Mémoire présenté devant l'Université de Paris-Dauphine pour l'obtention du Certificat d'Actuaire de Paris-Dauphine et l'admission à l'Institut des Actuaires le 8 février 2021

Par : Nada Berrada
Titre : Élaboration de zoniers en assurance MRH

Confidentialité : ☑ Non ☐ Oui (Durée : ☐ 1 an ☐ 2 ans)

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

Membres présents du jury de l'Institut des Actuaires :

Entreprise :
Nom : Mazars Actuariat
Signature :

Directeur de Mémoire en entreprise :
Nom : Axel Truy,
Fatima Benabdelkrim

Membres présents du Jury du Certificat d'Actuaire de Paris-Dauphine :

Signatures :

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

Secrétariat :

Signature du responsable entreprise

Signature du candidat

Bibliothèque :

Résumé

En assurance IARD, la zone géographique constitue l'un des critères de segmentation les plus couramment employés par les assureurs. En effet, l'aspect géographique du risque permet de prendre en compte les informations relatives à l'environnement où évolue le contrat. Il constitue également un moyen de lutter contre le phénomène d'anti-selection.

Afin de retranscrire la géographie du risque en une variable exploitable dans des modèles statistiques, les assureurs définissent un découpage du territoire selon différents niveaux de risque. Cette segmentation du portefeuille est appelée un zonier. Il permet de synthétiser l'effet géographique en un seul critère tarifaire qui représente une variable latente non mesurable et non observable directement dans un portefeuille.

Ce mémoire vise à modéliser un zonier pour un produit d'assurance multirisque habitation, à partir de données d'un portefeuille d'assureur IARD français et de données en libre-service. Nous nous intéresserons notamment à la création de zoniers pour les garanties vol et dégâts des eaux en France métropolitaine.

Ce mémoire s'articule ainsi autour de trois axes majeurs. Dans un premier temps, il sera question de construire une base de données structurée d'apprentissage à partir de données externes issues de l'*Open Data*. Dans un deuxième temps, nous implémenterons les méthodologies de construction des zoniers en intégrant des techniques de *Machine Learning*. Enfin, nous procéderons à une étape d'analyse critique des résultats obtenus suite à l'intégration des zoniers dans le portefeuille d'étude.

Mots-clés : *Assurance IARD*, *MRH*, *Machine Learning*, *Zonier*, *Random Forest*, *Assurance non vie*, *Assurance multirisque habitation*, *Cross Validation*, *Optimisation*, *Tarification*, *Big Data*, *Open Data*

Abstract

In non-life insurance, the geographic area is one the most widely used criteria for risk segmentation. Indeed, the geographical aspect of the risk enables insurers to use information relating to the environment in which the contract evolves. It contributes as well to the protection against adverse selection.

In order to transcribe the geography of the risk into a variable that can be used in statistical models, most insurers define a territory segmentation based on risk levels. This portfolio segmentation is called a zoning plan. The zoning variable created synthetizes the geographical effect in one unique pricing criteria. It is a latent variable that cannot be directly measured and that is unobservable in a portfolio.

This thesis aims to model a zoning plan for a household multi-risk insurance product of a French non-life insurer. The focus is on the creation of the zoning variable for theft and water damage warranties in metropolitan France.

This thesis is built around three focal points. First, the building of a structured dataset from external Open Data. Second, the implementation of the zoning variable methodology. Finally, a critical analysis of the obtained results following the integration of the zoning variables will take place.

Keywords : *P&C insurance*, *MRH*, *Household Multirisk Insurance*, *Zoning variable*, *Random Forest*, *Cross Validation*, *Optimization*, *Pricing*, *Big Data*, *Open Data*

Note de Synthèse

La segmentation tarifaire représente un enjeu stratégique pour les assureurs IARD car elle participe à la lutte contre l'anti-sélection, à la compétitivité et à la pérennité du portefeuille. En effet, la tarification d'un contrat d'assurance s'appuie sur des variables déclaratives qui permettent à l'assureur de discriminer le risque représenté par l'assuré. En particulier, dans le marché de l'assurance Multirisque Habitation, l'assureur dispose de nombreuses informations relatives au contrat (franchise, garanties inclues, etc.), à l'assuré (CSP, nombre d'enfants, etc.) ainsi qu'au logement assuré (type de logement, nombre de pièces, adresse, etc.). L'enjeu étant d'améliorer la segmentation du portefeuille et proposer aux assurés un tarif représentatif de leur risque intrinsèque. Parmi ces variables déclaratives, les variables spatiales (adresse, code INSEE, coordonnées GPS) font l'objet d'une attention particulière car elles représentent une composante géographique du risque, discriminante sur de nombreuses garanties d'assurance IARD. Par exemple, les garanties vol et dégâts des eaux couvertes par les contrats MRH sont particulièrement concernées par la localisation géographique. Cela s'explique, entre autres, par une disparité de la criminalité sur le territoire français pour la garantie vol et à certains facteurs géographiques comme la température et les précipitations qui sont directement discriminants pour la garantie dégâts des eaux.

La zone géographique occupe donc une place centrale dans les critères de segmentation, l'estimation de la sinistralité étant souvent influencée par l'environnement géographique, météorologique ou socio-démographique du risque. Dès lors, la plupart des assureurs définissent un découpage du territoire selon différents niveaux de risque. Cette segmentation du portefeuille est appelée zonier. Il s'agit d'une segmentation homogène du territoire en différentes zones plus ou moins larges qui quantifient le risque auquel est exposé l'assuré selon son lieu de résidence.

D'autre part, l'émergence de l'*Open Data* a favorisé l'essor de nouvelles approches visant à tirer parti du pouvoir explicatif de certains facteurs externes sur la sinistralité propre aux assureurs. Les approches classiques uniquement basées sur les observations internes à l'assureur, sont délaissées au profit d'approches mixtes qui prennent également en compte ces facteurs externes. Ainsi, l'émergence de nouvelles données en libre-service ainsi que l'apparition de nouvelles techniques d'apprentissage automatique entraînent les actuaires à revoir leurs modèles de tarification de produits d'assurance. Ceux-ci reposent sur l'étude du risque géographique qui s'appuie sur des choix stratégiques concernant la nature des données externes à intégrer dans le modèle de tarification ; les techniques de jointure ou de lissage spatial nécessaires pour l'interpolation les données sur tout le territoire étudié ; la maille de construction du zonier et les modalités de discrétisation de l'espace ou encore le nombre de zones de risque à construire.

Dans ce contexte, ce mémoire s'attachera à présenter la démarche de conception de zoniers en assurance multirisque habitation, portant sur les garanties vol et dégâts des eaux en France métropolitaine et à la maille de la commune (code INSEE). L'enjeu principal sera de montrer comment l'exploitation des données en libre-service recueillies permettent une amélioration des modèles tarifaires des

assureurs.

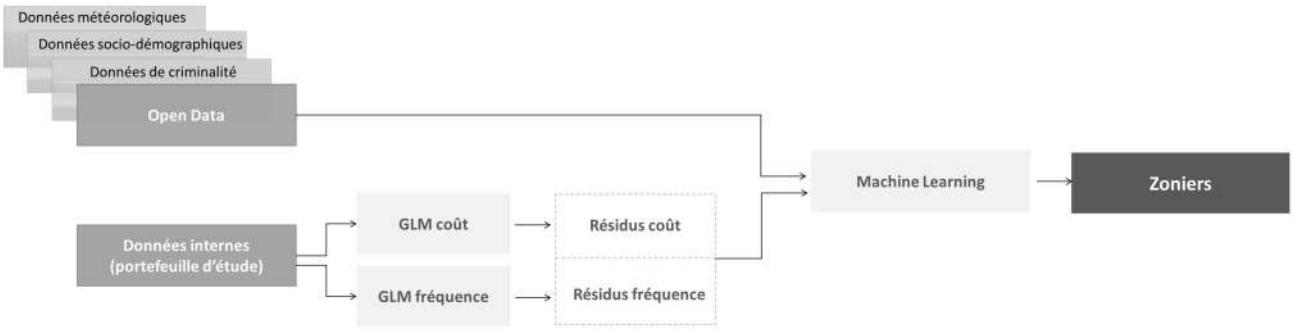


FIGURE 1: Étapes de construction d'un zonier

A l'image du schéma ci-dessus, nous avons suivi la démarche suivante afin de construire les zoniers :

Elaboration des modèles de coût et de fréquence des sinistres hors facteurs géographique

Cette première étape consiste à élaborer des modèles linéaires généralisés sur des variables préalablement sélectionnées ne prenant pas en compte les critères géographiques du risque. Cette étape est enrichie par la captation d'interaction entre les variables grâce à des méthodes de machine learning. Ces modèles nous permettront d'obtenir des résidus pour chacune des garanties considérées. Ceux-ci sont ensuite agrégés et lissés (par commune) afin d'obtenir une base de résidus des quatre modèles obtenus (ie. modèles de coût et de fréquence pour les garanties vol et dégâts des eaux) à la maille code INSEE.
La théorie de la crédibilité en assurance est à l'origine de la méthode de lissage utilisée ici. Cette étape permet de ne pas accorder trop de poids aux communes peu exposées dans le portefeuille d'étude. Les résidus lissés obtenus contiennent à priori l'information spatiale de la sinistralité. Pour s'en convaincre, nous avons établi des semivariogrammes dont les allures laissent présager qu'il demeure un effet géographique dans les résidus non expliqués par les modèles.

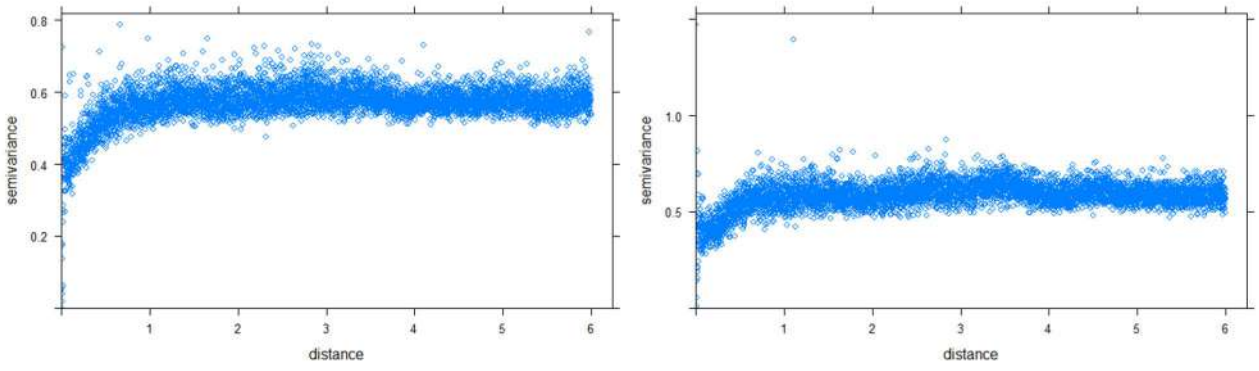


FIGURE 2: Semivariogramme empirique appliqué aux résidus des modèles de coût des garanties dégâts des eaux (à gauche) et vol (à droite)

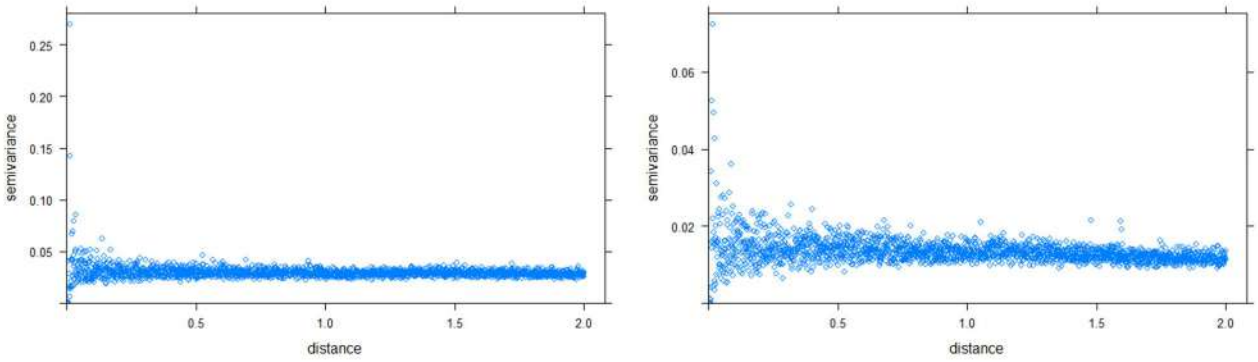


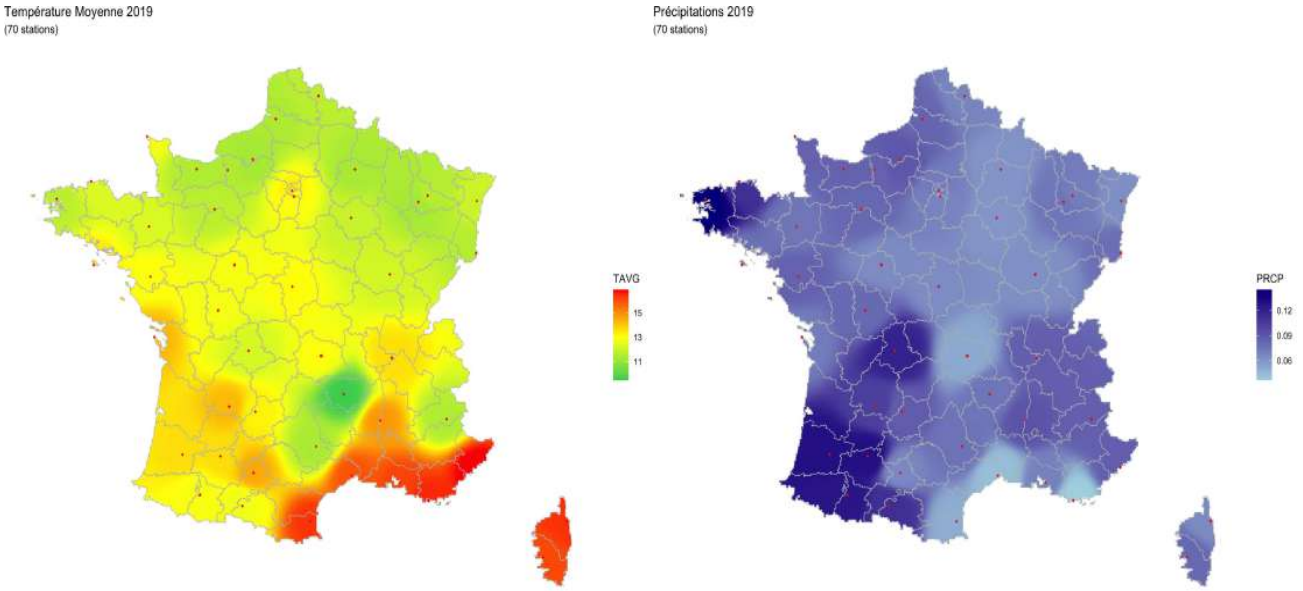
FIGURE 3: Semivariogramme empirique appliqué aux résidus des modèles de fréquence des garanties dégâts des eaux (à gauche) et vol (à droite)

Exploitation des données de l'*Open Data*

La donnée constitue la matière première des modèles tarifaires des assureurs. Ainsi, l'acquisition de celle-ci constitue l'enjeu principal de toute démarche de tarification. L'idée est ici de constituer la base de données externe qui permettra d'intégrer le facteur risque géographique aux modèles de départ à l'aide des données en libre-service. Les données externes exploitées sont issues de différentes sources et concernent plusieurs mailles géographiques, nous étudions en particulier :

- Des données de criminalité départementales provenant de l'Observatoire National de la Délinquance et des Réponses Pénales.
- Des données socio-démographiques à la maille INSEE issues du site officiel de l'Institut National de la Statistique et des Etudes Economiques.
- Des données à l'adresse de la station météo (coordonnées GPS) concernant des variables météorologiques qui proviennent du *National Centers for Environmental Information (NCEI)*.

Les données externes brutes nécessitent une étape préliminaire de retraitement. Par exemple, en ce qui concerne les données météorologiques, seules les données de certaines stations françaises sont disponibles sur le site du NCEI. Nous avons donc appliqué une méthode d'interpolation spatiale visant à couvrir l'ensemble du territoire français.



(a) Température moyenne (b) Précipitation

FIGURE 4: Cartes post-interpolation spatiale sur l'ensemble du territoire Français

L'ensemble des variables en libre-service retenues constituent la base de données externe qui permettra de modéliser les résidus dans l'étape suivante.

Explication des résidus des modèles par des variables géographiques externes

La modélisation des résidus contenant l'information géographique non traitée jusqu'ici constitue la première étape de la construction d'un zonier. Notre choix final s'est porté sur l'algorithme du *Random Forest* pour modéliser les résidus d'anscombe des GLM. Il s'agit d'une technique facile à interpréter et stable. Une étape supplémentaire d'optimisation des hyperparamètres a été implémentée pour renforcer la capacité explicative des modèles. A l'issue de cette étape, nous avons pu évaluer les attributs les plus importants dans la modélisation des résidus. Pour les modèles liés à la garantie dégâts des eaux, les forêts aléatoires attribuent un score d'importance significatif aux variables météorologiques. En revanche, pour la garantie vol, ce sont les variables de criminalité qui sont les plus importantes. Au sein de cette troisième partie, un apprentissage statistique par forêts aléatoires a permis une modélisation robuste de l'information spatiale contenue dans les résidus lissés, cette estimation aboutira à la création d'un zonier dans la partie suivante.

Construction et intégration du zonier créé dans le modèle de départ

Cette ultime étape est dédiée à la création d'une unique variable chargée de quantifier le risque géographique d'un logement. Il s'agit ici de rassembler l'ensemble de l'information extraite de l'*Open Data* en une seule et unique variable qui englobera le facteur risque géographique. Il convient donc de définir un nombre de classes de risque optimal et de recourir à des méthodes de clustering pour affecter une classe de risque à chaque commune. Le zonier ainsi construit est ajouté en tant que variable explicative des modèles de la première partie. L'apport du zonier sur le caractère prédictif d'un tarif comme son incidence sur le pouvoir de segmentation sont ici vérifiés par des mesures statistiques. Nous établissons ainsi une étude comparative des différents indicateurs statistiques qui permettent de juger de la qualité des modèles obtenus.

	VOL		DDE	
	AIC	Déviance	AIC	Déviance
GLM coût sans zonier	20 361	6 061	79 668	22 638
GLM coût avec zonier	20 173	5 916	79 282	22 354
GLM fréquence sans zonier	96 792	82 041	356 347	288 090
GLM fréquence avec zonier	93 032	78 272	349 705	281 440

FIGURE 5: Impact de l'introduction des zoniers comme nouvelles variables explicatives sur la modélisation de la fréquence et du coût des sinistres

Une nette amélioration des indicateurs est incontestable ce qui prouve l'importance de tenir compte du risque géographique dans la tarification. Pour finir, nous avons choisi d'analyser le modèle de coût de la garantie dégâts des eaux pour quantifier la plus-value de l'intégration du zonier dans le GLM. Pour cela, nous avons comparé les deux modèles suivants :

- Modèle 1 : Modèle GLM incluant les variables internes ainsi que l'ensemble des variables externes météorologiques. (Ajout de 10 variables)
- Modèle 2 : Modèle GLM incluant les variables internes ainsi que le zonier créé. (Ajout d'une unique variable)

	AIC	Déviance
GLM sans zonier	79 668	22 638
GLM avec variables externes (modèle 1)	79 655	22 616
GLM avec zonier (modèle 2)	79 282	22 354

FIGURE 6: Comparaison des différents modèles pour la garantie DDE

Ces résultats démontrent qu'utiliser le zonier construit permet d'obtenir de meilleurs indicateurs de performance qu'un GLM classique incluant des données internes et externes (amélioration de 0,5% de l'AIC). Ainsi, le zonier créé permet d'optimiser le modèle tout en gardant une bonne interprétabilité.

Au vu des résultats, nous pouvons conclure que l'ensemble des travaux réalisés dans le cadre de ce mémoire permettent d'obtenir une unique variable facilement interprétable qui apporte de l'information sur l'environnement géographique de l'assuré.

Synthesis note

Pricing segmentation is a strategic issue for P&C insurers because it contributes to the protection against adverse selection, the competitiveness and the sustainability of the portfolio. In fact, insurance companies provide insurance coverage based on identified risk variables. In the multi-risk home insurance market, variables such as the policyholder's socioprofessional category, its type of accommodation, the number of rooms included, the address etc. all contribute to improve the insurer's portfolio segmentation and to offer policyholders a premium that is representative of their intrinsic risk. Among these variables, the spatial variables (address, INSEE code, coordinates, etc.) must receive special attention as they represent a geographic component of the risk. In fact, they are discriminating on many property and casualty insurance coverages. For instance, theft and water damage warranties are both particularly affected by the geographical location of the risk. Concerning the theft guarantee, this is explained by the disparity of the crime rate on the french territory. As for the water damage guarantee, some geographical factors such as temperature and precipitation are directly discriminating.

The geographic area is a key segmentation criteria as the estimation of the loss exposure is often influenced by the geographic, meteorologic or sociodemographic environment of the risk. Henceforth, most insurers define their territory segmentation based on different levels of risk. This segmentation of the portfolio is called a zoning plan. It is a homogeneous segmentation of the territory in several zones varying in size which quantify the risk to which is exposed the insured based on his place of residence.

On the other hand, the emergence of Open Data has fostered the development of new approaches which aim is to extract the benefits of Open Data on risk segmentation and subsequently in the adjustment of premiums. Henceforth, the traditional approaches based solely on internal information are abandoned and replaced by mixed approaches that also take into account these external factors. Thus, the rise of Open Data and the emergence of new machine learning techniques has pushed actuaries to reconsider their insurance product pricing models. So far, a zoning variable is the most efficient way for insurers to translate the geographical structure of the risk into a variable that can be used in statistical models. The study of the geographic risk is based on strategic choices concerning the nature of the external data to be incorporated into the pricing model; the joining techniques and spatial smoothing methods necessary for the interpolation of the data over the entire study area; the grid of the zoning plan and the number of zones to build.

In this thesis, we will be presenting the approach that lead us to the creation of a zoning structure of both theft and water damage warranties in metropolitan France. The main challenge will be to show how the exploitation of the open data allows the improvement of the insurers' pricing models.

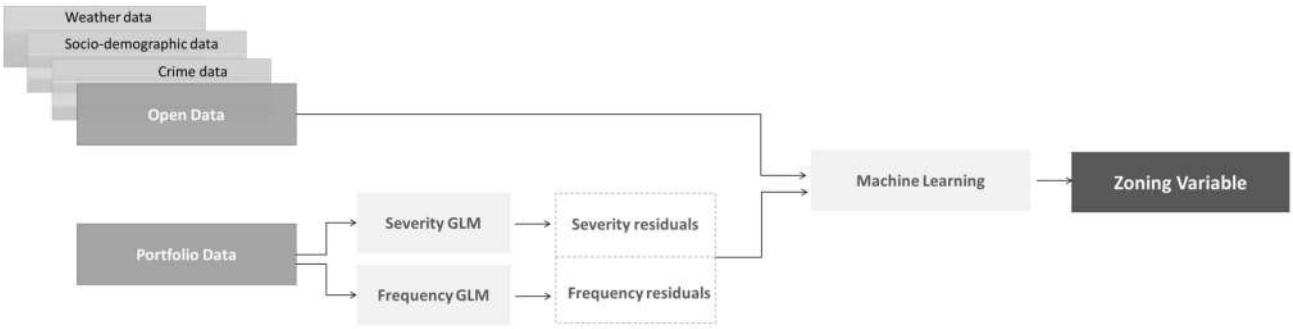


Figure 7: Implementation of a zoning system

As shown above, to elaborate zoning variables for both warranties studied in this thesis, the below approach was followed :

Elaboration of severity and frequency models excluding geographic factors

This first step consists in the elaboration of linear models generalized on preselected variables not taking into consideration the risk's geographic criteria. These models will define residuals for each considered warranty (theft and water damage). The aforementioned residuals are then aggregated and smoothed to obtain a residual dataset for the 4 models obtained at a granularity equivalent to INSEE codes (ie. severity and frequency models for theft and water damage warranty). The credibility theory in insurance is at the origine of the smoothing method used here. This step allows to give less importance to the municipalities that have minimal exposure in the portfolio. The obtained smoothed residuals contain a priori the spatial information of the loss exposure. As proof, a semivariogram was established and its appearance suggests that a geographic effect remains in the residuals that are not explained by the models.

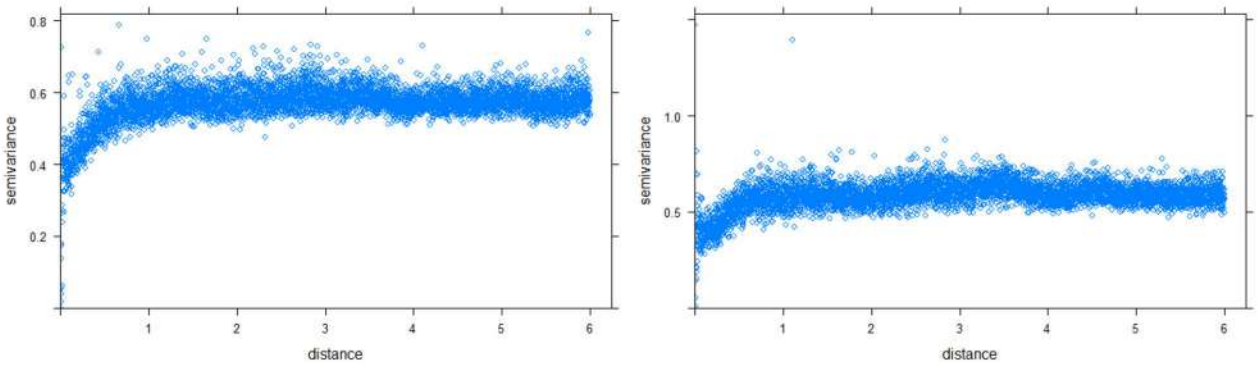


Figure 8: Empirical semivariogram applied to water damage warranty severity models' residuals.

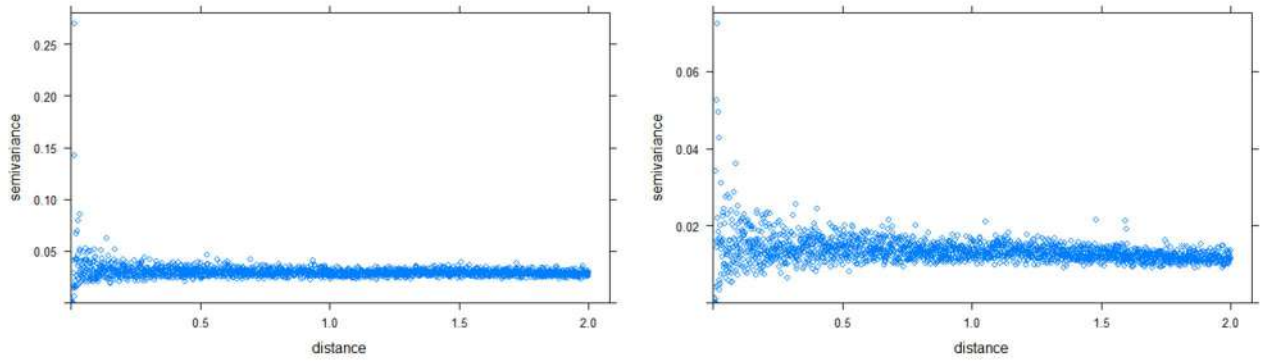


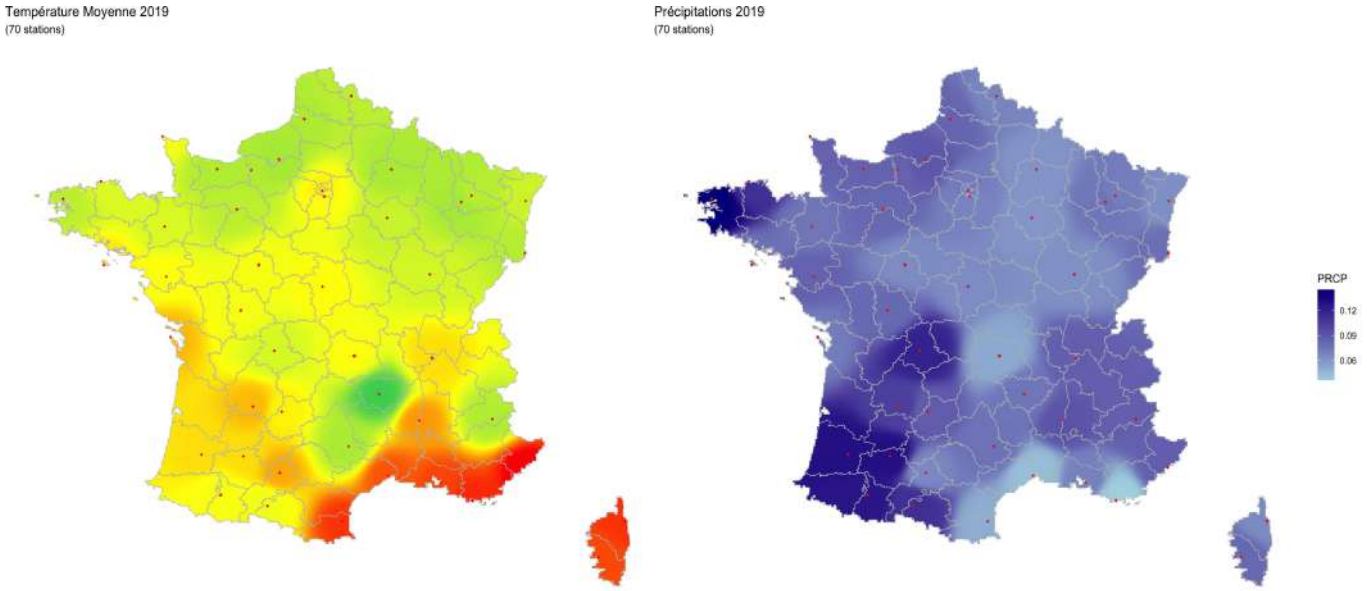
Figure 9: Empirical semivariogram applied to water damage warranty frequency models' residuals.

Open Data exploitation

Data is the insurers' pricing models' raw material. Thus, the acquisition of data is the main challenge for all pricing procedure. The idea here is to form the external database that will enable the integration of the geographic risk factor to the starting models using *Open Data*. The exploited external data comes from different sources and involve many geogrpahic levels. The focus is on :

- Departmental data on crime from the council guiding the national monitoring centre of delinquency and penal responses (ONDRP)
- Sociodemographic data at municipal level from the official website of the French National Institute of Economic and Statistical Information (INSEE)
- Longitude and latitude data concerning meteorologic variables from National Centers for Environmental Information (NCEI)

Raw external data require a preliminary step of reprocessing. For example, regarding meteorologic data, only the data of a certain number of french stations are available on the NCEI website. A spatial interpolation method was applied with the aim to cover the entire french territory and more specifically all the metropolitan France municipalities.



(a) Average temperature (b) Precipitation

Figure 10: Maps of the post spatial interpolation on the entire french territory

The set of retained Open Data variables form the external data base will enable the modeling of residuals in the next step.

Explanation of the residuals through external geographical data

The modeling of residuals containing geographic information forms the first step of the buidling of a zoning variable. The chosen algorithm to model GLM residuals is the *Random Forest*. It is a technique that is widely used as it is easy to interpret and stable. An additional step of hyperparameters optimization is essential to the accuracy of the results. At the end of this step, the most important attributes for the residuals modeling were evaluated. For the models linked to water damage warranties, random forests attribute a significant score to meteorologic variables. However, for theft warranty, the crime variables are more important. Within this third step, a random forest algorithm enabled a robust modeling of the spatial information contained in the smoothed residuals. This estimation will lead to the creation of a zoning variable in the next step.

Building and integration of the zoning variable generated in the starting model

This last step is dedicated to the creation of a unique variable in charge of quantifying the geographic risk of a household. The aim is to gather the information taken from Open Data in one unique variable that will incorporate the geographic risk factor. The built zoning variable is added as an explanatory variable of the first part's models. A comparative study of the different statistic indicators that allow to judge the quality of the obtained model is established.

	Theft		Water Damage	
	AIC	Deviance	AIC	Deviance
Severity GLM without zoning variable	20 361	6 061	79 668	22 638
Severity GLM with zoning variable	20 173	5 916	79 282	22 354
Frequency GLM without zoning variable	96 792	82 041	356 347	288 090
Frequency GLM with zoning variable	93 032	78 272	349 705	281 440

Figure 11: Impact of the zoning variable's insertion as a new explicative variable of the frequency and severity models

There is a clear improvment of the indicators which proves the importance of taking geographic risk into consideration for pricing.

Finally, the water damage severity model was analysed to quantify added value of the zoning variable's integration in the GLM. To do so, the following models were compared :

- Model 1 : GLM model the including internal variables and the set of external meteoroligic variables (addition of 10 variables)
- Model 2 : GLM model the including internal variables and the generated zoning variable (addition of a unique variable)

	AIC	Deviance
GLM without zoning variable	79 668	22 638
GLM with external variables (model 1)	79 655	22 616
GLM with zoning variable (model 2)	79 282	22 354

Figure 12: Comparison of the different models for water damage warranty

The results demonstrate that using the generated zoning variable allows a significant increase in the performance indicators when comparing to a GLM including internal and external data (0.48% improvement of the AIC). Thus, the generated zoning variable enables the optimization of the model while keeping a good interpretability.

To conclude, all the work carried out within this thesis allowed us to get a unique easily interpretable variable which provides information on the geographical environment of the risk.

Remerciements

Je tiens à remercier tout particulièrement mes tuteurs d'entreprise: Axel Truy, pour m'avoir fait confiance sur ce sujet d'étude que j'ai énormément apprécié. Ses conseils et son avis critique ont été très précieux. Fatima Benabdelkrim, pour son accueil, son encadrement et son sens de l'écoute qui ont rendu cette expérience particulièrement enrichissante. Mehdi Echchelh, pour ses réponses à mes interrogations techniques et surtout pour la bienveillance dont il a fait preuve, avec Fatima et Axel, depuis mon arrivée dans l'entreprise.

J'adresse notamment ma reconnaissance à ma tutrice académique de stage, Claire Lamon, pour son soutien, ses bons conseils et sa disponibilité qui m'ont permis d'avancer de manière stratégique tout au long de la rédaction de ce mémoire. Je remercie également toute l'équipe pédagogique de Dauphine.

Mes remerciements à Antoine Dourdy, Thierry-séphine Goma-Legernard et Marion Boivin Champeaux, de l'équipe actuariat de Mazars pour le travail et le temps qu'ils ont accordé à mon mémoire ainsi que leur soutien tout au long de mon stage.

Je souhaite, de plus, exprimer mes meilleurs sentiments aux autres stagiaires et à l'ensemble de l'équipe actuariat de Mazars, pour leur accueil et leur sympathie tout au long de mon stage. Je remercie en particulier Redouane Aalilou pour m'avoir fait partager ses connaissances dès mon arrivée.

Enfin, j'aimerais remercier mes parents ainsi que ma soeur pour leur soutien et leurs encouragements à chaque étape de mon parcours académique et de mon mémoire.

Table des matières

Introduction

Le développement de nouveaux algorithmes d'apprentissage ainsi que l'accès à de grandes quantités de données en libre-service ont représenté un tournant dans le monde de l'assurance. En particulier, dans un environnement concurrentiel, les assureurs doivent constamment améliorer et apporter de l'innovation dans leurs modèles de tarification afin d'adapter au mieux leurs primes pour être rentables et compétitifs sur le marché.

L'émergence de *l'Open Data* a favorisé l'essor de nouvelles approches visant à tirer parti du pouvoir explicatif de certains facteurs externes sur la sinistralité. Ainsi, les approches classiques de tarifications, basées uniquement sur les variables déclaratives (endogènes), sont complétées par des approches mixtes qui prennent également en compte ces facteurs externes. Ces nouvelles approches de tarification intègrent généralement des techniques de *Machine Learning* permettant une optimisation de la performance prédictive des algorithmes.

La composante géographique du risque représente un critère discriminant sur de nombreuses garanties d'assurances MRH, qui sera le périmètre d'étude pour ce mémoire. Dès lors, la plupart des assureurs définissent un découpage du territoire selon différents niveaux de risque. Cette segmentation du portefeuille, appelé zonier, est intégrée par les assureurs dans la tarification. Dans ce contexte, le présent mémoire a pour objectif de créer des zoniers robustes en utilisant des techniques de *data science* tout en gardant une bonne interpretabilité des résultats. Le périmètre de notre étude se limite aux garanties vol et dégâts des eaux qui sont généralement intégrés dans les contrats MRH.

Pour ce faire, nous suivrons une méthodologie visant à étudier l'impact de l'utilisation de nouvelles variables externes fournies à une maille géographique fine. Cette méthode est généralisable et n'est pas uniquement dédiée à l'assurance MRH. En effet, l'ensemble des techniques utilisées ainsi que l'interprétation des résultats pourront être appliqués à d'autres types d'assurance (comme la création d'un véhiculier en assurance auto).

L'étude se décompose alors en quatre parties qui constituent les différents chapitres :

- Contexte et présentation des étapes générales de réalisation du zonier.
- Implémentation des modèles de fréquence et de coût sans prise en compte des informations relatives à la zone géographique.
- Création de la base de données externe via les données recueillies de *l'open data*.
- Construction et intégration des zoniers dans les modèles de tarification.

Chapitre 1

Contexte général et enjeux du mémoire

1.1 Généralités sur l'assurance Multirisque Habitation

1.1.1 L'assurance IARD – zoom sur l'assurance MRH

L'assurance est l'opération par laquelle un assureur collecte des primes d'un ensemble d'assurés exposés à un risque et indemnise ceux d'entre eux qui subissent un sinistre grâce à la masse commune des primes collectées. Il est important de bien distinguer :

- Le souscripteur : personne qui signe le contrat et qui verse les primes.
- L'assuré : personne sur la tête de laquelle repose le risque assuré.
- Le bénéficiaire : personne au profit de laquelle a été souscrit le contrat.

Dans le cadre d'une opération d'assurance, le souscripteur s'engage à verser une prime et l'assureur s'engage à verser au bénéficiaire une prestation dépendant de la réalisation d'un risque aléatoire (survenance d'un sinistre). Ainsi, le versement de la prime intervient avant l'éventuel règlement de la prestation, c'est ce qu'on appelle l'inversion du cycle de production.

Les risques couverts en assurance peuvent concerner les risques liés à la vie ou la santé d'une personne (assurance vie ou prévoyance), ou les dommages matériels (assurance auto ou MRH). Dans ce contexte, le domaine de l'assurance peut être séparé en deux catégories, l'assurance de dommages et l'assurance de personnes.

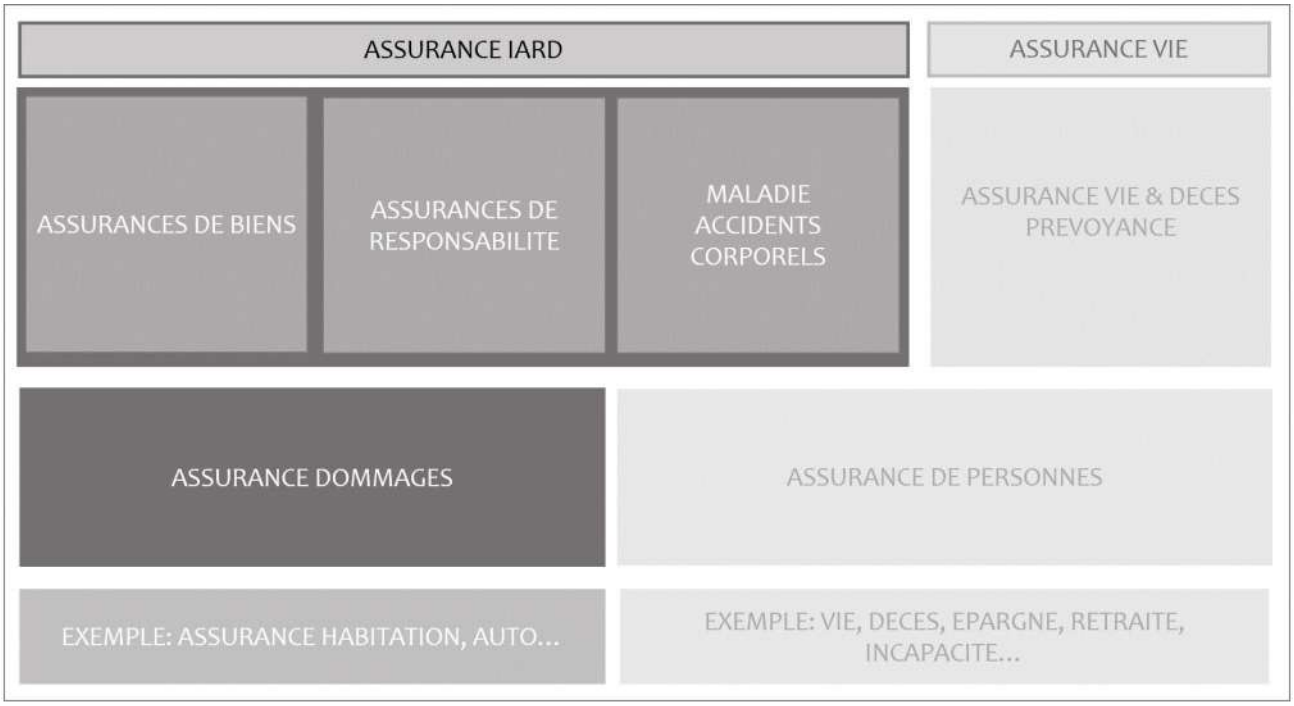


Figure 1.1: Les différents produits d'assurance

La distinction entre assurance dommages et assurance de personnes est principalement due aux risques garantis et à la différence fondamentale entre les obligations de l'assureur lors de l'exécution du contrat. En effet, les assurances de dommages garantissent les risques qui touchent aux biens de la personne suite à la survenance d'un évènement accidentel et involontaire alors que les assurances de personnes sont celles qui vont toucher directement à la personne assurée, et non à ses biens (accidents corporels, maladie, décès). De plus, dans le cadre de l'assurance dommages, le bénéficiaire doit être indemnisé du préjudice probablement subi et le montant des sinistres inconnus. En revanche, dans le cas de l'assurance de personnes, l'indemnisation qui sera versée au bénéficiaire est une somme forfaitaire déterminée au moment de la conclusion du contrat, indépendamment des préjudices subis.

Pour notre part, nous nous focaliserons sur l'assurance IARD (Incendie, Accidents et Risques Divers) qui englobe l'assurance dommage. Nous travaillerons plus spécifiquement sur l'assurance multirisque habitation, aussi appelée assurance MRH.

1.1.2 Vie d'un contrat en MRH

Le contrat d'assurance multirisque habitation est un contrat d'assurance qui permet de couvrir des risques tels que les incendies, dégâts des eaux et d'autres événements exceptionnels de la vie quotidienne.

Les étapes générales de la vie d'un contrat MRH sont les suivantes :

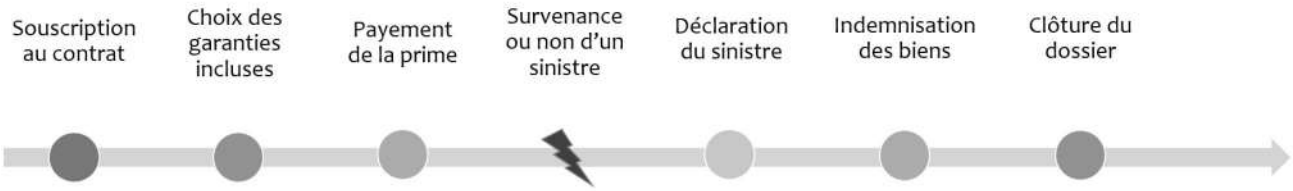


Figure 1.2: Vie d'un contrat MRH

Étape 1: Souscription au contrat

L'assurance multirisque habitation peut être obligatoire ou non selon le statut d'occupation du logement : alors qu'elle est fortement recommandée pour le propriétaire, le locataire a l'obligation de souscrire à un contrat d'assurance habitation. En effet, celui-ci doit obligatoirement s'assurer pour les risques locatifs afin de couvrir les dommages causés au logement par un dégât des eaux, une explosion ou encore un incendie.

Étape 2: Choix des garanties incluses dans le contrat

Le contrat se compose de deux types de garanties principales. D'une part, la garantie responsabilité civile privée permettant la protection du souscripteur ainsi que ses ayants droits ; de l'autre, les garanties du logement qui permettent de protéger le logement et les biens mobiliers en cas de dommages. Parmi les garanties proposées, nous retrouvons des garanties de base qui sont automatiquement incluses dans tous les contrats, assorties de garanties complémentaires qui sont optionnelles et qui diffèrent selon les assureurs. Par exemple, dans notre cas, nous considérerons l'étude d'une garantie de base, la garantie dégâts des eaux, mais également une garantie complémentaire, la garantie vol.

Les garanties les plus courantes d'un contrat MRH sont les suivantes :

a. Les garanties du logement

La garantie incendie-explosion : elle couvre les dégâts causés par le feu et la fumée ou encore les explosions.

La garantie bris de glace : elle couvre les dommages matériels (fissures, bris...) subis par les parties vitrées (portes, fenêtres, baies vitrées, les garde-corps, les parois séparatrices de balcons ...) ou le mobilier (écran TV, four...).

Les garanties évènements climatiques et catastrophes naturelles : elle prend en charge les dommages causés par des événements tels que les tempêtes, les tremblements de terre etc. à la suite de la parution d'un arrêté de catastrophes naturelles.

La garantie dégât des eaux : elle couvre les dégâts causés par l'eau aux biens immobiliers et mobiliers. Les dommages causés par les eaux peuvent être dues à des fuites, des ruptures et débordements de canalisations d'eau (douche, machine à laver, etc.), d'infiltrations au travers de toitures et des installations à l'intérieur des locaux etc.

La garantie vol : elle couvre les effractions et les vols survenus dans le logement assuré et prend en charge la disparition, la destruction ou la détérioration des biens mobiliers qui résultent de vols, tentatives de vol et d'actes de vandalisme commis dans les circonstances prévues au contrat.

b. Les garanties responsabilité civile

La garantie responsabilité civile vie privée : il s'agit d'une garantie qui n'est pas obligatoire et protège l'assuré et sa famille pour les dommages corporels, matériels ou immatériels causés à autrui. L'assureur se substitue au responsable pour indemniser la victime.

La garantie responsabilité civile risques locatifs : il s'agit de la responsabilité du locataire à l'égard du propriétaire pour risques locatifs. Elle est obligatoire pour le locataire et couvre ce dernier contre les dommages causés aux biens loués.

Étape 3: Paiement de la prime

Lorsqu'un assuré adhère à un contrat d'assurance multirisque habitation, celui-ci doit verser une prime d'assurance ou une cotisation. Une cotisation est le montant versé à la compagnie d'assurance, sur les bases fixées par les clauses du contrat. Celle-ci peut être payée mensuellement ou trimestriellement par exemple. Une prime d'assurance est le montant final qui est payé à échéance en fin d'année ; elle représente la somme de toutes les cotisations et est annuelle.

Étape 4: La survenance d'un sinistre

Le contrat d'assurance MRH a pour vocation de protéger les assurés d'un événement incertain, futur et aléatoire appelé le risque. Un risque est l'éventualité d'un évènement ne dépendant pas exclusivement de la volonté des parties et pouvant causer la perte d'un objet ou tout autre dommage, évènement contre la survenance duquel on s'assure.

Étape 5: La déclaration

L'assuré doit respecter un certain délai pour déclarer un sinistre à son assureur. En général, l'assuré doit déclarer le sinistre au plus tard dans les cinq jours à partir du jour où il en a connaissance. Toutefois, en cas de vol, ce délai est réduit à 2 jours ouvrés. A l'inverse, en cas de catastrophe naturelle, celui-ci est de 10 jours ouvrés à compter de la parution de l'arrêté portant reconnaissance de l'état de catastrophe naturelle.

Étape 6: L'indemnisation des biens mobiliers

A la suite de la survenance d'un sinistre, le recours à l'expertise est décidé par l'assureur qui doit indemniser celui-ci. Le rôle de l'expert est de déterminer les circonstances du sinistre, d'évaluer le montant des dégâts et de préconiser les modalités de réparation. En ce qui concerne le calcul de l'indemnité, en fonction de l'expertise effectuée, de l'ensemble des risques inclus dans le contrat et de l'ensemble des biens couverts, l'assureur propose un règlement en déduisant les deux éléments suivants prévus dans le contrat :

- La franchise : il s'agit du montant forfaitaire restant à la charge du souscripteur, qui s'ajoute donc en cas de sinistre à la prime payée par l'assuré dans l'analyse d'utilité éventuelle qu'il peut mener.
- La vétusté : l'indemnisation d'un bien tient compte de sa vétusté car les biens perdent souvent un peu de leur valeur au fil du temps. Le règlement ne sera alors qu'une portion de la valeur assuré.

Il y a réduction de l'indemnité lorsque que les montants du sinistre excédent la valeur assurée. Dans ce cas, l'indemnité est réduite dans la proportion de la sous déclaration entre la somme assurée et celle qui aurait dû l'être. C'est ce qu'on appelle la règle proportionnelle.

Les sinistres sont parfois réglés, en fonction de leurs montants, par application de conventions de règlements. Pour la garantie dégâts des eaux et incendie, les conventions CIDRE et IRSI * (Indemnisation et recours des sinistres immeuble) ont pour avantage de réduire considérablement le temps d'indemnisation des assurés grâce au respect d'un barème commun et à la prise en charge de

*Voir l'annexe A.1 pour plus d'informations sur les conventions CIDRE et IRSI

l'indemnisation de leurs assurés par les sociétés d'assurance. En d'autres termes, ces conventions permettent de régler plus rapidement le sinistre à l'assuré et de laisser les assureurs exercer les recours entre eux par compensation. Avant le 1^{er} juin 2018, en cas de sinistre dégâts des eaux ou incendie de faible ampleur (inférieur à 1600€), la convention CIDRE prévoyait en France que ce soit l'assureur de l'occupant et non celui du responsable qui prenait en charge l'indemnisation. Cela est dû à la forte fréquence de ces sinistres mais surtout du coût élevé de leur gestion. Depuis $le 1^{er}$ juin, tous les dommages inférieurs à 5000€ hors taxe rentrent dans le cadre de la convention IRSI. C'est l'assureur du local sinistré (le "gestionnaire" du dossier) qui s'occupe de gérer le sinistre dégât des eaux.

Étape 7: La prescription et le règlement de sinistre

La prescription est le délai au terme duquel il n'est plus possible d'agir. L'assureur et l'assuré sont alors libérés de leurs obligations : l'assureur n'a plus l'obligation d'indemniser un sinistre et l'assuré n'est plus contraint à payer les primes qui lui seraient réclamées. Le Code des Assurances fixe ce délai à deux ans à compter de la date du sinistre.

1.1.3 Chiffres clés et risques couverts dans le mémoire

L'assurance multirisque habitation, avec l'assurance automobile, constitue un des piliers de l'assurance de biens et de responsabilité.

Selon les dernières informations restituées du site de la Fédération Française de l'assurance https://www.ffa-assurance.fr/etudes-et-chiffres-cles/assurance-habitation-en-2018, en 2018, avec un total de 41,9 millions de contrats, le nombre de contrats Multirisque Habitation (MRH) a progressé de 2,0 %. Ce chiffre représente une augmentation de + 1,6 % pour les contrats occupants et + 4,2% pour les contrats non occupants par rapport à l'année d'avant. Pour rappel, l'assurance propriétaire non-occupant appartient à la catégorie des assurances de responsabilité civile. Il s'agit d'une garantie particulière souscrite par le bailleur d'un bien immobilier qu'il n'occupe pas (pendant une période de vacance entre deux locataires, pour se prémunir des risques non-couverts par l'assurance habitation de son locataire, pour se couvrir en cas de mise en jeu de sa responsabilité civile de propriétaire etc.). Parallèlement, le nombre de logements en 2018 est estimé à 36,8 millions, soit une hausse de 1,1%.

L'ensemble des cotisations des contrats MRH toutes garanties confondues enregistrent une hausse des cotisations de 3,5% et atteint une valeur estimé à 10,5 milliards d'euros. En ce qui concerne la prime moyenne, celle-ci a crû de 1,6% soit moins rapidement que l'indice FFB * qui a crû de 2,3% sur la même période.

Toutefois, cela n'a qu'un faible impact sur le coût de l'assurance habitation. La hausse des prix est également due à plusieurs années riches en catastrophes naturelles. En 2018, la fréquence globale des contrats MRH (y compris les catastrophes naturelles) est en hausse de 7,9% par rapport à 2017. Parallèlement, le coût moyen des sinistres est orienté à la hausse (+ 9,2%) après une baisse sensible en 2017 (- 7,2%). Selon la Fédération Française de l'Assurance, la fréquence des sinistres pour la garantie dégâts des eaux a connu une nette augmentation de +35,6% comparé aux autres garanties qui restent aux alentours de +7% par rapport à l'année précédente. De plus, la garantie vol a également connu une augmentation considérable de sa fréquence des sinistres avec une hausse de 9% en 2018 par rapport à l'année précédente.

Par ailleurs, l'entrée en vigueur de la loi Hamon le 26 juillet 2014, a impacté les prix de l'assurance

*L'indice de la Fédération Française du Bâtiment, ou indice FFB, est un indicateur basé sur le coût de la construction d'un immeuble. Réévalué tous les trimestres, il sert de référence aux compagnies d'assurance pour calculer le tarif des contrats d'habitation.

habitation à la hausse. En effet, cette loi, comportant notamment tout un volet concernant l'assurance habitation, vient renforcer les droits des particuliers et leur permettre de changer à tout moment d'assurance habitation après la première année de leur contrat.

1.2 L'apport de l'*Open Data* dans le monde de l'assurance

1.2.1 Les données en libre-service

Les données en libre-service, ou *open data*, sont destinées à la recherche, aux statistiques, à l'aide à la décision et à l'information du public. Elles se caractérisent par un droit d'accès, de consultation et de réutilisation des données librement. De nos jours, ces données offrent de nombreuses opportunités pour étendre le savoir humain et enrichir les connaissances.

En général, les données ouvertes sont mises à disposition à travers des sites web publics ou privés et sont maintenues par divers organismes et secteurs (la météorologie, l'environnement etc.).

Il existe aujourd'hui de nombreuses plateformes fiables qui permettent l'accès au grand public à tout type d'informations.

Les trois plateformes essentiellement utilisées dans le cadre de notre étude sont la base de données sociodémographique (INSEE), la base de données météorologique (NCEI) et la base de données de criminalité (ONDRP). Elles seront présentées de manière plus détaillée dans la suite de ce mémoire (cf. partie 2).

Pour chacune de ces plateformes de données en libre-service, nous avons procédé aux étapes suivantes afin de constituer notre base de données externe dans le but d'expliquer le risque lié à la zone:

- Sélection des variables
- Retraitement des données
- Identification des valeurs aberrantes et/ou mal renseignées

1.2.2 Utilité de l'open data dans le monde de l'assurance

Dans le monde de l'assurance, jusqu'à l'apparition de *l'open data*, les actuaires n'utilisaient que les données internes sur le profil de l'assuré pour tarifer un produit. L'apparition des données en libre-service a représenté une opportunité pour les assureurs d'être plus compétitif sur le marché en évaluant les risques de façon individualisée mais aussi de combattre le phénomène d'antisélection en intégrant des données externes supplémentaires dans le calcul de la prime. Cela a été le cas dans plusieurs secteurs de l'assurance et notamment dans le secteur de l'habitation.

En France, la protection des données personnelles est encadrée par la loi dite "Informatique et libertés" qui a été modifiée pour l'adapter aux dispositions du Règlement Général sur la Protection des Données (RGPD), applicable partout en Europe depuis le 25 mai 2018. Ce nouveau cadre juridique responsabilise les acteurs traitant les données personnelles des citoyens Européens. Le RGPD s'applique aux entreprises, aux organismes publics et aux associations quelles que soient leur taille ou leur activité, dès lors qu'ils traitent des données personnelles de personnes physiques se trouvant sur le territoire de l'Union européenne. Ce sont les autorités indépendantes de chaque Etat qui contrôlent l'application de la législation relative à la protection des données, en particulier, en France, il s'agit de la CNIL. Ainsi, trouver un équilibre entre individualisation des risques et vie privée est la délicate équation que doivent résoudre les assureurs exploitant les données personnelles de leurs clients.

Néanmoins, les données provenant du gouvernement et du secteur public sont des données auxquelles l'accès est totalement libre de droit, au même titre que l'exploitation et la réutilisation. Ces données offrent de nombreuses opportunités pour étendre le savoir humain et sont légalement et techniquement ouvertes. L'ouverture légale signifie que l'accès, l'exploitation, le partage et la modification des données sont légaux. Généralement, une licence prévue à cet effet, autorise l'accès libre et la réutilisation. De plus, en règle générale, ces types de données sont uniquement des données non personnelles. C'est-à-dire qu'ils ne comportent aucune information sur des individus pour des raisons de respect de la vie privée.

1.3 Principe de tarification et intégration du risque

1.3.1 Principe de tarification en IARD

La tarification des produits d'assurance consiste à calculer la prime pure d'un produit et de lui ajouter par la suite les différents chargements liés aux frais de fonctionnement et la marge de l'assureur.

L'estimation de la prime pure est capitale dans la construction de la prime d'une police d'assurance:

Figure 1.3: Etapes de tarification

La réalisation d'un tarif en assurance IARD peut s'appuyer sur l'analyse de la prime pure dans le cadre d'un modèle coût fréquence dans lequel l'effet des variables explicatives sur le niveau du risque est modélisé par des modèles de régression de type GLM.

Le modèle coût-fréquence :

Commençons par prendre en compte le cas d'un assuré. Pour calculer la prime pure, on considère deux variables supposées indépendantes :

- La fréquence de sinistres qui représente le nombre de sinistres survenus durant la période d'exposition.

$$Frequence = \frac{Nombre\ de\ sinistres}{Exposition} \tag{1.1}$$

- Le coût moyen qui représente la charge totale des sinistres divisée par le nombre de sinistres.

$$Cout = \frac{Charge\ totale\ des\ sinistres}{Nombre\ de\ sinistres} \tag{1.2}$$

Utiliser un modèle coût fréquence pour obtenir la prime pure revient à combiner ces deux variables, c'est-à-dire, à faire le produit de la fréquence et du coût moyen. **La prime pure annuelle** de l'assuré

considéré serait alors :

$$\begin{aligned} Prime\ Pure &= Fréquence \times Coût \\ &= \frac{Nombre\ de\ sinistres}{Exposition} \times \frac{Charge\ de\ sinistres}{Nombre\ de\ sinistres} \\ &= \frac{Charge\ de\ sinistres}{Exposition} \end{aligned} \tag{1.3}$$

Soit $S = \sum_{i=1}^{N} C_i$ avec N une variable discrète représentant le nombre de sinistres et C_i des variables continues positives représentant le coût des sinistres. Si l'on suppose que les C_i sont indépendants et identiquement distribués et que la fréquence est indépendante de la sévérité (i.e. N et C_i sont indépendants), alors on a:

$$E(S) = E(E(S|N) = E(\sum_{i=1}^{N} E(C_i|N) = E(N \times E(Ci)) = E(N) \times E(C) \tag{1.4}$$

De la même manière, si l'on élargit l'étude au cas d'un ensemble d'assurés dans un portefeuille, sous réserve d'indépendance entre la fréquence et le coût moyen, la prime pure vaudrait le produit des espérances de la fréquence et du coût moyen. C'est l'approche qui sera suivi dans ce mémoire.

1.3.2 Intégration du risque géographique

Afin de tenir compte du risque géographique, les compagnies d'assurances mettent en place des zoniers. Il s'agit d'une segmentation homogène du territoire en différentes zones plus ou moins larges qui quantifient le risque auquel est exposé l'assuré selon son lieu de résidence. Ce zonier correspond donc à une variable latente : elle ne peut pas être mesurée directement et est non observable dans le portefeuille. Elle est issue d'algorithmes ou de modélisations statistiques et synthétise plusieurs variables observées.

Nous travaillons ici sur les garanties vol et dégâts des eaux dont la fréquence de sinistre dépend du risque géographique. En effet, les statistiques nationales mettent en évidence une disparité des actes de cambriolages en France métropolitaine en faisant apparaître des zones fortement exposées au risque de cambriolage et à l'inverse d'autres qui le sont moins. Ainsi, ces publications donnent toute sa légitimité à l'application d'un critère géographique sur la garantie vol. D'autre part, pour la garantie dégâts des eaux, certains facteurs géographiques sont directement discriminants. C'est le cas de la température (rupture des canalisations dues au gel), la précipitation ou les spécificités locales pour la construction (matériaux utilisés, type d'habitation etc.).

1.4 Processus de création d'un zonier et enjeux

1.4.1 Étapes générales de construction d'un zonier

La première étape consiste à modéliser le risque purement déclaratif en s'appuyant sur des modèles traditionnels de tarification, en l'occurrence, le modèle GLM. Dans le cadre de ce mémoire, nous modélisons d'une part le coût de sinistres et de l'autre la fréquence de sinistre pour les garanties vol et dégâts des eaux (cela revient à réaliser quatre GLM différents, soit deux GLM par garantie).

L'étape suivante consiste à calculer des résidus qui constitueront par la suite la variable cible que nous chercherons à prédire. Ces résidus permettront de quantifier le risque géographique auquel est exposé l'assuré selon son lieu de résidence. En effet, intégrer directement une variable "Code INSEE" dans le modèle n'est pas envisageable étant donné le nombre conséquent de modalités de celle-ci. Utiliser une approche séquentielle est donc le meilleur moyen d'intégrer la dimension géographique du risque aux modèles. Ensuite, nous créerons les variables explicatives géographiques à partir de *l'open data*. Ces variables vont permettre d'implémenter un modèle prédictif des résidus construits à l'étape précédente. Les prédictions du modèle créé permettront de regrouper le territoire en zones de risques et de rattacher à chaque zone, une valeur du risque prédit. Enfin, la nouvelle variable représentant le risque géographique sera jointe à la base de départ pour obtenir de nouvelles prédictions, incluant les composantes géographiques du risque.

Le processus de création d'un zonier peut donc être décomposé en trois phases distinctes.

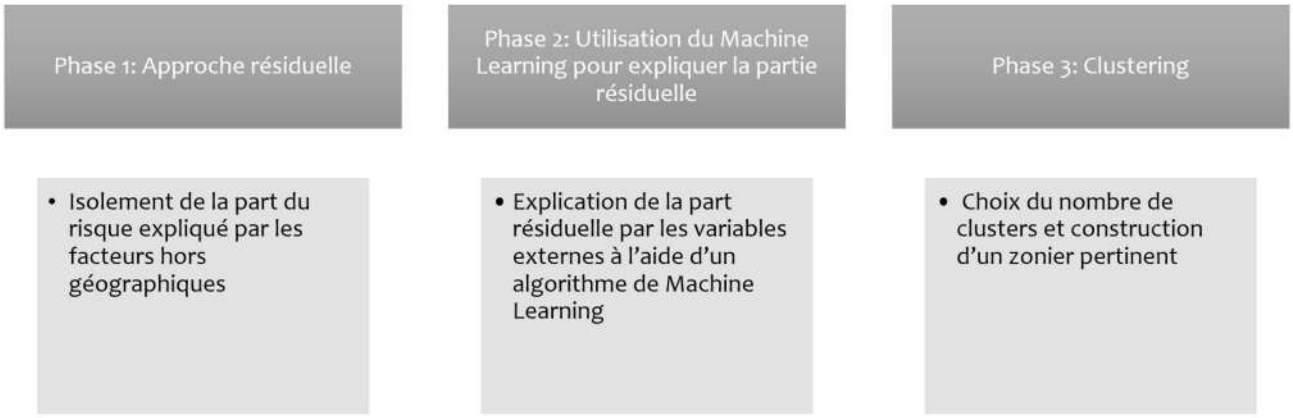


Figure 1.4: Étapes de construction d'un zonier

Après avoir modélisé le risque lié aux caractéristiques de l'habitation, nous cherchons à quantifier le risque résiduel lié à la zone géographique du logement assuré. Dans le cadre de ce mémoire, nous avons choisi de travailler sur une maille code INSEE*, toute l'étude va donc consister à affecter une classe de risque à chaque code INSEE.

1.4.2 Les problématiques opérationnelles et enjeux

Le manque d'exposition et de précision dans certaines régions

La base de données sur laquelle nous travaillons présente certaines hétérogénéités au niveau de l'exposition dans le territoire français (cf. Partie 2.2.2). Tout d'abord, il est important de préciser que dans le cadre de ce mémoire, nous faisons notre étude uniquement sur la France métropolitaine et la Corse. Certaines régions tels que l'Occitanie et l'île-de-France ne présentent aucun problème en

*Le code INSEE est un code numérique ou alphanumérique élaboré par l'Institut national de la statistique et des études économiques (INSEE). Il contient cinq chiffres qui correspondent à la concaténation du code département et de la codification sur trois chiffres de la commune ou de l'arrondissement municipal à Paris, Lyon et Marseille. Au 1er mars 2019, la France métropolitaine et les départements d'outre-mer sont découpés en 34 967 communes.

termes d'exposition car elles contiennent énormément de contrats. Cependant, d'autres régions comme le Grand Est sont beaucoup moins exposées et devront ainsi faire l'objet d'une attention particulière. En effet, un manque d'exposition dans une zone pourrait biaiser nos résultats étant donné qu'un unique contrat ne suffit pas à représenter une zone entière.

Afin de palier à ce manque d'exposition et de compenser la volatilité issue d'un faible nombre d'observation, il sera nécessaire d'avoir recours à une théorie de crédibilité afin de lisser spatialement les données. Le lissage spatial permettra ainsi de prendre en compte la dépendance spatiale entre les communes voisines qui ont plus de chance d'être exposées au même risque que des communes éloignées (cf. chapitre 4).

Choix du maillage

La construction d'un zonier peut se faire selon différentes mailles (cf. Figure 1.5). En effet, les assureurs peuvent constituer des zoniers à des mailles plus larges (par exemple : la maille département) comme à des mailles plus fine (par exemple : la maille IRIS ou adresse) selon les contraintes auxquelles ils doivent faire face. Ces contraintes peuvent être internes (disponibilité des données, abilité des systèmes d'information à traiter de la donnée etc.) ou liées à la localisation des risques (la criminalité contient certaines composantes très localisées comme le quartier résidentiel ou la distance à un commissariat, dans ce cas, un zonier plus fin serait plus adéquat, mais aussi d'autre composantes moins localisées).

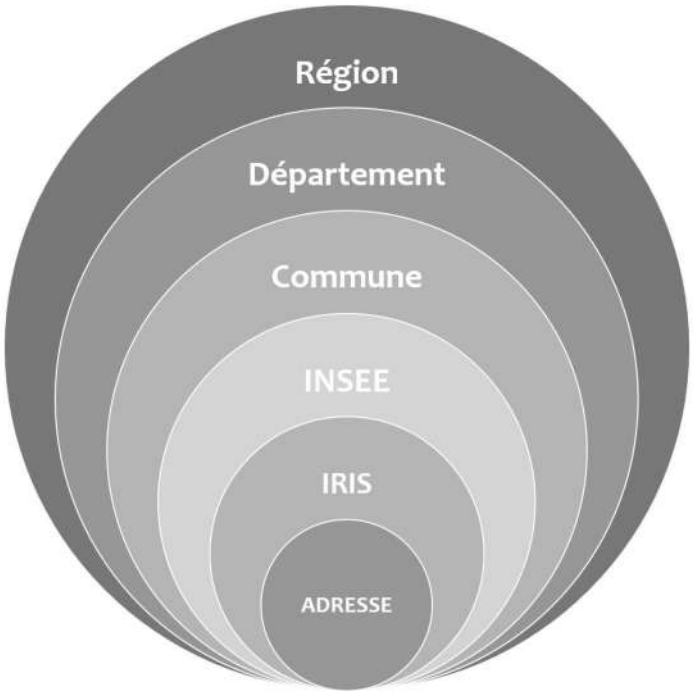


Figure 1.5: Maillage du territoire

Dans notre cas, nous étudions un portefeuille à la maille code INSEE par police d'assurance (il s'agit du code INSEE à la date de souscription). Nous avons donc fait le choix de travailler à la maille INSEE pour la suite de l'étude. Une maille plus fine n'aurait pas pu être considérée pour des raisons de disponibilité des données. Néanmoins, les risques considérés ne sont pas suffisamment localisés pour que la marge soit trop large.

De plus, les données externes analysées sont issues de différentes sources externes et concernent-elles aussi plusieurs mailles géographiques. Nous travaillons en particulier sur des données de criminalité de l'Observatoire National de la Délinquance et des Réponses Pénales (ONDRP) qui sont à la maille départements, des données sociodémographiques de l'Institut national de la statistique et des études économiques (INSEE) qui sont à la maille INSEE et des données météorologiques de la "National Climatique Data Center" qui sont à la maille la plus fine avec des informations sur la longitude et la latitude.

Afin de joindre nos bases de données internes comportant toutes les observations à la maille INSEE et nos données externes citées ci-dessus, il sera nécessaire d'ajouter une étape de retraitement qui nous permettra de travailler à la même maille et ainsi joindre nos bases.

Adaptabilité et robustesse du zonage

Pour construire et calibrer un modèle prédictif, il est important de s'assurer:

- du respect de la qualité de la donnée sur toute la chaîne de tarification (exactitude, robustesse, complétude).
- de la sensibilité au choix des modèles, hypothèses et jugements d'experts.
- de la traçabilité de l'information permettant d'assurer la simplicité à comprendre et à interpréter les résultats.
- de la stabilité dans le temps des modèles étant donnée que le nombre de communes en France est fluctuant.

Ces différents points évoqués sont alors au coeur de la problématique de converger vers un zonier qui permet de répondre à tous les changements temporels, tout en optimisant les différentes phases de production des calculs.

Notre base d'étude interne comprend un historique d'observations de 4 années (de 2015 à 2019). Nous choisissons d'établir un zonier en considérant l'ensemble des données de ces années afin de pouvoir généraliser le résultat final sur une période supérieure à un an. Cette approche permettra ainsi au zonier crée de ne pas être revu tous les ans.

Chapitre 2

Tarification en excluant les variables géographiques

2.1 Base d'étude interne

2.1.1 Périmètre de l'étude

Nous considérons le portefeuille d'un assureur qui commercialise un produit MRH depuis l'année 2015, soit un historique de données couvrant une période de 5 ans. Le jeu de donnée fournit des informations sur les assurés ainsi que les logements assurés et les codes INSEE qui leur sont associés. L'ensemble des logements assurés dans le portefeuille concernent soit des appartements soit des maisons de particuliers. Les immeubles entiers ainsi que les contrats professionnels ne sont pas considérés ici.

Le produit commercialisé par l'assureur se compose de plusieurs garanties. Nous nous focalisons dans la suite sur deux d'entre elles : les garanties dégâts des eaux et vols.

De plus, nous focalisons notre étude sur la France métropolitaine ainsi que la Corse et Monaco. L'analyse de la répartition des contrats réalisée dans le chapitre précédent nous a permis de constater une certaine hétérogénéité de la répartition des contrats dans tout le territoire Français. En effet, le portefeuille d'étude comprend un nombre conséquent de contrats dans les régions Ile-de-France et Provence-Alpes-Côte d'Azur mais une exposition moindre dans certaines régions telles que le Grand Est.

Enfin, les variables déclaratives considérées dans la base d'étude sont les suivantes :

- La nature du logement : maison, appartement.
- L'étage : au dernier étage, à un étage intermédiaire etc.
- La qualité : propriétaire, locataire.
- La superficie : nombre de pièces, dépendances, véranda, piscine, garage.
- Le montant du capital mobilier : estimation du montant des objets et meubles présents dans le logement.
- La franchise : le montant qui reste à la charge de l'assuré en cas de sinistre.
- ...

L'ensemble des variables tarifaires considérées dans le jeu de donnée peuvent être consultées en annexe A.2.

Fréquemment, les assureurs utilisent une structure tarifaire différente selon le type de bien (appartement/maison). La sinistralité et l'exposition au risque est différente selon le type de logement. En effet, généralement, selon les garanties, le type de l'habitation a souvent un impact sur la fréquence et le coût moyen des sinistres observés. De ce fait, nous avons cartographié la fréquence de sinistres en distinguant les appartements des maisons afin de mettre en évidence la différence de sinistralité entre le type d'habitation et la zone géographique. Les cartes ont été floutées et les légendes supprimées pour des raisons de confidentialité.

(a) Fréquence de sinistres - Appartements (b) Fréquence de sinistres - Maisons

Figure 2.1: Fréquence de sinistres selon le type d'habitation pour la garantie DDE en Ile-de-France

(a) Fréquence de sinistres - Appartements (b) Fréquence de sinistres - Maisons

Figure 2.2: Fréquence de sinistres selon le type d'habitation pour la garantie VOL en Ile-de-France

Dans la suite de l'étude, la distinction entre les appartements et les maisons sera indirectement intégrée via la variable *étage*. D'autre part, la variable *type* sera supprimée car elle est redondante.

2.1.2 Pré-retraitement de la base de données

Qualité des données

Plusieurs contrôles sur la qualité des données ont été effectués en amont de l'étude (contrôles de cohérence, d'exhaustivité et de pertinence). De plus, ces contrôles permettent à la fois d'éviter le phénomène *garbage in garbage out** mais aussi d'assurer une certaine robustesse du résultat.

Nous avons fait plusieurs vérifications sur les données traitées comme par exemple :

- Vérifier que les codes INSEE sont correctement renseignés (à 5 chiffres...)
- Vérifier que les charges de sinistres ont bien un sinistre associé.
- Vérifier que la somme de l'exposition d'un contrat est inférieure à 1 sur une période d'un an.

Regroupement des modalités sous-exposées

Des analyses des différentes expositions des modalités de chacune des variables nous ont permis de faire des retraitements sur notre jeu de donnée afin d'éviter que nos résultats soient biaisés. En effet, dans certains cas, l'exposition d'une ou plusieurs modalités d'une variable tarifaire peut être trop faible et biaiser les résultats des GLM. Ainsi, nous avons fait le choix pour les variables concernées (*nombre de pièces tarifées*, *catégorie socioprofessionnel* et *capital mobilier*), de regrouper certaines modalités entre elles afin d'obtenir une exposition assez élevée pour toutes les modalités des variables considérées.

En guise d'exemple, les logements ayant plus de 11 pièces (comprises dans la tarification) ont été regroupés en une seule modalité "11 ou plus" pour la garantie dégâts des eaux. La même approche a été utilisée pour regrouper les modalités des variables *capital mobilier* et *catégorie socioprofessionnel* (regroupement en 8 catégories différentes comme indiqué dans le tableau en annexe A.3) pour les deux garanties.

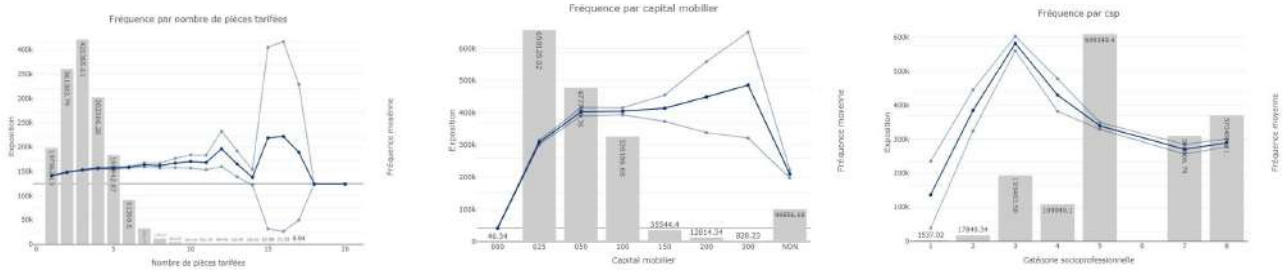


Figure 2.3: Statistiques descriptives pour le modèle de coût de la garantie DDE

*GIGO (garbage in, garbage out) est le concept selon lequel des données d'entrée défectueuses ou absurdes produisent des sorties absurdes.

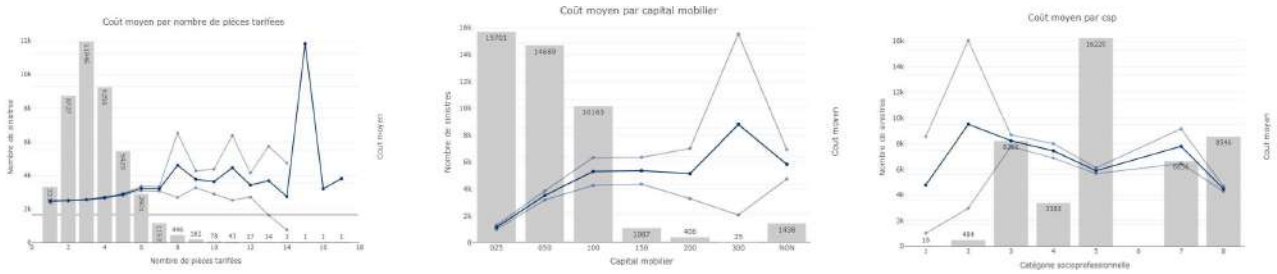


Figure 2.4: Statistiques descriptives pour le modèle de fréquence de la garantie DDE

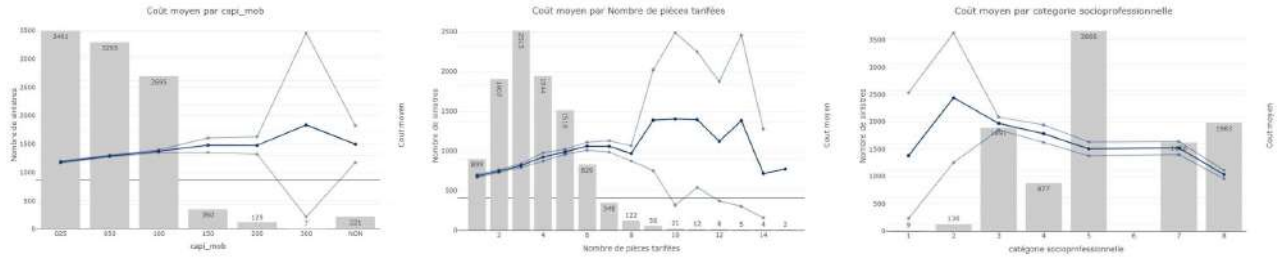


Figure 2.5: Statistiques descriptives pour le modèle de coût de la garantie VOL

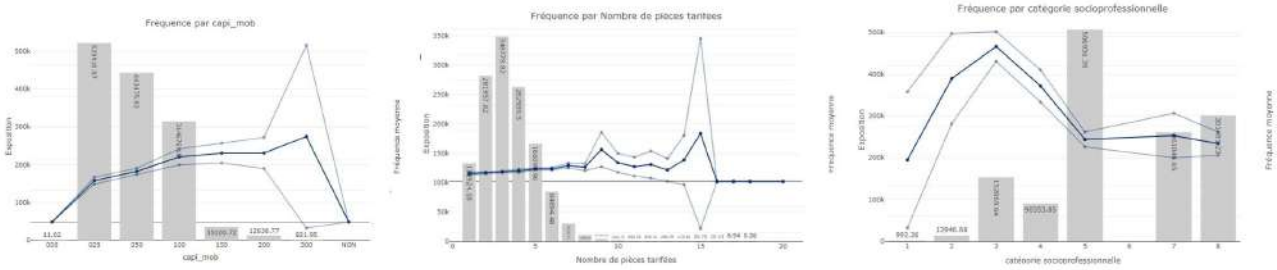


Figure 2.6: Statistiques descriptives pour le modèle de fréquence de la garantie VOL

2.2 Analyse descriptive

2.2.1 Statistiques descriptives

En souscrivant une assurance multirisque habitation, l'assuré protège ses biens mobiliers et immobiliers, son bâtiment principal, ses dépendances mais aussi les équipements contenus dans chaque bâtiment. Dans un premier temps, pour avoir une idée de la composition du portefeuille, nous avons calculé les proportions des modalités des variables tarifaires principales :

Variable Tarifaire	Pourcentage par modalité			
Type d'habitation	Appartement		Maison	
	66%		34%	
Qualité	Locataire		Propriétaire	
	61%		39%	
Usage	PNOM	PNONM	RP	RS
	1,8%	5,5%	90,5%	2,2%

Les significations des modalités de la variable *usage* sont les suivantes : PNOM : propriétaire non occupant meublé PNONM : propriétaire non occupant non meublé RP : résidence principale RS : résidence secondaire.

Dans un deuxième temps, nous avons illustré les effets des caractéristiques du logement assuré sur le coût et la fréquence. Par exemple, nous avons considéré ici les différentes modalités de la qualité de l'assuré ainsi que le type d'habitation. Les graphiques obtenus pour les autres variables tarifaires (*nombre de pièces tarifées, nombre d'enfants, piscine, véranda etc.* sont à consulter en annexe A.4 .

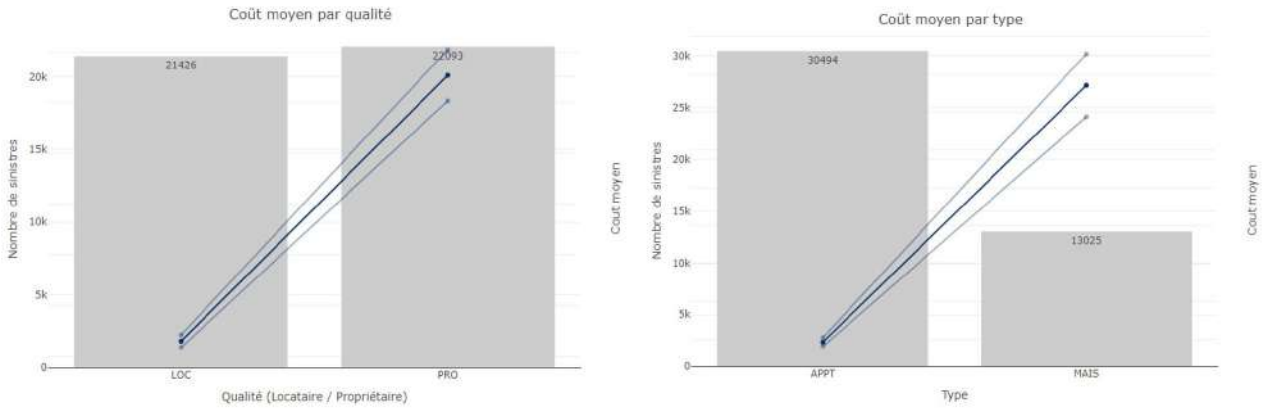


Figure 2.7: Coût moyen et exposition par modalité des variables tarifaires pour la garantie DDE

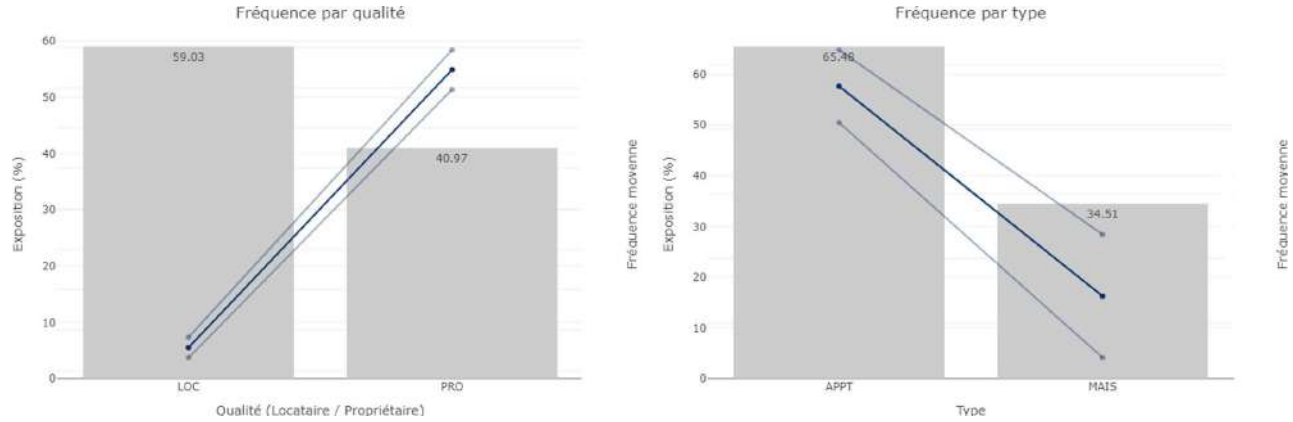


Figure 2.8: Fréquence de sinistre et exposition par modalité des variables tarifaires pour la garantie DDE

La même étude a été réalisée pour la garantie vol :

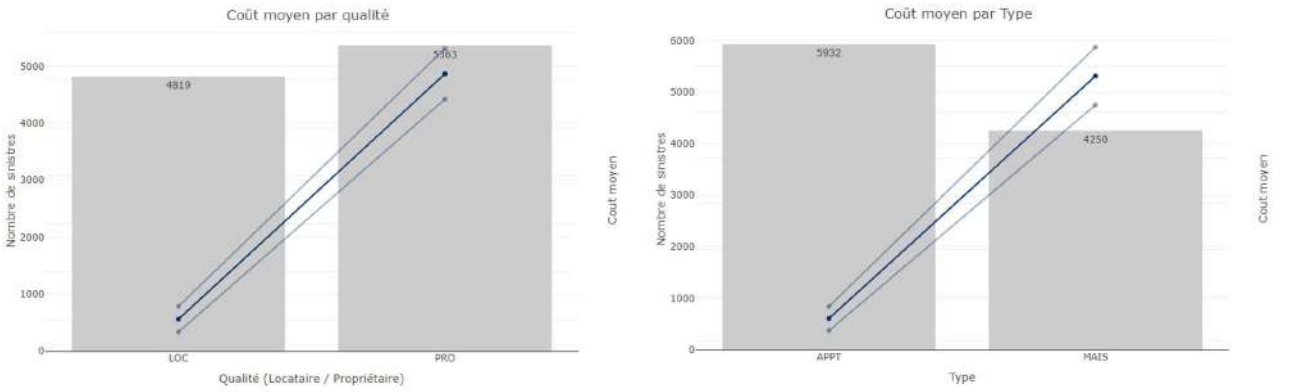


Figure 2.9: Coût moyen par modalité des variables tarifaires pour la garantie VOL

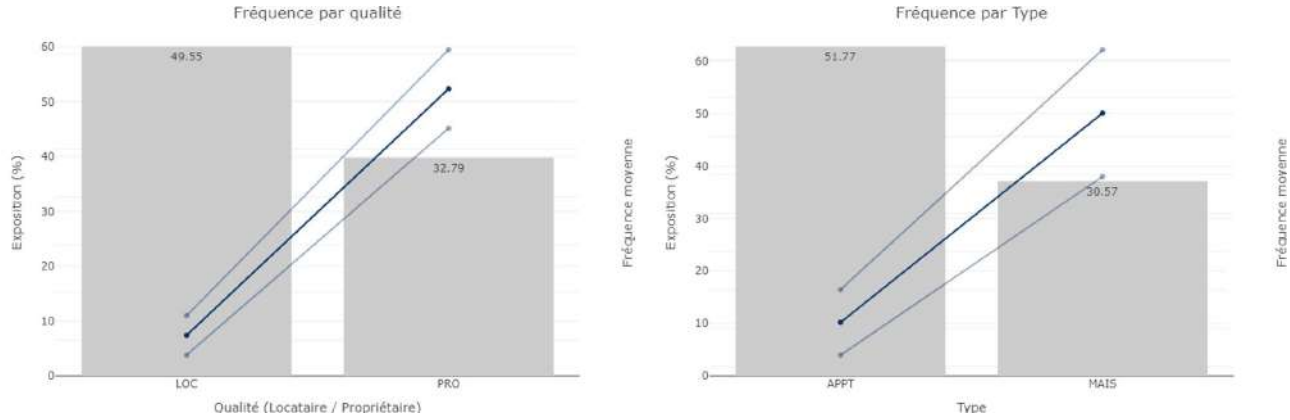


Figure 2.10: Fréquence de sinistre et exposition par modalité des variables tarifaires pour la garantie VOL

Pour les deux garanties, nous observons une hausse globale de la fréquence de sinistres selon la qualité de l'assuré. En effet, la modalité propriétaire apparaît comme plus risquée que la modalité locataire. Pour la garantie vol, nous pouvons faire l'hypothèse que cela est probablement lié au fait qu'un propriétaire investit plus dans le mobilier ou dans ses biens de manière générale qu'un locataire, susceptible de déménager à tout moment. C'est pourquoi nous retrouvons également un coût moyen des sinistres plus élevé pour les propriétaire (comparé aux locataires).

A partir de l'analyse de la sinistralité selon le type d'habitation, il est possible d'affirmer que dans le cas de la garantie dégâts des eaux, ce sont les appartements qui ont une fréquence de sinistres plus élevée malgré un coût moyen largement supérieur pour les maisons. En revanche, en ce qui concerne la garantie vol, ce sont les maisons qui sont bien plus concernées par le risque vol que les appartements que ce soit en termes de coût ou de fréquence. Cela est en adéquation avec le fait qu'en France, les systèmes d'alarme équipent cinq fois plus de maisons que d'appartements.

2.2.2 Cartographie - Répartition des contrats sur le territoire français

L'objectif de l'étude étant d'élaborer un zonier, il est nécessaire d'avoir une idée de la répartition des contrats présents dans le portefeuille d'étude sur le territoire français. Nous présentons ci-dessous la distribution du portefeuille sur la France métropolitaine qui met en exergue une forte densité sur les aires les plus peuplées.

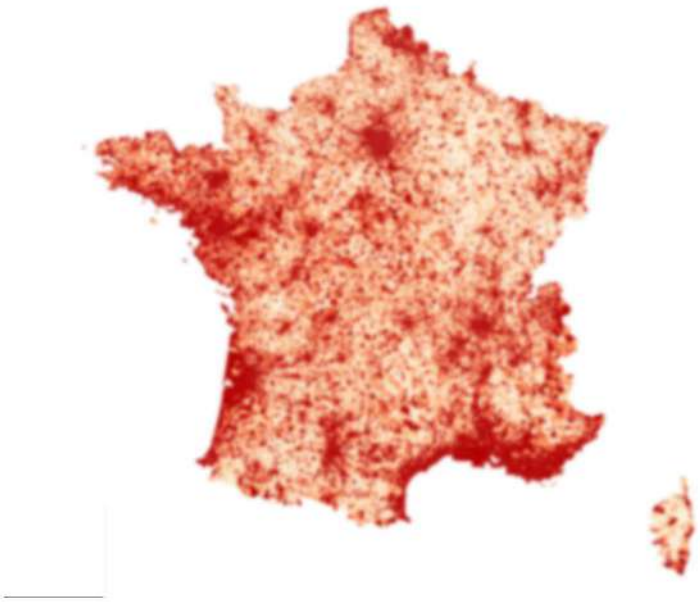


Figure 2.11: Répartition des contrats en France métropolitaine

De plus, en ciblant la région Ile-de-France, nous remarquons que la répartition des contrats est également hétérogène au sein d'une même région. Ainsi, certaines communes comportent beaucoup plus de contrats que d'autres. Cela est un point à ne pas négliger et devra être traité dans la suite de ce mémoire (cf section Théorie de crédibilité partie 4.1.3) ayant pour objectif de créer un zonier à la maille code INSEE.

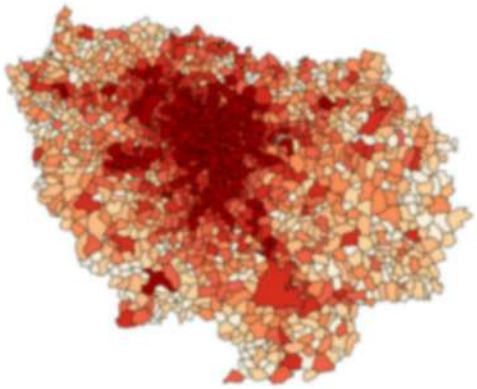


Figure 2.12: Répartition des contrats en île de France

2.3 Modélisation sans prise en compte des facteurs géographiques

2.3.1 Principe du modèle linéaire généralisé GLM

La construction d'un tarif en assurance IARD s'appuie généralement sur un modèle coût fréquence dans lequel l'effet des variables explicatives sur le niveau du risque est modélisé par des modèles de régression. Dans les statistiques, ces modèles de régression ont pour but de chercher à établir une relation entre une variable d'intérêt Y et plusieurs variables explicatives $X = {}^t(X_1, \ldots, X_p)$ $(p \in \mathbb{N}^*)$. L'idée générale étant d'appliquer au tarif une discrimination en fonction des caractéristiques de l'assuré

(type d'habitation, nombre de pièces, nombre d'enfants, qualité etc.), la modélisation doit prendre en compte les effets des variables explicatives dans l'estimation de la prime pure.

Il existe une multitude de méthodes de régression: les modèles linéaires, les modèles linéaires généralisés, les lissages (noyaux, splines, régression locale), arbres de régression, réseaux de neurones etc. Néanmoins, les modèles linéaires généralisés (GLM) sont ceux qui sont le plus souvent utilisés par les assureurs, et ceux que nous appliquerons pour nos modélisations.

2.3.2 Caractérisation d'un GLM

Le modèle linéaire généralisé est caractérisé par les trois quantités suivantes que nous expliciterons par la suite:
1. La **variable réponse** Y, composante aléatoire à laquelle est associée une loi de probabilité
2. les **variables explicatives** X_1, ..., X_p représentant les prédicteurs
3. la **fonction de lien** g qui décrit la relation entre la combinaison linéaire des variables explicatives et l'espérance de la variable réponse Y.

Ainsi le modèle s'écrit:

$$g(\mathbb{E}[Y|X]) = \beta_0 + \sum_{i=1}^{p} \beta_i X_i \tag{2.1}$$

Loi de la variable réponse Y

On considère que la variable à expliquer Y admet une distribution issue d'une structure de famille exponentielle. On dit que la loi de Y appartient à une famille exponentielle si elle est dominée par une mesure de référence et si la vraisemblance de Y calculée en y par rapport à cette mesure s'écrit de la façon suivante :

$$f_Y(y;\omega,\phi) = \exp\left(\frac{y\omega - b(\omega)}{a(\phi)} + c_\phi(y)\right)$$

Cette formulation inclut les lois usuelles suivantes : gaussienne, gaussienne inverse, gamma, poisson, binomiale etc. Le paramètre ω est appelé le paramètre naturel de la famille exponentielle. De plus, pour certaines lois, la fonction $a(\phi) = \phi$. Dans ce cas, ϕ est appelé paramètre de dispersion, il s'agit d'un paramètre de nuisance qui intervient par exemple lorsque les variances des lois gaussiennes sont inconnues, mais vaut 1 pour les lois à un paramètre comme la loi de poisson, de bernoulli etc.

Soit Y une variable aléatoire dont la loi de probabilité appartient à la famille exponentielle, alors

$$\begin{cases} \mathbb{E}(Y) = b'(\omega) \\ \mathbb{V}(Y) = b''(\omega) \times a(\phi) \end{cases}$$

Voici quelques exemples de lois de probabilité appartenant à la famille exponentielle:

Lois appartenant à la famille exponentielle					
Distribution	ω	$b(\omega)$	$a(\phi)$	$\mathbb{E}(Y) = b'(\omega)$	$\mathbb{V}(Y) = b''(\omega) \times a(\phi)$
Gaussienne	μ	$\frac{\omega^2}{2}$	$\phi = \sigma^2$	$\mu = \omega$	σ^2
Poisson	$ln(\lambda)$	$\lambda = \exp(\omega)$	1	$\lambda = \exp(\omega)$	$\lambda = \exp(\omega)$
Gamma	$\frac{-1}{\mu}$	$-\ln(-\omega)$	$\frac{1}{\nu}$	$\mu = \frac{-1}{\omega}$	$\frac{\mu^2}{\nu}$

Figure 2.13: Exemple de lois usuelles appartenant à la famille exponentielle

Pour notre étude, dans un premier temps, nous allons modéliser le coût moyen des sinistres (modèle de coût), les lois généralement utilisées dans ce cas sont les lois gamma ou lognormale. Dans notre cas, nous opterons pour la loi lognormale suite à certaines analyses que nous expliciterons par la suite. Dans un deuxième temps, nous serons amenés à modéliser le nombre de sinistres (modèle de fréquence), l'utilisation d'une distribution de Poisson est alors préconisée.

Prédicteur linéaire

Les observations des variables explicatives sont organisées dans la matrice X appelée *Design Matrix*. Soit β, un vecteur de k (=p+1) paramètres. Le prédicteur linéaire, composante déterministe du modèle est le vecteur défini par

$$\eta = X\beta$$

Le β_0 que l'on observe dans la formule de la prédiction est appelé intercept, il représente la classe de référence. Le profil de référence sera celui qui regroupe toutes les variables explicatives de référence (ie.les modalités les plus exposées dans la base de donnée).

On interprète les β_i de la manière suivante : Si $\beta_i > 0$, alors on peut dire qu'un individu présentant la modalité i a tendance à avoir une sinistralité plus importante que l'individu de référence. Au contraire, si $\beta_i < 0$ cela nous permet de détecter des individus avec un profil moins risqué que celui de la classe de référence.

Fonction de lien

Cette quantité exprime une relation fonctionnelle entre la variable à expliquer et les variables explicatives.

Soit $\mu_i = \mathbb{E}(Y_i);\ i = 1, \ldots, n$. On pose$\forall i = 1, \ldots, n,\ \eta_i = g(\mu_i)$ où g, la fonction de lien, est supposée monotone et différentiable. Ceci revient donc à écrire : $\forall i = 1, \ldots, n,\ g(\mu_i) = x_i\beta$. La fonction de lien qui associe la moyenne μ_i au paramètre naturel ω_i est appelée fonction de lien canonique. Dans ce cas, $\forall i = 1, \ldots, n,\ g(\mu_i) = \omega_i = x_i\beta$

De plus, il est important de souligner que l'interprétation des résultats d'un modèle linéaire généralisé dépend de la fonction de lien choisie. Par exemple, si la fonction de lien choisie est la fonction identité, alors le modèle sera additif :

$$\mathbb{E}[Y|X] = \beta_0 + \sum_{i=1}^{p} \beta_i X_i$$

En revanche, si la fonction de lien utilisée est le logarithme, le modèle sera multiplicatif :

$$\ln\left(\mathbb{E}\left[Y|X\right]\right) = \beta_0 + \sum_{i=1}^{p} \beta_i X_i \iff \mathbb{E}\left[Y|X\right] = \exp\left(\beta_0 + \sum_{i=1}^{p} \beta_i X_i\right)$$

En particulier, pour un client i, nous obtiendrons un estimateur de la forme :

$$\mu_i = \exp\left(x_i\beta\right) = \exp\left(\sum_{j=1}^{p} x_{ij}\beta_j\right) = \prod_{j=1}^{p} \exp\left(x_{ij}\beta_j\right) \tag{2.2}$$

Cette structure du modèle permet de dissocier les effets de chacune des variables, et donc de construire un modèle directement interprétable.

2.3.3 Estimation des coefficients

Une fois le modèle posé, c'est-à-dire, après avoir défini les quantités présentées ci-dessus (variable à expliquer, variables explicatives et fonction de lien), il convient d'estimer les coefficients associés à chacune des variables explicatives du modèle. Il s'agit ici d'estimer le vecteur des paramètres de notre modèle :

$$\beta = (\beta_1, \beta_2, \ldots, \beta_p)^t$$

Pour ce faire, différentes méthodes telles que la méthode des moindres carrés ou la méthode du maximum de vraisemblance peuvent être utilisées. Néanmoins, la méthode des moindres carrés n'est pas applicable dans un grand nombre de situations pour le modèle linéaire généralisé (sauf pour des fonctions de lien canonique, i.e. identité). La méthode préconisée est donc celle d'estimation du maximum de vraisemblance. On cherche ici à maximiser la log-vraisemblance du modèle linéaire généralisé. Par indépendance des observations, la vraisemblance d'un échantillon $Y = (Y_1, ..., Y_n)$ s'écrit :

$$\beta \longrightarrow \mathcal{L}(y;\beta) = \prod_{i=1}^{n} f_{Y_i}(y_i;\omega_i)$$

d'où la log vraisemblance :

$$\beta \longrightarrow ln(\mathcal{L}(y;\beta)) = l(y;\beta) = \sum_{i=1}^{n} \ln(f_{Y_i}(y_i;\omega_i))$$

L'estimateur par maximum de vraisemblance associé vérifie donc :

$$\beta \longrightarrow \hat{\beta}_{MV} \in argmax\ \mathcal{L}(Y;\beta) = argmax\ l(Y;\beta)$$

Grâce aux propriétés de la fonction de lien et au fait que f appartienne à une famille exponentielle, il est possible de dériver selon le paramètre β. L'équation à résoudre pour obtenir les coefficients estimés devient donc :

$$\partial_\beta\left(l(Y_1, \ldots, Y_n;\beta)\right) = 0$$

Nous obtenons ainsi l'estimation des paramètres du modèle:

$$\hat{\beta}_{MV} = (\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p)^t$$

Il n'existe pas de formule fermée pour l'expression des coefficients. Il est cependant possible de montrer que le problème associé à la détermination de $\hat{\beta}_{MV}$ est un problème d'optimisation convexe qui peut être traité par un algorithme de type Newton-Raphson, que nous ne détaillerons pas ici*.

2.3.4 Le principe de l'offset et la notion d'exposition

Dans le cadre de l'élaboration du modèle de fréquence, il est important d'être attentif à une variable particulière qui est l'exposition. En effet, dans une optique de tarification d'un contrat, nous cherchons à prédire le nombre de sinistres moyen qui pourraient survenir l'année suivante. Or, les durées d'expositions des polices sont souvent différentes. Ces durées d'expositions représentent les durées effectives des contrats, par exemple, un assuré ayant eu deux sinistres sur une période d'un an n'aura pas le même impact sur le modèle qu'un assuré ayant eu deux sinistres sur une période de 6 mois. Ainsi, il est nécessaire de pondérer la fréquence de sinistre par l'exposition afin de pouvoir analyser la fréquence annuelle de sinistres des contrats. En général, une police ayant une durée d'exposition d'un an aura en moyenne deux fois plus de sinistres qu'une police ayant une durée d'exposition de seulement 6 mois. Pour intégrer ce phénomène d'échelle au modèle linéaire généralisé l'écriture du modèle est remplacée par :

$$g\left(\mathbb{E}\left[\frac{Y}{\theta}|X\right]\right) = \beta_0 + \sum_{i=1}^{p} \beta_i X_i$$

où θ correspond à l'*offset*.

Donc si le nombre de sinistres est modélisé à l'aide d'une fonction de lien logarithmique, le modèle s'écrit :

$$\ln(\mathbb{E}\left[Y|X\right]) = \beta_0 + \sum_{i=1}^{p} \beta_i X_i + \ln(\theta)$$

L'*offset* a donc pour objectif de *normaliser* le nombre de sinistres. Il correspond à la variable *Exposition* et ne représente pas une variable explicative du modèle, aucun coefficient de régression n'est associé à celle-ci.

2.3.5 Qualité d'ajustement et choix du modèle

Dans cette section, nous cherchons le modèle qui explique le mieux la variable réponse Y et celui qui utilise le moins de variables explicatives possibles pour avoir un modèle de prédiction robuste et éviter le sur-ajustement† (overfitting). Il s'agit ici d'évaluer la qualité d'ajustement du modèle sur la base des différences entre observations et estimations. Plusieurs critères sont proposés, entre autres: la déviance, le R^2, le khi-deux de Pearson, les critères AIC et BIC.

*L'algorithme de Newton-Raphson approxime le log de la fonction de vraisemblance dans un voisinage du paramètre initial par une fonction polynomiale qui a la forme d'une parabole concave.

†Le sur-ajustement se produit lorsque l'algorithme sur-apprend (overfit), autrement dit, le modèle se rapproche tellement de la fonction qu'il prête trop attention au bruit. Cela peut causer des fluctuations aléatoires dans la fonction. Ainsi, l'overfitting est caractérisé par une erreur de type variance très élevée. Par conséquent, lorsque le modèle reçoit de nouvelles données, il ne pourra pas généraliser les prédictions.

Déviance

Afin d'évaluer la qualité d'ajustement d'un modèle linéaire généralisé, il est courant de calculer la déviance, aussi appelée la statistique des écarts. Elle se définit comme suit :

$$\mathcal{D}(\mathcal{M}) = -2 \times (\mathcal{L} - \mathcal{L}_{sat})$$

où $\mathcal{L}$ fait référence à la valeur maximisée du maximum de vraisemblance du modèle étudié et $\mathcal{L}_{sat}$ représente la log-vraisemblance du modèle saturé, c'est-à-dire le modèle le plus complexe (ie. le modèle possédant autant de paramètres que d'observations et estimant exactement les données). La déviance permet ainsi de comparer la log-vraisemblance du modèle estimé à celle d'un modèle parfait en terme d'adéquation aux données : le modèle saturé. Par conséquent, plus la déviance est faible, meilleur est le modèle en terme d'ajustement. Lorsque le modèle étudié est exact, la déviance suit asymptotiquement une loi du khi-deux à n-k degrés de liberté ce qui permet de construire un test de rejet ou d'acceptation du modèle selon que la déviance est jugée significativement ou non importante.

Le khi-deux de Pearson

Il s'agit de la statistique définie par

$$\chi^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{var(\hat{y}_i)}$$

avec $\hat{y}_i = g^{-1}(x_i \hat{\beta}_{MV})$

Elle est utilisée pour comparer les valeurs observées y_i à leur prévision par le modèle. Comme la déviance, cette statistique est distribuée asymptotiquement selon une loi du khi-deux à n-k degrés de liberté si le modèle étudié est exact. Cela nous permet de construire un test de rejet ou d'acceptation du modèle: on rejette le modèle étudié au niveau α si la valeur observée de χ^2 est supérieure au $1-\alpha$ quantile de la loi $\chi^2(n-k)$.

L'usage est souvent de comparer ces statistiques (déviance et statistique khi-deux) avec le nombre de degrés de liberté. Le modèle est jugé satisfaisant si les rapports $\frac{\mathcal{D}}{ddl}$ ou/et $\frac{\chi^2}{ddl}$ sont inférieurs à 1. Cependant, ces statistiques ne prennent pas en compte le nombre de paramètres intégrés dans le modèle c'est pourquoi d'autres indicateurs tels que l'AIC et le BIC sont également utilisés.

Critère AIC et BIC

Les critères AIC et BIC fonctionnent de la manière suivante : plus la log-vraisemblance est grande, meilleur est le modèle en terme d'ajustement. Cependant, étant donné que la vraisemblance augmente avec la complexité du modèle, choisir le meilleur modèle revient à choisir le modèle saturé. Ainsi, pour éviter *l'overfitting*, une stratégie consiste à pénaliser la vraisemblance par une fonction du nombre de paramètres.

Le critère d'information d'Akaike (ou AIC) s'écrit comme suit:

$$AIC = 2p - 2\ln(\mathcal{L})$$

où p est le nombre de paramètres à estimer du modèle et$\mathcal{L}$ est le maximum de la fonction de vraisemblance du modèle. Avec ce critère, la déviance du modèle ($-2log(\mathcal{L})$ est pénalisée par deux fois le nombre de paramètres.

Le critère de choix de modèle BIC (Critère d'information bayésien) pour un modèle de dimension p est défini par:

$$BIC = p\ln(N) - 2\ln(\mathcal{L})$$

avec $\mathcal{L}$ la vraisemblance du modèle estimée, N le nombre d'observations dans l'échantillon et p le nombre de paramètres libres du modèle.

Le meilleur modèle est celui qui possède le plus petit AIC ou BIC.

2.3.6 Sélection de variables

Nous cherchons à trouver le meilleur modèle parmi plusieurs à partir de l'échantillon dont nous disposons. L'objectif étant de choisir le plus petit nombre de variables expliquant le mieux la donnée souhaitée, nous pouvons simplifier notre modèle en choisissant de ne garder qu'un sous-ensemble des variables explicatives (les plus significatives*). Dans le cadre des modèles linéaires généralisés, la loi de l'estimateur du maximum de vraisemblance n'est connue qu'asymptotiquement. Aussi les procédures de test vont être menées dans un cadre asymptotique. Nous allons dans la suite considérer un problème de test qui permettent de déterminer si les différentes variables explicatives du modèles sont pertinentes ou pas.

Test de Wald

La théorie du maximum de vraisemblance nous donnant le comportement asymptotique des estimateurs, il est possible de tester la significativité des variables explicatives. Pour cela, le test de Wald est généralement utilisé.

Celui-ci permet de comparer deux modèles emboîtés et de déterminer si un sous-ensemble de variables explicatives est suffisant pour expliquer la réponse Y.

Soient $\mathcal{M}_0$ et $\mathcal{M}_1$ deux modèles tels que $\mathcal{M}_0 \subset \mathcal{M}_1$. On suppose que $\mathcal{M}_0$ et $\mathcal{M}_1$ ont les mêmes distributions et que $\mathcal{M}_1$ possède plus de variables explicatives que $\mathcal{M}_0$.

On veut tester :

- H_0 : les paramètres ajoutés au modèle $\mathcal{M}_1$ sont nuls
- H_1 : il existe un des paramètres supplémentaires du modèle $\mathcal{M}_1$ non nul.

L'hypothèse H_0 est rejetée dès que la statistique dépasse le fractile d'ordre 1-α de la loi asymptotique.

Le test que nous venons d'étudier permet de sélectionner un modèle parmi deux modèles emboîtés. Or, il est possible de définir un grand nombre de modèles qui ne sont pas forcément emboîtés à partir de p variables explicatives. On a donc recours à des méthodes basées sur la comparaison de critères qui permettent de comparer des modèles qui ne sont pas forcément emboîtés les uns dans les autres. Ce sont les méthodes pas à pas que nous présentons ci-dessous.

Méthodes usuelles de sélection de variables: les méthodes pas à pas

Il existe plusieurs méthodes différentes basées sur la comparaison des critères présentés ci-dessus pour sélectionner les variables explicatives à ajouter au modèle. En pratique, nous commençons par choisir un critère de sélection de modèles, par exemple, les critères AIC (Akaike Information Criterion) et BIC (Bayesian Information Criterion). Puis, nous choisissons l'une des stratégies suivantes:

a. La méthode descendante (backward)

Nous partons du modèle complet qui comporte toutes les variables explicatives et nous cherchons,

*Une variable est significative lorsque l'ajout (ou le retrait) de celle-ci permet d'améliorer la qualité du modèle au sens d'un critère défini.

à chaque étape de l'algorithme, la variable la plus pertinente à retirer selon le critère choisi. Parmi les ensembles de variables testés pendant l'algorithme, nous retenons le meilleur au vu du critère (par exemple, si le critère choisi est l'AIC, alors l'algorithme s'arrêtera lorsque l'AIC est minimisé).

b. La méthode ascendante (forward)

Nous partons de l'ensemble vide de variables (intercept uniquement) et nous cherchons, à chaque étape de l'algorithme, la variable la plus pertinente selon le critère choisi. Nous itérons ainsi l'algorithme jusqu'à intégrer toutes les variables.

c. La méthode progressive (stepwise)

Enfin, cette dernière méthode peut être vu comme une combinaison des deux dernières : à chaque étape, l'algorithme retire ou rajoute les variables au modèle en cherchant à minimiser le critère considéré.

Sélection de variables grâce aux méthodes de *Machine Learning*

Il existe, en plus des méthodes usuelles, de nombreuses méthodes de *Machine Learning* qui permettent la sélection des variables les plus influentes. Cependant, celles-ci ne sont généralement pas utilisées en pratique car dans les secteurs réglementés (banque, assurance etc.), l'interprétabilité du modèle est une exigence pour passer en production. Or, les méthodes de *Machine Learning*, malgré leur efficacité et leur résultats souvent très satisfaisant, sont souvent qualifiées de boites noires et ne permettent pas une bonne traçabilité des décisions prises par l'algorithme. Néanmoins, elles permettent d'obtenir des informations supplémentaires comme l'interaction entre les variables, qui ne sont pas obtenues en utilisant les méthodes usuelles citées ci-dessus.

La méthode de *Machine Learning* que nous avons choisi d'utiliser pour améliorer la connaissance de nos données est l'un des incontournable dans le monde de la data science aujourd'hui: le XGBoost (Extreme Gradient Boosting). Il s'agit d'une implémentation *open source* optimisée de l'algorithme d'arbres de boosting de gradient. Il utilise des arbres de décision pour résoudre des problèmes de classification, de classement et de régression. Il s'agit donc d'un algorithme d'apprentissage supervisé dont le principe et de combiner les résultats d'un ensemble de modèles plus simples afin de fournir une meilleur prédiction.

Corrélation entre les variables (V de cramer)

Pour aider à la décision dans le choix des variables à intégrer dans le modèle, l'étude des corrélations est obligatoire. En effet, les modèles linéaires généralisés ne prennent pas en compte les intéractions entre les variables. Il faut veiller à ce que les variables explicatives ne soient pas trop corrélées entre elles, cela pourrait fausser les résultats du modèle, ce dernier étant confronté à une redondance d'information.

Les données à disposition ne sont pas toutes numériques, il n'est donc pas possible d'utiliser les coefficients de corrélation usuels tel que le coefficient de corrélation de Pearson pour mettre en évidence une relation entre les variables explicatives. Ainsi, nous utilisons le V de Cramer qui représente une bonne alternative pour avoir de l'information sur l'intensité des relations entre les variables qualitatives.

Le V de Cramer se calcule à partir du χ^2. Comme ce dernier, il dépend de la taille de l'échantillon et du degré de liberté et se calcule de la manière suivante:

$$V\ de\ Cramer = \sqrt{\frac{\chi^2}{N \times DDL}} \tag{2.3}$$

avec N la taille de l'échantillon et DDL correspond au degré de liberté du tableau.

Le V de Cramer correspond à une valeur entre 0 et 1. Un V de Cramer qui s'approche de 1 indique une forte corrélation entre les variables. En revanche, si le V de cramer est proche de 0, les variables

sont indépendantes.

2.3.7 Validation de modèle

La validation du modèle consiste à vérifier, une fois le modèle mis en œuvre, la performance et la cohérence de celui-ci. De nombreux tests sont proposés afin d'évaluer si le modèle constitue un bon prédicteur ou non. Ils concernent l'étude graphique des résidus et le calcul d'indicateurs objectifs pour évaluer si le modèle a généré des prédictions pertinentes pour la variable étudiée.

Analyse des résidus

L'étude des résidus nous permet de vérifier l'adéquation du modèle avec les données. Les résidus doivent être sans tendance et répartis de manière homogène autour de 0. Leur observation permet à la fois de vérifier les hypothèses d'indépendance et d'homoscédasticité sur le terme d'erreur mais aussi de détecter des cas particuliers, comme les outliers (valeurs aberrantes) qui peuvent avoir un impact significatif sur la qualité du modèle.

Les résidus peuvent être calculés de différentes manières, les trois principales sont le rapport de l'observé sur le prédit r_i, les résidus de pearson r_i^p et les résidus de déviance r_i^d.

$$r_i = \frac{Y_i}{\hat{Y}_i} \tag{2.4}$$

$$r_i^p = \frac{Y_i - \hat{Y}_i}{\sqrt{Var_{\hat{Y}_i(Y_i)}}} \tag{2.5}$$

$$r_i^d = \sqrt{d_i} signe(Y_i - \hat{Y}_i) \tag{2.6}$$

où d_i représente la contribution de l'observation i à la déviance $\mathcal{D}$ et fournit un diagnostic individuel concernant la linéarité du modèle. Ainsi, nous pouvons écrire:

$$\mathcal{D} = \sum_{i=1}^{n} r_i^d$$

De la même manière, r_i^p représente la contribution de l'observation i à la statistique de pearson χ^2

$$\chi^2 = \sum_{i=1}^{n} r_i^p$$

Les indicateurs de performance

a. L'erreur quadratique moyenne (RMSE)

Cet indice fournit une indication par rapport à la dispersion ou la variabilité de la qualité de la prédiction. Le RMSE représente la mesure d'écart de prédiction par rapport à la valeur réelle:

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

avec y_i les valeurs observées, $\hat{y}_i$ les valeurs prédites et n le nombre d'observations. Si l'on compare deux estimateurs sans biais, le meilleur est celui qui présente le RMSE le plus faible.

b. L'indice de Gini

Le coefficient de Gini est un indicateur qui permet de mesurer la capacité de segmentation d'un modèle. Son calcul se base sur un rapport entre deux aires et sa valeur est comprise entre 0 (égalité parfaite) et 1 (inégalité totale). Le coefficient de Gini permet de comparer des modèles et de tester l'apport de nouvelles variables sur la capacité à segmenter les risques. Il est calculé à partir de la courbe de Lorenz. Si on note $\mathcal{L} : x \longrightarrow \mathcal{L}(x)$ la courbe de Lorenz, alors l'indice de Gini qui lui est associé se définit de la manière suivante :

$$G = 1 - 2\int_0^1 \mathcal{L}(x)dx$$

Toutefois, sa limite porte sur la distance entre les prédictions et les valeurs observées. Ainsi, deux modèles peuvent obtenir des valeurs de Gini similaires bien que l'un s'approche mieux des observations que l'autre.

2.3.8 Les limites du GLM et l'apport du *Machine Learning*

Le modèle linéaire généralisé (GLM) est l'outil principal utilisé en tarification par les assureurs. Cependant, certaines étapes de son implémentation peuvent s'avérer incommodes (tests de corrélations etc.). D'un autre coté, les modèles de *Machine Learning* ne sont pratiquement pas utilisés par les assureurs du fait d'une théorie initiale parfois complexe. Néanmoins, ils peuvent être un moyen efficace de challenger les résultats des modèles linéaires généralisés ou de leur apporter quelques améliorations. Dans les secteurs réglementés (banque, assurance, médicale), l'interprétabilité des modèles utilisés est une exigence pour passer en production. Ainsi, l'ensemble des nouvelles techniques de *Machine Learning* et des modèles utilisés doivent être totalement explicables et interprétables. Le trade-off avec ces nouveaux modèles est qu'en gagnant en performance, ils perdent en interprétabilité. Or, dans le secteur de l'assurance, le besoin d'interprétabilité est nécessaire pour se conformer aux exigences réglementaires (ACPR).

En prenant en compte toutes ces informations, nous avons décidé de combiner en pratique à la fois la précision et la rigueur des méthodes usuelles de sélection de variables (backward, forward et stepwise) aux informations supplémentaires apportées par une méthodes de *Machine Learning* qui est le XGBoost afin d'améliorer significativement les modèles implémentés. En effet, le XGBoost nous a à la fois permis de faire un choix quant aux variables à écarter de l'étude lorsqu'elles étaient trop corrélées entre elles, mais aussi d'intégrer au modèle les interactions* significatives entre les variables.

2.4 Mise en place des modèles de coût et de fréquence

En pratique, nous avons modélisé la fréquence et le coût moyen des sinistres avec un Modèle Linéaire Généralisé puis en intégrant les interactions entre les variables obtenues à l'aide d'un algorithmes d'apprentissage prédictif: l'extreme gradient boosting. Enfin, les modèles ont été comparés du point de vue de leurs performances et de leur interprétabilité.

Pour rappel, afin de choisir un modèle GLM, il faut:

*Deux variables interagissent si l'effet de l'une sur la variable à expliquer diffère suivant les valeurs de l'autre. Il est évident que l'ajout d'une interaction augmente la dimension explicative du modèle.

- choisir la loi de $Y|X = x$ dans la famille exponentielle des GLM
- choisir une fonction de lien inversible g

Ensuite, il faudra estimer les paramètres $\beta = (\beta_0, \beta_1, ..., \beta_p)^t$. Une fois cette estimation réalisée, on disposera d'une estimation de $\eta(x)$ ainsi que de $E[Y|X = x] = g^{-1}(\eta(x))$ (avec g la fonction de lien).

Dans le cadre de notre étude, nous avons modélisé le coût et la fréquence comme suit :

Y	Fréquence de sinistres	Coût des sinistres
Loi	Poisson	Log Normale
Fonction de lien	log(x)	log(x)

Figure 2.14: Structure choisie des modèles de coût et de fréquence

De plus, nous avons fait l'hypothèse que les données internes sont décorrélées de l'effet géographique.

2.4.1 Vérification de l'hypothèse d'indépendance

Calcul des coefficients de corrélations

Afin d'avoir une première intuition quant à la véracité de l'hypothèse d'indépendance, nous avons commencé par calculer trois mesures différentes pour chacune des garanties : le coefficient de Pearson, le coefficient de Kendall et le coefficient de Spearman. Plus ces coefficients sont proches de 0, plus la dépendance entre les variables est faible. Nous obtenons les résultats ci-dessous pour les garanties dégâts des eaux et vol :

Pearson	Kendall	Spearman
-0.00036	0.00213	0.00321

Figure 2.15: Coefficients de corrélation pour la garantie dégâts des eaux

Pearson	Kendall	Spearman
0.01139	0.03184	0.04711

Figure 2.16: Coefficients de corrélation pour la garantie vol

Les résultats obtenus montrent une faible corrélation entre les deux variables que ce soit pour la garantie vol ou dégâts des eaux. Ainsi, l'hypothèse d'indépendance entre les variables d'intérêts reste plausible.

Copule d'indépendance

Dans un deuxième temps, afin de valider cette hypothèse d'une manière plus rigoureuse, nous avons comparé la représentation en deux dimensions de la copule* empirique à la copule indépendante.

*Une copule est une distribution multivariée dont les lois marginales sont uniformes sur et qui possède une structure

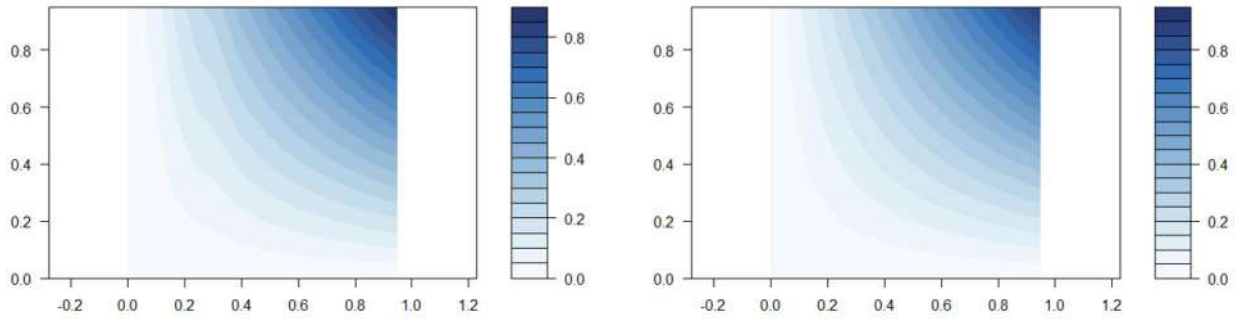


Figure 2.17: Copule empirique (à gauche) et copule indépendante (à droite) pour la garantie dégâts des eaux

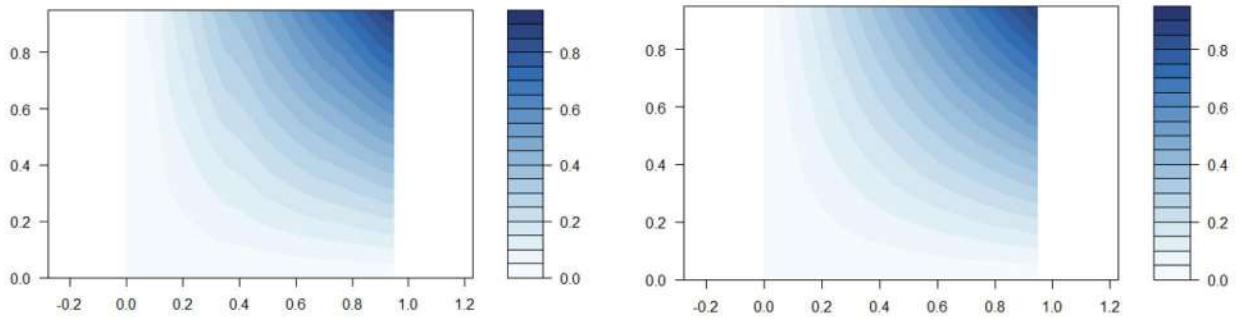


Figure 2.18: Copule empirique (à gauche) et copule indépendante (à droite) pour la garantie vol

La copule indépendante présente des courbures lisses et convexes. Ainsi, plus la copule empirique s'en approche, plus l'hypothèse d'indépendance est crédible. Au vu de nos résultats et de la proximité entre les deux représentations des copules, nous pouvons supposer l'indépendance du coût et de la fréquence dans la suite de notre étude.

Nous avons cherché à affiner notre analyse en testant deux autres approches. La première était d'analyser les scatter plot. En effet, une indépendance entre la fréquence et le coût des sinistres montrerait une uniformisation des points dans le *scatter plot*. Or, les résultats affichaient des points non répartis uniformément (certains endroits étaient plus dense que d'autres):

de dépendance particulière. Les copules permettent ainsi de créer des distributions multivariées dont les marginales sont de lois quelconques et dont les propriétés de dépendances peuvent être très variées.

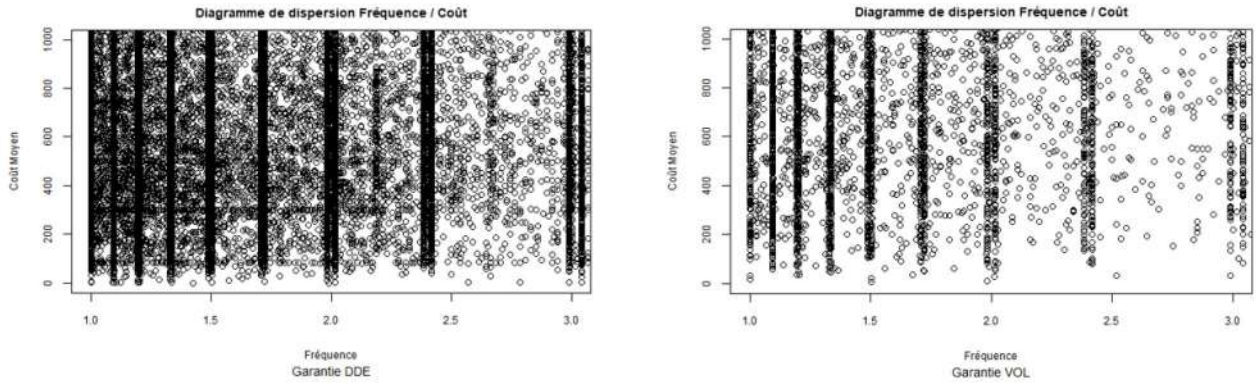


Figure 2.19: Scatter plots pour la garantie DDE (figure de gauche) et la garantie VOL (figure de droite)

Il ne semble pas y avoir une parfaite indépendance entre les deux variables. Cependant, il est possible de se rapprocher de cette hypothèse en utilisant les copules. Ainsi, nous avons considérés les résultats obtenus à travers les coefficients de corrélations et les représentations des copules pour pouvoir valider notre hypothèse.

En pratique, nous avons suivi à chaque fois le même processus afin d'établir indépendamment les modèles de coût et de fréquence:

Figure 2.20: Étapes d'implémentation d'un GLM

Dans la suite, nous avons utilisé le logiciel R afin de lancer les modèles tout en appliquant les points ci-dessous:

- Pour établir chacun des modèles, nous avons décidé de séparer le jeu de données en une base d'apprentissage (80%) et une base de test (20%). Ainsi, en pratique, nous nous baserons sur des données d'apprentissage pour capturer toutes propriétés et corrélations présentes dans le *Training Set*. Puis, dans un deuxième temps, nous vérifierons que le modèle obtenu est généralisable en testant celui-ci sur le *Testing Set* (données qui ne sont pas présentes dans le *Training Set*).
- Nous avons fait certaines hypothèses concernant chacun des modèles. En pratique nous utiliserons la loi de Poisson pour le modèle de fréquence et la loi lognormale pour le modèle de coût et la fonction de lien log qui permet d'avoir un tarif multiplicatif.
- Pour chaque test statistique effectué, un seuil de 5% a été considéré.
- Nous avons procédé à l'optimisation des différents paramètres de chaque algorithme via des procédures de validation croisées, dont le but est de minimiser l'erreur moyenne quadratique de validation (RMSE).
- Nous avons masqué les valeurs des coûts, des fréquences et des coefficients estimés sur l'ensemble des figures présentées pour des raisons de confidentialité.

2.4.2 Implémentation du modèle de coût

Séparations des sinistres graves et attritionnels

La séparation des sinistres graves et attritionnels est indispensable car la modélisation de ces deux types de sinistres ne se fait pas de la même manière. En effet, les sinistres graves ont une faible fréquence, ce qui représente un challenge pour leur analyse, notamment du fait du manque d'observations de ceux-ci.

La modélisation des sinistres dits graves fait souvent l'objet d'une attention particulière et utilise la théorie des valeurs extrêmes, qui ne sera pas abordée ici. Dans le cadre de ce mémoire, nous nous limiterons à l'analyse des sinistres hors graves.

Il s'agit donc ici de déterminer un seuil pertinent qui séparera les sinistres graves (supérieurs au seuil fixé) des sinistres attritionnels (inférieurs à ce seuil). Pour ce faire, nous avons tracé les *mean excess plots* suivants:

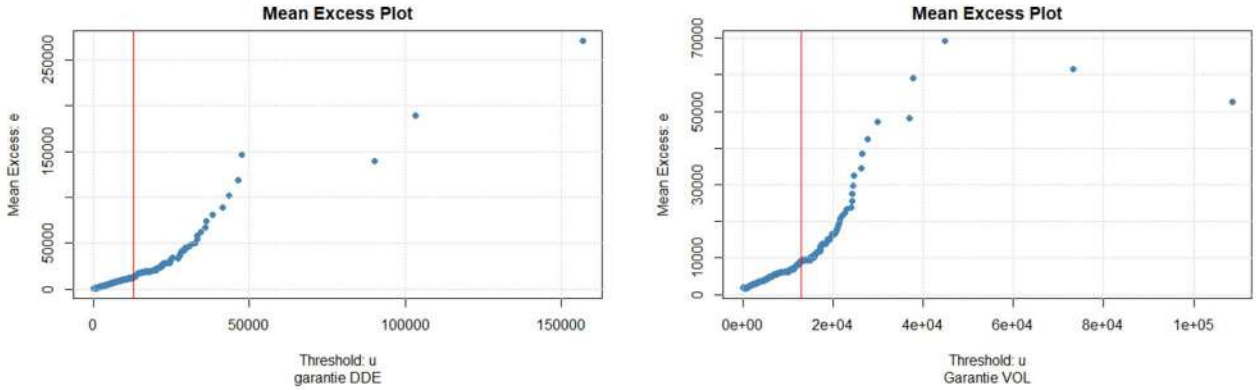


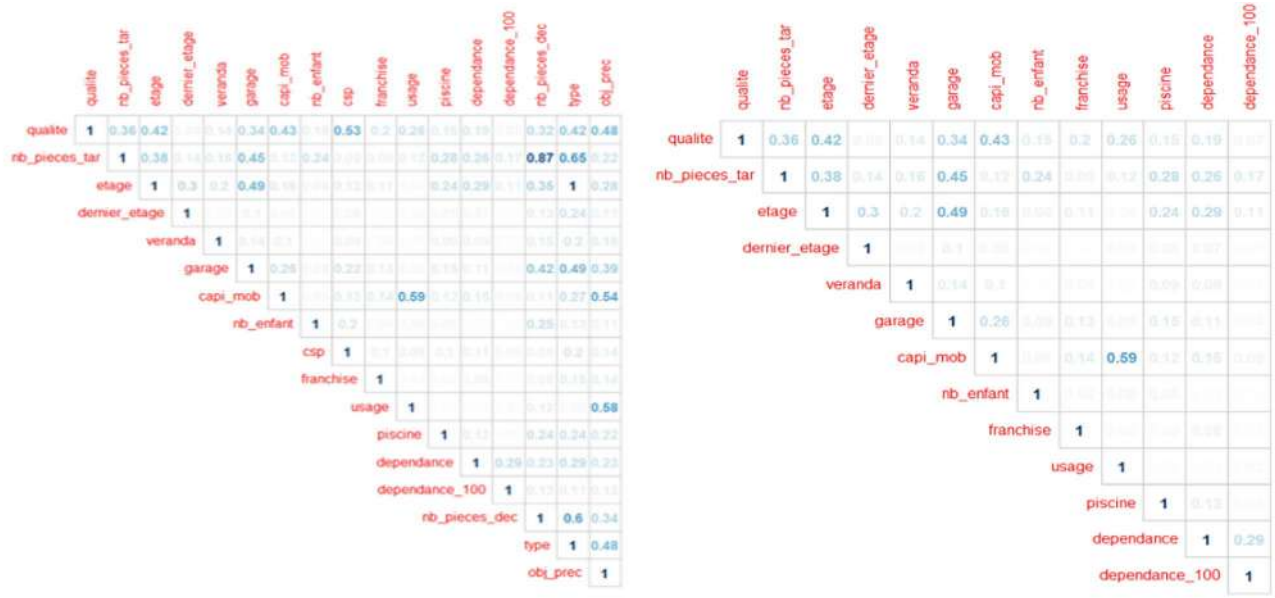
Figure 2.21: Mean Excess Plots (à gauche: Garantie DDE, à droite: Garantie VOL)

Au vu des résultats, le seuil de la séparation entre les sinistres graves et attritionnels se situe à 5 800 € pour la garantie vol et 2 800 € pour la garantie dégâts des eaux. Bien que ces seuils reflètent une vision purement statistique, nous avons soumis ces analyses à des experts métiers qui ont choisi de significativement réviser à la hausse le seuil pour converger vers 13 000 € sur les deux garanties. Ce jugement d'expert pourrait faire l'objet d'analyses complémentaires dans des études ultérieures. Les sinistres graves représentent $0,9\%$ des valeurs dans le cas de la garantie vol et $0,3\%$ pour la garantie DDE et ne seront donc pas traités dans le cadre de ce mémoire. Ces sinistres écrêtés constituent 9,28% de la charge totale des sinistres pour la garantie vol et 6,64% pour la garantie dégâts des eaux.

Corrélation des variables (V de cramer)

Dans un premier temps, nous avons calculé le V de cramer qui nous a permis de nous rendre compte des corrélations entre nos variables de départ.

En appliquant cette méthode aux deux garanties considérées et en décidant d'écarter de l'étude les variables qui ont une corrélation supérieure à 0,5, nous obtenons les résultats suivants :



(a) avant suppression des variables corrélées (b) après suppression des variables corrélées

Figure 2.22: Analyse des corrélations pour la garantie DDE - V de Cramer

Les V de Cramer obtenus nous indiquent de fortes relations entre le *nombre de pièces tarifées* et le *nombre de pièces déclarées*; le *type de logement* et la variable *étage*; le *capital mobilier* et la variable *objet précieux*. Nous décidons donc d'écarter les variables *nombre de pièces déclarées*, *type* et *objet précieux* pour la suite de l'étude.

(a) avant suppression des variables corrélées (b) après suppression des variables corrélées

Figure 2.23: Analyse des corrélations pour la garantie vol - V de Cramer

Pour les même raisons que pour l'analyse précédente, nous retirons de la base de donnée les variables suivantes: *Type, Nombre de pièces déclarées, Objets précieux.* Nous relevons également une forte corrélation entre la variable *qualité* et plusieurs autres variables du jeu de donnée. Nous remarquons à travers la méthode de *Machine Learning* XGBboost appliquée à toutes les données (en incluant les variables corrélées entre elles) que l'importance de la variable *qualité* ne vient qu'en dixième position*: Par conséquent, nous choisissons de poursuivre l'étude sans cette dernière variable.

Les V de Cramer obtenus après avoir retiré les variables mentionnées ci-dessus sont tous inférieurs à 0,5 à l'exception des variables *Usage* et *Capital Mobilier* qui seront traitées à l'issu de l'étape de transformation en dummies†. Le but étant d'obtenir des variables explicatives n'ayant pas énormément de liens entre elles, ces résultats semblent être satisfaisants pour la suite de l'étude.

Une fois cette étape réalisée, nous avons transformé les variables qualitatives en dummies pour obtenir un modèle contenant moins de variables. Puis, pour chacune des variables transformée, nous avons choisi d'écarter à chaque fois la modalité qui est la plus représentée dans notre portefeuille. En effet ces variables écartées représenteront l'individu de référence. Ce choix aura un impact uniquement sur l'interprétation des résultats.

Pour s'assurer qu'aucun effet important n'a été oublié lors de la construction du modèle, nous avons vérifié que l'ajout des variables non sélectionnées pour la modélisation n'améliorerait pas de manière significative la qualité du modèle. Cela est établi en comparant les AIC des modèles obtenus en intégrant une variable plutôt qu'une autre dans l'ensemble des variables explicatives.

Détermination de la distribution du coût et choix de la fonction de lien

L'implémentation du GLM nécessite de faire une hypothèse quant à la distribution des coûts des sinistres. L'étude de la distribution des charges des deux garanties nous a permis de nous rendre compte d'un biais de distribution provenant non pas des conventions CIDRE/IRSI mais plutôt des forfaits

*Se référer à la partie "Graphiques XGBoost" de l'annexe A.5 pour voir les résultats

†Une variable *dummy* est une variable qui prend comme valeur 0 ou 1 pour indiquer la présence ou non d'une variable catégorielle

d'ouverture des sinistres. En effet, les charges de sinistres contiennent une partie règlement et une partie provision qui est fixée de manière forfaitaire par les gestionnaires. Les procédures de règlements et de conciliation sont mises en place au fil du temps ce qui explique que pour les survenances récentes, une partie de la charge représente de l'estimé et doit donc faire l'objet d'un retraitement. Ainsi, il convient d'appliquer des coefficients de vieillissement qui matérialisent l'écart entre les normes de provisionnement et le coût réel ou de retraiter directement la donnée. Dans notre cas, nous avons fait le choix de supprimer les sinistres qui contribuent à la formation des pics afin d'obtenir des courbes lisses et ainsi établir nos conjectures quant aux hypothèses de distribution des coûts. Nous avons retiré les valeurs des montants sur les figures ci-dessous pour des raisons de confidentialité.

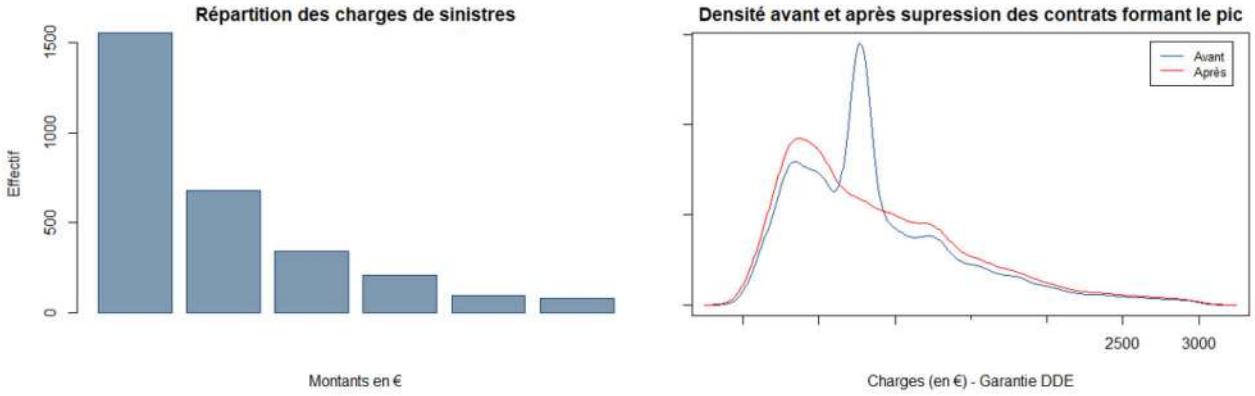


Figure 2.24: Répartition des charges de la garantie DDE

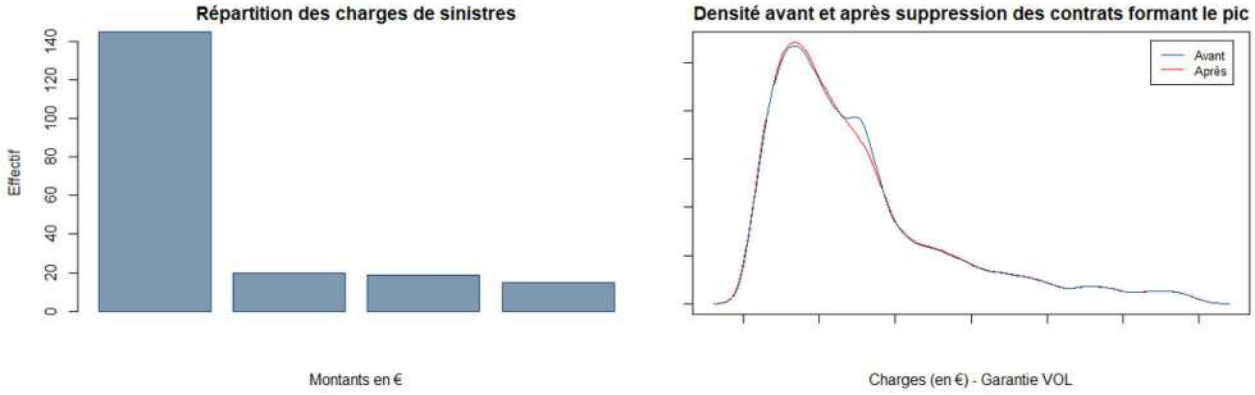


Figure 2.25: Répartition des charges de la garantie VOL

Nous avons testé plusieurs combinaisons possibles de lois (Gamma, Normale etc.) et de fonctions de liens (identité, log etc.). Nous avons choisi la déviance et l'AIC comme critères de sélection afin de choisir le modèle qui correspond le mieux à nos données.

Modèle	Déviance	Déviance/ddl	AIC
LogNormale avec lien identité	22 691	0,712	79 761
LogNormale avec lien log	22 688	0,711	79 757
Gamma avec lien log	21 085	0,661	501 159
Gamma avec lien inverse	21 043	0,660	501 088

Table 2.1: Comparaison des modèles de coût pour la garantie dégât des eaux

Modèle	Déviance	Déviance/ddl	AIC
LogNormale avec lien identité	6 059	0,771	20 381
LogNormale avec lien log	6 058	0,771	20 380
Gamma avec lien log	5 652	0,719	132 388
Gamma avec lien inverse	5 658	0,720	132 397

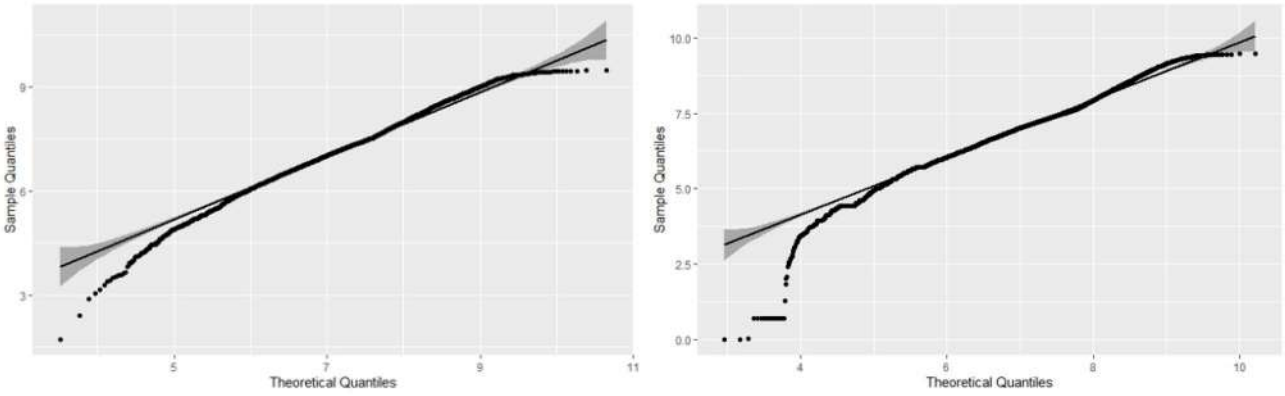
Table 2.2: Comparaison des modèles de coût pour la garantie vol

Au vu des résultats, pour notre modèle de coût, nous choisissons de conserver le modèle lognormal avec une fonction de lien logarithmique malgré le fait qu'il n'ai pas une déviance minimale. Cela facilitera l'interprétation des résultats. Néanmoins, ce modèle est satisfaisant car le rapport déviance sur degrés de liberté est inférieur à 1 et il minimise l'AIC.

Remarque: En pratique, afin d'utiliser la loi lognormale, nous commençons par convertir la variables explicative représentant le coût en logarithme du coût. Puis nous utilisons une loi Gaussienne pour la modélisation de cette variable transformée.

Adéquation de la loi:

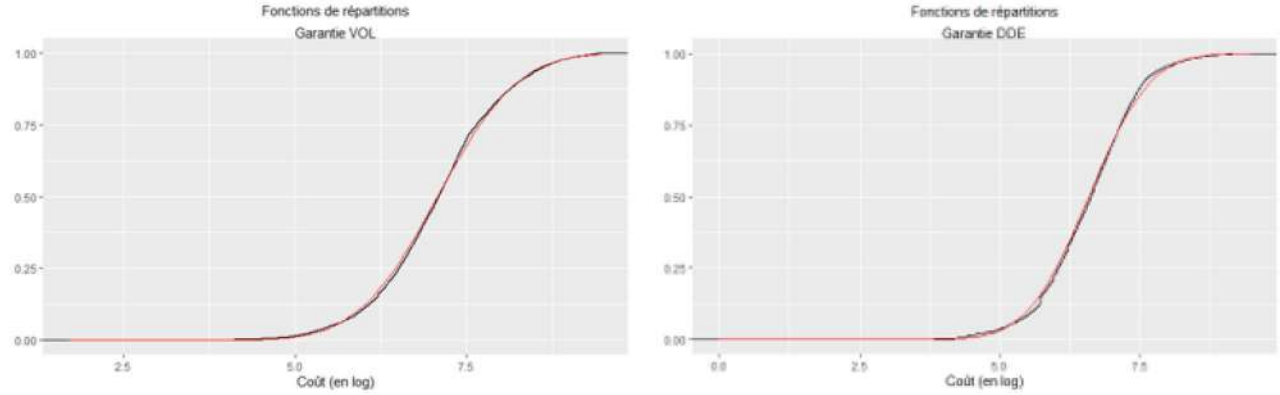
Dans un premier temps, le diagramme Quantile-Quantile (QQ-plot) nous permet d'apprécier l'adéquation de la distribution observée à la loi choisie. Le graphique obtenu ci-dessous correspond aux résultats du QQ-plot qui compare les quantiles théoriques (en abscisse) aux quantiles observées (en ordonnée). Les points étant positionnés suivant la première diagonale, nous pouvons en conclure que la distribution théorique choisie est pertinente pour chacune des garanties considérée. Par ailleurs, nous soulignons une surreprésentativité de la distribution à gauche qui pourrait inciter à complexifier l'approche dans des études ultérieures.



(a) Garantie vol (b) Garantie dégâts des eaux

Figure 2.26: Adéquation de la loi au modèle de coût

Pour appuyer notre choix, nous avons établi un deuxième test qui consiste à comparer les fonctions de répartitions empiriques et théoriques. Selon les résultats ci-dessous, la même conclusion est faite : la loi lognormale est appropriée pour modéliser le coût des sinistres pour chacune des garanties considérées.



(a) Garantie vol (b) Garantie dégâts des eaux

Figure 2.27: Fonctions de répartitions empiriques et théoriques (loi)

Pour finir, nous avons analysé les résidus afin de confirmer notre choix et valider les hypothèses concernant la loi choisie et la fonction de lien. Les graphiques ci-dessous mettent en évidence que le modèle lognormale (à gauche) est mieux adapté que le modèle Gamma (à droite) pour la garantie DDE :

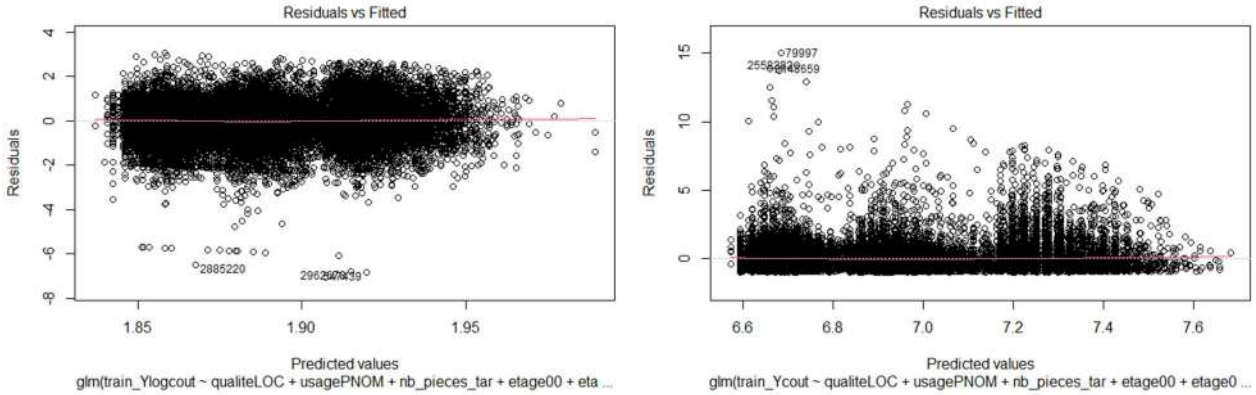


Figure 2.28: Graphique des résidus en fonction des valeurs ajustées pour la garantie dégâts des eaux

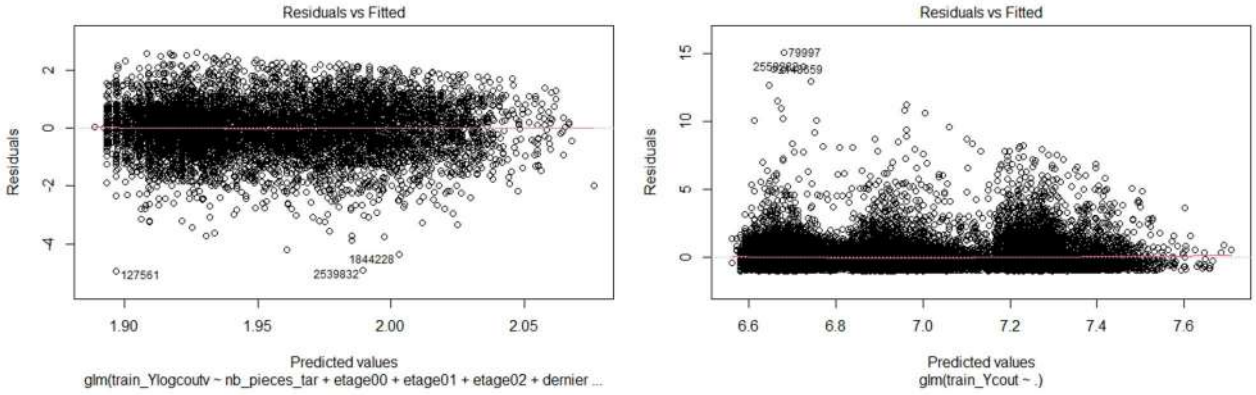
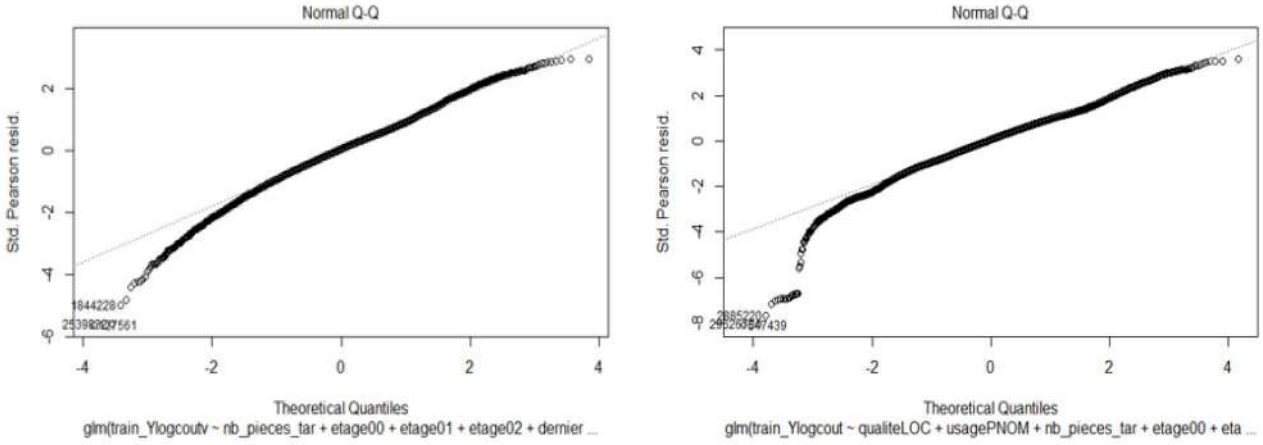


Figure 2.29: Graphique des résidus en fonction des valeurs ajustées pour la garantie vol

De plus, l'analyse des droites de Henry des valeurs résiduelles des deux garanties étudiées permettent de vérifier l'hypothèse selon laquelle les valeurs résiduelles sont normalement distribuées. La droite de Henry des valeurs résiduelles doit suivre approximativement une ligne droite, ce qui est le cas dans les résultats obtenus :



(a) Garantie vol (b) Garantie dégats des eaux

Figure 2.30: Droite de Henry des résidus

En complément, nous avons tâcher de mettre en place un test de Kolmogorov Smirnov pour tester H0: $\mathcal{F} = \mathcal{F}_0$ contre H1: $\mathcal{F} \neq \mathcal{F}_0$ avec $\mathcal{F}_0$ la fonction de répartition de la loi lognormale. Cependant, en appliquant le test sur toute la base de donnée, nous obtenons une petite p-valeur impliquant ainsi le rejet de l'hypothèse nulle au niveau 0,5%.

Au final, les résultats des trois tests graphiques établis ci-dessus nous permettent de tirer une seule et unique conclusion quant à la distribution des coût de sinistres. Il est donc légitime de poursuivre notre étude en choisissant une loi Log Normale et une fonction de lien logarithmique.

Sélection des variables

Les résultats des GLM avant sélection de variable et consultables en annexe A.6 sont issus de la fonction *anova* permettant de réaliser un test de significativité des variables explicatives. Ce test

a pour but de tester l'hypothèse de nullité d'un coefficient*. Ainsi, une petite p-valeur implique le rejet de l'hypothèse nulle et donc une significativité de la variable considérée. Dans notre cas, nous remarquons que plusieurs variables possèdent une p-valeur supérieure au seuil alpha fixé ($\alpha = 5\%$). Nous procéderons donc à une étape de sélection de variables afin de réduire le modèle en supprimant les termes non significatifs.

Nous avons utilisé les méthodes Backward et Stepwise pour entamer une première sélection de variables. La comparaison des AIC de chacun des modèles obtenus, à savoir 79749 pour la méthode Backward et pour la méthode Stepwise, nous conduit à choisir un de ces deux derniers modèles pour comparer la significativité des variables.

Dans un deuxième temps, nous avons décidé d'améliorer le modèle obtenu en intégrant les interactions de variables les plus significatives obtenues à l'aide du XGBoost. En effet cette méthode de *Machine Learning*, contrairement aux méthodes usuelles (*backward, forward et stepwise*) permettent de récupérer non seulement les variables les plus significatives mais aussi les interactions entre les variables qui sont importantes dans la modélisation. Ainsi, l'application du XGBoost nous a permis d'obtenir les résultats suivants:

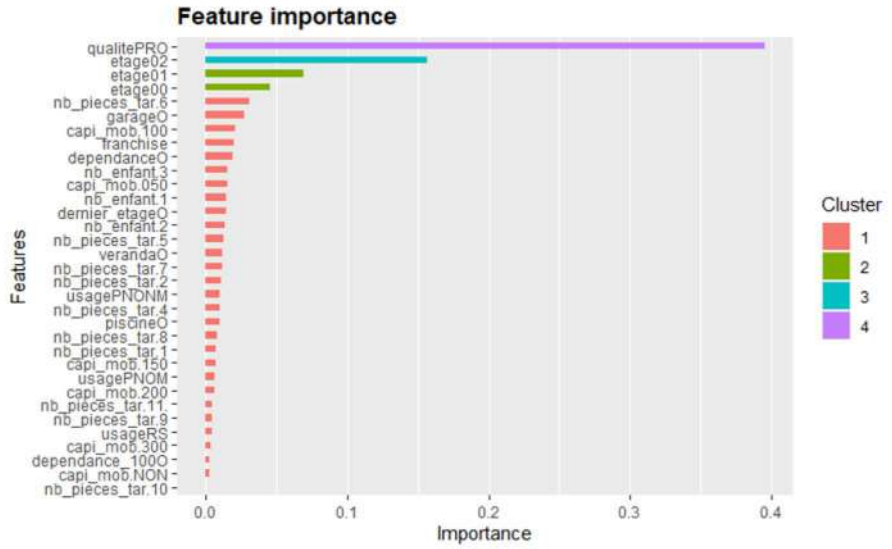


Figure 2.31: Importance des variables à travers la méthode XGBoost

*Pour rappel, les p-valeurs obtenues dans les modèles présentés correspondent aux p-valeurs du test:

- $H_0 : \beta_j = 0$
- $H_1 : \beta_j \neq 0$

	Var1	Var2	Gain.Percentage
	All	All	All
1	etage02	qualitePRO	29.9600%
2	etage01	etage02	13.3400%
3	nb_pieces_tar.6	qualitePRO	7.5700%
4	etage00	etage01	6.4100%
5	dernier_etageO	qualitePRO	1.8700%

Figure 2.32: Interactions entre les variables par la méthode XGBoost

Les résultats ci-dessus ont été obtenu via la librairie xgboost qui indique directement le gain en calculant la moyenne de la réduction de la fonction coût (pour les données d'entraînement) quand une variable est utilisée pour une division au niveau d'un noeud.

In fine, nous obtenons les 3 modèles ci-dessous à comparer:
Modèle 1 : sans sélection de variables.
Modèle 2 : après sélection de variables par la méthode Stepwise sans interactions entre les variables.
Modèle 3 : après sélection de variable par la méthode Stepwise avec intéractions entre les variables.

Choix du meilleur modèle

Toutes les métriques citées dans la section précédente sont à utiliser ensemble pour mieux comprendre et caractériser la qualité d'un modèle de prédiction.Ainsi, nous avons établi un tableau afin de comparer et choisir le meilleur modèle entre les trois modèles obtenus.

Critères d'évaluation de la qualité d'ajustement des modèles			
Critère	Modèle	Garantie dégâts des eaux	Garantie vol
Déviance			
	GLM1	22688	6071.8
	GLM2	22694	6068.9
	GLM3	22638	6061.1
AIC			
	GLM1	79757	20380
	GLM2	79744	20360
	GLM3	79668	20361
Deviance/DDL			
	GLM1	0.711	0.772
	GLM2	0.711	0.770
	GLM3	0.710	0.771

Pour les garanties dégâts des eaux et vol, le modèle pour lequel l'AIC et la déviance sont les plus faibles est le GLM3 qui correspond au modèle incluant les interactions entre les variables jusqu'au niveau 3. Nous remarquons que la déviance du 3^{eme} modèle pour la garantie dégâts des eaux est légèrement supérieure à celle du modèle 2. Toutefois, cet écart reste à la marge

De plus, en ce qui concerne le contrôle de la légitimité du modèle, nous devons également nous intéresser au rapport entre la déviance et le nombre de degrés de liberté des résidus. Nous admettons que le modèle est pertinent si le rapport de la déviance sur les degrés de liberté n'est pas grand devant 1.

Critères d'évaluation du pouvoir de prédiction des modèles			
Critère	Modèle	Garantie dégâts des eaux	Garantie vol
RMSE			
	GLM1	1078.642	1735.152
	GLM2	1078.222	1734.078
	GLM3	1076.3641	1736.757
Ratio			
	GLM1	1.0087	1.0244
	GLM2	1.0090	1.0246
	GLM3	1.0087	1.0258
Estimation du coût moyen			
	GLM1	1050.571	1835.838
	GLM2	1050.825	1836.215
	GLM3	1051.285	1838.389

Les métriques présentes dans le tableau ci-dessus nous permettent de comparer les prédictions du GLM aux valeurs réellement observées. Ainsi, plus le RMSE est faible, meilleur est la pouvoir de prédiction du modèle. Plus le ratio, qui correspond au rapport des valeurs prédites sur les valeurs observées, est proche de 1, meilleur est le modèle. Par conséquent, les résultats nous poussent à choisir le dernier modèle, c'est à dire, le modèle correspondant à celui incluant les intéractions entre les variables.

Prédictions et interprétation des résultats

La fonction de lien log permet une meilleure interprétabilité des coefficients. Par exemple, si le coût des sinistres d'une habitation doit être estimé, le caractère multiplicatif de la fonction exponentielle rend les coefficients bien plus facilement interprétables. La fonction de lien étant la fonction logarithmique, un coefficient négatif indique un risque moins fort que pour l'individu de référence et inversement un coefficient positif indique un risque plus fort. L'exponentielle du coefficient du GLM peut alors être utilisée en tant qu'indicateur de l'influence de la modalité par rapport à la classe de référence.

Le profil de référence du modèle que nous considérons ici (correspondant au profil le plus exposé au risque) est le suivant :

Modalités concernant le logement	Modalités concernant l'assuré
Résidence principale	Locataire
Appartement	CSP : 5 (employés)
3 pièces tarifées	Pas d'enfant
Avec franchise	
Pas de piscine, ni véranda, ni dépendance, ni garage	
Le capital mobilier : *capi_mob* 025	

Figure 2.33: Profil de référence des modèles de coût

En considérant cela, nous obtenons les résultats suivants:

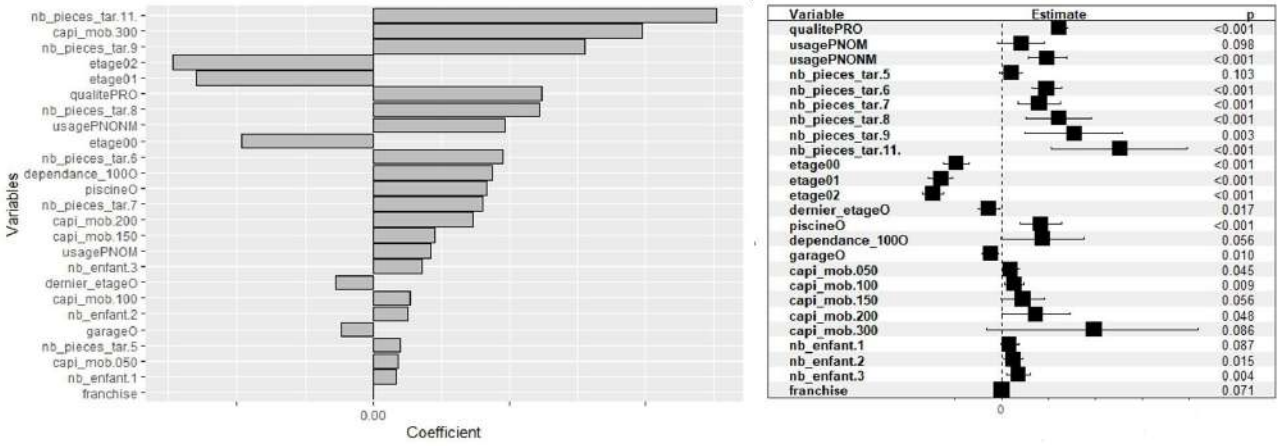


Figure 2.34: Coefficients estimés pour la garantie DDE

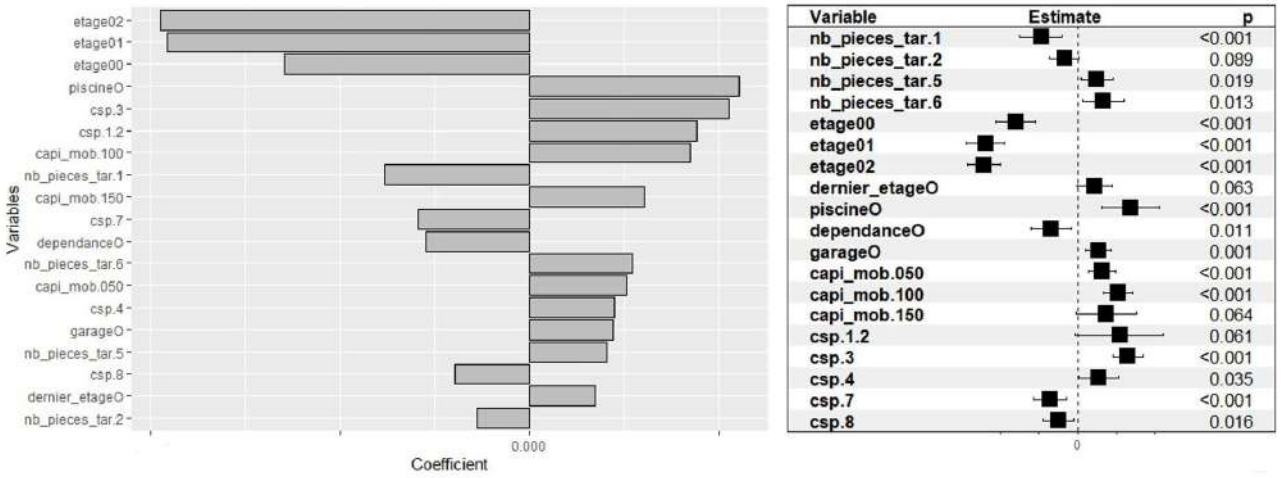


Figure 2.35: Coefficients estimés pour la garantie VOL

La dernière colonne p donne la p-valeur asymptotique du test de Wald. Lorsque la p-valeur est très petite, cela indique que les facteurs sont significatifs. Le nombre de pièces tarifés, le capital mobilier ainsi que l'étage où est situé le logement sont les variables les plus significatives lorsqu'on considère la garantie dégâts des eaux. En revanche, pour la garantie vol, ce sont les variables *étage* et *piscine* qui ressortent. Cela est en adéquation avec le fait que ce sont les maisons qui sont le plus concernées par le risque vol que les appartements.

Des graphiques supplémentaires ont été réalisés et permettent d'appréhender l'impact des modalités de chacune des variables explicatives sur la variable à expliquer. Ceux-ci sont regroupés par type de garantie en annexe A.7 .

Contrairement aux lois Gamma et Inverse-Gaussienne, la loi Log Normale n'appartient pas à la famille exponentielle. Le modèle Log Normale est obtenu en considérant une régression linéaire classique sur le logarithme de la variable d'intérêt :

$$\ln(Y) = \beta_0 + \sum_{i=1}^{p} \beta_i X_i + \epsilon$$

avec $\epsilon \sim \mathcal{N}(0, \sigma^2)$. Cette façon détournée de poser le modèle, tire sa véracité du fait qu'une variable aléatoire $Y \sim \mathcal{LN}(\mu, \sigma^2)$ si et seulement si $\ln(Y) \sim \mathcal{N}(\mu, \sigma^2)$. Cependant, nous remarquons que :

$$\begin{cases} \mathbb{E}[Y] = \exp\left(\mu + \dfrac{\sigma^2}{2}\right) \neq \exp\left(\mathbb{E}[\ln(Y)]\right) \\ \mathbb{V}[Y] = \exp\left(2\mu + \sigma^2\right) \times \left(\exp(\sigma^2) - 1\right) \neq \exp\left(\mathbb{V}[\ln(Y)]\right) \end{cases}$$

Au final, étant donné que la variable d'intérêt est modélisée selon un modèle lognormal, pour passer des estimations faites à partir du modèle sur $\ln(Y)$ à des prédictions sur le coût Y, il est nécessaire de multiplier la prédiction par $\exp(\sigma^2/2)$.

2.4.3 Implémentation du modèle de fréquence

A l'image de l'approche adoptée pour le modèle de coût, le modèle de fréquence sera réalisé en respectant le protocole en 4 étapes : la détermination des corrélations entre les variables à travers le V de Cramer; la détermination de la distribution de la fréquence et choix de la fonction de lien; la sélection des variables explicatives et le choix puis validation du meilleur modèle. L'ensemble de ces étapes sont regroupés en annexe A.8, nous nous focalisons ici uniquement sur les résultats ainsi que leurs interprétations.
Remarque : Pour le modèle de fréquence, par manque de temps, nous n'avons pas utilisé l'algorithme XGBoost pour améliorer la sélection de variables au sein de nos modèles, ce travail supplémentaire peut faire l'objet d'une piste d'amélioration.

Prédictions et interprétation des résultats

Le profil de référence du modèle de fréquence est le même que celui du modèle de coût. Les résultats de la régression de Poisson sont alors présentées dans les figures ci-dessous.

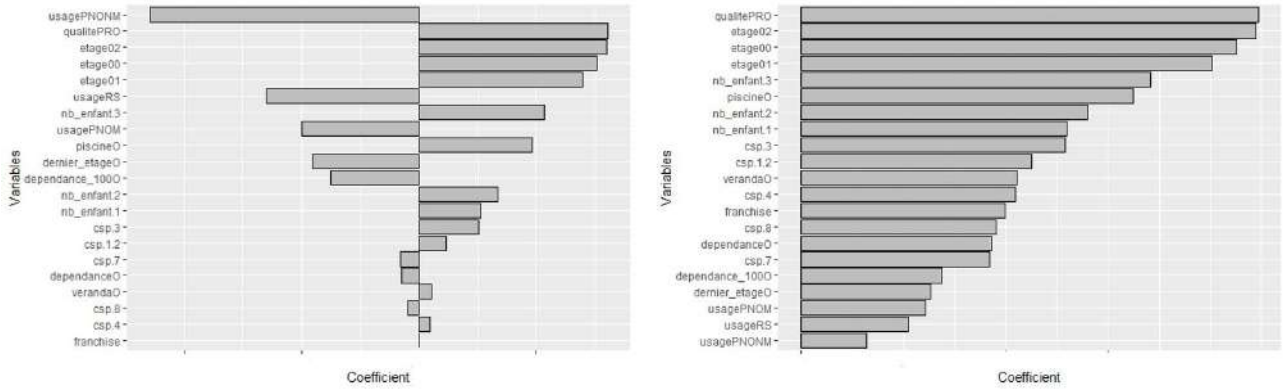


Figure 2.36: Coefficients estimés pour la garantie DDE

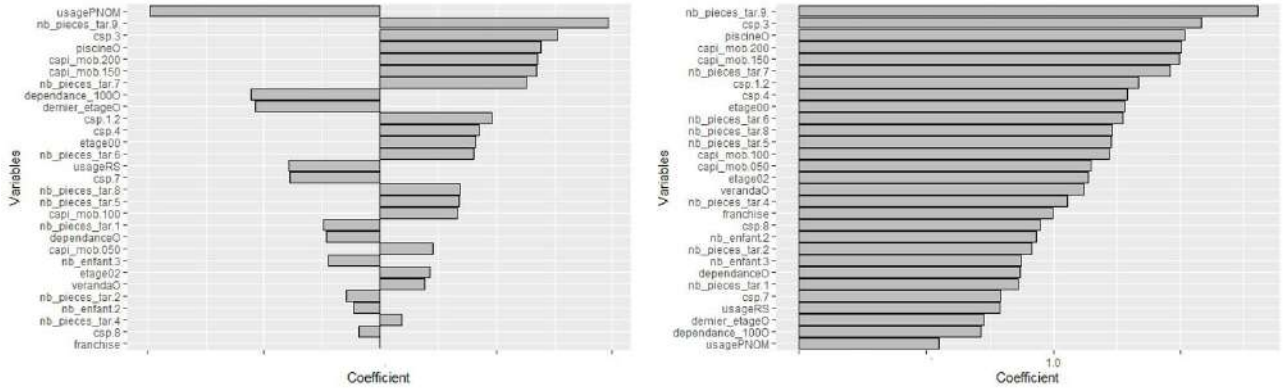


Figure 2.37: Coefficients estimés pour la garantie VOL

Les paramètres s'interprètent de la même manière que dans le modèle de coût (par rapport au profil de référence spécifié ci-dessus). Les coefficients ont été masqués pour des raisons de confidentialité.

Chapitre 3

Intégration des variables géographiques à partir de l'*Open Data*

Dans ce chapitre, nous créons une base de données externe issue des informations de l'*Open Data*. Nous précisons les variables qui la composent, l'échelle à laquelle elles sont observées ainsi que leurs provenances.

3.1 Constitution d'une base externe à partir de l'*Open Data*

3.1.1 Base météorologique : NCEI

Présentation de la base et des variables sélectionnées

Ce mémoire s'articule autour de la construction d'un zonier pour deux garanties, en particulier, la garantie dégâts des eaux. Afin de constituer ce dernier, nous allons nous intéresser plus particulièrement à l'impact des données météorologiques sur la fréquence et la sévérité de sinistre. Dans ce contexte, nous avons récupéré des données en libre service afin de constituer une base externe regroupant les données météorologiques. Ces données météorologiques récupérées sont issues du *National Center for Environmental Informations* (NCEI) qui est le résultat de la fusion entre le plus grand centre mondial de collecte de données météorologiques: le *National Climate Data Center* (NCDC), avec les deux centres de données: le *National Geophysical Data Center* et le *National Oceanic Data Center* (NODC).

Les données présentes dans cette base proviennent d'observations de stations météorologiques terrestres par des observateurs du *National Weather Service* et sont accessibles sur le site suivant: https://www.ncdc.noaa.gov/. En ce qui concerne la France métropolitaine, nous avons accès aux données de 72 stations réparties de manière homogène sur l'ensemble du territoire Français.

Figure 3.1: Répartition des stations en France

De plus, le centre fourni des données historiques concernant le climat jusqu'à plusieurs dizaines d'années en arrière. Nous avons donc pu récupérer l'ensemble des données de 2015 à 2019 afin de couvrir la période sur laquelle nous modélisons les contrats de la base d'étude interne.

Enfin, les nombreux jeux de données proposés sont accessibles à différentes périodicités. Nous travaillons ici sur les données journalières climatiques filtrées sur les stations présentes en France métropolitaine. En effet, celles-ci comportent des données plus complètes concernant les variables de température et de précipitation. Néanmoins, afin de couvrir l'ensemble des informations à notre disposition, nous avons ajouté une variable supplémentaire correspondant au nombre de jours sur l'année où la température est inférieure à $0°C$ pour la station considérée. Cette variable nous permettra de prendre en compte les éventuels dégâts survenus à cause du gel hivernal. Elle n'est disponible que dans la base de données mensuelle issue du même site et sera traitée séparément.

Ainsi, nous entamons la constitution de le base de données externe météorologique en incluant les 5 variables suivantes : Température minimale *TMIN*, Température maximale *TMAX*, Température moyenne *TAVG*, Précipitation *PRCP* et Nombre de jours où la température est inférieure à $0°C$ *NBT0*.

Retraitements effectués

Le jeu de données récupéré nécessite certains retraitements avant d'être exploitable. En effet, généralement, les bases de données en libre service contiennent des valeurs manquantes ou incohérentes qui nécessitent d'être retraitées avant de pouvoir être exploitées. Par exemple, après avoir analysé les données des stations, nous remarquons qu'une des stations présente des valeurs globalement différentes par rapport aux autres. Il s'agit de la station du mont Aigoual qui est un sommet situé dans le Sud du Massif central, à la limite entre les départements du Gard et de la Lozère. Il culmine à 1565 mètres d'altitude et présente des caractéristiques qui sont susceptible de biaiser les résultats, notamment après l'étape d'interpolation. C'est pourquoi nous avons décidé de retirer ce point de l'étude.

Traitements des valeurs manquantes

Nous avons eu recours à des méthodes qui nous ont permis d'associer une valeur cohérente à chacune des valeurs manquantes de notre jeu de données. Pour cela, plusieurs possibilités se sont offertes à nous : remplacement des valeurs manquantes par la moyenne des valeurs de la variable considérée, application d'une méthode d'interpolation etc. Pour notre part, nous avons choisi de retraiter les données pas à pas. Pour les variables liées à la température: T_{MIN}, T_{MAX} et T_{AVG} , nous avons appliqué la relation suivante pour compléter les informations manquantes:

$$T_{AVG} = \frac{T_{MIN} + T_{MAX}}{2} \tag{3.1}$$

Ce choix de retraitement a été établi à la suite d'une analyse effectuée en amont sur les données: nous avons utilisé un indicateur d'écart afin de vérifier que les données effectivement renseignées respectaient bien la formule ci-dessus. En calculant la racine de la moyenne des carrée des erreurs (RMSE), nous obtenons un écart minime (moins de $1°C$). L'approche considérée est donc pertinente.

En ce qui concerne les données de température manquantes restantes et les valeurs de précipitations - plus complètes et de variance moins élevée que les données de température - nous avons opté pour une approche différente en remplaçant les valeurs manquantes par la valeur associée à la veille du jour concerné. Cela a été décidé après avoir analysé la distribution des valeurs manquantes dans le jeu de données.

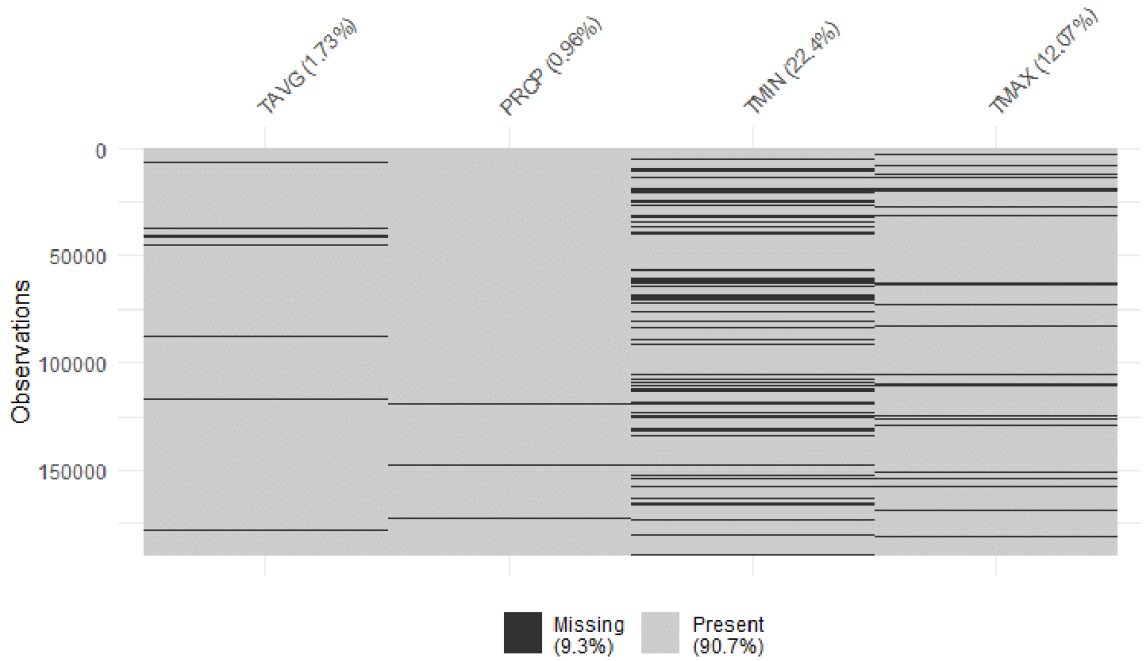


Figure 3.2: Analyse des valeurs manquantes

Les valeurs manquantes ne se trouvent pas sur plusieurs jours successifs: les quatre variables considérées ne comportent pas de bandes noires apparentes sur l'ensemble des observations. De ce fait, la méthode utilisée semble être cohérente pour cette variable.

Une fois ce travail effectué, nous avons complété l'information recueillie en ajoutant à notre base de données la variable indiquant le nombre de jours où la température est inférieure à $0°C$. Pour cela, nous avons utilisé les données mensuelles fournies par le *NCEI* et avons procédé aux mêmes étapes d'analyse et de retraitement que la base journalière pour obtenir une base de données complète sans valeurs manquantes.

Pour finir, une analyse de la distribution des valeurs retraitées (min, max, moyenne et quantiles) a été effectué pour s'assurer de la proximité des résultats post-retraitement et des données brutes.

Traitements temporels

Les données fournies pour l'ensemble des années allant de 2015 à 2019 sont journalières et trop précises pour notre étude. Nous avons donc calculé une moyenne par saison* (été/hiver) sur 5 ans pour chaque variable considérée. Cette étape permettra une meilleur interprétation des résultats. Ce retraitement nous a ainsi permis de réduire la dimension temporelle de notre base. Nous obtenons à ce stade une base comprenant des valeurs pour chacune des 72 stations considérées. Une analyse préalable visant à s'assurer de l'absence d'années atypiques a été réalisée en amont du retraitement.

De la même manière, le retraitement des données mensuelles a été effectué de telle sorte que nous puissions étudier le nombre de jours moyens où la température est inférieure à $0°C$ sur les 5 dernières années pour chaque station.

Traitements géographiques

Nous rappelons que le but est ici d'associer une valeur à chaque commune de la France métropolitaine. Or, à ce stade, nous n'avons des valeurs qu'à des points ponctuels qui correspondent aux 72 stations considérées. Il convient donc à présent d'interpoler les résultats obtenus sur tout le territoire Français afin d'obtenir des valeurs pour chaque code INSEE. Pour ce faire, il est légitime d'utiliser une méthode d'interpolation spatiale.

De manière générale, les méthodes *IDW* et *Krigeage* s'adaptent plutôt bien aux variations de terrain. Les autres méthodes sont généralement plus sensibles aux variations. En conséquence, nous avons comparé les méthodes *IDW* et *Krigeage* que nous présentons plus en détail dans la section suivante.

a. Méthodes d'interpolation spatiale

La pondération inverse à la distance (en anglais, *inverse distance weighting* ou *IDW*) est une méthode d'interpolation spatiale permettant d'assigner une valeur à tout point d'un espace à partir d'un ensemble d'observations. Cette méthode est très utilisée dans plusieurs domaines et notamment dans la météorologie.

L'interpolation de pondération par l'inverse de la distance détermine les valeurs de cellule via la moyenne pondérée d'un ensemble de points d'échantillonnage. Cette pondération est une fonction d'inverse de la distance. En d'autres termes, le poids des points voisins diminue lorsque la distance augmente.

L'IDW repose donc principalement sur l'inverse de la distance élevée à une puissance mathématique :

$$u(\mathbf{x}) = \frac{\sum_{k=1}^{N} w_k(\mathbf{x})^p u_k}{\sum_{k=1}^{N} w_k(\mathbf{x})^p}, w_k(\mathbf{x}) = \frac{1}{d(\mathbf{x}, \mathbf{x}_k)} \tag{3.2}$$

avec $\mathbf{x}$ le point à interpoler, x_k un point d'interpolation (connu), u_k la valeur de la fonction u au point x_k, d une distance donnée (opérateur de mesure) du point d'interpolation x_k au point à interpoler $\mathbf{x}$, N le nombre total de points connus utilisés dans l'interpolation, w le poids et p est un nombre positif réel appelé le paramètre de puissance.

*Les mois ont été séparés de la manière suivante: été (mars-août) et hiver (septembre-février).

Le paramètre de puissance permet de contrôler la significativité des points connus sur les valeurs interpolées en fonction de leur distance par rapport au point en sortie. Le choix de paramétrer une puissance plus élevée induit une concentration sur les points les plus proches qui auront ainsi plus d'influence dans le calcul et la surface obtenue sera moins lisse. En revanche, une valeur de puissance moins élevée accorde plus d'influence aux points distants, ce qui génère une surface plus lisse.

L'outil d'interpolation IDW (Pondération par l'inverse de la distance) que nous venons de présenter est considéré comme une méthode d'interpolation déterministe, car elle est directement basée sur des formules qui déterminent le lissage de la surface résultante. En outre, une seconde famille de méthodes d'interpolation comprend des techniques géostatistiques (telle que les méthodes de krigeage) qui sont basées sur des modèles statistiques comprenant l'auto-corrélation. Le krigeage est une technique géostatistique de modélisation spatiale permettant, à partir de données dispersées, d'obtenir une représentation homogène des informations étudiées. Cette technique permet, à l'aide des mesures des stations, d'estimer les valeurs hors station en prenant en compte les distances entre les données (les stations de mesure), les distances entre les données et la cible (le point pour lequel nous volons estimer la mesure) et la structure spatiale (grâce à l'analyse variographique).

Pour interpoler les données, nous avons commencé par nous intéresser à la méthode de krigeage qui est plus avancée que la méthode IDW. Les résultats de l'analyse variographique présentés en Annexe B.2 , nous ont permis de conclure qu'une méthode d'interpolation de type IDW est suffisante pour l'étape d'interpolation dans notre étude.

Le choix des hyperparamètres a été effectué en comparant les résultats obtenus aux cartes météorologiques de France accessibles en libre service aujourd'hui. Un choix plus rigoureux des hyperparamètres pourrait éventuellement faire l'objet d'une piste d'amélioration.

Nous rappelons que l'idée général de cette méthode consiste à calculer la valeur d'un point en effectuant la moyenne des valeurs des points connus situés dans le voisinage pondérées par l'inverse de la distance au point calculé: plus les points sont proches, plus la pondération affectée est forte. En pratique, il s'agit d'associer une valeur à chaque point d'une grille que nous superposons à notre carte de France. Cette valeur correspond à la valeur moyenne des stations situés dans le voisinage pondérées par l'inverse de la distance au point calculé. L'étape suivante consiste à récupérer l'ensemble de l'information par code INSEE, nous avons donc affecté à chacun des codes, la moyenne des valeurs de chaque point de la grille inclue dans la commune considérée. Par exemple, sur le schéma ci-dessous représentant les limites de Paris, le point rouge correspond à la station la plus proche pour laquelle nous avons une observation $T_{AVG} = 18,8°C$ et les valeurs dans la grille représentent les résultats obtenues après interpolation (notez que plus on s'éloigne de la station, plus les valeurs dans la grille diminuent). Enfin, la valeur finale affectée à la commune grisée est la moyenne de l'ensemble des données au sein de cette commune, soit $18,6°C$.

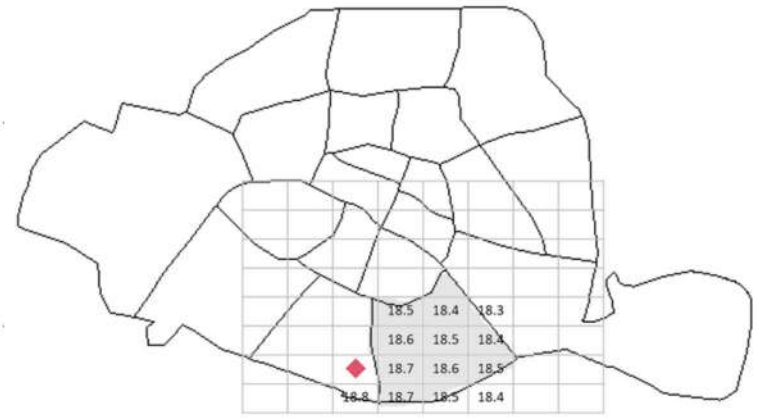


Figure 3.3: Méthode d'interpolation: IDW

En ce qui concerne la variable retraitée séparément, à savoir, le nombre de jours où la température est inférieure à $0°C$, nous avons procédé à une analyse départementale pour attribuer une valeur à l'ensemble des codes INSEE en France:

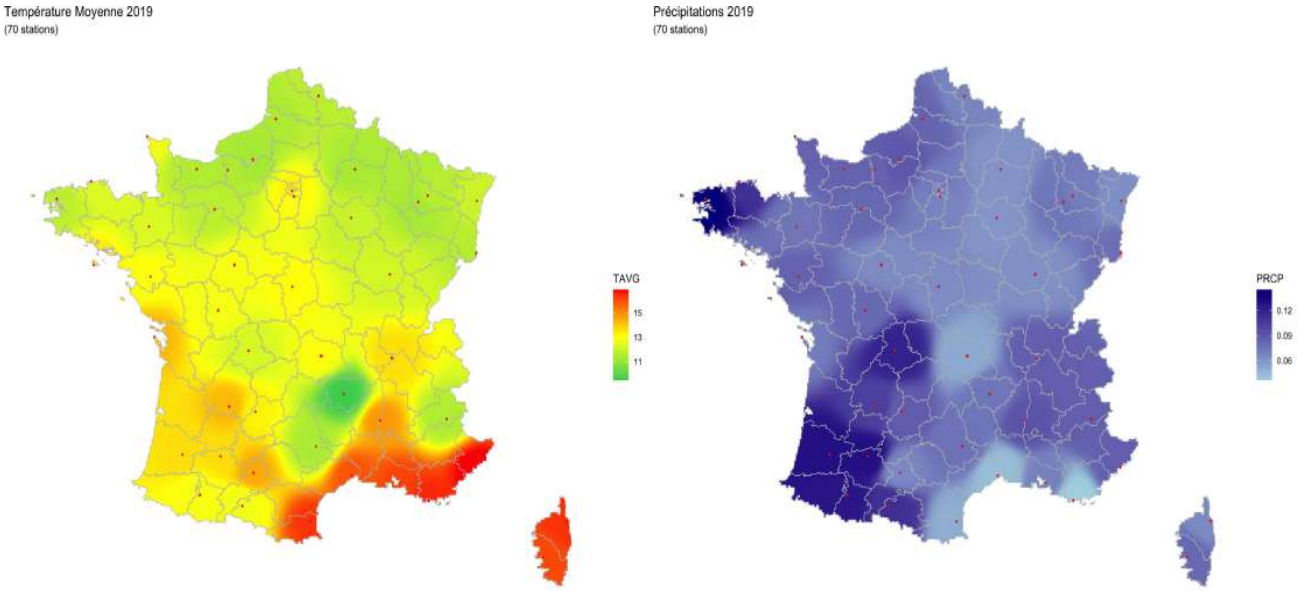
- Si le département contient une seule station, nous affectons à tous les codes INSEE inclus dans celui-ci, la valeur de la station.
- Si le département contient plus d'une station, nous affectons à chaque code INSEE la valeur moyenne de celles-ci.
- Si le département ne contient aucune station, nous choisissons d'attribuer aux codes INSEE la valeur moyenne de l'ensemble des stations dans la région considérée.

In fine, nous obtenons une base de données météorologique fiable et complète comportant toutes les informations à la maille code INSEE.

b. Cartographie

Afin de représenter graphiquement les disparités géographiques météorologiques en France métropolitaine, nous avons recours à un fond de carte, un ensemble de données géolocalisées et un système de coordonnées cartographiques ainsi qu'une projection qui déterminent la façon dont les données géographiques sont représentées. Pour plus d'informations sur notre approche et sur les éléments énumérés ci-dessus, se référer à l'annexe B.3 .

A partir de la base de donnée retraitées, nous avons établi à l'aide du logiciel R des cartes représentant les températures moyennes et les précipitations en France métropolitaine. Les résultats ci-dessous représentent les cartes obtenues après interpolation par la méthode IDW.



(a) Température moyenne (b) Précipitation

Figure 3.4: Cartes post-interpolation spatiale sur l'ensemble du territoire Français

Dans certaines zones, nous nous écartons légèrement des relevés météofrance (par exemple, pour la température : Ardennes ; frontière suisse de la Haute-Savoie et pour la pluviométrie : chaînes montagneuses Corse ; frontière espagnole). Néanmoins, il s'agit de zones assez peu peuplées. L'intégration de données de stations (payantes par exemple) pourrait contribuer à améliorer notre modèle.

3.1.2 Bases socio-démographiques : INSEE

Présentation de la base et des variables sélectionnées

L'ensemble des données socio-démographiques externes qui seront utilisées dans le but d'expliquer le risque lié à la zone géographique ont été récupérées sur le site de l'Institut National de la Statistique et des Etudes Economiques. L'INSEE est chargé de la production, de l'analyse et de la publication des statistiques officielles en France. L'accès à ses informations est gratuit dans sa quasi-totalité.

Dans le cadre de notre étude, nous avons extrait les données de deux bases différentes du site de l'INSEE, à savoir, la base comparateur des territoires et la base de logement. Il existe de nombreuses variables au sein de ces deux base de données. Toutes n'ont pas été conservées. En effet, nous avons commencé par faire une première sélection des variables en éliminant celles qui n'apporteraient aucune information sur le risque considéré. Les variables ci-dessous sont issues de la base de comparateur des territoires et ont été pré-sélectionnées pour la suite de l'étude :

Ensemble des données socio-démographiques recueillies : Comparateur des territoires	
Nom de la variable	Signification
P17_POP	Population en 2017
NAIS1217	Nombre de naissances entre le 01/01/2012 et le 01/01/2017
DECE1217	Nombre de décès entre le 01/01/2012 et le 01/01/2017
P17_MEN	Nombre de ménages en 2017
P17_LOG	Nombre de logements en 2017
P17_RP	Nombre de résidences principales en 2017
P17_RSECOCC	Nombre de résidences secondaires et logements occasionnels en 2017
P17_LOGVAC	Nombre de logements vacants en 2017
P17_RP_PROP	Nombre de résidences principales occupées par propriétaires en 2017
NBMENFISC17	Nombre de ménages fiscaux en 2017
MED17	Médiane du niveau de vie en 2017
TP6017	Taux de pauvreté en 2017
P17_POP1564	Nombre de personnes de 15 à 64 ans en 2017
P17_CHOM1564	Nombre de chômeurs de 15 à 64 ans en 2017
P17_ACT1564	Nombre de personnes actives de 15 à 64 ans en 2017

https://www.insee.fr/fr/statistiques/2521169

De même, des données relatives aux logements issues de la deuxième base de données ont permis d'avoir l'ensemble des informations suivantes (qui regroupent plusieurs variables dans chaque catégorie).

Ensemble des données socio-démographiques recueillies: Base des logements
Ancienneté d'emménagement
Catégorie socioprofessionnelle
Catégorie de logement
Emplacement réservé de stationnement
Installations sanitaires
Logement
Ménage
Mode de chauffage
Nombre de pièces
Nombre de voitures
Période d'achèvement de la construction
Population des ménages
Résidence principale
Statut d'occupation du logement (propriétaire, locataire, sous locataire ou logé gratuitement)
Taille du ménage (nombre de personnes du logement)
Type de logement

https://www.insee.fr/fr/statistiques/4515532?sommaire=4516107

Pour les bases de l'INSEE, le recensement repose sur une collecte d'information annuelle qui concerne successivement tous les territoires communaux pendant une période de cinq ans. Les communes comportant plus de 10000 habitants réalisent tous les ans une enquête par sondage auprès d'un échantillon

d'adresses qui englobent un certain pourcentage de leurs logements. Les communes de moins de 10 000 habitants réalisent une enquête de recensement portant sur toute la population, à raison d'une commune sur cinq chaque année. En cumulant les enquêtes sur cinq années, les données prises en comptes couvrent 100% des habitants des communes de moins de 10 000 habitants et environ 40 % de la population des communes comportant plus de 10 000 habitants. Pour assurer l'égalité de traitement entre les communes, les informations collectées sont ramenées à une même date. Cette date de référence est fixée au 1er janvier de l'année médiane des cinq années d'enquête pour obtenir une meilleure robustesse des données. Dans la base récupérées, les cinq dernières enquêtes de recensement complètes ont été réalisées de 2015 à 2019. Ainsi, L'ensemble des données exploitées sont issues des bases de 2017 (l'année médiane des années d'enquêtes).

Retraitements effectués

Les données socio-démographiques brutes proposées sont disponibles pour toutes les zones des niveaux géographiques suivants : France métropolitaine, département, région, commune (code INSEE) et arrondissement municipal pour Paris, Lyon et Marseille. Ainsi, les données sont bien disponible à la maille code INSEE et n'auront pas à subir un retraitement d'agrégation ou de dissociation.

Traitement des valeurs manquantes

Dans chacune des deux bases issues du site de l'INSEE, certaines variables sont dédiées aux communes appartenant aux départements d'outre-mer. Ainsi, les valeurs qui leur sont associées sont manquantes dans les bases considérées. Un premier retraitement a donc été effectué visant à éliminer les valeurs manquantes de la base en supprimant les variables uniquement dédiée aux Outre-mer.

Une fois ce traitement établi, une analyse des codes INSEE est nécessaire pour vérifier la pertinence des codes dans la base. Cette étape est indispensable pour limiter au maximum la perte d'information lors l'agrégation des bases externes (à travers le code INSEE).

Traitement des codes INSEE

En France, les communes de Lyon, Marseille et Paris sont divisées respectivement en 9, 16 et 20 arrondissements municipaux. Certains jeu de données peuvent donc contenir un unique code pour toute la commune alors que d'autres comportent différents codes pour chaque arrondissement de celle-ci. La base d'étude interne comporte des codes INSEE différents pour chaque arrondissement, une attention particulière devra être apportée aux bases de données utilisant la codification suivante:

Paris	75056
Marseille	13055
Lyon	69123

Table 3.1: Codification des communes

Il est nécessaire de s'assurer de l'homogénéité des codes dans toutes les base de données externes afin de garder le maximum d'informations lors de l'étape de concaténation de l'ensemble des bases de données externes. Dans un deuxième temps, nous avons vérifié que les codes INSEE comportent bien tous 5 chiffres. En effet, il est courant que les codes commençant par 0 sont automatiquement modifiés lors de l'importation du jeu de données sur un logiciel et se retrouvent tous avec 4 chiffres. Enfin, nous avons également vérifié que les codes INSEE correspondant à la Corse commencent bien par 2A ou 2B.

Traitement des effectifs

Certaines variables inclues dans les bases extraites font référence à des effectifs. Pour être in-

terprétables, ces variables devront être converties en taux afin de ne pas avoir un effet caché quant au nombre d'habitant par commune.

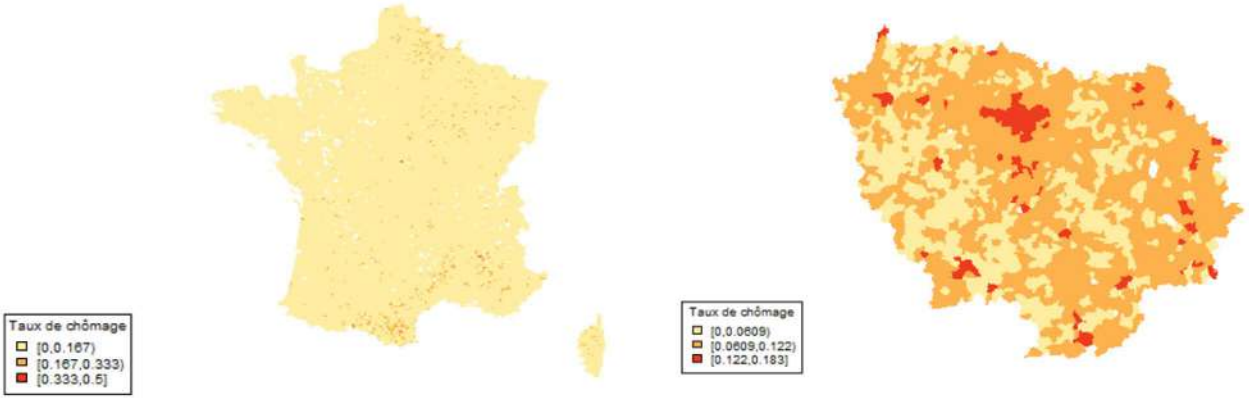
Dans la base de comparateur des territoires comme dans la base des logements, il s'agit d'effectifs de résidences principales, de ménages, de logements ou encore de personnes. Ainsi, pour effectuer le retraitement, il convient de diviser chacun des effectifs par des données cohérentes selon le cas étudié pour obtenir des taux facilement exploitables. Par exemple, dans la base de comparateur de territoire, le retraitement effectué pour la variable *Nombre de chômeurs entre 15 et 64 ans* est le suivant:

$$Taux\ de\ chômeurs\ entre\ 15\ et\ 64\ ans = \frac{Nombre\ de\ chômeurs\ entre\ 15\ et\ 64\ ans}{Part\ de\ la\ population\ entre\ 15\ et\ 64\ ans}$$

Remarque : Les effectifs supérieurs à 500 peuvent généralement être utilisés en toute confiance. Cependant, les effectifs inférieurs à 200 dans les bases de données brutes doivent être maniés avec précaution car ils peuvent ne pas être significatifs du fait de l'imprécision liée au sondage. En conséquence, les comparaisons entre territoires de petite taille sont à proscrire.

Cartographie

Pour avoir une vision globale de certaines données, nous avons représenté une des variables externes socio-démographiques sur la carte de France. Nous avons établi une carte permettant de visualiser la répartition du nombre de chômeurs entre 15 et 64 ans sur l'ensemble des territoire Français. Pour cela nous avons représenté via R la variable préalablement convertie en taux sur les cartes de France et de l'île-de-France.



(a) Taux de chômage en France

(b) Taux de chômage en île-de-France

Figure 3.5: Représentation du taux de chômage en France

3.1.3 Base de criminalité : ONDRP

Présentation de la base et des variables sélectionnées

Nous avons récupéré les données en libre service accessibles sur le site de l'Observatoire National de la Délinquance et des Réponses Pénale (ONDRP) afin de constituer une base externe regroupant les données de criminalité pour l'étude de la garantie vol. Ces données correspondent aux nombres de crimes et délits enregistrés mensuellement par les services de police et de gendarmerie et couvrent les données des 24 dernière années. Nous avons donc établi une moyenne sur l'ensemble de la plage de

donnée disponible.

L'ONDRP a pour missions la production et la diffusion de statistiques sur la délinquance, les réponses pénales, ainsi que tout autre question liée à la sécurité. La base étudiée présente un grand nombre de variables. L'ensemble de ces dernières ne vont pas toutes permettre d'expliquer le nombre de sinistres. Seules les variables les plus pertinentes en termes de sinistralité vont être retenues. Ainsi, un grand nombre de variables disponibles mais n'ayant aucun caractère influant sur la sinistralité ont été écartés en amont de l'étude. Nous conservons les 20 variables suivantes pour la constitution de la base de criminalité :

Ensemble des données de criminalité recueillies	
Nom de la variable	Signification
c1	Cambriolages de locaux d'habitations principales
c2	Cambriolages de résidences secondaires
c3	Cambriolages de locaux industriels, commerciaux ou financiers
c4	Cambriolages d'autres lieux
c5	Autres vols simples contre des particuliers dans des locaux privés
c6	Autres vols simples contre des particuliers dans des locaux ou lieux publics
c7	Violations de domicile
c8	Recels
c9	Vols d'automobiles
c10	Vols de véhicules motorisés à 2 roues
c11	Vols à la roulotte
c12	Vols d'accessoires sur véhicules à moteur immatriculés
c13	Escroqueries et abus de confiance
c14	Autres coups et blessures volontaires criminels ou correctionnels
c15	Menaces ou chantages dans un autre but
c16	Vols à la tire
c17	Vols à l'étalage
c18	Usage de stupéfiants
c19	Autres destructions et dégradations de biens privés
c20	Destructions et dégradations de véhicules privés

https://static.data.gouv.fr/resources

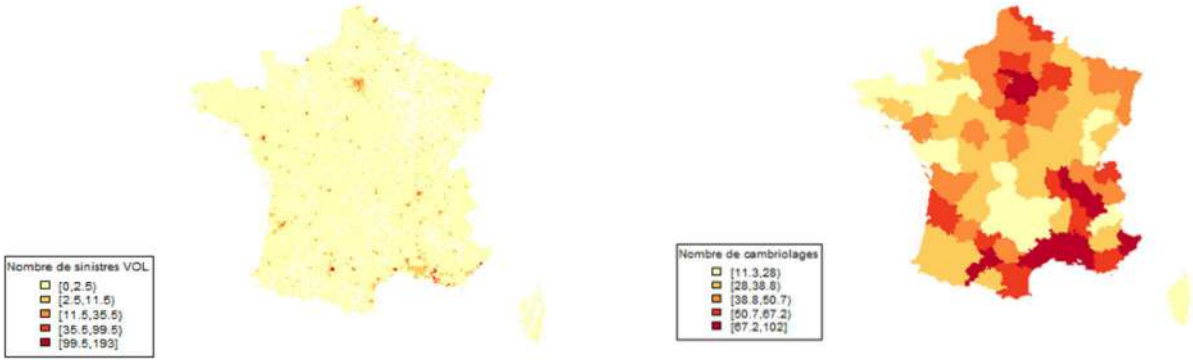
Retraitements effectués

Les données proposées sont présentées selon une maille départementale. Nous avons retraité celles-ci en dupliquant les données disponibles sur l'ensemble des communes au sein des différents départements. Ainsi, nous avons obtenu une base de données à la maille code INSEE. De plus, comme toutes données en libre service, l'utilisation et l'interprétation de ces informations doit tenir compte de certains éléments. Tout d'abord, seuls sont pris en compte les crimes et délits qui ont été enregistrés pour la première fois par les forces de sécurité afin d'éviter de comptabiliser deux fois une même infraction traitée successivement par des services différents. De plus, il est important de souligner

que les infractions sont comptabilisées selon l'administration qui les a constatées et enregistrées. Or, une infraction n'est pas toujours déclarée sur le territoire où elle a été commise. C'est par exemple le cas pour les compagnies de gendarmerie départementale ou les services de la Police aux frontières qui ont généralement une compétence sur plusieurs départements (ou sur plusieurs régions). Il est donc important de garder en tête qu'il s'agit ici du nombre de crimes et délits enregistrés par un service et non le nombre de crimes et délits qui ont réellement eu lieu sur le territoire où ils sont situés. De même, les données comptabilisées comprennent les infractions enregistrées une année donnée. Or, il est possible que certaines d'entre elles se soient déroulées l'année précédente, voire exceptionnellement plus tôt, et être prises en compte l'année donnée. A l'inverse, certains délits ayant eu lieu cette année pourront être enregistrés ultérieurement.

Représentation cartographique

Ce mémoire a pour vocation la constitution d'un zonier pour deux garanties dont la garantie vol. Afin de constituer ce dernier, nous allons nous intéresser plus particulièrement à la fréquence de sinistre car en pratique c'est sur la fréquence que la variable zone est la plus discriminante en matière de vol. Le portefeuille étudié est restreint à la France métropolitaine, nous avons dans un premier temps représenté l'ensemble des informations sur l'ensemble de la France métropolitaine avant de nous focaliser sur l'île de France. La figure de gauche représente le nombre de sinistres vol en France dans le portefeuille considéré. Celle-ci peut être comparée au nombres de cambriolages que nous avons représenté sur la figure de droite.



(a) Nombre de sinistres VOL

(b) Nombre de cambriolages

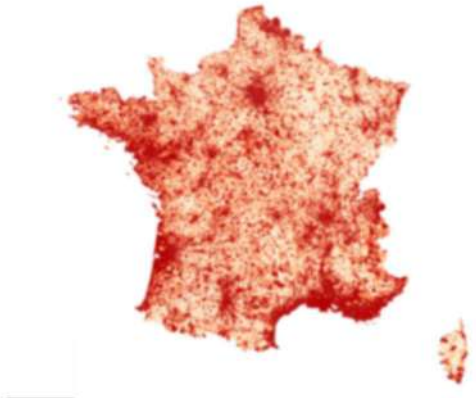
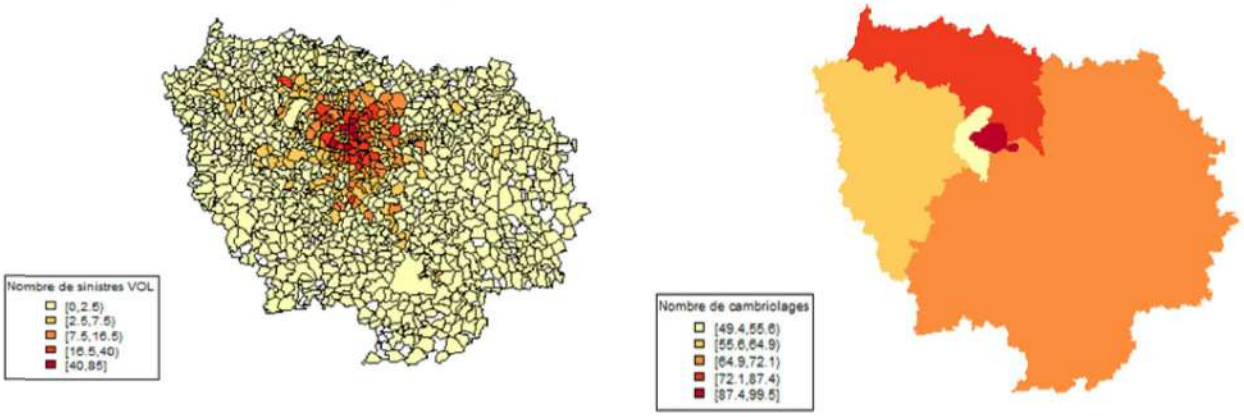


Figure 3.7: Répartition des contrats en France

Dans le portefeuille étudié, la garantie vol n'est souscrite qu'en option. Ces cartes nous permettent de nous rendre compte que certaines zones sont plus à risque que d'autres.

Pour rendre compte de l'hétérogénéité de la sinistralité sur le territoire et la nécessité de faire l'étude à une maille fine (code INSEE), nous nous focalisons à présent sur l'île-de-France.



(a) Nombre de sinistres VOL (b) Nombre de cambriolages

Figure 3.8: Représentation cartographiques en IDF

Nous représentons également la répartition des contrats souscrits en île de France.

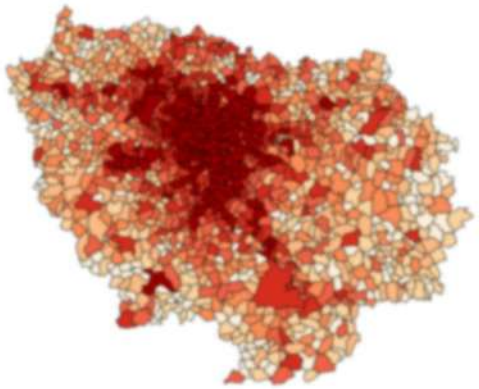


Figure 3.9: Répartition des contrats en IDF

3.2 Agrégation des bases

L'idée est ici de combiner l'ensemble de ces données pour former une base externe unique permettant d'englober l'ensemble des informations récupérées.

Une fois cette base créée, la prochaine étape consiste à l'agréger aux résidus des différents modèles de coût et de fréquence constitués en amont. En effet, l'idée est de pouvoir expliquer ces résidus à travers toute l'information récupérée dans la base de données externe finale et ainsi, inclure la dimension géographique à l'étude.

Chapitre 4

Construction du zonier

L'élaboration d'un zonier repose sur la création d'une unique variable chargée de quantifier le risque géographique d'une habitation. Il s'agit ici de rassembler l'ensemble de l'information extraite de l'*open data* en une seule et unique variable qui englobera le facteur risque géographique. Cette nouvelle donnée, appelée zonier, peut être construite de différentes manières. Ce chapitre propose de modéliser les résidus des modèles GLM bâtis dans le chapitre 2 à l'échelle de la commune. Pour cela, il est d'usage d'exploiter les variables qui constituent la base de données externe créée dans le chapitre 3. Ces variables externes constitueront les variables explicatives de la modélisation.

La construction du zonier repose donc sur différentes étapes synthétisées dans le graphique ci-dessous.

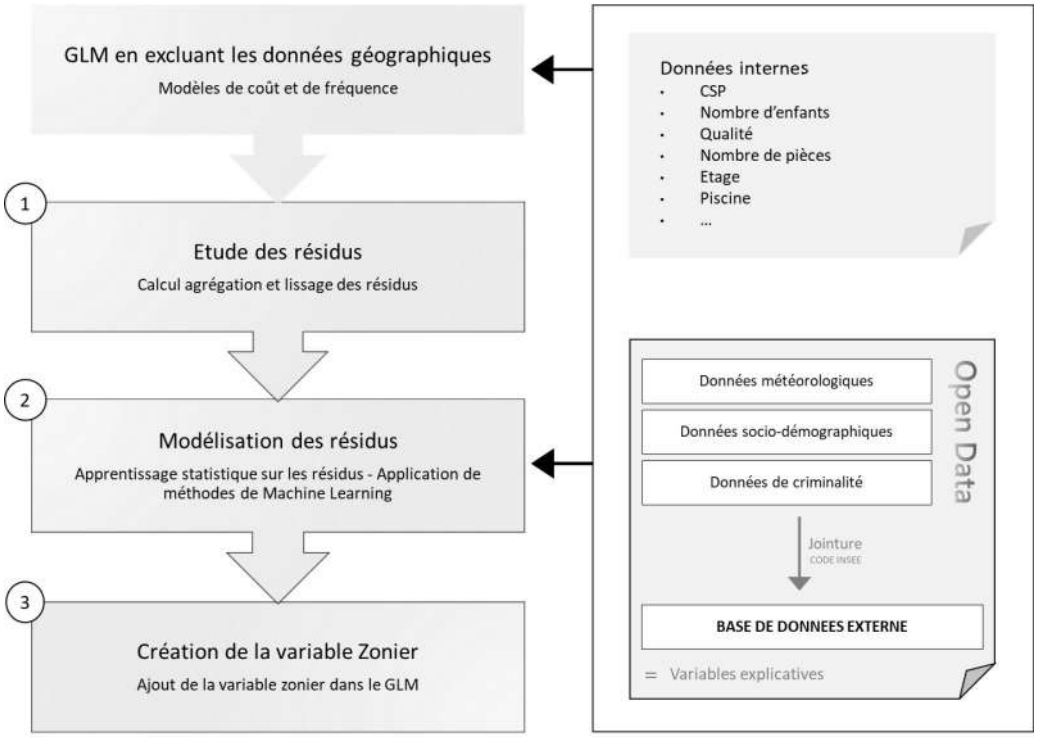


Figure 4.1: Processus de construction du Zonier

4.1 Étude autour des résidus

4.1.1 Définitions des résidus

Le calcul des résidus est une étape primordiale dans le processus de création d'un zonier. Ils permettent de mesurer l'écart entre l'observation y_i et la prédiction du modèle $\hat{y}_i$ avec i=1..p et p le nombre d'observations de l'échantillon étudié. Les résidus peuvent être définis de différentes manières et présentent des caractéristiques qui sont récapitulées dans le tableau ci-dessous.

Type	Définition	Caractéristiques
Résidus additifs	$R_i = y_i - \hat{y}_i$	Ce type de résidu est couramment utilisé et est simple à exprimer. Cependant, il ne correspond pas à la structure multiplicative des GLM utilisés dans le cadre de cette étude (utilisation de la fonction de lien logarithmique).
Résidus de Pearson	$R_{p_i} = \frac{y_i - \hat{y}_i}{\sqrt{\mathbb{V}[\hat{y}_i]}}$	Cette approche permet d'obtenir des résidus d'une même variance (homoscédastiques), pouvant être prédits par des modèles simples mais ne disposent pas de la structure multiplicative des GLM.
Résidus de déviance	$R_{d_i} = sign(y_i - \hat{y}_i)$ $\sqrt{2(\ln(\mathcal{L}_{estime}) - \ln(\mathcal{L}_{sature}))}$	Ce type de résidu est basé sur l'expression de la vraisemblance des données à prédire. Cependant, ils ne sont pas inversibles, en d'autres termes, ils permettent d'obtenir une valeur de résidu R_{d_i} à partir de y et $\hat{y}$ mais pas de retrouver explicitement la valeur de y à partir du résidu R_{d_i} et de $\hat{y}$.
Résidus d'Anscombe	Loi de Poisson: $R_{A_i} = \frac{3}{2}(y^{2/3} - \frac{(\hat{y}_i - 1/6)^{2/3}}{(\hat{y}_i{}^{1/6})})$ Loi Normal: $R_{A_i} = y_i - \hat{y}_i$	Ces résidus sont une approximation des résidus de déviance, dont la formule est inversible.

Figure 4.2: Les différents types de résidus

Au vu des propriétés des résidus présentés ci-dessus (BERAUD-SUDREAU G (2017)[2]), nous poursuivons l'étude avec les résidus d'Anscombe* qui semblent être les plus adaptés à notre cas.

*Les modèles de coûts utilisés ici sont des modèles lognormaux. Or, dans notre cas, nous étudions non pas le coût mais le logarithme du coût qui suit une loi Normale. C'est pourquoi nous définissons dans cette séction les résidus d'Anscombes associés à la loi Normale.

4.1.2 Agrégation des résidus et analyse des résultats

A l'issue de l'étape précédente qui consiste à calculer les résidus d'Anscombe de chaque modèle, nous disposons d'un résidu par lieu de risque. Il s'agit alors de consolider ces derniers au niveau de la commune, en retenant la moyenne pondérée par l'exposition. On considère alors l'estimateur du risque spatial résiduel suivant pour la commune i:

$$r_i^c = \frac{\sum_{j=1}^{n} e_j r_j^{ind}}{\sum_{j=1}^{n} e_j} \tag{4.1}$$

avec n le nombre d'observations dans la commune i, e_j l'exposition de l'observation j et r_j^{ind} le résidu associé à l'observation j (individu j). Nous obtenons ainsi une valeur de résidu par code INSEE en tenant compte de l'exposition.

Analyse variographique :
Une fois les résidus agrégés à la maille INSEE calculées, il est plausible que ces résidus ne représentent que du bruit. Afin de savoir s'il demeure un signal géographique dans ceux-ci, nous réalisons une analyse variographique* sur les résidus qui nous permettra de démontrer qu'ils contiennent bien une part restante de signal géographique.

Il s'agit d'un modèle de covariance ne dépendant que de la distance entre les observations. Le semivariogramme dépend donc uniquement du vecteur de translation h entre les points s et s+h qui contient de l'information sur la distance entre ces deux points. Il est défini de la manière suivante:

$$\gamma(h) = \frac{1}{2} Var[r(s) - r(s+h)] \tag{4.2}$$

avec r(s), la valeur du résidu d'anscombe pour le point s de coordonnées (x,y) et r(s+h) la valeur du résidu dans le voisinage à une distance h.

Dans notre étude, un semivariogramme est appliqué directement sur les résidus issus des modèles de fréquence et de coût. Si les résidus ne contenaient que du bruit, alors la tendance du semivariogramme serait relativement constante. En revanche, une croissance de la semivariance en fonction de la distance jusqu'à un certain palier nous permettrait de confirmer qu'il demeure un effet géographique dans les résidus non expliqués par les modèles L'existence d'une auto-corrélation spatiale entre les résidus serait alors une hypothèse acceptable.

*Rappel: Le (semi)variogramme est un outil géostatistique permettant de caractériser l'auto-corrélation spatiale de la variable étudiée (se référer au chapitre 3 pour plus d'informations)

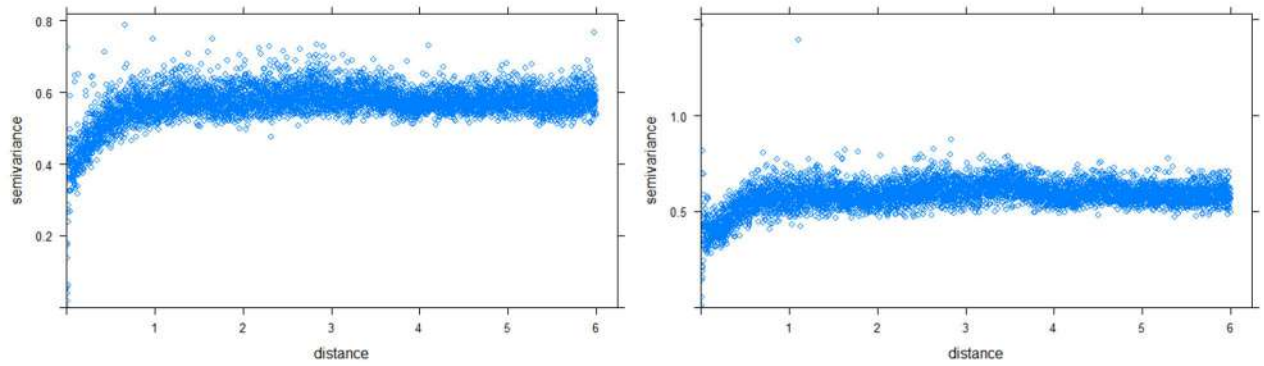


Figure 4.3: Semivariogramme empirique appliqué aux résidus des modèles de coût des garanties dégâts des eaux (à gauche) et vol (à droite)

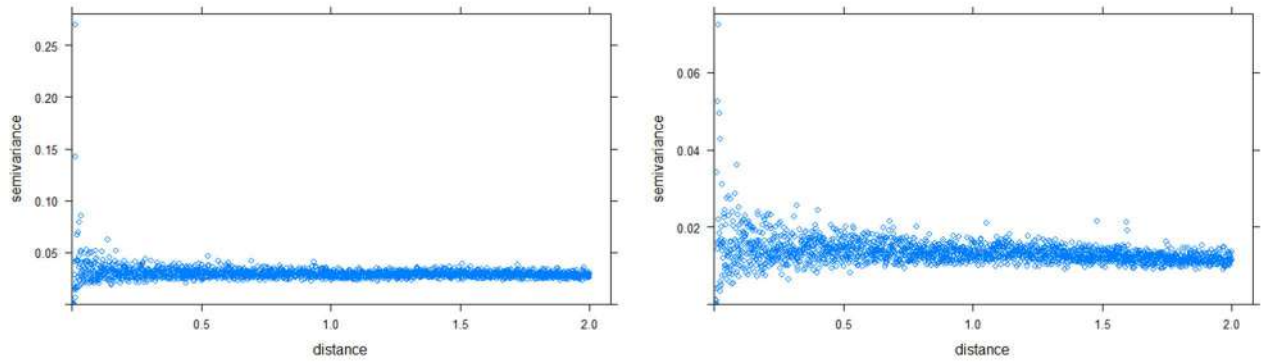


Figure 4.4: Semivariogramme empirique appliqué aux résidus des modèles de fréquence des garanties dégâts des eaux (à gauche) et vol (à droite)

Les figures ci-dessus nous permettent de visualiser les semivariogrammes empiriques appliqués aux résidus des modèles de coût et de fréquence des deux garanties considérées. Les figures laissent apparaître une croissance de la semivariance jusqu'à une certaine distance (environ 1 km) avant de se stabiliser à un certain palier. L'allure des semivariogrammes laisse donc présager qu'il existe une potentielle auto-corrélation géographique des résidus. Ces faibles valeurs (aux alentours de 1 km) concernant la distance à considérer pourraient être expliqué par le fait que nous travaillons sur des risques localisés.

4.1.3 Lissage des résidus

A ce stade, les résidus calculés ne tiennent pas compte de la structure spatiale des données. En effet, les résidus sont déterminés par code INSEE sans prendre en compte l'information des communes voisines ce qui peut conduire à des résidus très différents pour des communes pourtant très proches. De plus, ils sont déterminés à partir d'une quantité d'information inégale entre les communes. En effet, l'exposition totale des contrats sur le territoire français est très hétérogène selon les communes.

Ainsi, en raison du manque d'exposition dans certaines communes de la base d'étude, il est nécessaire de prendre davantage en compte les valeurs des communes voisines dans le calcul du risque.

En conséquence, les communes les plus proches auront plus de chance d'avoir un risque spatial similaire que les communes les plus éloignées. Cette étape permet de pondérer les résidus par l'exposition et ainsi enlever le bruit sur les codes INSEE qui sont faiblement exposés. Pour cela, nous optons de procéder à un lissage basé sur la théorie de la crédibilité. Il s'agit d'un cas d'application du modèle de Bulhman Straub.

Théorie de crédibilité

Nous nous baserons dans cette section sur les travaux de LOIRET C (2016)[4].
Soit r_i^* le résidu lissé de la commune i avec i=1..p et p le nombre total de communes.

$$r_i^* = Z(e_i)r_i + (1 - Z(e_i))\frac{\sum_j e_j r_j f(d_{ij})}{\sum_j e_j f(d_{ij})}, \forall i = 1..p \tag{4.3}$$

avec r_i le résidus de Anscombe moyen calculé pour la commune i; e_i l'exposition de la commune i; d_{ij} la distance en km entre la commune i et j et f la fonction de distance. On considère $f(d_{ij}) = \frac{1}{d_{ij}^m}$ avec m le paramètre de distance à déterminer. Une valeur importante du paramètre de distance pénalise les communes les plus lointaines de la commune i.

On considère $Z(e_i) = \frac{e_i}{e_i+a}$ avec a le paramètre de crédibilité à déterminer. Ainsi, plus la valeur de a est importante, plus la fonction de crédibilité Z pénalisera les zones faiblement exposées. En effet : $\lim_{a\to\infty} Z = 0$. Il est naturel de considérer que les communes les plus exposées possèdent une information plus fiable que celles qui sont moins exposées. Le paramètre de crédibilité a peut donc être vu comme un coefficient de tolérance quant à la fiabilité des informations de sinistralité des communes.

Choix des hyperparamètres:

Le choix des hyperparamètres dépend du portefeuille d'étude et plus spécifiquement de la taille de celui-ci. En effet, l'exposition par commune a un impact prépondérant dans le calcul des résidus lissés. Nous avons choisi différents couples d'hyperparamètres de lissage (a,m) pour les modèles de coûts et de fréquences. En ce qui concerne le paramètre de crédibilité a, sachant que plus la valeur de a est importante, plus la fonction de crédibilité pénalisera les zones faiblement exposées, nous avons opté pour une valeur de a=100. Cela revient à considérer un facteur de crédibilité de 50% pour une zone pour laquelle l'exposition est de 100.

D'autre part, le paramètre de distance m permet schématiquement de déterminer le rayon au sein duquel les voisins seront considérés pour le calcul des résidus lissés. Une valeur inférieure de m pour les modèles de coût permettra de prendre en compte plus de données (et donc agrandir le périmètre des voisins). Le choix final de m=1 pour les modèles de coûts et m=3 pour les modèles de fréquence se déduit des semi-variogrammes présentés ci-dessus. En effet, nous remarquons que ceux-ci présentent une autocorrélation entre les résidus jusqu'à une distance d'environ 1 km. Ainsi, il faudrait d'avantage prendre en compte les valeurs des résidus voisins se trouvant à une distance inférieure à celle-ci et pénaliser l'effet des valeurs des résidus plus lointains. Un facteur m=1 (respectivement 3) pour les résidus du modèle de coût (resp. fréquence) impliquerait de considérer qu'à partir de 2 km (resp. 1,25), nous considérerons une part inférieure à 50% de la donnée voisine dans le calcul.

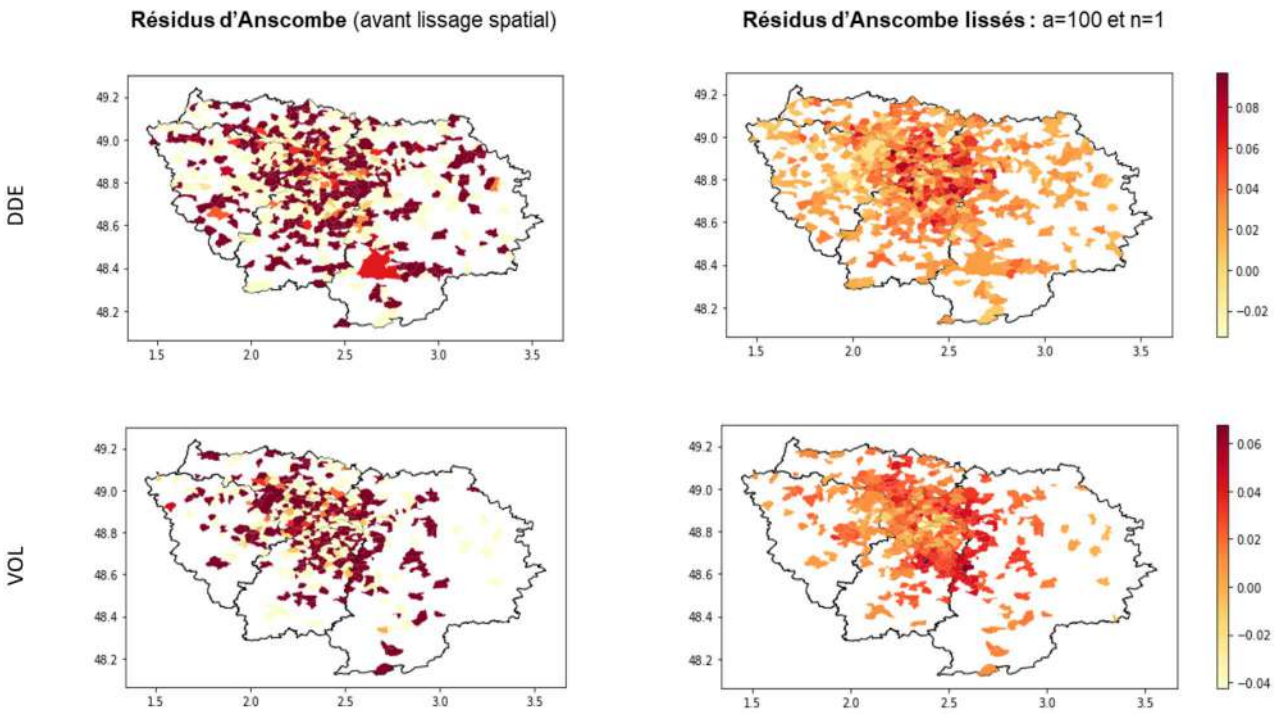


Figure 4.5: Lissage des résidus en île-de-France - modèle de coût

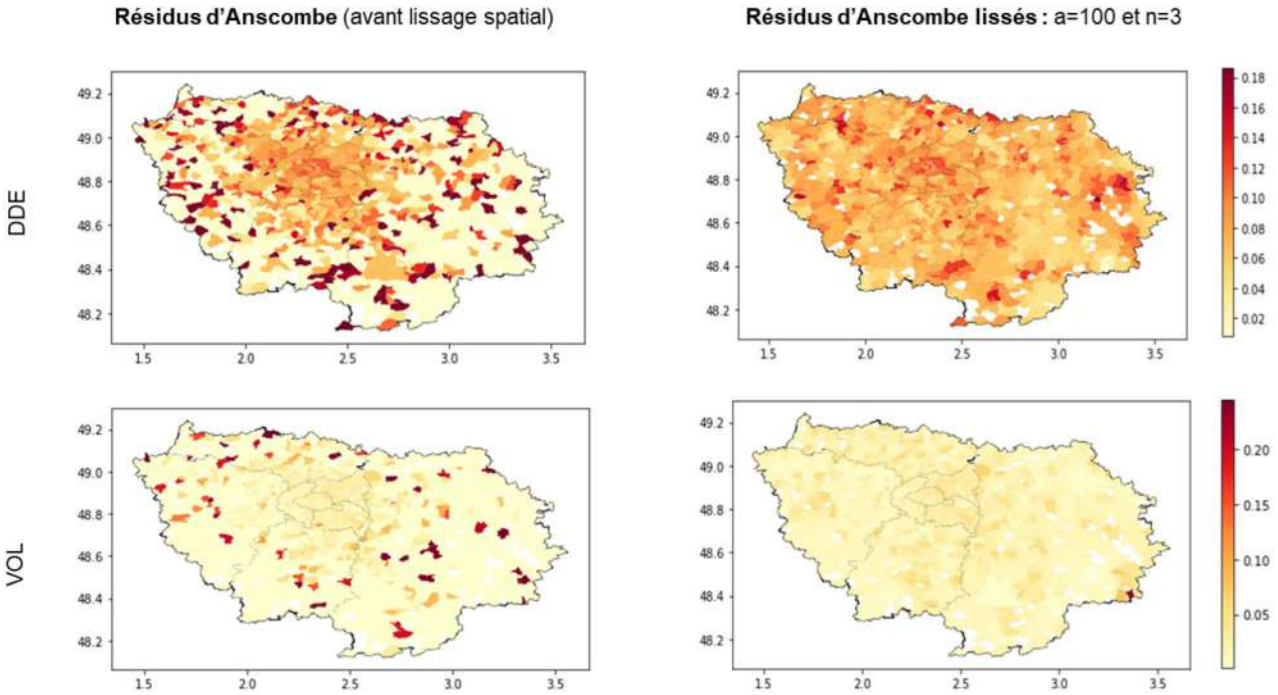
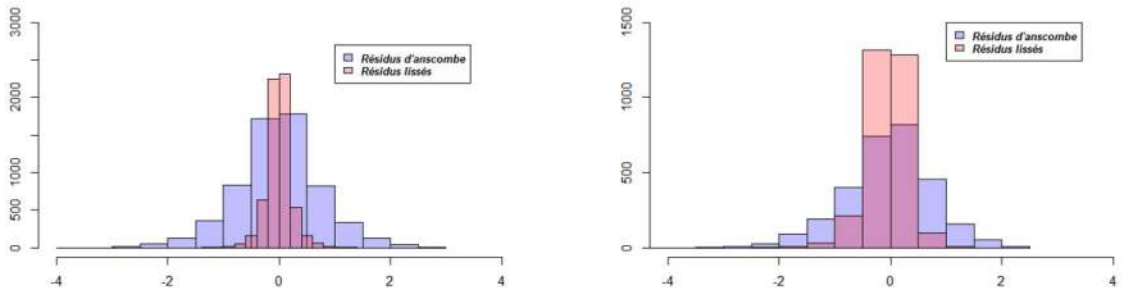


Figure 4.6: Lissage des résidus en île-de-France - modèle de fréquence

Plus la zone est rouge et foncée, plus le risque au sein de cette zone est important. Nous notons que les résidus d'Anscombe avant le lissage spatial ont été retraité afin d'obtenir des cartes comparables de même échelle. De manière additionnelle, nous avons tracé la distribution des résidus avant et après lissage pour les modèles de coûts.



(a) Garantie DDE (b) Garantie VOL

Figure 4.7: Distribution des résidus du modèle de coût

4.1.4 Jointure des données externes aux résidus

Afin de pouvoir expliquer les résidus par les variables externes extraites de l'*open data*, il est indispensable de joindre la base de données externe créée précédemment (voir chapitre 3) aux résidus lissés calculés à la maille code INSEE. Pour cela, il suffit d'effectuer une jointure par code INSEE pour obtenir une base comportant l'ensemble des données externes retraitées et un résidu associé à chaque commune. Une fois cette étape effectuée, nous nous attendons à obtenir des données manquantes, c'est à dire, des communes qui n'ont aucune valeur associée. Ces valeurs manquantes peuvent être liées à plusieurs causes (suppression ou modification de communes avec le temps etc.). Nous procédons donc à une analyse de l'ensemble des données manquantes à travers le package visdat sur R (se référer à l'annexe A.9 pour la visualisation des résultats). Le calcul des valeurs manquantes en terme d'exposition permet de rendre compte de la significativité des valeurs manquantes dans le portefeuille d'étude. En effet, les communes les plus exposées possèdent une information plus fiable que celles qui sont moins exposées. Ainsi, si les données manquantes correspondent à des communes peu exposées, alors leur impact ne sera pas conséquent dans notre étude. En revanche, si l'exposition est très significative, alors les résultats seront biaisés et peu fiables.

Dans notre cas, les pourcentages de manquants en terme d'exposition valent dans le cas des résidus des modèles de fréquences, 3,59% pour la garantie DDE et 3,49% pour la garantie vol. En ce qui concerne les résidus des modèles de coût, nous obtenons respectivement 1,94% et 2,16% de manquants pour les garanties DDE et vol.

4.2 Modélisation des résidus à l'aide des données géographiques

4.2.1 Méthodes de machine learning

Les arbres de décisions

Les arbres de décisions sont une classe d'algorithmes d'apprentissage se basant sur la représentation hiérarchique de la structure des données sous forme des séquences de décisions en vue de la prédiction d'un résultat ou d'une classe.

Ces derniers permettent en effet de représenter un ensemble de choix sous la forme d'un arbre. Le principal intérêt d'une telle structure en informatique décisionnelle est sa hiérarchisation des décisions que nous pouvons représenter de la manière suivante.

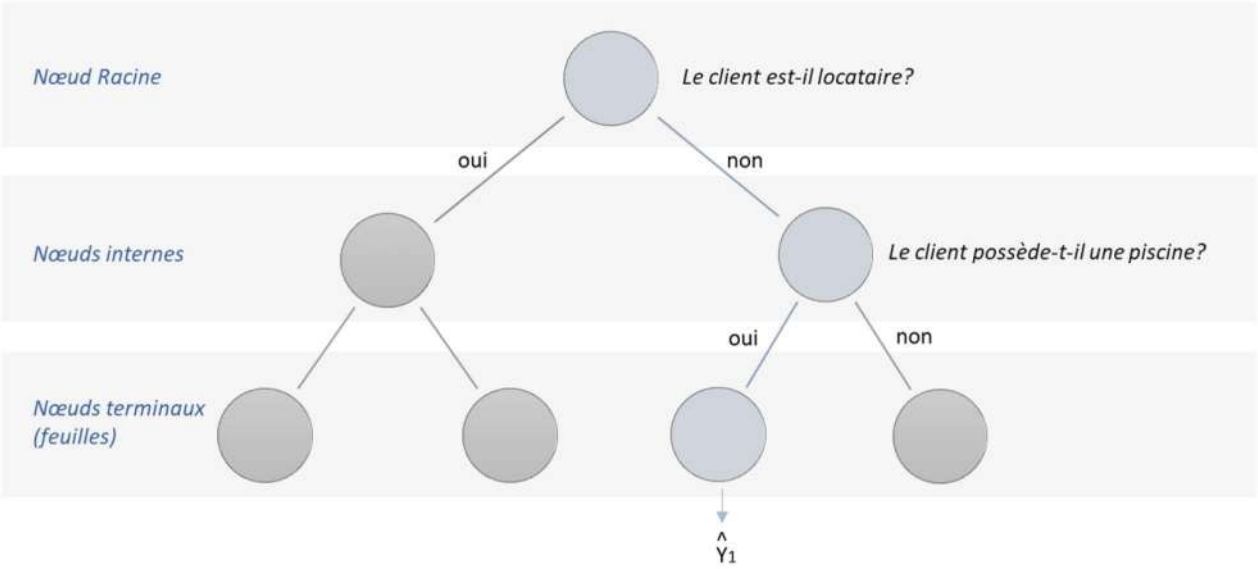


Figure 4.8: Fonctionnement d'un arbre de décision

L'ensemble des noeuds représentés peuvent se diviser en trois catégories : Les noeuds racine (l'accès à l'arbre se fait par ce noeud); les noeuds internes (les noeuds présentant des descendants) et les noeuds terminaux (aussi appelé feuilles) qui n'ont pas de descendant. Pour construire un arbre de décision, il suffit donc de prédéfinir les facteurs discriminants. Ceux-ci auront pour noeuds des tests portant sur ces facteurs (Le client est-il locataire ? possède-t-il une piscine?) et pour feuilles les choix finaux (les prédictions $\hat{y}$). L'intérêt d'un arbre étant sa simplicité, il est naturel de limiter le nombre de noeuds qu'il possède. Cela permet de trouver le plus rapidement possible le résultat de la prédiction. La construction de l'arbre se fait donc de manière progressive en cherchant toujours à trouver quel attribut permet de maximiser le gain d'information à chaque étape. C'est ainsi que nous introduisons la notion d'entropie. Celle-ci correspond dans notre cas au nombre de candidats appartenant aux différentes classes suite au test. Par exemple, un test qui divise une classe en deux classes non finales de même taille possède une entropie élevée car cela engendrerais une multitude de tests à conduire avant d'arriver aux résultats attendus. Au contraire, l'entropie est nulle lorsque nous arrivons à une feuille et tous les objets appartiennent à la même classe.

La construction d'un arbre se fait de manière récursive. Initialement, l'arbre est vide. Puis, l'algorithme sélectionne le test qui engendre l'entropie minimale, et en fait sa racine. Ensuite, il calcule pour chaque facteur discriminant (ou *feature*), le gain d'information que l'on obtiendrait avec ce choix puis réitère le processus pour choisir quel test se fera à chaque noeud. Enfin, si tous les exemples qu'il reste à traiter appartiennent à la même classe (on obtient donc une entropie nulle), ou s'il n'y a plus de test à effectuer, nous sommes arrivés à une feuille et l'algorithme se termine.

Les arbres de décision, malgré leur simplicité, présentent trois inconvénients majeurs, qui font qu'ils ne peuvent être utilisés tel quel :

- L'overfitting (il se produit pour de nombreuses raisons, notamment la présence de bruit et l'absence d'instances représentatives).
- L'erreur due au biais (il se produit lorsque trop de restrictions sont placées sur les fonctions cibles).
- L'erreur de variance (elle fait référence à la variation d'un résultat en fonction des modifications apportées à l'ensemble des données de départ).

Afin de modéliser les résidus des modèles obtenus, plusieurs possibilités d'algorithmes de *Machine Learning* s'offrent à nous. Nous nous focalisons dans la suite sur la méthodes des forêts aléatoires (*Random Forest*) qui, de part sa simplicité, offre une meilleure transparence sur l'utilisation des données d'entrainement.

Les forêts aléatoires

Les forêts aléatoires (ou *Random Forest*) figurent parmi les principales extensions des arbres de décisions et constituent un outil fondamental dans l'apprentissage automatique. Cette méthode peut etre utilisée pour des problèmes de regression (variable cible continue) comme pour des taches de classification (cas discret). Elle couvre donc une grande partie des problèmes de *Machine Learning*.

Le *Random Forest* est généralement considéré comme l'algorithme d'apprentissage le plus précis dans la mesure où il permet de pallier l'instabilité des arbres de décision.

L'idée général du *Random Forest* est de créer un grand nombre d'arbres de décisions indépendants de façon aléatoire, à partir de différents sous-ensembles de données de l'ensemble de données initial. En considérant différents sous-ensembles, le risque d'erreur est considérablement réduit et le problème de surapprentissage (ou *overfitting*) qui intervient lorsque l'arbre construit s'adapte trop à l'échantillon considéré est également évité. Plus l'on dispose d'arbres, plus la forêt sera fiable.

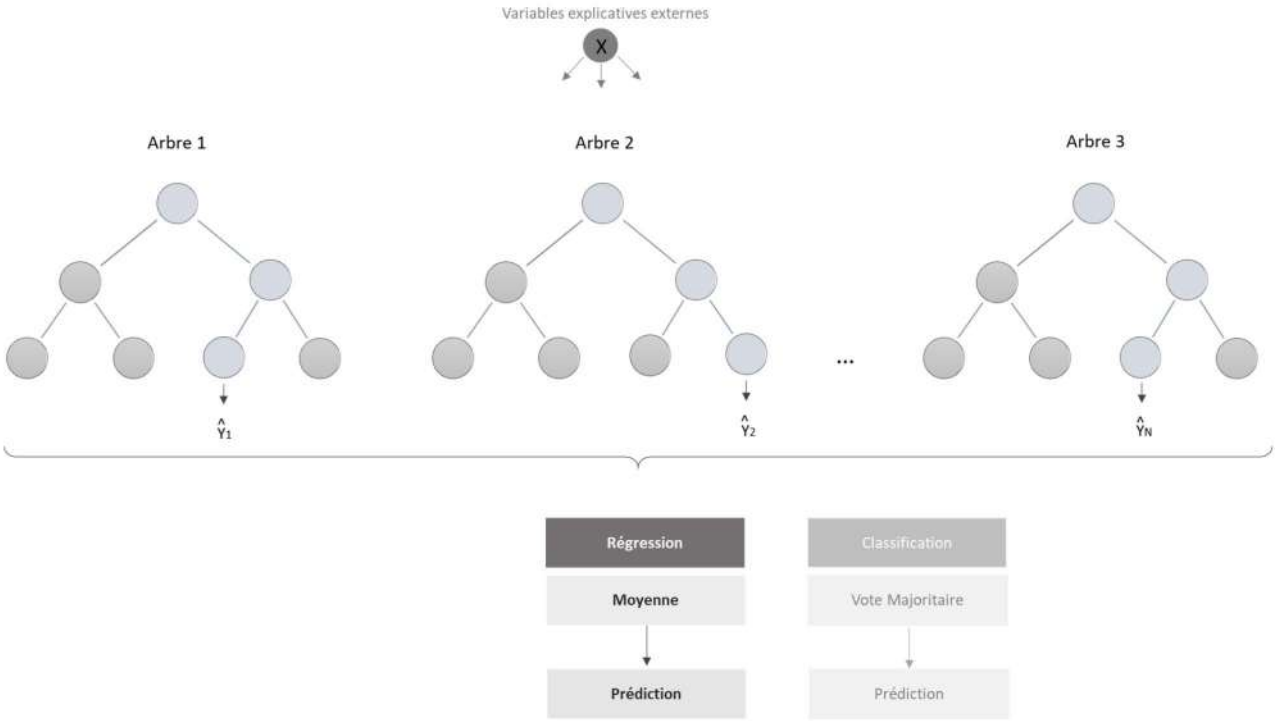


Figure 4.9: Schéma représentatif de l'agrégation des prédictions avec l'algorithme de random forest

Les forêts aléatoires peuvent être composées de plusieurs dizaines voire centaines d'arbres, le nombre d'arbre, comme les autres hyperparamètres, est ajustable par cross validation. En effet, lorsque nous abordons un problème d'apprentissage automatique, nous nous assurons de séparer nos données en une base d'apprentissage et une base de test. Dans la méthode de cross validation par K-Fold, l'idée est de diviser la base d'apprentissage définie en un nombre K de sous-ensembles, appelés *folds*. Cette technique consiste à entraîner et tester le modèle sur des échantillons du dataset de départ de la manière suivante :

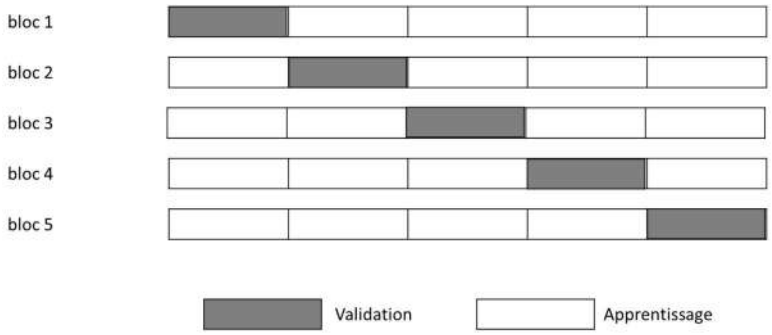


Figure 4.10: Schéma explicatif de la cross validation

Dans le cadre d'une régression en créant un grand nombre d'arbres indépendants et en prenant la moyenne des résultats obtenus, la stabilité et la précision de l'algorithme d'apprentissage automatique sont considérablement améliorés.

La méthode des forêts aléatoires est donc similaire au *bagging (Bootstrap Aggregating)* dans la manière d'agréger les modèles. Cependant la différence se situe dans la construction des arbres qui seront par la suite agrégés. En effet, cette méthode se distingue du *bagging* par le fait qu'avant la

division de chaque noeud, à la place de sélectionner la division optimale parmi les divisions possibles basées sur toutes les variables explicatives, on tire aléatoirement un certain nombre m de variables explicatives, et on considère les divisions possibles basées sur ce sous-ensemble.

L'algorithme associé à cette méthode est le suivant.

Algorithme Forêts Aléatoires

Entrées: $\mathbf{x}$ l'observation à prévoir, d_n l'échantillon, N le nombre d'arbres, m le nombre de variables candidates pour découper un noeud
Pour k=1..N:
1. Tirer un échantillon bootstrap dans d_n
2. Construire un arbre sur cet échantillon bootstrap, chaque coupure est sélectionnée en minimisant la fonction de coût sur un ensemble de m variables choisies au hasard parmi les p variables. On note $h_k(.)$ l'arbre construit.

Sortie:
L'estimateur $h(\mathbf{x}) = \frac{1}{N}\sum_{k=1}^{N} h_k(\mathbf{x})$.

L'algorithme permet de découper le dataset en plusieurs sous-ensembles aléatoirement constitués d'échantillons. Nous tirons au hasard dans la base d'apprentissage N échantillons avec remise où chaque échantillon contient n points. Pour chaque échantillon k, nous construisons un arbre selon un algorithme: à chaque fois qu'un noeud doit être séparé, nous tirons au hasard une partie des attributs (m parmi les p attributs) et nous choisissons le meilleur découpage parmi ce sous-ensemble. Il entraîne ensuite un modèle sur chaque sous-ensemble (les coupures sont choisies de manière à minimiser une fonction de coût particulière). A chaque étape, il cherche le facteur discriminant qui minimise la variance des noeuds fils en régression*. Les arbres sont ainsi construits jusqu'à atteindre une règle d'arrêt.

Les résultats des modèles sont ensuite combinés en calculant la moyenne des résultats obtenus† ce qui nous donne une prédiction finale fiable. De cette manière on construit un modèle robuste à partir de plusieurs modèles qui ne sont pas forcément aussi robustes.

Remarque: lorsque m diminue, la tendance est à se rapprocher d'un choix aléatoire des variables de découpe des arbres. Ainsi, si m diminue, la corrélation entre les arbres diminue également, ce qui entraîne une baisse de la variance de l'estimateur agrégé. A contrario, choisir les axes de découpe des arbres de manière (presque) aléatoire se traduit par une moins bonne qualité d'ajustement des arbres sur l'échantillon d'apprentissage, d'où une augmentation du biais pour chaque arbre ainsi que pour l'estimateur agrégé.

Pour des informations plus détaillées sur les forêts aléatoires, le lecteur est invité à se reporter au livre de HASTIE T, TIBSHIRANI R, FRIEDMAN J (2006) [1]

4.2.2 Modélisation des résidus et optimisation des hyperparamètres

Si la construction du zonier nécessite d'établir une prédiction des résidus des GLM, la méthode de modélisation de ceux-ci n'est pas figée. Notre choix final s'est porté sur l'algorithme du *Random Forest* pour modéliser les résidus des GLM. Le *Random Forest* présente certaines caractéristiques qui font de lui l'algorithme le plus adapté pour résoudre ce problème de régression. Il s'agit d'une technique

*Pour un problème de classification, c'est l'indice de Gini des noeuds fils qui devra être minimisé
†Pour un problème de classification, il s'agira ici d'un système de vote

facile à interpréter, stable, qui présente en général des résultats précis. En pratique, les résultats ont été générés grâce à des algorithmes disponibles dans la librairie *scikit-learn*.

Présentation de Azure

La plateforme Windows Azure constitue un élément fondamental de la stratégie Cloud de Microsoft. Elle permet permet de louer des services comme des machines virtuelles, de l'espace de stockage, des services réseaux etc. Parallèlement, Ray est un framework open-source qui fournit un moyen de modifier le code python existant pour tirer parti de l'exécution parallèle à distance. En d'autres termes, il permet de lancer les instruction en parallèle depuis python et d'assurer la communication entre les machines.

Les applications distribuées modernes, telles que celles utilisées pour entraîner des modèles de *deep learning*, doivent évoluer de manière dynamique pour répondre aux exigences de vitesse de calcul et réduire efficacement les coûts. Dans cette partie, nous expliquons la manière dont nous avons exploité de telles applications distribuées à l'aide de Ray sur Azure.

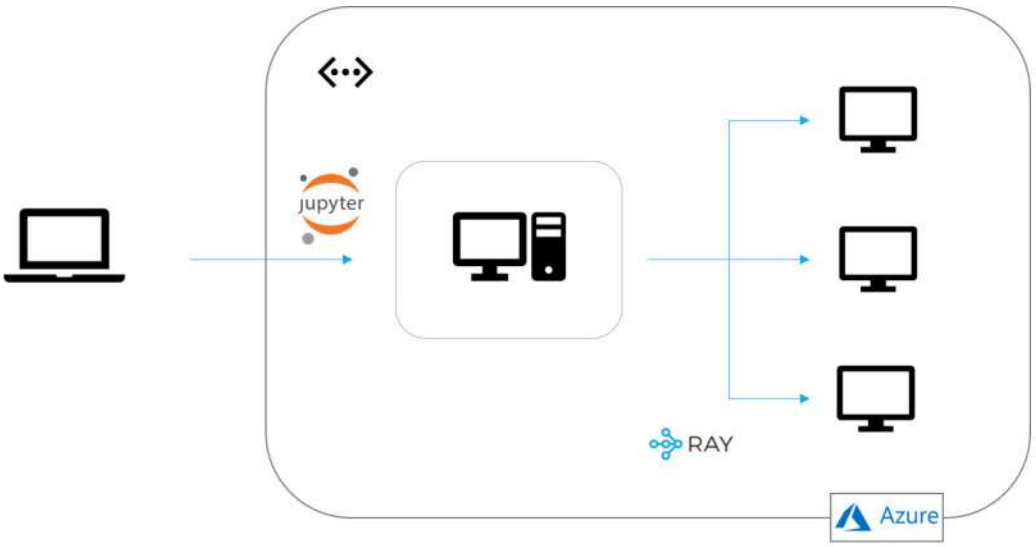


Figure 4.11: Fonctionnement de Ray sur Azure

Dans le schéma ci-dessus, la machine encadré représente le noeud principal, c'est sur cette machine que nous implémentons l'algorithme à travers le notebook jupyter. Celle-ci envoie les instructions à exécuter en parallèle et récupère ensuite les résultats. Certains packages basés sur Ray comme *Tune* offrent la possibilité d'accélérer le réglage d'hyperparamètres pour les modèles d'apprentissage automatique. C'est dans cette optique que nous l'utilisons ici.

Optimisation des hyperparamètres

En général, les modèles de *machine learning* doivent être calibrés pour donner les meilleurs résultats possibles. En d'autres termes, cela revient à ajuster les hyperparamètres pour optimiser les performances de notre modèle.

Les hyperparamètres optimaux sont généralement impossibles à déterminer à l'avance, il est donc indispensable de calibrer le modèle en s'appuyant davantage sur les résultats expérimentaux que sur la théorie. Ainsi, une étape d'optimisation des hyperparamètres a été réalisée à partir d'une méthode d'optimisation qui consiste à essayer de nombreuses combinaisons différentes et évaluer les performances de chaque modèle. En effet, le choix des hyperparamètres peut s'avérer très coûteux en terme de temps si les tests sont fait un par un. C'est pourquoi, la méthode d'optimisation, appelée le *Grid-*

Search, va nous permettre de tester une série de paramètres et de comparer les performances pour en déduire le meilleur paramétrage.

La méthode du *Grid Search* consiste donc à déterminer un ensemble de valeurs à tester puis à créer un modèle pour chaque combinaison de paramètres de la manière suivante:

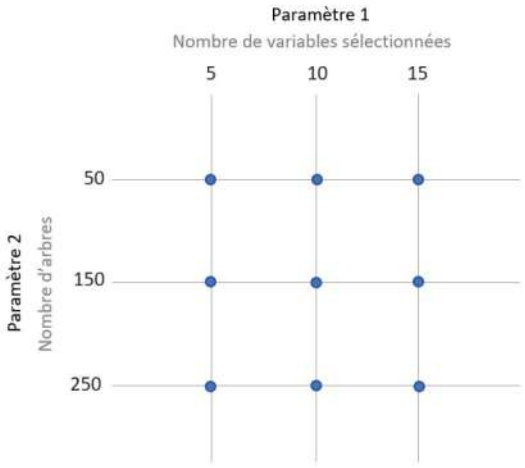


Figure 4.12: Principe du GridSearch

L'étape suivante consiste à tester les 9 modèles obtenus sur le dataset par validation croisée* (le k-fold). Nous testons donc tous les couples de paramètres et nous construisons un indicateur de performance sur la base d'apprentissage et sur la base de validation. Il s'agit ici d'un problème de régression, la comparaison des performances se fera donc à travers une étude menée sur l'évolution du score en fonction du choix des hyperparamètres. Les scores obtenus correspondent aux scores de précision (ou *accuracy score*) et permettent d'évaluer les modèles. Il s'agit de la précision moyenne sur les données d'apprentissage (train) et de test (test).

Les forêts aléatoires prennent en entrée plusieurs hyperparamètres : dans le cadre de ce mémoire, nous nous focalisons sur la calibration des hyperparamètres suivants: le nombre d'arbres dans la forêt, le nombre de variables choisies à chaque noeud, la profondeur maximale des arbres, la taille minimale (nombre d'observations) que doit avoir un noeud parent pour être éventuellement subdivisé et la taille minimale (nombre d'observations) que doit avoir un noeud fils après une subdivision pour être conservé.

Les hyperparamètres optimaux seront différents selon les modèles et les garanties étudiés. En effet, pour les zoniers de coût moyen, nous ne considérons que les contrats sinistrés, il y a donc beaucoup moins de données à prendre en compte que dans les modèles de fréquence où toutes les polices sont considérées.

Ainsi, la méthode Grid Search nous permet d'optimiser les paramètre de chacun des modèles. Cependant, cette méthode présente des limites dans la mesure où les différents paramètres à tester sont défini à l'avance par l'utilisateur. Afin de choisir les valeurs à tester des différents hyperparamètres considérés, nous utilisons les courbes de validation† qui permettent de visualiser pour chacun des hyperparamètres les scores d'entraînement et les scores de validation selon les valeurs considérés.

*Rappel: La validation croisée est une méthode qui consiste à partitionner la base de données afin de constituer des bases d'apprentissages et de test.

†Aussi appellée les *validation curves*, ces courbes permettent de comparer les scores d'entraînement et les scores de validation pour différentes valeurs d'un paramètre donné.

Par exemple, pour le paramètre *Nombre de variables sélectionnées*, nous obtenons les courbes de validation suivantes pour chacun des modèles:

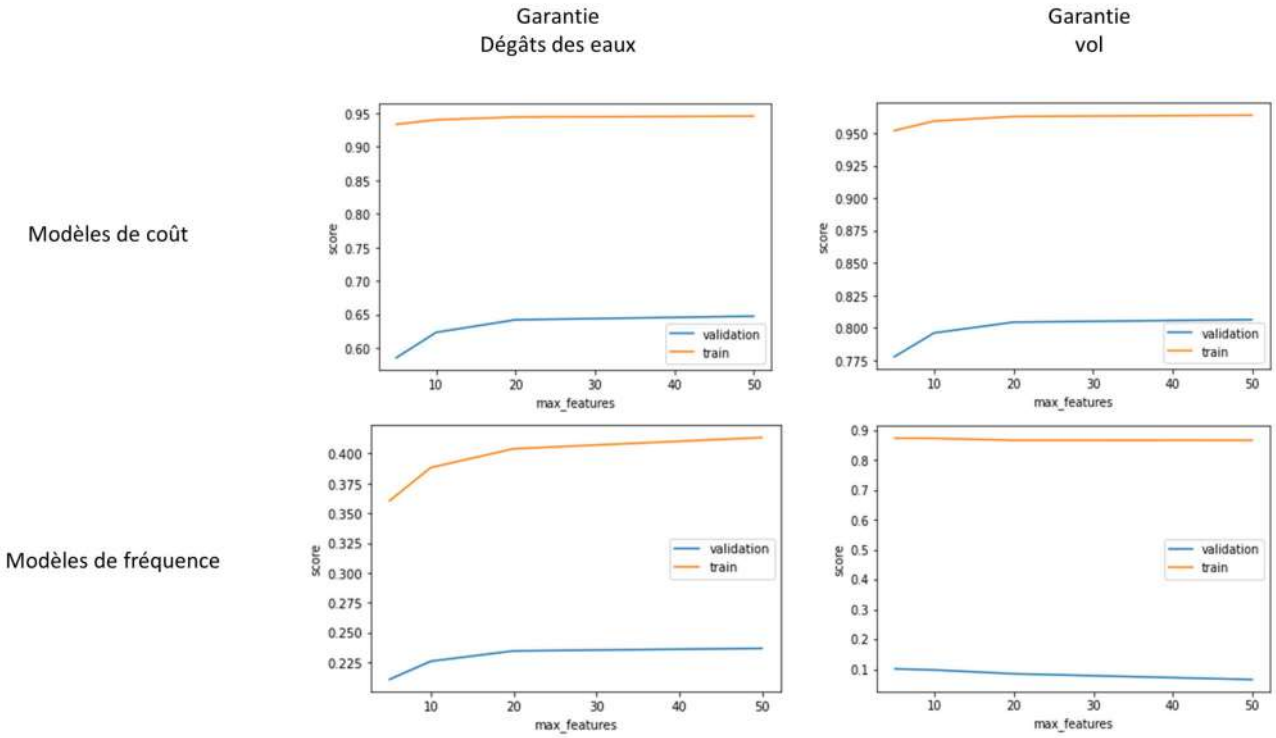


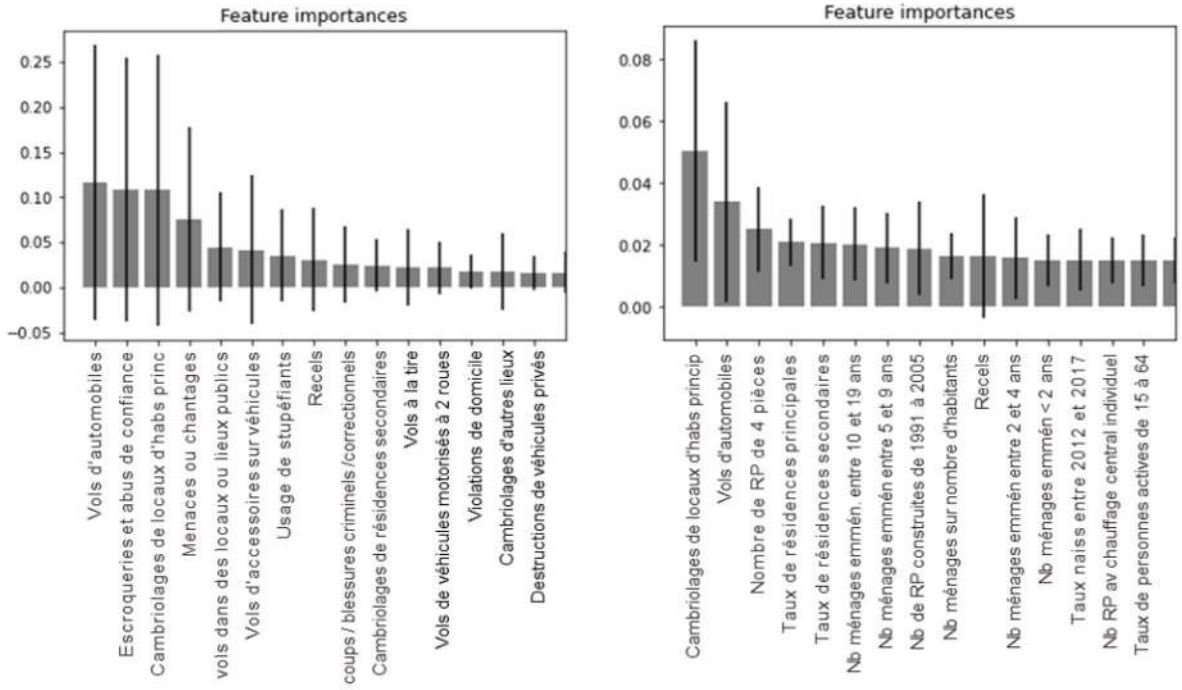
Figure 4.13: Validation curves de l'hyperparamètre Nombre de variables seléctionnées

Les *learning curves* obtenues sont données à titre illustratif car elles répondent à une optimisation unidimensionelle. En effet, l'optimisation de l'hyperparamètre *max_features* suppose que les autres hyperparamètres sont fixés. Néanmoins, l'approche *grid search* est bien un programme numérique d'optimisation multidimensionel. Les courbes unidimensionelle ont uniquement été analysées pour définir des intervalles de recherche. Par exemple, la figure correspondant au modèle de coût de la garantie dégâts des eaux (en haut à gauche) nous permet de constater qu'au delà de 20 variables, le score de la base de test se stabilise. Nous conserverons donc pour les tests des valeurs aux alentours de 15 variables. De la même manière, nous optons pour différentes valeurs à tester des 5 hyperparamètres. Ceux-ci constitueront les combinaisons des *grid search*. Cependant, cette approche empirique contient certaines limites dont la maille de la grille qui pourrait être à réduire et l'extremum qui pourrait n'être que locale. Une fois les paramètres optimaux des *Random Forest* obtenus, il convient de calibrer le modèle sur la base d'apprentissage. Le modèle de prédiction peut alors être appliqué à n'importe quel point disposant des variables explicatives nécessaires.

Importance des variables

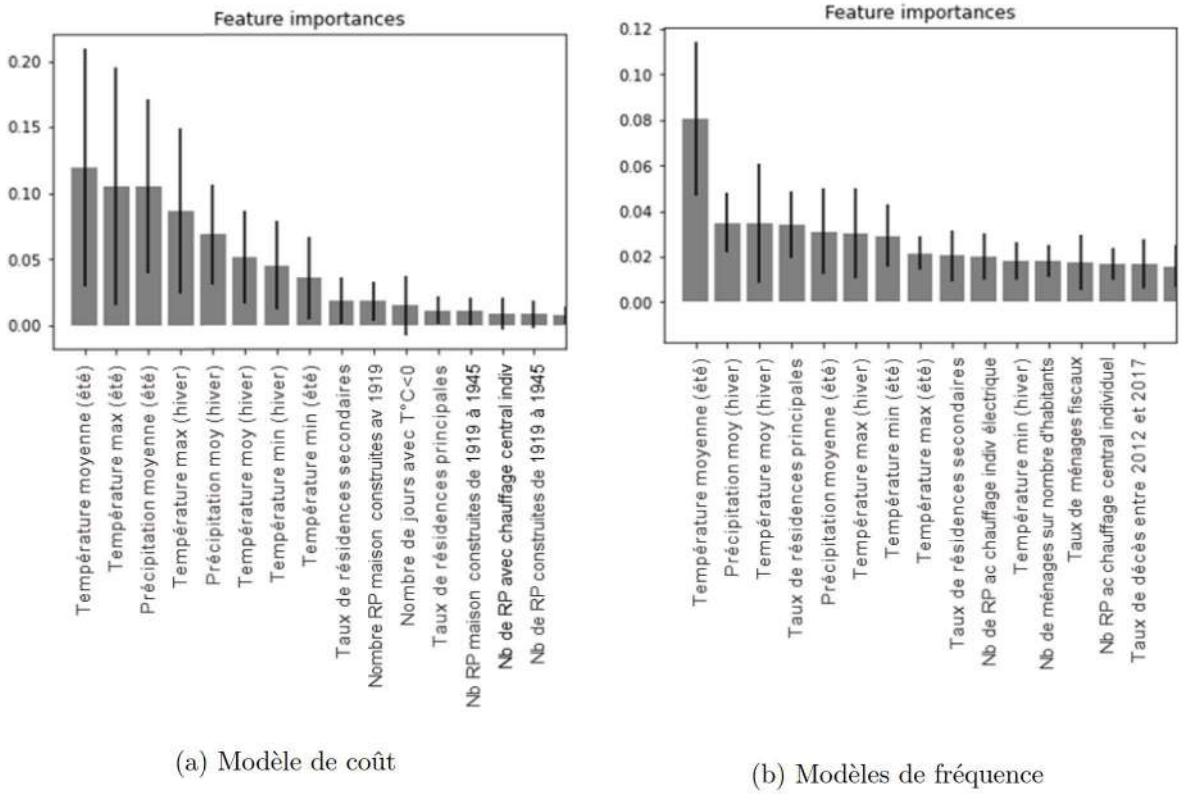
Les attributs du meilleur modèle peuvent être évalués pour voir leur impact dans la construction des arbres (mesure de Gini).
*Définition Gini : Le changement dans l'impureté (ou gain d'information) dans chaque noeud cumulé sur tous les arbres de la forêt**.

*se réferer à la section 2.3.7 pour plus d'informations sur le calcul de l'indice de Gini



(a) Modèle de coût (b) Modèle de fréquence

Figure 4.14: Importance des attributs pour la garantie vol



(a) Modèle de coût (b) Modèles de fréquence

Figure 4.15: Importance des attributs pour la garantie DDE

Les graphiques ci-dessus ont été obtenus via la librairie *Sklearn* qui propose en particulier une implémentation de forêt aléatoire avec une feature importance intégrée au modèle, appelée *mean decrease in impurity (MDI)*, aussi appelée *Gini importance*. Il s'agit de calculer la réduction de l'impureté des noeuds (pondérée par la proportion de données d'entraînement passant par ce noeud) moyennée sur tous les arbres de l'ensemble. Ce calcul se rapproche de la *gain importance* du XGBOOST.

Les variables qui ressortent en premier pour le modèle de coût de la garantie vol sont : vols d'automobiles, escroqueries et abus de confiance, cambriolages de locaux d'habitations principales et menaces ou chantages dans un autre but. Pour le modèle de fréquence ce sont les cambriolages de locaux d'habitations principales, les vols d'automobiles et le taux de résidences principales de 4 pièces qui sont les plus importantes.

En ce qui concerne la garantie dégâts des eaux, les variables météorologiques sont celles qui sont les plus importantes (température en été et en hiver et précipitations) pour les deux modèles. Nous retrouvons également le taux de résidences secondaires, le taux de résidences principales de type maison construites avant 1919 et le taux de résidences principales avec chauffage central individuel.

Dans les deux cas, les variables les plus discriminantes ne sont pas incohérentes avec les a priori que nous pouvons avoir sur le facteur géographique de la sinistralité. Par exemple, la vétusté des immeubles ainsi que les basses températures en hiver pouvant causer le gel des canalisations et provoquer des fuites et des inondations lors du dégel sont des variables que nous estimons importantes à retrouver dans nos modèles. Il en est de même pour les vols, escroqueries et cambriolages pour la garantie vol.

4.3 Classification des résidus lissés

Cette étape consiste à réaliser des regroupements de communes homogènes en terme de risque. Pour cela, nous ferons appel aux techniques de classification automatique. Les classes obtenues définiront les modalités du critère zone. Nous obtiendrons ainsi un zonier, c'est-à-dire une partition du territoire en zone de risques homogènes.

4.3.1 Méthodes de classification

La construction du critère zone ne peut s'arrêter au lissage des résidus et nécessite une étape supplémentaire. En effet, il existe plusieurs dizaines de milliers de communes et il ne peut être envisagé d'inclure un tel facteur comme variable d'un GLM. Il convient alors de regrouper en classes les effets géographiques. L'idée est de faire en sorte que deux communes appartenant à la même classe soient les plus semblables possibles et deux communes appartenant à des classes différentes soient les plus différentes possibles.

Pour ce faire, nous avons le choix d'utiliser différentes techniques de classification automatique (ou *clustering*). Il s'agit de techniques d'apprentissage non supervisé reposant sur des algorithmes qui repèrent les similarités dans les données pour pouvoir ensuite les structurer. Par exemple, ils permettent d'étudier les similarités entre individus, ce qui rend possible leur division en différents groupes. Cette partition des individus est appelée *clustering*.

Il existe plusieurs familles d'algorithmes de classification, par exemple :

- Les méthodes d'agrégation autour de centres mobiles : algorithmes conduisant directement à des partitions.

– Les méthodes hiérarchiques : algorithmes fournissant une hiérarchie de partitions.

Nous présentons par la suite la méthode des *k-means*, qui est la méthode centroïde la plus classique et que nous utiliserons dans la suite, ainsi que la CAH (ou Classification Ascendante Hierarchique), qui appartient à la famille des méthodes hiérarchiques.

Principe de la méthode des k-means

La méthode des *k-means* est un outil de classification classique qui permet de répartir un ensemble de données en classes homogènes. Cette méthode permet de regrouper les objets en k clusters distincts et repose sur la minimisation de la somme des distances euclidiennes au carré entre chaque objet et le centroïde (le point central) de son cluster.

Comme pour toute les méthodes de classification, il est nécessaire, dans un premier temps, de choisir comment mesurer la similarité entre deux individus. Pour cela, il convient de choisir une fonction distance, ici, nous prenons par exemple la distance euclidienne*, en imaginant les n observations comme des points de l'espace des réels en dimension p.

Nous cherchons ensuite à attribuer un cluster à chaque point, de façon à ce que la somme des distances euclidiennes au carré entre chaque point et le centroïde de son cluster, soit minimale.

L'algorithme associé à la méthode des *k-means* est donc un algorithme itératif qui a pour objectif de trouver des groupes en faisant en sorte de minimiser l'inertie intra-classe : l'algorithme minimise donc la fonction critère égale à la somme des distances entre chaque observations et le centroïde du cluster.

Soient $X_1, ..., X_n$ l'ensemble fini des n observations (ou objets) représentés par d caractéristiques,
$P_k = \{j | X^{(j)} \in k\}$ un ensemble non vide d'observations représentant un cluster k,
$m_k = \frac{1}{card(P_k)} \sum_{j \in P_k} X^{(j)}$ le barycentre d'un cluster k,
$C_k = \sum_{j \in P_k} d(X^{(j)}, m_k)$, la cohérence d'un cluster k (avec d la distance définie précédemment).
$\min_p \sum_{i=1}^{k} \sum_{j \in P_k} d(X^{(j)}, m_k)$ la fonction critère à minimiser et
$P = P_1, ..., P_k$, une partition des données à k clusters, appelée un *clustering*.

L'algorithme fonctionne de la manière suivante:

Algorithme k-means

Entrées: k le nombre de clusters
Principe:
1. Attribuer un cluster à chacune des n observations de façon aléatoire
2. Calculer le centroïde m_k de chaque cluster
3. Calculer la distance euclidienne pour chaque observation avec les centroïde de chacun des clusters
4. Attribuer à l'observation le cluster le plus proche de lui
5. Calculer la somme de la variabilité intra-cluster
6. Réitérer les étapes 2 à 5, jusqu'à atteindre un équilibre (la convergence est atteinte lorsqu'il n'y a plus aucun changement de clusters, c'est-à-dire quand le centroïde reste immobile ou lorsqu'il y a stabilisation de la somme de la variabilité intra-classe)
Sortie: Les k clusters finaux

*Dans un espace à deux dimensions la distance euclidienne entre deux points peut être estimée de la manière suivante: $d = \sqrt{(x_B - x_A)^2 + (y_B - y_A)^2}$

Les classes finales obtenues à l'aide de l'algorithme dépendent des objets choisis pour l'initialisation. C'est pourquoi certains algorithmes de k-means itèrent plusieurs fois le processus avec des initialisations différentes, dans le but de garder la partition qui minimise le plus la variance intra-classe (somme des distances entre les individus d'une même classe).

Nous nous sommes basés ici sur les travaux de PARIENTE J (2017) [5].

Choix du nombre de clusters: L'algorithme n'est pas capable de déterminer le nombre de classes optimal, c'est pourquoi il laisse un paramètre libre : le nombre de clusters k. Différents critères permettent d'estimer le nombre de clusters optimal en minimisant la distance intra-classes et en maximisant la distance inter-classes. Pour cela, trois méthodes sont généralement employées :

- Elbow method (la méthode du coude) : basée sur la minimisation de la somme des carrés des écarts à l'intérieur des clusters.
- Average silhouette method : basée sur la maximisation du paramètre appelé *average silhouette*.
- Gap statistic method : basée sur la comparaison de la variation totale intra-cluster pour différentes valeurs de k avec leurs valeurs attendues sous une distribution de référence nulle des données.

Principe de la CAH

Cette méthode consiste à créer, à chaque étape, une partition obtenue en agrégeant 2 à 2 les formes les plus proches. Dans ce cas, l'algorithme ne fournit pas une partition en k clusters mais une hiérarchie de partitions sous la forme d'arbres appelés dendrogrammes. L'avantage de ces arbres est qu'ils sont capable de donner une idée du nombre de clusters existant effectivement dans la base de départ.

Les méthodes de clustering hiérarchiques ascendantes forment pas à pas des connexions entre individus et utilisent une matrice de distances entre individus pour trouver le regroupement le plus proche d'un autre. Nous ne nous attarderons pas ici sur cette méthode. Néanmoins, pour plus d'informations, le lecteur est invité à se reporter aux travaux de FERRIER C (2016)[3].

4.3.2 Choix et application de la méthode de clustering aux résidus

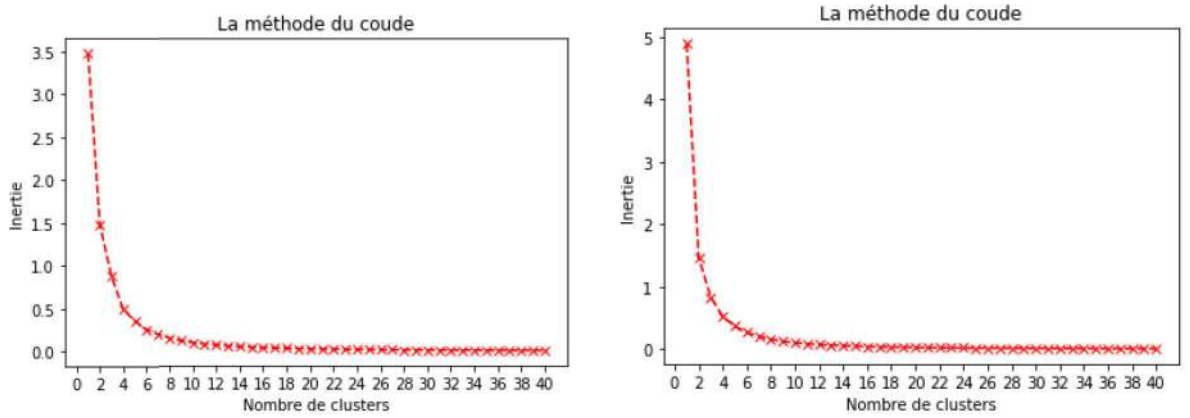
La classification *k-means* présente notamment l'avantage d'être efficace, facile à interpréter et d'avoir un faible temps de calcul. En effet, comparée à l'utilisation d'autres méthodes de classification, une technique de classification *k-means* est rapide et efficace en termes de coût de calcul. Par ailleurs, l'algorithme *k-means* s'adapte aux divers changements des données : un objet affecté à une classe au cours d'une itération peut éventuellement changer de classe à l'itération suivante. Or, cela est impossible dans la cas de la classification ascendante hiérarchique où une affectation est irréversible. De plus, dans la méthode des *k-means*, en multipliant les points de départ et les répétitions, il est possible d'explorer plusieurs solutions possibles. En outre, cet algorithme converge en général très rapidement ce qui fait de lui la technique de partitionnement la mieux adaptée aux vastes recueils de données. Ainsi, du fait de sa flexibilité, sa simplicité, mais surtout de la volumétrie des bases de données que nous disposons, la méthode des *k-means* semble être la plus appropriée à nos jeux de données. Cependant, l'inconvénient de cette méthode est qu'elle ne permet pas d'avoir le nombre cohérent de classes, ni de visualiser la proximité entre les classes ou les objets.

Malgré le fait que nous n'ayons pas d'idée a priori sur le nombre de classes à effectuer, la volumétrie

de la base de données à classifier ne nous permet pas d'appliquer la méthode de Classification Ascendante Hiérarchique. En effet, la complexité algorithmique de la cette méthode (de l'ordre du cubique) devient prohibitive à l'instant où la taille du jeu de donnée excède le millier d'observations. Or, dans notre cas, nous possédons des bases de dimension allant jusqu'à environ 36 000 observations (les lignes et colonnes correspondent respectivement au nombre de communes de France et au nombre de variables explicatives de la base de donnée). Par ailleurs les résultats d'une CAH ne sont pas satisfaisants au-delà de quelques centaines d'observations car les derniers niveaux de regroupement cumulent les erreurs d'écart à l'optimum des partitions obtenues à tous les niveaux précédents. En conséquence, nous mettons en oeuvre dans la suite la méthode des *k-means*. Afin de choisir le nombre optimal k de clusters, nous utilisons ici la méthode du coude.

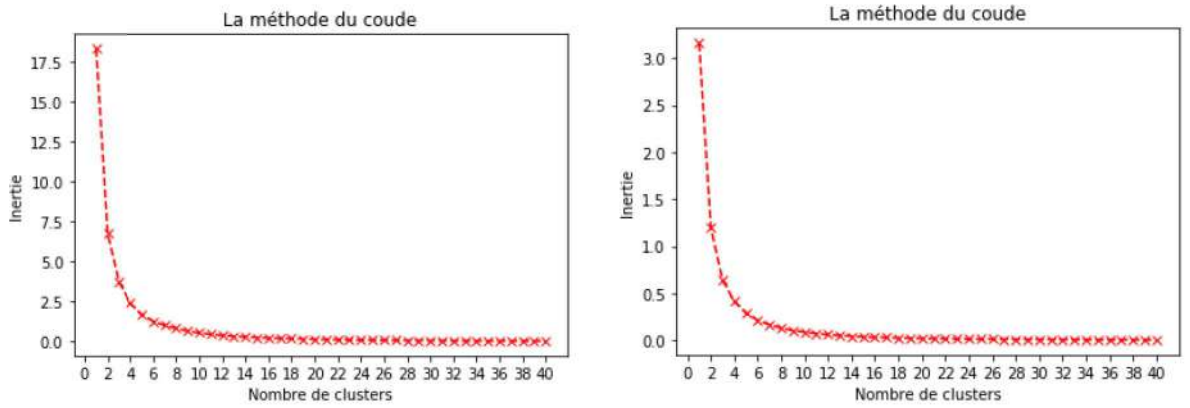
Définition: La méthode du coude est une méthode qui fonctionne en implémentant l'algorithme des k-means avec différentes valeurs de k. Pour chacune d'entre elles, nous notons l'inertie intra-classe obtenue. Lorsque le nombre k de clusters augmente, l'inertie intra-classe diminue. Cependant, il est possible de déterminer une valeur optimale du nombre de clusters en observant la valeur de k au-delà de laquelle la diminution est plus faible.

En appliquant cette méthode à nos jeux de données, nous obtenons pour chacun des modèles les résultats suivants.



(a) Choix du nombre de clusters - Garantie DDE (b) Choix du nombre de clusters - Garantie VOL

Figure 4.16: Application de la méthode du coude - Modèles de coût



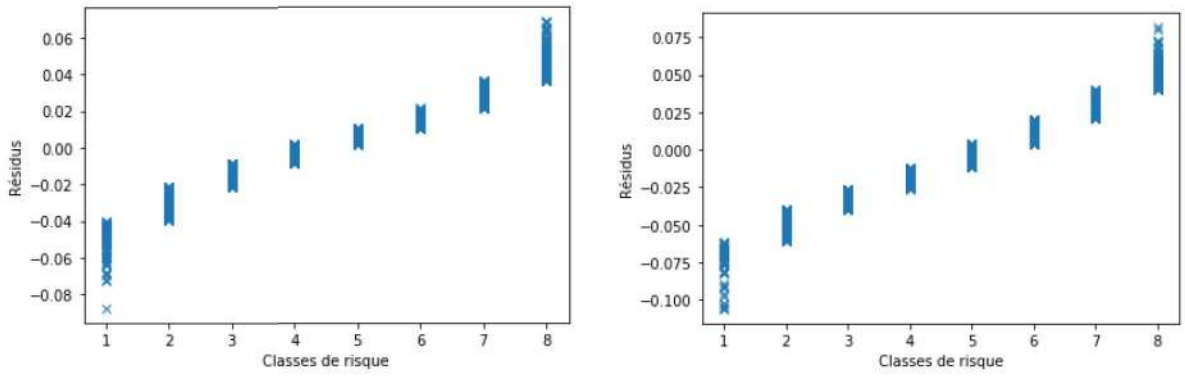
(a) Choix du nombre de clusters - Garantie DDE (b) Choix du nombre de clusters - Garantie VOL

Figure 4.17: Application de la méthode du coude - Modèles de fréquence

Le nombre de classe optimal déduit de ces graphiques est aux alentours de 8 zones pour l'ensemble des modèles. D'un point de vue métier, cette approche purement statistique pourrait être complété en segmentant ou en regroupant certaines classes.

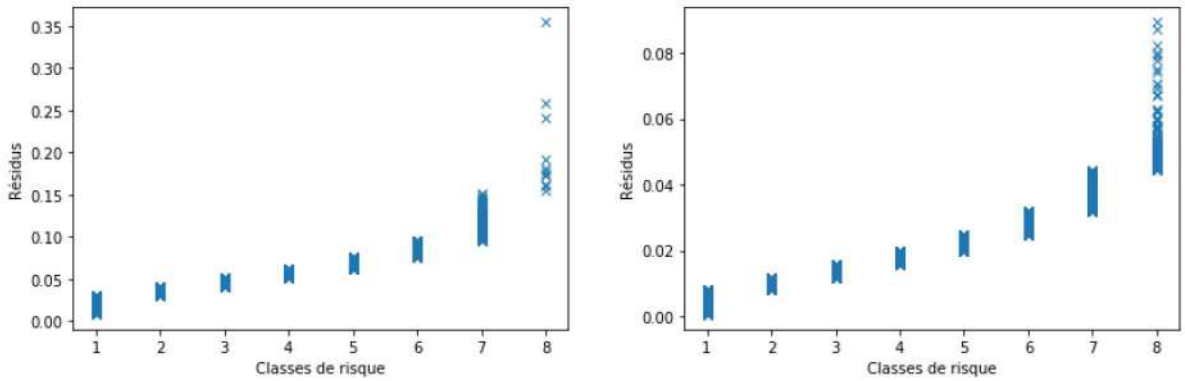
Nous poursuivons donc l'étude en choisissant de faire la séparation en 8 classes de risques. Dans les figures ci-dessous, nous présentons les valeurs que prennent les résidus par cluster pour les garanties dégâts des eaux et vol.

Remarque : Les classes les plus risquées sont celles dont le nombre associé est le plus élevé (La classe numéro 1 est la moins risquée et la classe numéro 8 est la plus risquée).



(a) Classes de risques - Garantie DDE (b) Classes de risques - Garantie VOL

Figure 4.18: Classes de risques - Modèles de coût



(a) Classes de risques - Garantie DDE (b) Classes de risques - Garantie VOL

Figure 4.19: Classes de risques - Modèles de fréquence

Nous poursuivons l'étude avec un nombre de classes fixés à 8. Cependant, il serait intéressant de tester d'autres nombres de classes de risques (par exemple 10, 12 ou 14) afin de comparer les résultats. Par manque de temps, cette approche n'a pas été traitée ici et pourrait faire l'objet de travaux ultérieurs.

4.4 Intégration de la variable zonier dans la base GLM de départ

L'intégration des variables spatiales au sein d'un modèle tarifaire constitue un enjeu majeur de la tarification. Cette section vise à mesurer l'apport des zoniers que nous venons de construire dans le

modèle de tarification.

4.4.1 Modélisation avec les facteurs géographiques

L'intégration du zonier nécessite une étape de jointure des données internes à la variable *zones de risques* pour chacun des quatre modèles obtenus :
Modèle 1: Dans le cas du modèle de coût de la garantie DDE, cette étape entraîne des valeurs manquantes qui représentent 1,94% en terme d'exposition dans la base d'étude. Ces valeurs manquantes correspondent à des codes INSEE de la base interne qui n'avaient pas de variables externes associées. Ces codes INSEE se retrouvent donc sans zone de risque associée.
Modèle 2: La jointure dans le cas du modèle de coût de la garantie vol entraîne 2.16% de manquants en terme d'exposition.
Modèle 3: Pour le modèle de fréquence de la garantie DDE, nous obtenons 3.59% de manquants en terme d'exposition.
Modèle 4: Pour le modèle de fréquence de la garantie VOL, nous obtenons 2.56% de manquants en terme d'exposition.

Au vu des faibles pourcentages de valeurs manquantes obtenues, nous faisons le choix d'attribuer aux codes INSEE ne disposant pas de zone de risque, la valeur de la zone de risque englobant le plus de communes. Par exemple, dans le cas de la garantie DDE et la garantie VOL, pour le modèle de coût, la zone de risque regroupant le plus grand nombre de communes correspond à la zone 4 et pour le modèle de fréquence, c'est la zone 2. Toutefois, cette approche constitue une limite forte qui nécessiterait d'être mise en avant dans l'optique de l'implémentation pratique du tarif car cela pourrait engendrer des sujets de rentabilité (par exemple dans le cas de développement commercial dans certaines zones aujourd'hui peu exposées). Nous avons illustré ci-dessous le nombre de communes par zone de risque pour l'ensemble des modèles.

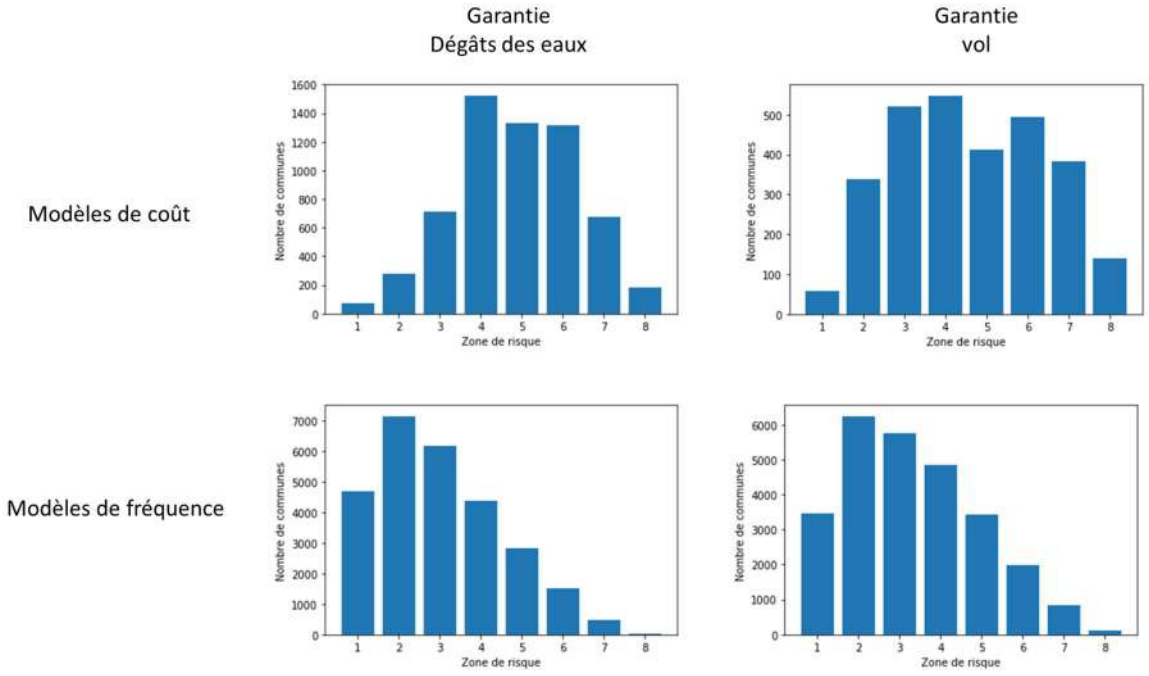
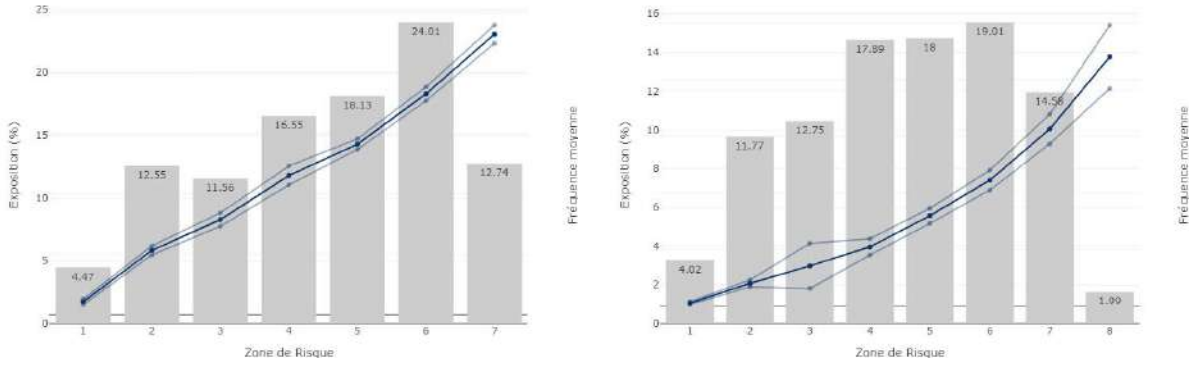


Figure 4.20: Nombre de communes par zone de risque

Nous remarquons sur la figure ci-dessus que l'effectif des communes appartenant à la zone de

risque 1 n'est pas conséquente devant les autres classes. Ainsi, afin d'obtenir de meilleurs résultats dans les GLM, nous choisissons (en consultation avec les experts du portefeuille) de regrouper les deux premières zones de risque. Cette caractéristique appuie également l'idée mentionnée ci-dessus d'éventuellement tester différents nombre de classes (8,10, 12 et 14 classes) afin de comparer les GLM obtenus et choisir le zonier optimal.

Pour conclure l'étape de construction des zoniers, nous avons représenté la fréquence de sinistres par zone de risque pour chacune des deux garanties ainsi que les cartes nous permettant d'identifier les zones à risque sur l'ensemble du territoire Français.

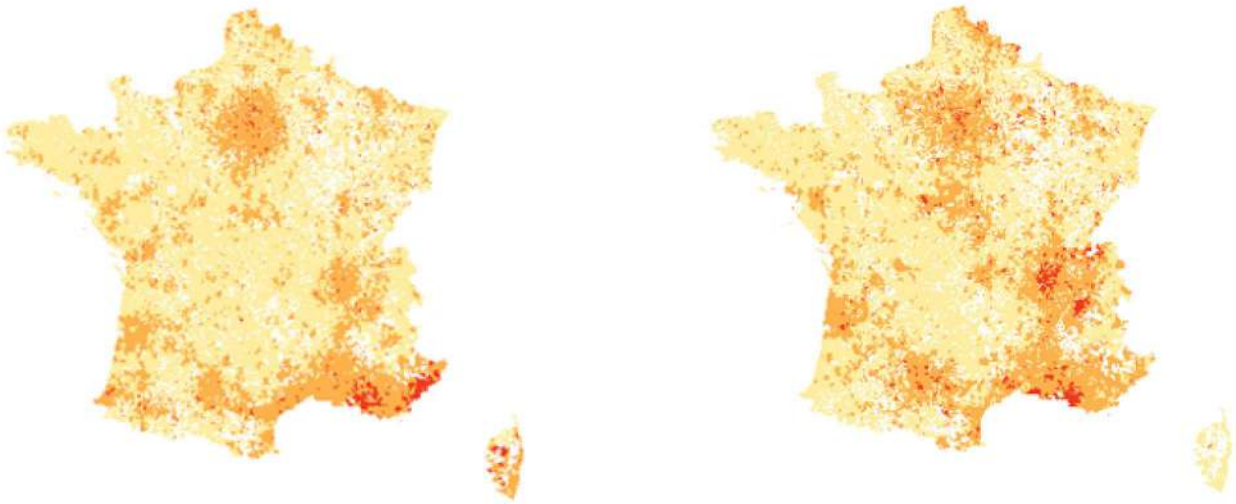


(a) Garantie DDE (b) Garantie VOL

Figure 4.21: Fréquence de sinistres par zone de risque - Zoniers fréquence

Nous remarquons que plus la zone est risquée, plus la fréquence de sinistre est élevée ce qui nous permet de valider les zoniers obtenus.

Pour finir, nous avons cartographié les zoniers obtenus ce qui nous a permis d'en extraire les zones les plus à risques. Nous représentons par exemple ici les zoniers de fréquence obtenus pour chacune des deux garanties.



(a) Zonier - garantie DDE (b) Zonier - garantie VOL

Figure 4.22: Cartographie des Zoniers

4.5 Interprétation des résultats

4.5.1 Comparaison des indicateurs avec et sans données géographiques

Les résidus modélisés puis discrétisés ont permis la création de quatre zoniers :

- Un zonier coût et un zonier fréquence pour la garantie vol construit à l'aide des données de criminalité, de la base de logement de l'INSEE et des données socio-démographiques de la base de comparateur de territoire de l'INSEE.
- Un zonier coût et un zonier fréquence pour la garantie dégâts des eaux construit à l'aide des données météorologiques, de la base de logement de l'INSEE et des données socio-démographiques de la base de comparateur de territoire de l'INSEE.

Ces zoniers sont ensuite introduits en tant que nouvelles variables explicatives dans les modélisations GLM du coût et de la fréquence des sinistres pour les deux garanties considérées. Cette étape nécessite certains retraitements que nous avons explicité dans la partie précédente (retraitement des valeurs manquantes).

Dans le cas des quatre modèles crées, l'ensemble des modalités de la nouvelle variable introduite ont des petites p-valeurs (inférieure à 5%) ce qui permet d'affirmer que le zonier implémenté est statistiquement significatif pour la modélisation du coût et de la fréquence des sinistres pour les deux garanties.

Afin de comparer les différents modèles obtenus et pour évaluer leur qualité, nous avons synthétisé dans le tableau ci-dessous les valeurs des indicateurs statistiques pour chacun des modèles.

	VOL		DDE	
	AIC	Déviance	AIC	Déviance
GLM coût sans zonier	20 361	6 061	79 668	22 638
GLM coût avec zonier	20 173	5 916	79 282	22 354
GLM fréquence sans zonier	96 792	82 041	356 347	288 090
GLM fréquence avec zonier	93 032	78 272	349 705	281 440

Figure 4.23: Impact de l'introduction des zoniers comme nouvelles variables explicatives sur la modélisation de la fréquence et du coût des sinistres.

Nous remarquons que l'AIC et la déviance sont nettement améliorés dans les GLM comprenant les zoniers. En ce qui concerne les modèles de coûts, nous obtenons une amélioration de la deviance de 2,4% ainsi qu'une amélioration de 0,9% de l'AIC pour la garantie vol. Pour la garantie dde, l'AIC est amélioré de 0,5% et la déviance de 1,3%. Si l'on considère à présent les modèles de fréquence, nous obtenons une amélioration de la déviance et de l'AIC respectivement de 4,6% et 3,9% pour la garanties vol. Nous voyons l'AIC s'améliorer de 1,9% et la déviance de 2,3% pour la garantie dde. Ces chiffres prouvent l'importance de tenir compte du risque géographique dans la tarification.

Afin de quantifier la plus-value de l'intégration du zonier dans le GLM, nous avons comparé les deux modèles suivants pour la garantie dégâts des eaux:

- Modèle 1: Modèle GLM incluant les variables internes ainsi que l'ensemble des variables externes météorologiques. (Ajout de 10 variables)
- Modèle 2: Modèle GLM incluant les variables internes ainsi que le zonier crée. (Ajout d'une unique variable)

Nous avons rassemblé dans le tableau ci-dessous les indicateurs de performance des modèles obtenus.

	AIC	Déviance
GLM sans zonier	79 668	22 638
GLM avec variables externes (modèle 1)	79 655	22 616
GLM avec zonier (modèle 2)	79 282	22 354

Figure 4.24: Comparaison des différents modèles pour la garantie DDE

4.5.2 Analyse des résultats

L'intégration du zonier dans les GLM permet non seulement une meilleure explication de la fréquence et le coût des sinistres mais aussi l'apport d'interprétabilité. En effet, comme indiqué dans le premier chapitre, l'interprétation des résultats des GLM est indispensable pour un assureur cherchant à expliquer le tarif octroyé à un assuré. Ici, l'environnement géographique est capté grâce à une seule variable simple de compréhension et d'utilisation. Nous pouvons déterminer de manière extrêmement limpide le degré de risque d'une commune alors que l'ajout de plusieurs données géographiques dans un modèle ne le permet pas aussi facilement. L'ensemble des valeurs sur les figures présentées dans cette section ont été masquées par soucis de confidentialité (coût, fréquence, coefficients estimés etc.). Ces résultats sont basés sur un individu de référence correspondant au profil le plus exposé au risque.

Modalités concernant le logement	Modalités concernant l'assuré
Résidence principale	Locataire
Appartement	CSP : 5 (employés)
3 pièces tarifées	Pas d'enfant
Avec franchise	
Pas de piscine, ni véranda, ni dépendance, ni garage	
Le capital mobilier : *capi_mob* 025	
Zone de risque: 4	

Figure 4.25: Profil de référence des modèles de coût

Les résultats des GLM sont les suivants* :

*Les résultats des GLM obtenus incluant la variable *Zone de Risque* ont nécessité un pré retraitement consistant à

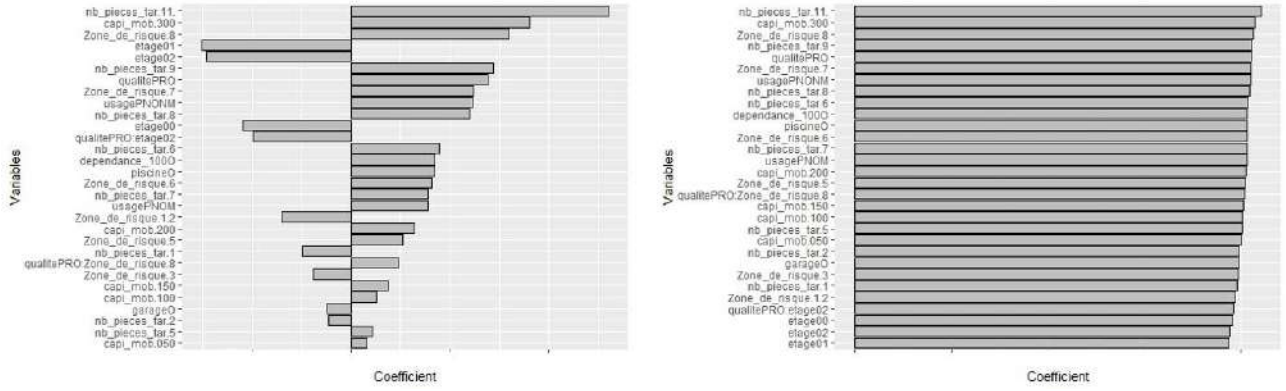


Figure 4.26: Coefficients estimés - modèle de coût de la garantie DDE

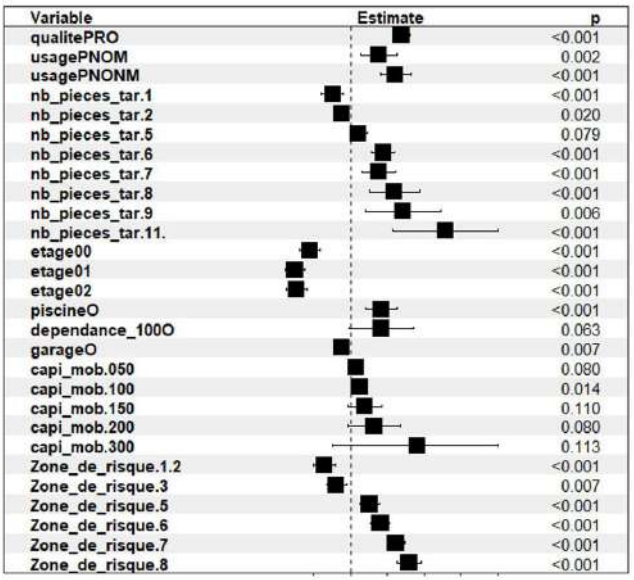


Figure 4.27: Coefficients estimés détaillés - modèle de coût de la garantie DDE

regrouper les zones 1 et 2 en une seule zone afin de garantie l'obtention de résultats cohérents et fiables. Ainsi, nous nous retrouvons avec 7 classes de risques.

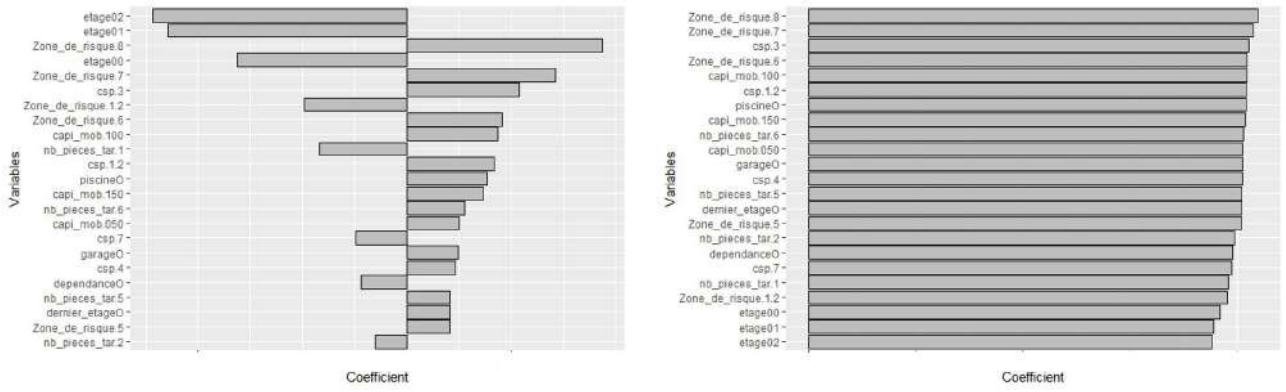


Figure 4.28: Coefficients estimés - modèle de coût de la garantie VOL

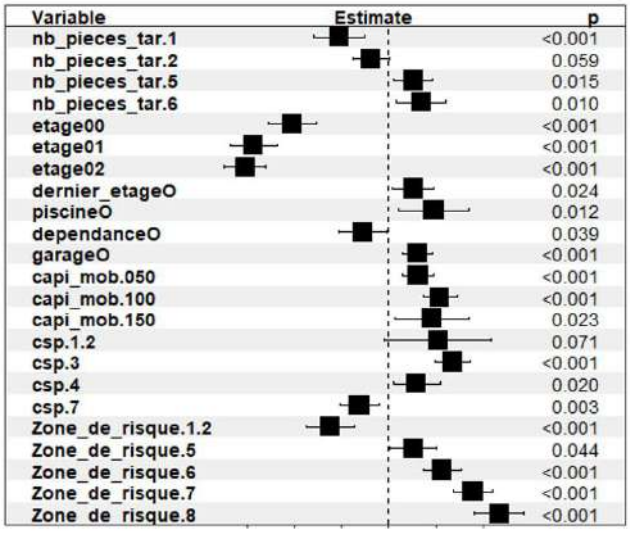


Figure 4.29: Coefficients estimés détaillés - modèle de coût de la garantie VOL

En suivant la même approche explicitée dans le chapitre 2 (section *Interprétation des résultats*), nous pouvons clairement déduire des graphiques ci-dessus que plus le logement assuré est situé dans une zone risquée, plus le coût moyen des sinistres augmentera.

De même, le profil de référence pour les modèles de fréquence est le suivant :

Modalités concernant le logement	Modalités concernant l'assuré
Résidence principale	Locataire
Appartement	CSP : 5 (employés)
3 pièces tarifées	Pas d'enfant
Avec franchise	
Pas de piscine, ni véranda, ni dépendance, ni garage	
Le capital mobilier : *capi_mob* 025	
Zone de risque: 2	

Figure 4.30: Profil de référence des modèles de fréquences

En considérant cette référence, nous pouvons à présent présenter les résultats obtenus:

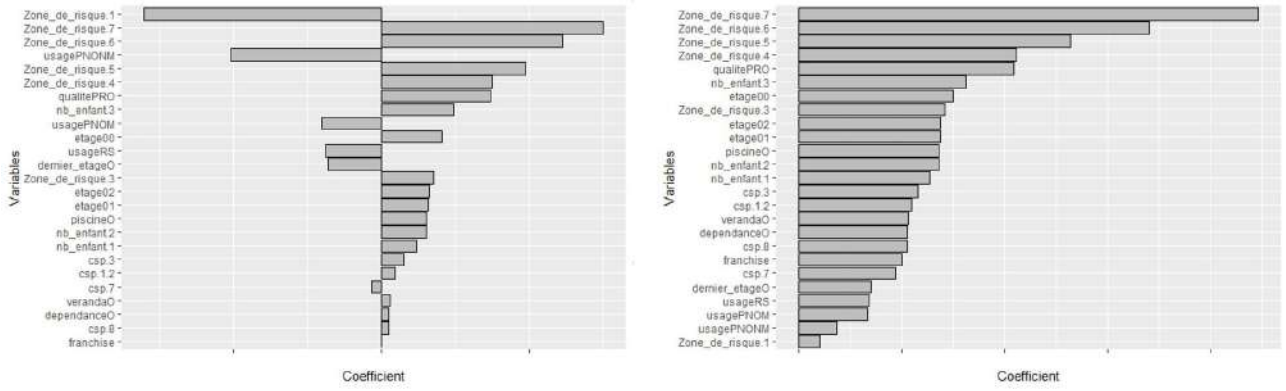


Figure 4.31: Coefficients estimés - modèle de fréquence de la garantie DDE

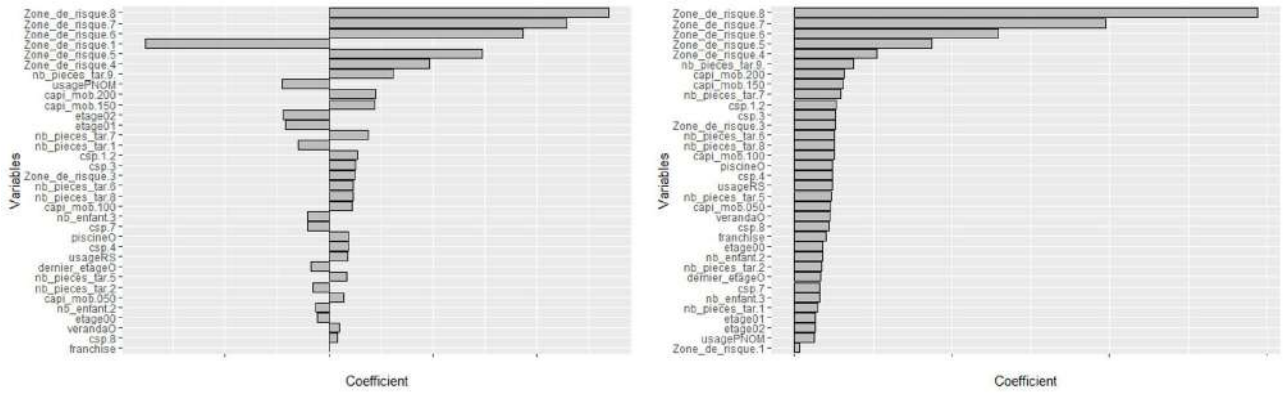


Figure 4.32: Coefficients estimés - modèle de fréquence de la garantie VOL

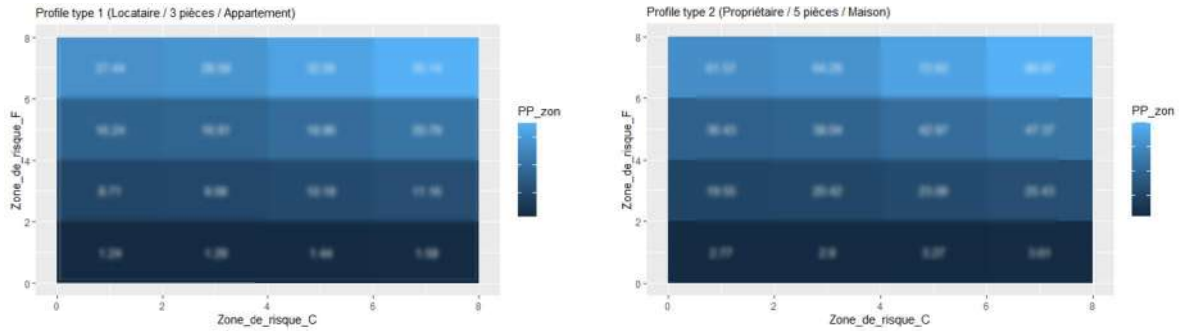
Nous interprétons les résultats en tenant compte du profil de référence explicité ci-dessus. Par exemple, pour la garantie dégâts des eaux, si l'on considère la variable *Zone_de_risque.7*, exp(1,5)=4,48; ainsi, le nombre de sinistres pour un logement appartenant à la zone de risque 7 est 4,48 fois plus élevé qu'un logement dans la zone de risque 2. Pour rappel, les zones de risques sont classés de la moins risquée à la plus risquée.

Analyse détaillée de profils types

Nous avons réalisé une analyse plus détaillée sur deux profils types du portefeuille d'étude.

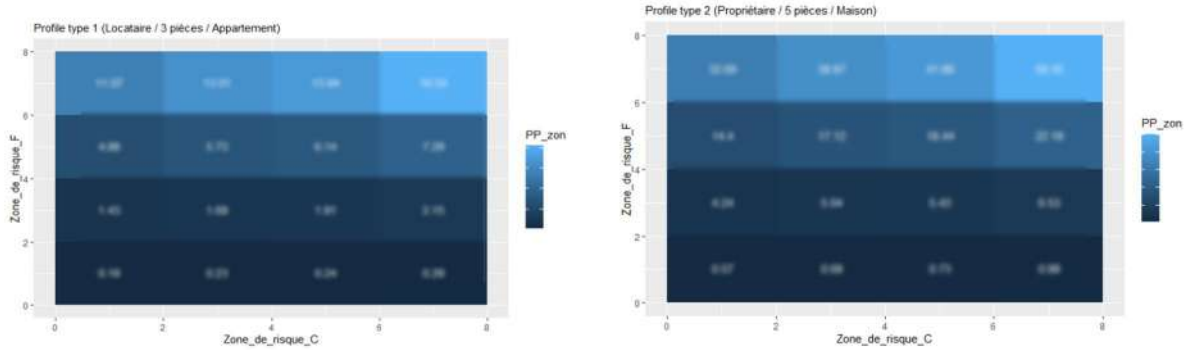
- **Profil 1 :** Le premier profil correspond à un locataire dont le logement est un appartement qui comporte 3 pièces.
- **Profil 2 :** Le deuxième profil est un propriétaire vivant dans une maison qui comporte 5 pièces.

Nous avons représenté ci-dessous des matrices permettant de comparer la prime pure obtenue après intégration des zoniers respectivement liés au coût et à la fréquence pour chaque garantie. Celles-ci nous ont permis de tirer des conclusions sur l'apport d'information de chacun d'entre eux.



(a) Profil 1: locataire - Appartement - 3 pièces (b) Profil 2: Propriétaire - Maison - 5 pièces

Figure 4.33: Analyse de profils types - Garantie DDE



(a) Profil 1: locataire - Appartement - 3 pièces (b) Profil 2: Propriétaire - Maison - 5 pièces

Figure 4.34: Analyse de profils types - Garantie VOL

Les matrices obtenues nous permettent de conclure que plus les zones sont risquées à la fois sur le modèle de coût et le modèle de fréquence, plus la prime pure augmente. Nous déduisons également de ces matrices que la variabilité est plus importante sur la fréquence que sur le coût pour les deux garanties.

Pour conclure ce chapitre, l'intégration d'un zonier dans les modèles GLM améliore nettement les résultats tout en garantissant une bonne compréhension et une bonne interprétation du lien entre la sinistralité et la zone de risque considérée.

Conclusion

Dans le contexte concurrentiel du marché de l'assurance multirisque habitation, la localisation géographique occupe une place majeure. En effet, la localisation du risque influe sur la sinistralité. Elle doit donc faire l'objet d'une attention particulière en étant finement évaluée pour être intégrée comme critère dans la tarification des contrats d'assurance. De plus, les assureurs segmentent de plus en plus leurs tarifications afin d'éviter les effets d'antisélection.

Ainsi, les données internes (recueillis lors de la souscription du contrat) fournies par l'entreprise ainsi que celles récupérées de l'open data ont fait l'objet d'une analyse critique approfondie. Nous avons ainsi travaillé à partir des éléments disponibles, après en avoir mesuré la qualité statistique.

L'enjeu de ce mémoire était de définir un découpage optimal du territoire selon différents niveaux de risque pour les garanties dégâts des eaux et vol. Cette segmentation du portefeuille est appelé zonier. La mise en place de cette variable géographique *zonier* a donc pour objectif d'être intégrée dans le modèle de tarification, afin de :

- Optimiser les modèles de tarifications.
- Résumer l'effet géographique en un seul critère tarifaire (ou en un nombre restreint de variables).
- Apporter une connaissance du risque individuel plus précise, permettant donc une relation client plus pertinente et plus juste. C'est une façon de se démarquer de la concurrence.

Ce mémoire démontre dans un premier temps l'importance de tenir compte du risque géographique dans la tarification. Par exemple, on améliore de 4% l'AIC d'un modèle de fréquence pour la garantie vol en intégrant la variable zonier aux données internes. Dans un deuxième temps, ce mémoire démontre qu'utiliser le zonier crée a de meilleurs indicateurs de performance qu'un GLM classique incluant des données internes et externes (amélioration de 0,5% de l'AIC).

Au vu des résultats, nous pouvons conclure que l'ensemble des travaux réalisés dans le cadre de ce mémoire permettent d'obtenir une unique variable facilement interprétable qui apporte de l'information sur l'environnement géographique de l'assuré. En effet, l'intégration de plusieurs variables géographiques dans un GLM ne permet pas cette facilité de compréhension. En effet, il est plus difficile de repérer si une ville est plus risquée qu'une autre dans le modèle incluant 10 nouvelles variables météorologiques.

En définitif, l'analyse des indicateurs nous permet d'en déduire que le GLM avec le zonier a non seulement de meilleurs indicateurs de performance qu'un GLM n'incluant que les données internes mais il est aussi plus performant qu'un GLM classique avec des données internes et externes. Ainsi, le zonier crée permet d'optimiser le modèle tout en gardant une bonne interprétabilité.

Le but de ce mémoire étant d'apporter un réel intérêt professionnel, les différentes techniques

employées à chaque étape de la réalisation des zoniers ont fait l'objet d'analyses rigoureuses visant à apporter pertinence, exhaustivité et exactitude. Ces techniques ont permis d'aboutir à une nette amélioration de la performance des modèles ainsi qu'à l'obtention de zoniers stables et adaptables dans le temps. Cette amélioration est due à l'approche utilisée derrière chaque étape du processus :

La conciliation entre expérience professionnelle et data science :

Le rapprochement entre les résultats des méthodes avancées de *machine learning* et les jugements d'experts ont permis de faire des choix stratégiques tout au long de l'avancement de l'étude. Cela a été le cas pour le choix du seuil de séparation entre les sinistres graves et attritionnels, ou encore pour le choix des hyperparamètres dans l'étape de lissage des résidus.

L'utilisation de techniques de data science avancées pour apporter de la robustesse aux modèles étudiés :

L'optimisation des hyperparamètres des modèles de *Machine Learning* par des techniques avancées à travers Ray sur Azure ont abouti à une nette amélioration des résultats.

La combinaison entre méthodes classiques et apport des méthodes de *Machine Learning* :

Par exemple, dans la partie visant à implémenter les modèles de coûts et de fréquence, nous avons optimisé le modèle en ajoutant les interactions entre les variables les plus significatives, une information apportée par la méthode du XGBoost implémentée.

Les travaux réalisés nous ont permis de souligner un certain nombre de pistes d'améliorations. En particulier :

- L'approfondissement du modèle tarifaire (modèles de tarifications alternatifs (ex: tweedie), modélisation des sinistres graves)
- Les approches alternatives (approche par lissage géographique (ex: krigeage), établissement de zoniers de prime pure)
- Le travail sur les résidus (prise en compte de la temporalité des résidus, maillage plus fin (code IRIS))
- Le travail sur le paramétrage (optimisation des hyperparamètres des modèles de *Machine Learning* et des méthodes de lissage)
- La prise en compte des modifications perpétuelles des codes INSEE dans le temps (retraitements des codes INSEE des jeux de données considérés via des tables de mouvement de communes)

Bibliographie

Livres

[1] HASTIE T, TIBSHIRANI R, FRIEDMAN J (2006), *The elements of statistical learning : data mining, inference, and prediction*, Springer

Mémoires

[2] BERAUD-SUDREAU G (2017), *Construction d'un zonier en MRH à l'aide d'outils de data-science*, le CNAM

[3] FERRIER C (2016), *Le zonier en tarification IARD : approche comparative de deux techniques de construction d'un critère de segmentation géographique en assurance habitation*, ISFA

[4] LOIRET C (2016), *Refonte du tarif Multirisque Habitation : construction de micro-zoniers et intégration de la sinistralité passée à l'adresse*, ISFA

[5] PARIENTE J (2017), *Modélisation du risque géographique en assurance habitation*, Université Paris Dauphine

Annexe A

Analyse de la base de données et modélisation

A.1 Les conventions CIDRE et IRSI

Avant le 1 juin 2018: En cas de sinistre dégâts des eaux ou incendie de faible ampleur (inférieur à 1600 €), la convention CIDRE prévoyait en France que ce soit l'assureur de l'occupant et non celui du responsable qui prenait en charge l'indemnisation. Cela est dû à la forte fréquence de ces sinistres mais surtout du coût élevé de leur gestion. Dans le cadre d'application de cette convention, la franchise n'était pas appliquée. Après le 1 juin 2018 : Depuis le 1 juin, tous les dommages inférieurs à 5000 € hors taxe rentrent dans le cadre de la convention IRSI. C'est l'assureur du local sinistré (le "gestionnaire" du dossier) qui s'occupe de gérer le sinistre dégât des eaux. Avant l'établissement de cette convention, le traitement d'un sinistre impliquait plusieurs intervenants ce qui causait souvent de nombreux litiges. Désormais, ce gestionnaire a pour objectif de s'assurer de la réalité du dégât, d'organiser les modalités de recherche de fuite, de désigner un expert et de procéder à l'évaluation des dommages. Ensuite, il doit désigner l(es) assureur(s) qui doi(ven)t prendre en charge les dégâts en fonction des nouveaux barèmes par tranches. En effet, l'IRSI établit deux tranches de sinistres en fonction des dommages matériels et autres frais : La première tranche inclus les dommages inférieurs à 1600 € (sinistre indemnisé par l'assureur gestionnaire avec abandon de recours). Dans ce cas, il y a une prise en charge globale par l'assureur gestionnaire (donc l'assureur de l'immeuble sinistré). La deuxième tranche, elle, regroupe les dommages supérieurs à 1600 € et inférieurs à 5000 € (sinistre indemnisé par l'assureur gestionnaire avec recours). Si le dégât entre dans cette tranche, un seul expert est désigné par l'assureur gestionnaire. Son rapport est opposable aux divers intervenants (assureurs des uns et des autres) et ce sera à l'assureur du bien sinistré (donc au gestionnaire) de prendre en charge le sinistre.

A.2 Variables Tarifaires

Variables liées au contrat	Niveau de franchise
	Année
Variables sur le bien à assurer	Type d'habitation (maison ou appartement)
	Usage (Résidence secondaire, principale etc.)
	Nombre de pièces
	Dépendances
	Étage
	Dernier étage
	Véranda
	Piscine
	Garage
	Zone géographique (Adresse)
Variable sur le contenu à assurer	Valeur des biens mobiliers
	Objet précieux
Variables concernant le souscripteur	Locataire ou propriétaire
	CSP
	Nombre d'enfants
Variables liées au modèle coût fréquence	Exposition
	Nombre de sinistres
	Charges de sinistres

Table A.1: Ensemble des variables tarifaires

A.3 CSP

CSP		Nomenclature de la PCS-ESE 2003/ 2017 (Base Interne)	
1	Agriculteurs exploitants	11	Agriculteurs sur petite exploitation
		12	Agriculteurs sur moyenne exploitation
		13	Agriculteurs sur grande exploitation
2	Artisans, commerçants et chefs d'entreprise	21	Artisan
		22	Commerçant et assimilés
		23	Chefs d'entreprise de 10 salariés ou plus
3	Cadres et professions intellectuelles supérieures	31	Professions libérales
		33	Cadres de la fonction publique
		34	Professeurs, professions scientifiques
		35	Professions de l'information, des arts et des spectacles
		37	Cadres administratifs et commerciaux d'entreprise
		38	Ingénieurs et cadres techniques d'entreprise
4	Professions intermédiaires	42	Professeurs des écoles, instituteurs et assimilés
		43	Professions intermédiaires de la santé et du travail social
		44	Clergé, religieux
		45	Professions intermédiaires administratives de la fonction publique
		46	Professions intermédiaires administratives et commerciales des entreprises
		47	Techniciens
		48	Contremaîtres, agents de maîtrise
5	Employés	52	Employés civils et agents de service de la fonction publique
		53	Policiers et militaires
		54	Employés administratifs d'entreprise
		55	Employés de commerce
		56	Personnels des services directs aux particuliers
6	Ouvriers	62	Ouvriers qualifiés de type industriel
		63	Ouvriers qualifiés de type artisanal
		64	Chauffeurs
		65	Ouvriers qualifiés de la manutention, du magasinage et du transport
		67	Ouvriers non qualifiés de type industriel
		68	Ouvriers non qualifiés de type artisanal
		69	Ouvriers agricoles
7	Retraités	71	Anciens agriculteurs exploitants
		72	Anciens artisans, commerçants et chefs d'entreprise
		74	Anciens cadres
		75	Anciennes professions intermédiaires
		77	Anciens employés
		78	Anciens ouvriers
8	Autres personnes sans activité professionnelle	81	Chômeurs n'ayant jamais travaillé
		83	Militaires du contingent
		84	Élèves, étudiants
		85	Personnes diverses sans activité professionnelle de moins de 60 ans (sauf retraités)
		86	Personnes diverses sans activité professionnelle de 60 ans et plus (sauf retraités)

Figure A.1: Catégories socio-professionnelles

A.4 Analyse descriptive

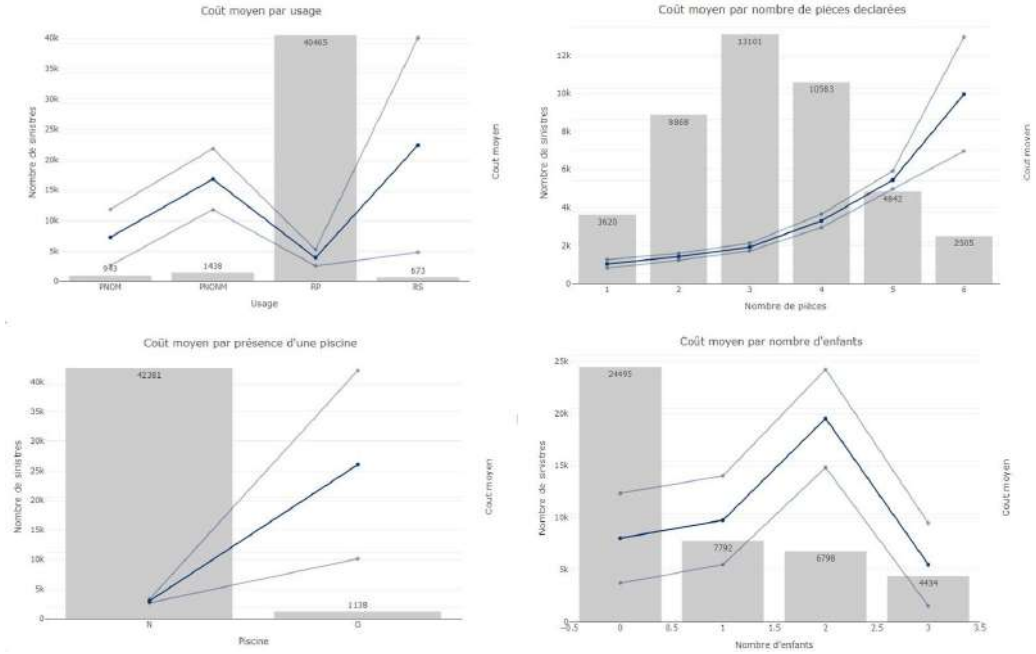


Figure A.2: Statistiques descriptives pour le modèle de coût de la garantie DDE

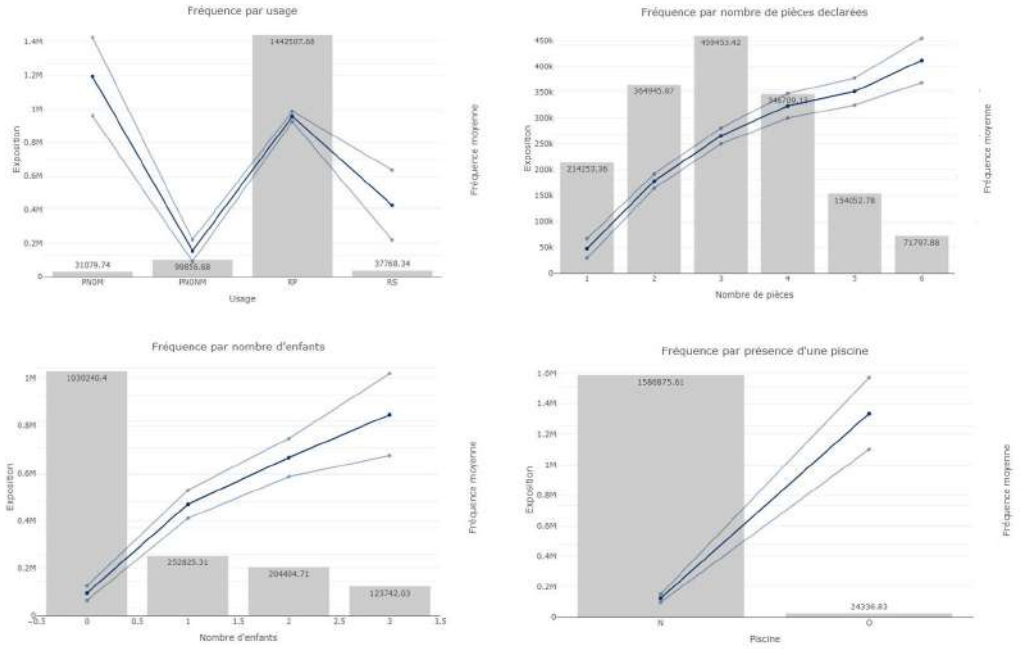


Figure A.3: Statistiques descriptives pour le modèle de fréquence de la garantie DDE

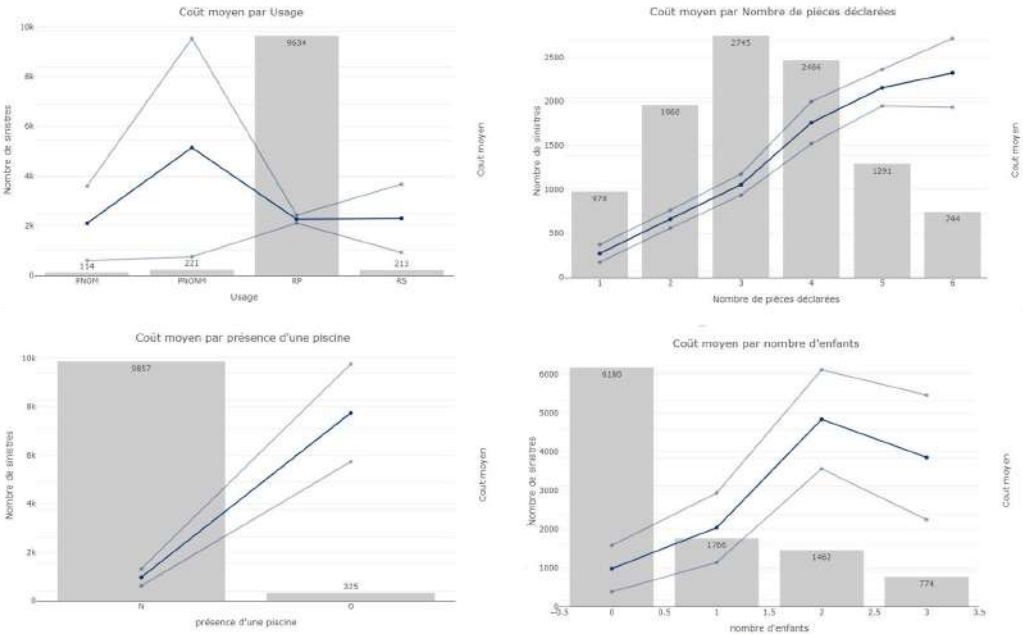


Figure A.4: Statistiques descriptives pour le modèle de coût de la garantie VOL

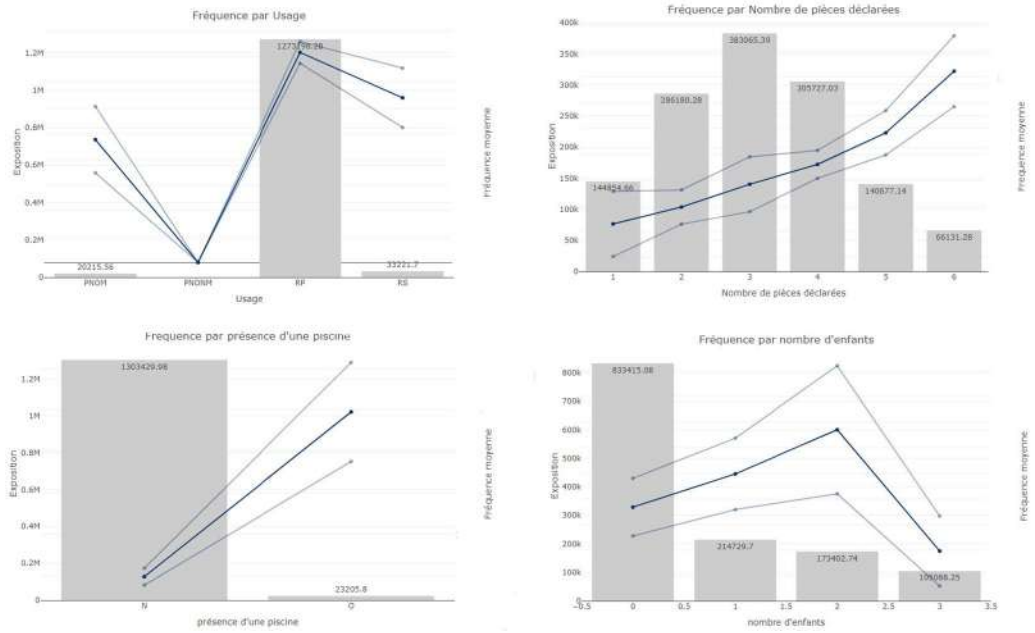


Figure A.5: Statistiques descriptives pour le modèle de fréquence de la garantie VOL

A.5 Graphiques XGBoost

Pour la sélection de variables du modèle de coût, nous utilisons les résultats suivants qui incluent toutes les variables explicatives (dont celles qui sont corrélées):

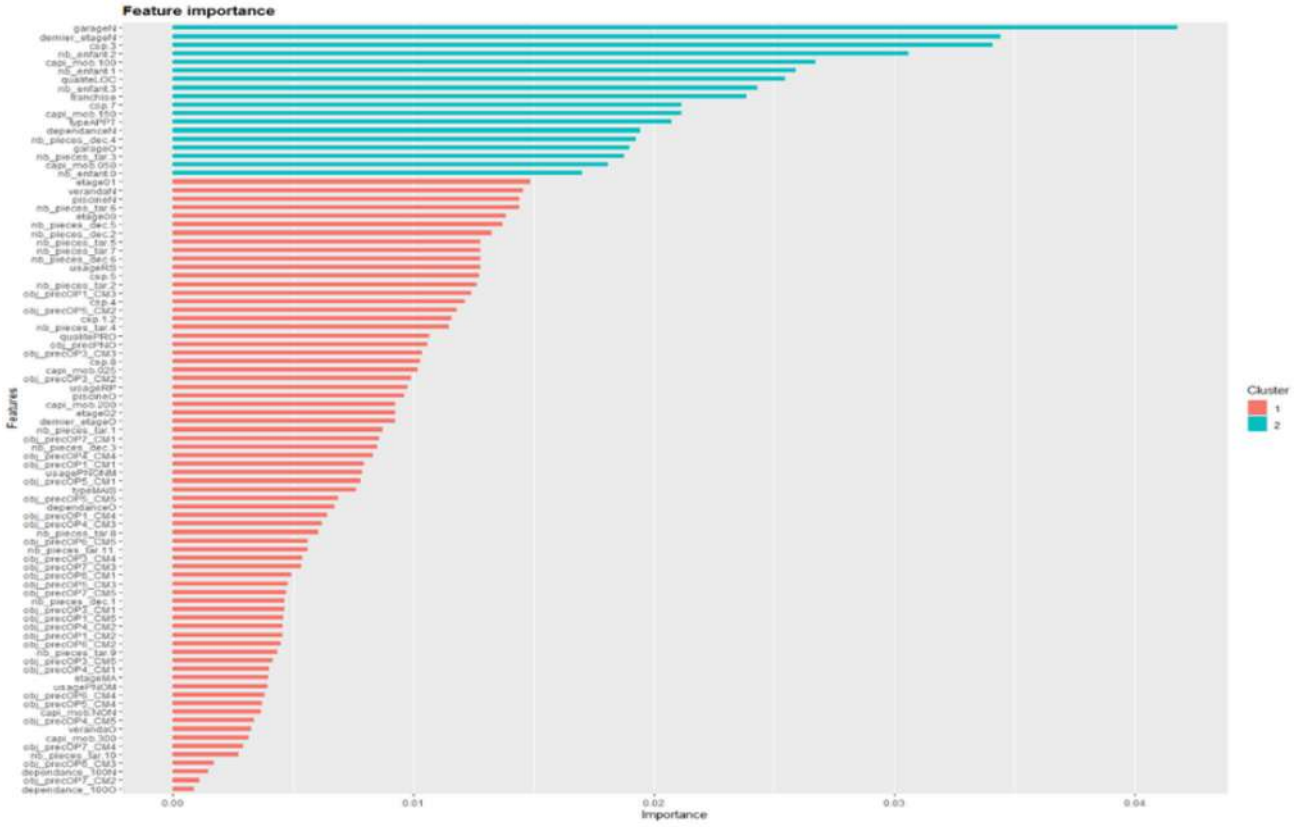


Figure A.6: Importance des variables par XGBoost pour la garantie dégât des eaux

Les variables les plus importantes dans le cas de la garantie dégat des eaux correspondent aux variables *Garage*,*Dernier_etage*, *CSP*,*Nombre d'enfants*, *Capital_Mobilier* et *Qualite*.

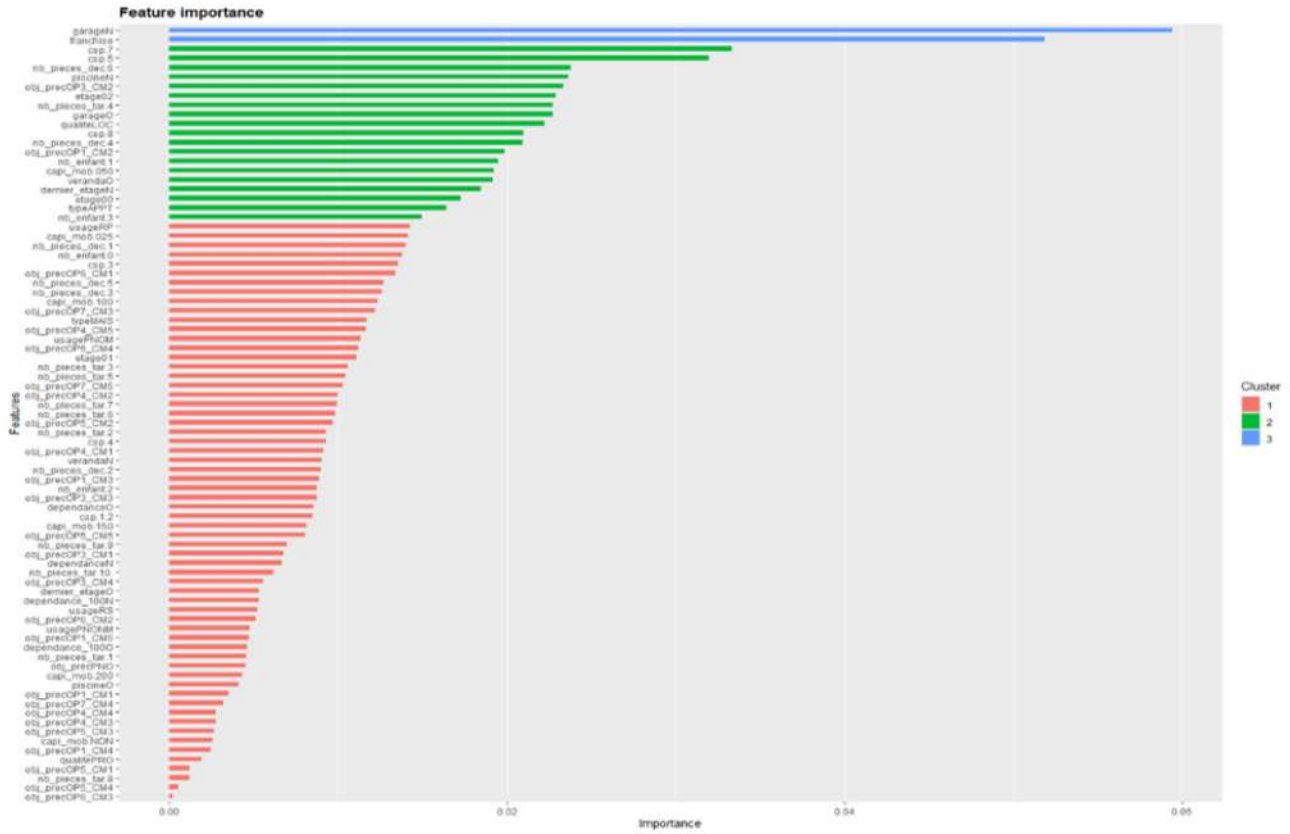


Figure A.7: Importance des variables par XGBoost pour la garantie vol

Les variables les plus importantes dans le cas de la garantie vol correspondent aux variables *Garage,franchise*, *CSP,Nombre de pièces déclarés* et *piscine*.

A.6 Implémentation du modèle de coût

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)	
NULL			31920	23695		
qualitePRO	1	573.39	31919	23121	< 2.2e-16	***
usagePNOM	1	0.81	31918	23120	0.2867153	
usagePNONM	1	2.30	31917	23118	0.0722367	.
usageRS	1	1.73	31916	23116	0.1188495	
nb_pieces_tar.1	1	10.96	31915	23105	8.702e-05	***
nb_pieces_tar.2	1	36.44	31914	23069	8.280e-13	***
nb_pieces_tar.4	1	15.70	31913	23053	2.640e-06	***
nb_pieces_tar.5	1	1.99	31912	23051	0.0941538	.
nb_pieces_tar.6	1	73.07	31911	22978	< 2.2e-16	***
nb_pieces_tar.7	1	43.74	31910	22934	4.493e-15	***
nb_pieces_tar.8	1	31.18	31909	22903	3.586e-11	***
nb_pieces_tar.9	1	17.91	31908	22885	5.259e-07	***
nb_pieces_tar.10	1	2.72	31907	22883	0.0504841	.
nb_pieces_tar.11.	1	18.43	31906	22864	3.600e-07	***
etage00	1	3.88	31905	22860	0.0195763	*
etage01	1	2.42	31904	22858	0.0649305	.
etage02	1	129.20	31903	22729	< 2.2e-16	***
dernier_etage0	1	4.27	31902	22724	0.0142559	*
veranda0	1	0.10	31901	22724	0.7056566	
piscine0	1	9.70	31900	22715	0.0002215	***
dependance0	1	0.22	31899	22714	0.5815016	
dependance_1000	1	3.21	31898	22711	0.0337562	*
garage0	1	5.08	31897	22706	0.0075392	**
capi_mob.050	1	0.12	31896	22706	0.6787955	
capi_mob.100	1	2.39	31895	22704	0.0669015	.
capi_mob.150	1	1.84	31894	22702	0.1078017	
capi_mob.200	1	2.53	31893	22699	0.0594336	.
capi_mob.300	1	2.06	31892	22697	0.0887631	.
capi_mob.NON	0	0.00	31892	22697		
nb_enfant.1	1	0.14	31891	22697	0.6609588	
nb_enfant.2	1	1.62	31890	22695	0.1307744	
nb_enfant.3	1	4.95	31889	22690	0.0083370	**
franchise	1	2.35	31888	22688	0.0692731	.

(a) GLM avant sélection de variable

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)	
NULL			31920	23695		
qualitePRO	1	573.39	31919	23121	< 2.2e-16	***
usagePNOM	1	0.81	31918	23120	0.2867075	
usagePNONM	1	2.30	31917	23118	0.0722320	.
nb_pieces_tar.5	1	14.92	31916	23103	4.652e-06	***
nb_pieces_tar.6	1	100.19	31915	23003	< 2.2e-16	***
nb_pieces_tar.7	1	51.13	31914	22952	< 2.2e-16	***
nb_pieces_tar.8	1	33.13	31913	22919	8.851e-12	***
nb_pieces_tar.9	1	18.44	31912	22900	3.568e-07	***
nb_pieces_tar.11.	1	18.46	31911	22882	3.518e-07	***
etage00	1	2.04	31910	22880	0.0907614	.
etage01	1	3.80	31909	22876	0.0207622	*
etage02	1	143.44	31908	22732	< 2.2e-16	***
dernier_etage0	1	4.20	31907	22728	0.0150893	*
piscine0	1	9.88	31906	22718	0.0001945	***
dependance_1000	1	2.57	31905	22716	0.0575493	.
garage0	1	3.88	31904	22712	0.0194559	*
capi_mob.050	1	0.17	31903	22712	0.6263997	
capi_mob.100	1	2.19	31902	22710	0.0790213	.
capi_mob.150	1	1.67	31901	22708	0.1251073	
capi_mob.200	1	2.18	31900	22706	0.0797856	.
capi_mob.300	1	1.99	31899	22704	0.0940527	.
nb_enfant.1	1	0.40	31898	22703	0.4551813	
nb_enfant.2	1	2.53	31897	22701	0.0592768	.
nb_enfant.3	1	6.14	31896	22695	0.0032940	**
franchise	1	2.31	31895	22692	0.0713178	.

(b) GLM après sélection de variable

A.7 *Plot effects*

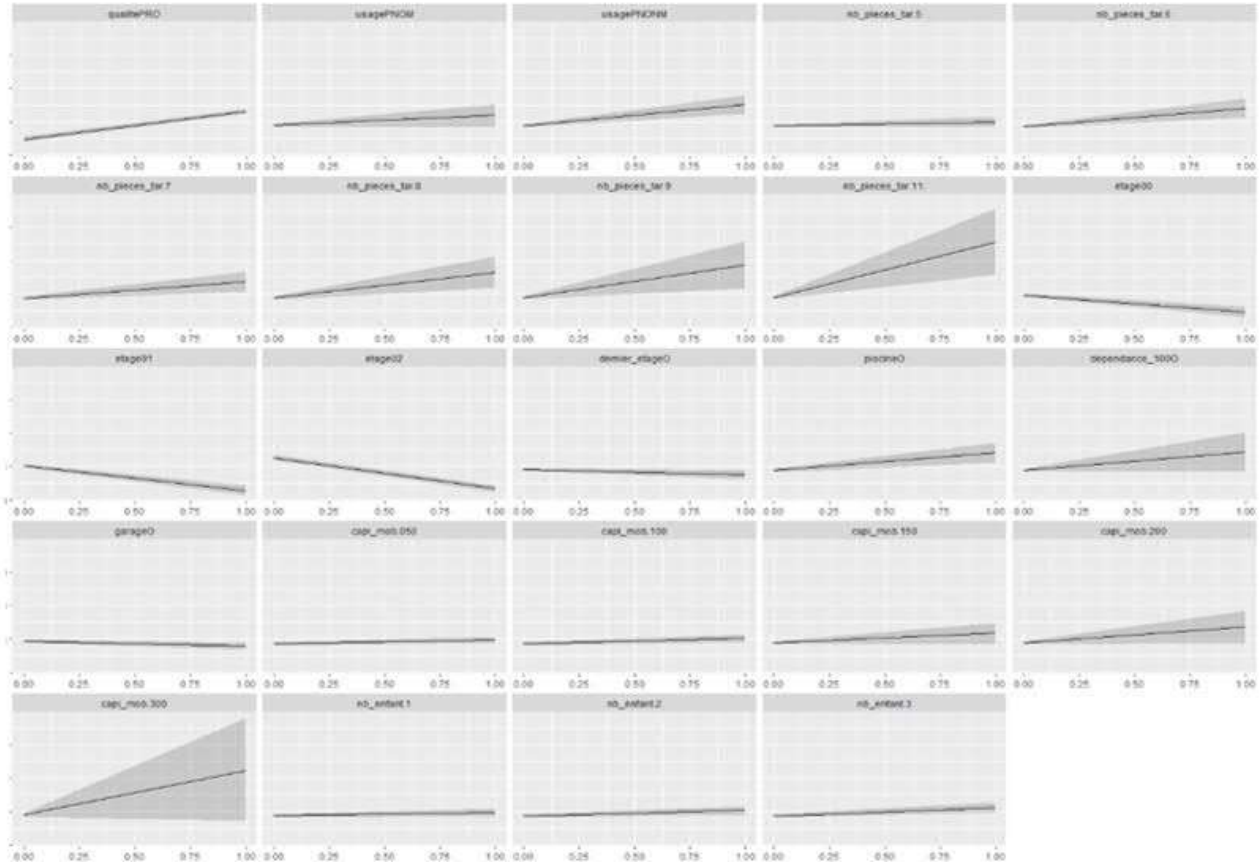


Figure A.9: Plot effect garantie DDE

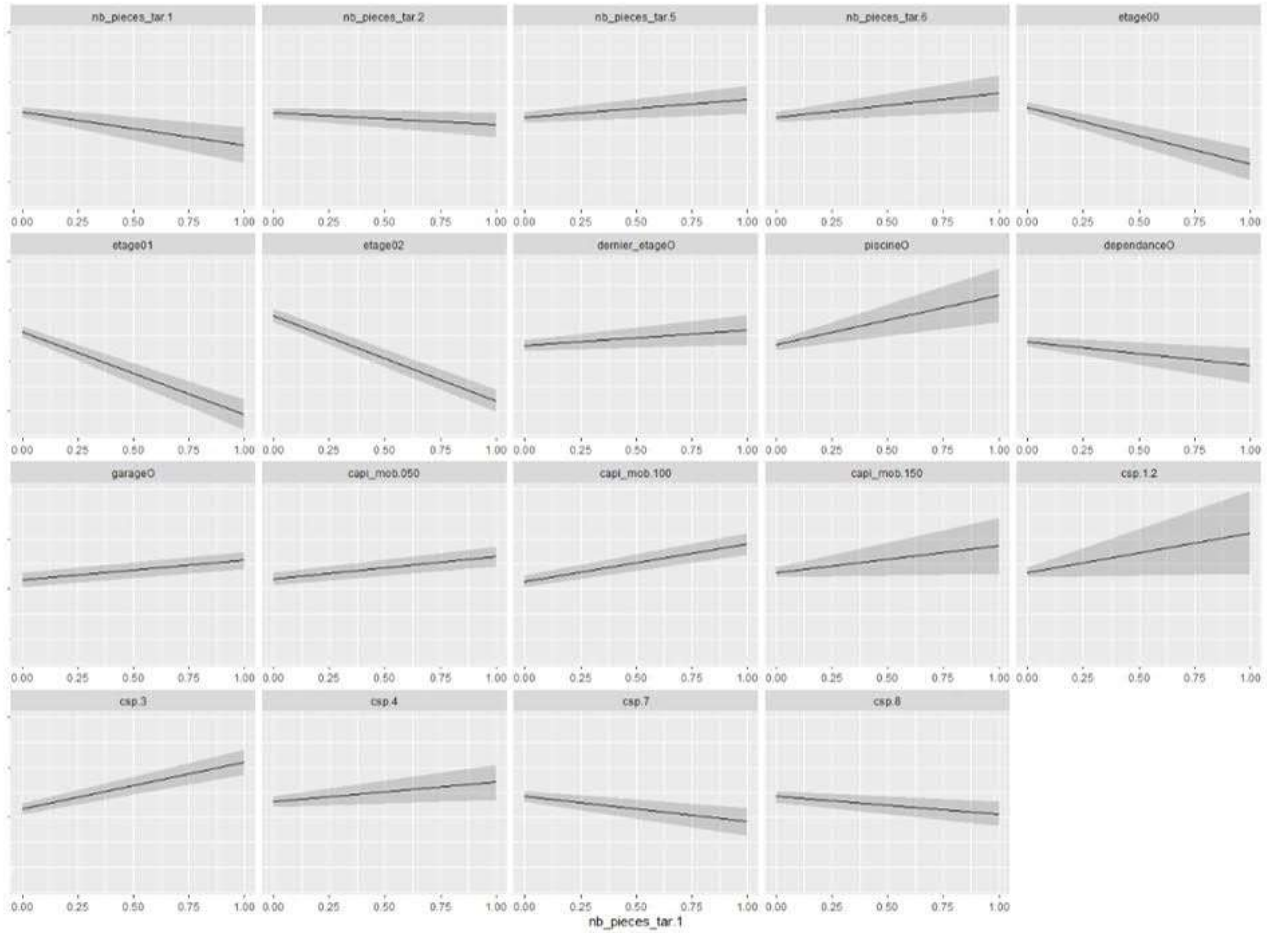


Figure A.10: Plot effect garantie VOL

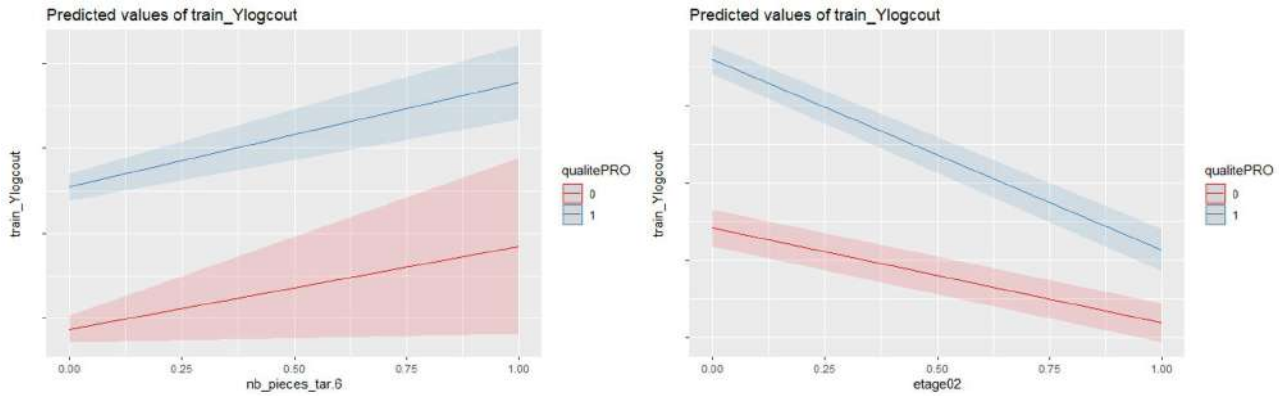


Figure A.11: Plot effect avec intéractions garantie DDE

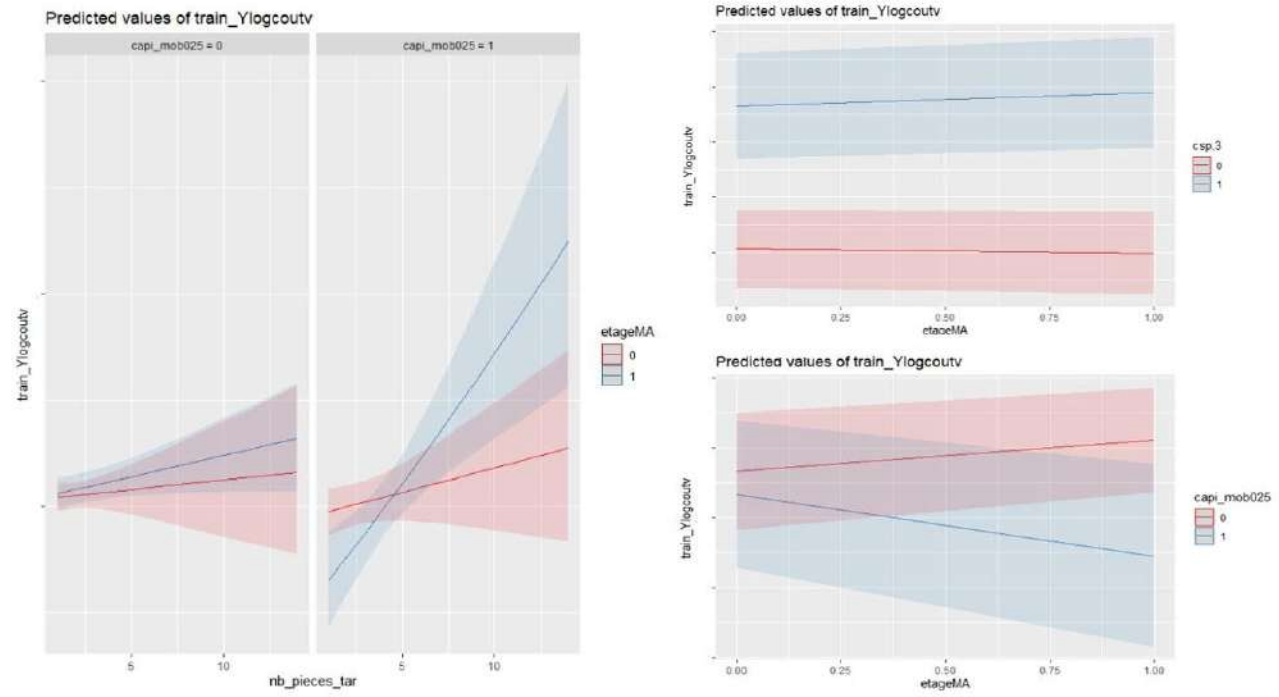
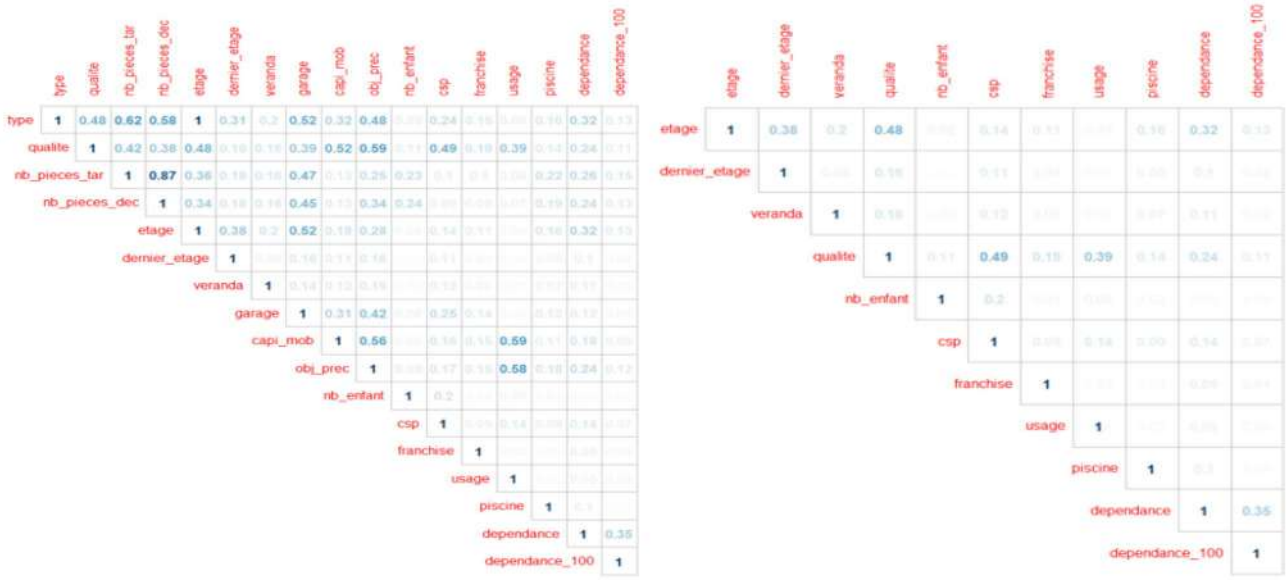


Figure A.12: Plot effect avec intéractions garantie vol

A.8 Implémentation du modèle de fréquence

Corrélation des variables (V de cramer)



(a) avant suppression des variables corrélées (b) après suppression des variables corrélées

Figure A.13: Analyse des corrélations pour la garantie DDE - V de Cramer

En utilisant la même approche que pour le modèle de coût, afin de ne pas avoir des variables explicatives trop corrélées entre elles et éviter une redondance d'information, nous poursuivons l'étude en supprimant les variables suivantes: *Type, Nombre de pièces déclarées, Objets précieux, Qualité et Garage.*

En appliquant la même analyse à la garantie vol, nous obtenons les résultats suivants pour les V de Cramer:

(a) avant suppression des variables corrélées (b) après suppression des variables corrélées

Figure A.14: Analyse des corrélations pour la garantie vol - V de Cramer

Ainsi, nous retirons de la base de donnée les variables: *Type, Nombre de pièces déclarées, Objets précieux, Qualité et Garage.*

Les V de Cramer obtenus après avoir retiré les variables mentionnées ci-dessus sont tous inférieurs à 0,5 à l'exception des variables *Usage* et *Capital Mobilier* qui seront traitées à l'issu de l'étape de transformation en dummies. Le but étant d'obtenir des variables explicatives n'ayant pas énormément de liens entre elles, ces résultats semblent être satisfaisants pour la suite de l'étude.

Regroupement des modalités

Suite à l'analyse des expositions des différentes modalités des variables tarifaires, nous avons regroupé certaines modalités entre elles afin d'avoir une exposition plus globale plus significative et ainsi obtenir des résultats non biaisés. En guise d'exemple, les logements ayant plus de 11 pièces (comprises dans la tarification) ont été regroupées en une seule modalité *11 ou plus* pour la garantie dégâts des eaux. L'approche utilisée est détaillée dans la section ci-dessus (section implémentation des modèles de coût).

Détermination de la distribution de la fréquence et choix de la fonction de lien

La fréquence des sinistres est traditionnellement appréhendée à l'aide de la loi de Poisson ou la loi binomiale négative. Pour notre part, afin de modéliser la fréquence de sinistre (nombre de sinistres sur exposition), nous avons choisi d'utiliser le modèle de Poisson avec la variable exposition en offset*. En effet, en théorie des probabilités et en statistiques, la loi de Poisson est une loi de probabilité discrète qui décrit le comportement du nombre d'événements se produisant dans un intervalle de temps fixé et indépendamment du temps écoulé depuis l'événement précédent.

Dans l'ajustement de notre modèle, nous devons nous assurer de l'adéquation de la loi de Poisson

*Étant donnée que nous utilisons une fonction de lien logarithmique, nous prenons en pratique log(exposition) en offset.

à nos données en vérifiant l'hypothèse d'équidispersion de celle-ci (homogénéité du portefeuille par rapport au risque). En effet, nous pouvons soupçonner un problème de sur-dispersion des données si nous observons :

- Une large déviance par rapport au nombre de degrés de liberté.
- Une variance supérieure à la moyenne

Dans la pratique, nous avons calculé l'espérance et la variance de nos données afin de vérifier que ces deux valeurs sont bien égales. Pour la garantie dégâts des eaux, nous obtenons une espérance de 0.01256 et une variance de 0.01317. Pour la garantie vol, nous obtenons une espérance de 0.00325 et une variance de 0.00330. Au vu des résultats précédents, le choix d'une loi de Poisson pour la modélisation des données semble cohérent. Selon les résultats ci-dessus, la loi de Poisson semble être appropriée pour modéliser la fréquence des sinistres pour chacune des garanties considérées.

Sélection des variables

Après implémentation des GLM pour les deux garanties via le logiciel R, nous remarquons que certaines variables des GLM n'ont pas une faible p-valeur ce qui nous pousse à penser qu'à priori, certaines variables explicatives ne sont pas significatives. Ainsi, nous procédons à une étape de sélection de variable automatique qui nous a permis d'obtenir un modèle réduit.

Coefficients:

	Std. Error	z value	Pr(>|z|)	
(Intercept)	0.0235084	-189.440	< 2e-16	***
qualitePRO	0.0146224	55.012	< 2e-16	***
usagePNOM	0.0376275	-13.275	< 2e-16	***
usagePNONM	0.0307174	-37.279	< 2e-16	***
usageRS	0.0435170	-14.908	< 2e-16	***
etage00	0.0195740	38.651	< 2e-16	***
etage01	0.0185355	37.627	< 2e-16	***
etage02	0.0158745	50.326	< 2e-16	***
dernier_etageO	0.0167145	-27.206	< 2e-16	***
verandaO	0.0255485	2.164	0.03049	*
piscineO	0.0344856	14.004	< 2e-16	***
dependanceO	0.0240365	-3.012	0.00259	**
dependance_100O	0.0746904	-5.017	5.26e-07	***
nb_enfant.1	0.0150685	17.324	< 2e-16	***
nb_enfant.2	0.0160477	21.023	< 2e-16	***
nb_enfant.3	0.0190110	28.078	< 2e-16	***
csp.1.2	0.0499816	2.344	0.01907	*
csp.3	0.0163118	15.661	< 2e-16	***
csp.4	0.0215088	2.102	0.03554	*
csp.7	0.0175264	-4.532	5.84e-06	***
csp.8	0.0155953	-3.030	0.00245	**
franchise	0.0001005	-4.104	4.06e-05	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 296271 on 2771060 degrees of freedom
Residual deviance: 288090 on 2771039 degrees of freedom
AIC: 356347

Number of Fisher Scoring iterations: 7

Figure A.15: GLM après sélection de variable

Nous comparons les deux modèles suivants:

- Modèle 1: sans étape de sélection de variable.

- Modèle 2: avec sélection de variable par la méthode stepwise.

Choix du meilleur modèle

Critères d'évaluation de la qualité d'ajustement des modèles			
Critère	Modèle	Garantie dégâts des eaux	Garantie vol
Déviance			
	GLM1	289593	82044
	GLM2	288090	82041
AIC			
	GLM1	357860	96796
	GLM2	356347	96792

Lorsque nous comparons les prédictions de chacun des modèles aux valeurs réellement observées, nous obtenons:

Critères d'évaluation du pouvoir de prédicion des modèles			
Critère	Modèle	Garantie dégâts des eaux	Garantie vol
RMSE			
	GLM1	0.1125906	0.1142737
	GLM2	0.1125914	0.1142738
Ratio			
	GLM1	0.9603277	1.002730
	GLM2	0.9602723	1.002393

A.9 Données manquantes

Analyse des données manquantes suite à la jointure de la base de données externe créée aux résidus.

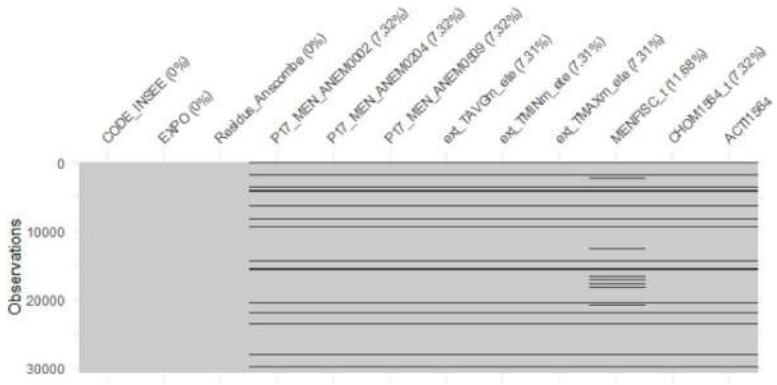


Figure A.16: Valeurs manquantes du modèle de fréquence de la garantie DDE

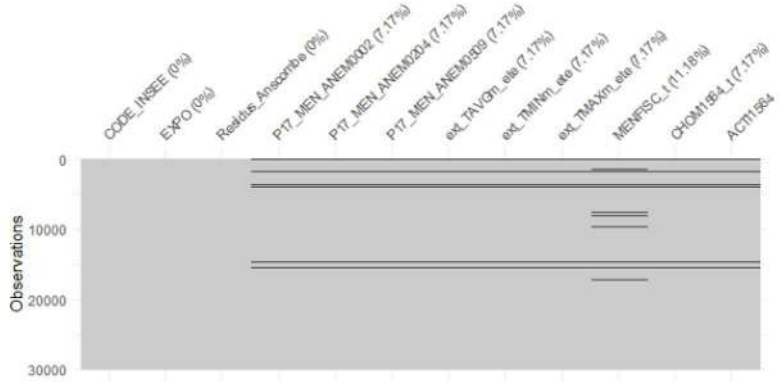


Figure A.17: Valeurs manquantes du modèle de fréquence de la garantie VOL

Figure A.18: Valeurs manquantes du modèle de coût de la garantie DDE

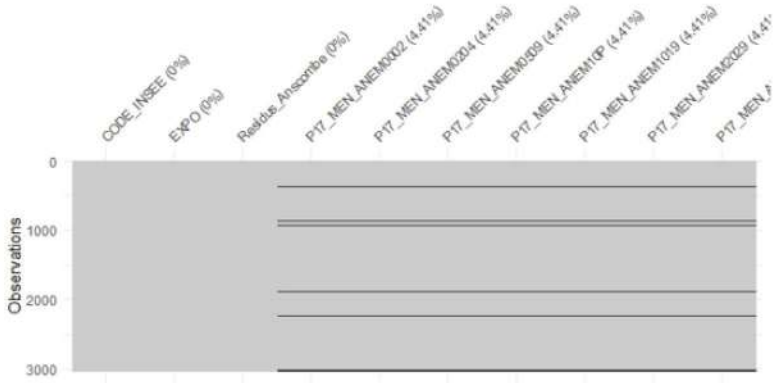


Figure A.19: Valeurs manquantes du modèle de coût de la garantie VOL

Les données manquantes sont représentées par les bandes noires sur les figures ci-dessus. Nous remarquons que celles-ci sont plus ou moins éparpillée dans la base de données et qu'elles concernent en moyenne 7% des communes dans les modèles de fréquence et environ 4% dans les modèles de coût.

Annexe B

Méthodes de cartographie et interpolation spatiale

B.1 Interpolation par Krigeage - *Théorie*

Le krigeage est l'approche géostatistique la plus utilisée pour réaliser des interpolations spatiales. Il permet à partir de données dispersées, d'obtenir une représentation homogène des informations étudiées. Dans l'analyse de la température et des précipitations, seul un certain nombre de stations nous fournissent les données. Le krigeage a pour but d'estimer les valeurs hors station à l'aide des mesures de chacune des stations. Nous pourrons alors créer une carte étendant les relevés au territoire Français.

Une des spécificités du krigeage est qu'il prend en compte non seulement la distance entre le point à prédire et les stations de mesure mais aussi les distances entre les données (les stations de mesure en elles-mêmes). Les méthodes de krigeage utilisent ces informations pour attribuer un poids à chaque échantillon avant de réaliser les prédictions.

L'idée générale du krigeage est donc de prévoir la valeur de la variable étudiée en un point non échantillonné x_0 par une combinaison linéaire de données ponctuelles connues. La formule générale utilisée par cette méthode d'interpolation consiste en une somme pondérée des données :

$$z(x_0) = a + \sum_{k=1}^{N} w_k z(x_k) \tag{B.1}$$

avec z_k la valeur relevée au point x_k, w_k le poids inconnu associé à chacune des valeurs observées à l'emplacement k, x_0 l'emplacement de prévision représentant le point à interpoler, a une constante et N le nombre total de valeurs connus.

Il s'agit ici simplement de déterminer la valeur des poids w_k de la combinaison linéaire. Avec la méthode IDW, la pondération w_k, dépendait uniquement de la distance par rapport à l'emplacement de prévision. En revanche, avec la méthode de krigeage, les pondérations ne s'appuient pas seulement sur la distance entre les points voisins connus et l'emplacement de prévision, mais aussi sur l'organisation spatiale générale des points. Par conséquent, le krigeage donne de bons résultats, même pour les zones de forte déclivité ainsi que pour les zones couvrant des pentes abruptes et douces à la fois.

Pour effectuer une prévision par krigeage, deux tâches sont nécessaires. La première est la découverte

des règles de dépendance et la deuxième correspond à la formulation des prévisions. Afin de réaliser ces deux tâches, il faut procéder en plusieurs étapes :

Figure B.1: Démarche pour effectuer une prévision par krigeage

Étape 1: Estimation du variogramme qui permet d'évaluer les auto-corrélations spatiales

Pour appliquer l'interpolation par krigeage, il est important d'explorer la structure spatiale des données afin de vérifier si celles-ci sont bien auto-corrélées. L'analyse variographique permet de mener à bien cette étude. L'outil principal utilisé pour cette analyse est le semi-variogramme qui illustre l'évolution de la semi-variance en fonction de la distance entre les mesures et permet ainsi d'étudier le lien spatial entre les données. Ils'agit d'un modèle de covariance dépendant de la distance entre les observations et est défini de la manière suivante :

$$\gamma(h) = \frac{1}{2} Var(Z(\mathbf{x}+h) - Z(\mathbf{x})) \tag{B.2}$$

avec **x** le vecteur des coordonnées, Z la variable régionalisée* étudiée et h est le vecteur distance†. Pour rappel, la variable régionalisée est vue comme une réalisation d'une fonction aléatoire $Z(\mathbf{x})$ et toute valeur $z(x_k)$ correspond à une réalisation d'une variable aléatoire $Z(x_k)$. Cette première étape consiste donc à estimer le semi-variogramme à partir des données disponibles, c'est-à-dire les $z(x_k)$ pour k = 1 .. N.

Étape 2: Ajustement d'un modèle

La prochaine étape consiste à ajuster un modèle selon les points formant le semi-variogramme empirique obtenu. La modélisation des semi-variogrammes est une étape clé située entre la description spatiale et la prévision spatiale. Pour ajuster un modèle au semi-variogramme empirique, il convient de sélectionner une fonction comme modèle (sphérique, circulaire, exponentiel etc.). Chaque modèle est conçu de façon à s'adapter plus précisément à différents types de phénomènes. Par exemple, une fonction de type sphérique indique une réduction progressive de l'auto-corrélation spatiale (équivalant à une augmentation de semi-variance) jusqu'à une certain distance, au-delà de laquelle l'auto-corrélation est de 0. D'un autre coté, un modèle exponentiel est utilisé lorsque l'auto-corrélation spatiale se réduit exponentiellement avec l'accroissement de la distance.

Étape 3: Prédiction des valeurs inconnues

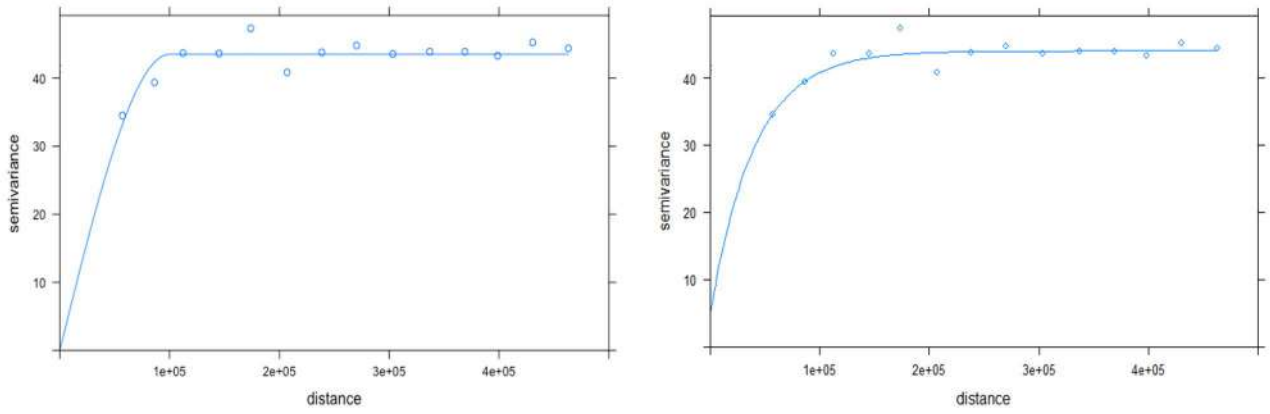
Après avoir modélisé l'auto-corrélation spatiale dans les données, la prochaine étape consiste à déterminer une prévision en utilisant le modèle ajusté. Pour ce faire, le krigeage génère des pondérations à partir des valeurs relevées avoisinantes afin de prédire des emplacements non mesurés. Comme pour la méthode IDW, les valeurs connues les plus proches des emplacements non mesurés ont l'influence la plus forte. Toutefois, les facteurs de pondération du krigeage pour les points relevés avoisinants proviennent du semi-variogramme développé en examinant les données de nature spatiale contrairement à la méthode IDW qui utilise une formule simple basé sur la distance.

*Dans le domaine de la géostatistique, une variable régionalisée (VR) est toute fonction mathématique déterministe destinée à modéliser un phénomène présentant une structure plus ou moins prononcée dans l'espace et/ou dans le temps

†Rappel: la norme euclidienne d'un vecteur $h = (x_h, y_h)$ est $|h| = \sqrt{x_h^2 + y_h^2}$

B.2 Interpolation par krigeage - *Pratique*

Pour interpoler les données, nous commençons par nous intéresser à la méthode de krigeage qui est plus avancée que la méthode IDW. Dans ce contexte, nous entamons les étapes de la figure 3.3 en estimant le semi-variogramme empirique et nous choisissons les fonctions de type sphérique et exponentielle pour essayer d'ajuster au mieux notre modèle.



(a) Modèle sphérique

(b) Modèle exponentiel

Figure B.2: Représentation du semi-variogramme sur les données de température moyenne

Les semi-variogrammes ci-dessus représentent des diagrammes affichant les semi-variances moyennes sur l'axe des ordonnées et la distance sur l'axe des abscisses ainsi que le modèle ajusté pour les données de température moyennes. Chaque point représente une paire d'emplacements.

L'auto-corrélation quantifie un principe élémentaire de géographie : les choses qui sont plus proches se ressemblent davantage que celles qui sont plus éloignées. Ainsi, dans chacun des graphiques, les paires d'emplacements qui sont proches (c'est-à-dire à l'extrémité gauche sur l'axe des abscisses) devraient présenter des valeurs similaires (en bas de l'axe des ordonnée). En revanche, plus les paires d'emplacements s'éloignent les unes des autres (à l'extrémité droite sur l'axe des abscisses), plus elles deviennent dissemblables et ont un écart plus important (en haut de l'axe des ordonnées). En conséquence, nous nous attendons à voir des valeurs plus élevée sur la partie droite du graphique et des valeurs plus faible sur la partie gauche.

Les résultats obtenus permettent de faire l'étude du comportement spatial de la variable régionalisée examinée. A une certaine distance, nous remarquons que le modèle se stabilise. La distance à laquelle le modèle commence à s'aplanir est appelée la portée. L'atteinte d'un plateau indique qu'à partir d'une certaine distance, il n'y a plus de dépendance spatiale entre les données. Les emplacements d'échantillons séparés par une distance inférieure à la portée sont auto-corrélés spatialement, en revanche, les emplacements séparés par une distance supérieure à la portée ne le sont pas. Dans notre cas, la petite valeur de la portée illustre une faible auto-corrélation spatiale des points d'échantillonnages mesurés.

B.3 Représentation spatiale

Les représentations cartographiques Afin de représenter graphiquement les disparités géographiques météorologiques en France métropolitaine, nous avons recours à :

- Un fond de carte sur lequel nous allons ajouter nos propres données.
- Un ensemble de données géolocalisées, qui peuvent s'appliquer soit à un point, soit à une zone, qui devront être jointes à notre fond de carte.
- Un système de coordonnées cartographiques et une projection qui déterminent la façon dont les données géographiques sont représentées

Un fond de carte permet de représenter graphiquement les caractéristiques d'une région géographique d'intérêt et peut être rangé dans deux catégories:

1. Les représentations vectorielles

Dans ce cas, les caractéristiques de la région d'intérêt sont représentées comme un ensemble de formes qui peuvent être des points ou des tracés (polygones, lignes etc.). Ce format est particulièrement adapté à la représentation de zones géographiques bien délimitées telle que les limites administratives (pays, régions, départements, communes).

2. Les représentations matricielles

Dans ce cas, les caractéristiques de la région d'intérêt sont stockées sous la forme d'un ensemble de pixels. A chaque pixel est associé une valeur, permettant de décrire les caractéristiques. Par exemple, la température peut être représentée sous la forme de pixels allant du bleu au rouge.

Pour notre part, nous avons combiné ces représentations pour créer un fond de carte plus complexe. En effet, nous avons combiné des *données vectorielles* représentant les limites des communes de France et un fond *matricielle* représentant les données de température moyenne et de précipitation.

La notion de référentiel cartographique Afin de représenter les données météorologiques souhaitées sur une carte de France, nous allons fusionner les données ponctuelles (latitude/longitude) issues du site *NCEI* avec les limites des communes de la France métropolitaine. Pour ce faire, il est impératif de s'assurer que les systèmes de coordonnées et la projection est la même.

Pour obtenir des fonds de cartes reprenant les limites des communes de France (sous forme de données vectorielles), nous avons utilisé les données *GEOFLA* mises à disposition par l'*IGN*. Ces données distribuées utilisent le *système cartographique Lambert-93* représentant le format légal en France. D'autre part, les données météorologiques sont récoltées à partir du site de *NCEI* et sont diffusées dans le *système WGS84*. Dans cette situation, il est impératif de faire matcher les deux systèmes de coordonnées utilisés (Lambert 93 et WGS84) sans quoi la représentation sera erronée.

Les systèmes cartographiques:

Si l'on considère un point sur la Terre, ses coordonnées dépendront d'un élément principal: l'ellipsoïde de référence, c'est-à-dire la forme de la sphère qui va représenter la Terre. En effet, la Terre n'étant pas une sphère parfaite, elle présente de nombreuses irrégularités selon les endroits considérés. Pour les

calculs, des formes simplifiées qui s'approchent de la forme de la Terre sont généralement utilisées: des ellipsoïdes de révolution. En fonction des objectifs d'utilisation et de la représentation voulue, le choix de l'ellipsoïde doit être adapté. Parmi les ellipsoïdes les plus utilisées, nous pouvons citer le WGS84 (celui utilisé dans le système WGS84) ou l'IAG-GRS80 (celui utilisé dans le système Lambert-93). Le deuxième élément indispensable est le point d'origine. Selon les cas, l'origine sera déterminée soit par la position du centre de l'ellipsoïde, soit par un point de référence à la surface de la Terre. Le regroupement de ces deux éléments [Ellipsoïde + Point d'origine] est désigné par le terme de système de coordonnées géographiques.

A ce système géographique est ajouté un paramètre supplémentaire qui permet d'obtenir une représentation cartographique. En effet, une carte est une représentation plane du globe terrestre, il convient donc d'appliquer une projection aux coordonnées pour pouvoir les afficher dans un plan. Pour ce faire, nous avons le choix de entre différents types de projections: une projection cylindrique, projection conique etc. qui donneront un rendu différent sur la carte (taille et forme différentes des pays).

L'ensemble [Ellipsoïde + Point d'origine + Projection] est désigné par le terme système de référence spatiale (ou SRS). Il est référencé sous la forme d'un code EPSG à 4 chiffres. Ainsi, des coordonnées en latitude/longitude doivent toujours être renseignées avec le triplet [Ellipsoïde + Point d'origine + Projection] utilisé. Les principaux codes EPSG à connaître sont le EPSG:4326 et le EPSG:2154 respectivement basés sur le système WGS84 et le système Lambert-93.