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 22 Septembre 2021

Par : Laurine BABUT
Titre : Réduction du nombre de questions posées à la souscription d'un contrat d'assurance habitation par l'apport de données externes

Confidentialité : Oui - (Durée : 2 ans)

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

Membres présents du jury de l'Institut des Actuaires :
Sophie BOURDET
Romain LAILY
Bertrand DESCHAMPS
Signature :

Entreprise :
Allianz France
Signature :

Membre présent du jury de l'EURIA :
Daniel BOIVIN
Signature :

Directeur de mémoire en entreprise :
Claire LAMON
Signature :

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

Signature du responsable entreprise :

Signature du candidat :

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

L'assurance habitation (MRH) est indispensable dans la vie quotidienne, elle est notamment obligatoire en France dans certaines situations. Les enjeux pour l'assureur sont multiples : il s'agit notamment de proposer des tarifs avantageux et des offres adaptées afin de se démarquer de la concurrence, de favoriser la pérennité du portefeuille et d'accompagner l'assuré lors de la souscription et en cas de survenance d'un sinistre.

Ce mémoire s'inscrit dans une volonté d'obtenir un tarif concurrentiel, précis et personnalisé tout en allégeant le nombre de questions posées lors de la souscription de l'assuré. Ainsi, l'assureur pourra proposer à son assuré un tarif adapté à son profil de risque de façon simple et rapide. Dans cet objectif, les assureurs tendent à introduire dans leur modélisation des données externes qui sont des facteurs non demandés à la souscription (par opposition aux données internes).

Après avoir présenté le périmètre de l'étude ainsi que les nombreuses données externes récoltées dans ce mémoire, la première approche est de prédire directement des données, habituellement demandées à l'assuré lors de la souscription, grâce à des méthodes de *Machine Learning*. En prédisant des données internes à partir d'autres données internes et externes, l'objectif est de pouvoir les retirer du questionnaire de souscription.

La deuxième approche abordée est d'intégrer directement les données externes à la modélisation GLM utilisée pour la tarification chez Allianz. Ainsi, une modélisation sans donnée externe, une modélisation avec des données externes en plus des données internes et enfin une modélisation avec uniquement des données externes ont été mises en place afin non seulement de quantifier l'optimisation des modèles de tarification grâce aux données externes mais aussi de conclure sur des questions posées à la souscription qui pourraient être retirées du questionnaire

Mots clefs: Multirisque Habitation, données externes, modèles linéaires généralisés, *Machine Learning*.

Abstract

Household insurance is essential in daily life, it is mandatory in France in some situations. The challenges for the insurer are multiple : it is especially a question of proposing advantageous rates and adapt offers to stand out from the competition, to further the sustainability of the portfolio, and to accompany the customer during the subscription and in case of a claim.

This thesis is part of a desire to obtain a competitive, precise, and personalized price while reducing the number of questions asked when the insured subscribes. Thus, the insurer will be able to offer his client a price adapted to his risk profile simply and quickly. To achieve this, insurers tend to introduce external data into their modeling, which are factors that are not requested at subscription (as opposed to internal data).

After presenting the scope of the study and the numerous external data collected in this thesis, the first approach is to directly predict data, usually asked to the insured during the subscription, thanks to Machine Learning methods. By predicting internal data from other internal and external data, the goal is to be able to remove them from the subscription survey.

The second approach is to integrate external data directly into the GLM modeling used for pricing at Allianz. Thus, a model with no external data, a model with external data in addition to internal data, and finally a model with only external data were set up in order not only to quantify the optimization of the pricing models thanks to external data but also to conclude on questions asked at subscription that could be removed from the survey.

Keywords: Household, external data, generalized linear models, Machine Learning

Synthèse

Contexte

L'assurance Multirisque Habitation (MRH) est indispensable dans la vie quotidienne, elle est notamment obligatoire en France dans certaines situations. Elle permet d'assurer les bâtiments, leurs contenus ainsi que la responsabilité civile des occupants de l'habitation. Les enjeux pour l'assureur sont multiples : il s'agit notamment de proposer des tarifs avantageux et des offres adaptées afin de se démarquer de la concurrence, de favoriser la pérennité du portefeuille et d'accompagner l'assuré lors de la souscription et en cas de survenance d'un sinistre.

Un assureur doit donc mettre régulièrement à jour son tarif afin de rester compétitif tout en assurant la rentabilité de son portefeuille. Cet enjeu est d'autant plus important que le cycle du produit est inversé par rapport à d'autres produits, En effet l'assureur vend un contrat moyennant une prime avant de connaître son coût. En assurance dommages, le coût du préjudice est probable et le montant des sinistres inconnus. Ainsi, l'un des enjeux de l'assureur est de prédire le coût des sinistres afin de tarifer au mieux son produit d'assurance. Pour ce faire, l'assureur calcule la prime pure qui correspond à l'engagement de l'assureur envers l'assuré et reflète la perte attendue moyenne par l'assureur. Cette prime pure est calculée en fonction de différentes caractéristiques liées à l'assuré, au bien assuré, ainsi qu'aux garanties pour lesquelles l'assuré souscrit.

Problématique

L'objectif de l'assureur est de segmenter au mieux sa tarification afin de s'adapter aux différents profils de risques des assurés. Ainsi, l'assureur pose plusieurs questions à l'assuré lors de la souscription du contrat afin de pouvoir lui attribuer une prime reflétant au mieux son profil de risque. Ces nombreuses questions peuvent s'avérer contraignantes pour l'assuré qui peut y voir une perte de temps.

Afin de réduire ce nombre de questions posées à la souscription, les assureurs tendent à introduire des données externes, ce sont des facteurs qui ne sont pas demandés à la souscription (par opposition aux données internes). Nous pouvons, par exemple citer la zone géographique qui constitue l'un des critères de segmentation les plus couramment

employés par les assureurs. En effet, l'estimation de la sinistralité en assurance habitation est souvent influencée par l'environnement géographique, météorologique, sociodémographique, ... du risque. Ces facteurs externes, disponibles en *Open Data*, permettent d'apporter une meilleure connaissance du risque assuré et peuvent même aller jusqu'à remplacer cette donnée. L'enjeux de ce mémoire est donc d'apprécier l'apport des données externes pour tarifer un contrat MRH dans l'objectif de réduire le nombre de questions posées à la souscription.

Périmètre de l'étude

La base de données du portefeuille utilisée dans ce mémoire s'étend sur 10 ans d'historique, de 2011 à 2020. Elle recense l'ensemble des données connues par Allianz liées à l'assuré, au bien assuré, au contrat d'assurance ainsi qu'à la sinistralité. La vision des sinistres de cette base de données est au 31 décembre 2020. L'année 2020 a fait l'objet d'une étude particulière suite à la crise sanitaire du Covid-19 qui à conduit à la supprimer de l'étude.

En assurance multirisque habitation, les facteurs influençant la sinistralité varient en fonction du type d'habitation, ainsi un tarif différent sera calculé pour les maisons, chalets en bois, châteaux manoirs d'une part et les appartements et lofts d'autres part. Ce mémoire se focalisera sur le tarif relatif aux maisons, chalets en bois et châteaux manoirs.

De plus, les modélisations tarifaires sont réalisées pour chaque garantie du contrat. En assurance habitation, la principale garantie recensant la fréquence de sinistralité la plus élevée est le dégât des eaux. Ainsi, ce mémoire traite uniquement de la modélisation tarifaire sur cette garantie.

Méthodologie

Dans un premier temps, ce mémoire vise à présenter les différentes données qui vont être utilisées : des données internes à Allianz France et externes. Ce large panel de facteurs externes est collecté en *Open Data* ainsi qu'auprès de différents services internes chez Allianz. Ces données sont divisées en deux types de données :

- Des données de coûts, comme le coût de la main d'œuvre du plombier, le cours du cuivre. Ces données seront utilisées uniquement pour les modèles de coût moyen et sont intégrées via la date de survenance du sinistre.
- Des données géographiques, comme des données démographiques, socio-économiques, météorologiques. Ces informations sont intégrées aux données internes via l'adresse du logement assuré. Certaines données sont directement obtenues à partir de la latitude et de la longitude de l'adresse de l'assuré, d'autres seront récupérées à d'autres mailles telles que le quartier (l'IRIS), la commune (l'INSEE) ou encore le département.

Le retraitement de ces facteurs externes peut s'avérer complexe et technique. Tout d'abord, l'ajout de données géographiques passe par le géocodage du portefeuille d'assurés. Ensuite certaines données récoltées sont le fruit de travaux développés en interne telle que l'exploitation de données via la reconnaissance d'images *Google*. D'autre part, certaines données collectées ne peuvent pas directement être rattachées aux adresses du portefeuille comme par exemple, le retraitement des données météorologiques. Ces données sont collectées de manière journalière et par station météorologique. Il convient donc de les agréger de manière temporelle et de les interpoler spatialement (via une méthode de lissage géospatiale) afin d'avoir un indicateur météorologique pour l'ensemble du territoire français.

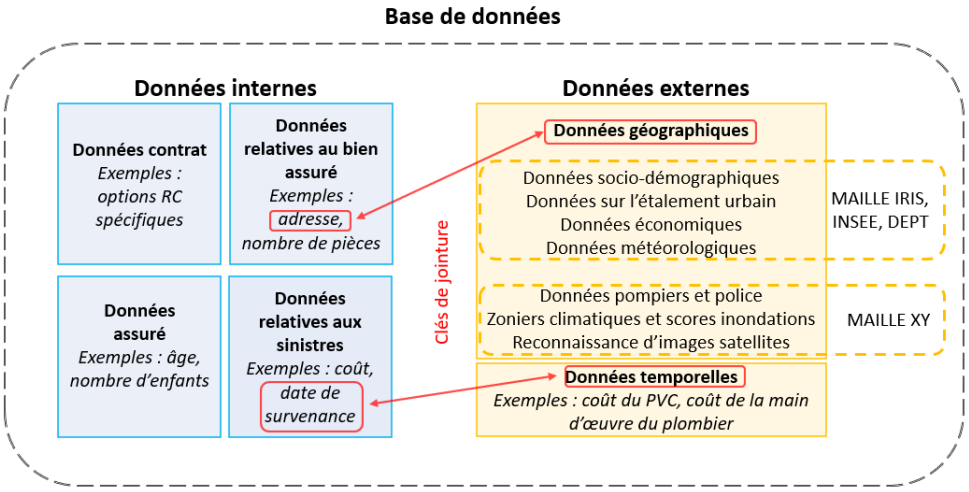

FIGURE 1 – Présentation de l'ensemble de la base de données

L'objectif du mémoire est d'alléger le questionnaire de souscription par l'apport de données externes. Pour ce faire, deux approches vont être utilisées :

- La prédiction de questions posées à la souscription via des méthodes de *Machine Learning*,
- L'ajout de facteurs externes dans les modèles de fréquence et coût moyen afin d'optimiser les modèles tarifaires et remplacer (partiellement ou complètement) les données internes classiquement utilisées par les données externes.

La première approche du mémoire est d'utiliser des algorithmes de prédictions statistiques afin de prédire des données latentes correspondant à des données habituellement demandées à l'assuré lors de la souscription du contrat. Par définition une variable latente ne peut pas être mesurée directement et est non observable dans notre échantillon. Elle est issue d'algorithmes ou de modélisations statistiques et synthétise plusieurs variables observées. Ainsi, ces données latentes vont être prédites à partir de données internes et externes sur la base de données de 2011 à 2017. Deux méthodes ont été testées :

— L'algorithme *Random Forest*, qui est une méthode combinant plusieurs arbres de décisions avec une approche de *Boostrap Aggregation* aussi appelée *Bagging*. L'ensemble des échantillons d'observations pour chaque arbre est construit avec une méthode de *Bootstrap*, qui peut s'assimiler à un tirage avec remise dans la base de données totale. L'algorithme des forêts aléatoires va entraîner plusieurs arbres de décisions et faire la moyenne des prédictions de l'ensemble des arbres dans le cas d'une régression et prendre la classe majoritairement prédite dans le cas d'une classification. Ainsi l'avantage de cette méthode est de réduire la variance de prédiction car chaque arbre est indépendant de l'autre.
— L'algorithme du *Gradient Boosting* (GBM) repose sur des méthodes de *Boosting* d'arbres. A la différence du *Bagging*, le *Boosting* construit ses arbres de décision par récurrence. Chaque arbre créé prend en compte la prédiction de l'arbre précédent en pénalisant les observations dont la prédiction est fausse. L'idée générale du *Gradient Boosting* est de construire des arbres successifs avec comme objectif d'améliorer l'erreur initiale en se basant sur la descente du gradient de la fonction de perte pour améliorer les erreurs. Cette méthode permet de diminuer à la fois le biais et la variance mais elle présente un risque de surapprentissage.

L'objectif de cette première approche est de comparer ces méthodes de *Machine Learning* et de retenir celle qui permettra de prédire au mieux nos données afin de pouvoir les substituer aux données demandées lors de la souscription du contrat. Ces méthodes ont été testées sur 5 variables internes : la profession (CSP), la nature du logement (résidence principale/secondaire), l'âge, le nombre de pièces et les capitaux assurés. Parmi ces variables, deux ont été retenues (les plus performantes) et testées ensuite sur les modèles GLM qui sont réalisés sur la base de données 2018 et 2019.

La deuxième approche consiste à intégrer directement des données externes à la modélisation GLM. Les GLM ont été utilisés pour la modélisation car Allianz utilise le logiciel Emblem qui facilite leur mise en place. De plus les GLM sont les modèles les plus répandus et utilisés en tarification non-vie. Trois modélisations sont faites pour la fréquence et le coût moyen de la garantie dégâts des eaux sur les données de 2018 et 2019, une modélisation sans donnée externe, une modélisation avec l'apport de données externes en plus des données internes et enfin une modélisation avec uniquement des données externes. L'objectif est non seulement de quantifier l'optimisation des modèles de tarification grâce aux données externes mais aussi de conclure sur des questions posées à la souscription qui pourraient être retirées du questionnaire.

Résultats

La première approche présentée nous a permis de retenir deux variables bien prédites qui sont l'âge et la nature. La méthode retenue pour leurs prédictions est la méthode du *Gradient Boosting* qui s'est montrée plus performante que la méthode *Random Forest*. La méthode du *Gradient Boosting* a été retenue en comparant six indicateurs de

performance pour la nature : l'AUC (*Area Under the Curve*), l'AUC PR (*Area Under the Curve Precision Recall*), la sensibilité, la précision, la spécificité et l'exactitude. Pour l'âge, la RMSE à servi d'indicateur de performance dans un premier temps, puis les âges prédits et observés ont été divisés en tranches correspondant à la segmentation du tarif commercial actuel d'Allianz sur lequel le F1 score, le F1 score pondéré, l'exactitude ainsi que l'exactitude à une classe près (les classes sont ordonnées) ont été calculés.

	AUC	AUC PR	Sensibilité	Précision	Spécificité	Exactitude
Random Forest	0,96	0,83	0,6332	0,8426	0,9826	0,9318
Gradient Boosting	0,97	0,88	0,7113	0,8702	0,9844	0,9494

TABLE 1 – Performance des modèles de *Machine Learning* pour la prédiction de la variable nature

	RMSE	F1 score	F1 score pondéré	Exactitude	Exactitude à une tranche près
Random Forest	2,58	0,6447	0,6786	0,6773	0,8990
Gradient Boosting	2,42	0,6738	0,7042	0,6987	0,9009

TABLE 2 – Performances des modèles de *Machine Learning* pour la prédiction de la variable âge

Les variables latentes ont ensuite été prédites sur la base 2018 et 2019 et testées dans le modèle de données internes. La déviance des modèles en présence de ces variables latentes comparée à celle des modèles avec les données récupérées à la souscription montre que les modèles sont tout aussi performants lors de la substitution par les variables latentes.

	Modèle de sévérité	**Modèle de fréquence**
Âge et nature récupérés à la souscription	-8,38	-4,91
Âge et nature prédits par *Machine Learning*	-8,13	-4,90

TABLE 3 – Écart de déviance (%) entre les modèles de données internes et les modèles sans variable sur la base de test

Les modélisations GLM ont montré que les modèles avec uniquement des données externes sont bien moins performants que les modèles avec des données internes. L'indicateur de performance utilisé pour comparer les modèles est la déviance qui a été calculée sur la base de test. Cependant les modèles mixtes combinant à la fois des données internes et externes se révèlent aussi performants que les modèles de données internes.

	Modèle de sévérité	**Modèle de fréquence**
Écart de déviance (%) entre le modèle de données internes et le modèle sans variable	- 8,38	-4,91
Écart de déviance (%) entre le modèle de données externes et le modèle sans variable	-2,85	-2,64
Écart de déviance (%) entre le modèle de données internes et externes avec le modèle sans variable	-8,29	-5,13

TABLE 4 – Analyse des déviances sur la base de test des différents modèles

Ainsi, en comparant ces deux modélisations, on observe cinq variables présentes dans les modèles de données internes qui ont été substituées par des données externes dans les modèles de données mixtes (internes et externes) ; la CSP de l'assuré, l'ancienneté du bâtiment, la nature du logement, la présence d'alarme au sein du logement ainsi que l'état civil de l'assuré. La catégorie socio-professionnelle par exemple est expliquée par les variables taux d'ouvriers de la classe supérieure et taux de chômeurs par zone géographique pour le modèle de fréquence et le taux de personnes de professions intermédiaires dans le modèle de coût moyen. De même l'état civil se voit substitué par le taux de personnes divorcées. La présence d'une alarme peut quant à elle s'expliquer grâce à la variable renseignant sur la distance minimale au commissariat de police ou la gendarmerie le plus proche ainsi que l'aire urbaine. La variable nature s'explique dans les deux modèles par le taux de résidences principales par zone géographique. Enfin la variable ancienneté du bâtiment est expliquée par la variable taux de bâtiments construits avant 1990 par zone géographique pour les modèles de fréquence et de coût moyen.

Conclusion

A travers ce mémoire, deux approches ont été établies avec le même objectif de réduire le nombre de questions posées à la souscription en apportant de la donnée externe. Les résultats obtenus pour ces deux méthodes sont satisfaisants, en effet chacune des méthodes nous permet de pouvoir retirer des variables internes grâce à l'apport de données externes tout en conservant la performance des modèles de tarification. De plus cette étude a permis de montrer que malgré un grand nombre de données externes variées collectées, ces dernières ne suffisent pas pour calculer un tarif sans question.

La première méthode nous permet de pouvoir retirer deux variables du questionnaire de souscription puisque cette méthode s'applique à l'ensemble du produit MRH d'Allianz pour les maisons, chalets en bois et châteaux manoirs. La deuxième méthode quant à elle nous permet de pouvoir retirer cinq variables internes des modèles pour la garantie dégâts des eaux. La prochaine étape serait de réaliser la deuxième méthode sur les autres garanties du produit habitation d'Allianz France afin de vérifier que les données externes permettent de remplacer les mêmes données internes tout en conservant la performance des modèles. Ainsi, les questions posées à la souscription pour récupérer les variables internes pourraient être retirées du questionnaire.

Synthesis

Context

The household insurance essential in daily life, it is mandatory in France in some situations. It allows to insure the buildings, their contents as well as the civil liability of the occupants of the dwelling. The challenges for the insurer are multiple : it is especially a question of proposing advantageous rates and adapted offers in order to stand out from the competition, to further the sustainability of the portfolio and to accompany the customer during the subscription and in case of a claim.

An insurer must therefore regularly update its rates in order to remain competitive while ensuring the profitability of its portfolio. This issue is all the more important in insurance because the product cycle is reversed compared to other products. Indeed, the insurer sells a contract for a premium before knowing its cost. In damage insurance, the cost of the harm is likely and the amount of the claims is unknown. Thus, one of the insurer's challenges is to predict the cost of claims in order to price his insurance product as well as possible. To achieve this, the insurer calculates the pure premium, which corresponds to the insurer's commitment to the insured and reflects the insurer's average expected loss. This pure premium is calculated according to various characteristics related to the insured, the insured property, as well as the guarantees for which the insured subscribes.

Problematic

The insurer's goal is to segment at best its pricing in order to adapt to the different risk profiles of the insured. Thus, the insurer asks several questions to the insured during the subscription of the contract in order to be able to assign a premium that best reflects his risk profile. These numerous questions can be restrictive for the insured, who may see them as a waste of time.

In order to reduce the number of questions asked at the subscription, insurers tend to introduce external data, which are factors that are not asked at subscriptions (as opposed to internal data). We can, for example, quote the geographical area which is one of the segmentation criteria most commonly used by insurers. Indeed, the estimation of

the claims in household insurance is often influenced by the geographical, meteorological, socio-demographic environment, ... of the risk. These external factors, available in open data, allow to bring a better knowledge of the insured risk and can even replace this data. The challenge of this thesis is to appreciate the contribution of external data to price an household contract with the aim of reducing the number of questions asked at the time of subscriptions.

Scope of the study

The portfolio database used in this thesis covers 10 years of historical, from 2011 to 2020. It contains all the data known by Allianz related to the insured, the insured property, the insurance contract and the claims experience. The vision of the claims in this database is as of December 31, 2020. The year 2020 was the subject of a special study following the Covid-19 health crisis, which led to the deletion of the study.

In multi-risk household insurance, the factors influencing claims vary according to the type of dwelling, so a different rate will be calculated for houses, wooden cottages and manor houses on the one hand and apartments and loft apartments on the other. This thesis will focus on the rate for houses, wooden chalets and manor houses.

In addition, pricing modeling is performed for each warranty of the contract. In household insurance, the main warranty, with the highest claims frequency is water damage. Thus, this thesis deals only with the pricing modeling for this warranty.

Methodology

In a first place, this thesis aims at presenting the different data that will be used : internal data at Allianz France and external data. This large panel of external factors is collected in Open Data as well as from different internal services. This data is divided into two types of data :

- Cost data, such as plumber's labor cost, copper price. These data will only be used for average cost models and are integrated via the date of the claim.
- Geographic data, such as demographic, socio-economic, weather data. These information are integrated with internal data via the address of the insured dwelling. Some data are obtained directly from the latitude and longitude of the insured's address, while other data will be retrieved from other grids such as the neighborhood (IRIS), the commune (INSEE) or the department.

The reprocessing of these external factors can be complex and technical. First, the addition of geographic data requires geocoding of the portfolio. Secondly, some of the data collected is the result of work developed internally, such as the use of data from Google image recognition. On the other hand, some of the data collected cannot be directly linked

to the addresses in the portfolio. For example, the reprocessing of meteorological data. This data is collected on a daily basis and by station. It is therefore meteorological to aggregate them in a temporal manner and to interpolate them spatially (via a geospatial smoothing method) in order to have a meteorological indicator for the entire French territory.

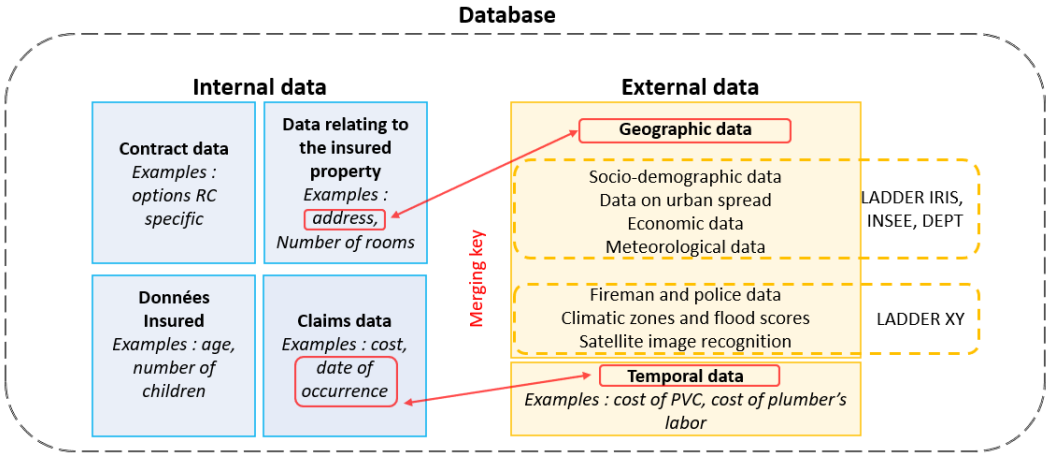

FIGURE 2 – Presentation of the whole database

The objective of the thesis is to lighten the subscription questionnaire by providing external data. To do so, two approaches will be used :

- Predicting asked questions at the subscription using Machine Learning methods,
- Adding external factors to the frequency and averaged cost models in order to optimize the pricing models and replace (partially or completely) the internal data traditionally used by external data.

The first approach of the thesis is to use statistical prediction algorithms to predict latent data corresponding to data usually requested from the insured at the time of the subscription of a contract. By definition, a latent variable cannot be measured directly and is not observable in our sample. It is derived from algorithms or statistical models and synthesizes several observed variables. Thus, these latent data will be predicted from internal and external data on the database from 2011 to 2017. Two methods were tested :

— The Random Forest algorithm, which is a method combining several decision trees with a Boostrap Aggregation approach also called Bagging. The sample of observations for each tree is constructed with a Bootstrap method, which can be likened to a random draw from the total database. The random forest algorithm will lead to several decision trees and make an average of the predictions of all the trees in the case of a regression and take the most predicted class in the case of a classification. Thus, the advantage of this method is to reduce the prediction variance because each tree is independent of the other.

— The Gradient Boosting (GBM) algorithm is based on tree boosting methods. Unlike Bagging, Boosting builds its decision trees by recurrence. Each tree created takes into account the prediction of the previous tree by penalizing the observations whose prediction is false. The general idea of Gradient Boosting is to build successive trees with the objective of improving the initial error based on the descent of the gradient of the function of loss to improve the errors. This method allows to decrease both the bias and the variance but it presents a risk of overlearning.

The objective of this first approach is to compare these Machine Learning methods and to select the one that will best predict our data in order to be able to substitute them for the data requested at the time of the contract subscription. These methods were tested on 5 internal variables : profession (CSP), type of housing (primary/secondary residence), age, number of rooms and insured capital. Among these variables, two were selected (the best performing ones) and then tested on the GLM models that will be carried out on the 2018 and 2019 data base.

The second approach is to integrate external data directly into the GLM modeling. GLMs were used for the modeling because Allianz uses Emblem software which facilitates their implementation. Moreover, GLMs are the most widespread models used in non-life pricing. Three modeling will be used for the frequency and average cost of water damage coverage on 2018 and 2019 data : a model with no external data, a model with external data in addition to internal data and finally a model with only external data. The objective is not only to quantify the optimization of pricing models thanks to external data but also to conclude on subscription questions that could be removed from the survey.

Résultats

The first approach presented allowed us to retain two well predicted variables which are age and nature. The method chosen for their prediction is the Gradient Boosting method, which proved to be more efficient than the Random Forest method. The Gradient Boosting method was selected by comparing six performance indicators for nature : AUC (Area Under the Curve), AUC PR (Area Under the Curve Precision Recall), sensitivity, precision, specificity and accuracy. For age, the RMSE was first used as a performance indicator, and then the predicted and observed ages were divided into slices corresponding to the segmentation of the current Allianz commercial pricing on which the F1 score, the weighted F1 score, the accuracy as well as the accuracy to within one class (the classes are ordered) were calculated.

	AUC	AUC PR	Sensitivity	Precision	Specificité	Accuracy
Random Forest	0,96	0,83	0,6332	0,8426	0,9826	0,9318
Gradient Boosting	0,97	0,88	0,7113	0,8702	0,9844	0,9494

TABLE 5 – Performance of the Machine Learning models for the prediction of the nature variable

	RMSE	F1 score	F1 weighted score	Accuracy	Accuracy to the nearest slice
Random Forest	2,58	0,6447	0,6786	0,6773	0,8990
Gradient Boosting	2,42	0,6738	0,7042	0,6987	0,9009

TABLE 6 – Performance of the Machine Learning models for the prediction of the age variable

The latent variables were then predicted on the 2018 and 2019 basis and tested in the internal data model. The deviance of the models in the presence of these latent variables compared to the models with the data retrieved at subscription shows that the models perform equally well during the substitution by with the latent variables.

	Severity model	**Frequency model**
Age and nature recovered at subscription	-8,38	-4,91
Age and nature predicted by Machine Learning	-8,13	-4,90

TABLE 7 – Deviance gap (%) between models with internal data and models without variable on the test base

GLM modeling has shown that models with only external data perform much less than models with internal data. The performance indicator used to compare the models is the deviance that was calculated on the test basis. However, the mixed models combining both internal and external data perform as well as the internal data models.

	Modèle de sévérité	**Modèle de fréquence**
Deviance gap (%)between the model with internal data and the model without variable	- 8,38	-4,91
Deviance gap (%) between the model with external data and the model without variable	-2,85	-2,64
Deviance gap (%) between the mode with internal and external data model with the model without variable	-8,29	-5,13

TABLE 8 – Deviance analyses on the test base for the different models

Thus, when comparing these two modelling, we observe five variables present in the internal data models that have been substituted by external data in the mixed data models (internal and external) ; the insured's CSP, the age of the building, the nature of the dwelling, the presence of alarms in the dwelling and the insured's civil status. For example, the socio-occupational category, is explained by variables of upper class workers and unemployed people by geographical area in the frequency model and the rate of people in intermediate occupations in the average cost model. Likewise, marital status is substituted by the rate of divorced persons. The presence of an alarm can be explained by the variable indicating the minimum distance to the nearest police station or gendarmerie, as well as the urban area. The nature variable is explained in both models by the rate of primary residences by geographical area. Finally, the building age variable is explained by the variable of buildings built before 1990 by geographical area for the frequency and average cost models.

Conclusion

Through this thesis, two approaches have been established with the same objective of reducing the number of questions asked at subscription by providing external data. The results obtained for these two methods are satisfactory, in fact each method allows us to remove internal variables thanks to the contribution of external data while maintaining the performance of the pricing models. Moreover, this study has shown that, despite the large amount of external data collected, it is not sufficient to calculate a tariff without questions.

The first method allows us to remove two variables from the subscription survey since this method applies to the whole Allianz household product for houses, wooden chalets and manor houses. The second method allows us to remove five internal variables from the models for water damage coverage. The next step would be to carry out the second method on the other guarantees of the Allianz France household insurance product in order to verify that the external data can replace the same internal data while main-

taining the performance of the models. In this way, the questions asked at the time of subscriptions to recover the internal variables could be removed from the survey.

Remerciements

Tout d'abord, je tiens à remercier Judie DARMON, responsable de l'équipe *Pricing* MRH d'Allianz France, qui m'a permis de réaliser cette année de Master 2 en alternance au sein de son équipe et de m'avoir fait confiance pour la réalisation de l'étude présentée dans ce mémoire.

Un grand merci à ma tutrice entreprise Claire LAMON pour sa disponibilité, son expertise et ses conseils, ainsi que l'encadrement dont j'ai bénéficié durant mon année d'alternance.

Je remercie également l'équipe pédagogique de l'EURIA et particulièrement Franck VERMET pour son encadrement et sa disponibilité, ainsi que ma tutrice académique Laurence ABIVEN pour sa réactivité et ses précieux conseils.

Je tiens à adresser ma reconnaissance à toute l'équipe *Pricing* d'Allianz France avec qui j'ai pu échanger tout au long de mon étude.

Enfin je souhaite remercier l'ensemble de ma famille pour leur soutien et leurs encouragements tout au long de mon parcours académique ainsi que durant la rédaction de ce mémoire.

Table des matières

Résumé **i**

Abstract **iii**

Synthèse **v**

Synthesis **xiii**

Introduction **1**

1 Présentation générale **3**
1.1 L'assurance en France . . . 3
1.1.1 L'assurance de Biens et Responsabilités . . . 3
1.1.2 L'assurance Multirisque Habitation . . . 4
1.2 Présentation d'Allianz France . . . 5
1.3 Le produit MRH d'Allianz France . . . 7
1.4 Le processus d'indemnisation . . . 9
1.4.1 La garantie DDE . . . 9
1.4.2 Les conventions CIDRE et IRSI . . . 9
1.5 Le principe de tarification . . . 10
1.5.1 La prime pure et la prime commerciale . . . 10
1.5.2 La segmentation client . . . 12
1.5.3 L'apport des données externes . . . 13
1.6 Méthodologie du mémoire . . . 15

2 Constitution de la base de données **17**
2.1 Périmètre de l'étude . . . 17
2.2 Présentation des données en portefeuille . . . 19
2.2.1 Constitution de la base de données internes . . . 19
2.2.2 Analyse du covid . . . 21
2.2.3 Statistiques descriptives . . . 26
2.3 Présentation des données externes . . . 29
2.3.1 Données géographiques . . . 29
2.3.1.1 Géocodage du portefeuille . . . 30

2.3.1.2 Données socio-démographiques . . . 30
2.3.1.3 Données relatives à l'étalement urbain . . . 31
2.3.1.4 Données économiques . . . 33
2.3.1.5 Données relatives aux pompiers et à la police/gendarmerie 33
2.3.1.6 Données météorologiques . . . 35
2.3.1.7 Zoniers climatiques et scores inondations . . . 37
2.3.1.8 Reconnaissance d'images satellites . . . 39
2.3.2 Données externes temporelles . . . 40
2.4 Base de données finale . . . 41

3 Création de données latentes 43
3.1 Théorie . . . 45
3.1.1 Méthode de *Machine Learning* . . . 45
3.1.1.1 Algorithme SMOTE . . . 45
3.1.1.2 Arbres CART . . . 46
3.1.1.3 *Random Forest* . . . 47
3.1.1.4 *Gradient Boosting* . . . 48
3.1.2 Paramétrages des modèles . . . 49
3.1.3 Évaluation des modèles . . . 51
3.1.3.1 Régression . . . 51
3.1.3.2 Classification binaire . . . 52
3.1.3.3 Classification multiclasses . . . 53
3.2 Modélisation de la nature du logement . . . 55
3.2.1 Comparaison et validation des modèles de *Machine Learning* . . . 56
3.2.2 Interprétation des résultats . . . 59
3.3 Modélisation de l'âge de l'assuré . . . 60
3.3.1 Comparaison et validation des modèles de *Machine Learning* . . . 60
3.3.2 Interprétation des résultats . . . 64

4 Modélisation GLM 65
4.1 Théorie des GLM . . . 66
4.1.1 Composantes d'un GLM . . . 66
4.1.2 Ajustement d'un GLM . . . 67
4.1.3 Qualité d'ajustement d'un GLM . . . 68
4.1.3.1 Fonction *Mean Excess* . . . 68
4.1.3.2 Test de Kolmogorov-Smirnov . . . 69
4.1.3.3 Test du khi-deux . . . 70
4.1.3.4 Analyse des corrélations . . . 70
4.1.3.5 Méthode *Forward* . . . 71
4.1.3.6 Test de Wald . . . 71
4.1.4 Validation du modèle . . . 72
4.1.4.1 Calcul de la déviance . . . 72
4.1.4.2 Calculs des résidus . . . 72
4.2 Mise en place des modèles GLM . . . 73

4.2.1 Étude des sinistres graves 74
4.2.2 Adéquation des distributions de fréquence et coût moyen à des lois 75
4.2.3 Étude des corrélations 77
4.2.4 Résultats et qualité d'ajustement du GLM 79
4.3 Validation du GLM 86
4.3.1 Analyse des résidus 86
4.3.2 Analyse de la déviance sur la base de test 88
4.4 Interprétation des résultats 90
4.4.1 Modélisation avec des données internes 90
4.4.2 Modélisation avec des données externes 93
4.4.3 Modélisation avec des données internes et externes 96
4.5 Réduction du nombre de questions à la souscription 98
4.5.1 Étude des variables latentes dans les modèles mixtes 98
4.5.2 Sélection de données externes pour remplacer les données internes . 102

Conclusion **103**

A Démonstrations **105**

B Étude des résidus des modèles de fréquence **107**

Liste des abréviations **109**

Bibliographie **113**

Introduction

Initiée pour assurer les échanges maritimes de marchandises, l'assurance dommages est aujourd'hui l'assurance la plus développée dans le monde, elle permet d'assurer tout ce qui ne relève pas de la vie humaine.

En assurance dommages, l'assureur doit mettre régulièrement à jour son tarif afin de rester compétitif tout en assurant la rentabilité de son portefeuille. Cet enjeu est d'autant puis important que le cycle de produit est inversé. En effet l'assureur vend un contrat moyennant une prime avant de connaître son coût. En assurance dommages, le coût du préjudice est probable et le montant des sinistres inconnus. Ainsi, l'un des enjeux de l'assureur est de prédire le coût des sinistres afin de tarifer au mieux son produit d'assurance. Pour ce faire, l'assureur calcule la prime pure correspondant à l'engagement de l'assureur envers l'assuré. Elle reflète la perte moyenne attendue par l'assureur. Cette prime pure est calculée en fonction de différentes caractéristiques liées à l'assuré, au bien assuré, ainsi qu'aux garanties pour lesquelles l'assuré souscrit. Ces données récoltées auprès de l'assuré sont appelées données internes.

Le fondement premier de l'assurance est de mutualiser les risques. Le principe de mutualisation stipule que chaque assuré qui souscrit un même contrat d'assurance est exposé de manière indépendante mais identique aux risques pris en compte dans son contrat. Ainsi, si un assureur mutualise complètement son portefeuille, chaque assuré payera la même prime quel que soit son profil de risque. Cependant les assureurs tendent aujourd'hui à segmenter de plus en plus leurs portefeuilles, c'est-à-dire qu'ils vont créer des segments correspondants à des profils de risques et chaque assuré d'un même segment payera une prime identique. L'un des enjeux des assureurs dommages est de proposer à leurs assurés un tarif adapté et rapide. L'objectif est donc de calculer un tarif le plus segmenté possible qui reflétera au mieux le profil de risque de l'assuré tout en lui posant le minium de questions afin de réduire le temps de souscription. Pour ce faire, les assureurs tendent à introduire des données externes dans leurs modélisations, ce sont des facteurs qui ne sont pas demandés à la souscription (par opposition aux données internes).

C'est dans ce contexte que ce mémoire a été réalisé, ainsi la problématique soulevée est la suivante : **Les données externes peuvent-elles remplacer des données internes tout en conservant la même performance de modélisation ?**

Afin de répondre à cette problématique, il convient de collecter et retraiter un large panel de données externes, grâce à des techniques de *Data Engineering*. Deux approches seront abordées dans ce mémoire. La première est d'utiliser des modèles de *Machine Learning* pour prédire directement des données qui résultent habituellement de questions posées à l'assuré lors de la souscription de son contrat ; puis de tester ces données prédites sur les modèles de tarification. La deuxième est d'intégrer directement les données externes à la modélisation tarifaire afin de savoir si un modèle composé uniquement de données externes est aussi, voir plus, performant qu'un modèle avec des données internes. Le cas échéant, de démontrer que certaines données internes peuvent être substituées par des données externes sans nuire à la performance des modèles tarifaires.

Chapitre 1

Présentation générale

1.1 L'assurance en France

En France, l'assurance se divise en deux grandes branches : l'assurance de personnes et l'assurance dommages. L'assurance dommages a été initiée dans l'objectif d'assurer les échanges maritimes de marchandises. Aujourd'hui, l'assurance dommages permet d'assurer tout ce qui ne relève pas de la vie humaine. Au contraire, l'assurance de personnes est relative à tout ce qui touche la personne assurée. Elle permet d'assurer une personne pour : la vie, le décès, les accidents corporels, la maladie, l'incapacité ou encore l'invalidité. Ce mémoire porte sur la branche d'assurance non-vie et en particulier sur l'assurance dommages.

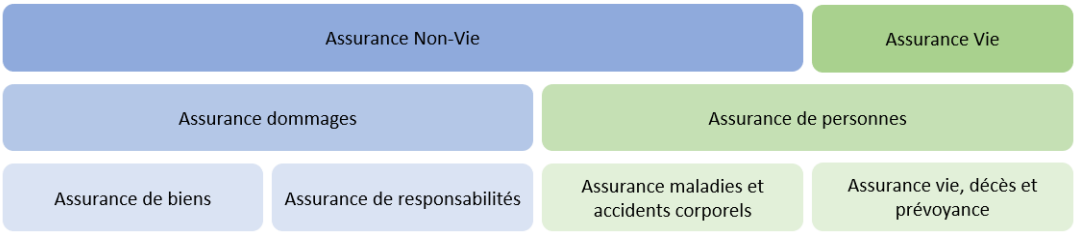

FIGURE 1.1 – L'assurance en France

1.1.1 L'assurance de Biens et Responsabilités

L'assurance de biens et responsabilités aussi appelée assurance dommages est l'assurance la plus développée dans le monde mais elle ne représente que 25,7 % des primes d'assurances et 22,9 % des prestations en France en 2019. En assurance dommages, le coût du préjudice est probable et le montant des sinistres inconnus.

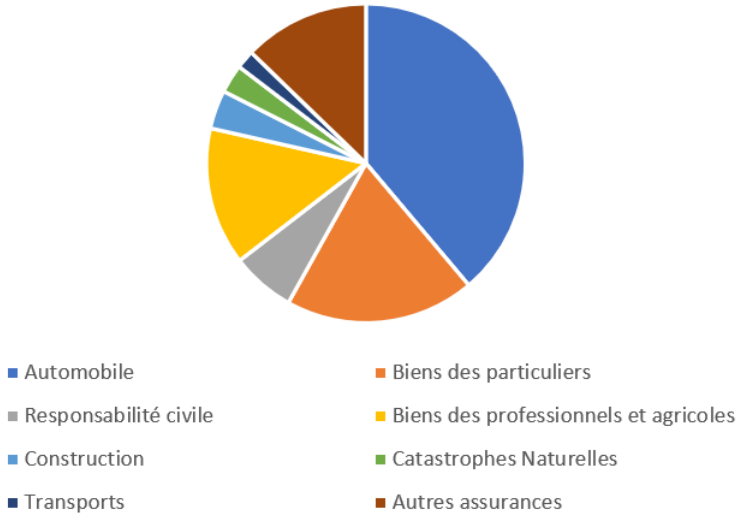

FIGURE 1.2 – Divisons de l'assurance de Biens et Responsabilités

Bien que l'assurance automobile domine le marché de l'assurance dommages avec 55.6 % des primes, l'assurance de biens des particuliers (inclus dans l'assurance Multirisque Habitation) représente 30.5 %, soit presque un tiers de l'assurance dommages en terme de primes d'assurance. Devant une telle part de marché, il devient primordial en tant qu'assureur de tarifer ses contrats de manière optimale afin d'être rentable et compétitif, mais également de s'assurer d'un excellent suivi client pour se démarquer sur le marché. C'est l'enjeu de ce mémoire qui va se focaliser sur l'assurance de biens des particuliers et plus spécifiquement l'assurance Multirisque Habitation.

1.1.2 L'assurance Multirisque Habitation

Au sein de l'assurance de biens des particuliers, l'assurance MRH [1] fait partie des assurances indispensables de la vie quotidienne. Le contrat MRH couvre les risques auxquels sont exposés votre logement et les personnes vivant sous votre toit. L'assurance MRH possède donc des garanties dommages aux biens et responsabilités civiles des occupants de l'habitation, effective même lorsque les habitants ne sont pas à leur domicile.

L'assurance MRH est obligatoire pour les locataires, on l'appelle assurance « risques locatifs » mais non obligatoire pour les propriétaires sauf lorsque le logement est situé dans une copropriété. Lorsque le logement est loué, dans un immeuble en copropriété, le locataire doit souscrire une assurance « risques locatifs » et le propriétaire une assurance « MRH non occupant ». Ainsi, le caractère obligatoire de l'assurance MRH constitue un enjeu important pour les assureurs puisque le nombre d'assurés sera important et par conséquent une tarification segmentée et adaptée à chaque assuré est indispensable

1. Multirisque Habitation

afin d'obtenir un tarif attractif tout en obtenant un portefeuille rentable. Comme évoqué précédemment, le mémoire porte sur l'assurance MRH.

Les principales garanties de l'assurance MRH en France sont :
- Incendie : permet de couvrir les dégâts causés par le feu et la fumée,
- Tempête, grêle, neige (TGN) : permet de couvrir les dommages causés par ces évènements climatiques,
- Vol : permet de couvrir les effractions et vols ayant eu lieu dans le logement assuré,
- Dégâts des eaux : permet de couvrir les dégâts causés par l'eau,
- Responsabilité civile : permet de couvrir les dommages corporels, matériels et immatériels causés à autrui par les personnes assurés,
- Bris de glaces : permet de couvrir les dommages subis par les parties vitrées du logement.
- Catastrophes naturelles (CATNAT) : permet de couvrir les dégâts causés par une catastrophe naturelle

	Fréquence		Coût Moyen	
	2018	2019	2018	2019
Incendie	0.55%	0.51%	7419€	8272€
Tempêtes, grêles, neige	1.01%	1.14%	2193€	2123€
Vol	0.90%	0.90%	1843€	1893€
Dégâts des eaux	3.57%	3.33%	1135€	1082€
Responsabilité Civile	1.04%	1.01%	1099€	1103€
Bris de glace	0.71%	0.72%	428€	449€

TABLE 1.1 – Moyenne de la fréquence et du coût moyen des sinistres pour les principales garanties MRH d'après les chiffres de la FFA

D'après les chiffres de la FFA [2], on peut voir qu'en termes de fréquence de sinistralité pour l'assurance MRH, la garantie dégâts des eaux enregistre la fréquence de sinistralité la plus élevée : 3,33 % en 2019. Toutefois, le coût moyen de la garantie dégâts des eaux n'est pas le plus élevé. L'importante fréquence de sinistralité pour la garantie des dégâts des eaux se retrouve aussi dans le portefeuille MRH d'Allianz France et ce mémoire va se focaliser sur cette garantie.

1.2 Présentation d'Allianz France

Encouragé par la révolution industrielle, le Groupe Allianz a été fondé en Allemagne, à Munich, en 1890 dans l'objectif d'assurer le transport et les accidents dus aux activités industrielles. Le Groupe Allianz est aujourd'hui le *leader* [3] en assurance de biens et responsabilités. Implanté dans plus de 70 pays, le Groupe Allianz assure plus de 86 millions

2. Fédération Française de l'Assurance
3. Le numéro 1 mondial selon le classement Interbrand des meilleures marques 2019

de clients à travers le monde et représente un chiffre d'affaires de 130,6 milliards d'euros en 2018.

FIGURE 1.3 – Zones d'implantation du Groupe Allianz

En 2019, Allianz France qui est une entité du Groupe Allianz, compte plus de treize milliards d'euros de chiffre d'affaires et plus de cinq millions de clients. Allianz France emploie plus de 10000 salariés et se divise en deux domaines principaux : l'assurance de personnes et l'assurance non-vie. En terme d'assurance MRH, Allianz France se classe 8ème du marché français en termes de chiffre d'affaires hors taxes en 2019 d'après l'Argus de l'assurance.

Allianz France est divisée en plusieurs directions. Au sein de la branche non-vie, on trouve notamment la Direction Technique P3D (Particuliers, Pricing, Data et Pilotage).

La Direction Technique P3D est divisée en plusieurs équipes en fonction des produits. Il y a les produits destinés aux particuliers tels que les produits MRH et Auto et les produits destinés aux professionnels tels que les produits Flottes, Midcorp et Pros. La direction a pour objectifs :

- Le développement des produits Auto et MRH,
- Le développement du tarif des affaires nouvelles (*Standard Products Pricing*),
- Le suivi du portefeuille (Retail, Pros, Midcorp),
- Le pilotage de l'activité technique,
- La gouvernance des données (*Data*, Innovation et Gouvernance).

C'est au sein du service *Standard Products Pricing* que la tarification des produits de la branche assurance de biens et responsabilités est réalisée. En effet, ce service est en charge de :

- La modélisation du tarif technique et commercial des risques non-vie,
- L'adéquation du tarif et le respect des normes édictées par le Groupe Allianz,
- Le suivi de la concurrence et de la sensibilité clients pour optimiser les tarifs commerciaux.

1.3 Le produit MRH d'Allianz France

Bien que le marché actuel de l'assurance habitation soit dominé par les mutuelles, Allianz France possède des résultats très encourageants. Le produit MRH d'Allianz France représente plusieurs millions de contrats et plusieurs centaines de millions d'euros de chiffres d'affaires.

Le produit MRH d'Allianz France a évolué au cours du temps. La suite de ce mémoire portera sur le produit actuellement en vente puisque les questions à la souscription ainsi que les possibilités de réponses par l'assuré diffèrent selon les produits.

Actuellement, le produit MRH d'Allianz France est un produit annuel à tacite reconduction qui est composé de garanties obligatoires et d'autres facultatives. Les garanties facultatives peuvent être prises comme options en fonction du besoin des assurés. L'assurance MRH d'Allianz France n'est donc pas constituée de plusieurs formules comme le proposent certains concurrents.

Les garanties obligatoires sont donc immédiatement présentes dans le contrat lorsqu'un assuré souscrit un contrat MRH chez Allianz France. Les garanties facultatives sont prises en supplément lors de la souscription d'un contrat. Certaines d'entre elles, telles que le vol et le vandalisme, le bris de glace, l'assistance ou encore la panne sont proposées aux assurés avec un système de niveau. Plus le niveau est élevé, plus l'assuré est couvert. Par exemple pour la garantie facultative vol et vandalisme, l'assuré peut choisir de ne pas souscrire la garantie, d'opter pour la garantie avec le niveau 1 ou de prendre la garantie avec le niveau 2. Si l'assuré souscrit la garantie avec le niveau 1, seuls les vols à l'intérieur de l'habitation et le vandalisme à l'intérieur et à l'extérieur de l'habitation seront assurés. S'il choisit la garantie avec le niveau 2, alors en plus du niveau 1, les vols à l'extérieur de l'habitation seront assurés.

Garanties obligatoires	Granties facultatives
Incendie	Vol
Tempêtes, grêle, neige	Bris de glace
Dégâts des eaux	Assistance
Attentats	Dommages électiques
Responsabilité Civile	Panne
Catastrophes naturelles/technologiques	Protection juridique
Défense pénale et recours suite à un accident	Installations extérieures
Biens emportés en voyage/Villégiature	Objets de valeur
	Scolaire
	RC spécifique

TABLE 1.2 – Garanties obligatoires et facultatives du produit MRH d'Allianz France

La suite de ce mémoire se focalisera sur la garantie obligatoire dégâts des eaux et uniquement pour les maisons, chalets en bois, châteaux manoirs. En effet, certaines questions à la souscription diffèrent en fonction du type de bien à assurer.

Pour souscrire un contrat MRH à Allianz France, il y a plusieurs réseaux de distributions :

- Les agents qui sont dans des agences Allianz, en charge de vendre les produits d'assurance d'Allianz,
- Les courtiers, qui sont des personnes ou des groupes de personnes, n'appartenant pas à Allianz, qui aident les assurés à trouver l'assurance qui leur convient le mieux en termes de garanties et de tarifs. Lorsqu'une police d'assurance est vendue via un courtier, alors Allianz doit reverser une commission au courtier,
- La souscription en ligne, soit à travers les comparateurs d'assurance en ligne, soit directement à partir du site internet d'Allianz.

L'enjeu du mémoire qui va être détaillé par la suite, est d'alléger le questionnaire de souscription afin de faciliter le temps de souscription par l'assuré tout en récupérant les mêmes informations. Ainsi, il est important de présenter les questions posées à l'assuré lors de la souscription de son contrat d'assurance MRH chez Allianz France. Les questions sont :

- Quelle est la nature du logement ?
- Quel est le nombre de pièces du logement à assurer ?
- Quelle est l'adresse du logement à assurer ?
- Quelle est la qualité juridique (propriétaire, locataire,...) du souscripteur ?
- Quelle est la date de naissance du souscripteur ?
- Quelle est la CSP [4] du souscripteur ?
- Quelle est la typologie du bâtiment à assurer ?
- Quel est le montant de capitaux à assurer au sein du logement ?
- Quel est le montant d'objets de valeurs au sein du logement ?
- Quelle est la surface des dépendances du logement à assurer ?
- Le logement à assurer dispose t-il d'un insert ou d'une cheminée ?
- Est-ce que le logement dispose d'une alarme ou d'une télésurveillance ?
- Est-ce que le souscripteur à déjà eu des sinistres au sein de ce logement durant les trois dernières années ? Si oui, quelles garanties ont été impactées ?
- Quel niveau de franchise le souscripteur souhaite-t-il ?
- Le bâtiment à assurer est-il en cours de construction ?
- Quelle est l'ancienneté du bâtiment à assurer ?

D'autres questions sont ensuite posées pour savoir si le souscripteur souhaite prendre des garanties facultatives et pour certaines d'entre elles le niveau de protection souhaité.

4. Catégorie Socio-Professionelle

Les réponses à l'ensemble des questions ci-dessus ainsi que celles aux questions supplémentaires sur la prise d'options permettent d'attribuer au souscripteur un profil de risque et donc lui attribuer un tarif.

1.4 Le processus d'indemnisation

1.4.1 La garantie DDE

La garantie dégâts des eaux (DDE) couvre l'ensemble des dommages dus à une action de l'eau telle que :

- Les fuites d'eau qui ont lieu lorsqu'il y a un écoulement d'eau accidentel,
- Les débordements qui ont lieu lorsqu'un contenant d'eau est en excès et que l'eau se répand ailleurs qu'au sein de ce contenant, comme le débordement d'une piscine suite à d'importantes pluies,
- Les infiltrations qui ont lieu lorsque l'eau pénètre dans des endroits non propices à être au contact de l'eau,
- Les inondations. Il existe quatre types d'inondations :

  - Les inondations par ruissellement qui ont lieu dès lors que les eaux de pluie ne peuvent plus ou pas s'infiltrer dans le sol. Elles sont donc généralement la conséquence d'une pluie d'intensité très élevée ou de plusieurs jours de pluie. Elles ont généralement lieu dans des zones pentues, ce qui favorise l'accumulation d'eau,
  - Les inondations par débordement de cours d'eau qui ont lieu dès lors qu'un cours d'eau sort de son lit,
  - Les inondations par submersion marine qui ont lieu dès lors que des grandes vagues entraînent l'eau des mers sur les terres. Elles résultent généralement des grandes marées accompagnées de tempêtes et ont lieu au niveau des littoraux,
  - Les inondations par remontées de nappes qui ont lieu dès lors que le niveau d'eau d'une nappe phréatique remonte jusqu'à la surface du sol.

A noter que dès lors qu'une inondation fait l'object d'un arrêté ministériel, cette dernière n'est plus couverte par la garantie dégâts des eaux mais pas la garantie catastrophes naturelles.

1.4.2 Les conventions CIDRE et IRSI

Comme nous avons pu le voir précédemment, la garantie DDE est la garantie ayant la fréquence de sinistralité la plus importante pour l'assurance MRH. Dans l'objectif de faciliter les procédures d'indemnisation ainsi que le remboursement des sinistres, des conventions ont été mises en place en France et signées par la majorité des assureurs.

La première est la convention CIDRE [5] : Cette convention a été créée en 1970 et a fait ensuite l'objet de différentes mises à jour. La dernière a été publiée en 2002 et a perduré jusqu'au 31 mai 2018. Elle s'appliquait pour un sinistre dégâts des eaux s'étendant dans un immeuble, dans des maisons individuelles lorsque celles-ci sont mitoyennes ou encore entre une maison individuelle et un immeuble en copropriété. La convention s'appliquait pour les sinistres compris entre 250 € et 1600 € et stipulait que l'assureur de l'occupant du logement ayant eu un dégât des eaux devrait rembourser les dégâts du sinistre lié au contenu et à l'embellissement du logement. De plus l'assureur de l'occupant devrait prendre en charge la recherche de la fuite. L'assureur du bâtiment ou le propriétaire du logement devrait quant a lui, prendre en charge les dégâts causés aux parties privatives et communes du bâtiment. De plus, la convention affirmait qu'il n'y a aucun recours possible contre l'assureur du responsable du dégât des eaux.

A compter du 1er juin 2018, la convention CIDRE a été remplacée par la convention IRSI. [6] Elle s'applique au même périmètre de logement que la convention CIDRE mais, à la différence de la convention CIDRE qui s'appliquait uniquement aux sinistres dégâts des eaux, cette dernière s'applique en plus aux sinistres incendies. Elle s'applique à deux tranches de sinistres, la première contient les sinistres inférieurs à 1600 € et la deuxième aux sinistres compris entre 1600 € et 5000 €. Dans le cas d'un sinistre inférieur à 1600€, c'est l'assureur de l'occupant qui prend en charge l'ensemble du sinistre incluant le contenant et le contenu, à la différence de la convention CIDRE. La recherche de fuites est aussi à la charge de l'assureur de l'occupant et comme pour la convention CIDRE il n'y a aucun recours possible contre l'assureur du responsable du dégât des eaux. Pour les sinistres compris entre 1600 € et 5000 €, les conditions qui s'appliquent sont les mêmes que la convention CIDRE avec la possibilité de faire un recours contre l'assureur du responsable du dégât des eaux.

Ces deux conventions seront prises en considération lors de l'étude réalisée dans ce mémoire. En effet, la modélisation de la sinistralité sera réalisée sur les années 2018 et 2019, ainsi, les conventions sont appliquées sur ces sinistres.

1.5 Le principe de tarification

1.5.1 La prime pure et la prime commerciale

Un contrat d'assurance est un engagement réciproque entre un assureur et un assuré sur une période fixée en fonction d'évènements aléatoires bien définis. En général, en assurance non-vie, la durée d'un contrat est d'une année avec tacite reconduction. L'assuré s'engage à verser une prime envers l'assureur et l'assureur s'engage à verser une prestation à l'assuré en cas de dommage. A la différence d'autres produits vendus, les produits d'assurance ont un cycle de production inversé, c'est-à-dire que le produit, ici le contrat d'assurance, est vendu avant de connaître le coût de sa production, en l'espèce le

5. Convention d'Indemnisation Directe et de Renonciation à Recours en dégâts des Eaux
6. Indemnisation et Recours des Sinistres d'Immeuble

coût du dommage. Ainsi l'un des enjeux de l'assureur est de prédire le coût des sinistres afin de tarifer au mieux son produit d'assurance.

Pour ce faire l'assureur calcule la prime pure correspondant à l'engagement de l'assureur envers l'assuré, elle reflète la perte moyenne attendue par l'assureur, elle peut aussi être désignée comme étant le coût du risque.

Notons S la variable aléatoire qui correspond à la charge totale de l'assureur que l'on comptabilise pendant une période donnée T. Cette charge totale de sinistres correspond à :

$$S = \begin{cases} 0 & \text{si le nombre de sinistres obtenus est nul} \\ \sum_{i=1}^{N} X_i & \text{sinon} \end{cases} \tag{1.1}$$

où :

- N est une variable aléatoire donnant le nombre de sinistres obtenus durant la période T. N est à valeurs dans $\mathbb{N}$
- X=$(X_i)_{i\geq 1}$ est une suite de variables aléatoires dans $\mathbb{R}^+$.
- X_i est la variable aléatoire donnant le coût du i-ème sinistre.

Supponsons que :
- $\forall i > 0$, les variables N et X_i sont indépendantes.
- La suite de variables aléatoires $X = (X_i)_{i\geq 1}$ est iid. [7]

Selon les hypothèses ci-dessus, nous obtenons la prime pure telle que :

$$\mathbb{E}(S) = \mathbb{E}(N).\mathbb{E}(X_1) \tag{1.2}$$

E(S) correspondant à la prime pure, on a alors littérairement :

$$Prime\ Pure = Frequence * Coût\ Moyen \tag{1.3}$$

Ainsi, pour obtenir la prime pure, cette approche revient à modéliser la fréquence de sinistres ainsi que le coût moyen. C'est cette méthode qui va être suivie dans ce mémoire.

La fréquence de sinistres correspond au nombre de sinistres survenus durant une période donnée. Cette période correspond à l'exposition.

$$Frequence = \frac{Nombre\ de\ sinistres}{Exposition} \tag{1.4}$$

7. Indépendante et Identiquement Distribuée

Le coût moyen correspond à la charge de sinistres moyenne.

$$Coût\ Moyen = \frac{Charge\ totale\ de\ sinistres}{Nombre\ de\ sinistres} \tag{1.5}$$

Une fois que la prime pure est calculée, en rajoutant les frais liés à la commercialisation du produit, les frais de gestion, les frais généraux ainsi que les taxes, on obtient la prime commerciale, c'est à dire la prime réellement payée par l'assuré pour la souscription de son contrat d'assurance.

$$Prime\ Commerciale = Prime\ Pure + Frais + Taxes + Marge \tag{1.6}$$

1.5.2 La segmentation client

Le fondement premier de l'assurance est de mutualiser les risques. Le principe de mutualisation stipule que chaque assuré qui souscrit un même contrat d'assurance est exposé de manière indépendante mais identique aux risques pris en compte dans son contrat. Ainsi, si un assureur mutualise complètement son portefeuille, chaque assuré payera la même prime quel que soit son profil de risque, en effet il sera alors considéré que chaque assuré a le même profil de risque.

Cependant les assureurs tendent aujourd'hui à segmenter de plus en plus leurs portefeuilles, c'est-à-dire qu'ils vont créer des segments correspondants à des profils de risques et chaque assuré d'un même segment payera une prime identique. La segmentation présente un double objectif, le premier étant de responsabiliser l'assuré, en effet par exemple un assuré qui a des antécédents de sinistres payera une prime plus élevée qu'un assuré qui n'a jamais eu de sinistre ou encore dans le cas de l'assurance MRH un assuré qui possède une alarme aura une prime pour la garantie vol plus faible que celui qui n'en possède pas. Plus l'assuré tend à prévenir un potentiel sinistre et moins sa prime sera élevée. Le deuxième objectif de la segmentation est d'empêcher le phénomène d'antisélection. En effet, l'assurance est un domaine de plus en plus concurrentiel et sans segmentation, l'assureur risque d'avoir en portefeuille uniquement les assurés ayant un profil de risque plus propice à avoir une charge de sinistres importante. En effet une personne ayant un profil de risque non risqué aura tendance à s'assurer chez l'assureur qui segmente le plus et ainsi aura une prime reflétant son risque qui sera faible par déduction. Au contraire l'assuré qui à un profil de risque très risqué aura tendance à s'assurer auprès de l'assureur qui mutualise complétement car il payera la prime moyenne qui est donc largement inférieure à la prime reflétant son profil de risque. Ainsi l'assureur qui ne mutualise pas dans un marché concurrentiel se retrouvera avec uniquement des profils très risqués qu'il ne pourra compenser par des profils peu risqués.

	Assurés	Assureur
Charge	$\mathbb{E}(S)$	$S - \mathbb{E}(S)$
Charge moyenne	$\mathbb{E}(S)$	0

TABLE 1.3 – Charge de l'assureur et des assurés dans le cas de la mutualisation

	Assurés	Assureur
Charge	$\mathbb{E}(S/X)$	$S - \mathbb{E}(S/X)$
Charge moyenne	$\mathbb{E}(S)$	0

TABLE 1.4 – Charge de l'assureur et des assurés dans le cas de la segmentation

Par la suite, la modélisation sera faite à partir d'une segmentation non totale, c'est à dire uniquement sur la base des variables tarifaires que l'on aura sélectionnées. On aura alors $X = \{X_1, X_2, ..., X_p\}$ avec p le nombre de variables retenues pour la modélisation.

1.5.3 L'apport des données externes

Afin d'associer à chaque assuré un profil de risque, l'assureur doit poser un certain nombre de questions à l'assuré lors de la souscription de son contrat. Plus la segmentation est importante et plus le nombre de questions augmente ce qui peut être contraignant pour l'assuré qui peut y voir une perte de temps et une certaine complexité. Ainsi l'enjeu de l'assureur est d'avoir un tarif le plus segmenté possible en réduisant au maximum le nombre de questions posées à l'assuré. Pour ce faire l'assureur va utiliser des données externes qui vont lui permettre d'avoir des informations qu'il n'aura pas besoin de demander à l'assuré pour permettre d'alléger le questionnaire à la souscription sans réduire le nombre d'informations dont il dispose pour segmenter son tarif. Ainsi, de nombreuses données liées au logement peuvent être récupérées par le biais de données externes avec l'unique connaissance de l'adresse de l'habitation.

Ces données externes géographiques sont récupérées à différentes mailles plus ou moins précises à travers différents types de sources en *Open Data* mais également grâce aux travaux d'autres équipes au sein d'Allianz.

Des données à la maille du département ont été récupérées. La France compte 96 départements si on ne prend pas en compte les DOM [8]. Les départements sont représentés par un numéro à deux chiffres sauf pour la Corse qui est séparée en deux, le nord est nommé 2B et le sud 2A.

8. Département et région d'Outre Mer

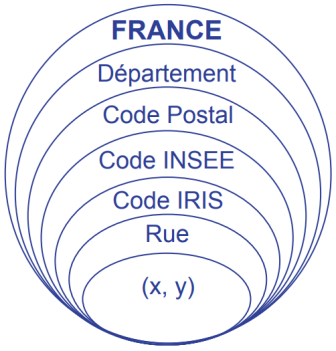

FIGURE 1.4 – Différentes mailles des données externes

Plus précis que le département, des données à la maille INSEE ont été récupérées. Le code INSEE est un code à cinq chiffres. La France est divisée en 34 883 codes INSEE en 2019 si on ne prend pas en compte les DOM.

Le code IRIS [9] est un découpage de la France plus fin que le code INSEE. Les communes avec un nombre d'habitants supérieur à 10 000 et une majorité des communes ayant un nombre d'habitants entre 5 000 et 10 000 sont découpées en IRIS. Le code IRIS est un code composé de neuf chiffres, les cinq premiers chiffres correspondent au code INSEE. La France est divisée en 48 590 IRIS en 2019 si on ne prend pas en compte les DOM. Chaque IRIS est composé d'environ 2 000 habitants. Si une commune n'est pas découpée en IRIS, on l'appellera une commune non-irisable et son IRIS sera alors son code INSEE suivi de quatre zéro. Dès lors qu'une commune est irisable, son code IRIS correspond à son code INSEE suivit de quatre autres chiffres. Par exemple, prenons la commune d'Arbigny qui est composée de 453 habitants, cette dernière est non-irisable. Son code INSEE est 01016 et son code IRIS est alors 010160000. La commune de Gex, quant à elle, est composée de 11 687 habitants, elle est irisable. Son code INSEE est 01173 et elle est divisée en trois codes IRIS qui sont alors 011730101, 011730102, 011730103. La commune de Bordeaux, présentée ci-dessous, se décompose en 88 IRIS.

9. Ilots Regroupés pour l'Information Statistique

FIGURE 1.5 – Découpage en IRIS de la commune de Bordeaux

D'autres données externes, par exemple des données de coûts peuvent aussi être récupérées. L'ensemble de ces données sera présenté dans le deuxième chapitre de ce mémoire. Les différentes sources de données externes utilisées dans la suite du mémoire sont issues :

- Des données publiques disponibles en *Open Data* sur différents sites en ligne,
- Des données récupérées dans différents services internes d'Allianz.

1.6 Méthodologie du mémoire

L'objectif du mémoire est d'alléger le questionnaire de souscription en apportant des données externes. Pour ce faire, deux approches vont être utilisées.

La première approche, présentée dans le troisième chapitre, va être de prédire directement des données latentes à partir de données internes et externes grâce à des méthodes de *Machine Learning*. Ces modélisations sont faites sur une période d'étude des contrats en portefeuille entre 2011 et 2017. Les deux données latentes prédites dans ce mémoire sont l'année de naissance de l'assuré principal à partir de laquelle l'âge sera calculé, et la nature du logement (résidence principale ou secondaire). Ces deux données correspondent aux réponses à deux questions posées lors de la souscription. Ainsi, une bonne prédiction permettrait de retirer ces questions à la souscription. Deux méthodes de *Machine Learning* sont utilisées dans ce mémoire, l'algorithme des Forêts Aléatoires et le *Gradient Boosting*. Ces données seront ensuite testées dans des modèles GLM afin de comparer la performance des modèles avec ces données latentes prédites et celle réalisée avec les données directement récupérées à la souscription.

La deuxième approche va être de comparer la performance de différents modèles GLM pour prédire la fréquence et le coût moyen pour la garantie dégâts des eaux sur la période

2018 et 2019. De part son caractère atypique l'année 2020 est retirée de la modélisation et fait l'objet d'une étude à part entière dans le chapitre 2 :

— Modèles GLM avec uniquement des données internes,
— Modèles GLM avec un mixte de données internes et externes,
— Modèles GLM avec uniquement des données externes.

Ainsi, l'objectif de cette deuxième approche est multiple. Il s'agit d'abord de répondre à la question de savoir si un modèle composé uniquement de données externes est aussi, voire plus, efficace qu'un modèle avec des données internes. Le cas échéant, il s'agit de démontrer que certaines données internes peuvent être substituées par des données externes sans nuire à la robustesse des modèles tarifaires.

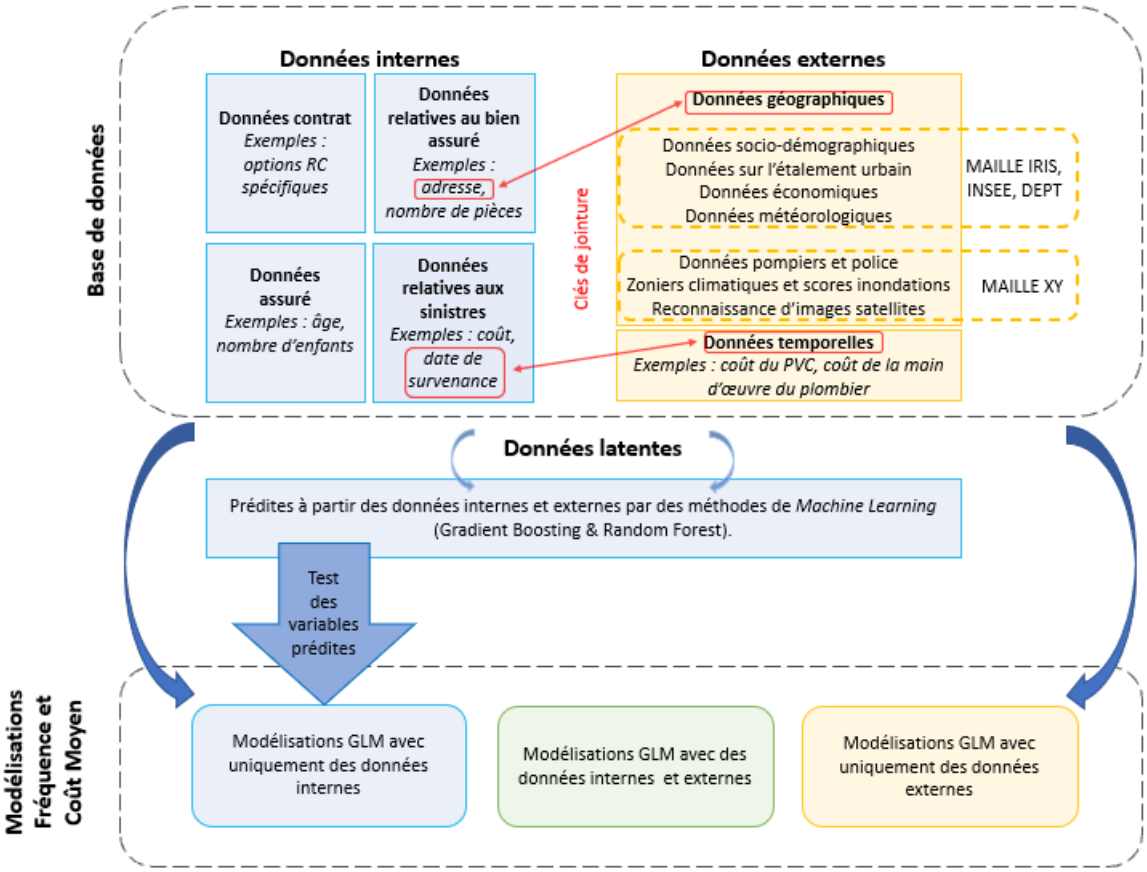

FIGURE 1.6 – Schéma récapitulatif de la méthodologie du mémoire

Chapitre 2

Constitution de la base de données

Ce deuxième chapitre vise à présenter le périmètre de l'étude ainsi que les données utilisées dans ce mémoire et qui sont les fondations des travaux de modélisations des chapitres suivants. En effet l'enjeu de ce mémoire est non seulement d'optimiser les modèles de tarification mais aussi de réduire le nombre de questions posées à la souscription. Pour ce faire, nous disposons de données internes qui sont par définition récupérées directement auprès de l'assuré. Elles sont généralement collectées à travers des questions posées lors de la souscription ou dans le cas d'un avenant (c'est-à-dire une déclaration de l'assuré pour un changement de risque, comme un déménagement, la naissance d'un enfant, ...). De nombreuses autres données, dite externes sont collectées et retraitées à partir de sources extérieures à l'assuré afin d'enrichir cette base de données internes et potentiellement de remplacer certaines de ces données internes. Les chiffres présentés dans cette partie ont été retraités pour des raisons de confidentialité.

2.1 Périmètre de l'étude

Les données internes d'Allianz France ont été recensées sur l'ensemble des contrats Multirisque Habitation en portefeuille entre 2011 et 2020. Le périmètre de l'étude de ce mémoire se porte ainsi un historique de 10 ans.

L'ensemble de ces données va être séparé en fonction de l'année.

- La période de 2011 à 2017 est destinée à la prédiction de variables latentes qui sera détaillée dans la troisième partie de ce mémoire.
- La période de 2018 à 2019 est destinée aux modélisations GLM pour prédire la fréquence et le coût moyen qui seront détaillées dans le chapitre 4 de ce mémoire.
- L'année 2020 va être retirée de l'étude de part son caractère atypique du à la crise sanitaire. L'année 2020 a fait l'objet d'une étude particulière qui sera présentée dans la suite de ce chapitre.

L'ensemble des données est anonymisé et concerne uniquement les habitations de type maisons, chalets en bois et châteaux manoirs. Les appartements, les lofts ainsi que les

caravanes ont été supprimés de l'étude. Par ailleurs, nous nous focalisons uniquement sur l'assurance du particulier, c'est-à-dire que les contrats des copropriétaires non occupants ne sont pas pris en compte dans l'étude de même que les contrats mono. Les contrats mono sont des contrats où l'assuré s'assure uniquement contre une garantie qui n'est pas liée à son logement, par exemple l'option vélo ou une responsabilité civile spécifique. Les contrats présents dans la base de données résultent de souscriptions via plusieurs canaux de distribution comme le réseau des agents, les courtiers ou encore via l'AEL (Agence En Ligne) comme présenté dans la première partie de ce mémoire.

Dans ce mémoire, l'objectif est d'optimiser les modèles tarifaires avec l'apport de données externes et notamment de données liées à l'adresse du risque. Aussi, il est important de rappeler que les biens assurés sont localisés en France métropolitaine (y compris la Corse) et à Monaco, qu'il s'agisse d'une résidence principale ou secondaire. Afin de visualiser la répartition des assurés présents dans le portefeuille sur le territoire français, la carte ci-dessous représente l'exposition de notre portefeuille pour les années 2011 à 2020. Par exemple lorsque l'exposition est de 18, cela peut représenter deux assurés qui ont été en portefeuille de 2011 à 2020.

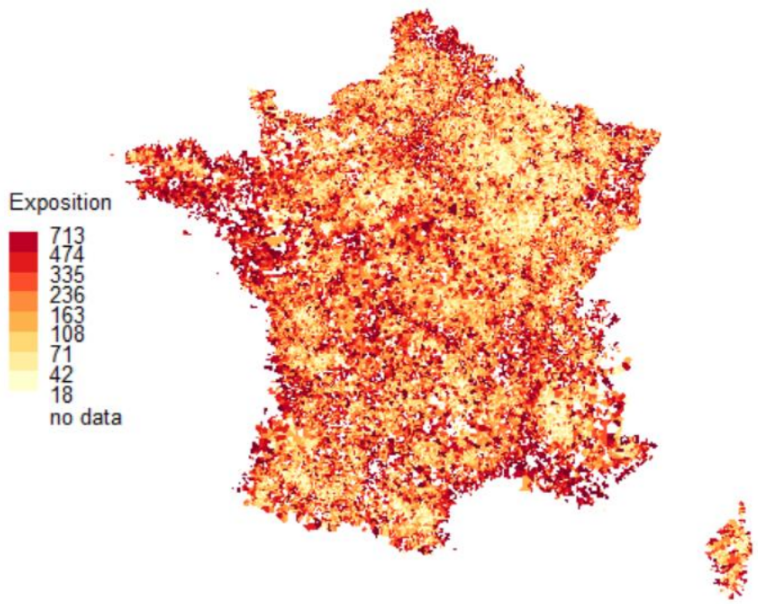

FIGURE 2.1 – Exposition de 2011 à 2020 du portefeuille MRH d'Allianz France par IRIS

Ainsi, on peut voir sur cette carte que les assurés sont répartis sur l'ensemble du territoire français. On note cependant une forte concentration au niveau de la Bretagne et une plus faible concentration au niveau de la région du Grand Est.

2.2 Présentation des données en portefeuille

2.2.1 Constitution de la base de données internes

Les données internes d'Allianz France présentes dans le portefeuille MRH regroupent l'ensemble des données qui sont récupérées à la souscription ainsi que lors de la survenance d'un sinistre. Elles se divisent en quatre catégories :

- Les données relatives au contrat,
- Les données relatives au bien assuré,
- Les données relatives à l'assuré,
- Les données relatives à l'historique de sinistralité.

La base de données est une base contrat, c'est-à-dire qu'une ligne correspond à un contrat sur une période donnée. Une période est définie par une date de début et une date de fin et, pour chaque période, l'exposition est calculée. L'exposition correspond au nombre de jours pendant lesquels le contrat est actif en portefeuille divisé par le nombre de jours dans l'année car l'exposition est calculée de manière annuelle. La décomposition d'une base contrat est présentée ci-dessous.

L'objectif de ce mémoire étant d'apprécier l'intérêt de l'ajout de données externes pour améliorer la performance des modèles ainsi que pour réduire le nombre de questions à la souscription, seuls les contrats relatifs au produit actuellement en vente en MRH ont été conservés. Ce produit a été présenté dans le chapitre précédent. L'étude de ce mémoire se focalise sur la garantie dégâts des eaux, ainsi comme cette garantie est obligatoire, l'ensemble des contrats dispose de cette garantie et donc l'exposition totale correspond à l'exposition de la garantie dégâts des eaux.

Lors de la souscription d'un contrat, une ligne est automatiquement créée dans la base de données. Pour un même contrat, plusieurs lignes peuvent être présentes dans la base, en effet une nouvelle ligne est générée pour un même contrat mais avec une période et donc une exposition différente dès lors qu'il y a :

- Un avenant, c'est-à-dire une modification dans le contrat initial, par exemple si l'assuré souhaite ajouter une option à son contrat,
- Une nouvelle année, en effet chaque 1er janvier une nouvelle ligne est automatiquement générée. L'exposition de nos bases étant annuelle, une période ne peut excéder une année. Ainsi l'exposition sur une ligne donnée ne peux être supérieure à 1,
- La survenance d'un sinistre, en effet dès lors qu'un sinistre a lieu une nouvelle ligne est créée de telle sorte que sur une période donnée il n'y ait pas plus d'un sinistre,
- Un renouvellement, en effet dès lors que le contrat arrive à son terme (ce sont des contrats annuels), une nouvelle ligne est créée.

Prenons l'exemple d'un assuré, ayant comme numéro de contrat 454545 et ayant été assuré par Allianz entre le 3 mai 2017 et le 2 février 2019. Dans nos bases de données nous aurons :

N° assuré	Début de période	Fin de période	Echéance	Sinistre	Expo
454545	03/05/2017	31/12/2017	03/05/2017	-	0,66
454545	01/01/2018	02/05/2018	03/05/2018	-	0,33
454545	03/05/2018	25/06/2018	03/05/2018	25/06/2018	0,15
454545	26/06/2018	31/12/2018	03/05/2018	-	0,52
454545	01/01/2019	02/02/2019	-	-	0,09

TABLE 2.1 – Découpage de la base de données pour un contrat donné en fonction des différentes périodes

Concernant les données liées au contrat, elles renseignent sur les options prises par l'assuré en plus du socle de garanties obligatoires proposé par Allianz France. En effet, comme mentionné dans la partie précédente, l'assuré peut ajouter des options dans son contrat MRH afin de se couvrir au mieux en fonction de ses besoins. Dans cette catégorie, on trouve par exemple la prise d'option Responsabilité Civile (RC) spécifiques (option RC chevaux, option RC chiens,...). Ces données propres au contrat décrivent également le réseau par lequel le contrat à été souscrit (agent, courtier, ...), les modalités de paiement (mensuelles, annuelles, ...), l'ancienneté du contrat...

Chaque année, l'ancienneté du contrat est recalculée à partir des dates de début de période présentées précédemment. L'âge est aussi recalculé à partir de l'année de naissance de l'assuré et de l'année en cours.

Le deuxième type de données internes concerne l'ensemble des informations liées aux biens assurés c'est à dire l'habitation ainsi que son contenu. Nous pouvons citer par exemple le nombre de pièces du logement, le montant d'objets de valeur à assurer au sein du logement ou encore la présence d'une piscine. Dans cette catégorie, nous pouvons également trouver l'adresse exacte du logement. A partir de cette information, le portefeuille MRH a été géocodé afin d'obtenir la longitude et la latitude correspondant à chaque adresse du portefeuille, l'objectif étant de pouvoir relier le portefeuille avec certaines données externes liées à la géographie de l'habitation qui seront présentées dans la suite de ce mémoire.

De nombreuses données permettent de décrire l'assuré lui-même ainsi que les personnes résidant au sein du logement assuré. Dans le cas d'une famille ou d'un couple, ce sont les informations concernant la personne qui souscrit le contrat qui sont renseignées dans cette catégorie. On trouve par exemple, l'âge, la catégorie socio-professionnelle ou encore le nombre d'enfants de l'assuré.

Enfin, nous avons les données liées aux sinistres. L'ensemble de l'historique de sinistralité par contrat est renseigné dans les bases internes d'Allianz France. On trouvera dans cette catégorie la date de survenance des sinistres, le coût des sinistres ainsi que le nombre de sinistres pour chaque garantie pour un contrat donné sur une période donnée. La date de survenance des sinistres permettra de lier les données externes temporelles relatives au coût du sinistre.

Le schéma ci-dessous résume les différentes parties de la base de données internes d'Allianz France.

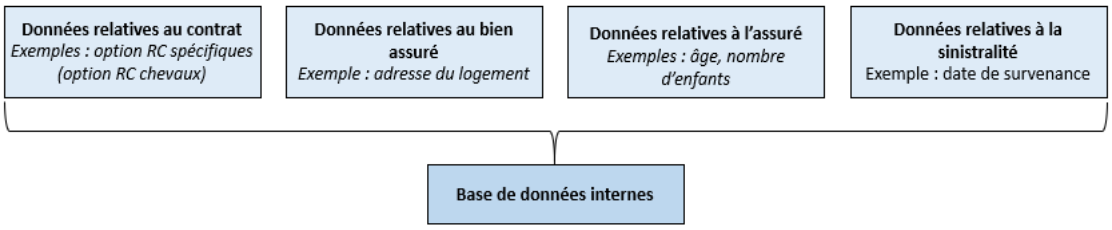

FIGURE 2.2 – Résumé de la structure de la base de données internes

2.2.2 Analyse du covid

La base de données retraitée correspond à des dates allant de 2011 à 2020. Nous souhaitons modéliser la fréquence et le coût moyen de la garantie dégâts des eaux sur la période 2018 à 2020. Cependant, 2020 étant une année particulière pour les assureurs du fait du changement de comportement des personnes durant la crise sanitaire, il est important de l'étudier. En effet, faire une modélisation avec une année atypique ne serait pas approprié car :

- Le modèle ne serait pas valable dans le temps,
- Certaines variables pourraient être sélectionnées dans les modèles alors qu'elles ne sont que le reflet d'un comportement atypique. Par exemple, les personnes exerçant la profession de cadre, habituellement sur leur lieu de travail, sont restées confinées à leur domicile et ont exercé leur profession en télétravail. Aussi, la survenance d'un dégât des eaux est détectée plus rapidement et le coût du sinistre est alors plus faible.

Les deux graphiques ci-dessous représentent les courbes de la fréquence et du coût moyen des sinistres dégâts des eaux du portefeuille MRH d'Allianz France pour les années 2018, 2019, et 2020. Elles permettent d'observer une périodicité de la fréquence et du coût moyen des sinistres. La fréquence de sinistralité dégâts des eaux ne semble pas être impactée par la crise sanitaire. Cependant, le coût moyen de la sinistralité dégâts des eaux décroît de manière importante entre juillet et décembre 2020. Cela peut s'expliquer par le fait que, suite aux différents confinements, les assurés étaient contraints de rester

chez eux et de favoriser le télétravail. Ainsi, lors de la survenance d'un sinistre dégâts des eaux, les assurés pouvaient le détecter au plus tôt afin de le stopper et de limiter les dégâts et donc le coût du sinistre. Toutefois, la vision des sinistres utilisés dans ce mémoire est une vision au 31 décembre 2020, c'est à dire que l'ensemble des sinistres ayant eu lieu en 2020 mais non déclarés à l'assureur en 2020 ne seront pas pris en compte dans la base de données. De plus un sinistre survenu et déclaré fin 2020 peut ne pas être clôturé et/ou faire l'objet d'une aggravation et donc impacter le coût moyen. En effet, un sinistre survenu en 2018 et vu en 2020 a un vieillissement de 2 ans alors qu'un sinistre survenu et vu en 2020 a un vieillissement de moins d'un an. Cet écart de vieillissement du sinistre peut impacter le montant des coûts moyens survenus fin 2020. Ainsi cette diminution du coût moyen peut s'avérer être le reflet de sinistres non encore déclarés ou de l'effet vision de la sinistralité. Une étude plus approfondie sur l'année 2020 a été menée grâce à des données de mobilités récupérées sur *Google* afin d'analyser le comportement des personnes en terme de mobilité au cours de l'année 2020.

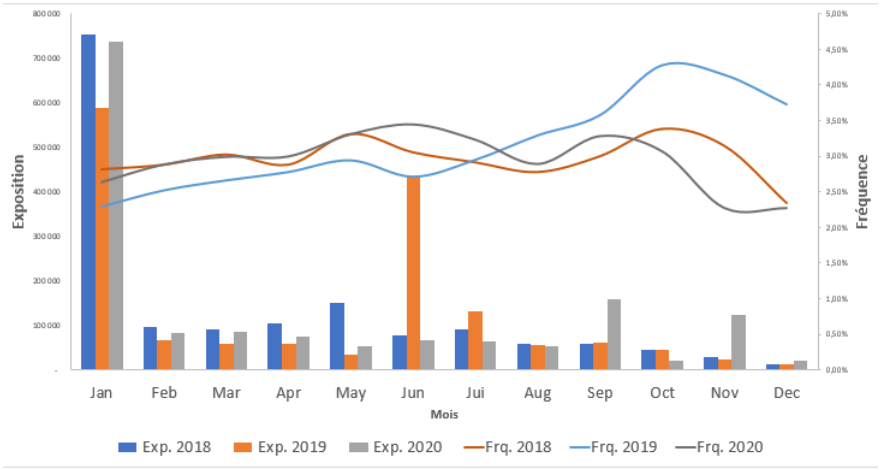

FIGURE 2.3 – Fréquence de sinistralité pour 2018, 2019 et 2020

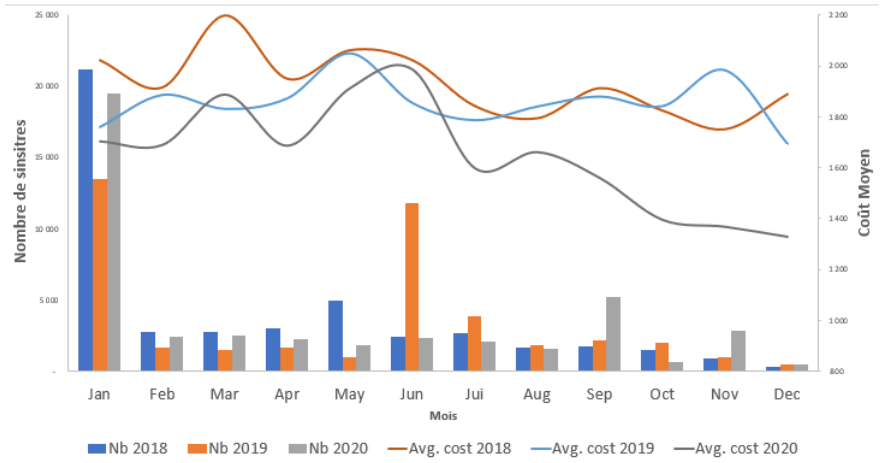

FIGURE 2.4 – Coût moyen des sinistres pour 2018, 2019 et 2020

En fonction des lois du gouvernement, l'année 2020 a été découpée en 4 périodes :
- *00-2020_No_restriction* correspond au début de 2020 lorsqu'il n'y avait pas de restriction (du 01/01/2020 au 16/03/2020)
- *01-Lockdown* correspond au premier confinement jusqu'à la deuxième phase de déconfinement (du 17/03/2020 au 01/06/2020)
- *02-Lockdown_end* correspond à la période entre les deux confinements de 2020 (du 02/06/2020 au 29/10/2020)
- *03-Lockdown_2* correspondant au deuxième confinement de l'année 2020 (du 30/10/2020 au 31/12/2020)

Le graphique ci-dessous représente la tendance de mobilité des personnes par rapport à leur lieu d'habitation. La tendance de mobilité est représentée en pourcentage par rapport à une base qui représente la mobilité habituelle des personnes. On peut voir qu'avant le premier confinement, la tendance de mobilité est proche de la base. Lors du premier confinement très strict, on peut voir que par rapport à la tendance de base, les personnes sont davantage restées chez elles. Entre les deux confinements, les personnes ont eu une mobilité proche de la tendance de base, cela peut s'expliquer par un retour à la normale où chaque personne reprend ses habitudes. Enfin lors du deuxième confinement, on note une tendance des personnes à rester chez elles inférieure à celle du premier confinement mais bien au-dessus de la tendance de base. En effet, le deuxième confinement permettait à certains de se déplacer pour aller travailler. Cependant, ceux pour qui le télétravail était possible devaient rester chez eux, ce qui explique la tendance supérieure à la base des personnes à restées dans leur lieu d'habitation.

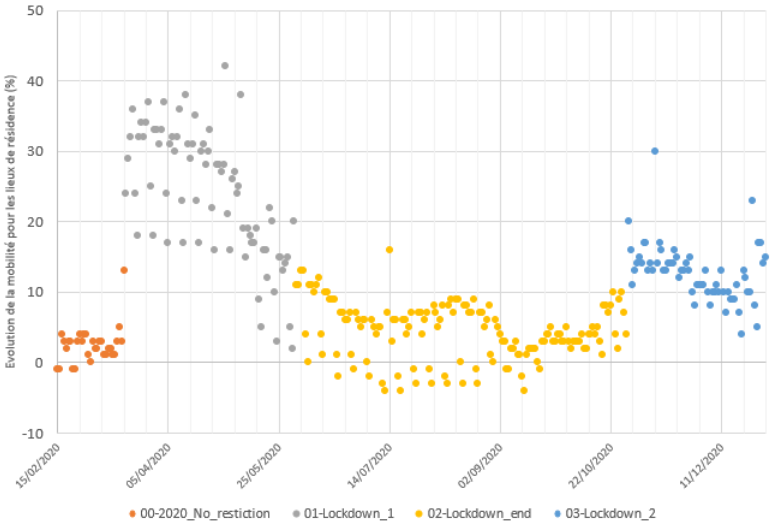

FIGURE 2.5 – Tendance de mobilité pour les lieux d'habitation

Le graphique ci-dessous représente la tendance de mobilité pour les lieux de travail, c'est à dire la tendance de mobilité des personnes à se rendre sur leur lieu de travail par rapport à une base représentant la tendance de mobilité habituelle des personnes à se rendre sur leur lieu de travail. On peut voir comme sur le graphique précédent, que la tendance de mobilité des personnes à se rendre sur leur lieu de travail avant le premier confinement est conforme à la base. Lors du premier confinement, un grand nombre de personnes était en activité partielle et un grand nombre également contraint de télétravailler. Ainsi on peut voir une chute importante de la tendance de mobilité pour les lieux de travail lors du premier confinement. Durant la phase entre les deux confinements, le télétravail restait très encouragé par le gouvernement pour l'ensemble des personnes qui étaient en mesure de télétravailler, ce qui explique la tendance de mobilité légèrement inférieure à la base. Enfin lors du deuxième confinement, les personnes ne pouvant pas télétravailler pouvaient se rendre sur leur lieu de travail contrairement au premier confinement, ce qui explique la tendance de mobilité légèrement inférieure à la base pour le deuxième confinement de 2020.

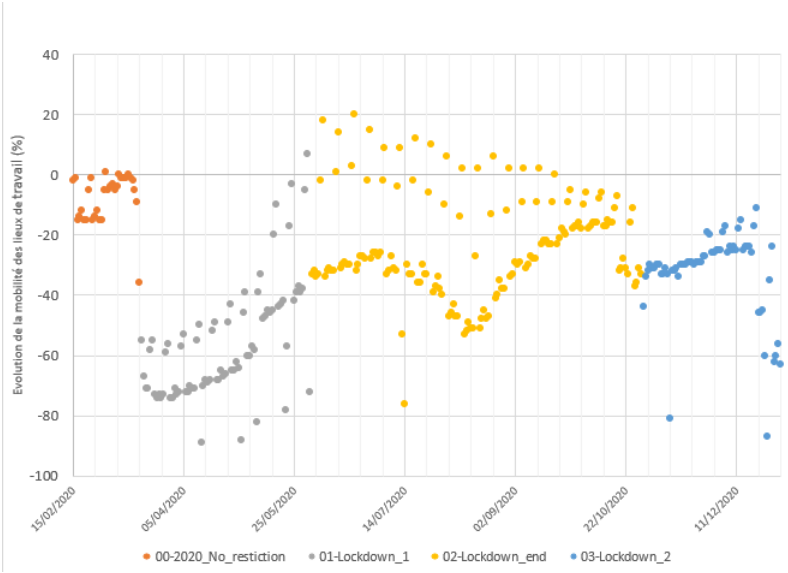

FIGURE 2.6 – Tendance de mobilité pour les lieux de travail

Ainsi, on peut conclure que le découpage de l'année 2020 en 4 périodes reflète la différence de comportement des assurés. Regardons à présent la fréquence et le coût moyen de la garantie dégâts des eaux selon ces quatre périodes ainsi que 2018 et 2019 en fonction de certains facteurs explicatifs.

Par exemple, le graphique ci-dessous présente le coût moyen par période selon la CSP. On remarque que les enseignants ont un coût moyen en dégâts de eaux particulièrement bas sur les périodes de confinement en 2020 (en rouge pour le premier confinement et en rose pour le deuxième) par rapport aux périodes hors confinement en 2020 et par rapport au niveau de 2018 et 2019.

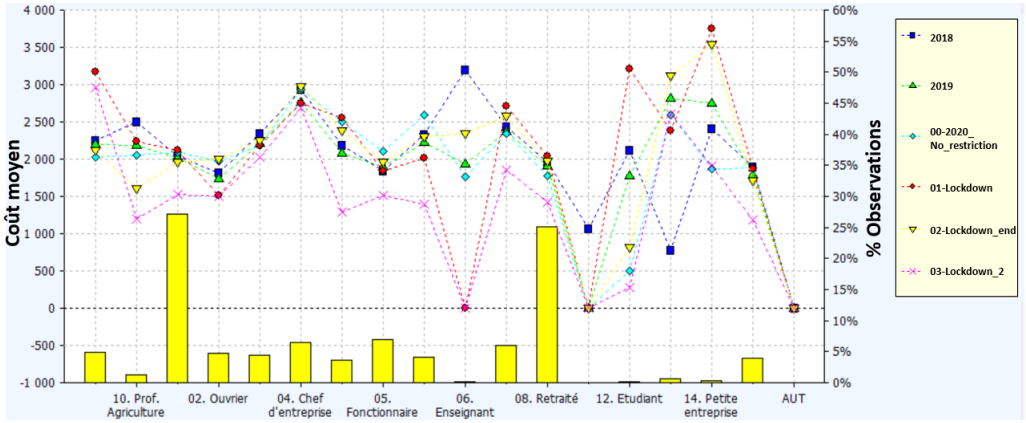

FIGURE 2.7 – Coût moyen des sinistres en fonction de la CSP pour différentes périodes

Le graphique ci-dessous présente la fréquence par période selon le nombre de pièces du logement. On remarque que cette fréquence est particulièrement basse sur la période du deuxième confinement par rapport aux périodes hors confinement en 2020 et par rapport au niveau de 2018 et 2019.

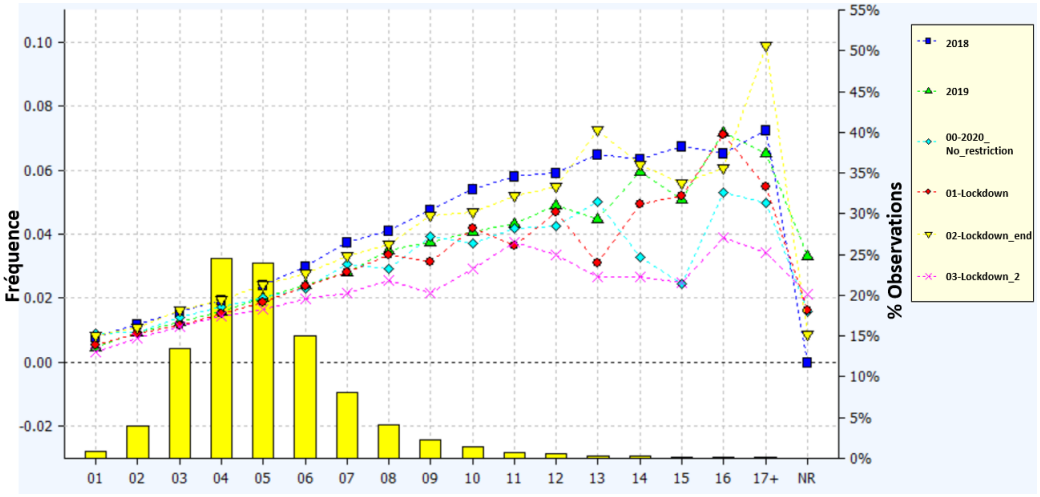

FIGURE 2.8 – Fréquence des sinistres en fonction du nombre de pièces pour différentes périodes

Toutefois, conserver les 2 périodes hors confinement en 2020 (avant mars et pendant l'été) pourrait également créer un biais. En effet, les graphiques 2.3 et 2.4 témoignent d'une certaines périodicité. Une étude de séries temporelles approfondit pourrait le démontrer. Ainsi, nous prenons la décision de supprimer entièrement l'année 2020 pour la suite de ce mémoire.

2.2.3 Statistiques descriptives

Afin de mieux visualiser la répartition du portefeuille de l'étude, l'ensemble des graphiques ci-dessous permet d'observer la fréquence et le coût moyen observés dans le portefeuille MRH d'Allianz France pour le produit actuellement en vente pour les maisons sur les années 2018 et 2019. Le portefeuille est composé de 88 % de résidences principales.

La variable région renseignée dans le portefeuille permet de diviser la France en 13 parties. On observe sur le graphique ci-dessous la fréquence et le coût moyen en fonction des régions. Tout d'abord ce graphique permet de voir que les assurés sont répartis sur l'ensemble du territoire français. La région Île-de-France se démarque avec un coût moyen et une fréquence de sinistres plus élevée. Cela peut s'expliquer par le coût élevé des logements parisiens.

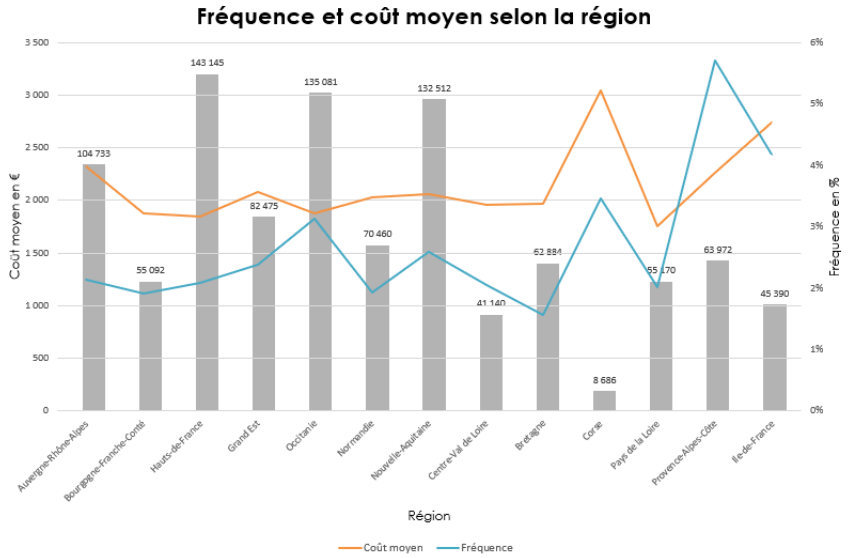

FIGURE 2.9 – Fréquence et coût moyen en fonction de la variable région

Le graphique ci-dessous permet quant à lui d'observer la fréquence et le coût moyen en fonction du nombre de pièces. Le portefeuille est composé majoritaire de logements de 4 ou 5 pièces. Plus le nombre de pièces est élevé et plus la fréquence de sinistralités ainsi que le coût moyen sont élevés. Cela peut s'expliquer par le fait que plus un logement comporte de pièces, plus il y a de pièces d'eau qui peuvent être à l'origine d'un sinistre dégâts des eaux. De plus, la taille du logement est proportionnelle au coût du sinistre.

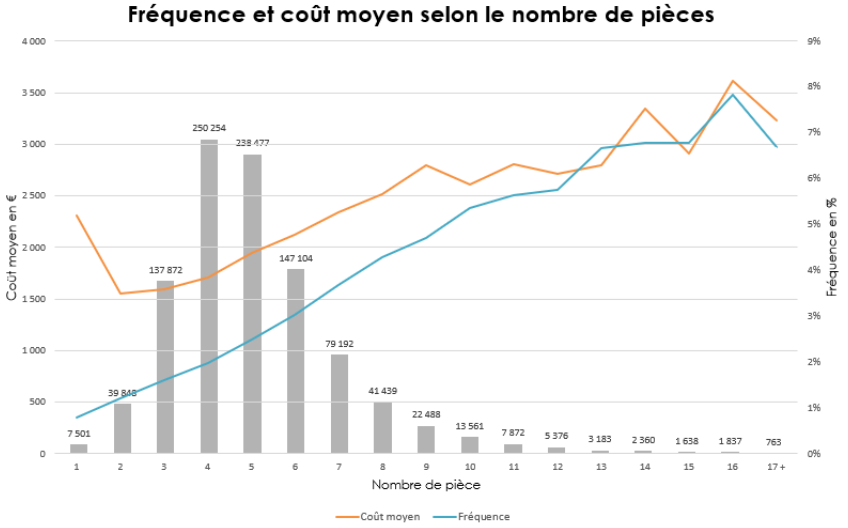

FIGURE 2.10 – Fréquence et coût moyen en fonction de la variable nombre de pièces

Dans le graphique ci-dessous, on observe la fréquence et le coût moyen en fonction de la qualité juridique. Le portefeuille est composé de 73 % de propriétaires et 22 % de locataires. Cette variable est importante dans les modèles de tarification car l'assurance habitation est obligatoire en France en fonction de la qualité juridique de l'assuré. En effet, l'assurance MRH est obligatoire pour les locataires et non obligatoire pour les propriétaires sauf lorsque le logement se situe dans une copropriété. On peut observer sur le schéma ci-dessous que la fréquence ainsi que le coût moyen des sinistres sont bien plus élevés pour les propriétaires que pour les locataires. Le coût moyen peut s'expliquer par le fait que les propriétaires ont probablement plus de biens chez eux. Par ailleurs pour la fréquence cela découle du fait qu'une partie des dégâts des eaux causés dans un logement assuré n'est pas à la charge du locataire mais à celle du propriétaire, par exemple lorsqu'une canalisation se casse.

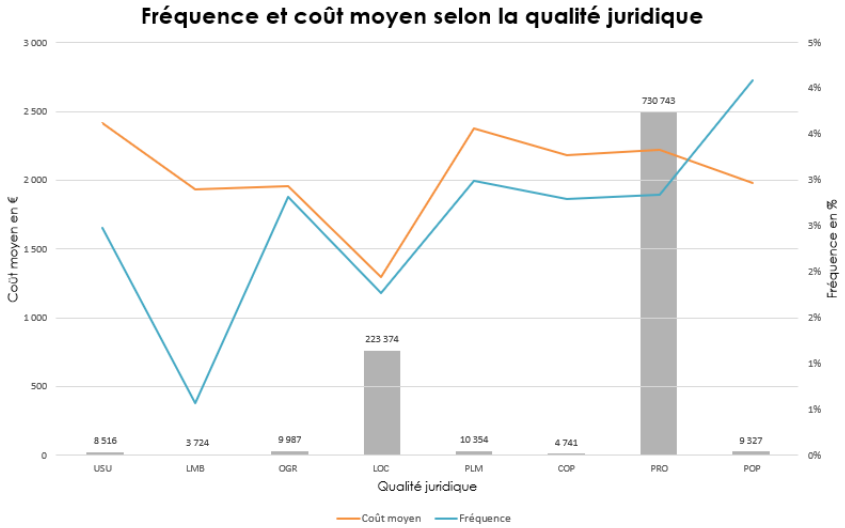

FIGURE 2.11 – Fréquence et coût moyen en fonction de la variable qualité juridique

Enfin, le graphique ci-dessous renseigne sur la fréquence et le coût moyen en fonction de la catégorie socio-professionnelle de l'assuré. On observe que les retraités ainsi que les employés/salariés sont les plus représentés dans le portefeuille. Les ouvriers, étudiants et fonctionnaires ont une fréquence et un coût moyen plus faibles que les professions libérales et les chefs d'entreprises.

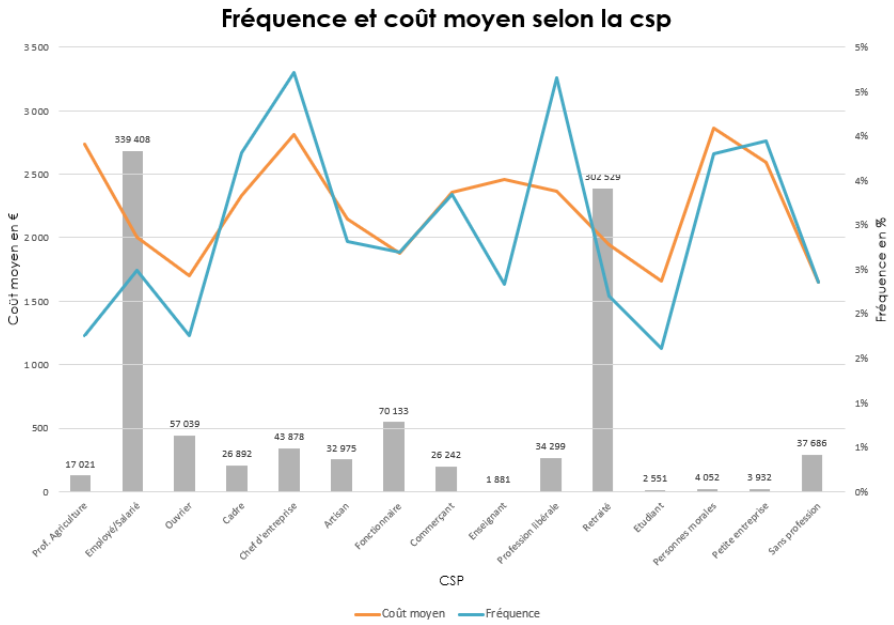

FIGURE 2.12 – Fréquence et coût moyen en fonction de la variable CSP

2.3 Présentation des données externes

L'assureur ne possède pas directement d'informations décrivant et illustrant l'environnement du bien à assurer. De nombreux facteurs peuvent alors être récoltés et retraités. On parle de données externes. Dans ce mémoire, elles sont collectées en *Open Data* ainsi qu'auprès de différents services internes chez Allianz. Ces données sont divisées en deux types de données :

— Des données géographiques, comme des données démographiques, socio économiques, météorologiques. Ces informations sont intégrées aux données internes via l'adresse du logement assuré,

— Des données de coûts, comme celui de la main d'œuvre du plombier ou du cours du cuivre. Elles seront utilisées uniquement pour les modèles de coûts moyens et sont intégrées via la date de survenance du sinistre.

2.3.1 Données géographiques

En IARD, la zone géographique constitue l'un des critères de segmentation les plus couramment employés par les assureurs. En effet, l'estimation de la sinistralité est souvent influencée par l'environnement géographique, météorologique, socio-démographique, ... du risque. Généralement, l'assureur dispose uniquement de l'adresse du lieu où est exposé le risque dans ses données internes.

2.3.1.1 Géocodage du portefeuille

L'ensemble des adresses de risque du portefeuille MRH a été géocodé. Ainsi pour chaque adresse de logement assuré nous avons la longitude, la latitude ainsi que l'IRIS. A partir de l'IRIS, nous pouvons en déduire l'INSEE puis le département.

Lors du géocodage, une variable concernant la qualité de géocodage est ajoutée en fonction de la précision du géocadage réalisée pour une adresse donnée. Les taux de fusion *tx_niv_fusion*, c'est à dire les pourcentages d'exposition du portefeuille géocodé pour une précision donnée sont renseignés dans le tableau ci-dessous :

Géocodage latitude/longitude (XY)	expo_dde	tx_niv_fusion
Centre du bâtiment	2763343,56	60,75%
Adresse	357234,58	7,85%
Interpolation de l'adresse	412015,14	9,05%
Mixe interpolation de l'adresse et centre de la rue	71883,93	1,58%
Centre de la rue	859071,29	18,89%
Centre du lieu-dit-habité	2,89	0,00%
Centre du chef-lieu de la commune	74011,66	1,63%
Centre du code postale	11047,63	0,25%
Monaco	12,41	0,00%

TABLE 2.2 – Statistiques sur la qualité de géocodage des adresses du portefeuille MRH sur les périodes de 2011 à 2020 (arrondi au centième)

Dans la suite du mémoire l'objectif va être de récupérer les données externes à la maille la plus fine. Lorsque l'information est manquante pour une maille donnée, alors les données seront agrégées à une maille supérieure et ainsi de suite. L'ensemble des adresses de risque du portefeuille MRH d'Allianz France contient au moins le département, ainsi à la maille du département l'ensemble des données externes peuvent être jointes avec le portefeuille.

Maille	expo_dde	tx_niv_fusion
IRIS	4461839,18	98,09%
INSEE	84452,99	1,86%
DEPT	2330,93	0,05%

TABLE 2.3 – Statistiques sur les mailles disponibles dans le portefeuille MRH

2.3.1.2 Données socio-démographiques

Les données socio-démographiques ont été récupérées sur le site de l'INSEE. Elles ont étés collectées par l'INSEE en 2017 et publiées en 2020. L'ensemble de ces données

a été mise sous forme de taux afin de prendre en compte les différences de population et de nombre de logements par zone géographique.

Prenons la variable *tx_ csp_ retraite* qui correspond au taux de ménages dont l'assuré principal est à la retraite par zone géographique. Cette dernière est issue de deux variables externes récupérées en *Open Data* à l'IRIS qui sont le nombre de ménages dont l'assuré principal est à la retraite par IRIS et le nombre de ménages par IRIS.

Notons a_i, le nombre de ménages dont la personne de référence est à la retraite pour un IRIS i et b_i le nombre de ménages dans l'IRIS i.

1. La variable est calculée à l'IRIS. Pour un IRIS donné i, on a : $tx_csp_retraite_i = \frac{a_i}{b_i}$
2. La variable est ensuite calculée à l'INSEE à partir de la variable calculée à l'IRIS. Pour un INSEE v, on a : $tx_csp_retraite_v = \frac{\sum\limits_{i \in v} a_i}{\sum\limits_{i \in v} b_i}$
3. La variable est calculée au département à partir de la variable calculée à l'IRIS. Pour un département d, on a : $tx_csp_retraite_d = \frac{\sum\limits_{i \in d} a_i}{\sum\limits_{i \in d} b_i}$

Ainsi, comme exposé dans le tableau ci-dessus, la variable tx_csp_retraite est renseignée à la maille IRIS pour 98,09 % de l'exposition du portefeuille, à la maille INSEE pour 1,86 % de l'exposition du portefeuille et enfin à la maille du département pour 0,05% de l'exposition du portefeuille. Le taux de manquants de cette variable est nul. Ce même raisonnement a été appliqué à l'ensemble des variables socio-démographiques.

2.3.1.3 Données relatives à l'étalement urbain

L'étalement urbain reflète l'augmentation de la superficie d'une commune et la diminution de sa densité de population. Afin de quantifier l'étalement urbain, plusieurs données externes ont étés collectées :

— La densité moyenne par IRIS, récupérée sur le site de l'INSEE. Elle se calcule comme la superficie de l'IRIS divisée par le nombre d'habitants de l'IRIS. Le nombre d'habitants est récupéré sur le site de l'INSEE alors que la superficie de chaque IRIS a été calculée par une autre équipe d'Allianz France.
— L'altitude au centroïde de chaque IRIS, récupérée auprès d'une autre équipe d'Allianz,
— Le taux de surface en eau par IRIS calculé à partir de données provenant du site du gouvernement (base Corine),
— L'aire urbaine, récupérée sur le site de l'INSEE et à la maille de l'INSEE.

Pour calculer le taux de surface en eau, la superficie des zones humides (*zone_ humide*) ainsi que la superficie des surfaces en eau (*surf_ eau*) ont étés récupérées en hectare par

IRIS (dans la base Corine). La superficie de chaque IRIS en mètre carré a ensuite été récoltée. Ainsi pour un IRIS donné i :

$$tx_surf_eau_i = \frac{zone_humide_i + surf_eau_i}{\frac{superficie_i}{10000}}$$

La densité, l'altitude ainsi que le taux de surface sont obtenus à la maille de l'IRIS puis agrégés à la maille INSEE et département comme évoqué pour les données socio-démographiques. L'aire urbaine quant à elle apporte des informations sur les communes ainsi que sur les communes voisines. Par sa définition, cette donnée ne peut pas être agrégée. En effet, toute chose étant égale par ailleurs, une petite commune dans l'aire urbaine de Paris n'est pas exposée au même risque qu'une « même » petite commune au milieu de la Creuse. Les deux cartes ci-dessous, réalisées via le logiciel R, représentent respectivement la densité de population par IRIS et le taux de surface en eau par IRIS. Sur la première carte, on observe les grandes villes en couleur rouge/orange pour lesquelles la densité de population est élevée telles que Paris, Marseille, Bordeaux et Lyon. Sur la deuxième carte, on voit se dessiner les fleuves et cours d'eau en bleu foncé.

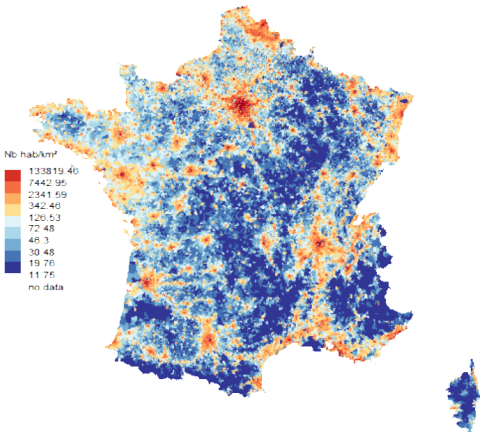

FIGURE 2.13 – Densité de population par IRIS

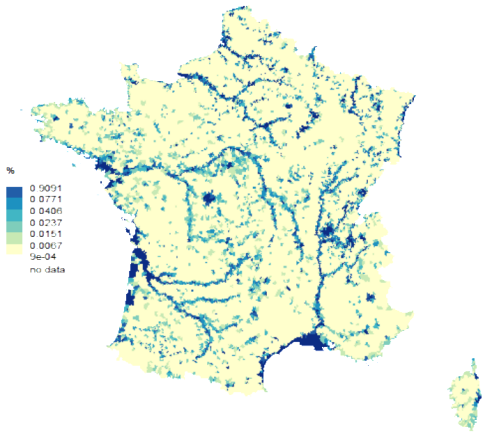

FIGURE 2.14 – Taux de surface en eau par IRIS

2.3.1.4 Données économiques

Les données économiques permettent d'apporter des informations de marché. Des données payantes du CALLON telles que le coût de reconstruction au mètre carré des maisons ont été achetées par Allianz. Elles ont étés collectées par le CALLON en 2018 à la maille du département. D'autres données ont étés collectées sur le site du gouvernement tel que le prix moyen au mètre carré d'un logement. Cette donnée a été collectée à la maille INSEE par le gouvernement en 2019 et publiée en 2020. Les deux cartes ci-dessous, réalisées sur R, représentent respectivement le coût de reconstruction au mètre carré des maisons par département et le prix moyen au mètre carré d'un logement par INSEE. On peut noter que la région Île-de-France ainsi que le Sud Est de la France sont les zones les plus chères.

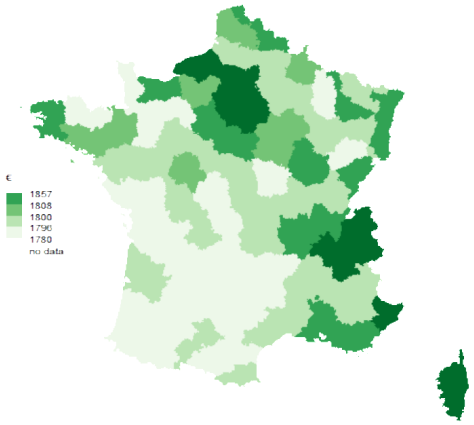

FIGURE 2.15 – Coût de reconstruction au mètre carré des maisons par département

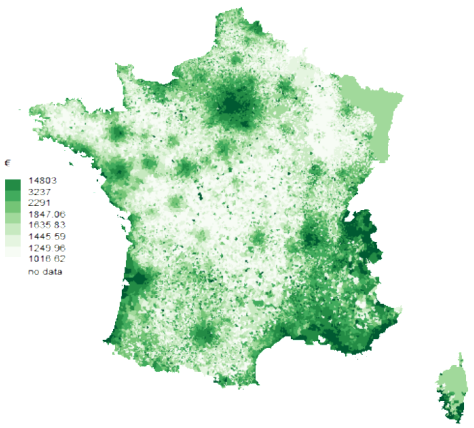

FIGURE 2.16 – Prix moyen au mètre carré d'un logement par INSEE

2.3.1.5 Données relatives aux pompiers et à la police/gendarmerie

Plusieurs données relatives aux pompiers ont été collectées sur le site du gouvernement et calculées. Une première donnée est par exemple le taux d'intervention des pompiers pour un sinistre dégâts des eaux par département. Pour calculer ce taux, le nombre d'interventions des pompiers pour un sinistre dégâts des eaux et inondation (nb_interv_dde) ainsi que le nombre de logements ($nb_logements$) par département ont été récupérés. Pour un département d donné on a alors :

$$tx_interv_dde_d = \frac{nb_interv_dde_d}{nb_logements_d}$$

Un deuxième travail concerne la récupération des adresses des casernes de pompiers ainsi que des commissariats de police et des gendarmeries. Elles ont étés collectées en *OpenStreetMap* puis géocodées par l'équipe dédiée afin d'obtenir pour chaque adresse la longitude et la latitude.

Les distances entre l'adresse du logement et chacune des adresses des casernes de pompiers ont été calculées, puis la distance minimale a été retenue pour chacune des adresses du portefeuille MRH d'Allianz France.

Pour les commissariats de police et les gendarmeries, ce sont les distances entre le XY (longitude, latitude) du centroïde de l'IRIS du logement assuré et les différentes adresses XY des commissariats de police et les gendarmeries qui ont été calculées, puis de même que pour les casernes de pompiers, la distance la plus petite a été retenue pour chacune des adresses du portefeuille. Nous aurions pu calculer la distance minimale entre l'adresse du logement et l'adresse des commissariats de police et des gendarmeries comme pour les casernes de pompiers. Nous avons calculé la distance au centroïde de l'IRIS du logement pour des contraintes de temps machine.

Pour calculer la distance angulaire en radian S entre deux adresses A et B sur une sphère, connaissant la longitude et la latitude de chacune des deux adresses, la formule suivante a été utilisée : notons R le rayon de la sphère (en mètres), φ_A, φ_A et λ_A, λ_B les latitudes et longitudes des adresses A et B.

$$S = R \cdot arccos(sin\varphi_A \cdot sin\varphi_B + cos\varphi_A \cdot cos\varphi_B \cdot \cos(\lambda_A - \lambda_B)) \tag{2.1}$$

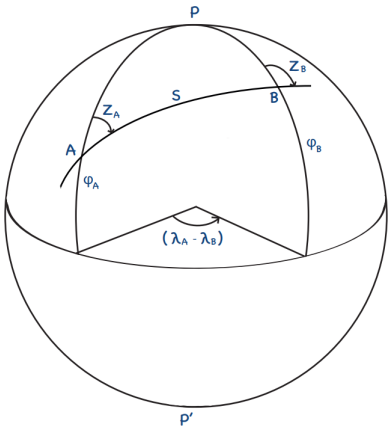

FIGURE 2.17 – Représentation de la distance angulaire entre deux points d'une sphère

Les cartes ci-dessous, réalisées via le logiciel R, permettent d'observer respectivement la moyenne des distances minimales entre une caserne de pompiers et un logement assuré par IRIS et la moyenne des distances minimales entre un commissariat de police ou une gendarmerie et le centroïde de l'IRIS d'un logement. On peut noter qu'à Paris, il y a une forte concentration de casernes de pompiers ainsi que de commissariats de police ou de gendarmerie car les distances minimales calculées sont plus petites.

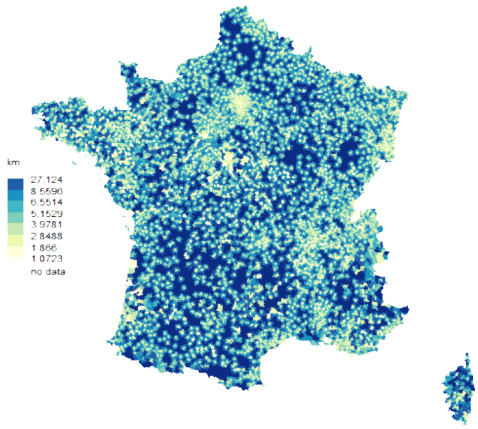

FIGURE 2.18 – Distance minimale entre une caserne de pompiers et un logement assuré

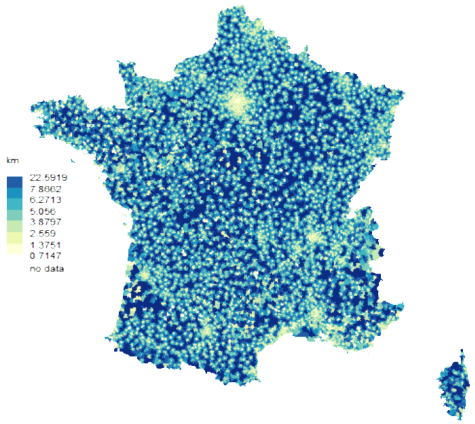

FIGURE 2.19 – Distance minimale entre un commissariat de police ou gendarmerie et le centroïde de l'IRIS d'un logement assuré

2.3.1.6 Données météorologiques

Les données météorologiques ont été récupérées sur le site de la NOAA (*National Oceanic and Atmospheric Administration*) en *Open Data*. La NOAA est une agence américaine qui dispose de stations météorologiques à travers le monde et notamment en France (environ 160 stations météorologiques). L'ensemble des données pour les stations en France a été récupérée du 1er janvier 2015 au 7 décembre 2020. Il s'agit de données journalières avec le détail du nombre d'heures d'observations. Parmi ces données, nous avons des données météorologiques telles que la température moyenne, minimale et maximale, la vitesse maximale du vent, ... Nous avons également des indicateurs qui prennent les valeurs 0 ou 1 selon s'il y a de la pluie, une tempête, de la grêle, ... au cours de la journée d'observation.

Ces données font l'objet d'un retraitement particulier qui est double. En effet, nous souhaitons obtenir des données agrégées temporellement et pour l'ensemble du territoire français :

- Les données sont relevées de manière journalière. Il convient de définir un segment de temps sur lequel agréger ces données. Nous allons choisir un découpage annuel et saisonnier (hiver/été).
- Les données météorologiques sont disponibles pour environ 160 stations françaises. Il convient donc d'utiliser des méthodes d'interpolations spatiales afin d'interpoler ces données pour chaque commune/quartier français.

Le premier retraitement est de convertir ces données aux formats français. Par exemple, les données sur la température, initialement en Degré Fahrenheit, sont converties en Degré Celsius. Ensuite, une agrégation temporelle a été réalisée afin d'obtenir pour chaque station ces variables selon trois segments temporels. En effet, pour chaque variable et

pour chaque station, une agrégation temporelle est effectuée en faisant :

1. La moyenne sur les 6 années,
2. La moyenne sur les mois d'été (avril, mai, juin, juillet, août) sur les 6 années,
3. La moyenne sur les mois d'hiver (septembre, octobre, novembre, décembre, janvier, février, mars) sur les 6 années.

Cette agrégation temporelle tient compte du nombre de jours et d'heures d'observations. Cela est possible car ces données journalières sont accompagnées du nombre d'heures d'observations dans la journée. En effet, les stations météorologiques ne récupèrent pas toujours l'information sur l'ensemble d'une journée. Ainsi, lors du retraitement des données, il est important de tenir compte du nombre d'heures d'observations via une pondération pour ne pas biaiser ces facteurs météorologiques pour les stations n'ayant pas des relevés complets (24h/24h).

Prenons par exemple la variable correspondant à la température moyenne $TEMP_k$ pour une station donnée k. On notera $TEMP_EXPO_{k,i}$ le nombre d'heures d'observation de la journée i pour la station k. Ainsi, la température moyenne sur les 6 années pour une station donnée k sera :

Soit A = [01/01/2015 - 07/12/2020],

$$\forall i \in A,\ TEMP_k = \frac{\sum_i (TEMP_{k,i} \cdot TEMP_EXPO_{k,i})}{\sum_i TEMP_EXPO_{k,i}} \tag{2.2}$$

Afin d'obtenir la donnée correspondant à la température moyenne en été $TEMP_été_k$ pour une station donnée k, la même formule s'applique en prenant uniquement les observations journalières ayant un mois appartenant aux mois considérés comme étant l'été (avril, mai, juin, juillet, août). La même méthode est appliquée pour les données hiver avec septembre, octobre, novembre, décembre, janvier, février et mars comme mois pris en compte.

A partir des données existantes, de nouvelles variables sont créées telles que le nombre de jours moyens par an où la température est inférieure à 0°C. En effet, en dessous de 0°C les canalisations cassent et un dégât des eaux peut en découler, ainsi cette variable peut être intéressante dans nos modèles.

Après avoir réalisé cette étape d'agrégation temporelle, une interpolation spatiale a été réalisée afin d'obtenir des données météorologiques par IRIS et non plus par station météorologique, le but étant de pouvoir attribuer à chaque logement de notre portefeuille une donnée météorologique. Pour chaque station, la longitude, la latitude ainsi que l'altitude étaient renseignées dans les données récupérées sue le site de la NOAA. La longitude et la latitude du centroïde de chaque IRIS ont été calculées.

Notons P_A, la valeur de la donnée météorologique au centroïde de l'IRIS A, W_{Ak} le poids de la donnée météorologique de la station k sur la donnée météorologique en A et enfin P_k la donnée météorologique de la station k. Ainsi nous avons :

$$P_A = \frac{\sum\limits_k W_{Ak} \cdot P_k}{\sum\limits_k W_{Ak}} \tag{2.3}$$

Le poids est défini à partir de la fonction de *biweight* qui tient compte des stations météorologiques uniquement dans un rayon D autour du centroïde de l'IRIS A. Cette fonction admet donc que les données récupérées aux stations météorologiques au-delà du rayon D n'ont aucune influence. La distance D a été définie arbitrairement à 150 kilomètres, afin d'avoir au moins une station météorologique dans chaque IRIS.

Notons d_{Ak}, la distance entre le centroïde de l'IRIS A et la station météorologique k calculée à partir de l'équation (2.1) vue précédemment. Ainsi à partir de la fonction de *biweight* :

$$W_{Ak} = \begin{cases} \left[1 - (\frac{dAk}{D})^2\right]^2 & \text{si } d_{Ak} < D \\ 0 & \text{sinon} \end{cases} \tag{2.4}$$

2.3.1.7 Zoniers climatiques et scores inondations

L'ensemble des données de cette partie a été récupéré auprès d'une autre équipe d'Allianz France. Elles sont le résultat d'une étude approfondie menée par cette équipe et vont être utilisées dans ce mémoire. La construction de ces données va être présentée succinctement car elles font l'objet d'une étude à part entière.

Les zoniers vents

Les zoniers vents ont été construits dans le cadre des travaux de thèse d'Alexandre Mornet réalisés chez Allianz. Ils ont été revus en 2019 et sont uniquement disponibles à la maille du département.

De nombreuses données ont été exploitées afin de constituer ces zoniers tels que :

- Des données de la NCDC (*National Climatic Data Center*) regroupant des données telles que les vitesses du vent moyennes et maximales par station météorologique. Ces données sont disponibles pour 90 stations entre 1973 et 1997 et 130 stations entre 1998 et 2012,
- Des données de Météo France regroupant des données telles que la vitesse maximale et moyenne du vent, la direction des vents instantanée, la vitesse du vent maximale instantanée, la vitesse moyenne du vent à 2 mètres,
- Des données liées aux sinistres du portefeuille Allianz entre 2004 et 2016,
- Des données historiques telles que les principales tempêtes Lothar et Martin en 1999, Klaus en 2009 et Xynthia en 2020.

A partir de ces données, sept zones géographiques ont été définies pour la fréquence du vent et sept zones pour la sévérité du vent correspondant à des classes de risques.

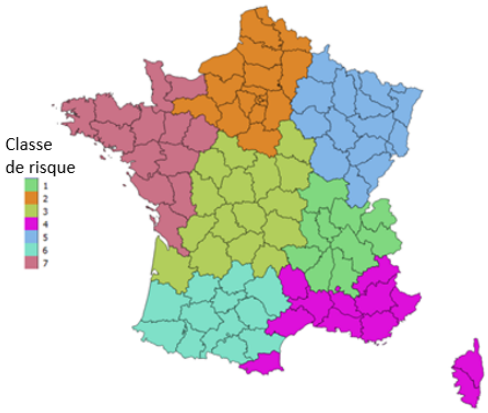

FIGURE 2.20 – Zonier vent - Fréquence

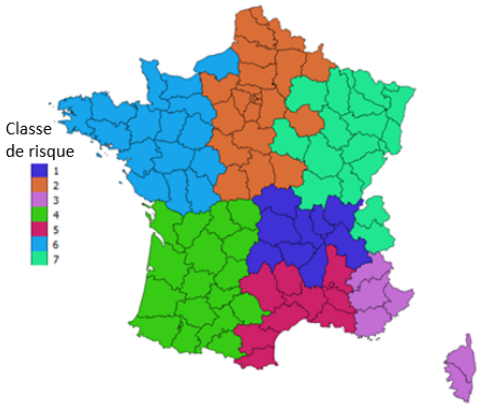

FIGURE 2.21 – Zonier vent - Sévérité

Le zonier tempête RMS

A partir du logiciel payant RMS, un modèle tempête destiné à prédire un niveau de prime pure pour l'ensemble des INSEE du territoire français a été développé au sein d'Allianz France en 2019. En input du modèle, il y a pour chaque police : l'adresse de l'assuré, le type d'habitation et la somme assurée. Le logiciel RMS est ensuite composé de trois modèles :

- Le premier est un modèle de risques qui va générer un évènement stochastique pour chaque INSEE en examinant la taille, la vitesse du vent, la zone géographique, ... de chaque INSEE.
- Le deuxième est un modèle de vulnérabilité qui va estimer le montant des évènements provoqués par le modèle de risque en fonction de différentes données liées au logement telle que l'année de construction.
- Le modèle financier qui va calculer les pertes financières évènement par évènement à partir des montants calculés par le modèle de vulnérabilité.

L'output du modèle est un zonier RMS représentant la prime pure pour une maison de 100 000 euros pour chacun des INSEE du territoire français.

Les scores inondations

L'ensemble des scores inondations est défini à la maille XY. En effet la proximité à un point d'eau permet de définir un score précis qui ne peut pas l'être à une maille plus grande tels que l'IRIS ou l'INSEE. Deux scores ont ainsi étés définis :

- JBA, qui est un score d'inondation par débordement qui prend en compte les défenses mises en place par l'homme telles que les digues.
- R, qui est un score d'inondation par ruissellement.

Le score d'inondation par ruissellement a été développé par un expert au sein d'Allianz tandis que le score par débordement provient des cartes fournies par l'organisme JBA. Ces deux scores sont définis selon cinq niveaux en fonction de l'exposition du logement assuré au risque inondation :

— 0 : Pas exposé
— 1 : Faiblement exposé
— 2 : Moyennement/Faiblement exposé
— 3 : Moyennement/Fortement exposé
— 4 : Fortement exposé

Le score convergence

Le score convergence permet de quantifier les endroits où l'eau stagne ce qui favorise les inondations. Rappelons que si une inondation fait l'objet d'un arrêté ministériel, alors cette dernière n'est plus couverte par la garantie dégâts des eaux mais pas la garantie catastrophe naturelle.

2.3.1.8 Reconnaissance d'images satellites

Une équipe au sein d'Allianz France a développé un outil de reconnaissance d'images satellites *Google* qui a été utilisé dans ce mémoire. Cet outil se présente de la manière suivante :

— Étape (1) : Renseignement de l'adresse du logement pour lequel on souhaite avoir des informations,
— Étape (2) : A partir de l'image du logement récupérée sur *Google Street View*, un algorithme va nous donner les informations présentées dans la capture d'écran ci-dessous. Chaque information est accompagnée d'un indice de confiance calculé à partir de l'entropie. Plus le nombre d'étoiles jaunes est important et plus l'indice de confiance est élevé.

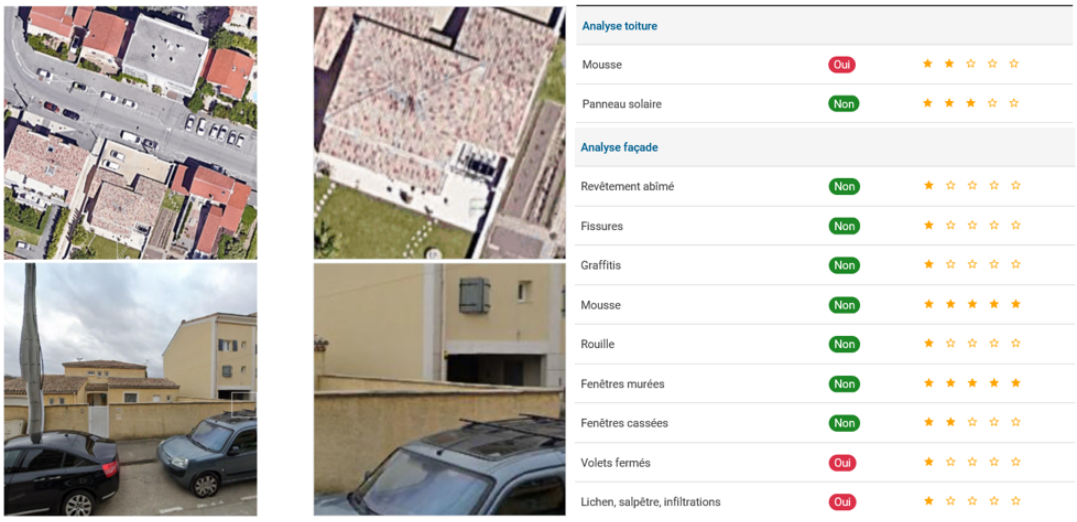

FIGURE 2.22 – Output de l'outil de reconnaissance d'images satellites d'Allianz

L'utilisation de cet outil pour l'ensemble des adresses du portefeuille étant difficile d'un point de vu IT et accessibilité, l'équipe IT d'Allianz nous a fourni directement une base de données correspondant à l'ensemble des données pour les adresses en portefeuille entre 2018 et 2020. Ces données seront utilisées uniquement pour les modélisations GLM du chapitre 4.

Nous récupérons ainsi, par exemple, la présence de mousses sur le toit ou encore la présence de fissures sur le mur du logement, ce qui peut favoriser l'infiltration de l'eau.

2.3.2 Données externes temporelles

Des données temporelles relatives à des coûts ont été récupérées par mois sur une période allant de 2012 à 2019 directement sur le site de l'INSEE. L'objectif de ces données est d'apporter de l'information supplémentaire aux modèles de coût moyen lors de la modélisation GLM. Ces données mensuelles ont été fusionnées a la base de données internes avec la date de survenance des sinistres. Les données de coûts ont été récupérées dans des bases différentes et ont été harmonisées en base 100 selon la même année de référence 2019. Parmi ces données, il y a le coût des travaux de plomberie en France métropolitaine, c'est à dire le prix moyen d'une heure de main d'oeuvre d'un plombier toutes taxes comprises. On peut voir que le prix moyen des services de plomberie augmente au cours des années, ce qui peut être une variable explicative pour les modèles de coût moyen.

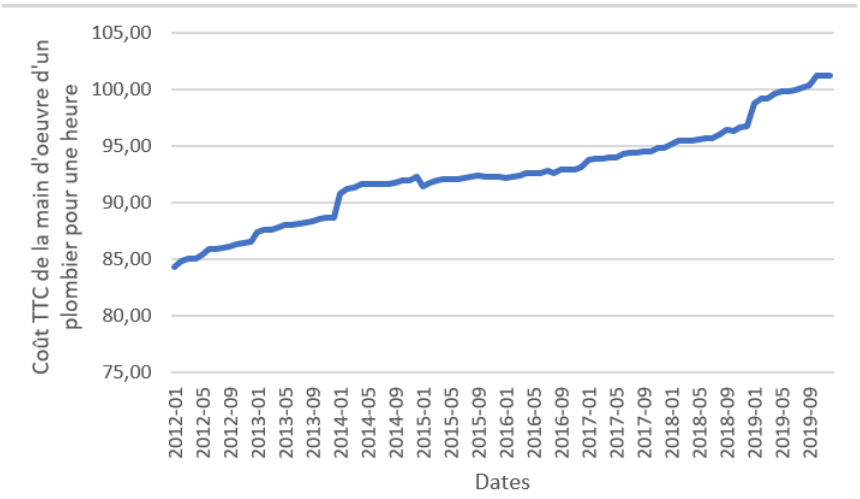

FIGURE 2.23 – Coût de la main d'oeuvre du plombier au cours du temps

Ainsi, de nombreuses données externes ont été récupérées en *Open Data* et ont ensuite été retraitées pour pouvoir être ajoutées au portefeuille MRH d'Allianz France. D'autres données ont directement été récupérées en interne auprès d'autres équipes d'Allianz. Le schéma ci-dessous résume les composantes des données externes.

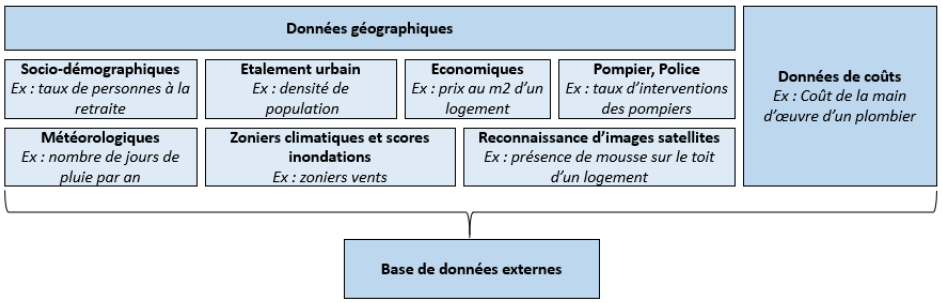

FIGURE 2.24 – Résumé de la structure de la base de données externes

2.4 Base de données finale

Une fois la base de données constituée et retraitée avec les données internes et externes, les données vont être séparées en deux bases de données en fonction des années. En effet, comme évoqué précédemment la base de données est une base de données contrat annuelle et donc chaque contrat dispose à minima d'une nouvelle ligne chaque année

dans la base de donnée. La première base de données correspondra aux lignes ayant une année de début de période entre 2011 et 2017 et sera utilisée pour les prédictions des variables latentes par méthode de *Machine Learning*. La deuxième base de données correspondra quant à elle aux lignes ayant une année de début de période entre 2018 et 2019. Comme étudié précédemment, l'année 2020 étant atypique, elle ne sera pas utilisée dans ce mémoire.

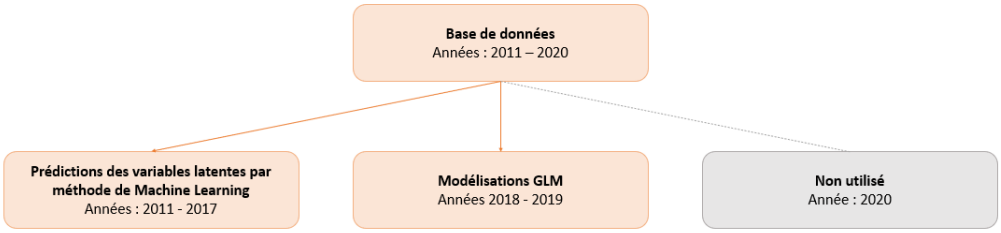

FIGURE 2.25 – Synthèse du découpage de la base de données

Chapitre 3

Création de données latentes

L'objectif de cette partie est d'utiliser des modèles de *Machine Learning* afin de prédire des variables latentes à partir de données internes et externes. Par définition une variable latente ne peut pas être mesurée directement et est non observable dans notre échantillon. Elle est issue d'algorithmes ou de modélisations statistiques et synthétise plusieurs variables observées. Ces données ont été présentées dans le chapitre précédent. Ces variables latentes correspondent directement aux réponses de questions posées à l'assuré lors de la souscription de son contrat. Ainsi, si la prédiction de ces variables est performante, elles pourront être supprimées du questionnaire posé à l'assuré lors de la souscription d'un contrat puisqu'elles seront directement prédites à partir d'autres variables. L'assuré gagnera ainsi du temps et sera moins découragé lors de la souscription de son contrat et l'assureur aura de son côté l'information des questions supprimées via les variables latentes créées. En effet chaque question posée à l'assuré lors de la souscription d'un contrat est importante, elles permettent à l'assureur de pouvoir segmenter son tarif et ainsi avoir un tarif le plus adapté au profil de risque de l'assuré. Les chiffres présentés dans cette partie ont été retraités pour des raisons de confidentialité.

Les modèles de *Machine Learning* vont être réalisés sur une base de données allant de 2011 à 2017 dans l'objectif de pouvoir ensuite prédire les données latentes à partir de ces modèles sur la base de modélisation des GLM de 2018 à 2019 pour pouvoir tester leur performance. Cette base de données est une base représentant 1 909 981 lignes et 151 colonnes. Chaque colonne correspond à une variable explicative pour les modèles et chaque ligne à un contrat donné. La base présentée dans la partie précédente a été retraitée de telle sorte que si un contrat n'a pas subi d'avenant ayant modifié une ou plusieurs variables explicatives, alors le contrat est représenté par une seule ligne afin de ne pas avoir de doublons. Le nombre de lignes important va permettre d'avoir des modèles robustes mais le temps d'exécution des programmes est long.

Plusieurs variables latentes ont été testées :
- La CSP,
- Les capitaux assurés,

— La nature du logement,
— La qualité juridique,
— L'année de naissance de l'assuré principal.

L'objectif métier est dans un premier temps de prédire deux données latentes afin de pouvoir retirer deux questions du questionnaire de souscription MRH d'Allianz France. Les méthodes de *Machine Learning* présentées ci-dessous ont été réalisées sur l'ensemble de ces variables et ce sont l'année de naissance de l'assuré ainsi que la nature du logement qui ont été retenues pour ce mémoire. En effet, c'est pour ces deux variables que les prédictions ont été les plus performantes.

La variable année de naissance correspond à l'année de naissance de l'assuré principal. Cette variable est une variable continue qui est renseignée dans le questionnaire de souscription par l'assuré. A partir de cette variable, l'objectif est d'en déduire l'âge qui est un critère tarifaire. Il est important de noter que le tarif commercial d'Allianz France ne varie pas en fonction de l'âge mais en fonction de tranches d'âge (des tanches de 10 ans). Nous représentons directement l'âge selon les mêmes tranches que celles utilisées dans le tarif commercial à partir de la base de 2011 à 2017 dans le graphique ci-dessous. L'objectif de prédiction de l'année de naissance ne nécessite pas une précision à l'année près mais plutôt à la catégorie/tranche près. Il est important de remarquer qu'en terme de nombre d'observations ces tranches sont globalement équilibrée (hormis la tranche de 24 ans et moins).

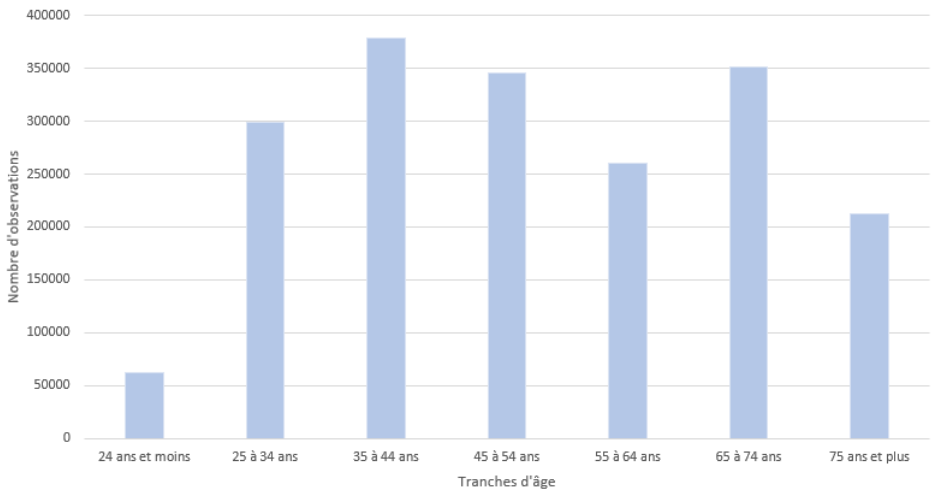

FIGURE 3.1 – Nombre d'observations par tranches d'âge dans la base de données 2011-2017

La variable nature correspond à la nature du logement, à savoir résidence principale ou secondaire. C'est une variable binaire ayant pour modalités P lorsque le logement assuré est une résidence principale et S lorsque le logement assuré est une résidence secondaire.

C'est une variable tarifaire dans le produit MRH d'Allianz France. On remarque que nous avons essentiellement des résidences principales en portefeuille.

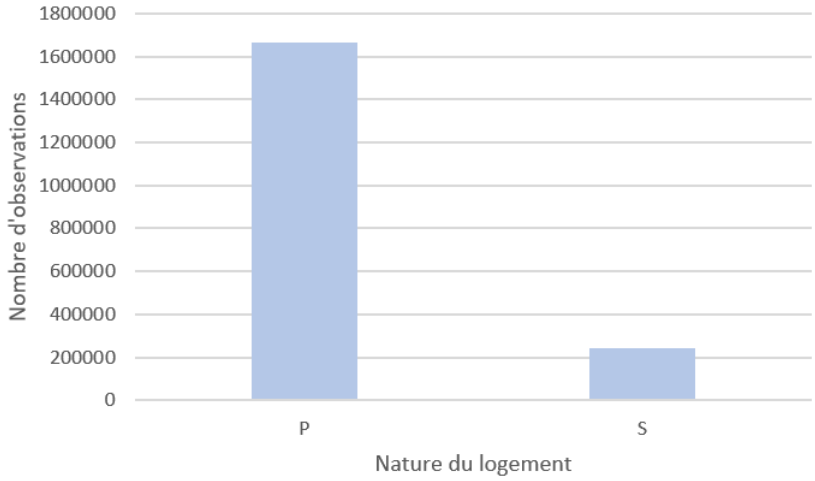

FIGURE 3.2 – Nombre d'observations par nature de logement dans la base de données 2011-2017

3.1 Théorie

L'enjeu de cette section est de présenter les différentes méthodes de*Machine Learning* mises en pratique dans ce mémoire.

3.1.1 Méthode de *Machine Learning*

3.1.1.1 Algorithme SMOTE

L'algorithme SMOTE (*Synthetic Minority Over-sampling TEchnique*) est un algorithme proposé par Niesg V. Chawla en 2002. Cet algorithme est utilisé lorsque la variable que l'on cherche à prédire a des modalités ayant un nombre d'observations non équilibré ce qui favorise le surapprentissage des modèles. Ainsi, cet algorithme va permettre de générer de manière artificielle de nouvelles observations appartenant à la classe ayant un faible nombre d'observations appelée classe minoritaire dans le cas d'une classification binaire, en s'appuyant sur la méthode des k plus proches voisins. L'objectif est de créer un nombre d'observations artificielles dans la classe minoritaire permettant d'obtenir le même nombre d'observations que dans la classe majoritaire.

Pour créer une observation dans le cas où l'ensemble des variables explicatives sont continues, la méthode des k plus proches voisins consiste à calculer la distance euclidienne entre les variables explicatives d'une observation de la classe minoritaire avec celle des autres observations de la classe minoritaire et de sélectionner les k observations ayant

la distance euclidienne la plus petite. Les k plus proches voisins de l'observation i sont ensuite ordonnés et stockés. Un nombre entier aléatoire j est ensuite généré entre 1 et k et c'est le j-ème plus proche voisin qui est alors retenu pour la suite de l'algorithme ainsi que le k-ème plus proche voisin. Un nombre entier aléatoire nommé r est ensuite généré entre 0 et 1. La nouvelle observation de la classe minoritaire s'obtient ainsi :

$$x_{art} = x_j + r * (x_k - x_j) \tag{3.1}$$

Une évolution de cet algorithme, appelé SMOTE-NC permet de traiter des problèmes de déséquilibre de données lorsque les variables explicatives sont un mélange de variables continues et catégorielles. Pour ce faire, la médiane des écarts de l'ensemble des variables continues de la classe minoritaire est calculée. Lors du calcul de la distance euclidienne, deux cas se présentent pour les variables catégorielles :

— Si les deux observations ont la même modalité, alors la valeur 0 est incluse dans le calcul de la distance euclidienne,
— Si les deux observations n'ont pas la même modalité, alors la valeur de la médiane calculée précédemment est incluse dans le calcul de la distance euclidienne.

Cependant, cet algorithme ne peut pas être appliqué si l'ensemble des variables explicatives sont catégorielles.

3.1.1.2 Arbres CART

Les arbres de décision CART (*Classification And Regression Trees*) sont des algorithmes proposés par le statisticien américain Leo Breiman en 1984 adapté aussi bien pour résoudre des problèmes de régression que de classification.

Un arbre de décision CART permet de prédire un résultat ou une classe à chaque observation de la base de donnée initiale. Chaque observation est représentée par p variables explicatives et chaque niveau de l'arbre correspond à un noeud. Le dernier noeud correspond quant à lui à une feuille. A chaque noeud de l'arbre, les modalités d'une des p variables explicatives vont permettre de séparer les observations en deux noeuds découlant du précédent. C'est la variable explicative donnant la meilleure division pour un noeud donné qui est retenue. Le partitionnement des observations se fait de telle sorte que les observations soient séparées en deux ensembles disjoints. Le critère de division utilisé varie en fonction du problème à résoudre.

Dans le cas d'une classification (par exemple pour modéliser la nature), l'objectif est d'obtenir un noeud le plus homogène possible, c'est à dire que l'ensemble des observations d'un même noeud ait la même modalité pour la variable à expliquer. Pour ce faire, c'est l'indice de Gini qui est utilisé.

Notons $\hat{p}_{ij}$ la proportion d'observations de la classe j dans le noeud i, et n le nombre de modalités différentes de la variable à expliquer Y ainsi :

$$Indice\,de\,Gini = 1 - \sum_{j=1}^{n} \hat{p}_{ij}^2 \tag{3.2}$$

Lorsque l'indice de Gini est égal à 0, l'ensemble des observations de ce noeud a alors la même variable réponse Y, on dira que le noeud est pur.

Dans le cas d'une régression (par exemple pour modéliser l'âge), on va chercher à minimiser la variance intra-groupe. Pour un noeud i contenant T observations notées $(x_1, x_2, ...x_T)$, notons $\bar{y}_i$ la moyenne des variables réponses des observations présentes dans le noeud i. Ainsi la variance intra-groupe pour le noeud i est définie de la manière suivante :

$$V(i) = \frac{1}{T} \cdot \sum_{t=1}^{T} (y_t - \bar{y}_i)^2 \tag{3.3}$$

Les différents critères d'arrêts de division de l'arbre sont :

— Le noeud est pur,

— Le nombre d'observations minimales pour un noeud est atteint (ce nombre est un paramètre défini dans l'algorithme).

Le point fort de l'algorithme d'arbres CART est que le temps machine est rapide, cependant l'une des grandes faiblesses de cet algorithme est qu'il présente une forte variance de prédiction.

3.1.1.3 *Random Forest*

L'algorithme des forêts aléatoires a été proposé par les statisticiens américains Leo Breiman et Adèle Cutler en 2001. Cet algorithme combine plusieurs arbres de décisions avec une approche de *Boostrap Aggregation* aussi appelé *Bagging*. L'algorithme des forêts aléatoires va entraîner plusieurs arbres de décision et chaque arbre de décision sera entraîné sur un sous ensemble différent parmi l'échantillon d'apprentissage afin de réduire la variance de prédiction. En effet les prédictions obtenues par l'algorithme des arbres de décisions varient de manière importante lors de petites modifications dans l'échantillon d'apprentissage.

L'algorithme des forêts aléatoires se décompose en plusieurs étapes. Notons l'échantillon d'apprentissage $Z = (X_i, Y_i)_{i\in[1:n]}$ avec X un vecteur de taille p correspondant aux variables explicatives quantitatives et qualitatives et Y correspondant à la variable réponse. La première étape, appelée *Tree Bagging*, consiste à réaliser B tirages avec remise de n observations sur l'échantillon d'apprentissage. Ainsi, B échantillons seront obtenus à la fin de cette première étape. Le nombre de tirages B est un paramètre de l'algorithme.

La deuxième étape, appelée *Feature Sampling* consiste à réaliser un tirage aléatoire de taille m sur les variables explicatives pour chaque échantillon.

Un arbre de décision est ensuite entraîné pour chaque échantillon précédemment réalisé. Chaque échantillon est donc différent par ses observations ainsi que ses variables explicatives. Pour un échantillon donné b, la prédiction obtenue sera notée $\hat{f}_b(x)$ et dépendra de $x = (x^1, ..., x^m)$, les m variables explicatives sélectionnées pour cet échantillon.

Les B prédictions obtenues sont ensuite récupérées et la prédiction finale de l'algorithme des forêts aléatoires s'obtient :

— Dans le cas d'une régression, en faisant la moyenne des variables réponses obtenues :

$$\hat{f_B} = \frac{1}{B} \sum_{b=1}^{B} \hat{f}_b(x)$$

— Dans le cas d'une classification, en prenant la classe qui a été prédite en majorité, appelée vote majoritaire :

$$\hat{f_B} = \underset{j}{\operatorname{argmax}}\, card\{b \mid \hat{f}_b(x) = j\}$$

L'algorithme *Random Forest* à un temps machine plus important que les arbres CART mais il permet de diminuer la variance de prédiction grâce au nombre important d'arbres indépendants construits.

3.1.1.4 *Gradient Boosting*

L'algorithme du *Gradient Boosting* (GBM) repose sur des méthodes de *Boosting* d'arbres de décision. A la différence du *Bagging*, le *Boosting* construit ses arbres de décision par récurrence. Chaque arbre créé prend en compte la prédiction de l'arbre précédent en pénalisant les observations dont la prédiction est fausse. Le premier arbre est construit sur l'échantillon d'apprentissage comprenant n observations, chaque observation se voit attribuer un poids qui est dans un premier temps identique. Ensuite chaque nouvel arbre construit attribue des poids différents aux observations en fonction des prédictions de l'arbre précédent. Plus une observation a une prédiction mauvaise et plus sa pondération dans l'arbre suivant sera élevée. Le nombre d'itérations, c'est à dire le nombre d'arbres construits est un paramètre du modèle.

L'idée générale du *Gradient Boosting* est de construire des arbres successifs avec comme objectif d'améliorer l'erreur initiale en se basant sur la descente du gradient de la fonction de perte pour améliorer les erreurs.

Notons l'échantillon d'apprentissage $(x_i, y_i)_{i\in[1:n]}$ correspondant respectivement à un vecteur de taille p représentant les variables explicatives et à la variable réponse et notons $L(y_i, F(x))$ la fonction de perte associée. L'algorithme de *Gradient Boosting* se déroule de la manière suivante :

1. Initialisation du modèle par une valeur constante :

$$F_0(x) = \underset{\gamma}{\operatorname{argmin}} \sum_{i=1}^{n} L(y_i, \gamma)$$

.

2. Pour m = 1 à M (M est le nombre d'arbres fixé comme paramètre du modèle)

   (a) Calculer les pseudos-résidus pour chaque observation i de la base d'apprentissage :

$$r_{i,m} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x))}$$

   (b) Construire un arbre de régression CART avec comme variable réponse les pseudos-résidus. On obtient alors J_m feuilles (noeuds terminaux).
   $\forall j \in [1 : J_m], R_{j,m}$ est la j-ième feuille de l'arbre m.

   (c) Pour tout $j \in [1 : J_m]$, on va calculer $\gamma_{j,m}$ en résolvant :

$$\gamma_{j,m} = \underset{\gamma}{\operatorname{argmin}} \sum_{x_i \in R_{i,j}} L(y_i, F_{m-1}(x_i) + \gamma)$$

   (d) Le modèle devient :

$$F_m(x) = F_{m-1}(x) + \nu \sum_{j=1}^{J_m} \gamma_{j,m} \mathbb{1}(x \in R_{j,m})$$

Bien que l'algorithme du *Gradient Boosting* permette de réduire le biais ainsi que la variance de prédiction, sa principale limite est le risque de surapprentissage.

3.1.2 Paramétrages des modèles

La base de données totale est séparée en deux :
— Une base d'apprentissage correspondant à 80 % de la base de données totale sur laquelle le modèle va apprendre,
— Une base de test correspondant à 20 % de la base de données totale sur laquelle les prédictions vont être réalisées.

Lors de la réalisation des modèles sur Python, des hyperparamètres doivent être renseignés. Afin de trouver les hyperparamètres optimaux pour un modèle donné, la méthode de la *cross validation* (validation croisée) est utilisée. Cette méthode consiste à découper la base d'apprentissage en k sous bases (*k-folds*) de manière aléatoire. Pour chaque combinaison d'hyperparamètres que l'on souhaite tester, nous allons entraîner un modèle sur *k-1 folds* et le valider (c'est à dire calculer une erreur de prédiction dans le cas d'une régression, ou un taux de bien prédit dans le cas d'une classification) sur le *k-fold* restant. Le processus est répété jusqu'à que chaque *k-fold* soit utilisé comme base de validation. La moyenne des erreurs de prédiction ou de taux de bien prédit sur les *k-folds* sera retenue pour cette combinaison d'hyperparamètres. Les paramètres optimaux correspondant a ceux pour lesquels la moyenne des RMSE est la plus petite dans le cas d'une régression et ceux pour lesquels le F1-score est le plus grand pour les modèles de classification. L'ensemble de ces métriques seront détaillés dans la suite de ce chapitre.

Le schéma ci-dessous présente la *cross validation* dans le cas où k = 5. C'est ce k qui a été utilisé pour les *cross validation* de ce mémoire.

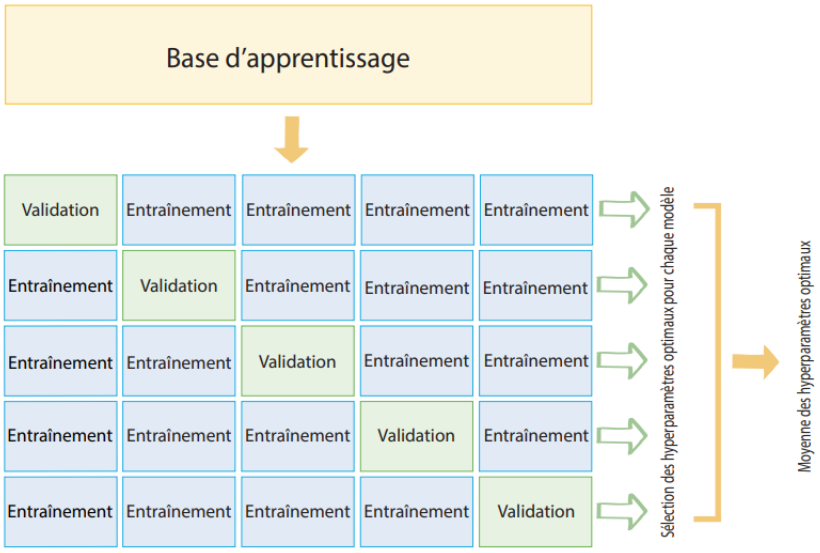

FIGURE 3.3 – *Cross validation* pour k = 5

Ainsi, une fois les hyperparamètres optimaux sélectionnés, le modèle sera entraîné sur l'ensemble de la base d'apprentissage et testé sur la base de test.

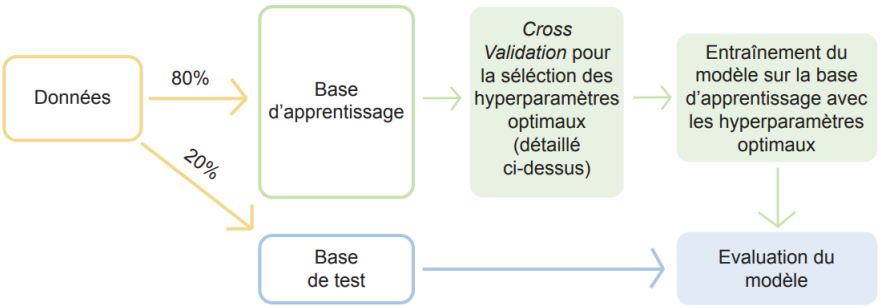

FIGURE 3.4 – Schéma récapitulatif des étapes du *Machine Learning*

Les différent hyperparamètres testés par *cross validation* pour les modèles *Random Forest* de classification et de régression sont :

- Le *min_samples_split*, correspondant au nombre d'observations minimal pour qu'un noeud soit divisé, a été testé entre 2 et 20 avec un pas de 2.
- Le *n_estimators*, correspondant au nombre d'arbres a été testé entre 50 et 500 avec un pas de 20.

Les différents hyperparamètres testés par *cross validation* pour les modèles *Gradient Boosting* de classification et de régression sont :

- Le *min_samles_split*, correspondant au nombre d'observations minimal pour qu'un noeud soit divisé a été testé entre 2 et 20 avec un pas de 2,
- Le *n_estimators*, correspondant au nombre d'arbres a été testé entre 50 et 500 avec un pas de 20,
- Le *learning_rate*, correspondant au coefficient ν du modèle qui permet de réduire la contribution de chaque arbre, a été testé entre 0,01 et 0,1 avec un pas de 0,01,
- La *max_depth*, correspondant à la profondeur maximale de l'arbre, a été testé entre 8 et 32 avec un pas de 8.

3.1.3 Évaluation des modèles

Une fois le modèle entraîné sur la base d'apprentissage, il est utilisé pour faire des prédictions sur la base de test. Afin de quantifier les performances du modèles, différentes métriques vont être calculées sur la base de test.

3.1.3.1 Régression

Afin de valider et comparer les modèles de *Machine Learning* utilisant la régression, la RMSE (*Root Mean Square Error*) à été utilisée. L'objectif est de minimiser la RMSE.

Notons n, la taille de l'échantillon, Y_i la valeur de l'observation i et $\hat{Y}_i$ la prédiction de Y_i par le modèle, ainsi :

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}} \tag{3.4}$$

3.1.3.2 Classification binaire

Par définition, la classification binaire correspond à une classification pour laquelle les observations sont divisées en deux classes. Les deux classes seront notées : 0 (négatif) et 1 (positif).

La matrice de confusion ci-dessous permet de visualiser les prédictions obtenues en fonction des observations.

- Vrai positif (VP) : Une observation positive a été prédite positive
- Vrai négatif (VN) : Une observation négative a été prédite négative
- Faux positif (FP) : Une observation négative a été prédite positive
- Faux négatif (FN) : Une observation positive a été prédite négative

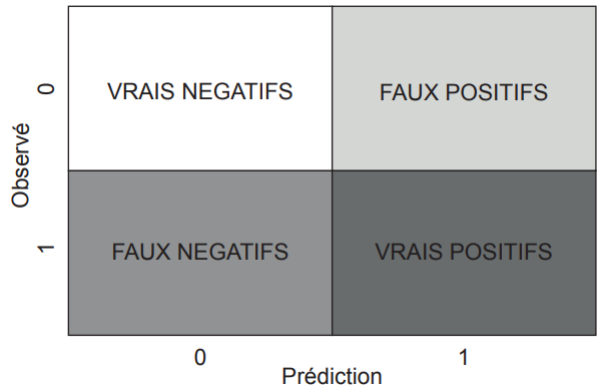

FIGURE 3.5 – Matrice de confusion pour une classification binaire

A partir de la matrice de confusion, quatre métriques vont être calculées afin d'évaluer la performance du modèle :

— **L'exactitude** correspondant à la probabilité qu'une observation soit correctement prédite :

$$excatitude = \frac{VP + VN}{VP + VN + FP + FN}$$

— **La précision** correspondant à la probabilité qu'une observation positive soit correcte :

$$precision = \frac{VP}{VP + FP}$$

— **La spécificité** correspond à la probabilité qu'une observation négative soit correctement prédite

$$specificite = \frac{VN}{VN + FP}$$

— **La sensibilité** correspond à la probabilité qu'une observation positive soit correctement prédite

$$sensibilite = \frac{VP}{VP + FN}$$

Deux autres métriques sont généralement utilisés pour quantifier la qualité de prédiction d'un modèle de classification binaire.

La courbe ROC et l'AUC
L'ensemble des indicateurs calculés précédemment dépendent d'un seuil de classification donné. En général ce seuil est fixé à 0,5 dans Python. La courbe ROC représente pour chaque seuil de classification entre 0 et 1, le taux de vrais positif (sensibilité) en fonction du taux de faux positifs (1 – spécificité). L'AUC mesure l'aire sous la courbe ROC, elle permet de quantifier la qualité de prédiction. Plus l'AUC sera proche de 1 et meilleure sera la prédiction. En effet, la valeur 1 est obtenue dans le cas où il n'y a aucun faux positif.

La courbe Précision-Rappel (PR) et l'AUC PR
La courbe PR est utilisée dans le cas où il y a un déséquilibre des observations entre les deux classes observées. La courbe PR représente pour chaque seuil de classification entre 0 et 1, la sensibilité (que l'on appelle aussi rappel) en fonction de la précision. Contrairement à la sensibilité et à la spécificité utilisées pour la courbe ROC, le calcul de la précision prend en compte les déséquilibres de données. L'AUC PR représente l'aire sous la courbe Précision-Rappel et permettra de quantifier la qualité de prédiction d'un jeu de données déséquilibré.

3.1.3.3 Classification multiclasses

On parle de classification multiclasses lorsque les observations sont réparties dans un nombre de classes supérieure à 2.

Soient $X_1, ...X_n$, les n classes de la variable que l'on cherche à prédire. Pour tout $i \in [1 : n]$ et $j \in [1 : n]$, $M_{i,j}$ correspond au nombre d'observations de la classe X_i prédites dans la classe X_j.

La matrice de confusion ci-dessous permet de visualiser les prédictions des classes obtenues en fonction des classes observées.

Prédit \ Observé	X_1	X_2	...	X_n
X_1	$M_{1,1}$	$M_{1,2}$	...	$M_{1,n}$
X_2	$M_{2,1}$	$M_{2,2}$	...	$M_{2,n}$
...	...	...	...	...
X_n	$M_{n,1}$	$M_{n,2}$	...	$M_{n,n}$

TABLE 3.1 – Matrice de confusion dans le cas d'une classification multiclasses

Afin de calculer les métriques correspondant à l'ensemble de la base, trois sous métriques par classe sont calculées dans un premier temps :

1. Calcul de la précision qui correspond pour une classe donnée i au nombre de bien prédit de la classe i par rapport au nombre de prédits de la classe i.

$$precision_i = \frac{M_{i,i}}{\sum_{j=1}^{n} M_{i,j}} \tag{3.5}$$

2. Calcul de la spécificité qui correspond pour une classe donnée i au nombre de bien prédit i par rapport au nombre observé de la classe i

$$sensibilite_i = \frac{M_{i,i}}{\sum_{j=1}^{n} M_{j,i}} \tag{3.6}$$

3. Calcul du *F1_score* pour la classe i qui correspond à la moyenne harmonique entre la précision et la spécificité de la classe i

$$F1_score_i = \frac{2 * (sensibilite_i * precision_i)}{sensibilite_i + preicison_i} \tag{3.7}$$

Les trois métriques permettant d'évaluer la performance du modèle total sont ensuite calculées :

- **L'exactitude**, correspondant au nombre de biens prédits exacts total par rapport au nombre d'observations totales dans l'échantillon.

$$exactitude = \frac{\sum_{i=1}^{n} M_{i,i}}{\sum_{i=1}^{n}\sum_{j=1}^{n} M_{i,j}} \tag{3.8}$$

- **Le F1 score** correspondant à la moyenne des F1_score par classe.

$$F1_score = \frac{1}{n} \cdot \sum_{i=1}^{n} [F1_score_i] \tag{3.9}$$

- **Le F1 score pondéré** correspondant à la moyenne pondérée par le nombre d'observation par classe des F1_score par classe afin de prendre en compte les déséquilibres dans les données.

$$F1_score_pondere = \frac{1}{\sum_{i=1}^{n}\sum_{j=1}^{n} M_{i,j}} \cdot \sum_{i=1}^{n} \left[F1_score_i * \sum_{j=1}^{n} M_{j,i}\right] \tag{3.10}$$

Lorsque les classes sont ordonnées, il est possible de calculer l'exactitude à une classe près. Le taux de biens prédits total à une classe près *exactitudeT1* se calcule de la manière suivante :

$$Pour\ i\ =\ 1 : exactitudeT1_i = \frac{M_{1,1} + M_{2,1}}{\sum_{j=1}^{n} M_{j,i}} \tag{3.11}$$

$$Pour\ i\ =\ n : exactitudeT1_i = \frac{M_{n-1,n} + M_{n,n}}{\sum_{j=1}^{n} M_{j,i}} \tag{3.12}$$

$$Pour\ i \in [2 : n-1] : exactitudeT1_i = \frac{M_{i-1,i} + M_{i,i} + M_{i+1,i}}{\sum_{j=1}^{n} M_{j,i}} \tag{3.13}$$

Ainsi, pour le modèle total, l'exactitude à une classe près vaut :

$$exactitudeT1 = \frac{\sum_{i=1}^{n} exacitudeT1_i}{n} \tag{3.14}$$

3.2 Modélisation de la nature du logement

La nature du logement dispose de deux modalités qui correspondent aux résidences principales et aux résidences secondaires, c'est donc une classification binaire qui est réalisée sur cette variable.

Les modalités vont être renommées comme ci-dessous lors de la mise en place des méthode de *Machine Learning*.

- 0 (négatif) : Le logement assuré est une résidence principale
- 1 (positif) : Le logement assuré est une résidence secondaire

Ainsi, nous obtenons :

- Vrai positif (VP) : Un logement secondaire a été prédit comme secondaire
- Vrai négatif (VN) : Un logement principal a été prédit comme principal
- Faux positif (FP) : Un logement principal a été prédit comme secondaire

- Faux négatif (FN) : Un logement secondaire a été prédit comme principal

Pour rappel, les variables explicatives sont celles collectées dans le chapitre 2. Parmi ces données il y a des données internes telles que le nombre de pièces et des données externes telles que le taux de résidences principales.

3.2.1 Comparaison et validation des modèles de *Machine Learning*

Nous allons modéliser la variable nature à l'aide de l'algorithme *Random Forest* et d'un *Gradient boosting*. L'objet de cette section et de comparer ces deux méthodes et de définir l'approche optimale.

Random Forest

Lors de la *cross validation*, les hyperparamètres retenues pour ce modèle sont :

- $min_samles_split = 2$
- $n_estimators = 200$

La matrice de confusion obtenue pour le modèle *Random Forest* avec les hyperparamètres optimaux est représentée dans le schéma ci-dessous.

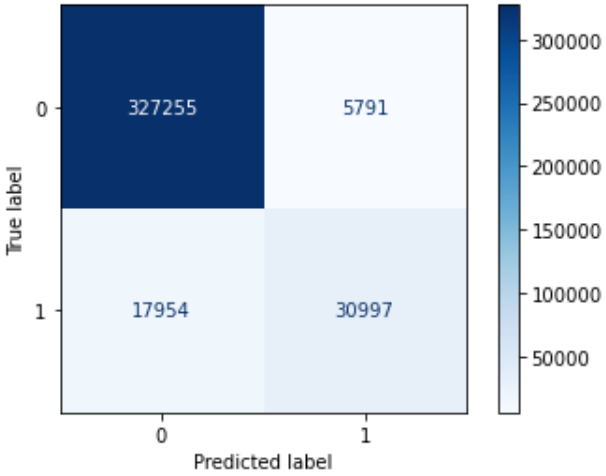

TABLE 3.2 – Matrice de confusion pour le *Random Forest*

$$exactitude = \frac{VP + VN}{VP + VN + FP + FN} = 93,78\%$$

$$precision = \frac{VP}{VP + FP} = 84,26\%$$

$$specificite = \frac{VN}{VN + FP} = 98,26\%$$

$$sensibilite = \frac{VP}{VP + FN} = 63,32\%$$

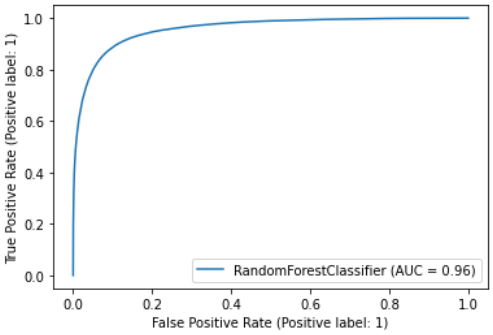

FIGURE 3.6 – Courbe ROC *Random Forest*

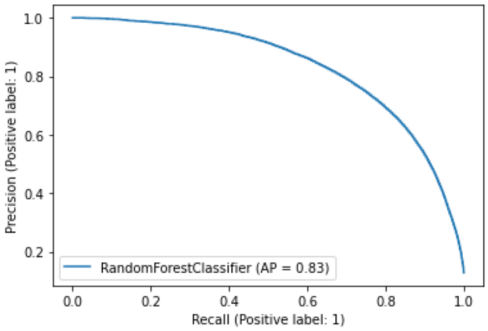

FIGURE 3.7 – Courbe Précision-Rappel *Random Forest*

L'AUC de ce modèle vaut 0,96 et l'AUC PR vaut 0,83.

Python calcule l'importance de chaque variable explicative dans le modèle créé. La base de données ayant un nombre important de variables, seules les 15 plus importantes seront renseignées dans le graphique ci-dessous.

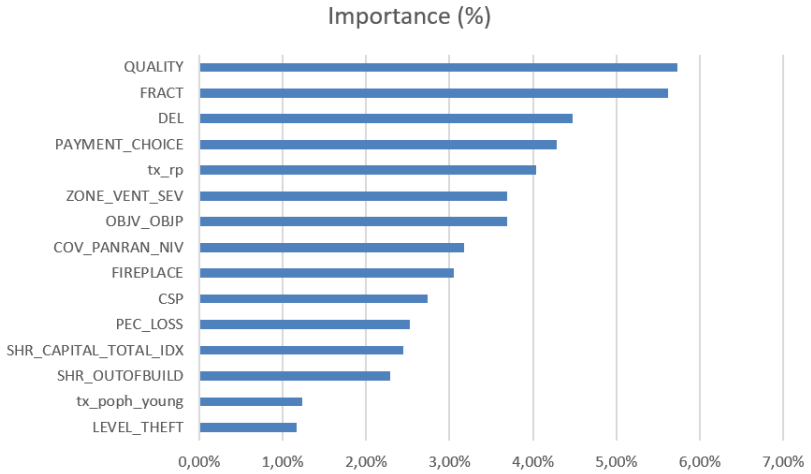

TABLE 3.3 – Importance des variables dans le modèle *Random Forest* pour la prédiction de la nature

Gradient Boosting

Les hyperparamètres retenus pour ce modèle sont :

- $min_samles_split = 10$
- $n_estimators = 500$
- $learning_rate = 0{,}04$
- $max_depth = 32$

La matrice de confusion obtenue pour le modèle de *Gradient Boosting* avec les hyper-paramètres optimaux est représentée dans le schéma ci-dessous.

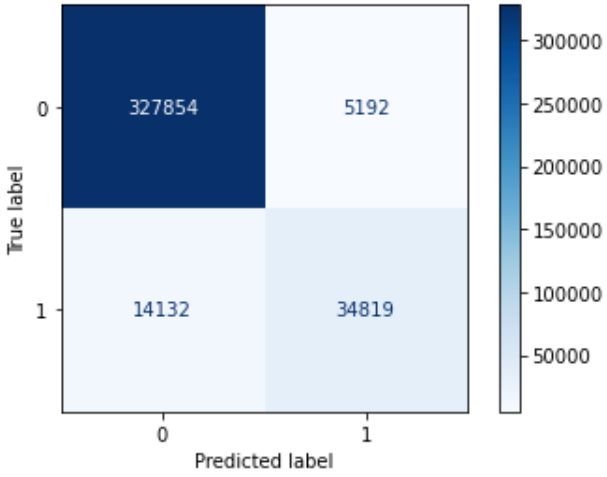

TABLE 3.4 – Matrice de confusion pour le *Gradient Boosting*

$$exactitude = \frac{VP + VN}{VP + VN + FP + FN} = 94,94\%$$

$$precision = \frac{VP}{VP + FP} = 87,02\%$$

$$specificite = \frac{VN}{VN + FP} = 98,44\%$$

$$sensibilite = \frac{VP}{VP + FN} = 71,13\%$$

L'AUC de ce modèle vaut 0,97 et l'AUC PR vaut 0,88.

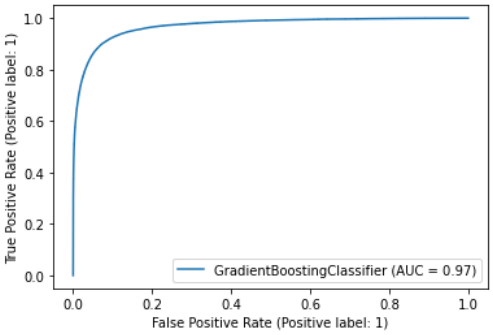

FIGURE 3.8 – Courbe ROC *Gradient Boosting*

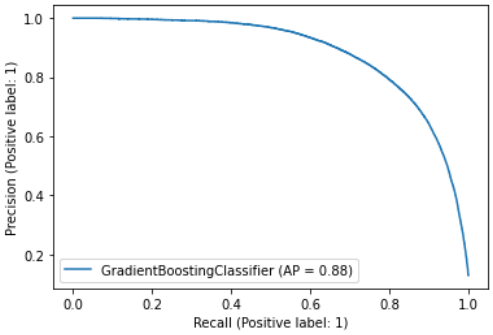

FIGURE 3.9 – Courbe Précision-Rappel *Gradient Boosting*

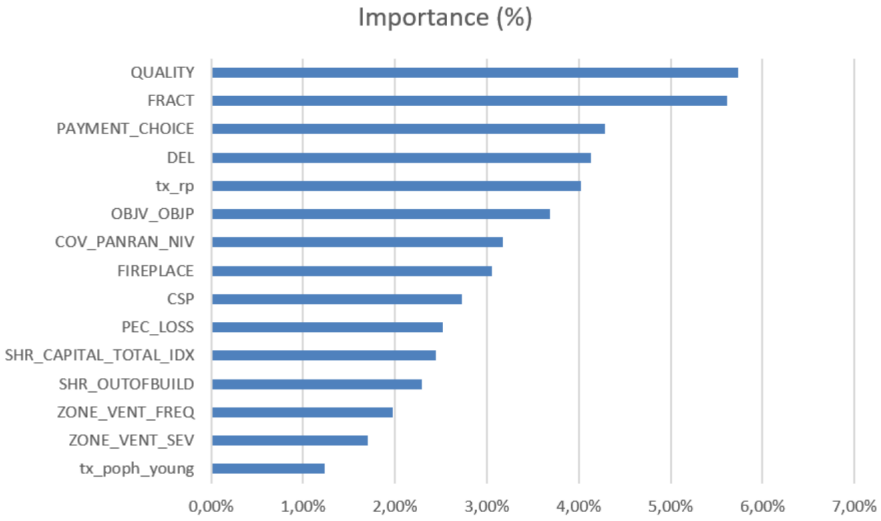

FIGURE 3.10 – Importance des variables dans le modèle *Gradient Boosting*

3.2.2 Interprétation des résultats

En regardant les résultats des deux modèles, on observe une exactitude de prédiction de 93,78 % pour le modèle *Random Forest* et de 94,94 % pour le modèle *Gradient Boosting*. Les deux résultats sont très élevés ce qui peut refléter une prédiction performante. Cependant, le calcul de l'exactitude ne prend pas en compte le déséquilibre des données observé pour la variable nature puisqu'il se focalise sur les bonnes prédictions au global. Ainsi, le modèle peut surapprendre et prédire uniquement la classe majoritaire tout en ayant une exactitude de prédiction élevée. Il est donc préférable de regarder des métriques telles que la précision et l'AUC PR. La précision vaut 84,26 % pour le modèle RF [1] et

1. *Random Forest*

87,02 % pour le modèle GBM [2]. L'AUC PR vaut 0,83 pour le modèle RF et 0,88 pour le modèle GBM. Ces deux métriques reflètent deux modèles performants pour la prédiction de la variable nature, bien que le modèle GBM soit plus performant que le modèle RF.

Les variables d'importances des deux modèles se rejoignent. La variable la plus importance pour les deux modèles est la variable qualité juridique (propriétaire, locataire,...). Cela est cohérent, en effet les résidences secondaires ont tendance à être plus des propriétés que des locations. Dans les deux modèles on observe parmi les variables les plus importantes, les variables externes *tx_rp* et *tx_poh_youg*, représentant respectivement le taux de résidences principales par zone géographique et le taux de population jeune par zone géographique. Ainsi, les données externes ont permis une meilleure performance des modèles. Ces variables sont cohérentes avec la variable que l'on cherche à prédire, en effet la variable correspondant au taux de résidences principales reflète directement la variable nature et la variable sur le taux de population jeune dans une zone géographique s'explique par le fait que les jeunes ont moins tendance à acheter des résidences secondaires.

Ainsi, on retiendra le modèle GBM qui est le plus performant et qui permet de prédire la variable latente nature de manière satisfaisante. Cette variable latente sera ensuite prédite sur la base de 2018 à 2019 et testée dans les modélisations du chapitre suivant.

3.3 Modélisation de l'âge de l'assuré

La modélisation de l'âge de l'assuré s'est faite en prédisant l'année de naissance de l'assuré puis en calculant la différence entre l'année de la date de début de période et l'année de naissance.

Il est nécessaire de noter que le tarif commercial ne varie pas énormément d'une classe d'âge voisine à l'autre. Toute chose égale par ailleurs, pour une classe d'âge donnée, le tarif commercial appliqué pour deux personnes ayant un âge appartement à cette classe sera identique, et celui pour deux personnes ayant un âge différent d'une classe sera très proche. Il peut donc être intéressant de calculer le taux de biens prédits global exact mais aussi le taux de biens prédits à une classe près. En effet, l'âge a été prédit à partir d'une régression sur l'année de naissance de l'assuré puis réparti sous forme de tranches d'âge. Ainsi, les tranches sont ordonnées.

3.3.1 Comparaison et validation des modèles de *Machine Learning*

Random Forest

De même que pour le modèle de prédiction de la nature, différents hyperparamètres ont été testés par la méthode de *cross validation* présentée précédemment.

2. *Gradient Boosting Machine*

Les hyperparamètres retenus pour ce modèle sont :

- $min_samles_split = 2$
- $n_estimators = 250$

Une fois les années de naissance prédites par le modèle, les âges des assurés ont été calculés. La RMSE obtenue pour l'algorithme de RF vaut 2,58. Les âges ont ensuite été répartis dans des classes correspondants à la segmentation utilisée pour le tarif commercial d'Allianz France et différentes métriques ont été calculées par classe ainsi que sur le modèle global.

		Observé						
		24 ans et moins	25 à 34 ans	35 à 44 ans	45 à 54 ans	55 à 64 ans	65 à 74 ans	75 ans et plus
Prédit	24 ans et moins	4179	80	6	4	6		1
	25 à 34 ans	4370	31824	2298	1398	626	92	37
	35 à 44 ans	3480	22989	60411	8018	3952	1372	438
	45 à 54 ans	726	4333	4771	61283	16127	3770	1107
	55 à 64 ans	59	432	1304	4471	42092	4592	849
	65 à 74 ans	14	43	92	392	5269	35603	16743
	75 ans et plus	11	2	13	82	2175	6733	23328

TABLE 3.5 – Matrice de confusion pour la prédiction des classes d'âge par *Random Forest*

	24 ans et moins	25 à 34 ans	35 à 44 ans	45 à 54 ans	55 à 64 ans	65 à 74 ans	75 ans et plus
Précision	97,73%	78,30%	60,01%	66,53%	78,24%	61,22%	72,12%
Spécificité	32,55%	53,30%	87,69%	81,01%	59,92%	68,25%	54,89%
F1_score	48,83%	63,43%	71,26%	73,06%	67,87%	64,55%	62,34%

TABLE 3.6 – Métriques calculées pour chaque classe d'âge sur la prédiction *Random Forest*

Trois métriques sur la performance globale du modèle ont été calculées :

$$exactitude = 67,73\%$$

$$exactitudeT1 = 89,90\%$$

$$F1_score = 64,47\%$$

$$F1_score_pondere = 67,86\%$$

Comme évoqué précédemment, Python calcule l'importance de chaque variable explicative dans le modèle créé. La base de données ayant un nombre important de variables, seules les douze plus importantes seront renseignées dans le graphique ci-dessous.

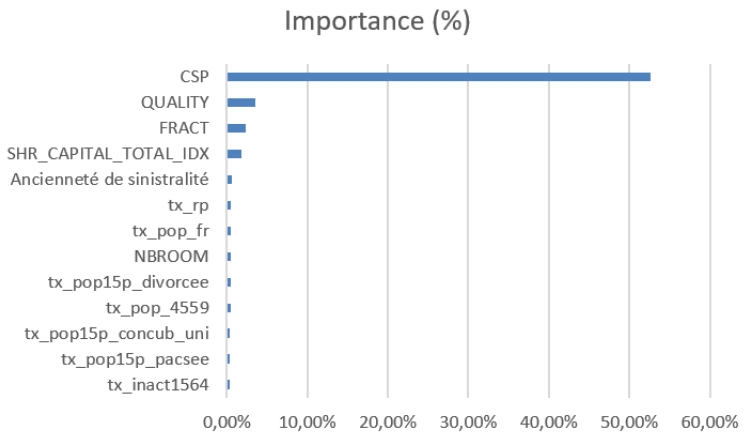

FIGURE 3.11 – Variables importantes du modèle de *Random Forest* pour la prédiction de l'âge

Gradient Boosting

Pour le GBM, les hyperparamètres retenus sont :

- $min_samles_split = 10$
- $n_estimators = 300$
- $learning_rate = 0{,}04$
- $max_depth = 32$

LA RMSE obtenue pour ce modèle vaut 2,42.

Prédit \ Observé	24 ans et moins	25 à 34 ans	35 à 44 ans	45 à 54 ans	55 à 64 ans	65 à 74 ans	75 ans et plus
24 ans et moins	4969	221	32	44	16	0	0
25 à 34 ans	3599	32442	1584	2014	907	123	48
35 à 44 ans	3315	22334	61963	9381	4120	1235	334
45 à 54 ans	865	4052	4563	60642	13900	3475	935
55 à 64 ans	67	599	650	3112	44335	3480	1011
65 à 74 ans	17	50	91	387	5681	36963	14599
75 ans et plus	7	5	12	68	1288	6886	25576

TABLE 3.7 – Matrice de confusion pour la prédiction des âges par *Gradient Boosting*

	24 ans et moins	25 à 34 ans	35 à 44 ans	45 à 54 ans	55 à 64 ans	65 à 74 ans	75 ans et plus
Précision	94,07%	79,68%	60,34%	68,57%	83,25%	63,96%	75,57%
Sensibilité	38,70%	54,34%	89,94%	80,16%	63,11%	70,86%	60,17%
F1_score	54,84%	64,61%	72,23%	73,92%	71,80%	67,24%	67,00%

TABLE 3.8 – Métriques calculées pour chaque classe d'âge prédites par *Gradient Boosting*

Comme pour le *Random Forest*, trois métriques sur la performance globale du modèle ont été calculées :

$$exactitude = 69,87\%$$

$$exactitudeT1 = 90,09\%$$

$$F1_score = 67,38\%$$

$$F1_score_pondere = 70,42\%$$

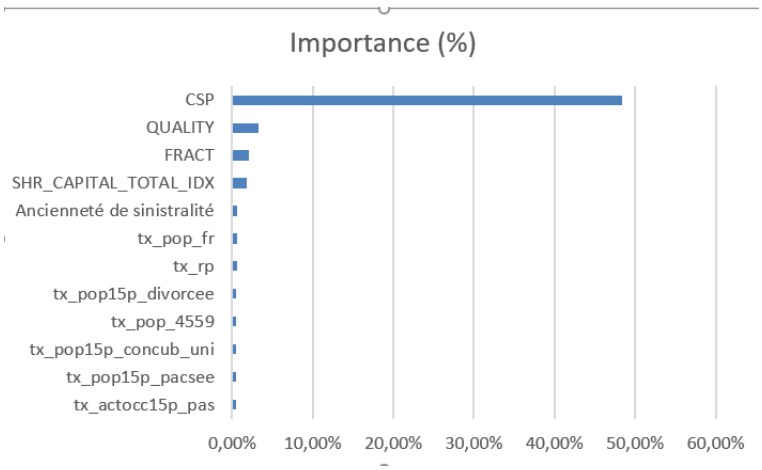

TABLE 3.9 – Importance des variables dans le modèle GBM

3.3.2 Interprétation des résultats

La RMSE obtenue pour le modèle de régression GBM vaut 2,42 et est ainsi inférieure à celle obtenue pour le modèle de régression RF. L'objectif étant de minimiser la RMSE, c'est le GBM qui est le plus performant en regardant cette métrique.

On observe que l'exactitude des modèles vaut respectivement 67,33 % pour le modèle RF et 69,87 % pour le modèle GBM. Ces résultats ont été tirés vers le bas par la classe "24 ans et moins" qui a un nombre d'observations plus faible que les autres classes et donc les modèles ont légèrement surappris sur les autres classes. Lorsque l'on regarde l'exactitude à une classe près, on observe 89,90 % pour le modèle RF et 90,09 % pour le modèle GBM. Cette métrique est très élevée, ainsi les modèles permettront de prédire de manière performante l'âge à une classe près. L'ensemble des métriques étant supérieur pour le modèle GBM, c'est ce dernier qui va être retenu pour la prédiction de l'âge.

Les graphiques représentant les variables importantes du modèle, c'est à dire les variables qui ont le plus été utilisées par le modèle pour sa prédiction fait ressortir dans les deux modèles la CSP. En effet, l'une des modalités de la CSP est la modalité retraitée qui correspond aux assurés étant à la retraite. Ainsi, cette modalité permet d'apporter une information capitale sur l'âge de l'assuré puisque par définition un retraité aura plus de 62 ans. La deuxième modalité est la qualité juridique (propriétaire, locataire,...), cette variable est cohérente, en effet, les personnes jeunes ont moins tendance à être propriétaires, ce qui peut être une information importante pour la prédiction de l'âge par le modèle. Parmi les variables externes, bien qu'elles soient bien moins importantes pour le modèle que les variables internes citées précédemment, il y a le taux de personnes pacsées pour une zone géographique donnée. Cela permet d'apporter une précision sur l'âge de l'assuré puisque les personnes pacsées sont en général des personnes jeunes qui se pacsent dans un premier temps, avant de s'engager par le mariage. En effet d'après l'INSEE, 20% des personnes en couple entre 25 et 34 ans sont pacsées, 5% chez des personnes en couple entre 45 et 59 ans et enfin moins de 1% pour les personnes en couple de plus de 60 ans.

Le modèle retenu pour la prédiction de l'âge est donc le modèle GBM. Les années de naissance vont être ensuite prédites pour l'ensemble du portefeuille de 2018 et 2019 afin de tester cette nouvelle variable latente dans la modélisation qui sera faite dans le chapitre suivant. Dans la modélisation, c'est la variable continue âge qui est prise bien que la grille tarifaire commerciale tarifaire varie en fonction de classes d'âge.

Chapitre 4

Modélisation GLM

Cette quatrième partie s'inscrit dans la continuité de la précédente avec le même objectif qui est de réduire le nombre de questions à la souscription en intégrant des données externes. La troisième partie de ce mémoire avait pour objectif de prédire directement des données latentes à partir de données internes et externes. Cette deuxième approche va être de comparer la performance de différents modèles GLM pour prédire la fréquence et le coût moyen pour la garantie dégâts des eaux sur une période de 2018 à 2019. Pour rappel, l'année 2020 est retirée de la modélisation à cause de son caractère atypique. Les chiffres présentés dans cette partie ont été retraités pour des raisons de confidentialité.

Pour ce faire, six modèles vont être réalisés :
- Un modèle de coût moyen et un modèle de fréquence avec uniquement des données internes,
- Un modèle de coût moyen et un modèle de fréquence avec uniquement des données externes,
- Un modèle de coût moyen et un modèle de fréquence avec à la fois des données internes et des données externes.

L'objectif de cette deuxième approche est multiple. Il s'agit d'abord de répondre à la question de savoir si un modèle composé uniquement de données externes est aussi voire plus efficace qu'un modèle avec des données internes. Le cas échéant, de démontrer que certaines données internes peuvent être substituées par des données externes sans nuire à la robustesse des modèles tarifaires. La modélisation permettra aussi de tester les variables latentes prédites sur 2018 et 2019 à partir des modèles créés dans le chapitre précédent.

Les GLM ont été implémentés sur *Emblem*, un logiciel de tarification utilisé par Allianz France. Avant d'implémenter les GLM sur *Emblem*, la base totale a été séparée en deux sous bases :
- Une base d'apprentissage correspondant à 80 % de la base totale,
- Une base de test correspondant à 20 % de la base totale.

L'ensemble des modélisations est réalisée sur la base d'apprentissage et sera ensuite testé sur la base de test.

4.1 Théorie des GLM

Les GLM ont été établis par Jonh Nelder et Robert Wedderburn en 1972 et ont pour objectif d'étudier les relations entre l'espérance d'une variable aléatoire réponse Y$\in \mathbb{R}^p$ et un ensemble de n variables explicatives $X = (X_1, ..., X_k)$. La loi de probabilité de la variable à expliquer Y doit appartenir à la famille exponentielle. Ils nous permettent d'exprimer $\mathbb{E}[Y|X = x]$. L'une des hypothèses fortes du GLM est que les variables explicatives du modèle ne doivent pas être corrélées.

4.1.1 Composantes d'un GLM

Les GLM sont composés de trois éléments :
- Une composante aléatoire,
- Une composante déterministe,
- Une fonction de lien.

La composante aléatoire
La composante aléatoire correspond à la variable aléatoire Y. Comme évoqué précédemment, la loi de probabilité associée à la variable aléatoire Y que l'on notera $f_{\theta,\phi}$ doit appartenir à la famille exponentielle, c'est à dire que pour tout y, $f_{\theta,\phi}$ est de la forme :

$$f_{\theta,\phi}(y) = exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right] \tag{4.1}$$

où a,b et c sont des fonctions connues et dérivables telle que :
- b est trois fois dérivable et inversible
- $\theta \in \mathbb{R}$ est le paramètre naturel de la famille exponentielle lié aux deux premiers moments et $\theta \in \mathbb{R}_*^+$ est le paramètre de dispersion

La composante déterministe La composante déterministe permet d'associer à chaque observation $i \in \{1, ..., k\}$ un prédicteur linéaire que l'on notera η_i tel que :

$$\forall i \in \{1, ...k\},\ \eta_i = \beta_0 + \beta_1 \cdot x_{i,1} + ... + \beta_p \cdot x_{i,p} + \epsilon_i \tag{4.2}$$

La fonction de lien
La fonction de lien g est une fonction monotone et dérivable qui permet de relier la variable réponse à l'ensemble des variables explicatives telle que :

$$g(\mathbb{E}[Y|X = x]) = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k + \epsilon \tag{4.3}$$

Ce qui équivaut à :

$$\mathbb{E}[Y|X = x] = g^{-1}(\beta_0 + \beta_1 X_1 + ... + \beta_k X_k + \epsilon) \tag{4.4}$$

Les fonctions de liens les plus fréquemment utilisées en fonction de la loi de $Y|X = x$ sont résumées dans le tableau ci-dessous.

Loi de $Y\|X = x$	Support	Fonction de lien
Normale	$\mathbb{R}$	$g(x) = x$
Bernoulli	{0,1}	$g(x) = ln(\frac{x}{1-x})$
Gamma	$\mathbb{R}^+$	$g(x) = \frac{1}{x}$
Poisson	$\mathbb{N}$	$g(x) = ln(x)$

TABLE 4.1 – Fonction de lien utilisé en fonction de la loi de $Y|X = x$

Classiquement, en assurance dommages, une loi de Bernoulli peut être privilégié pour modéliser des taux de résiliations. Une loi Gamma ou Log-normale est souvent choisie pour modéliser des coûts moyens de sinistres alors qu'une loi de Poisson permet de modéliser la fréquence de sinistres. Nous verrons dans la partie suivante quelles lois seront retenues afin de modéliser la fréquence et le coût moyen de la garantie dégâts des eaux.

4.1.2 Ajustement d'un GLM

L'objectif de la mise en place d'un GLM est d'estimer pour tout $j \in \{0, ...k\}$ les paramètres β_j en maximisant la fonction de log-vraisemblance. La fonction de lien g étant connue, nous pourrons alors exprimer $\mathbb{E}[Y|X = x]$. La fonction de log-vraisemblance pour n observations indépendantes est donnée par :

$$L(\beta) = ln(\prod_{i=1}^{n} f_{\theta_i,\phi}(y_i)) = \sum_{i=1}^{n} ln(f_{\theta_i,\phi}(y_i)) = \sum_{i=1}^{n} l_{\theta_i,\phi}(y_i) \tag{4.5}$$

Ainsi, à partir des équations (4.1) et (4.5), on obtient pour la i-ème observation :

$$l_{\theta_i,\phi}(y_i) = \frac{y\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi) \tag{4.6}$$

Commençons par exprimer l'espérance et la variance de la variable réponse Y_i correspondant à la i-ème observation. Dans cet objectif, les dérivées partielles première et seconde sont calculées :

$$\frac{\partial l}{\partial \theta_i} = \frac{y_i - b'(\theta_i)}{a(\phi)} \tag{4.7}$$

$$\frac{\partial^2 l}{\partial \theta_i^2} = \frac{-b''(\theta_i)}{a(\phi)} \tag{4.8}$$

Le maximum de la fonction de log-vraisemblance est atteint lorsque la dérivée première partielle est nulle. On peut alors écrire :

$$\mathbb{E}\left(\frac{\partial l}{\partial \theta_i}\right) = 0 \Rightarrow \mathbb{E}(Y_i) = b'(\theta) = \mu_i \tag{4.9}$$

Sous les conditions de régularité de la loi exponentielle, on a :

$$\mathbb{E}\left(\frac{\partial^2 l}{\partial \theta_i^2}\right) + \mathbb{E}\left[\left(\frac{\partial l}{\partial \theta_i}\right)^2\right] = 0 \Rightarrow \mathbb{E}\left(\frac{-b''(\theta_i)}{a(\phi)}\right) + \mathbb{E}\left[\left(\frac{Y_i - \mathbb{E}(Y_i)}{a(\phi)}\right)^2\right] = 0 \tag{4.10}$$

Par linéarité de l'espérance on obtient :

$$\frac{-b''(\theta_i)}{a(\phi)} + \frac{\mathbb{E}[(Y_i - \mathbb{E}(Y_i))^2)]}{a(\phi)^2} = 0 \Rightarrow Var(Y_i) = -b''(\theta_i) \cdot a(\phi) \tag{4.11}$$

Afin d'obtenir les équations de vraisemblance en fonction de $\beta = (\beta_0, ...\beta_p)$ et $\eta = X\beta$ le prédicteur linéaire à n composantes, il faut calculer la fonction de log-vraisemblance ci-dessous pour tout $i \in [1 : n]$ et tout $j \in [0 : p]$

$$\frac{\partial l}{\partial \beta_j} = \frac{\partial l}{\partial \theta_i} \cdot \frac{\partial \theta_i}{\partial \mu_i} \cdot \frac{\partial \mu_i}{\partial \eta_i} \cdot \frac{\partial \eta_i}{\partial \beta_j} \tag{4.12}$$

Ainsi il est alors possible de calculer chaque terme de l'équation ci-dessus :

- $\frac{\partial l}{\partial \theta_i} = \frac{y_i - b'(\theta_i)}{a(\phi)}$
- $\frac{\partial \theta_i}{\partial \mu_i} = (\frac{\partial \mu_i}{\partial \theta_i})^{-1} = (b''(\theta_i)) = \frac{a(\phi)}{Var(Y_i)}$
- $\frac{\partial \mu_i}{\partial \eta_i} = x_{ij}$ avec x_{ij} correspondant au terme de la i-ème observation et de la j-ème variable explicative de X
- $\frac{\partial \eta_i}{\partial \beta_j} = \frac{\partial g(\mu_i)}{\partial \beta_j}$ car η_i dépend de la fonction de lien et donc $\eta_i = g(\mu_i)$

Pour tout $j \in [0, p]$, les équations de vraisemblances en fonction des β sont alors de la forme :

$$\sum_{i=1}^{n} \frac{Y_i - \mu_i}{Var(Y_i)} \cdot \frac{\partial g(\mu_i)}{\partial \beta_j} \cdot x_{ij} = 0 \tag{4.13}$$

Les équations de vraisemblance obtenues sont des équations non linéaire en β. Dans la suite de ce mémoire, des solutions approchées des équations de vraisemblances vont être calculées à partir du logiciel de tarification *Emblem* dans lequel nous allons implémenté des GLM et nous obtiendrons ainsi une estimation des coefficients β.

4.1.3 Qualité d'ajustement d'un GLM

4.1.3.1 Fonction *Mean Excess*

La méthode du *Mean Excess* [1] a été utilisée afin de déterminer le seuil à partir duquel un sinistre sera considéré comme un sinistre grave et donc sera supprimé de la modélisation de ce mémoire. En effet, il est préférable de modéliser distinctement sinistres attritionnels et sinistre graves puisqu'ils ne suivent généralement pas la même loi de distribution. Dans ce mémoire, nous modélisons le coût moyen de la garantie dégâts

1. Moyenne des excès

des eaux uniquement pour les sinistres attritionnels. Cette méthode est une méthode graphique qui va nous permettre de déterminer l'allure de la queue de distribution. La Fonction du *Mean Excess* est définie par :

$$e(u) = \mathbb{E}[X - u | X > u] \tag{4.14}$$

On notera $\widehat{e}_n(u)$, l'estimateur empirique de la fonction *Mean Excess* tel que :

$$\widehat{e}_n(u) = \frac{\sum_{i=1}^{n} max(0, x_i - u)}{\sum_{i=1}^{n} \mathbb{1}_{x_i > u}} \tag{4.15}$$

Le numérateur représente la somme des excès, c'est à dire la partie des coûts de sinistres dépassant le seuil u et le dénominateur représente le nombre de sinistres dépassant le seuil u. Lorsque l'on trace la fonction *Mean Excess*, cela revient à tracer pour chaque seuil u, l'estimation empirique de la fonction *Mean Excess* en fonction de u. Plusieurs cas sont possibles pour l'allure de la fonction du *Mean Excess* à partir du seuil u, parmi eux :

- Si à partir d'un seuil u, la fonction *Mean Excess* est affine, alors les excès suivent une loi de Pareto généralisée de paramètres ξ et σ.
- Si à partir d'un seuil u, la fonction *Mean Excess* est horizontale alors les excès suivent une loi exponentielle de paramètres λ.

4.1.3.2 Test de Kolmogorov-Smirnov

Le test de Kolmogorov-Smirnov est un test non paramétrique consistant à tester si un échantillon d'observations suit une loi connue donnée en utilisant sa fonction de répartition.

Soient $X_1, X_2..., X_n$, n variables aléatoires i.i.d [2] définies dans un espace probabilisé Ω. On notera F_n la fonction de répartition empirique associée telle que :

$$\forall x \in \mathbb{R},\ F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{X_i \leq x} \quad \text{avec } \mathbb{1} \text{ la fonction indicatice} \tag{4.16}$$

Notons F la fonction de répartition continue de la loi théorique connue. Rappelons que la fonction de répartition d'une variable aléatoire X est définit de la manière suivante :

$$\forall x \in \mathbb{R},\ F_X(x) = P(X \leq x) \tag{4.17}$$

L'hypothèse nulle du test de Kolmogorov-Smirnov est :

$$H_0 : \text{l'échantillon suit la loi théorique connue}$$

2. Indépendantes et Identiquement Distribuées

H_1 : l'échantillon ne suit pas la loi théorique connue

La statistique du test de Kolmogorov-Smirnov est :

$$D_{KS} = \sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \tag{4.18}$$

Le test de Kolmogorov-Smirnov consiste à comparer la valeur obtenue pour la statistique de test à une valeur théorique que l'on trouve dans la table de Kolmogorov-Smirnov que l'on notera $D_\alpha(n)$ avec α la probabilité d'erreur et n la taille de l'échantillon. Si $D_{KS} > D_\alpha(n)$, on rejette l'hypothèse H_0. Sous H_0, plus D_{KS} est petite et plus la loi choisie est adaptée à la distribution de loi inconnue.

4.1.3.3 Test du khi-deux

Le test du khi-deux a été réalisé afin de tester si la loi de poisson était adaptée à la modélisation de la fréquence de sinistralité.

Soit n, le nombre d'observations de l'échantillon. Notons $(O_i)_{i\in[1:n]}$, l'effectif observé et $(T_i)_{i\in[1:n]}$, l'effectif théorique suivant une loi connue.

L'hypothèse du test du khi-deux est :

H_0 : La loi de probabilité théorique est adaptée à la distribution observée

H_1 : La loi de probabilité théorique n'est pas adaptée à la distribution observée

La statistique de test du test du khi-deux est alors :

$$\chi^2_{cal} = \sum_{i=1}^{n} \frac{(O_i - T_i)^2}{T_i} \tag{4.19}$$

La statistique de test obtenue est ensuite comparée avec la valeur de la loi du khi-deux pour un niveau de confiance α et un degré de liberté valant $n-1$.

— Si $\chi^2_{cal} \leq \chi^2_{\alpha,n-1}$, on ne rejette pas H_0 au niveau α

— Si $\chi^2_{cal} > \chi^2_{\alpha,n-1}$, on rejette H_0 au niveau α

4.1.3.4 Analyse des corrélations

L'une des hypothèses fortes qui doit être respectée lors de la mise en place d'un GLM est que les variables explicatives ne doivent pas être corrélées entre elles. Afin de vérifier cette hypothèse, la matrice de dépendance du V de Cramer qui s'obtient à partir de la statistique du khi-deux à été calculée.

$$V = \sqrt{\frac{\chi^2}{n.min(i-1, j-1)}} \tag{4.20}$$

Avec n, la taille de l'échantillon, i et j le nombre de modalités des deux variables pour lesquelles on étudie la dépendance et χ^2 est la statistique du khi-deux.

Plus le V est proche de zéro et moins les variables sont dépendantes entre elles. A l'inverse, plus le V est proche de 1 et plus les variables sont liées entre elles.

V de Cramer	*Force du lien statistique*
0	Indépendance
]0 : 0,1]	Très faible
]0,1 : 0,2]	Faible
]0,2 : 0,4]	Modérée
]0,4 : 0,8]	Forte
]0,8 : 1]	Colinéarité

TABLE 4.2 – Force du lien statistique en fonction des valeurs du V de Cramer

4.1.3.5 Méthode *Forward*

La méthode *Forward* est une méthode ascendante basée sur le calcul de l'AIC[3]. L'objectif final de la méthode *Forward* est de sélectionner les variables explicatives afin d'obtenir un modèle qui minimise l'AIC. L'AIC se calcule de la manière suivante :

$$AIC = -2l + 2p \tag{4.21}$$

Avec p le nombre de variables explicatives dans le modèle et l la log-vraisemblance maximisée.

Les trois étapes de la méthode Forward sont :

1. Estimer le modèle M_0 qui ne contient aucune variable explicative,
2. Construire pour tout $k \in [0 : p-1]$, les $p-k$ modèles à partir de M_k auquel on rajoute une variable explicative, et sélectionner le modèle ayant l'AIC le plus petit qui sera nommer M_{k+1},
3. Choisir le modèle ayant le plus petit AIC parmi les modèles $M_0, M_1, ...M_p$.

4.1.3.6 Test de Wald

Le test de Wald va permettre de tester la significativité des coefficients associés aux variables significatives sélectionnées par la méthode *Forward*.

3. *Akaike Information Criterion*

Soient les hypothèses : $H_0 : \beta_j = 0$ et $H_1 : \beta_j \neq 0$. La statistique de test du test de Wald est donnée par :

$$T = \frac{\hat{\beta}_j - \beta_j}{\sqrt{V(\beta_j)}} \tag{4.22}$$

Sous H_0, la statistique de test est alors donnée par : $T = \frac{\hat{\beta}_j}{\sqrt{V(\beta_j)}} \underset{n\to+\infty}{\longrightarrow} \mathcal{N}(0,\, 1)$

Si $|T| \geq q_{1-\frac{\alpha}{2}}$, alors on rejette H_0 au niveau α.

4.1.4 Validation du modèle

4.1.4.1 Calcul de la déviance

L'analyse de la performance des modèles est réalisée à partir du calcul de la déviance. La déviance, notée D permet de calculer la différence entre la vraisemblance du modèle estimé et la vraisemblance du modèle saturé correspond au meilleur modèle possible. Ainsi l'objectif est de minimiser D afin d'avoir un modèle estimé au plus proche du modèle saturé. La formule de la déviance est :

$$D = -2(L_{sature} - L_{estime}) \tag{4.23}$$

4.1.4.2 Calculs des résidus

Le résidu de déviance d'une observation y_i ayant pour prédiction $\hat{y}_i$ est défini de la manière suivante :

$$\hat{\epsilon}_{D,i} = signe(y_i - \hat{y}_i) \cdot \sqrt{d_i} \text{ avec } D = \sum_{i=1}^{n} d_i \tag{4.24}$$

où d_i représente la contribution de l'observation i à la déviance obtenue D. Dans le cas où l'hypothèse de distribution est une loi de Poisson :

$$d_i = 2\left(ln\left(\frac{y_i}{\hat{y}_i}\right) - (y_i - \hat{y}_i)\right)$$

Dans le cas où l'hypothèse de distribution est une loi Gamma :

$$d_i = 2\left(ln\left(\frac{y_i}{\hat{y}_i}\right) - \frac{y_i - \hat{y}_i}{\hat{y}_i}\right)$$

4.2 Mise en place des modèles GLM

Dans cette partie, nous modélisons la fréquence et le coût moyen de la garantie dégâts des eaux. Dans le cas du modèle de fréquence, la variable à expliquer correspond au nombre de sinistres survenus durant une période. Dans le cas du modèle de coût moyen, la variable à expliquer correspond à la charge de sinistres moyenne. De plus, les variables explicatives sont celles présentées dans le chapitre 2. Comme évoqué tout au long de ce mémoire, selon les modèles nous utiliserons soit uniquement des données internes, uniquement des données externes ou un mixte des deux.

D'autres part la modélisation GLM ne va pas être réalisée sur l'ensemble de la sinistralité. En effet, l'objectif est d'obtenir une distribution des coûts de sinistres qui suive une loi connue afin de mettre en place des GLM. Dans cette volonté, la distribution des coûts de sinistres a été tracée sur le graphe ci-dessous. Nous observons un pic important qui vient déformer la distribution des coûts de sinistres. Ce pic correspond à 1461 sinistres qui ont un coût de sinistre qui vaut 1350 €. Cela s'explique par le forfait d'ouverture d'Allianz France pour la garantie dégâts des eaux. Le forfait d'ouverture correspond à un coût fictif que va renseigner l'agent avant de connaître le coût exact du sinistre. En effet, lorsqu'un assuré appelle le service d'indemnisation afin de signaler son sinistre dégâts des eaux, le service indemnisation n'est pas en mesure de connaître le coût du sinistre avant d'avoir mené une expertise. Ainsi, il renseigne 1350 € comme montant, ce montant est calculé chaque année par les collaborateurs de l'équipe indemnisation.

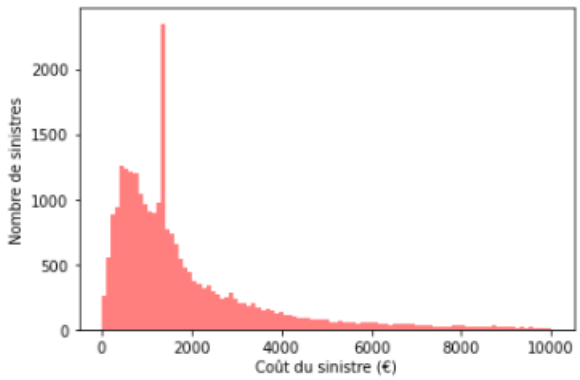

FIGURE 4.1 – Distributions des coût moyen avant suppression des pics

Ce pic biaisant la distribution des coûts et ne représentant pas le coût de sinistres exact, l'ensemble des 1461 sinistres concernés par le forfait d'ouverture vont être supprimés de notre base d'étude. Dans le cas de la réalisation du tarif technique, des coefficients de calage sont ajoutés après la réalisation des modèles afin de prendre en compte ces sinistres supprimés. De la même manière et comme évoqué dans le premier chapitre, une sur-représentation des sinistres forfaitaires payés dans la cadre des conventions CIDRE et

IRSI pourrait déformer la distribution des coûts de sinistres. Il serait alors plus compliqué de modéliser les coûts moyens par une loi simple comme la loi Gamma.

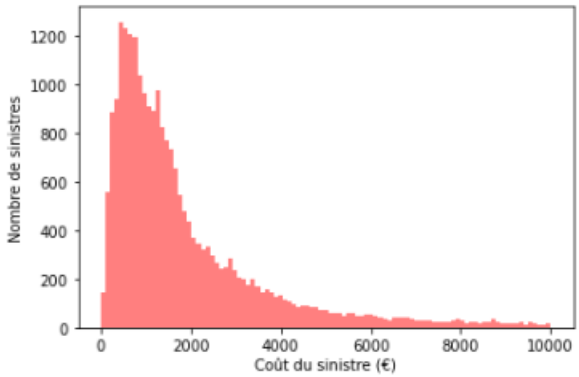

FIGURE 4.2 – Distribution des coût moyen après suppression des pics

4.2.1 Étude des sinistres graves

Il est important de modéliser séparément les sinistres attritionnels et les sinistres graves. Par définition, les sinistres graves sont des sinistres avec un coût moyen élevé et une fréquence faible. Il faut déterminer un seuil de coût de sinistres à partir duquel un sinistre sera considéré comme un sinistre grave. Il existe plusieurs méthodes afin de déterminer ce seuil. L'objectif de ce mémoire n'est pas porté sur la détermination du seuil de grave mais il est cependant important de retirer les sinistres graves de notre modélisation afin d'avoir une distribution des coûts de sinistres qui suive une loi simple et connue.

Il existe deux méthodes pour traiter les sinistres graves :

- La première est la méthode de la troncature, c'est à dire que pour tout sinistre supérieur au seuil déterminé u, alors le coût du sinistre est divisé en deux parties, la partie du coût du sinistre inférieur ou égal au seuil u va être modélisée dans le modèle attritionnel et la partie du coût du sinistre supérieur au seuil u sera traitée dans le modèle de grave.
- La deuxième méthode est la méthode de la séparation, c'est-à-dire que dès lors que le coût d'un sinistre est supérieur au seuil u, alors la totalité du coût du sinistre sera traitée dans le modèle de grave.

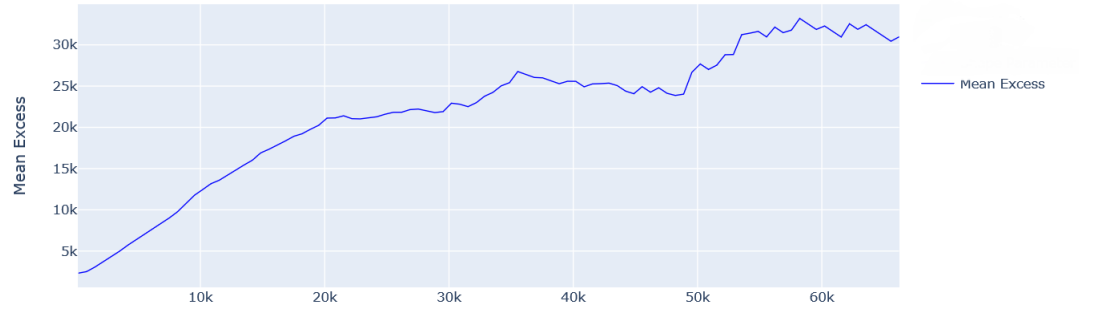

FIGURE 4.3 – *Mean Excess* des coûts de sinistres après suppression du pic à 1350 €

Dans le cas des coûts de sinistre dégâts des eaux du portefeuille MRH d'Allianz France, il est observé un changement de tendance au seuil de 20 000 euros à partir duquel la fonction est linéaire. C'est ce seuil qui va être retenu comme seuil de sinistres graves. La méthode retenue pour traiter les sinistres graves dans ce mémoire est la méthode de la séparation. Ainsi, tout sinistre supérieur à 20 000 euros sera supprimé entièrement de la modélisation.

4.2.2 Adéquation des distributions de fréquence et coût moyen à des lois

L'enjeux de cette section est de déterminer la loi la plus adéquate pour modéliser la fréquence et le coût moyen pour la garantie dégâts des eaux. Deux lois ont été testées pour modéliser le coût moyen. La loi Gamma définie pour tout x positif ainsi que la loi log-normale définie pour tout x strictement positif. Les coûts moyens étant nécessairement positifs, nous avons sélectionné ces lois en conséquence.

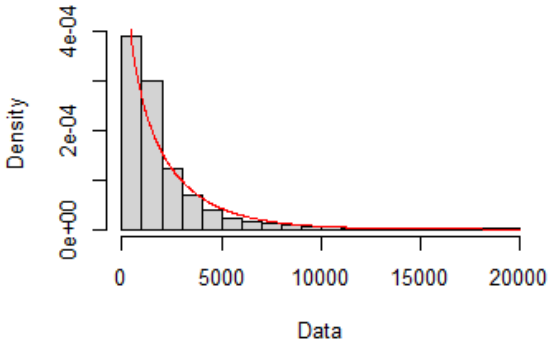

FIGURE 4.4 – Adéquation graphique de la loi gamma à la distribution des coûts moyens

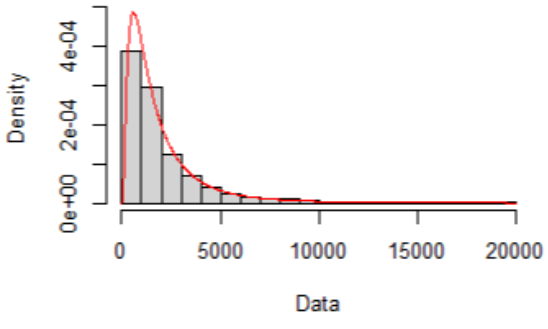

FIGURE 4.5 – Adéquation graphique de la loi log normale à la distribution des coûts moyens

Dans un premier temps, un test graphique a été réalisé, il s'agit de tracer le QQ-plot. Ce graphique permet de comparer deux distributions, ici la distribution des coûts moyens observés dans le portefeuille avec la distribution d'une loi connue (gamma et log-normale). L'objectif du QQ-plot est de tracer le nuage de points ayant comme ordonnée les quantiles empiriques et en abscisse les quantiles théoriques correspondant à ceux de la loi choisie. Si la distribution théorique représente bien les observations empiriques, alors le nuage de points sera aligné sur la droite d'équation $y = x$. Plus les déviations du nuage de points par rapport à cette droite sont importantes et moins la loi choisie sera adaptée à la distribution empirique.

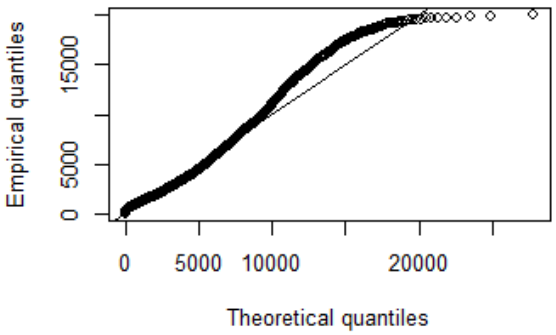

FIGURE 4.6 – QQ-plot loi gamma

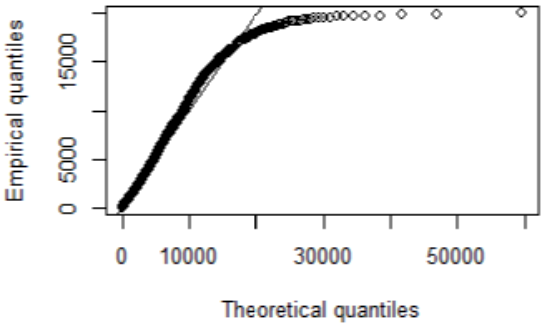

FIGURE 4.7 – QQ-plot loi log-normale

Il est observable que dans le cas de la loi log-normale, le nuage de point est plus aligné que dans le cas de la loi gamma.

Le test statistique de Kolmogorov-Smirnov a ensuite été réalisé afin de compléter l'étude graphique.

Loi gamma	Loi log-normale
$D_{KS} = 0,13959$	$D_{KS} = 0,028397$

TABLE 4.3 – Résultat du test de Kolmogorov-Smirnov

En prenant $\alpha = 5\%$ et $n = 24176$ correspondant à la taille de l'échantillon, on a d'après la table de Kolmogorow-Smirnov $D_{0.05}(24176) = 0{,}2927$.

Ainsi, $D_{KS} \leq D_{0.05}(24176)$, on ne rejette pas H_0 pour la loi gamma et pour la loi log-normale. Comme la distance D_{KS} est plus petite pour la loi log-normale, on peut conclure que la loi log-normale est plus adaptée pour modéliser la distribution des coûts moyens. Cependant, ce mémoire s'inscrit dans un objectif de production au sein d'Allianz France

et doit suivre certaines directives du Groupe Allianz. Dans cette volonté, la loi gamma va être retenue afin de s'adapter aux outils de tarification utilisés par le Groupe. La démonstration de l'appartenance de la loi Gamma à la famille exponentielle est réalisée en Annexe A.

La fonction de lien utilisée pour les modèles de coût moyen est la fonction logarithme pour des raisons de contraintes logiciels.

La loi de Poisson a été retenue afin de modéliser la fréquence. Ce choix a été corroboré grâce au test du khi-deux. De même que pour la loi Gamma, la démonstration de l'appartenance de la loi de Poisson à la famille exponentielle est réalisée en en Annexe A.

La fonction de lien utilisée pour les modèles de fréquence est la fonction logarithme car le paramètre canonique de la distribution de Poisson est le logarithme de l'espérance.

4.2.3 Étude des corrélations

L'étude des corrélations entre les variables explicatives est très importante lors de la mise en place des GLM. En effet, l'une de hypothèses forte du GLM est que l'ensemble des variables explicatives du modèle ne doivent pas être corrélées. Le corrélogramme ci-dessous représente le V de Cramer pour l'ensemble des variables explicatives du modèle de fréquence sur les données internes. Plus le bleu est foncé et plus la dépendance entre les deux variables est importante. On observe que la majorité des variables explicatives du modèle de fréquence de données internes ne sont pas corrélées entre elles. Le rond rouge en haut à droite montre une dépendance entre trois variables qui réfèrent toutes les trois à des anciennetés de sinistre, ainsi il faudra garder uniquement l'une des trois dans les modèles. En bas à droite, une autre dépendance est représentée, les deux variables concernées sont toute les deux liées au nombre de contrats (nombre de contrats en IARD et nombre de contrats total) chez Allianz, de même il faudra garder une seule de ces variables pour la modélisation.

FIGURE 4.8 – Corrélogramme des variables explicatives pour les modèles de données internes

Le corrélogramme pour l'ensemble des variables internes explicatives du modèle de coût moyen sur les données internes n'est pas représenté car il est semblable à celui de la fréquence. Le corrélogramme des variables explicatives pour les modèles de données externes ainsi que celui pour les modèles de données internes et externes étant très volumineux et par conséquent peu lisible, les corrélogrammes avec uniquement les variables retenues seront présentés dans la suite de ce mémoire. En effet, le nombre de données externes récoltées est de 162 variables explicatives.

4.2.4 Résultats et qualité d'ajustement du GLM

L'ensemble des GLM ont été implémentés sur l'outil *Emblem*. Les différentes étapes de réalisation d'un modèle GLM avec Emblem sont décrites dans le schéma ci-dessous.

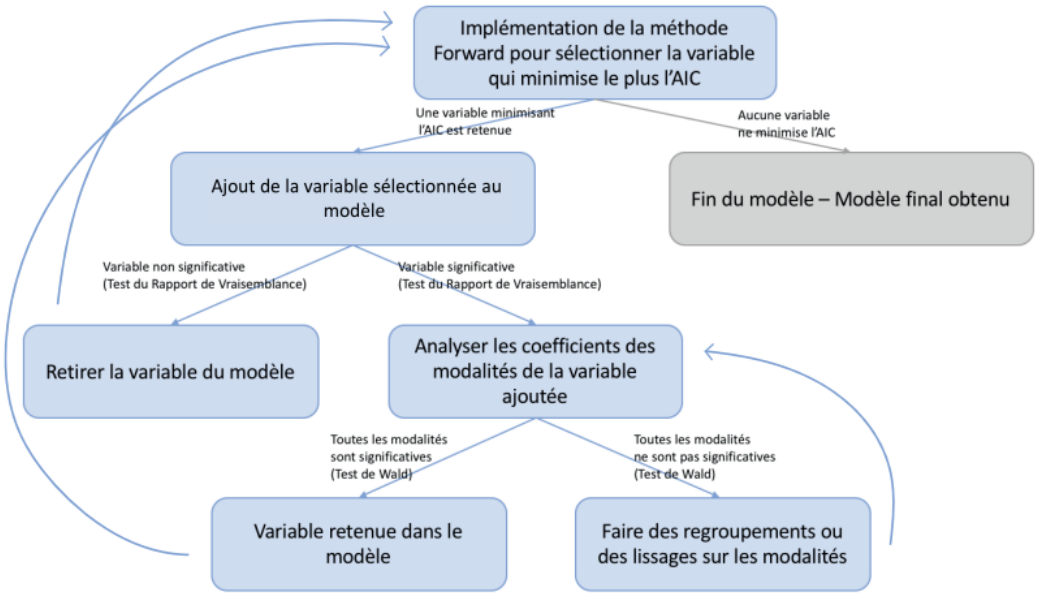

FIGURE 4.9 – Différentes étapes de la construction de GLM sur Emblem

L'ensemble des variables sélectionnées par la méthode *Forward* pour chacun des 6 modèles était significative lors du test de rapport de vraisemblance.

Le test de Wald à ensuite été réalisé afin de tester la significativité des coefficients associés aux variables significatives sélectionnées par la méthode *Forward*. Le logiciel de tarification Emblem nous donne la *Standard Error*[4] en % correspondant à l'écart type du coefficient estimé $\hat{\beta}_j$ divisé par ce dernier.

$$Standard\ Error\ (\%) = \left| \frac{Standard\ Error}{Valeur\ du\ parametre\ estime} \right| \tag{4.25}$$

4. Ecart type

Emblem utilise $\alpha = 4,56\%$ pour le test de Wald, ce qui correspond à $q_{1-\frac{\alpha}{2}} = 2$. Ainsi, sous l'hypothèse H_0, $|T| \geq 2$ et donc *Standard Error* (%) $< \frac{1}{2}$. T est la statistique de test présentée dans la partie théorique.

Ainsi, dans le logiciel Emblem :

- Si *Standard Error* (%) $< 50\%$, le coefficient est significatif, il apparaîtra alors en vert.
- Si *Standard Error* (%) $< 50\%$, le coefficient n'est pas significatif.
  - Si *Standard Error* (%) $\in [50\%; 75\%]$, il apparaîtra en gris.
  - Si *Standard Error* (%) $> 75\%$, il apparaîtra en rouge.

Lorsqu'une modalité n'est pas significative, il existe plusieurs méthodes pour contrer cela :

— Faire des regroupements entre plusieurs modalités. Par exemple pour une modalité ayant un nombre d'observations pas assez important ou dans le cas où deux modalités ont des coefficients très proches.
— Effectuer un lissage polynomial afin de prendre en compte la tendance des modalités d'une variable continue si elles suivent une tendance polynomiale.

Par exemple, dans le modèle de sévérité pour les données internes, la variable qualité juridique avait des modalités qui n'étaient pas toutes significatives. La qualité juridique est décomposée en 10 modalités dans les bases de données, certaines ayant un nombre d'observation nul. La *p_value* du test du rapport de vraisemblance pour cette variable est inférieure à 5%.

Les dix modalités sont :

- USU, usufruitiers
- LMB, locataires meublés
- OGR, occupants à titre gratuit
- PLM, propriétaires loueurs meublé
- COP, co-propriétaires
- PRO, propriétaires
- POP, propriétaires occupants partiels
- CNO, co-propriétaires non occupants
- AUT, autres
- NR, information manquante

Notons que les modalités CNO, AUT et NR ne seront pas prises en compte car leur nombre d'observations est égal à 0.

Nom	Ecart Type (%)	Poids (%)	Bêta (β)
USU	53,9	0,8	1,1450
LMB	210,1	0,1	0,8962
OGR	28,1	1,1	0,7959
LOC	4,8	14,6	0,6513
PLM	132,2	1,2	0,9534
COP	90,4	0,5	0,9028
PRO		80,3	
POP	28,8	1,5	0,8238

TABLE 4.4 – Variable qualité juridique sans regroupement dans le modèle sévérité de données internes

La modalité de référence est celle qui a le poids le plus important en nombre d'observations, elle est représentée en bleu dans le tableau ci-dessus. On peut voir que les modalités qui ne sont pas significatives en rouge ont des poids très faibles. L'objectif est donc de regrouper ces variables avec des variables ayant un coefficient proche d'une part mais aussi en apportant une analyse métier d'autre part. Dans ce cas là, il est cohérent de regrouper les copropriétaires, les propriétaires loueurs meublé, les propriétaires occupants partiels et les usufruitiers avec les propriétaires. De même que de regrouper les occupants à titre gratuit avec les locataires. Ainsi, ces deux groupes vont être constitués.

Nom	Ecart Type (%)	Poids (%)	Bêta (β)
PRO, POP, USU, COP, PLM		76,3	1,1450
LOC, OGR	7,4	23,3	0,7524
LMB	13,5	0,4	0,2675

TABLE 4.5 – Variable qualité juridique après regroupement dans le modèle sévérité de données internes

On peut voir qu'une fois les regroupements constitués, l'ensemble des modalités de la variable qualité juridique sont significatives.

Lorsque les variables sont continues de manière générale une approche polynomiale est réalisée. Par exemple pour la variable renseignant sur le nombre de pièces, une tendance croissante est observée pour le modèle de coût moyen. Ainsi un lissage polynomial sera réalisé afin de refléter cette tendance croissante. Notons que la modalité 17+ a pris la valeur 17 pour le lissage.

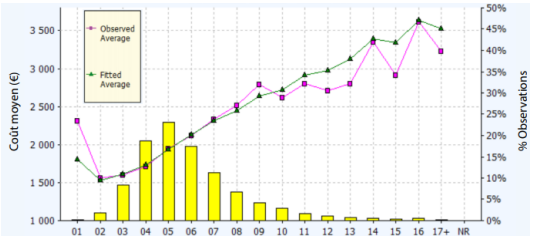

FIGURE 4.10 – Coût moyen observé et prédit en fonction des modalités de la variable nombre de pièces

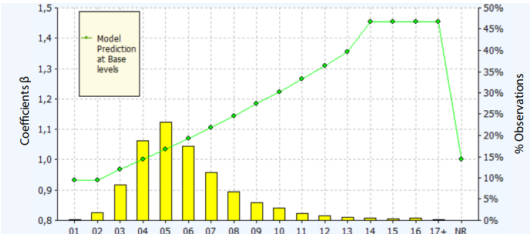

FIGURE 4.11 – Coefficients bêta attribués par le GLM après lissage polynomial pour la variable nombre de pièces

L'ensemble des variables sélectionnées pour les différents modèles est résumé dans les tableaux ci-dessous. Elles sont rangées par ordre d'importance et l'interprétation de ces résultats sera faite dans la section suivante.

Modèle de données internes	Modèle de données externes	Modèles de données mixtes
Qualité juridique	Précipitations été	Qualité
Région	Taux résidences $> 120m^2$	Nombre de pièces
Nombre de pièces	Aire urbaine	Age
CSP	Taux interventions dde	Objets de valeur
Nature	Coût reconst maison	Aire urbaine
Age	Coût du PVC	Piscine
Franchise	Taux csp prof intermédiaire	Franchise
Bâtiment en construction	Taux rés 1 ou 2 pièces	Précipitations annuelles
Segmentation client	Distance min pompiers	Choix de paiment
Piscine	Taux d'artisans	Type
Nombre de contrat AZ	Prix moyen m^2	Nombre de contrats AZ
Capitaux assurés	Mousse sur le toit	Taux csp prof intermédiaire
Choix du paiement	Cours du cuivre	Mousse sur le toit
Type	Nb de personnes par rés	Taux de personnes mariées
Option objets de valeur	Taux de rés avant 1990	Distance min pompiers
Objets de valeur	Taux de personnes mariées	Altitude
Garantie DEL	Coût du plombier	Panneaux solaires
Installations extérieures	Taux de familles 1 enfant	Cours du cuivre
Protéction juridique	Taux propriétaires	Coût reconst maison
Ancienneté de sinistres	Zonier tempête RMS	Zonier tempête RMS
Cheminée		Taux de familles 1 enfant
Ancienneté de client		Taux interventions dde
		Taux de rés avant 1990
		Taux de familles > 3 enfants
		Taux rés 1 ou 2 pièces
		Taux rés principale
		Coût du plombier

TABLE 4.6 – Variables retenues pour les modèles de coût moyen

Modèle de données internes	Modèle de données externes	Modèles de données mixtes
Région	Altitude	Ancienneté de sinistres
Ancienneté de sinistres	Température max hiver	Nombre de pièces
Nombre de pièces	Taux maisons	Age
Protocole	Aire urbaine	Qualité juridique
Qualité juridique	Taux résidences $> 120m^2$	Piscine
Age	Taux de bacheliers	Objets de valeur
Fractionnement de paiement	Distance min police	Altitude
Objets de valeur	Taux de rés avant 1990	Température max hiver
Piscine	Rouille	Franchise
CSP	Transports en commun	Aire urbaine
Franchise	Panneaux solaires	Transports en commun
Ancienneté de client	Taux ouvriers supérieurs	Rouille
Installations extérieures	Densité	Distance min police
Nature	Graffiti	Taux ouvriers supérieurs
Cheminée	Taux chômeurs	Taux chômeurs
Bâtiment en construction	Taux de personnes mariées	Densité
Segmentation client	Taux interventions dde	Taux résidences $> 120m^2$
Nombre d'enfants	Nb de jours Température < 0	Taux interventions dde
Ancienneté du bâtiment	Taux csp employées	Panneaux solaires
Alarme	Taux de personnes > 15 ans	Taux maisons
Garantie DEL	Mousse sur le toit	Taux de rés avant 1990
Etat civil	Taux rés principale	Score inondation
Dépendances	Score inondation	Mousse sur le toit
	Taux de familles > 3 enfants	Nb de jours Température < 0
	Taux propriétaires	Taux csp employées
	Nb de jours de pluie/neige	Taux de divorcées
		Taux rés principales

TABLE 4.7 – Variables retenues pour les modèles de fréquence

Le corrélogramme ci-dessous représente les dépendances entre les variables sélectionnées pour le modèle de fréquence sur les données externes. On observe que l'ensemble des variables sélectionnées ont une faible dépendance entre elles. Cependant avant de sélectionner ces variables, certaines variables externes avaient une dépendance forte entre elles.

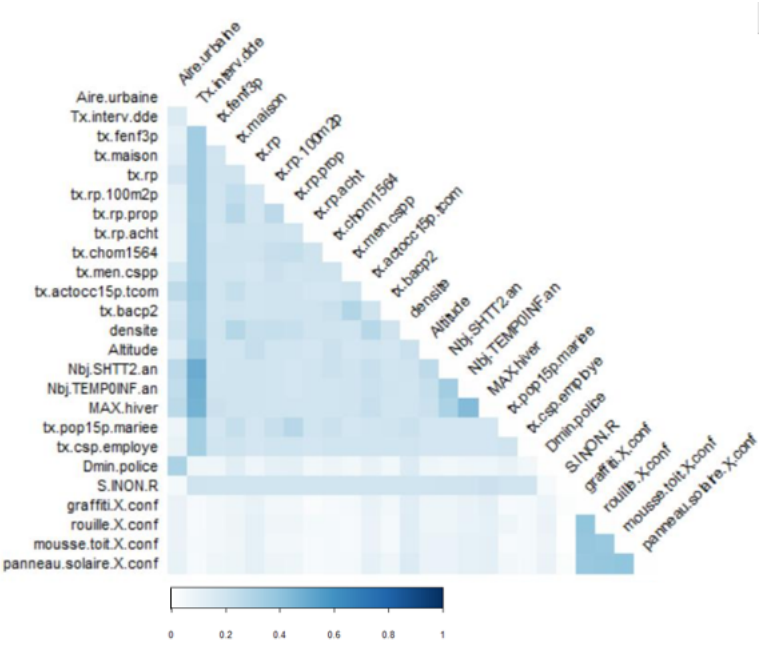

FIGURE 4.12 – Corrélogramme des variables explicatives sélectionnées pour le modèle de fréquence sur les données externes

Le corrélogramme ci-dessous représente les dépendances entre les variables sélectionnés pour le modèle de fréquence sur les données internes et externes. De même que pour le modèle de fréquence sur les données externes, on observe que l'ensemble des variables sélectionnées, exceptées les variables de coût ont une faible dépendance entre elles. Les variables de coût (coût de la main d'oeuvre du plombier et cours du cuivre) sont légèrement dépendantes mais cela n'a pas empêché le modèle GLM de converger. De plus ces variables ont été lissées par des polynômes dans le GLM ce qui a permis de réduire les dépendances.

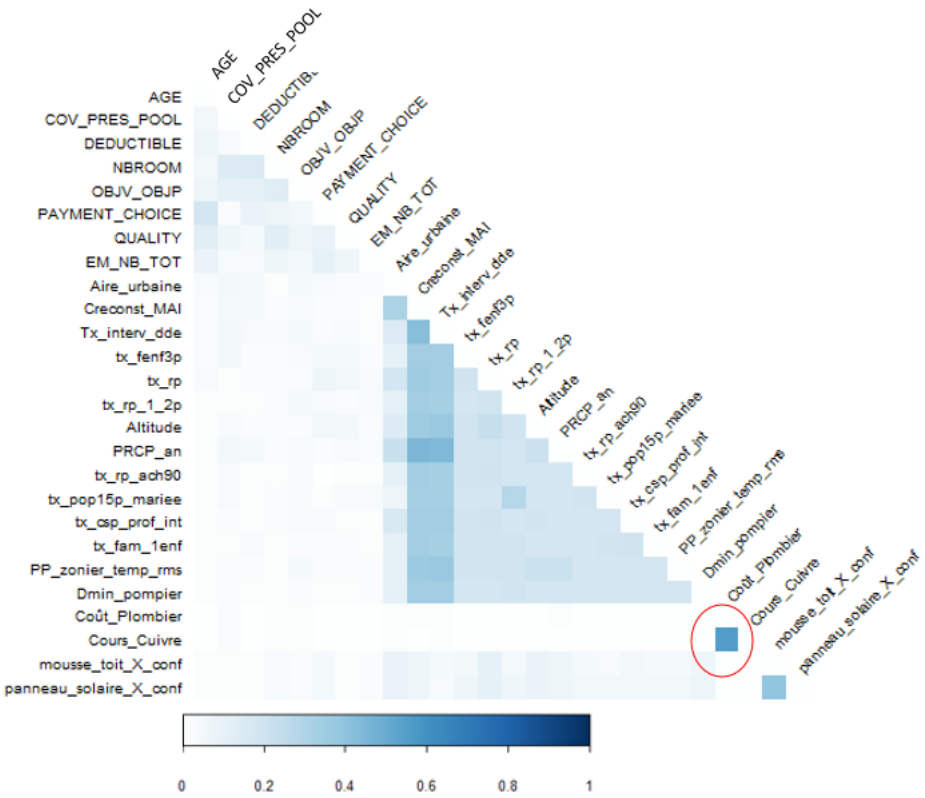

FIGURE 4.13 – Corrélogramme des variables explicatives sélectionnées pour le modèle de sévérité sur les données internes et externes

De même, les corrélogrammes des quatre autres modèles montrent que l'ensemble des variables sélectionnées ont une faible dépendance entre elles.

4.3 Validation du GLM

4.3.1 Analyse des résidus

L'analyse des résidus permet de vérifier la performance des modèles. Les résidus obtenus peuvent être analysé graphiquement en regardant la forme du nuage de points des résidus qui sont représentés en fonction des valeurs prédites :

— Les résidus doivent être centrés autour de zéro pour valider l'hypothèse de linéarité sur laquelle se base le modèle, si ce n'est pas le cas, il est nécessaire de réaliser d'autres graphiques représentant les résidus par rapport à chacune des variables explicatives, pour trouver la ou les variables responsables de cette non-linéarité,

— Les résidus ne doivent pas présenter une allure particulière pour vérifier l'hypothèse de l'homoscédasticité.

Les trois graphiques ci-dessous représentent les résidus de déviances des modèles de sévérité, respectivement sur les données internes, externes et mixtes. Ces graphiques montrent pour chaque valeur de coût moyen, le nombre d'observations concerné et le résidu de déviance calculé. L'ensemble des résidus des trois modèles sont compris entre -2 et 5. On remarque que dans les trois graphiques, les observations sont centrées entre 0 et -1. Cela signifie que les valeurs prédites sont légèrement supérieures aux valeurs observées. Pour les modèles de données internes et mixtes les résidus sont plus proches de 0 que de -1 à la différence du modèle de données externes.

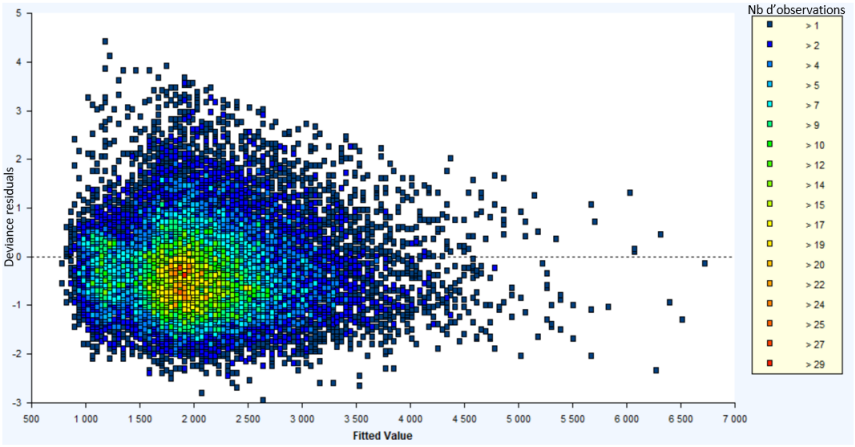

FIGURE 4.14 – Résidus de déviance pour le modèle de sévérité sur les données internes

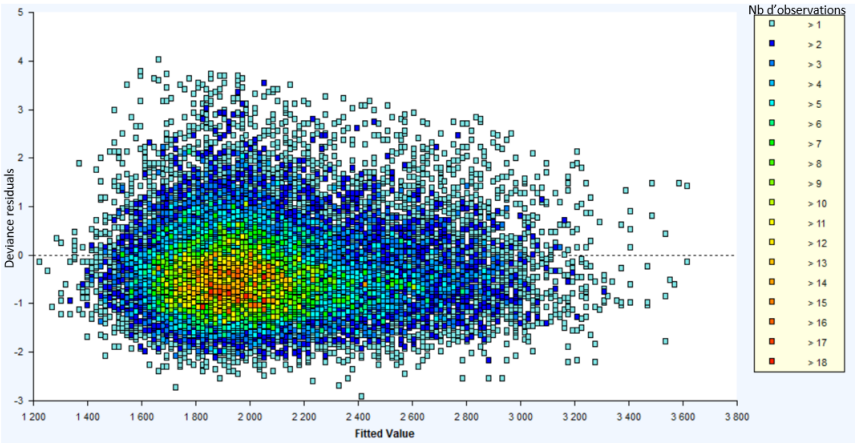

FIGURE 4.15 – Résidus de déviance pour le modèle de sévérité sur les données externes

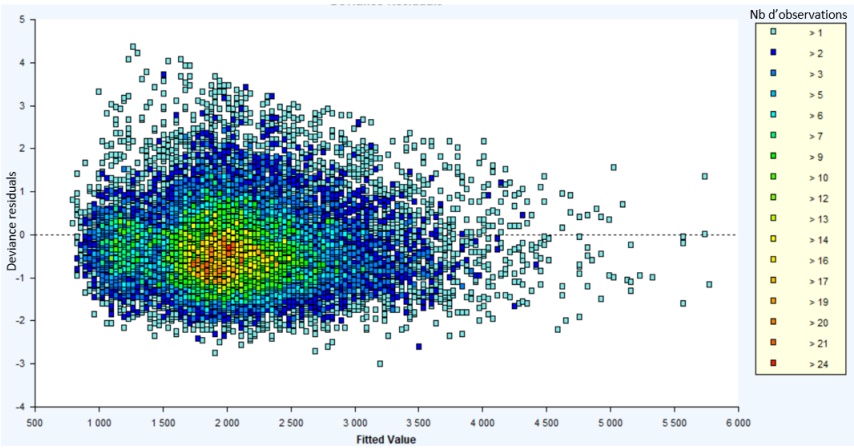

FIGURE 4.16 – Résidus de déviance pour le modèle de sévérité sur les données mixtes

Les résidus ont aussi été analysés sur les modèles de fréquence. L'ensemble des graphiques des résidus ainsi que leurs interprétations sont en Annexe B de ce mémoire. Notons que pour la fréquence ce sont les résidus *crunched* qui ont été tracés et que les conclusions vont dans le même sens que pour les modèles de coût moyen, c'est à dire que le modèle de données externes à des résidus moins centrés autour de 0 que pour les deux autres modèles ce qui reflète une moins bonne prédiction de la fréquence par le modèle.

4.3.2 Analyse de la déviance sur la base de test

Afin d'analyser la performance des différents modèles présentés précédemment, la déviance pour chacun des modèles a été calculée sur la base d'apprentissage et la base de test puis comparée à la déviance du modèle sans donnée, c'est à dire le modèle sans aucune variable ajoutée.

Dans les tableaux ci-dessous, on observe que l'ensemble des écarts de déviance calculées sont constants entre la base d'apprentissage et la base de test. Cela permet d'apprécier le fait que les modèles n'ont pas eu de surapprentissage. En effet, lors du surapprentissage, la déviance est bien meilleure sur la base d'apprentissage que sur la base de test.

	Apprentissage	**Test**
Modèle sans donnée	18006,19	4551,08
Déviance du modèle avec données internes	16332,19	4169,72
Ecart de déviance (%) entre le modèle de données internes et le modèle sans donnée	-9,30	- 8,38
Déviance du modèle avec données externes	17487,39	4421,56
Ecart de déviance (%) entre le modèle de données externes et le modèle sans donnée	-2,88	-2,85
Déviance du modèle avec données internes et externes	16471,32	4173,86
Ecart de déviance (%) entre le modèle de données internes et externes avec le modèle sans donnée	-8,52	-8,29

TABLE 4.8 – Analyse des déviances des différents modèles de sévérité

	Apprentissage	**Test**
Modèle sans donnée	194199,23	48436,95
Déviance du modèle avec données internes	184073,90	46056,7
Ecart de déviance (%) entre le modèle de données internes et le modèle sans donnée	-5,21	-4,91
Déviance du modèle avec données externes	189020,17	47158,77
Ecart de déviance (%) entre le modèle de données externes et le modèle sans donnée	-2,66	-2,64
Déviance du modèle avec données internes et externes	184096,85	45950,96
Ecart de déviance (%) entre le modèle de données internes et externes avec le modèle sans donnée	-5,20	-5,13

TABLE 4.9 – Analyse des déviances des différents modèles de fréquence

Lorsqu'on se focalise sur les déviances calculées pour l'ensemble des modèles sur la base de test, on observe que les déviances des modèles de données internes et des modèles de données mixtes ont des déviances similaires.

Concernant les modèles de fréquence, on observe une déviance nettement inférieure pour le modèle avec des données externes. L'écart de déviance par rapport au modèle sans variable calculé sur le modèle de fréquence avec des données externes est bien inférieur à celui observé sur les deux autres modèles de fréquence.

Pour les modèles de coût moyen, les observations vont dans le même sens que pour les modèle de fréquence, l'amélioration de la déviance est trois fois plus importante pour le modèle de coût moyen avec des données internes ainsi que celui avec des données mixtes comparé à celui avec des données externes.

Ainsi, on peut affirmer que malgré l'apport de nombreuses données externes, les modèles GLM avec uniquement des données externes sont bien moins performants que ceux avec des données internes ou des données mixtes. Les données externes ne suffisent pas pour la prédiction de la fréquence et du coût moyen. L'idée de construire un tarif avec la seule question de l'adresse à la souscription n'est donc pas envisageable. Par ailleurs les modèles mixtes sont aussi performants que les modèles de données internes. Ainsi, la substitution de certaines données internes par des données externes n'a pas d'impact sur la performance des modèles. Il est donc en revanche possible de réduire le nombre de questions à la souscription.

4.4 Interprétation des résultats

L'objectif de cette partie est d'analyser les résultats obtenus par les différents GLM mis en place.

4.4.1 Modélisation avec des données internes

Modèle de coût moyen sur les données internes

Le graphique ci-dessous permet de visualiser la manière dont chaque variable permet d'optimiser l'AIC du modèle. Dans le modèle de coût moyen sur les données internes, la variable optimisant de très loin l'AIC est la qualité juridique, puis vient la région et le nombre de pièces.

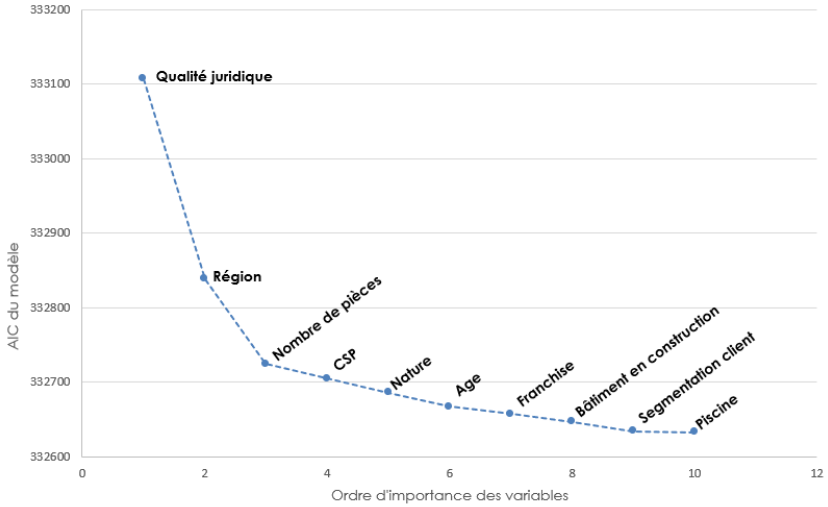

FIGURE 4.17 – Impact des 10 premières variables sélectionnées par la méthode *Forward* sur l'AIC du modèle de sévérité sur les données internes

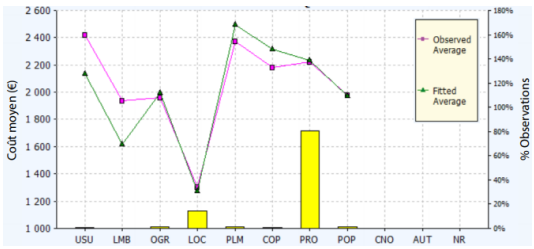

FIGURE 4.18 – Coût moyen observé et prédit pour le modèle de données internes en fonction de la variable qualité

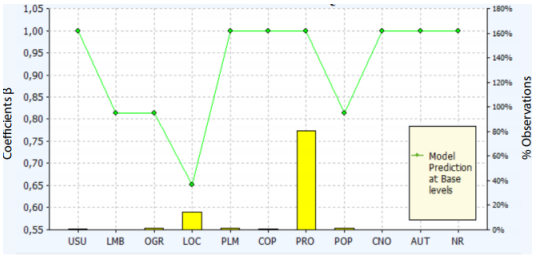

FIGURE 4.19 – Coefficients bêta associé aux modalités de la variable qualité juridique par le GLM

La variable la plus importante du modèle de sévérité sur les données internes est la variable qualité juridique. Le premier graphique ci-dessus nous permet de voir la tendance de la variable observée ainsi que de la variable prédite par le modèle après regroupement pour chacune des modalités présentées précédemment. Le deuxième graphique nous renseigne sur les coefficients β calculés par le modèle pour chacune des modalités de la variable qualité juridique. Aussi, on peut voir sur ce graphique, la répartition des observations selon les modalités. La modalité correspondant aux propriétaires est celle la plus représentée dans la base de données d'apprentissage, ainsi cette modalité est la modalité de référence et aura un coefficient β égal à 1. Les modalités regroupées avec la modalité propriétaire ont aussi un coefficient β qui vaut 1. La modalité locataire à un coefficient égal à 0,6513. Ainsi, toute chose égale par ailleurs, la sévérité d'un sinistre dégâts des eaux pour un locataire sera égal à 65,13 % de la sévérité d'un sinistre dégâts des eaux pour un propriétaire. L'une des explications est que lors d'un sinistre dégâts des eaux, les dégâts causés au contenant (en opposition au contenu du logement) sont à la charge du propriétaire et non du locataire ce qui réduit le coût pour le locataire.

Modèle de fréquence sur les données internes

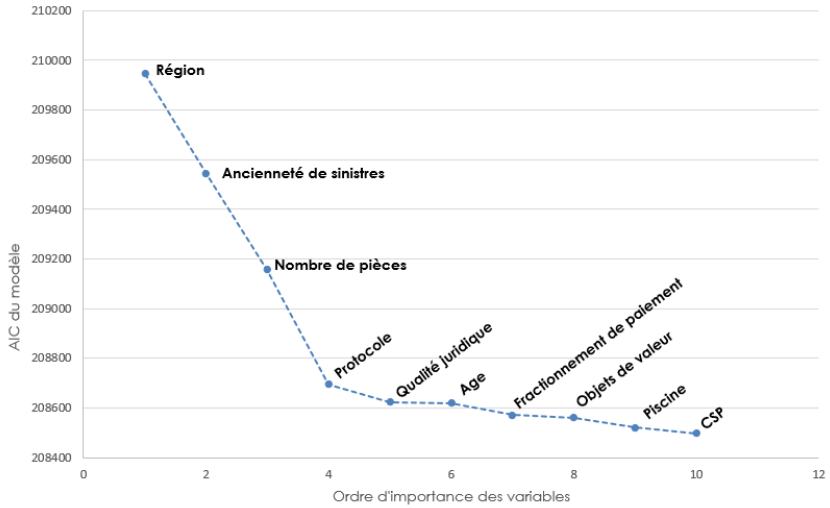

FIGURE 4.20 – Impact des 10 premières variables sélectionnées par la méthode *Forward* sur l'AIC du modèle de fréquence sur les données internes

Dans le modèle de fréquence sur les données internes, la variable optimisant de très loin l'AIC est la région, puis vient les antécédents de la sinistralité en dégâts des eaux et le nombre de pièces. Globalement, nous retrouvons des variables semblables à celles sélectionnées dans le modèle de coût moyen. Le fait que la région soit si significative témoigne de l'importance de la localisation géographique du bien assuré. L'apport de données externes permettant de décrire l'environnement géographique devient non seulement légitime mais optimise ainsi le modèle.

Nombre de pièces

La troisième variable la plus importante du modèle de fréquence pour les données internes est la variable correspondant au nombre de pièces du logement assuré, décomposée en 18 modalités correspondant respectivement au nombre de pièces du logement assuré de 1 à 16 et une modalité correspondant à 17 pièces ou plus. Il y a aussi la modalité NR qui regroupe l'ensemble des observations n'ayant pas d'information pour le nombre de pièces. On observe dans le graphique ci-dessous les coefficients β associés à chaque modalité par le GLM. La modalité de référence correspondant à celle ayant le plus d'observation (5 pièces) est attribuée à la modalité NR. Toute chose égale par ailleurs, plus un logement aura un nombre de pièces important et plus la fréquence d'un sinistre dégâts des eaux sera importante. Cela est cohérent, en effet de manière générale plus un logement a un nombre de pièces important et plus le nombre de pièces d'eau est important et ainsi le risque de survenance d'un dégâts des eaux augmente.

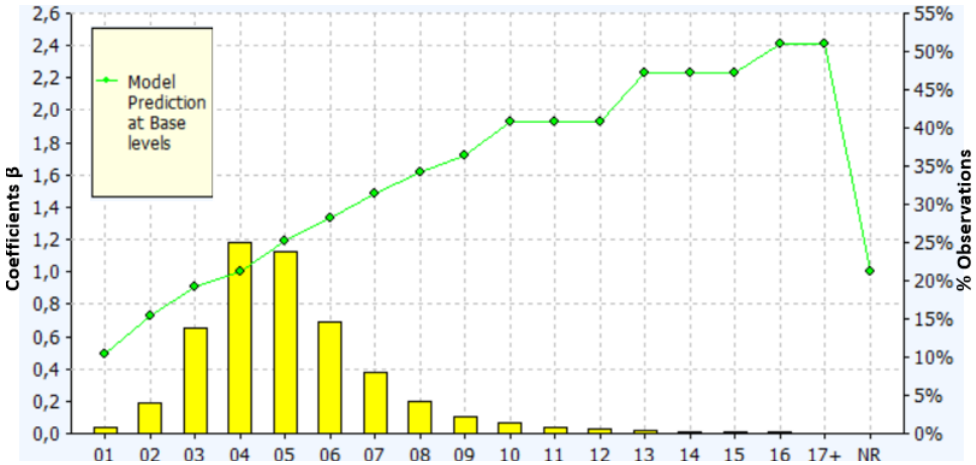

FIGURE 4.21 – Coefficients bêta associés aux modalités de la variables nombre de pièces par le GLM

4.4.2 Modélisation avec des données externes

Modèle de sévérité sur les données externes

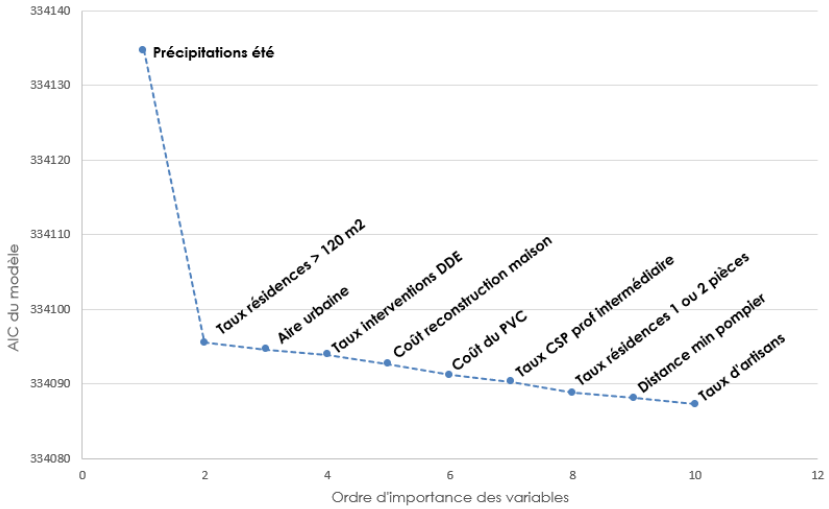

FIGURE 4.22 – Impact des 10 premières variables sélectionnées par la méthode *Forward* sur l'AIC du modèle de sévérité sur les données externes

La variable la plus importante du modèle de sévérité sur les données externes est la variable qui renseigne sur les précipitations en été, c'est à dire les précipitations moyennes en millimètre en été sur six ans d'historique. Cela peut s'expliquer par le fait que plus les pluies sont importantes, plus le risque d'inondation l'est. Si une inondation n'est pas

reconnue par un arrêté ministériel comme une catastrophe naturelle, alors cette derrière est un dégât des eaux.

Des variables de coûts sont aussi présentes parmi les variables les plus importantes du modèle de sévérité, on note le coût de reconstruction des maisons au mètre carré ainsi que le coût de PVC. En effet, lorsqu'un dégât des eaux se produit, si ce dernier est de faible ampleur alors un remplacement de la tuyauterie peut suffire et ainsi le coût du PVC est un très bon indicateur du coût du sinistre. Par ailleurs, lors de pluies violentes qui viennent à inonder des maisons entières, le coût de reconstruction des maisons permet d'apporter une information capitale pour le coût du sinistre. Les deux variables sont des variables continues qui ont été lissées avec un polynôme de degré 1. On observe une tendance croissante des coefficients β, ainsi toute chose égale par ailleurs, plus le coût de reconstruction des maisons ou celui du PVC augmente, plus le coût moyen du sinistre est élevé.

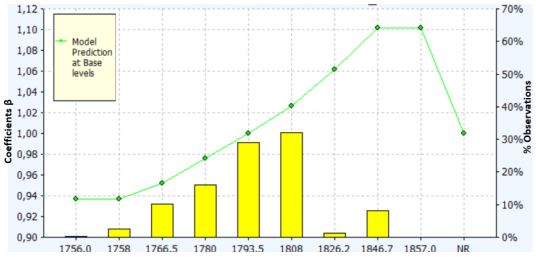

FIGURE 4.23 – Coefficients bêta associés aux modalités de la variable coût de reconstruction des maisons

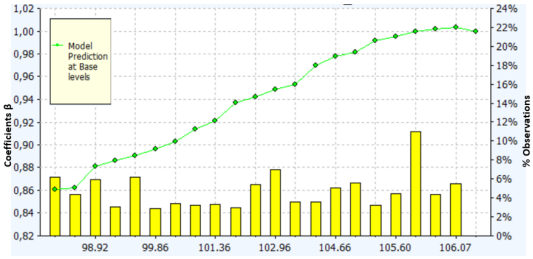

FIGURE 4.24 – Coefficients bêta associés aux modalités de la variable coût du PVC

Parmi les variables diminuant le plus l'AIC du modèle de coût moyen, on observe la variable correspondant à la distance minimale entre un logement et la caserne de pompiers la plus proche. Plus cette distance est grande et plus le coût du sinistre est élevé. En effet, plus le logement est loin d'une caserne de pompiers, plus le temps d'arrivée des pompiers est long et plus le dégât des eaux risque de s'amplifier. Dans le cas d'une inondation par exemple, il est important de pouvoir réagir rapidement.

Modèle de fréquence sur les données externes

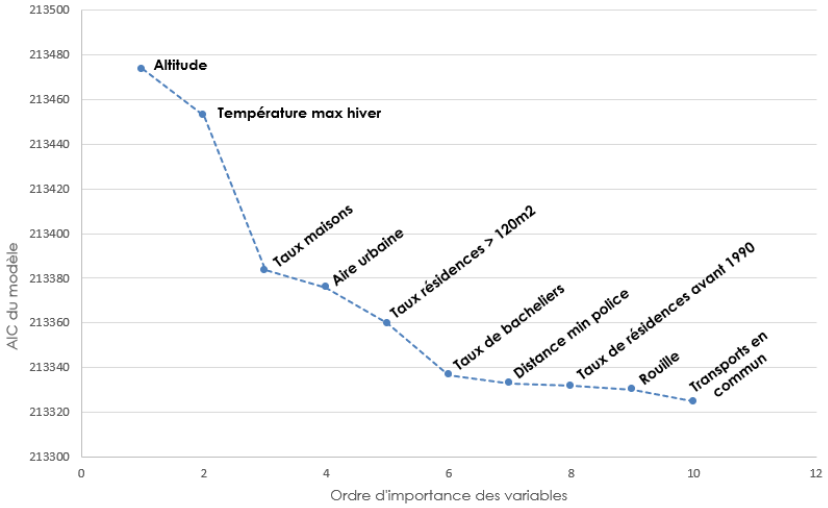

FIGURE 4.25 – Impact des 10 premières variables sélectionnées par la méthode *Forward* sur l'AIC du modèle de fréquences sur les données externes

Taux de résidences principales de plus de 120m²

La variable tx_rp_120m2p correspond au taux de résidences principales de plus de 120 mètres carrés parmi les résidences principales d'un IRIS, INSEE ou d'un département.

Pour une zone géographique donnée, on a :

$$tx_rp_120m2p = \frac{\text{Nombre de résidences principales de plus de } 120\text{m}^2}{\text{Nombre de résidences principales}} \tag{4.26}$$

Cette variable peut-être assimilée à la variable nombre de pièces présentées précédemment. En effet, plus le nombre de pièces est important et plus la superficie du logement le sera. Elle a été lissée avec un polynôme de degré 1. On peut visualiser que toute chose égale par ailleurs, plus le taux de logement supérieur à 120 mètres carrés dans pour une zone géographique est élevé et plus la fréquence du sinistre sera élevée. Les conclusions vont dans le même sens que pour la variable interne.

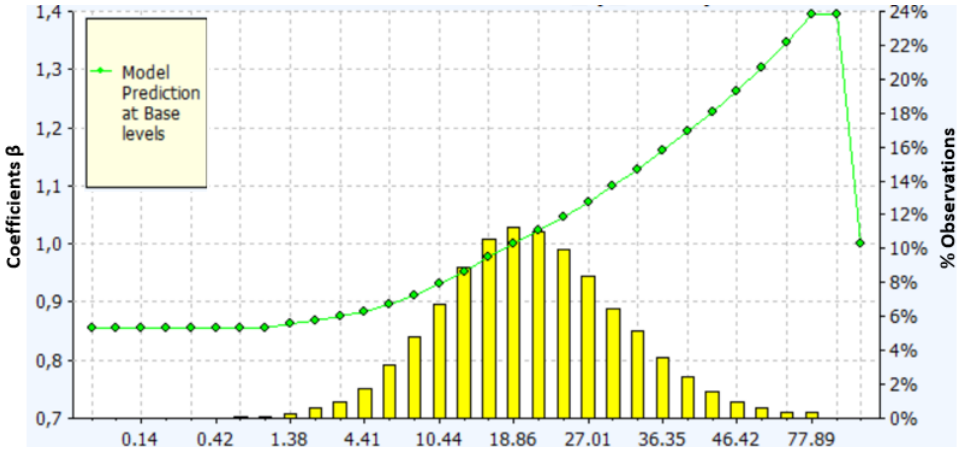

FIGURE 4.26 – Coefficients bêta lissés associés aux modalités de la variable continue du taux de résidence principale de plus de 120m² par le GLM

4.4.3 Modélisation avec des données internes et externes

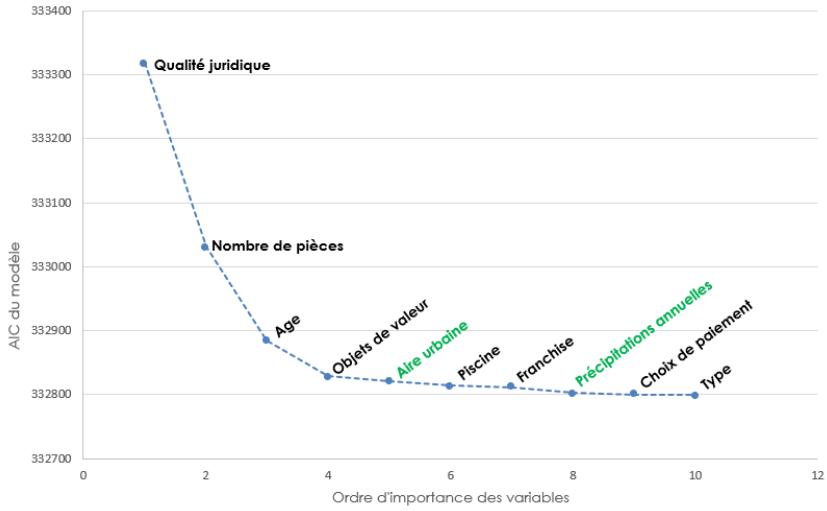

FIGURE 4.27 – Impact des 10 premières variables sélectionnées par la méthode Forward sur l'AIC du modèle de sévérité sur les données mixtes

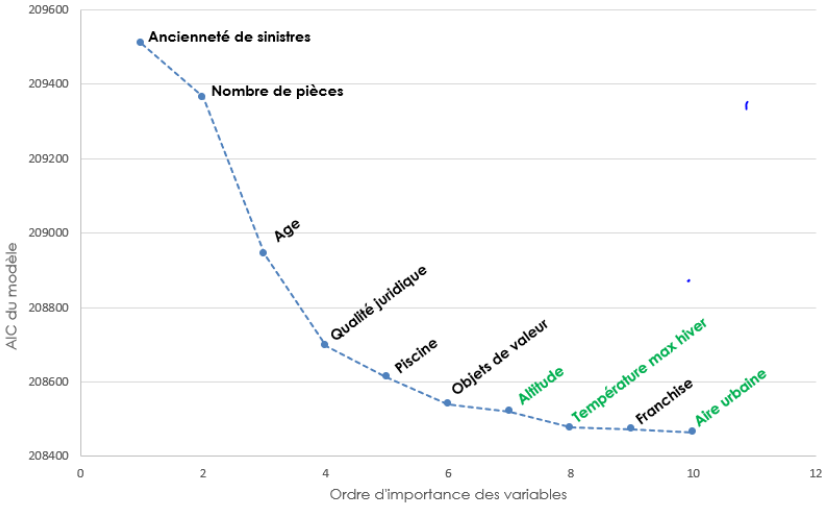

FIGURE 4.28 – Impact des 10 premières variables sélectionnées par la méthode *Forward* sur l'AIC du modèle de fréquence sur les données mixtes

Des données de températures ainsi que de précipitations sont présentes dans les variables explicatives des modèles mixtes. Concernant la température, la variable explicative présente dans le modèle de fréquence reporte le nombre de jours où la température est inférieure à zéro par zone géographique. Cette variable est cohérente puisqu'il est montré que lorsque la température est négative, les canalisations ont tendances à se casser et cela impacte alors directement la fréquence de sinistralité. De même, parmi les variables explicatives du modèle de fréquence on observe la variable renseignant sur la présence de mousses sur le toit ainsi que la présence de panneaux solaires. Lorsque de la mousse est présente sur le toit des logements, l'infiltration en eau est favorisé renforçant la possibilité d'obtenir un dégâts des eaux. De plus l'installation de panneaux solaires tend à fragiliser les toitures.

On observe que dans les modèles mixtes ce sont les mêmes variables que dans les modèles de données internes et de données externes qui ressortent de la sélection de variables *Forward* et qui sont significatives. Les variables sont très similaires entre le modèle de fréquence et le modèle de coût moyen, bien que le modèle de coût moyen prenne en compte des données de coût. Dans ces deux modèles on observe que la qualité juridique ainsi que le nombre de pièces et l'âge sont les variables les plus explicatives de la fréquence et du coût moyen. Cependant, la CSP qui était présente parmi les cinq variables les plus explicatives du modèle sévérité de données internes n'est plus présente dans le modèle de coût moyen sur les données mixtes. Cette dernière a été expliquée par des données externes telles que le taux de personnes exerçant une profession intermédiaire.

4.5 Réduction du nombre de questions à la souscription

4.5.1 Étude des variables latentes dans les modèles mixtes

Les modèles de prédictions réalisés dans le chapitre 3 ont permis de prédire l'âge et la nature sur les bases de données de 2018 et 2019. Ainsi, l'objectif de cette partie est d'analyser si les variables latentes créées permettent d'obtenir des modèles aussi performants qu'avec les variables internes récupérées à la souscription. Pour ce faire, les variables latentes ont été testées sur les modèles de données internes.

Les tableaux ci-dessous renseignent sur les déviances des différents modèles sur la base d'apprentissage et de test.

Âge récupéré à la souscription	Âge prédit par ML	Nature récupérée à la souscription	Nature prédite par ML	Ecart de déviance (%) sur la base d'apprentissage avec le modèle sans variable	Ecart de déviance (%) sur la base de test avec le modèle sans variable
X		X		-9,30	-8,38
X			X	-9,22	-8,14
	X	X		-9,28	-8,35
	X		X	-9,21	-8,13

TABLE 4.10 – Comparaison de la performance du modèle sévérité de données internes lors du remplacement de variables récupérées à la souscription par des variables latentes

Âge récupéré à la souscription	Âge prédit par ML	Nature récupérée à la souscription	Nature prédite par ML	Ecart de déviance (%) sur la base d'apprentissage avec le modèle sans variable	Ecart de déviance (%) sur la base de test avec le modèle sans variable
X		X		-5,21	-4,91
X			X	-5,19	-4,90
	X	X		-5,20	-4,91
	X		X	-5,18	-4,90

TABLE 4.11 – Comparaison de la performance du modèle fréquence de données internes lors du remplacement de variables récupérées à la souscription par des variables latentes

On observe dans un premier temps que les modèles sont stables puisque les écarts de déviances sont constants entre la base d'apprentissage et la base de test. De plus, les écarts de déviances sont constants sur la base de test lorsque l'on met les variables récupérées à

la souscription ou les variables latentes. Ainsi les variables latentes permettent d'obtenir des modèles aussi performants que les variables récupérées à la souscription.

L'ensemble des comparaisons suivantes seront réalisées entre le modèle de sévérité avec les âges et la nature récupérés à la souscription et le modèle de sévérité avec les âges et la nature prédits. Les comparaisons sont réalisées sur les modèles de données internes.

Lors que l'on observe les coefficients bêta calculés par le GLM pour les variables récupérées à la souscription et les variables prédites dans le modèle de données internes, on observe que les tendances sont similaires. De plus, on peut observer sur les graphiques ci-dessous que les observations sont réparties de manières équivalentes entres les différentes modalités, que ce soit pour la variable récupérées à la souscription ou la variable prédite.

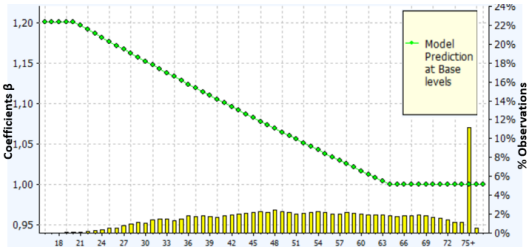

FIGURE 4.29 – Coefficients bêta de la variable âge à la souscription pour le modèle de sévérité

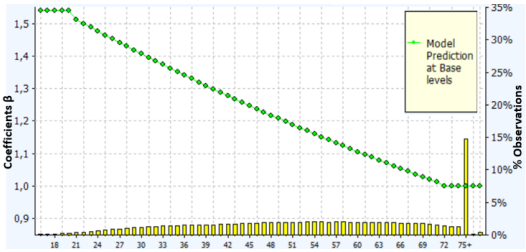

FIGURE 4.30 – Coefficients bêta de la variable âge prédite pour le modèle de sévérité

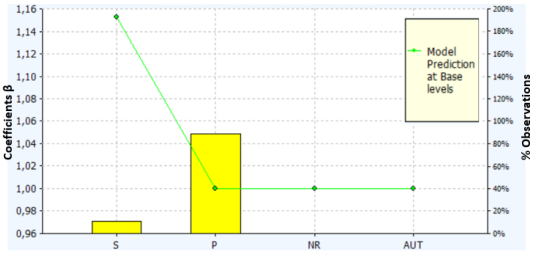

FIGURE 4.31 – Coefficients bêta de la variable nature à la souscription pour le modèle de sévérité

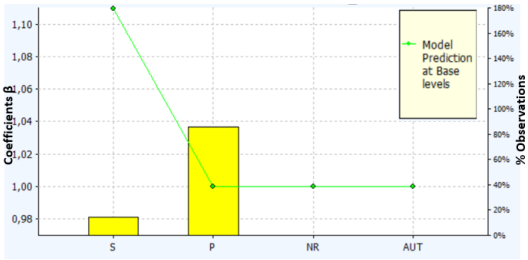

FIGURE 4.32 – Coefficients bêta de la variable nature prédite pour le modèle de sévérité

Le graphe ci-dessous représente les écarts de prédictions de la sévérité par rapport à la valeur observée en rose en fonction des différentes modalités de la variable nature. Le modèle de référence en bleu est celui avec les âges et la nature récupérés à la souscription. Le modèle représenté en vert correspond à celui avec les âges et la nature prédits par *Machine Learning*. On remarquera que les valeurs observées et les valeurs prédites

par le modèle avec les données récupérées à la souscription sont similaires. Les valeurs prédites par le modèle avec les variables latentes présentent un léger écart avec les valeurs observées.

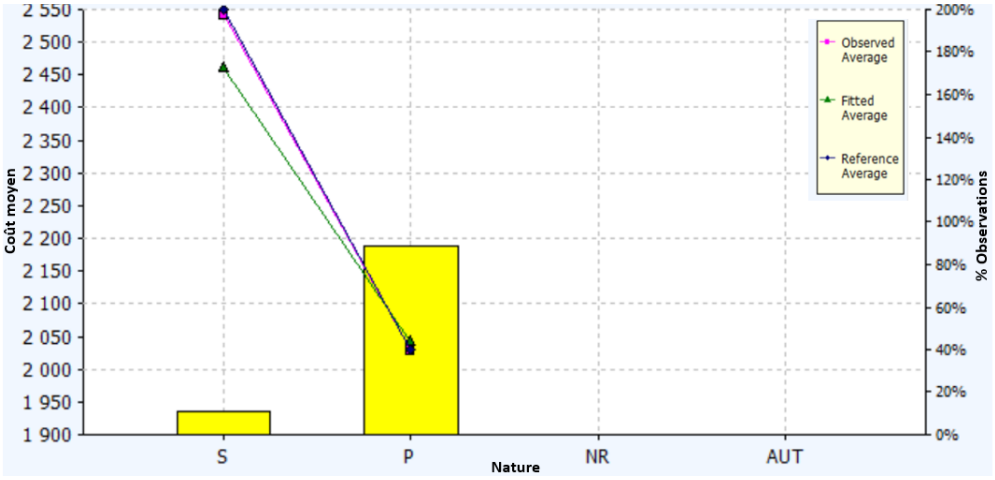

FIGURE 4.33 – Observation et prédictions de la sévérité en fonction de la nature pour le modèle de données internes avec les variables récupérées à la souscription (en bleu) et celui avec les variables latentes (en vert)

Le graphe ci-dessous représente les écarts de prédictions de la sévérité par rapport à la valeur observée en rose en fonction des différentes modalités de la variable âge. Le modèle de référence en bleu est celui avec les âges et la nature récupérés à la souscription. Le modèle représenté en vert correspond à celui avec les âges et la nature prédits par *Machine Learning*. On observe que les écarts de prédictions pour le modèle de sévérité sont équivalents lorsque l'on prend les variables récupérées à la souscription et les variables prédites.

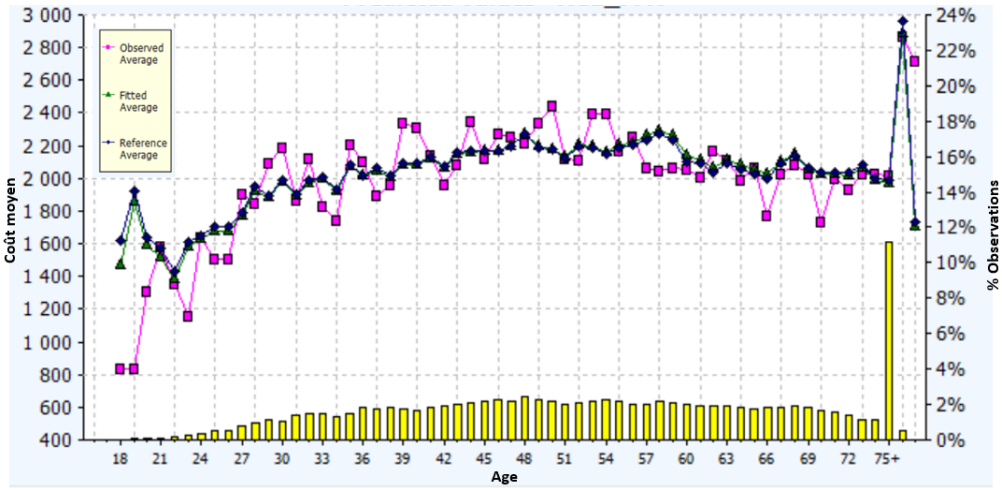

FIGURE 4.34 – Observation et prédictions de la sévérité en fonction de l'âge pour le modèle de données internes avec les variables récupérées à la souscription (en bleu) et celui avec les variables latentes (en vert)

La même analyse graphique a été réalisée sur les modèles de fréquences pour lesquels on observe des valeurs très proches entre les valeurs observées et les valeurs prédites par le modèle avec les variables récupérées à la souscription et les variables latentes.

Ainsi on peut conclure que les variables latentes prédites par méthode de *Machine Learning* permettent d'obtenir des modèles aussi performants que les variables récupérées à partir du questionnaire de souscription. De plus, l'avantage de la prédiction des variables latentes est qu'elles sont utilisables dans l'ensemble des modèles de chacune des garanties du produit MRH puisque les modèles de prédiction ont été entraînés sur la base de données totale sans la prise en compte de la sinistralité. Le même travail d'analyse de déviance pourra être réalisé sur les autres garanties avant la mise en production.

Cependant la limite de ces applications réside dans le fait que bien que les modèles de prédictions des variables latentes ont été entraînés sur la base de 2011 à 2017 et appliqués sur la base de données de 2018 à 2019, certains assurés dans les bases de données de 2018 et 2019 étaient aussi présents dans les bases de données de 2011 à 2017. Ainsi, certaines valeurs prédites sont appliquées sur les mêmes observations que celles sur lesquelles le modèle a été entraîné. Il faudrait donc prédire ces variables latentes sur des portefeuilles d'affaires nouvelles uniquement afin de confirmer la performance du modèle. Par ailleurs, rappelons que le modèle a été testé dans le chapitre 3 sur une base de test indépendante de la base d'apprentissage et que le modèle s'est avéré performant.

4.5.2 Sélection de données externes pour remplacer les données internes

L'objet de cette partie est d'analyser s'il y a des variables internes présentes dans le modèle de données internes et qui ne sont plus présentes dans le modèle de données mixtes. Cinq variables ont été identifiées. Ces cinq variables sont :

- La catégorie socio-professionnelle,
- L'état civil,
- La présence d'une alarme,
- La nature du logement,
- L'ancienneté du bâtiment.

Ces différentes variables ne sont plus nécessaires aux modèles de données mixtes, en effet elles ont étés expliquées par une ou plusieurs variables externes. La catégorie socio professionnelle par exemple est expliquée par les variables taux d'ouvriers de la classe supérieure ainsi que celle du taux de chômeurs par zone géographique pour le modèle de fréquence et le taux de personnes de profession intermédiaire dans le modèle de coût moyen. De même l'état civil se voit substitué par le taux de personnes divorcées. La présence d'une alarme peut quant à elle s'expliquer grâce à la variable renseignant sur la distance minimale au commissariat de police ou la gendarmerie le plus proche ainsi que l'aire urbaine. La variable nature s'explique dans les deux modèles par le taux de résidences principales par zone géographique. Enfin la variable ancienneté du bâtiment est expliquée par la variable taux de bâtiment construit avant 1990 par zone géographique pour les modèles de fréquence et de coût moyen.

On peut alors dire que l'apport de données externes dans la tarification GLM permet de réduire le nombre de variables internes nécessaires pour obtenir un modèle performant. Avec l'ensemble des données externes récupérées dans ce mémoire, on note la suppression de cinq variables internes résultant d'une question posée à la souscription pour chacune de ces variables.

Ainsi, ces variables étant expliquées par des données externes et n'étant plus nécessaires pour la performance des modèles GLM pour la garantie dégâts des eaux, le même travail devra être réalisé sur les autres garanties du produit MRH afin d'envisager ensuite de les supprimer du questionnaire et pouvoir l'alléger de cinq questions.

Conclusion

L'objectif de ce mémoire était d'étudier la faisabilité d'une réduction du nombre de questions posées par l'assureur lors de la souscription d'un contrat habitation tout en conservant la même performance de tarification dans un objectif d'optimisation du temps de souscription. Afin d'y parvenir, de nombreuses données externes ont été collectées : des données géographiques regroupant des données démographiques, socio-économiques et météorologiques issues de l'adresse de l'assuré ainsi que des données de coûts comme celui de la main d'œuvre d'un plombier ou le cours du cuivre intégrées via la date de survenance du sinistre.

Deux approches ont ensuite été mises en place. La première visait à utiliser des algorithmes de *Machine Learning* (*Random Forest* et *Gradient Boosting*), afin de prédire directement des données latentes correspondant à des questions posées à la souscription. Elle s'est révélée concluante puisque deux variables latentes, l'âge du souscripteur et la nature du logement assuré (principal ou secondaire) ont été bien prédites et leurs applications sur les modèles fréquence et coûts moyen ont permis à montrer qu'elle permettait d'obtenir des modèles aussi performants qu'avec les mêmes données issues d'informations recueillies auprès de l'assuré. Ainsi ces données latentes pourront donc être utilisées dans l'ensemble des modèles de chacune des garanties du produit MRH puisque les modèles de prédiction ont été entraînés sur une base de données totale, sans prise en compte de la sinistralité. Le temps d'exécution des algorithmes étant cependant très long, d'autres méthodes de *Machine Learning* pourraient être implémentées.

Il convient toutefois de souligner la limite de ces applications du fait que les modèles de prédictions des variables latentes ont été entraînés et testés sur la base de 2011 à 2017 et appliqués sur la base de données de 2018 à 2019. Certains assurés pouvant figurer dans chacune de ces bases, il pourrait être intéressant de prédire ces variables sur des portefeuilles d'affaires nouvelles uniquement pour confirmer totalement la stabilité et la robustesse dans le temps des modèles en présence de ces variables. A souligner cependant, comme indiqué au chapitre trois, que le modèle a été testé sur une base de test indépendante de la base d'apprentissage et que le modèle s'est révélé performant, en effet la RMSE obtenue pour la prédiction de l'âge vaut 2,42 et pour la nature l'AUC PR calculée vaut 0,88.

La deuxième approche consistait à intégrer directement les données externes à la modélisation GLM. Pour ce faire, trois modélisations de fréquence et de coût moyen pour la garantie dégâts des eaux ont été mises en place. Les résultats de performance sont clairs, malgré l'apport des nombreuses données externes variées récoltées dans le cadre de ce mémoire, les performances des modèles comportant uniquement des données externes sont bien inférieures à celles des modèles avec des données internes, collectées à la souscription, ou données mixtes. Plus précisément, le modèle de sévérité sur les données externes est trois fois moins performant et celui de la fréquence deux fois moins performant que les modèles sur les données internes ou mixtes. Il en résulte que la volonté de nombreux assureurs de se tourner vers une tarification où la seule question posée à l'assuré serait l'adresse du logement reste pour le moment inenvisageable avec les données récoltées. Toutefois, dans les modèles mixtes, certaines variables internes peuvent être remplacées par des variables externes tout en conservant une performance de modélisation identique. Ainsi cinq variables ont pu être identifiées pour la garantie dégâts des eaux : la CSP de l'assuré, la nature du logement, l'ancienneté du bâtiment, l'état civil et la présence d'une alarme au sein du logement. Il sera pertinent d'effectuer ce même travail sur les autres garanties du produit MRH afin d'envisager par la suite de supprimer cinq questions et ainsi poursuivre l'objectif d'alléger au maximum le questionnaire pour réduire la durée de la souscription.

Annexe A

Démonstrations

• Démonstration que la loi Gamma appartient à la famille exponentielle :
Soit Y $\sim \Gamma(\alpha, \lambda)$

$$f_Y(y) = \frac{\lambda^\alpha}{\Gamma(\alpha)} \cdot exp(-\lambda y) \cdot y^{\alpha-1} \cdot \mathbb{1}_{\mathbb{R}^+}(y) \tag{A.1}$$

$$f_Y(y) = exp\left[\frac{-y \cdot \frac{\lambda}{\alpha}}{\frac{1}{\alpha}} - \Big(ln(y\alpha) \cdot \alpha - ln(y) - ln\big(\Gamma(\alpha)\big)\Big)\right] \tag{A.2}$$

En posant les éléments ci-dessous, on obtient la même forme que la formule (4.1), ainsi la fonction de densité de la loi Gamma appartient à la famille exponentielle.

— $\theta = -\frac{\lambda}{\alpha}$
— $a(\phi) = \phi = \frac{1}{\alpha}$
— $b(\theta) = -ln(-\theta) = -ln(\frac{\lambda}{\alpha})$
— $c(y, \phi) = \frac{ln(\frac{y}{\phi})}{\phi} - ln(y) - ln\big(\Gamma(\frac{1}{\phi})\big) = ln(y\alpha) \cdot \alpha - ln(y) - ln\big(\Gamma(\alpha)\big)$

• Démonstration que la loi de Poisson appartient à la famille exponentielle :
Soit Y $\sim$ P(λ), alors :

$$f_Y(y) = \frac{\lambda^y}{y!} \cdot exp(-\lambda) = exp\left[\frac{y \cdot ln(\lambda) - \lambda}{1} - ln(y!)\right] \tag{A.3}$$

En posant les éléments ci-dessous, on obtient la même forme que la formule (4.1), ainsi la fonction de densité de la loi de Poisson appartient à la famille exponentielle.

— $\theta = ln(\lambda)$
— $a(\phi) = \phi = 1$
— $b(\theta) = exp(\theta) = \lambda$
— $c(y, \phi) = -ln(y!)$

Annexe B

Étude des résidus des modèles de fréquence

L'étude de l'analyse des résidus pour les modèles de fréquence a été réalisée sur les résidus *crunched*. Les résidus *crunched* sont formés lorsque les entrées de données sont regroupées : le résidu est calculé sur un groupe d'observation et non pas sur les observations individuelles comme le font les résidus de déviance. Les résidus *crunched* sont particulièrement utiles pour les modèles de fréquence où la plupart des observations ont une fréquence nulle et une poignée ont une fréquence non nulle. Les résidus *crunched* des graphiques ci-dessous sont des paquets de 500 observations.

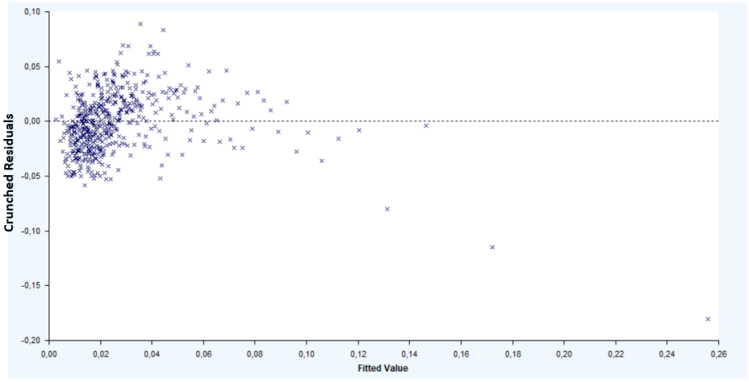

FIGURE B.1 – Résidus *crunched* pour le modèle de sévérité sur les données internes

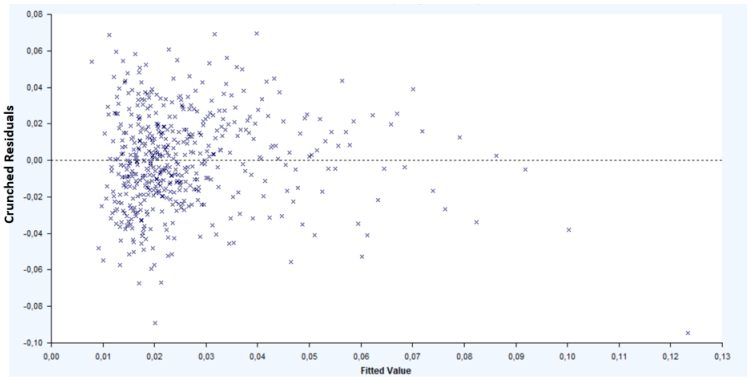

FIGURE B.2 – Résidus *crunched* pour le modèle de sévérité sur les données externes

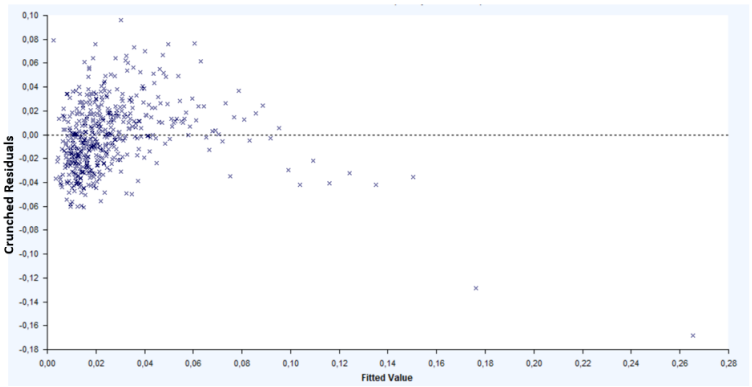

FIGURE B.3 – Résidus *crunched* pour le modèle de sévérité sur les données mixtes

Comme pour le coût moyen, une analyse graphique des résidus sur les modèles de fréquence montre que les résidus sont centrés autour de zéro pour le modèle sur les données internes et mixtes. Cependant sur le modèle de données externes les résidus sont plus dispersés, ce qui reflète une moins bonne prédiction de la fréquence de sinistralité par le modèle de données externes

Liste des abréviations

— AIC : *Akaike Information Criterion*

— AUC : *Area Under the Curve*

— AUC PR : *Area Under the Curve Precision Recall*

— CSP : Catégorie Socio-Professionnelle

— DDE : Dégâts des eaux

— FFA : Fédération Française de l'Assurance

— GLM : *Generalized Linear Model*

— IARD : Incendie Accidents et Risques Divers

— INSEE : Institut National de la Statistique et des Etudes Economiques

— IRIS : Ilots Regroupés pour des Indicateurs Statistiques

— MRH : Multirisque Habitation

— RC : Responsabilité Civile

— ROC : *Receiving Operating Characteristic*

— SMOTE : *Synthetic Minority Over-Sampling TEchnique*

Bibliographie

[1] HARDIN JAMES, W.(2001) : *Generalized Linear Models and Extensions.*

[2] M. DENUIT, A. CHARPENTIER(2010) : *Mathématiques de l'assurance non-vie.*

[3] CHARPENTIER, A.(2010) : *Statistique de l'assurance.*

[4] MORNET, A.(2015) : *Contributions à l'évaluation des risques en assurance tempête et automobile.*

[5] R. GENUER, JM. POGGI(2017) : *Arbres CART et Forêts aléatoires, Importance et sélection de variables.*

[6] GERON, A.(2017) : *Machine Learning avec Scikit-Learn.*

[7] RAKOTOMALALA, R.(2017) : *Machine Learning avec Scikit-Learn - Programmation Python.*

[8] PARIENTE, J.(2017) : *Modélisation du risque géographique en assurance habitation.*

[9] NOAA(2019) : *Données météorologiques :* `https://www.noaa.gov/`.

[10] LAMON, C.(2020) : *Cours sur les zoniers.*

[11] Gouvernement de la République française(2020) : *Données du gouvernements :* `https://www.data.gouv.fr/`.

[12] INSEE(2020) : *Données INSEE :* `https://www.insee.fr/`.

[13] FFA(2020) : *Etudes et chiffres clés :* `https://www.ffa-assurance.fr/etudes-et-chiffres-cles`.

[14] FERNANDEZ VILLALBA, V.(2020) : *Prise en compte des sinistres graves dans la modélisation de la sévérité en assurance Multirisque Habitation.*

[15] MARIE ROSE, J.(2021) : *Application sur R en tarification Non-Vie.*

Liste des tableaux

1 Performance des modèles de *Machine Learning* pour la prédiction de la variable nature . . . ix
2 Performances des modèles de *Machine Learning* pour la prédiction de la variable âge . . . ix
3 Écart de déviance (%) entre les modèles de données internes et les modèles sans variable sur la base de test . . . ix
4 Analyse des déviances sur la base de test des différents modèles . . . x
5 Performance of the Machine Learning models for the prediction of the nature variable . . . xvii
6 Performance of the Machine Learning models for the prediction of the age variable . . . xvii
7 Deviance gap (%) between models with internal data and models without variable on the test base . . . xvii
8 Deviance analyses on the test base for the different models . . . xviii

1.1 Moyenne de la fréquence et du coût moyen des sinistres pour les principales garanties MRH d'après les chiffres de la FFA . . . 5
1.2 Garanties obligatoires et facultatives du produit MRH d'Allianz France . . . 7
1.3 Charge de l'assureur et des assurés dans le cas de la mutualisation . . . 13
1.4 Charge de l'assureur et des assurés dans le cas de la segmentation . . . 13

2.1 Découpage de la base de données pour un contrat donné en fonction des différentes périodes . . . 20
2.2 Statistiques sur la qualité de géocodage des adresses du portefeuille MRH sur les périodes de 2011 à 2020 (arrondi au centième) . . . 30
2.3 Statistiques sur les mailles disponibles dans le portefeuille MRH . . . 30

3.1 Matrice de confusion dans le cas d'une classification multiclasses . . . 54
3.2 Matrice de confusion pour le *Random Forest* . . . 56
3.3 Importance des variables dans le modèle *Random Forest* pour la prédiction de la nature . . . 57
3.4 Matrice de confusion pour le *Gradient Boosting* . . . 58
3.5 Matrice de confusion pour la prédiction des classes d'âge par *Random Forest* 61

3.6 Métriques calculées pour chaque classe d'âge sur la prédiction *Random Forest* 61
3.7 Matrice de confusion pour la prédiction des âges par *Gradient Boosting* . . 63
3.8 Métriques calculées pour chaque classe d'âge prédites par *Gradient Boosting* 63
3.9 Importance des variables dans le modèle GBM 63

4.1 Fonction de lien utilisé en fonction de la loi de $Y|X = x$ 67
4.2 Force du lien statistique en fonction des valeurs du V de Cramer 71
4.3 Résultat du test de Kolmogorov-Smirnov 76
4.4 Variable qualité juridique sans regroupement dans le modèle sévérité de données internes 81
4.5 Variable qualité juridique après regroupement dans le modèle sévérité de données internes 81
4.6 Variables retenues pour les modèles de coût moyen 83
4.7 Variables retenues pour les modèles de fréquence 84
4.8 Analyse des déviances des différents modèles de sévérité 89
4.9 Analyse des déviances des différents modèles de fréquence 89
4.10 Comparaison de la performance du modèle sévérité de données internes lors du remplacement de variables récupérées à la souscription par des variables latentes 98
4.11 Comparaison de la performance du modèle fréquence de données internes lors du remplacement de variables récupérées à la souscription par des variables latentes 98

Table des figures

1 Présentation de l'ensemble de la base de données vii
2 Presentation of the whole database xv

1.1 L'assurance en France 3
1.2 Divisons de l'assurance de Biens et Responsabilités 4
1.3 Zones d'implantation du Groupe Allianz 6
1.4 Différentes mailles des données externes 14
1.5 Découpage en IRIS de la commune de Bordeaux 15
1.6 Schéma récapitulatif de la méthodologie du mémoire 16

2.1 Exposition de 2011 à 2020 du portefeuille MRH d'Allianz France par IRIS 18
2.2 Résumé de la structure de la base de données internes 21
2.3 Fréquence de sinistralité pour 2018, 2019 et 2020 22
2.4 Coût moyen des sinistres pour 2018, 2019 et 2020 23
2.5 Tendance de mobilité pour les lieux d'habitation 24
2.6 Tendance de mobilité pour les lieux de travail 25
2.7 Coût moyen des sinistres en fonction de la CSP pour différentes périodes . 25
2.8 Fréquence des sinistres en fonction du nombre de pièces pour différentes périodes 26
2.9 Fréquence et coût moyen en fonction de la variable région 27
2.10 Fréquence et coût moyen en fonction de la variable nombre de pièces . . . 27
2.11 Fréquence et coût moyen en fonction de la variable qualité juridique . . . 28
2.12 Fréquence et coût moyen en fonction de la variable CSP 29
2.13 Densité de population par IRIS 32
2.14 Taux de surface en eau par IRIS 32
2.15 Coût de reconstruction au mètre carré des maisons par département . . . 33
2.16 Prix moyen au mètre carré d'un logement par INSEE 33
2.17 Représentation de la distance angulaire entre deux points d'une sphère . . 34
2.18 Distance minimale entre une caserne de pompiers et un logement assuré . 35
2.19 Distance minimale entre un commissariat de police ou gendarmerie et le centroïde de l'IRIS d'un logement assuré 35
2.20 Zonier vent - Fréquence 38
2.21 Zonier vent - Sévérité 38
2.22 Output de l'outil de reconnaissance d'images satellites d'Allianz 40

2.23 Coût de la main d'oeuvre du plombier au cours du temps 41
2.24 Résumé de la structure de la base de données externes 41
2.25 Synthèse du découpage de la base de données 42

3.1 Nombre d'observations par tranches d'âge dans la base de données 2011-2017 44
3.2 Nombre d'observations par nature de logement dans la base de données 2011-2017 45
3.3 *Cross validation* pour k = 5 50
3.4 Schéma récapitulatif des étapes du *Machine Learning* 51
3.5 Matrice de confusion pour une classification binaire 52
3.6 Courbe ROC *Random Forest* 57
3.7 Courbe Précision-Rappel *Random Forest* 57
3.8 Courbe ROC *Gradient Boosting* 59
3.9 Courbe Précision-Rappel *Gradient Boosting* 59
3.10 Importance des variables dans le modèle *Gradient Boosting* 59
3.11 Variables importantes du modèle de *Random Forest* pour la prédiction de l'âge 62

4.1 Distributions des coût moyen avant suppression des pics 73
4.2 Distribution des coût moyen après suppression des pics 74
4.3 *Mean Excess* des coûts de sinistres après suppression du pic à 1350 € . . . 75
4.4 Adéquation graphique de la loi gamma à la distribution des coûts moyens 75
4.5 Adéquation graphique de la loi log normale à la distribution des coûts moyens 75
4.6 QQ-plot loi gamma 76
4.7 QQ-plot loi log-normale 76
4.8 Corrélogramme des variables explicatives pour les modèles de données internes 78
4.9 Différentes étapes de la construction de GLM sur Emblem 79
4.10 Coût moyen observé et prédit en fonction des modalités de la variable nombre de pièces 82
4.11 Coefficients bêta attribués par le GLM après lissage polynomial pour la variable nombre de pièces 82
4.12 Corrélogramme des variables explicatives sélectionnées pour le modèle de fréquence sur les données externes 85
4.13 Corrélogramme des variables explicatives sélectionnées pour le modèle de sévérité sur les données internes et externes 86
4.14 Résidus de déviance pour le modèle de sévérité sur les données internes . . 87
4.15 Résidus de déviance pour le modèle de sévérité sur les données externes . 87
4.16 Résidus de déviance pour le modèle de sévérité sur les données mixtes . . 88
4.17 Impact des 10 premières variables sélectionnées par la méthode *Forward* sur l'AIC du modèle de sévérité sur les données internes 90
4.18 Coût moyen observé et prédit pour le modèle de données internes en fonction de la variable qualité 91

4.19 Coefficients bêta associé aux modalités de la variable qualité juridique par le GLM . . . 91
4.20 Impact des 10 premières variables sélectionnées par la méthode *Forward* sur l'AIC du modèle de fréquence sur les données internes . . . 92
4.21 Coefficients bêta associés aux modalités de la variables nombre de pièces par le GLM . . . 93
4.22 Impact des 10 premières variables sélectionnées par la méthode *Forward* sur l'AIC du modèle de sévérité sur les données externes . . . 93
4.23 Coefficients bêta associés aux modalités de la variable coût de reconstruction des maisons . . . 94
4.24 Coefficients bêta associés aux modalités de la variable coût du PVC . . . 94
4.25 Impact des 10 premières variables sélectionnées par la méthode *Forward* sur l'AIC du modèle de fréquences sur les données externes . . . 95
4.26 Coefficients bêta lissés associés aux modalités de la variable continue du taux de résidence principale de plus de 120m^2 par le GLM . . . 96
4.27 Impact des 10 premières variables sélectionnées par la méthode Forward sur l'AIC du modèle de sévérité sur les données mixtes . . . 96
4.28 Impact des 10 premières variables sélectionnées par la méthode *Forward* sur l'AIC du modèle de fréquence sur les données mixtes . . . 97
4.29 Coefficients bêta de la variable âge à la souscription pour le modèle de sévérité . . . 99
4.30 Coefficients bêta de la variable âge prédite pour le modèle de sévérité . . . 99
4.31 Coefficients bêta de la variable nature à la souscription pour le modèle de sévérité . . . 99
4.32 Coefficients bêta de la variable nature prédite pour le modèle de sévérité . 99
4.33 Observation et prédictions de la sévérité en fonction de la nature pour le modèle de données internes avec les variables récupérées à la souscription (en bleu) et celui avec les variables latentes (en vert) . . . 100
4.34 Observation et prédictions de la sévérité en fonction de l'âge pour le modèle de données internes avec les variables récupérées à la souscription (en bleu) et celui avec les variables latentes (en vert) . . . 101

B.1 Résidus *crunched* pour le modèle de sévérité sur les données internes . . . 107
B.2 Résidus *crunched* pour le modèle de sévérité sur les données externes . . . 108
B.3 Résidus *crunched* pour le modèle de sévérité sur les données mixtes . . . 108