modélisation et en *fittant* les GLM, les résultats de l'étape 5 sont les suivants :

Modèle	**Fréquence**		**Coût moyen**	
	RMSE	**Gini**	**RMSE**	**Gini**
Akur8	0.1268	38.8%	5243	22.81%
Georev	0.1269	37.8%	5294	20.09%

TABLE 4.1 – Comparaison des performances entre Akur8 et GeoRev

Le zonier Akur8 permet dans le cas des deux modèles de meilleures performances que le zonier GeoRev. Ce sera le lissage retenu. Les graphiques suivants mettent en évidence l'impact de l'ajout de la variable zonier sur le modèle fréquence hors zonier. « Application de HZ » correspond à l'application du modèle HZ sur Z (orange), et « INSEE FREQ BN » correspond au modèle choisi (bleu). Le modèle a été sélectionné en équilibrant entre des performances optimales et la prévention du surapprentissage.

1. En d'autres termes, en améliore les performances.

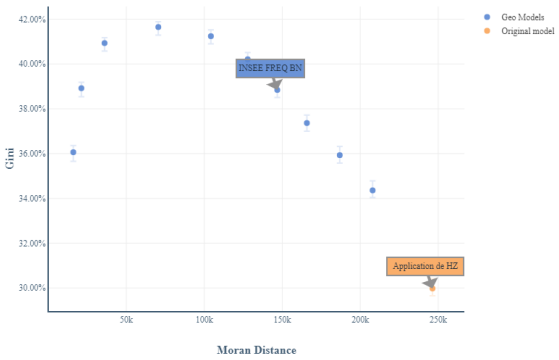


FIGURE 4.1 – Impact de l'ajout du zonier sur le Gini (fréquence)

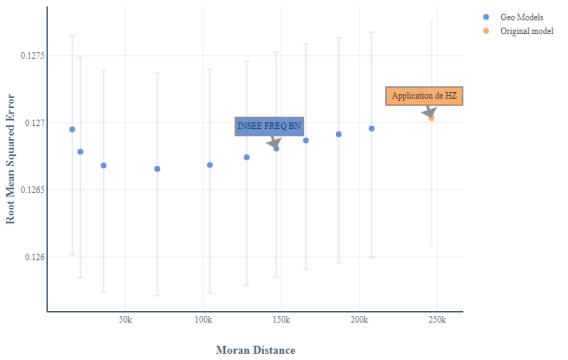


FIGURE 4.2 – Impact de l'ajout du zonier sur le RMSE (fréquence)

L'intégration de cette variable a un effet significatif : une augmentation de 9 points du Gini et une réduction du RMSE. Ceci renforce l'intérêt d'utiliser le zonier. La distance de Moran, indiquée en abscisse, est un indicateur d'autocorrélation spatiale. Elle représente le rayon d'une région circulaire où les points à l'intérieur de cette zone montrent une corrélation de plus de 50 % entre leurs distances relatives et les coefficients associés. En termes simples, si l'on se place dans un code INSEE donné, la distance de Moran indique la distance moyenne qu'il faudrait parcourir pour observer un changement significatif dans le coefficient de risque lié à la variable géographique.

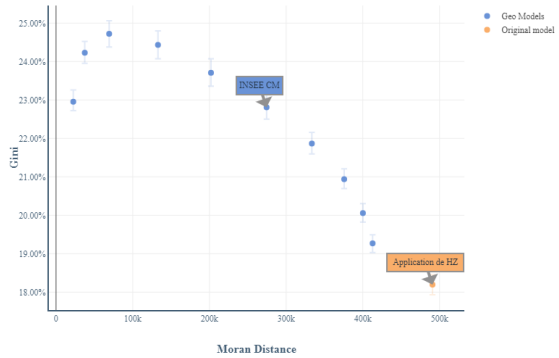


FIGURE 4.3 – Impact de l'ajout du zonier sur le Gini (coût moyen)

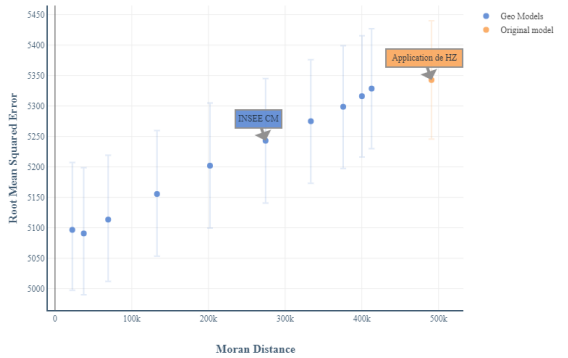


FIGURE 4.4 – Impact de l'ajout du zonier sur le RMSE (coût moyen)

Ces courbes mettent en évidence l'impact de l'ajout de la variable zonier sur le modèle coût moyen. Cet impact est tout aussi considérable que celui observé pour la fréquence.

Résultats cartographiques

Les résultats du zonier sont les suivants :

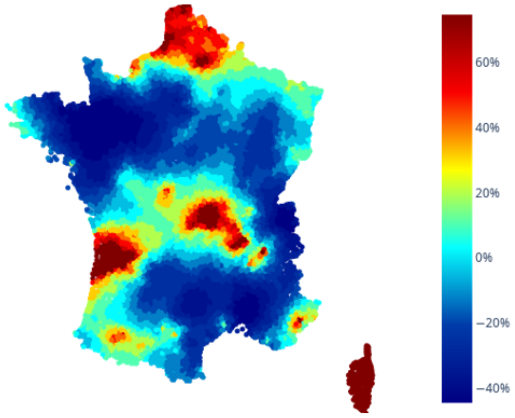


FIGURE 4.5 – Zonier fréquence, INSEE (coefficients)

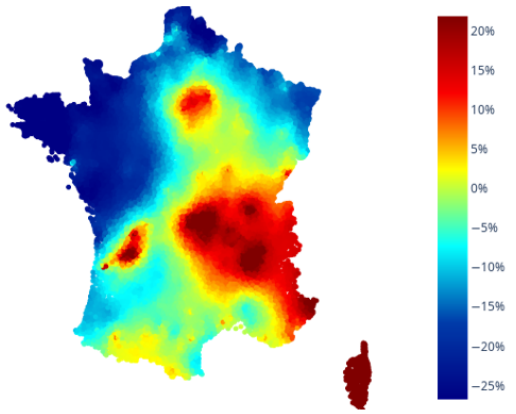


FIGURE 4.6 – Zonier coût moyen, INSEE (coefficients)

Le zonier relatif à la fréquence, illustré à gauche, révèle une concentration de coefficients plus élevés en Gironde et ses environs, dans la région Nord, en Corse, ainsi que dans le Sud-Est de la France. Ces zones coïncident avec les régions ayant le plus grand nombre de sinistres climatiques, comme identifié dans l'analyse de la section 1.2.2. Cette observation est donc en parfaite adéquation avec nos attentes. D'autre part, le zonier concernant le coût moyen, présenté à droite, montre des coefficients accentués dans le Sud et en Île-de-France. Cette tendance est également en phase avec les conclusions tirées de notre analyse de la sinistralité. Les résultats obtenus à l'échelle des IRIS corroborent ces observations. Les détails de leur zonier seront fournis en annexe. Il est important de souligner que l'intégration du zonier à la maille IRIS conduit à un Gini plus élevé. En effet, une agrégation à ce niveau de granularité permet d'affiner la segmentation du modèle, se traduisant par une augmentation de 12 points du Gini (au lieu de 9 comme la maille INSEE) et par une réduction légère de la RMSE.

4.3 Ajout des antécédents

L'introduction des antécédents dans le modèle influence significativement les performances de Gini, en particulier en ce qui concerne celui du modèle de fréquence. En effet, une amélioration de 2 points est observée, portant la valeur à 41.55 %. En ce qui concerne le coût moyen, l'effet est plus modeste, avec une légère augmentation à 22.82 %, contre 22.81 % sans les antécédents. Ainsi, dans le contexte des sinistres climatiques affectant les propriétaires de maison, la présence d'antécédents de ce type de sinistre semble avoir une influence plus significative sur la fréquence plutôt que sur le coût moyen. Les autres métriques ne montrent pas de variations significatives.

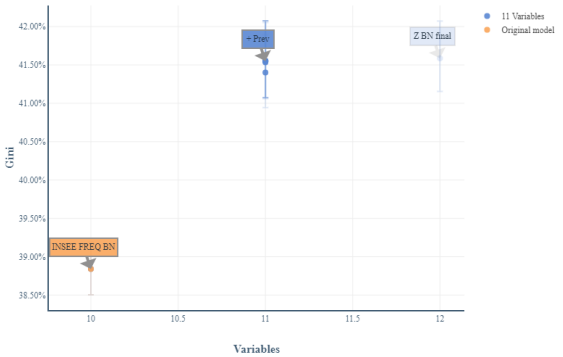


FIGURE 4.7 – Intégration des antécédents en bleu (fréquence)

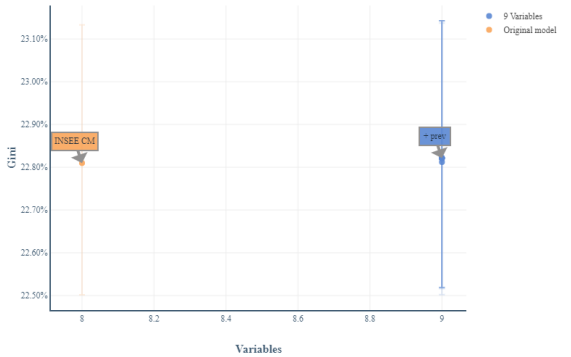


FIGURE 4.8 – Intégration des antécédents en bleu (coût moyen)

Notons également que les coefficients augmentent avec le nombre d'antécédents. Cette tendance est illustrée dans le graphique ci-dessous.

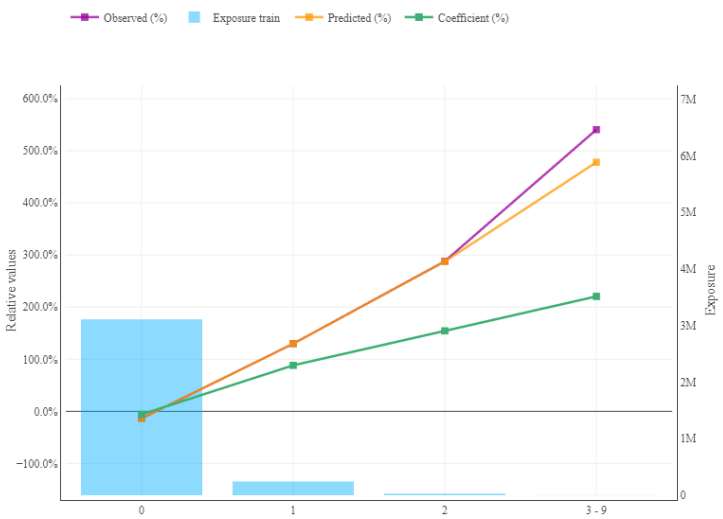


FIGURE 4.9 – Evolution de la fréquence en fonction de la variable des antécédents de sinistres climatiques des propriétaires de maisons sur une période de 10 ans. *L'axe des abscisses illustre le nombre d'antécédents.*

Cette observation et l'augmentation des performances de segmentation peuvent s'expliquer par plusieurs facteurs. En effet, au-delà de la simple localisation géographique d'un client, les antécédents pourraient refléter son niveau d'équipement, comme la présence d'un jardin ou d'une dépendance. De plus, ils peuvent également indiquer son comportement, qu'il soit plus ou moins préventif face aux risques climatiques. Il est essentiel de préciser que nous avons pris en compte les antécédents de sinistres sur une période de 10 ans.

Les observations ci-dessus sont en accord avec les anticipations formulées concernant cette variable au début de ce chapitre.

4.4 Ajout de la base DVF

Il est à noter que chaque variable a été intégrée séparément, sans jamais combiner plusieurs d'entre elles dans un même modèle.

4.4.1 Maille INSEE

Les résultats de l'intégration des données de la base de Demandes de Valeurs Foncières à la maille INSEE sont les suivants :

Modèle	**Fréquence**			**Coût moyen**		
	RMSE	**Gini**	**Déviance**	**RMSE**	**Gini**	**Déviance**
Sans données ouvertes	0.1267	41.55%	90 050	5243	22.82%	12.63
+ Valeur Foncière	0.1267	41.59%	90 030	5242	22.82%	12.63
+ Prix au m^2	0.1267	41.55%	90 040	5242	22.82%	12.63

TABLE 4.2 – Résultat de l'ajout de la base DVF, métriques, base de validation, maille INSEE

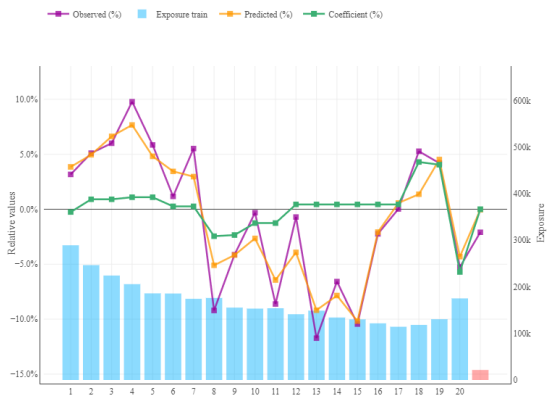


FIGURE 4.10 – Evolution de la fréquence (observée et prédite) en fonction de la zone de prix au m^2

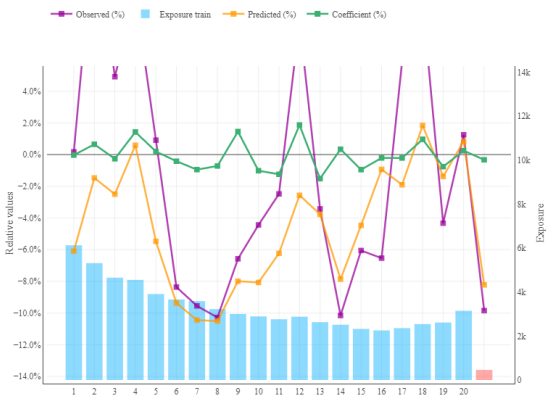


FIGURE 4.11 – Evolution du coût moyen (observé et prédit) en fonction de la zone de prix au m^2

Les spreads de coefficients pour la variable du prix au m^2 sont résumés dans le tableau suivant :

Spread	**Fréquence**	**Coût Moyen**
Spread 100	10.60%	3.44%
Spread 95	10.35%	2.73%

TABLE 4.3 – Spread de coefficients, ajout du prix au m^2

Analyse des résultats pour la fréquence

Pour le modèle de fréquence, l'ajout des données DVF n'a ni amélioré ni détérioré la précision du modèle en termes d'erreur quadratique moyenne (RMSE). En effet, elle reste égale à 0.1267. La variable relative à la valeur foncière a légèrement impacté le coefficient de Gini (de 41.55% à 41.59%) et la déviance, mais l'ajout du Prix au m^2 n'a pas eu d'effet notable.

Analyse des résultats pour le coût moyen

Pour le modèle coût moyen, la RMSE a baissé d'un point à l'ajout des données de la base DVF. Cela indique que leur ajout n'a pas eu d'impact significatif sur la précision du coût moyen. Le Gini est resté constant, ce qui veut dire que ces données n'ont pas contribué à segmenter davantage le modèle.

Analyse des résultats à la maille INSEE

Finalement, il semblerait que l'introduction des données ouvertes de la base DVF, à savoir la valeur foncière et le prix au m^2, n'ait pas conduit à des améliorations significatives des performances des modèles, que ce soit en termes de fréquence ou de coût moyen. Bien que l'ajout de la valeur foncière ait très légèrement amélioré le coefficient de Gini et réduit la déviance pour le modèle de fréquence, ces changements sont minimes.

De plus, lorsque nous analysons l'évolution de la fréquence et du coût moyen en fonction de la zone de prix sur les deux figures précédentes, nous réalisons que la tendance n'est pas lisible : il y a des variations, mais elles sont difficilement interprétables. À titre d'exemple, il n'est pas évident de conclure que le coût moyen des sinistres climatiques de PM est plus élevé dans les zones où le prix au m^2 est plus cher.

En analysant la dispersion des coefficients, nous observons que le prix au m^2 montre une variabilité plus prononcée dans le modèle de fréquence par rapport au coût moyen. Cette observation suggère encore une fois que le prix au m^2 pourrait avoir une influence plus fluctuante sur la fréquence des sinistres que sur leur coût moyen. Toutefois, cette variabilité est relativement modérée, indiquant une certaine stabilité de cette variable dans les deux modèles. D'autre part, la valeur foncière présente une dispersion des coefficients plus accentuée dans le modèle de fréquence comparativement au prix au m^2. Cela pourrait refléter une instabilité plus marquée de cette variable dans la prédiction de la fréquence

des sinistres. Néanmoins, en ce qui concerne le coût moyen, les dispersions des coefficients pour les deux variables sont assez similaires, suggérant une influence stable de ces variables sur la prédiction du coût moyen des sinistres, qui était en partie traduit par les métriques de performance.

Néanmoins, une question majeure émerge : pourquoi ces données influencent-elles davantage la fréquence que le coût moyen ? Plusieurs hypothèses pourraient être émises :

— Type de sinistre : La valeur foncière ou le prix au m^2 pourrait être lié à des zones où les sinistres sont plus fréquents, mais pas nécessairement plus coûteux. Par exemple, dans des zones à forte valeur foncière, il pourrait y avoir plus de sinistres mineurs qui augmentent la fréquence, mais qui n'ont pas un impact significatif sur le coût moyen (La zone 16 par exemple) ;

— Prévention : Dans les zones avec des valeurs foncières élevées, les propriétaires pourraient avoir davantage investi dans la prévention et l'entretien, réduisant ainsi le coût moyen des sinistres, mais pas nécessairement leur fréquence.

Toutefois, de façon surprenante, le volume de transactions immobilières, qui n'était pas initialement anticipé comme un facteur influent, a montré une certaine influence lorsqu'il a été intégré au modèle de coût moyen :

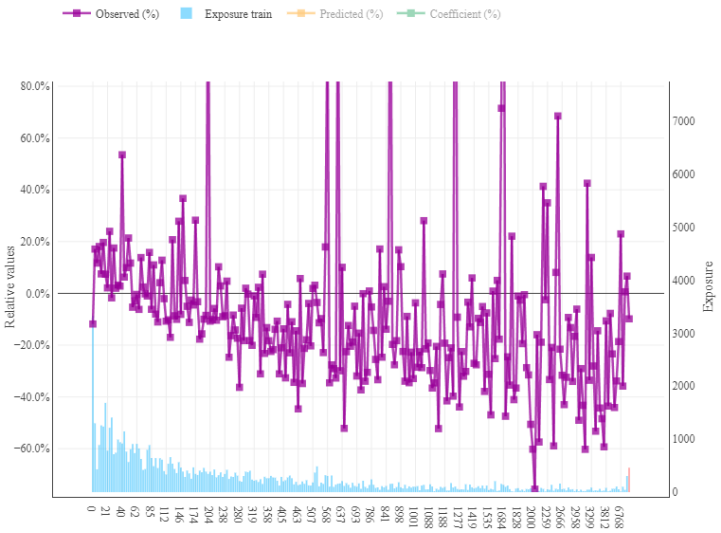


FIGURE 4.12 – Coût moyen des sinistres climatiques en fonction du nombre de transactions immobilières

Il semble y avoir une légère tendance à la baisse du coût moyen avec l'augmentation du nombre de transactions. Après une quantilisation en 10 intervalles, son intégration au modèle a légèrement amélioré les performances de Gini, surpassant très faiblement celles obtenues avec les valeurs foncières et le prix au mètre carré.

Modèle	**Fréquence**		**Coût moyen**	
	RMSE	**Gini**	**RMSE**	**Gini**
Sans DVF	0.1267	41.55%	5243	22.82%
+ Valeur Foncière	0.1267	41.59%	5242	22.82%
+ Prix au m^2	0.1267	41.55%	5242	22.82%
+ Nombre de transactions (INSEE)	**0.1267**	**41.54%**	**5242**	**22.87%**

TABLE 4.4 – Résultat de l'ajout de la base DVF, base de validation, maille INSEE 2

Ce résultat met en lumière l'importance d'explorer différentes variables et de ne pas se limiter aux données initialement envisagées. De plus, l'effet surprenant du nombre de transactions sur le modèle coût moyen à la maille INSEE peut découler de divers éléments. Nous pourrions supposer qu'un volume élevé de transactions dans une commune spécifique reflète la vitalité du marché immobilier local.

En effet, une forte activité peut indiquer une demande croissante, souvent due à des facteurs comme le développement régional, des infrastructures de qualité ou un emplacement géographique « prisé ». Par conséquent, cela pourrait réduire l'impact des climatiques et influencer le niveau de risque lié aux maisons de cette zone.

En conclusion, ces résultats montrent que même si ces données ouvertes peuvent fournir des données additionnelles non contenues dans les bases de données internes, elles peuvent ne pas avoir d'impact significatif sur la prévision de la fréquence ou du coût moyen des sinistres, du moins à l'échelle de l'INSEE. Néanmoins, il est possible que les données soient plus pertinentes à une autre échelle, comme celle de l'IRIS.

4.4.2 Maille IRIS

Les résultats de l'ajout des données de la base DVF à la maille IRIS sont :

Modèle	**Fréquence**		**Coût moyen**	
	RMSE	**Gini**	**RMSE**	**Gini**
Sans données ouvertes	0.1267	44.48%	5253	22.55%
+ Valeur Foncière	0.1266	44.50%	5252	22.57%
+ Prix au m^2	0.1266	44.51%	5252	22.56%
+ Nombre de transactions (IRIS)	0.1266	44.49%	5252	22.59 %

TABLE 4.5 – Résultat de l'ajout de la base DVF, base de validation, maille IRIS

Lorsque les données sont analysées à la maille IRIS, l'intégration des informations de la base DVF semble améliorer les performances du modèle de manière plus notable, bien que toujours modeste, par rapport à l'intégration à la maille INSEE. Cela pourrait s'expliquer par le fait que la maille IRIS offre une précision géographique plus fine que celle de l'INSEE, permettant de mieux saisir les variations locales. De plus, tout comme à la maille INSEE, le volume des transactions immobilières semble avoir un impact légèrement plus marqué sur le coût moyen que sur la fréquence. De même qu'à la maille INSEE, les tendances de la fréquence et du coût moyen en fonction de chacune des trois variables n'est pas lisible.

4.4.3 Synthèse des Résultats Concernant l'Intégration de la base DVF

L'intégration des données de la base DVF dans les modèles n'a pas entraîné d'amélioration significative des performances. Plusieurs facteurs peuvent être à l'origine de cette observation :

1. ***Redondance d'information*** : Il est possible que la base DVF apporte des renseignements déjà reflétés par d'autres variables du modèle (comme la variable du zonier), ou que ses renseignements intéragissent avec d'autres variables du modèle ;
2. ***Intégrité/Qualité des données*** : Les données de la base DVF pourraient ne pas être suffisamment qualitatives ou discriminantes pour les niveaux géographiques choisis ;
3. **Non pertinence** : Les informations immobilières ne pourraient tout simplement pas avoir de pertinence dans le cadre des climatiques.

Afin de vérifier la première hypothèse, il serait pertinent d'utiliser les informations issues de la base DVF pour d'autres type garanties, comme la garantie incendie ou la garantie vol. En effet, les sinistres liés aux incendies ont des conséquences importantes sur les biens immobiliers. Quant à la garantie vol, la valeur foncière, qui peut être perçue comme un indicateur de la richesse du ménage, pourrait jouer un rôle discriminant dans la survenue et la sévérité des vols.
En ce qui concerne l'hypothèse des intéractions, nous avons intégré quelques intéractions dans le modèle, mais celles-ci n'ont pas conduit à des améliorations notables. Quant à la deuxième hypothèse, une stratégie pourrait consister à agréger les données à une échelle départementale. Cependant, pour des données telles que les valeurs foncières, cette approche pourrait s'avérer moins pertinente en raison de la grande taille de cette maille.

Le prochain paragraphe sera centré sur l'utilisation des données MétéoNet.

4.5 Ajout de la base Météonet

Rappelons que suite à l'intégration de la base MétéoNet, les ensembles de données ont été considérablement réduits. Pour être précis, seuls les contrats de PM dont l'INSEE est référencé dans la base MétéoNet ont été retenus. Ainsi, la base du coût moyen ne compte plus que 7 924 entrées, tandis que celle de la fréquence s'élève à 478 350 entrées. En conséquence de cette réduction, nous avons adapté notre méthode de validation croisée en diminuant le nombre de *folds* à trois, dans le but d'assurer une certaine robustesse des résultats. Toutefois, malgré cette adaptation, la démarche générale de modélisation est restée inchangée, à savoir :

1. Sélection des variables basée sur le score d'importance et la corrélation ;
2. Modélisation sans données géographiques ;
3. Intégration du Zonier et des antécédents ;
4. Incorporation de la vitesse du vent.

Les résultats sont recensés dans le tableau suivant :

Modèle	**Fréquence**		**Coût moyen**	
	RMSE	**Gini**	**RMSE**	**Gini**
Sans MétéoNet	0.1229	45.44%	4913	17.29%
+ vmax	0.1228	45.69%	4912	17.34%
+ v5%	0.1228	45.70%	4910	17.52%

TABLE 4.6 – Résultat de l'ajout de la base Météonet, base de validation, maille INSEE

Où :

— *vmax a été construit en sélectionnant la vitesse maximale par INSEE ;*

— *v5% a été construit en prenant les 5% de vitesses de vent les plus élevées par INSEE, comme définie dans le chapitre 2.*

Spread	**Fréquence**	**Coût Moyen**
Spread 100 Vmax	29.50%	1.07%
Spread 100 V5%	42.49%	3.97%
Spread 95 Vmax	29.50%	1.07%
Spread 95 V5%	42.49%	3.97%

TABLE 4.7 – Spread de coefficients, ajout du prix au m^2

Le spread pour v5% est supérieur à celui de vmax. De plus, étant donné que v5% prend en compte un plus grand ensemble de données pour son calcul, il est plus enclin à offrir des résultats plus robustes. Par conséquent, nous privilégierons l'utilisation de v5% à vmax.

Nous notons que l'impact de la base Météonet sur les modèles est plus significatif que celui qu'a eu la base DVF sur les siens, en particulier pour le coût moyen. Cela suggère que les facteurs météorologiques, tels que la vitesse du vent, peuvent avoir une influence plus directe sur les sinistres climatiques que les données immobilières. Néanmoins, cet impact reste relativement modeste. Par conséquent, l'hypothèse suivante émerge : il est possible que le zonier capte déjà de manière efficace le risque lié à l'exposition aux vents violents. Ainsi, la variable issue de Météonet pourrait affiner cette estimation, entraînant un impact tarifaire accentué pour les deux derniers quantiles de v5% (zones 8 à 10) [1] :

1. Rappelons que les valeurs de v5% ont été divisées en 10 zones distinctes.

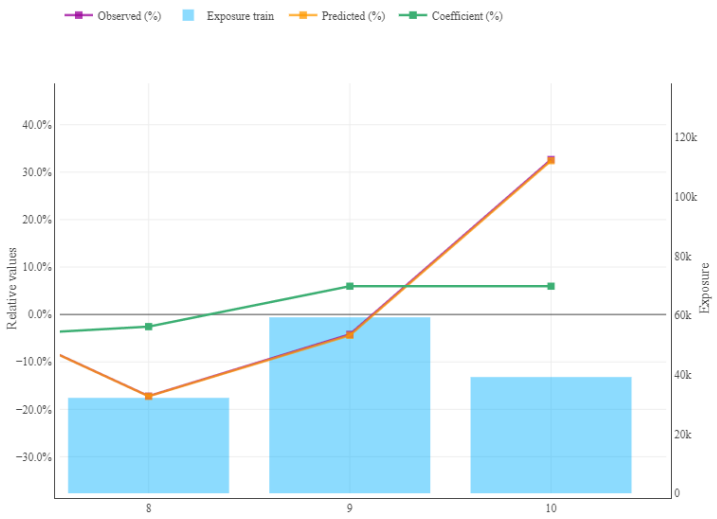


FIGURE 4.13 – Impact tarifaire des deux derniers quantiles de v5%. *Fréquence.*

La prochaine partie sera centrée sur l'impact de l'ajout des données ouvertes sur la prime pure, calculée en agrégeant la fréquence et le coût moyen. En effet, bien que l'impact sur ces deux derniers paramètres soit modeste, il est essentiel d'examiner les conséquences potentielles de leur agrégation.

4.6 Prime Pure

Afin d'évaluer l'impact des données ouvertes sur la prime pure des propriétaires de maison qu'ils soient occupants ou non, une agrégation des fréquences et des coûts moyens avec et sans l'intégration des données ouvertes a été effectuée. Le modèle actuellement en vigueur au sein de l'équipe sera également utilisé à titre de comparaison. Ce modèle sera référencé par le terme « ancien ».

Il est essentiel de préciser que ce modèle « Ancien » a été conçu en intégrant l'ensemble des profils disponibles (locataires d'appartements, etc.) à l'exception des PNO [1]. En effet, comme mentionné au chapitre 1, un modèle distinct a été spécifiquement conçu pour les PNO. Ces modèles ont été façonnés en utilisant un GLM basé sur une distribution TWEEDIE [2], et les antécédents de sinistres climatiques n'ont pas été pris en compte pour la modélisation des PNO.
Dans cette optique, nous avons choisi de séparer les propriétaires occupants de maison (POM) des propriétaires non-occupants (PNOM) afin d'appliquer le modèle « Ancien » le plus adapté à chaque profil. Il est à noter que les variables ayant le plus d'impact, identifiées lors de l'application des anciens modèles à nos bases, sont le nombre de pièces de la maison et l'année d'exposition.

4.6.1 Métriques de Performance

Les modèles renseignés comme « Nouveau avant X » correspondent au modèle obtenu après ajout des antécédents, juste avant l'ajout des données ouvertes de la base X.

Nombre de transactions (DVF)

Le tableau suivant récapitule les métriques de performance obtenues. Pour la base DVF, seule la variable relative au nombre de transactions est utilisée. La colonne « Profil » correspond au profil contenu dans la base de modélisation.

1. Propriétaires Non Occupants (appartements et maisons).
2. Et non une approche fréquence-coût moyen.

		Ancien	Nouveau (maille INSEE)		Nouveau (maille IRIS)	
			Avant DVF	Après DVF	Avant DVF	Après DVF
Profil	**Métrique**					
POM	**Gini**	39.95%	52.91%	52.93%	53.85%	53.85%
	RMSE	826.6	815.6	815.6	815.2	815.6
PNOM	**Gini**	23.38%	53.06%	53.03%	54.35%	54.37%
	RMSE	575.1	571.2	531.9	571.1	571.1
POM+PNOM	**Gini**	-	54.32%	54.37%	55.26%	55.45%
	RMSE	-	777.7	777.7	777.3	812.7

TABLE 4.8 – Comparaison des performances des modèles de prime pure, données DVF.

Dans le cas des POM, le coefficient de Gini a augmenté, passant de 39.95% à 52.91% (maille INSEE) et 53.85% (maille IRIS) avant l'intégration des données DVF. Cela indique une amélioration considérable de la capacité de segmentation du modèle qui avait déjà mis en place avant l'ajout des données DVF, par rapport à l'ancien modèle. De plus, la RMSE a légèrement diminué, passant de 826.6 à 815.6 (maille INSEE) et 815.2 (maille IRIS) avant l'intégration des données DVF, indiquant une meilleure précision du modèle. Dans le cas des PNOM, le coefficient de Gini a également augmenté de manière significative de 23.38% à 53.06% (maille INSEE) et 54.35% (maille IRIS) avant l'intégration des données DVF. La RMSE a quant à elle légèrement diminué de 575.1 à 571.2 (maille INSEE) et est restée stable à 571.1 (maille IRIS) avant l'intégration des données DVF.
Pour l'ensemble POM+PNOM, la maille IRIS est meilleure en termes de Gini, mais la RMSE est légèrement plus élevée après l'intégration des données DVF.

Finalement, l'intégration des données DVF n'a pas eu un impact significatif sur le Gini ou la RMSE pour les POM et PNOM, que ce soit à la maille INSEE ou IRIS. En effet, les valeurs sont presque identiques avant et après l'intégration des données DVF. Cette observation découle de leur impact assez faible sur les modèles de fréquence et de coût. Ainsi, c'est principalement le recours à une nouvelle méthode de modélisation (Modèle Avant DVF), plutôt que l'ajout de la base DVF, qui a entraîné une amélioration notable des performances des modèles comparativement au modèle ancien, en particulier, en termes de capacité discriminative (Gini).

MétéoNet

Une analyse similaire va être effectuée avec les données de MétéoNet.

		Ancien	Nouveau Avant MétéoNet	+ V5%	+V5% + DVF
Profil	**Métrique**				
POM	**Gini**	39.95%	53.15%	53.7%	53.7%
	RMSE	826.6	744.6	744.6	744.6
PNOM	**Gini**	21.9 %	52.17%	53.18%	53.18%
	RMSE	493.8	484	483.9	483.9
POM+PNOM	**Gini**	-	54.59%	54.96%	55.11%
	RMSE	-	703	702.9	702.9

TABLE 4.9 – Comparaison des performances des modèles de prime pure, données MétéoNet.

Pour les POM, le coefficient de Gini a augmenté, passant de 39.95% à 53.15% avant l'intégration des données de MétéoNet. Cela indique encore une fois une amélioration de la capacité de segmentation du modèle par rapport à l'ancien modèle. La RMSE a nettement diminué de 826.6 à 744.6 avant MétéoNet, indiquant une meilleure précision du modèle. L'intégration des données MétéoNet V5% s'est traduite par une légère amélioration du coefficient de Gini (+0.6 points). De même, pour l'ensemble PNOM, une hausse d'un point a été enregistrée.

Néanmoins, l'intégration de V5% n'a pas eu un impact significatif sur le coefficient de Gini ou la RMSE pour l'ensemble POM + PNOM. En effet, les valeurs sont presque identiques avant et après l'intégration des données V5%. De plus, l'ajout des données DVF (nombre de transactions) après MétéoNet V5% n'a également pas changé les performances.

Ainsi l'utilisation de nouvelles données (hors MétéoNet) a amélioré les performances des modèles par rapport au modèle « Ancien », mais comme précédemment, l'intégration des données MétéoNet V5% a eu un impact plus significatif que celle des données DVF. De plus, l'ajout des données DVF après MétéoNet V5% n'a pas du tout eu d'impact sur les performances.

4.6.2 Impact sur le total

Finalement, nous choisissons de regarder l'évolution de la prime pure totale prédite, notée var_pp, par profil, avec et sans données ouvertes :

$$\text{var\_pp} = \frac{\text{Montant total Prime Pure prédite} - \text{Montant total Prime Pure prédite du } \textbf{Modèle Ancien}}{\text{Montant total Prime Pure prédite du } \textbf{Modèle Ancien}},$$

Où Montant total Prime Pure prédite correspond à la prime pure prédite par le modèle élaboré tout au long de ce mémoire, avec ou sans données ouvertes.

Les résultats sont les suivants :

TABLE 4.10 – Variation de la prime pure prédite par profil, Cas de la base DVF, nombre de transactions

	PM4-	PM56	PM7+	PNOM
Zonier INSEE + Antécédents	+1.38 %	-8.35 %	-21.27 %	+79.49 %
+ nombre de transactions, INSEE	+1.42 %	-8.28 %	-21.16 %	+79.49 %
Zonier IRIS + Antécédents	+2.15 %	-7.49 %	-20.74 %	+79.92 %
+ nombre de transactions, IRIS	+2.20 %	-7.46 %	-20.67 %	+80.02 %

À titre d'exemple, la première ligne nous indique que la prime pure totale prédite des PM4- a connu une hausse de 1.38 % en utilisant la nouvelle modélisation par rapport au modèle Ancien.

TABLE 4.11 – Variation de la prime pure prédite par profil, Cas de la base MétéoNet

	PM4-	PM56	PM7+	PNOM
Zonier INSEE + Antécédents	+3.81 %	+6.96 %	-7.10 %	+67.88 %
+ MétéoNet v5%	+4.37 %	+6.99 %	-7.34 %	+68.17 %

Finalement, les variations de la prime pure vont refléter les changements dans la perception du risque des différents profils entre les différents modèles. Selon le profil et la résolution géographique (INSEE vs. IRIS), la prime pure peut augmenter ou diminuer. De plus, l'ajout de données spécifiques, comme le nombre de transactions, mais surtout v5%, peut affiner ces variations. À titre d'exemple, pour les profils de PM4-, l'augmentation de prime totale observée due à l'application du nouveau modèle avant données ouvertes (3.81%) suggère que le risque associé à ce profil est légèrement plus élevé avec les nos données par rapport à l'ancien modèle. Lorsque l'on ajoute v5%, cette augmentation est légèrement plus prononcée, passant à 4.37%.

De plus, ces résultats suggèrent que, pour le profil PNOM, le risque perçu avec la nouvelle modélisation est nettement plus élevé par rapport à l'ancien modèle (ordre de + 80%). Ainsi, cette hausse significative de la prime pure pour les PNOM indique une réévaluation majeure du risque associé à ce profil dans un contexte où ils ont été modélisés avec les POM avec une approche fréquence-coût moyen. En effet, nous remarquons encore une fois que les écarts sont davantage dus au changement de méthodologie qu'à l'ajout des données ouvertes.

Conclusion

Dans un contexte marqué par le changement climatique et l'essor des *Open Data* (données ouvertes), ce mémoire, réalisé au sein de l'équipe Actuariat tarification Affaires Nouvelles et remplacements (AN) de la direction MultiRisques Habitation (MRH) d'AXA France, a étudié l'impact de l'intégration des données ouvertes dans la modélisation de la garantie climatique des propriétaires de maison. Les données utilisées pour cette modélisation couvrent la période de 2016 à 2022. Le modèle collectif (fréquence-coût moyen) a été adopté, et les GLM ont été privilégiés face aux forêts aléatoires, parce que ces dernières, bien qu'efficaces, nécessitaient un temps d'exécution plus long.

Trois bases de données ouvertes ont été examinées : la base des arrêtés de catastrophes naturelles (CATNAT), celle des données météorologiques de MétéoNet et la base de Demandes de Valeurs Foncières (DVF). La base CATNAT avait été initialement envisagée afin d'améliorer les performances du modèle fréquence. Ainsi, l'étude s'est centrée en priorité sur les arrêtés relatifs aux tempêtes, étant donné qu'elles représentent plus de 60 % des sinistres climatiques du portefeuille. Cependant, après examen de ces arrêtés, il s'est avéré que les données n'étaient pas pertinentes. En effet, la base de modélisation commence en 2016 et il n'y a pas eu d'arrêtés CATNAT pour les tempêtes depuis 1989. De ce fait, l'analyse s'est recentrée sur les deux autres bases.

La base DVF, qui renseigne les valeurs foncières des maisons issues de transactions entre 2017 et 2022, a été exploitée à l'échelle communale (INSEE), et à celle de l'IRIS. Malgré son potentiel apparent pour affiner le modèle coût moyen, son apport s'est avéré relativement modeste. Plusieurs hypothèses ont été émises afin d'expliquer ce résultat, notamment, une remise en question de la qualité des données, et l'éventuelle non pertinence des informations immobilières dans le cadre de la modélisation des climatiques.

Les données relatives à la vitesse du vent issues de MétéoNet ont été intégrées, adaptées à la maille INSEE. Pour leur exploitation, les bases fréquence et coût moyen ont été filtrées afin de ne retenir que les contrats situés dans des zones INSEE disposant d'informations sur la vitesse du vent. Cette démarche a été adoptée car les données MétéoNet sont exclusivement localisées dans le Nord-Ouest et le Sud-Est de la France. Finalement, ces données ont contribué à une amélioration modeste des performances, et par conséquent des prédictions. Néanmoins, l'impact de ces données sur les performances du modèle reste plus élevé que celui des données de la base DVF. De plus, leur pertinence pourrait être discutée, notamment en raison de la limitation des données jusqu'en 2018. En effet, étant donné les changements rapides et imprévisibles induits par le réchauffement climatique ces dernières années, des données plus récentes pourraient être nécessaires afin d'obtenir une image fidèle des risques climatiques actuels. Cela met en exergue l'importance d'explorer d'autres bases météorologiques, de préférence plus récentes.

Finalement, ces résultats ouvrent des perspectives de recherche, comme l'exploration de la base DVF pour d'autres garanties, telles que celle relative à l'incendie, étant donné l'impact significatif de ce type de sinistre sur les biens immobiliers. De plus, nous pourrions envisager une modélisation spécifique selon le type de sinistre climatique, en complément d'*Open Data* de type météorologique, afin d'en étudier l'impact. En conclusion, ce mémoire a mis en lumière certains enjeux et des opportunités précises associés à l'intégration de données ouvertes dans la modélisation de la garantie climatique des propriétaires de maison en assurance habitation. De plus, il est important de souligner que c'est en réalité le modèle hors données ouvertes, qui a grandement amélioré la segmentation des propriétaires de maisons. Ce modèle a permis d'ajuster le risque, en particulier pour les maisons les plus équipées.

Il est espéré que ces découvertes guideront les démarches futures pour perfectionner les modèles et appréhender de manière renouvelée les risques climatiques pour les propriétaires de maisons, qu'ils soient occupants ou non.

Note de synthèse

Dans un contexte marqué par le changement climatique et l'essor des *Open Data* (données ouvertes), ce mémoire, réalisé au sein de l'équipe actuarielle Affaires Nouvelles et remplacements (AN) de la direction MultiRisques Habitation d'AXA France, étudie l'impact de l'intégration des données ouvertes dans la modélisation de la garantie climatique des propriétaires de maison, qu'ils soient occupants (PO) ou non (PNOM).
La garantie climatique couvre les sinistres liés aux événements climatiques tels que les tempêtes, les inondations, la grêle, et tout autre événement climatique n'ayant pas fait l'objet d'un arrêté de catastrophe naturelle. En effet, lorsqu'un sinistre lié à un évènement climatique est déclaré, une garantie climatique est automatiquement ouverte. Si plus tard l'évènement climatique concerné fait l'objet d'un arrêté de catastrophe naturelle, une garantie catastrophes naturelles (CATNAT) est ouverte pour l'indemnisation des dommages qui y sont liés.
Les tempêtes constituent la majorité des sinistres climatiques, représentant plus de 60 % des incidents en termes de nombre et de coûts. Par ailleurs, la sécheresse est traitée séparément, bénéficiant d'une garantie spécifique et activée lors d'un arrêté CATNAT Sécheresse.

Problématique

Nous avons observé que, entre 2016 et 2022, le ratio de Sinistres sur Cotisations (S/C) de la garantie climatique dépasse les 180 % pour les propriétaires de maisons (PM), qui représentent plus de 90 % du portefeuille. Plus précisément, les PNOM affichent les S/C climatiques les plus élevés dans toutes les régions, à l'exception de l'Ile-de-France. Néanmoins, ce n'est pas la catégorie de PM qui a les charges de sinistres climatiques les plus élevées. Cette observation suggère que leurs cotisations ne seraient particulièrement pas suffisantes pour couvrir leurs sinistres. Ainsi, les propriétaires de maison, qu'ils soient occupants ou non, sont exposés à un risque accru de sinistres climatiques, mettant en avant la nécessité d'améliorer la tarification de leur prime pure. Face à ces observations et à l'impact du changement climatique sur la fréquence et le coût des sinistres, il devient impératif de sophistiquer le modèle existant afin d'améliorer le ratio S/C des propriétaires de maisons, en prenant en compte les PNOM. En effet, du fait de la nature spécifique de ce profil, leur prime pure était initialement calculée séparément, conjointement avec les propriétaires non occupants d'appartements. Dans cette perspective, l'objectif de cette étude sera d'évaluer si l'utilisation d'*Open Data* peut contribuer à l'amélioration de la modélisation de leur prime pure. Si cette étude s'avère fructueuse, les bases de données utilisées seront testées sur d'autres garanties.

Méthodologie

La méthodologie de modélisation de la prime pure qui sera mise en pratique correspondra à celle du modèle collectif :

$$\text{Prime Pure} = \text{Fréquence} \times \text{Coût Moyen}$$

La fréquence des sinistres est principalement influencée par le comportement de l'assuré, tandis que le coût est davantage déterminé par les caractéristiques du bien assuré. De plus, il existe un décalage temporel entre ces deux concepts : la survenance du sinistre est connue dès sa déclaration à l'assureur, tandis que le coût réel n'est établi qu'à la fin du processus de règlement éventuel. C'est en prenant en compte ces deux aspects dans notre modèle que nous serons en mesure de mieux appréhender les

risques spécifiques aux Propriétaires de Maison, et de proposer une prime pure plus précise et adaptée à leur profil. Ainsi, plus la détermination de la prime pure sera précise, plus l'étude de l'impact des données ouvertes le sera.

Nous envisageons d'exploiter trois sources de données ouvertes pour l'étude. Chacune de ces bases de données sera soigneusement traitée et adaptée afin de répondre aux exigences spécifiques de notre étude. Voici un aperçu de notre approche pour chaque base :

— La base des arrêtés catastrophes naturelles, disponible sur le site du gouvernement français « data.gouv.fr ». L'exploitation de cette base vise à améliorer le modèle fréquence. Nous avons choisi de sélectionner uniquement les arrêtés spécifiquement liés aux tempêtes, étant donné que les sinistres liés aux tempêtes représentent plus de 60 % de notre portefeuille. L'objectif principal est de déceler une corrélation potentielle entre les sinistres de notre portefeuille et les décrets relatifs aux tempêtes ;

— La base de Demandes de valeurs foncières, également disponible sur le site du gouvernement. Dans la base de modélisation interne, nous disposons d'informations sur les caractéristiques des assurés et leur mobilier, mais pas sur les valeurs immobilières des biens assurés. Ainsi, l'exploitation de cette base vise à améliorer le modèle de coût moyen. La base DVF a nécessité un traitement des valeurs manquantes, des valeurs aberrantes et des doublons à l'aide de SAS. Nous avons choisi d'agréger les valeurs foncières et les prix au m^2 à la maille INSEE et à la maille IRIS. Deux méthodes de lissage ont été utilisées : celle des k-proches-voisins, basée uniquement sur la distance entre les zones géographiques, et celle de l'outil interne GeoRev (sur Python), qui prend en compte à la fois la distance et la crédibilité des données. La crédibilité, dans ce contexte, se réfère à la fiabilité des données pour une zone géographique spécifique, déterminée par le volume de transactions immobilières. Ainsi, après avoir quantifié les données en 20 zones avec les deux méthodes de lissage, nous avons opté pour les résultats de la méthode GeoRev en raison de sa pertinence accrue grâce à la prise en compte de la crédibilité. Les résultats de GeoRev sont donnés sur les cartes suivantes.

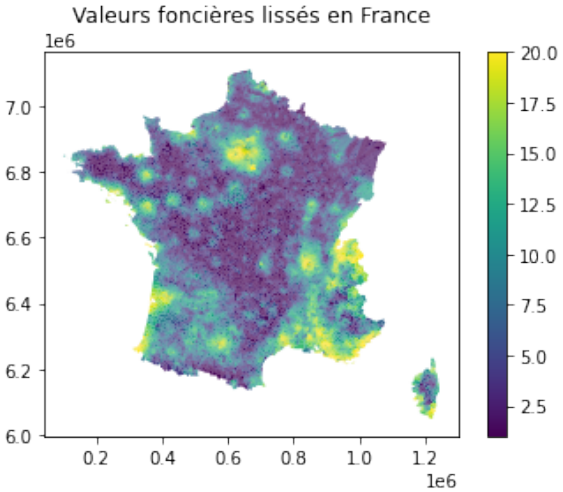


FIGURE 4.14 – Valeurs foncières lissées par IRIS (GeoRev)

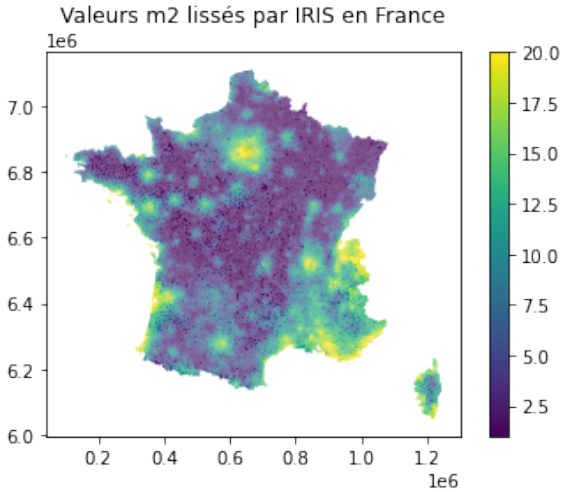


FIGURE 4.15 – Prix au m^2 lissés par IRIS (GeoRev)

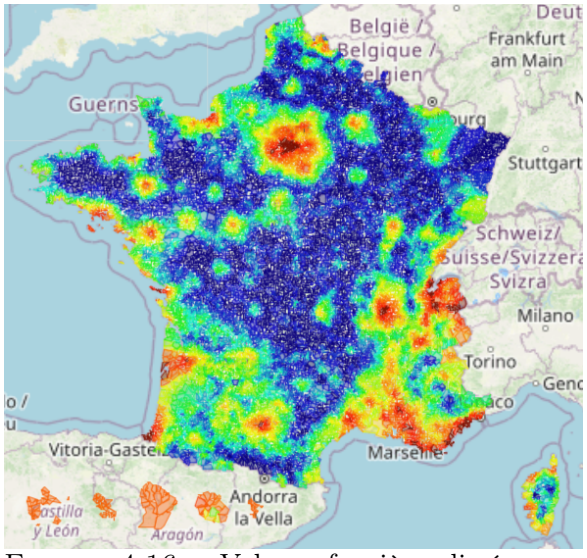


FIGURE 4.16 – Valeurs foncières lissées par INSEE (GeoRev), bibliothèque `maplot`

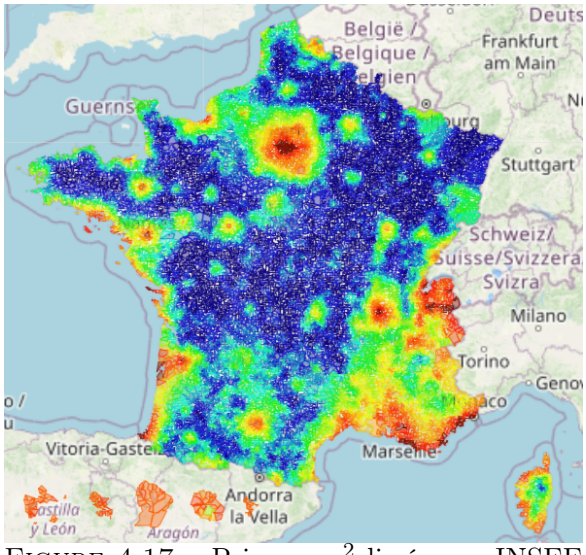


FIGURE 4.17 – Prix au m^2 lissés par INSEE (GeoRev), bibliothèque `maplot`

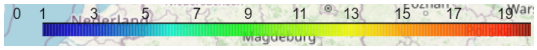


FIGURE 4.18 – Echelle

— La base météorologique de Météonet, qui contient des informations sur des paramètres météorologiques. L'exploitation de cette base vise également à améliorer le modèle fréquence. En effet, elle fournit des données sur la vitesse des vents de 461 stations situées au Nord-Ouest et au Sud-Est de la France pour la période s'étendant de 2016 à 2018. Après avoir traité les doublons et les valeurs manquantes, nous avons décidé d'agréger les données par code INSEE. Pour chaque code INSEE, nous avons calculé la moyenne des 5 % des vitesses de vent les plus élevées.

Ensuite, il convient d'expliciter la constructiion de la base de modélisation. Nous avons organisé notre base de modélisation selon l'image de risque. Cette organisation s'est basée sur une combinaison de diverses sources de données internes, telles que les bases contrats, les bases sinistres et les bases clients.
L'un des traitements clés a été la détermination d'un seuil pour les sinistres graves liés aux événements climatiques, en s'appuyant sur la théorie des valeurs extrêmes. Ce seuil, fixé à 41 000, a été choisi afin que nous puissions uniquement modéliser les sinistres dits « attritionnels », correspondant aux sinistres à montants de charge dits classiques et à des fréquences d'occurence importantes. La distinction entre sinistres graves et sinistres attritionnels était essentielle afin de garantir l'homogénéité du portefeuille et assurer une modélisation pertinente.
Les sinistres ont ensuite été ajustés temporellement grâce à la méthode Chain-Ladder. Nous avons également choisi l'année de survenance des sinistres comme variable explicative pour tenir compte de l'inflation.
Enfin, après avoir traité les valeurs manquantes et les valeurs aberrantes, il a été choisi de discrétiser certaines variables continues afin d'identifier les éventuels effets non linéaires dans la distribution des données. Ainsi, c'est après ce nettoyage qu'ont été générées les bases pour la fréquence et le coût moyen des PM. Une analyse des corrélations, et une analyse de l'importance des variables basée sur des priors bayésiens, ont été effectuées à l'aide du logiciel de tarification Akur8, afin de réduire le nombre de variables explicatives de chaque base.

Par la suite, nous avons intégré les bases de données ouvertes à nos bases de modélisation. Concernant l'étude de l'impact des données de la base MétéoNet, nous avons décidé de ne retenir dans les bases que les contrats liés aux codes INSEE disposant d'informations sur les vitesses de vent, afin de cibler plus précisément l'étude.

Enfin, deux méthodes de modélisation ont été considérées. D'une part, les Modèles Linéaires Généralisés (GLM) pénalisés de type LASSO, reconnus pour leur pertinence et couramment utilisés au sein de l'équipe. D'autre part, les forêts aléatoires, une technique d'apprentissage statistique, ont été adoptées pour enrichir l'analyse. Afin d'évaluer la qualité des modèles, les principales métriques de performance ayant été étudiées sont le Gini, la RMSE et la Déviance.
Un aspect central de nos modélisations est la prise en compte du « zonier ». En assurance Habitation, cette notion est essentielle car elle englobe les variables reflétant l'environnement géographique du contrat. Après avoir pris en compte cette dimension et les antécédents de sinistres, nous intégrerons les variables issues des données ouvertes aux modèles.

Résultats et interprétation

Premièrement, l'utilisation des GLM a été favorisée par rapport à celle des forêts aléatoires. En effet, les forêts aléatoires ont requis un temps d'exécution plus long, et n'ont pas montré de performances supérieures à celles du GLM, en particulier en ce qui concerne la RMSE. De surcroît, grâce à leur nature paramétrique, les GLM offrent une interprétation du modèle plus directe, plus lisible et intuitive.

Modèle	**Fréquence**		**Coût moyen**	
	RMSE	**Gini**	**RMSE**	**Gini**
GLM pénalisé	0.1266	29.23 %	5281	17.41 %
RfR	0.2145	31.24 %	5294	17.94 %

TABLE 4.12 – Comparaison des métriques d'évaluation entre les GLM et RfR sur la base de validation

Deuxièmement, la base CATNAT s'est avérée non pertinente. En effet, il n'y a pas eu d'arrêtés CATNAT pour les tempêtes depuis 1989 et étant donné que la base de modélisation commence en 2016, l'exploitation de cette base n'était pas pertinente.
Ensuite, l'intégration de la base DVF a apporté une amélioration très modeste aux performances des modèles et de leurs prédictions. Plusieurs éléments peuvent permettre d'expliquer ce résultat, notamment, une remise en question de la qualité des données, et l'éventuelle non pertinence des informations immobilières dans le cadre de la modélisation des climatiques.

Enfin, les données issues de Météonet ont contribué à une amélioration modeste des performances et donc des prédictions. Néanmoins, l'impact de ces données sur les performances du modèle reste plus élevé que celui des données de la base DVF. De plus, leur pertinence pourrait être discutée, notamment en raison de la limitation des données jusqu'en 2018. En effet, étant donné les changements rapides et imprévisibles induits par le réchauffement climatique ces dernières années, des données plus récentes pourraient être nécessaires afin d'obtenir une image fidèle des risques climatiques actuels. Une solution envisagée serait d'effectuer un lissage, en particulier pour les régions du Nord-Ouest (NO) et du Sud-Est (SE), et la compléter par des prédictions basées sur la théorie des séries temporelles, etc. Cette approche n'a pas été approfondie dans le cadre de ce mémoire.

Conclusion

En conclusion, l'intégration des données ouvertes a révélé des défis et des opportunités concernant la modélisation de la garantie climatique, spécifiquement pour le segment des Propriétaires de Maison. L'exploration approfondie de la base DVF pour d'autres garanties telles que l'incendie ou le vol, ainsi qu'une modélisation adaptée selon le type de sinistre climatique, constituent des axes de recherche potentiels pour l'avenir.

Executive Summary

In a context marked by climate change and the rise of Open Data, this thesis, carried out within the New Business and Replacements (AN) actuarial team of the Multi-Risk Home department of AXA France, studies the impact of integrating Open Data into the modelling of the climate guarantee for homeowners, whether they are occupants (PO) or not (PNOM).
The climate guarantee covers claims related to climatic events such as storms, floods, hails, and any other weather event that has not been the subject of a natural disaster decree. Indeed, when a claim related to a weather event is declared, a climate guarantee is automatically opened. If later the weather event in question is the subject of a natural disaster decree, a natural disaster guarantee (CATNAT) is opened for the compensation of related damages.
Storms constitute the majority of weather-related claims, accounting for over 60% of the incidents in terms of number and cost. Furthermore, drought is treated separately, benefiting from a specific guarantee.

Problem Statement

We observed that, between 2016 and 2022, the Claims to Premiums (C/P) ratio of the climate guarantee exceeds 180% for homeowners (PM), who represent more than 90% of the portfolio. More precisely, the PNOMs display the highest C/P climate ratios in all regions, except for Paris Region. However, this is not the PM category that has the highest weather-related claim charges. This observation suggests that their premiums would not be particularly sufficient to cover their claims. Thus, homeowners, whether occupants or not, are exposed to an increased risk of weather-related claims, highlighting the need to improve the pricing of their net premium. Given these observations and the impact of climate change on the frequency and cost of claims, it becomes imperative to refine the existing model to improve the C/P ratio of homeowners, taking into account the PNOMs. Indeed, due to the specific nature of this profile, their net premium was initially calculated separately, jointly with non-occupying apartment owners. In this perspective, the objective of this study will be to assess whether the use of *Open Data* can contribute to improving the modelling of their net premium. If this study proves fruitful, the databases used will be tested on other guarantees.

Methodology

The net premium modelling methodology to be implemented will correspond to that of the collective model :

$$\text{Net Premium} = \text{Frequency} \times \text{Average Cost}$$

The frequency of claims is mainly influenced by the behavior of the insured, while the cost is more determined by the characteristics of the insured property. Moreover, there is a temporal gap between these two concepts : the occurrence of the claim is known as soon as it is declared to the insurer, while the actual cost is established only at the end of the possible settlement process. By taking these two aspects into account in our model, we will be better able to understand the specific risks of Homeowners and propose a more precise and adapted net premium. Consequently, the more precise the determination of the net premium, the more the study of the impact of Open Data will be

We plan to exploit three Open Data sources for the study. Each of these databases will be carefully

processed and adapted to meet the specific requirements of our study. Here is an overview of our approach for each base :

— The natural disaster decree database, available on the French government website "*data.gouv.fr*". The exploitation of this base aims to improve the frequency model. We chose to select only decrees specifically related to storms, as storm-related claims represent over 60% of our portfolio. The main objective is to detect a potential correlation between the claims in our portfolio and the storm-related decrees ;

— The Real Estate Value Requests database, also available on the government website. In the internal modelling base, we have information on the characteristics of the insured and their furniture, but not on the real estate values of the insured properties. Thus, the exploitation of this base aims to improve the average cost model.
The DVF base required processing of missing values, outliers, and duplicates using SAS. We chose to aggregate property values and prices per m^2 to the zip and *IRIS* mesh. Two smoothing methods were used : the k-nearest neighbors method, based solely on the distance between geographical areas, and the internal GeoRev tool (on Python), which takes into account both distance and data credibility. Credibility, in this context, refers to the reliability of data for a specific geographical area, determined by the volume of real estate transactions.
Thus, after quantifying the data into 20 areas with the two smoothing methods, we opted for the results of the GeoRev method due to its increased relevance thanks to the consideration of credibility. The results of GeoRev are given on the following maps.

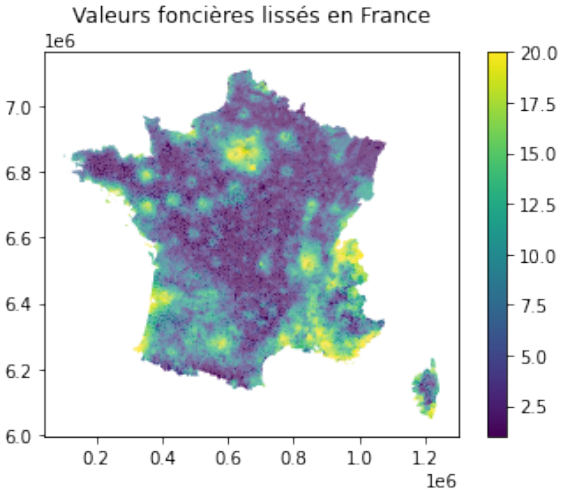


FIGURE 4.19 – Smoothed property values by IRIS (GeoRev)

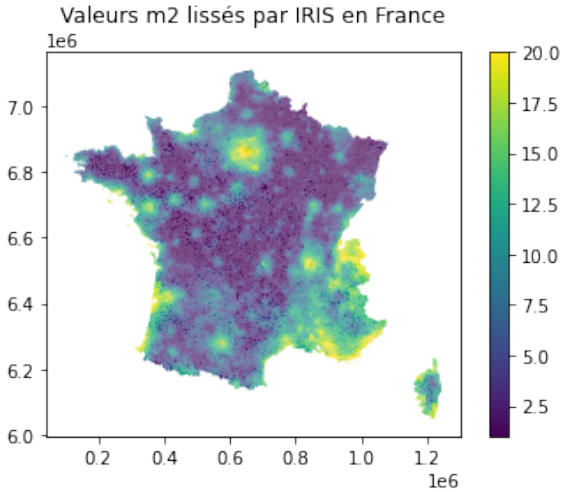


FIGURE 4.20 – Smoothed prices per m^2 by IRIS (GeoRev)

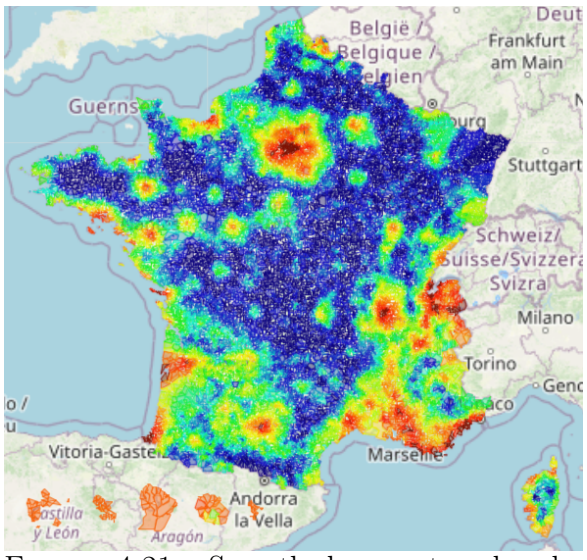


FIGURE 4.21 – Smoothed property values by INSEE (GeoRev), `maplot` library

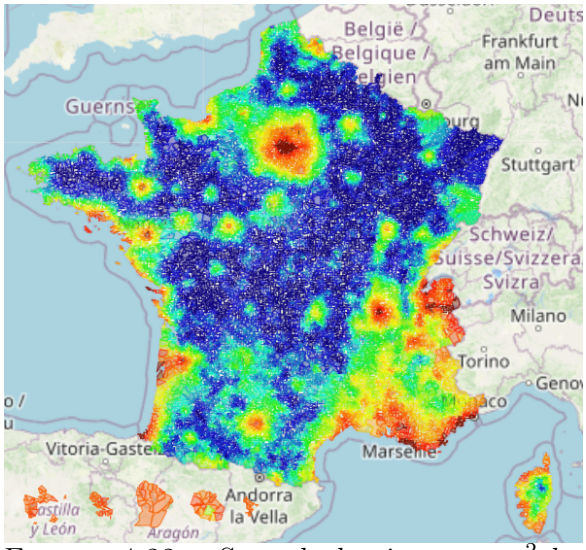


FIGURE 4.22 – Smoothed prices per m^2 by INSEE (GeoRev), `maplot` library

FIGURE 4.23 – Scale

— The MeteoNet meteorological database, which contains information on meteorological parameters. The exploitation of this base also aims to improve the frequency model. Indeed, it provides data on wind speeds from 461 stations located in the North-West and South-East of France for the period from 2016 to 2018. After processing duplicates and missing values, we decided to aggregate the data by zip code. For each zip code, we calculated the average of the 5% highest wind speeds.

Next, the construction of the modelling base should be explained. We organized our modelling base according to the risk image. This organization was based on a combination of various internal data sources, such as contract bases, claim bases, and customer bases.
One of the key treatments was determining a threshold for severe claims related to weather events, based on extreme value theory. This threshold, set at 41 000, was chosen so that we could only model so-called "attritional" claims, corresponding to claims with so-called conventional charge amounts and significant occurrence frequencies. The distinction between severe and attritional claims was essential to ensure portfolio homogeneity and ensure relevant modelling.
Claims were then adjusted over time using the Chain-Ladder method. We also chose the year of occurrence of the claims as an explanatory variable to account for inflation.
Finally, after processing missing values and outliers, it was decided to discretize certain continuous variables to identify any non-linear effects in the data distribution. Thus, it was after this cleaning that the bases for the frequency and average cost of PMs were generated. A correlation analysis, and an analysis of variable importance based on Bayesian priors, were carried out using the Akur8 pricing software, in order to reduce the number of explanatory variables for each base.

Subsequently, we integrated the Open Data databases into our modelling bases. Regarding the study of the impact of the MeteoNet database data, we decided to retain in the bases only the contracts linked to the zip codes with wind speed information, in order to more precisely target the study.

Finally, two modelling methods were considered. On the one hand, penalized LASSO Generalized Linear Models (GLM), known for their relevance and commonly used within the team. On the other hand, random forests, a statistical learning technique, were adopted to enrich the analysis. To evaluate the quality of the models, the main performance metrics studied were Gini, RMSE, and Deviance.
A central aspect of our models is the consideration of the zoning. In Home Multi-Risk Insurance, this concept is essential as it includes variables reflecting the geographical environment of the contract. After taking into account this dimension and the history of claims, we will integrate the variables from

Open Data into the models.

Results and Interpretation

Firstly, the use of GLMs was favored over that of random forests. Indeed, random forests required a longer execution time and did not show performance superior to that of the GLM, especially in terms of RMSE. Moreover, thanks to their parametric nature, GLMs offer a more direct, readable, and intuitive interpretation of the model.

Model	**Frequency**		**Average Cost**	
	RMSE	**Gini**	**RMSE**	**Gini**
Penalized GLM	0.1266	29.23%	5281	17.41%
RfR	0.2145	31.24%	5294	17.94%

TABLE 4.13 – Comparison of evaluation metrics between GLM and RfR on the validation base

Secondly, the CATNAT base proved to be irrelevant. Indeed, there have been no CATNAT decrees for storms since 1989 and since the modelling base starts in 2016, the exploitation of this base was not relevant.

Then, the integration of the DVF base brought a very modest improvement to the performance of the models and their predictions. Several elements can explain this result, notably questioning the quality of the data, and the possible irrelevance of real estate information in the context of modelling weather events.

Finally, the data from MeteoNet contributed to a modest improvement in performance and therefore predictions. However, the impact of this data on model performance remains higher than that of the DVF database data. Moreover, their relevance could be discussed, especially because of the limitation of the data until 2018. Indeed, given the rapid and unpredictable changes induced by global warming in recent years, more recent data may be needed to obtain an accurate picture of current weather risks. One solution considered would be to smooth, especially for the North-West (NO) and South-East (SE) regions, and complete it with predictions based on time series theory, etc. This approach was not explored in this thesis.

Conclusion

As a conclusion, the integration of Open Data revealed challenges and opportunities regarding the modelling of the weather guarantee, specifically for the Homeowners segment. The deep exploration of the DVF base for other guarantees such as fire or theft, as well as modelling adapted according to the type of weather event, are potential avenues of research for the future.

Annexe A

Indice FFB

L'indice **FFB** de la construction, associé à la **F**édération **F**rançaise du **B**âtiment, est un indicateur qui reflète l'évolution du secteur de la construction en France. Plus précisémment, « L'indice FFB du coût de la construction est calculé à partir du prix de revient d'un immeuble de rapport de type courant à Paris. Il enregistre les variations de coût des différents éléments qui entrent dans la composition de l'ouvrage. Ce calcul ne prend pas en compte la valeur des terrains »([1]). Il donne une indication sur la santé économique du secteur, qui est étroitement liée à la croissance économique globale, aux politiques de logement, aux taux d'intérêt et à d'autres facteurs macroéconomiques.
L'indice FFB joue un rôle essentiel dans le secteur de l'assurance, en particulier en MRH. Il permet d'indexer de nombreux éléments, tels que les primes ou les franchises. Une augmentation de l'indice se traduit généralement par une augmentation du coût moyen des sinistres.

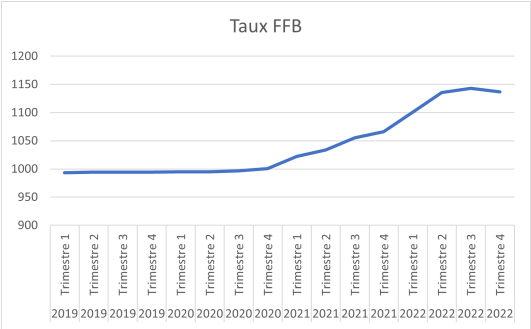


FIGURE A.1 – Evolution de l'indice FFB entre 2019 et 2022

Entre 2019 et 2020, l'indice est resté relativement stable, reflétant un marché de la construction équilibré. En 2021 nous notons une hausse significative est observée, principalement due à une pénurie généralisée de matériaux. Les augmentations des coûts des matériaux, comme la tuile, la brique, la ferraille et le bois, ont directement impacté l'indice. Cette pénurie a probablement été exacerbée par les perturbations de la chaîne d'approvisionnement liées à la pandémie de COVID-19.
En 2022, l'année a été marquée par une hausse brutale des coûts de l'énergie, en grande partie due à la guerre en Ukraine. La transformation de certains matériaux, qui nécessitent beaucoup d'énergie, a vu ses coûts augmenter, influençant davantage l'indice.

La tendance récente montre un ralentissement de la hausse de l'indice. Cela ne signifie pas que les coûts baissent, mais plutôt que leur augmentation est plus modérée ([2]).

1. [FFB, 2023]
2. [Bien, 2023]

Annexe B

Performances des modèles

B.1 Allure du Zonier IRIS

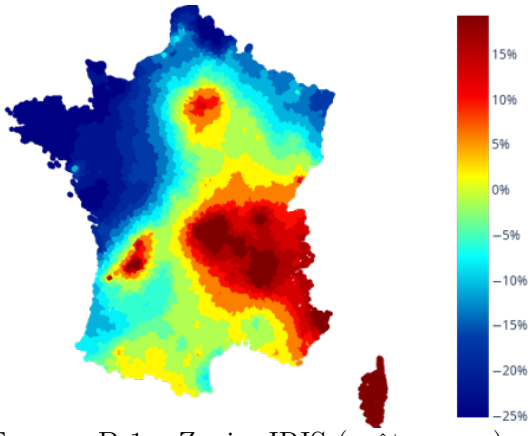


FIGURE B.1 – Zonier IRIS (coût moyen)

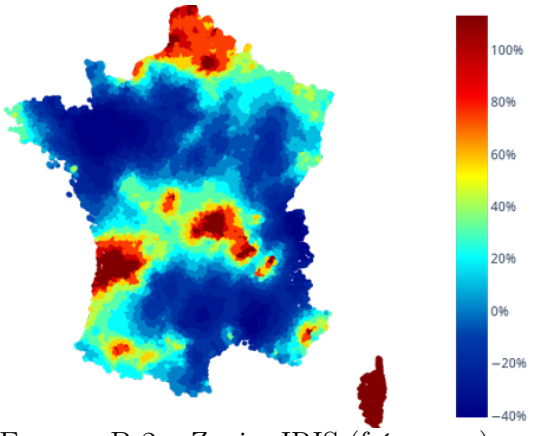


FIGURE B.2 – Zonier IRIS (fréquence)

Ce zonier permet de meilleures performances de segmentation que le zonier à la maille INSEE.

B.2 Recherche d'intéractions (GLM)

Une intéraction notable a été détectée dans le modèle coût moyen : celle entre la variable v5% et la variable relative à l'option aménagement de piscine. En effet, l'ajout de cette intéraction a eu pour conséquence une réduction de 10 unités de RMSE et une augmentation 0.7 points de Gini au modèle. Ce résultat peut s'interpréter de plusieurs façons :

— **Vulnérabilité** : Les maisons avec une piscine pourraient être plus vulnérables aux vents forts. Par exemple, les structures associées à une piscine (clôtures, équipements) pourraient être endommagées par des vents violents ;

— **Localisation** : Il est possible que les propriétés avec piscine soient plus fréquemment situées dans des zones où les vents forts sont plus courants ;

— **Effet cumulé** : L'intéraction pourrait indiquer que, bien que les vents forts et la présence d'une piscine puissent individuellement augmenter le risque de sinistre, leur effet combiné est particulièrement prononcé. Par exemple, lors d'un événement de vent fort, une propriété avec piscine pourrait subir des dommages à la fois à la piscine et à la maison, augmentant ainsi le coût total du sinistre.

Néanmoins, cette intéraction n'affichait pas de tendance claire d'un point de vue professionel, et son ajout au modèle induisait du surapprentissage. Elle n'a donc pas été rajoutée.

B.3 Données MétéoNet

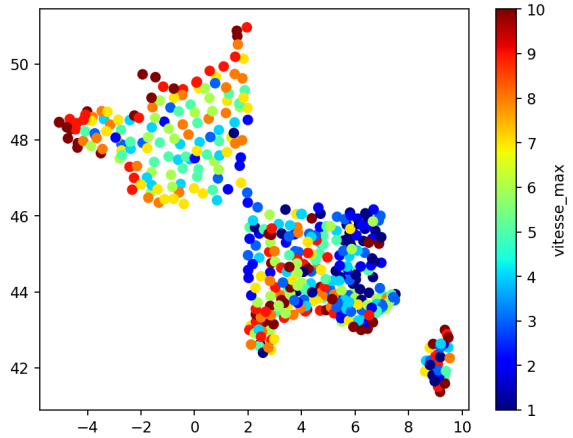


FIGURE B.3 – Répartition de la variable v5% (quantilisée)

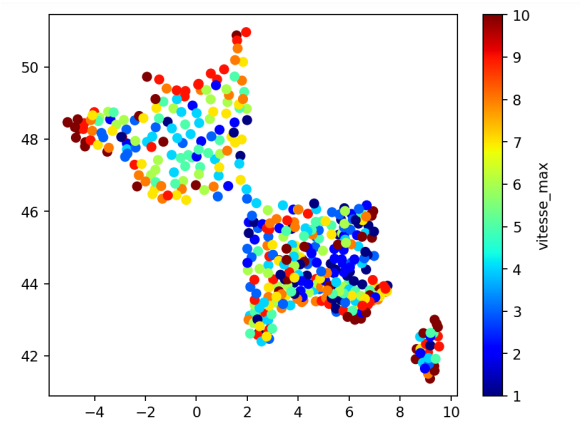


FIGURE B.4 – Répartition de la variable vmax (quantilisée)

Annexe C

Preuves

C.1 Loi binomiale négative

On définit l'espace probabilisé (Ω, A, P). Soit Y est une variable aléatoire définie dans cet univers. On note $\mathbb{E}[Y]$ l'espérance de cette variable aléatoire et $\mathrm{V}[Y]$ sa variance, que nous supposons bien définies.

Démontrons que la loi binomiale négative, utilisée pour le GLM fréquence, appartient bien à la famille exponentielle.

Soit Y une variable aléatoire définie sur l'espace probabilisé défini précédemment et suivant une loi binomiale négative. La loi binomiale négative est une distribution de probabilité discrète. Elle est définie comme le nombre d'échecs avant le r-ième succès (avec $r \in \mathbb{N}^*$) dans une séquence de Bernoulli indépendante, où la probabilité de succès est p. On a :

$$P(Y = y) = \binom{y + r - 1}{r - 1} p^r (1 - p)^y \quad \text{Où } y \geq 0 \text{ est le nombre d'échecs.}$$

En développant cette expression, nous obtenons :

$$P(Y = y) = \frac{(y + r - 1)!}{y! \times (r - 1)!} p^r (1 - p)^y$$

De plus, pour $q = 1 - p$:

$$\mathbb{E}[Y] = \frac{rq}{p}, \quad \mathbb{V}(Y) = \frac{rq}{p^2}$$

On définit les fonctions suivantes :

— $\theta = \log\left(\frac{p}{1-p}\right)$ (logarithme des cotes de p) ;

— $\phi = r$ (nombre de succès).

En utilisant ces substitutions, nous pouvons réécrire :

$$p = \frac{e^\theta}{1 + e^\theta}$$

Et pour $1 - p$:

$$1 - p = \frac{1}{1 + e^\theta}$$

En utilisant ces formules :

$$P(Y = y) = \binom{y + r - 1}{y} \left(\frac{e^\theta}{1 + e^\theta}\right)^r \left(\frac{1}{1 + e^\theta}\right)^y$$

En simplifiant, nous obtenons :

$$P(Y = y) = \binom{y + r - 1}{y} e^{r\theta} (1 + e^\theta)^{-r-y}$$

En utilisant les propriétés des logarithmes et des exponentielles, nous pouvons réécrire :

$$P(Y=y) = \binom{y+r-1}{y} \exp\Big(r\theta - (r+y)\log(1+e^{\theta})\Big)$$

Comparons cette expression à la forme canonique de la famille exponentielle (défini dans la partie 3.2.2) :

$$f_{(\theta;\phi)}(y) = \exp\left(\frac{y\theta - a(\theta)}{\phi} + c_{\phi}(y)\right) \tag{C.1}$$

où :

— θ désigne un paramètre réel, également appelé paramètre naturel ou canonique,
— ϕ symbolise le paramètre de dispersion (> 0),
— $a(\theta)$ est une fonction convexe de classe C^2,
— $c_{\phi}(y)$ est une fonction indépendante de θ.

Ainsi, nous pouvons identifier :

— $a(\theta) = r\log(1+e^{\theta})$;
— $c_{\phi}(y) = \log\left(\frac{(y+r-1)!}{y!(r-1)!}\right)$.

La forme obtenue est bien celle de la famille exponentielle, ce qui démontre que la loi binomiale négative appartient à cette famille.

C.2 Loi Inverse Gaussienne

On définit l'espace probabilisé (Ω, A, P). Soit Y est une variable aléatoire définie dans cet univers. On note $\mathbb{E}[Y]$ l'espérance de cette variable aléatoire et $\mathrm{V}[Y]$ sa variance, que nous supposons bien définies.

Soit Y une variable aléatoire définie sur l'espace probabilisé défini précédemment et suivant une loi inverse gaussienne. La loi inverse gaussienne, est définie par sa fonction de densité de probabilité :

$$f(y;\mu,\lambda) = \sqrt{\frac{\lambda}{2\pi y^3}} \exp\left(-\frac{\lambda(y-\mu)^2}{2\mu^2 y}\right) \tag{C.2}$$

où $y > 0$, $\mu > 0$ est l'espérance et $\lambda > 0$ est le paramètre de forme.

Pour exprimer cette loi dans la forme de la famille exponentielle, nous effectuons les substitutions suivantes, posons :

$$\theta = \frac{1}{\mu}$$
$$\phi = \frac{1}{\lambda}$$

Avec ces formules, la fonction de densité devient :

$$f(y;\theta,\phi) = \sqrt{\frac{\phi}{2\pi y^3}} \exp\left(-\frac{\phi(y-\frac{1}{\theta})^2}{2\theta^2 y}\right) \tag{C.3}$$

En utilisant les propriétés des exponentielles et des logarithmes, nous pouvons réécrire la fonction de densité sous la forme :

$$f(y;\theta,\phi) = \exp\left(\frac{1}{2}\log(\phi) - \frac{3}{2}\log(y) - \frac{\phi(y-\frac{1}{\theta})^2}{2\theta^2 y} - \frac{1}{2}\log(2\pi)\right) \tag{C.4}$$

En comparant cette expression à la forme canonique de la famille exponentielle, nous pouvons identifier :

$$b(\theta) = 0$$
$$c(y,\phi) = \frac{1}{2}\log(\phi) - \frac{3}{2}\log(y) - \frac{1}{2}\log(2\pi)$$

La forme obtenue est bien celle de la famille exponentielle, ce qui démontre que la loi inverse gaussienne appartient à cette famille.

Bibliographie

[EMH, 2018] (2018). Assurance multirisque habitation. economie.gouv.fr. [En ligne] Disponible sur : https://www.economie.gouv.fr/dgccrf/Publications/Vie-pratique/Fiches-pratiques/Assurance-multirisque-habitation.

[ibm, 2023] (2023). Qu'est-ce que l'algorithme des k plus proches voisins (knn) ? https://www.ibm.com/fr-fr/topics/knn. Accessed : 2023-07-30.

[ANGOUA, 2023] ANGOUA, Y. (2023). Provisionnement non-vie.

[Baaka *et al.*, 2019] Baaka, M., Koopmana, R., Snoekb, H. et Klous, S. (2019). A new correlation coefficient between categorical, ordinal and interval variables with pearson characteristics. *Nom de la revue, si connu.*

[Bien, 2023] Bien, J. M. (2023). Indice ffb. Consulté le 20 août 2023.

[BOUCHER, 2016] BOUCHER, B. L. (2016). Tarification i.a.r.d et open data.

[CHELBON, 2022] CHELBON, A. (2022). Refonte de la prime commerciale du perimetre dom en assurance mrh.

[DERKAOUI, 2021] DERKAOUI, S. (2021). Modélisation de la sinistralité grave dans le cadre des revalorisations de primes du partenariat.

[Etalab, 2016] Etalab (2016). Arrêtés de catastrophe naturelle en france métropolitaine.

[FFB, 2023] FFB (2023). Indices et index bâtiment. Consulté le [17 août 2023].

[France Assureurs, 2021a] France Assureurs (2021a). Changement climatique : quel impact sur l'assurance à l'horizon 2050 ? [En ligne] Disponible sur : https://www.franceassureurs.fr/lassurance-protege-finance-et-emploie/lassurance-protege/actualites-protege/changement-climatique-quel-impact-sur-lassurance-a-lhorizon-2050/.

[France Assureurs, 2021b] France Assureurs (2021b). Rapport annuel 2021. https://www.franceassureurs.fr/wp-content/uploads/fa-ra2021.pdf.

[France Assureurs, 2022] France Assureurs (2022). Changement climatique et transition écologique : les 5 propositions des assureurs. [En ligne] Disponible sur : https://www.franceassureurs.fr/espace-presse/les-communiques-de-presse/changement-climatique-et-transition-ecologique-les-5-propositions-des-assureurs/.

[Ministère de l'Économie, des Finances et de la Relance, 2022] Ministère de l'Économie, des Finances et de la Relance (2022). Ce qu'il faut savoir sur l'assurance habitation. [En ligne] Disponible sur : https://www.economie.gouv.fr/particuliers/assurance-habitation#:~:text=L'assurance%20habitation%20permet%20de,responsabilit%C3%A9%20civile%20%C2%AB%20vie%20priv%C3%A9e%20%C2%BB.

[Météo-France, 2019] Météo-France (2019). Meteonet. Consulté le : 10 août 2023.

[Paglia et Phélippé-Guinvarc'h, 2011] Paglia, A. et Phélippé-Guinvarc'h, M. V. (2011). Tarification des risques en assurance non-vie, une approche par modèle d'apprentissage statistique. *Bulletin français d'actuariat*, 12(Juillet-Décembre):24p. à paraitre.

[PERRIN, 2021] PERRIN, S. (2021). Refonte des elr sur la garantie dégât des eaux pour le produit habitation.

[Raillard, 2021] Raillard, N. (2021). Modélisation statistique des valeurs extrêmes. Année scolaire 2021 — 2022. Dernière mise à jour : 10/12/2021.

[SCHRYVE, 2018] SCHRYVE, L. (2018). Modélisation des sinistres graves en assurance multirisque habitation.

[scikit-learn Developers, 2023] scikit-learn Developers (2023). sklearn.ensemble.randomforestclassifier. Accessed : [17 août 2023].

[TOESCA, 2020] TOESCA, R. (2020). Zonier georev.

[VERMET, 2020] VERMET, F. (2020). Régression linéaire.

[Yann TRAONMILIN, 2018] Yann TRAONMILIN, A. R. (2018). Introduction aux statistiques bayésiennes.