

Mémoire présenté devant l'ENSAE Paris
pour l'obtention du diplôme de la filière Actuariat
et l'admission à l'Institut des Actuares
le 15/03/2021



Par : **Silvia Bucci**

Titre : **Étude et implémentation de techniques d'analyse de
sensibilité dans les modèles de tarification Non-Vie.
Application à la tarification à l'adresse.**

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

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

Membres présents du jury de la filière
Wissal SABBAGH

Entreprise : ADDACTIS France 
Signature :
Guillaume ROSOLEK 


*Membres présents du jury de l'Institut
des Actuares*

Directeur du mémoire en entreprise :
Nabil RACHDI
Signature : 

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

Signature du responsable entreprise

Secrétariat:



Signature du candidat

Bibliothèque:



Résumé

Le marché de l'assurance a beaucoup évolué depuis les deux dernières décennies grâce à la *digitalisation* du parcours de souscription, à l'exploitation du *Big Data* et aux techniques d'apprentissage automatique (forêt aléatoire, xgboost, CART, réseaux neurones, ...) qui fournissent à l'assureur une connaissance très fine du risque.

Cependant, dans la littérature actuarielle, peu d'articles vont au-delà du modèle linéaire généralisé (GLM) et peu d'assureurs utilisent les méthodes de *machine learning* comme modèle de tarification pour leur sinistralité, bien qu'on leur reconnaisse un meilleur pouvoir prédictif.

Il y a principalement trois raisons à ce choix :

- le manque *d'interprétabilité* dû à l'effet boîte noire des modèles de *Machine Learning*,
- la *complexité* dans la mise en place de ces modèles, car on s'éloignerait du mécanisme de calcul de tarification,
- et une contrainte *éthique* puisque l'utilisation du machine learning dans la tarification peut conduire à une "hyper-personnalisation du risque".

L'alternative la plus utilisée pour améliorer les modèles linéaires généralisés est l'ajout manuel des termes d'interaction parmi les variables explicatives dans l'équation tarifaire suivi par un test de significativité. Cette pratique est toutefois très limitée car les interactions à tester peuvent être nombreuses et demander un coût algorithmique non négligeable.

Pour pallier ceci, une méthodologie de détection d'interactions plus robuste que celle traditionnelle a été développée dans le cadre de ce mémoire, dans une vision plutôt inclusive et collaborative entre les modèles de machine learning et des modèles linéaires généralisés. Ainsi, en supposant que l'interaction statistique est une manifestation de la complexité des modèles *black box*, l'optimisation d'un GLM simple à l'aide des interactions bénéficie des gains opérationnels et de la performance des modèles d'apprentissage automatique.

Les algorithmes et les outils de détection employés (indices de Sobol et indices de SHAP) se basent sur les domaines de l'*analyse de sensibilité* et du plus récent *Explainable Artificial Intelligence (XAI)*. Ils sont les clés de relecture des modèles prédictifs, puisqu'ils visualisent et quantifient les impacts des variables d'entrée sur la sortie selon un "juste" partage.

Le périmètre d'étude est un produit d'assurance Multirisques habitation (MRH) pour la garantie dégâts des eaux où les données sont individualisées à l'adresse.

Mots-clés : Analyse de sensibilité, XAI, Interactions en assurance Non-Vie, Indice de Sobol, SHAP, MRH Multirisques Habitation, Tarification à l'adresse.

Abstract

Over the last two decades, the insurance market has changed thanks to the *digitalization* of the underwriting process, the use of *Big Data* and automatic learning techniques (random drill, xgboost, CART, neural networks, ...) which provide the insurer with a very detailed knowledge of the risk.

However, in the actuarial literature, few articles go beyond the Generalized Linear Model (GLM), and few insurers use Machine Learning methods as a pricing model for the claims experience, although they could improve the quality of the predictions. There are mainly three reasons for this choice:

- the lack of *interpretability* due to the black box effect of the *Machine Learning* models,
- the *complexity* in the implementation of these models, because it would move away from the rate calculation mechanism,
- and an *ethical* constraint since the use of the Machine Learning in underwriting can lead to a "hyper-personalization of the risk".

The most commonly used alternative to improve generalized linear models is the manual addition of interaction terms among the explanatory variables in the pricing equation followed by a significance test. However, this practice is very limited because the interactions to be tested can be numerous and require a significant algorithmic cost.

To overcome this, a more robust interaction detection methodology than the traditional one has been developed in the framework of this thesis. Thus, assuming that the statistical interaction is a manifestation of the complexity of *black box* models, the optimization of a simple GLM using interactions takes advantage from Machine Learning good properties.

The detection tools used (Sobol indices and SHAP indices) are based on the fields of *Sensitivity Analysis* and the most recent *Explainable Artificial Intelligence (XAI)*. They are the keys to re-reading predictive models, since they visualize and quantify the impacts of input variables on the output according to a "fair" split.

The scope of study is a multi-risk home insurance product (MRH) for water damage coverage where the data is individualized at the address.

Key-words: Sensitivity analysis, XAI, Interactions in P&C Insurance, Sobol index, SHAP, Home insurance, Adress Home Pricing.

Note de synthèse

Introduction

La popularité de l'apprentissage automatique et de l'exploitation du big data a modifié la modélisation prédictive pour de nombreuses applications commerciales. Néanmoins, dans la littérature actuarielle peu d'articles vont au-delà du modèle traditionnel GLM et peu d'assureurs utilisent les méthodes de Machine Learning comme modèle de tarification pour la sinistralité.

Trois raisons majeures amènent l'actuaire à ne pas avoir recours à ces nouvelles méthodes :

1. Effet boîte noire : les modèles de Machine Learning, tels que Random Forest, Xgboost, Réseaux de neurones ne sont pas **interprétables**. Or, cette notion d'interprétabilité est obligatoire d'après la loi (RGPD Raison 71) : un individu a le droit d'avoir une explication face à la décision du modèle, en assurance, le prix de la couverture. Les modèles de tarification doivent être ainsi transparents et faciles à communiquer à tous ;
2. Modifier le principe de la calculatrice basé sur une **structure multiplicative** signifierait changer radicalement la façon de travailler pour la plupart des départements d'actuariat et du département informatique ;
3. L'utilisation du Machine Learning dans la tarification peut conduire à une **personnalisation du risque** extrême ou à une discrimination, au détriment de la mutualisation des risques par exemple, sous la forme de primes extrêmement élevées. L'assureur perdrait un rôle social : celui de créer une solidarité entre les assurés.

Aujourd'hui ces techniques ont ainsi un rôle encore très marginal dans la tarification (dans la sélection des variables par exemple) et très rarement elles sont utilisées en tant que modèle de tarification principal.

Dans une vision plutôt inclusive et collaborative entre les modèles de Machine Learning et des Modèles Linéaires Généralisés (GLM), nous nous servons des modèles de type boîte noire pour détecter des **interactions** parmi les variables, sans le spécifier au préalable, puis nous testerons leur pertinence dans un GLM simple. Ainsi, les termes d'interactions, étant une expression de la complexité d'un modèle, permettent aux GLMs de bénéficier de bonnes propriétés des modèles plus sophistiqués.

Données

Le cadre d'application est le produit MRH de l'assurance du particulier *Smart Home Pricing* pour la garantie *Dégâts des eaux*. Sa particularité est que la tarification utilise des données météorologiques, économiques, climatiques, démographiques hyper-individualisées (à l'adresse et même au bâtiment) qui nécessitent la géolocalisation des biens assurés. Parmi les variables innovantes, nous utiliserons par exemple la présence de gel, le nombre de jours orageux ou le nombre d'artisans dans la commune.

Interactions et importance des variables

Par rapport à ce que la littérature propose sur la détection des interactions dans les modèles statistiques, notamment aux travaux de Antoine Guillot [\[8\]](#) en assurance Non-Vie, le mémoire étend la recherche et l'exploitation des effets combinés des variables d'entrées sur une sortie à deux domaines : l'analyse de sensibilité et l' *explainable artificial intelligence*.

GLM : un modèle interprétable, mais peu complexe

La détection de ces interactions par des méthodes innovantes a été motivée par une limite du GLM : ce modèle, malgré son haut degré d'interprétabilité, n'inclut pas des termes de degré supérieur (des polynômes multivariés), c'est-à-dire qu'il manque de la complexité engendrée par l'effet croisé de deux (ou plusieurs) variables.

Cependant, en tant que sophistication des modèles linéaires d'où ils héritent une structure linéaire, ils aboutissent à des prédictions disponibles sous forme de tableau, facilement traduisibles en plans tarifaires. C'est pour cela qu'aujourd'hui les GLMs, introduits par Nelder et Wedderburn (1972), sont la norme de l'industrie de l'assurance pour développer des modèles analytiques de tarification.

Il est possible en effet de décomposer la prédiction d'un GLM comme la somme des effets de chacune des variables du modèle :

$$g(\mathbb{E}(Y)) = \underbrace{\beta_0}_{\text{effet moyen}} + \underbrace{\beta_1 X_1}_{\text{effet variable } X_1} + \underbrace{\beta_2 X_2}_{\text{effet variable } X_2} + \dots + \underbrace{\beta_n X_n}_{\text{effet variable } X_n}$$

où g est la fonction de lien.

Selon les hypothèses du modèle, l'effet de chaque variable indépendante est constant quelque soit la valeur prise par les autres variables indépendantes.

Toutefois, l'effet de X_1 , ou de X_2 , ... ou de X_n peut varier en fonction des valeurs prises par une des autres variables indépendantes introduite dans le modèle. On dit dans ce cas qu'il y a une interaction entre ces deux variables.

Plus formellement, dans un modèle $Y = f(X_1, X_2, \dots, X_n)$, nous dirons qu'il y a une interaction entre X_1 et X_2 si l'effet marginal de X_1 , noté β , dépend de X_2 :

$$\frac{\partial Y}{\partial X_1} = \beta(X_2) \text{ où } \beta \text{ est une fonction.}$$

Méthode naïve d'ajout des interactions

Le modèle GLM n'inclut pas d'interactions statistiques, à l'exclusion de celle naturelle amenée par la fonction de lien.

Toutefois, on peut ajouter le terme croisé $X_1 * X_2$ à la main comme entrées du modèle :

$$g(\mathbb{E}(Y)) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_{n+1} X_1 * X_2$$

et puis tester sa significativité à l'aide du test du rapport de vraisemblance. En particulier, on teste l'hypothèse nulle : $H_0 : \beta_{n+1} = 0$

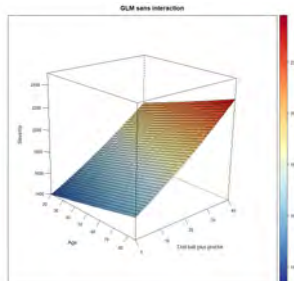


FIGURE 1 – Surface de réponse d'un GLM **sans interaction** dont les variables explicatives sont l'âge de l'occupant et la distance du bâtiment plus proche

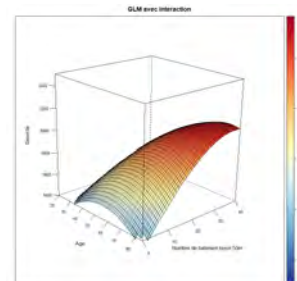


FIGURE 2 – Surface de réponse d'un GLM **avec l'interaction** entre l'âge de l'occupant et la distance du bâtiment plus proche

Cette procédure, sur laquelle les méthodes de sélection des variables *Forward/ Backward* se basent, a deux **limites principales** : calculatoire, car elle teste toutes les combinaisons possibles des variables et cela épuise la mémoire du logiciel en grande dimension ; et en terme de modèle d'interaction car dépendant du GLM. Par exemple, si la structure de lien utilisée ou la composante aléatoire n'est pas adaptée, des interactions pertinentes pourraient ne pas être significatives.

Méthode innovante de détection d'interactions

Au lieu de tester la significativité de l'ajout de chaque terme, nous préconisons une autre approche, introduisant des modèles qui utilisent intrinsèquement des interactions. En particulier, les modèles *Tree based* qui emploient comme modèle de base les arbres de régression ou de classification sont appréciés pour leur habilité à modéliser les effets d'interactions parmi les variables (Buchner et al., 2017 ; Schiltz et al., 2018).

Dans cette étude, les interactions seront détectées à partir de la **décomposition d'une quantité d'intérêt** :

- de la **variance** : les interactions seront quantifiées par les indices de Sobol. On s'appuiera sur la théorie de l'analyse de sensibilité et des principales contributions de Ilya Sobol [17]. La variance du modèle est décomposée en somme de fonctions de dimensions croissantes :

$$V := \mathbb{V}(Y) = \sum_{i=1}^n V_i + \sum_{1 \leq i < j \leq n} V_{ij} + \dots + V_{1\dots n}$$

où $\forall i, j = 1, \dots, n$

$$V_i = \mathbb{V}ar(\mathbb{E}(Y|X_i))$$

$$V_{ij} = \mathbb{V}ar(\mathbb{E}(Y|X_i, X_j)) - V_i - V_j$$

$$V_{1\dots n} = V - \sum_{i=1}^n V_i - \sum_{1 \leq i < j \leq n} V_{ij} - \dots - \sum_{1 \leq i_1 < i_{n-1} \leq n} V_{i_1 \dots i_{n-1}}$$

Chaque terme de cette somme, normalisée par la variance totale, est un indice de Sobol, dit *d'ordre k* si *k* est le nombre de variables par rapport auxquelles on conditionne l'espérance. La force de ces indices réside dans le fait que leur somme est égale à 1, donc ils sont très intuitifs.

Les interactions sont les indices de Sobol d'ordre 2 :

$$S_{i,j} = \frac{V_{i,j}}{V}, \forall i = 1, \dots, n \text{ avec } i \neq j \text{ et } n \text{ nombre de variables explicatives}$$

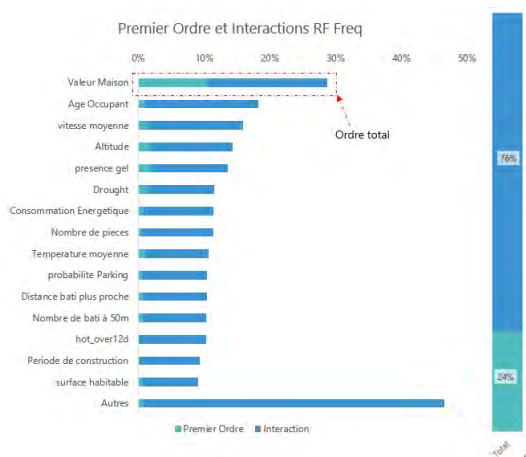


FIGURE 3 – Indice de Sobol du premier ordre et impacts des interactions dans le modèle de forêt aléatoire pour la fréquence : les effets principaux expliquent le 24% de la variance totale de la sortie du modèle, alors que les interactions contribuent au 76%. La variable la plus importante au sens de la décomposition fonctionnelle de Sobol est la valeur de la maison.



FIGURE 4 – Interaction (Indices de Sobol d'ordre 2) de la Valeur de la maison : interactions plus fortes dans le modèle de forêt aléatoire de fréquence.

- ou de la **valeur prédite** : les interactions seront quantifiées par les indices d'interaction SHAP. Cette notion est liée autant à la théorie des jeux qu'au plus récent domaine de l'Explainable Artificial Intelligence (XAI). L'idée est de décomposer la prédiction par la somme des importances des *n* variables explicatives pour pouvoir interpréter un modèle quelconque. Pour une observation $x = (x_1, \dots, x_n)$ la valeur prédite $f(x)$ se décompose de la façon suivante :

$$f(x) = \sum_{j=1}^n \underbrace{\varphi_j(\Delta^x)}_{\text{contribution de la variable explicative } X_j \text{ à la prédiction } f(x)} + \underbrace{\mathbb{E}(f(X))}_{\text{espérance des prédictions sur la base de apprentissage}}$$

$\varphi_j(\Delta^x)$ est la valeur SHAP pour la variable X_j du jeu coopératif ($N = \{X_1, \dots, X_n\}, \Delta^x$) :

$$\varphi_i(\Delta^x) = \sum_{S \in \mathcal{P}(N \setminus \{X_i\})} \frac{|S|!(p - |S| - 1)!}{p!} * [\Delta^x(S \cup \{X_i\}) - \Delta^x(S)]$$

où la quantité $[\Delta^x(S \cup \{X_i\}) - \Delta^x(S)]$ représente la contribution de la variable X_i à la coalition S (un élément de l'ensemble des parties de N).

Dans ce cadre, les **interactions sont les indices d'interaction SHAP** :

$$\Phi_{i,j} = \sum_{S \subseteq N \setminus \{X_i, X_j\}} \frac{|S|!(p - |S| - 2)!}{2(p - 1)!} \nabla_{i,j}(S)$$

avec $i \neq j$, n nombre de variables explicatives et

$$\begin{aligned} \nabla_{i,j}(S) &= f_x(S \cup \{X_i, X_j\}) - f_x(S \cup \{X_i\}) - f_x(S \cup \{X_j\}) + f_x(S) \\ &= f_x(S \cup \{X_i, X_j\}) - f_x(S \cup \{X_i\}) - [f_x(S \cup \{X_j\}) - f_x(S)] \\ f_x : \{X_{i_1}, \dots, X_{i_s}\} &\subset \{X_1, \dots, X_n\} \mapsto \mathbb{E}[f(X_1, \dots, X_n) | X_{i_1} = x_{i_1}, \dots, X_{i_s} = x_{i_s}] \\ \text{avec } \{i_1, \dots, i_n\} &\subset \{1, \dots, n\} \text{ et } s \in \{1, \dots, n\} \end{aligned}$$

Par analogie aux indices de Sobol du premier ordre et totaux, on peut séparer l'effet *individuel* d'une variable dans une prédiction de la valeur SHAP qui représente l'effet de cette variable dans toutes les coalitions auxquelles elle participe. Les effets interactions sont ainsi la partie de la valeur SHAP où l'on exclut l'effet de la variable seule :

$$\underbrace{\Phi_{i,i}}_{\text{effet individuel de la variable } X_i} = \underbrace{\varphi_i}_{\text{valeur SHAP de } X_i} - \underbrace{\sum_{j \neq i} \Phi_{i,j}}_{\text{Interactions de } X_i}$$

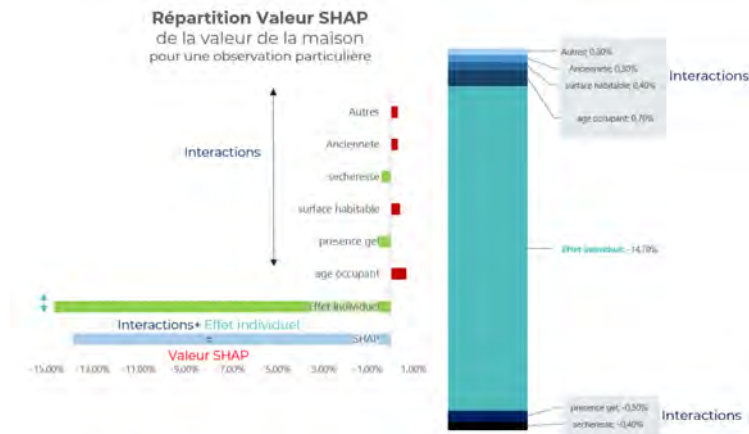


FIGURE 5 – Répartition de la valeur SHAP selon l'effet individuel de la variable "valeur de la maison" et ses effets d'interaction avec les autres variables explicatives. Le modèle considéré est la fréquence xgboost. La valeur SHAP de la maison pour cette observation particulière vaut -13,8%, c'est-à-dire que sa contribution diminue la prédiction (signe négatif). L'âge et l'ancienneté du client dans le portefeuille sont les seules interactions avec la valeur de la maison responsables de la hausse du nombre de sinistres.

Le besoin d'ouvrir ces modèles à l'aide de ces outils est dû à un problème d'explicabilité de certains modèles : si les modèles linéaires ou la régression logistique ont un nombre de paramètres limités et sont interprétés grâce à une structure additive *naturelle*, ce n'est pas le cas des modèles de type boîte noire tel que Random Forest ou xgboost.

L'avantage principal de passer par les domaines de l'analyse de sensibilité et de l'*Explainable Artificial Intelligence* est son agnosticité au modèle, qui permet de généraliser la représentation additive d'une quantité d'intérêt pour n'importe quel modèle. De plus, ces deux techniques, par la façon selon laquelle on les a introduites, sont complémentaires : les indices de Sobol sont *globaux*, c'est-à-dire qu'ils quantifient l'importance sur l'ensemble des sorties, alors que la valeur SHAP est *locale*, autrement dit elle exprime la contribution des coalitions des variables pour une observation spécifique.

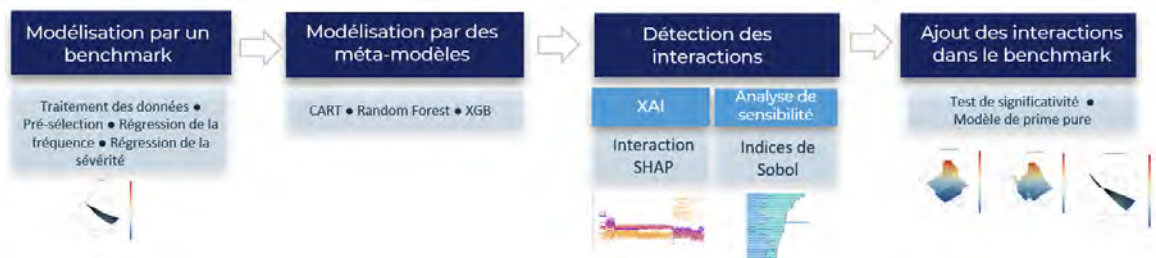
Application à la tarification à l'adresse

Étapes de l'étude

Afin de mesurer le gain que les interactions engendrent, nous avons comparé la performance des modèles GLM avec et sans ajout des termes d'interaction.

En particulier, dans un premier temps, nous avons construit un modèle simple, dit *Benchmark*, qui sera la base de comparaison, et des modèles plus avancés, dits *Métamodèles*. Ensuite, à partir de ces derniers, nous avons détecté les interactions des couples de variables. Enfin, nous avons ajouté ces termes dans le Benchmark et quantifié leur gain.

Grâce à son agnosticité au modèle de départ, cette méthodologie peut concerner n'importe quel modèle *Benchmark* :

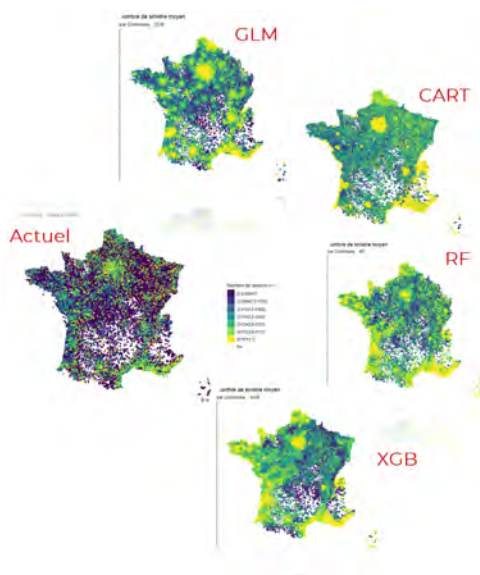


Effet de la Dimension et Pré-Sélection des variables

Les estimateurs des indices de Sobol et SHAP de type Monte Carlo (Strumbelj et Kononenko [21]) peuvent souffrir du fléau de la dimension et du type de variable (catégorielle, quantitatives, continues,...).

En effet, le temps d'exécution d'un algorithme non optimisé est $\mathcal{O}(n \times T(f) \times M)$, avec n : nombre de variables, $T(f)$: complexité du modèle et M : nombre d'itérations Monte Carlo. Afin d'utiliser des estimateurs robustes sans pourtant épuiser la mémoire vive du logiciel (R ou python), nous avons pré-sélectionné environ 40 variables explicatives, en réduisant la base utilisée dans certains cas à un sous-échantillon représentatif et/ou des estimateurs qui convergent plus rapidement, tel que *TreeShap*.

Construction des modèles de fréquence et sévérité avant l'ajout d'interactions



Nous souhaitons proposer à l'assureur un modèle linéaire généralisé amélioré à l'aide des interactions, pour ajouter de la complexité tout en gardant une structure tarifaire simple à mettre en place.

D'abord, nous avons calibré les GLMs de fréquence et sévérité.

Ensuite, nous avons paramétré trois modèles de Machine Learning : un arbre de régression, une forêt aléatoire (Random Forest) et un Extreme Gradient Boosting.

Ces derniers ne sont pas ceux retenus pour le calcul de prime pure, mais seront les modèles où l'on détectera les interactions.

Nous avons constaté que les quatre modèles de la fréquence segmentent le risque de façon différente : la comparaison sur la carte de France du nombre de sinistres moyen prédits, à gauche, montre que les modèles de Machine Learning prennent en compte le risque géographique de façon plus fine et lissée, surtout dans la zone rurale.

Détection et ajouts des interactions

CART : un premier modèle sophistiqué interprétable

L'algorithme *Glouton* utilisé par le modèle CART (*Classification and Regression Tree*) est par construction un modèle interprétable. Il construit un arbre en plusieurs étapes en découpant à chaque étape la population en deux groupes, maximisant la variance inter-classes (les groupes sont des sous-ensembles dont des sorties sont

les plus dispersées possibles).

Chaque division de la population décrit une interaction introduite par le modèle :

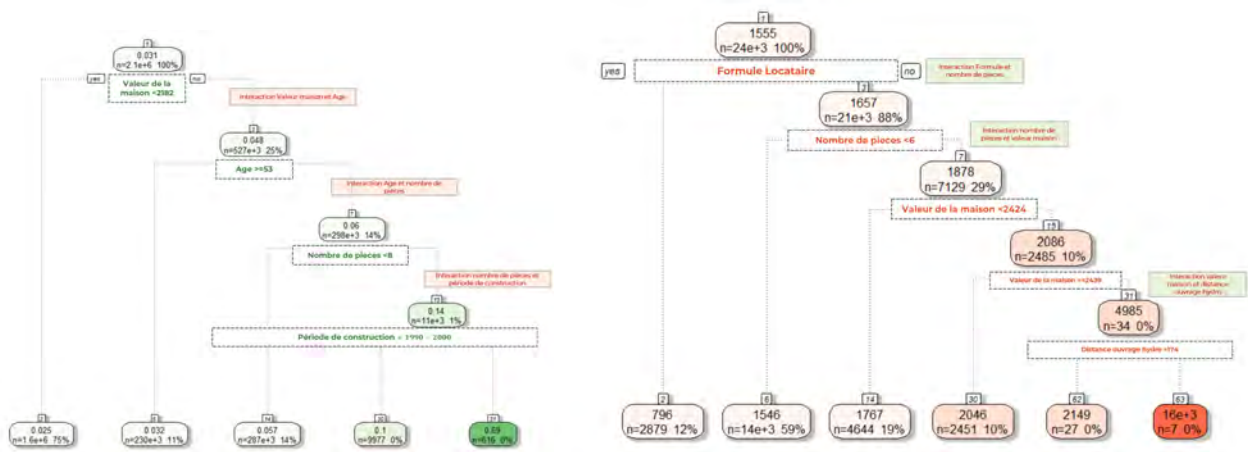


FIGURE 6 – CART sur la fréquence (gauche) et sur la sévérité (droite) : les interactions sont détectées par les étapes de partition de la base d'apprentissage. Par exemple, dans la deuxième étape du modèle de fréquence, on introduit l'interaction entre la valeur de la maison (division de la racine) lorsqu'elle est supérieure à 2182 et l'âge.

Interactions SHAP

Nous avons ensuite calculé les interactions SHAP pour les modèles de forêt aléatoire et xgboost pour détecter leurs interactions. Afin de ne pas introduire dans le modèle final des interactions artificielles ou de très petits effets, les interactions ont été visualisées graphiquement.

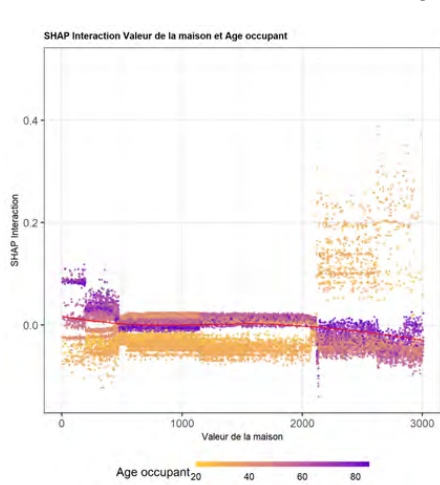


FIGURE 7 – Visualisation des interactions par individu de la base de validation, en abscisse la valeur de la maison et la couleur selon l'âge de l'occupant. L'intensité moyenne (en valeur absolue) est de 2,1%, la plus forte du portefeuille. La portée de cette interaction est à la fois locale et à la fois globale. Nous remarquons que la contribution des profils seniors habitant dans les zones décentrés (valeur de la maison inférieure) est positive, alors que les jeunes tendent à déclarer plus de sinistres dans des logements qui valent plus (des maisons en centre ville par exemple).

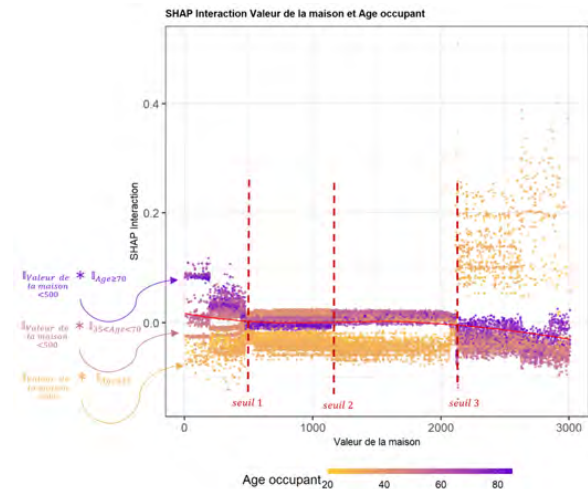


FIGURE 8 – Création des termes d'interactions pour le GLM à partir des graphes d'interactions SHAP.

Un choix naturel pour intégrer les interactions locales dans le GLM est de couper le domaine de définition des variables en interaction en intervalles. Pour le CART ces intervalles sont définis par construction, alors que pour les interactions SHAP, nous avons défini des seuils selon les graphes d'interaction (fig. 13.8).

Nous avons remarqué que les interactions détectées avec le modèle de *random forest* sont très faibles par rapport à celles introduites par *xgboost* :

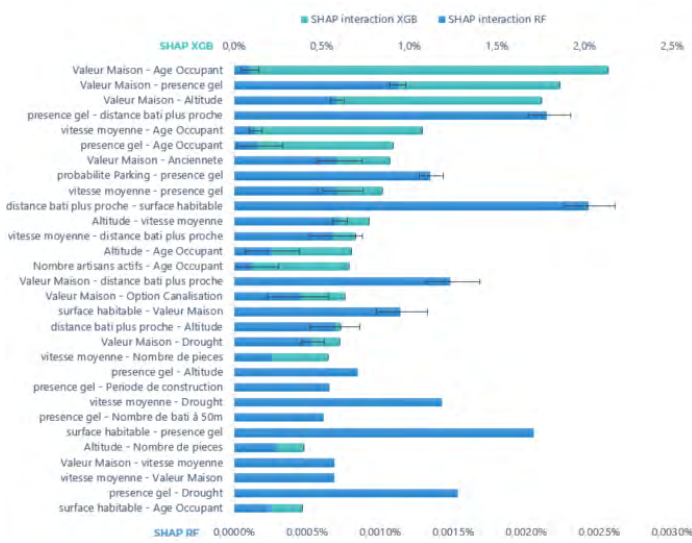


FIGURE 9 – Comparaison des effets d’interactions SHAP pour la fréquence : le modèle *xgboost* a des valeurs beaucoup plus importantes que le modèle de forêt aléatoire et dans la prédiction il prend en compte les interactions de la valeur de la maison avec l’âge, la présence du gel, l’altitude, ainsi que des interactions moins fortes. Le modèle *random forest* quant à lui prend en compte dans la prédiction les interactions de la distance du bâtiment le plus proche et de la surface habitable.

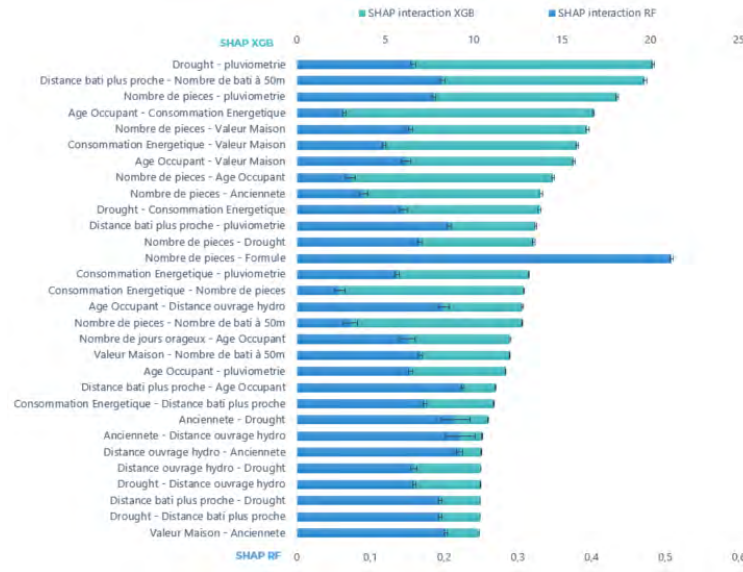


FIGURE 10 – Comparaison des effets d’interactions SHAP pour la sévérité : les interactions de forêt aléatoire sont très faibles (moyenne 0,17) par rapport à celles intégrées par *xgboost* (moyenne à 5,9). Dans la prédiction, *xgboost* prend en compte les interactions de la valeur de la maison avec l’âge, le nombre de pièces, la consommation énergétique ainsi que des interactions entre variables climatiques (pluviométrie et sécheresse).

Interactions selon l’analyse de Sobol

Les indices de Sobol d’ordre 2 identifient des interactions globales.

Au vu de ces indices, les interactions de la fréquence sont beaucoup plus fortes que celles de sévérité. Les modèles sous-jacents de fréquence sont ainsi plus complexes au sens de l’analyse de sensibilité de Sobol.

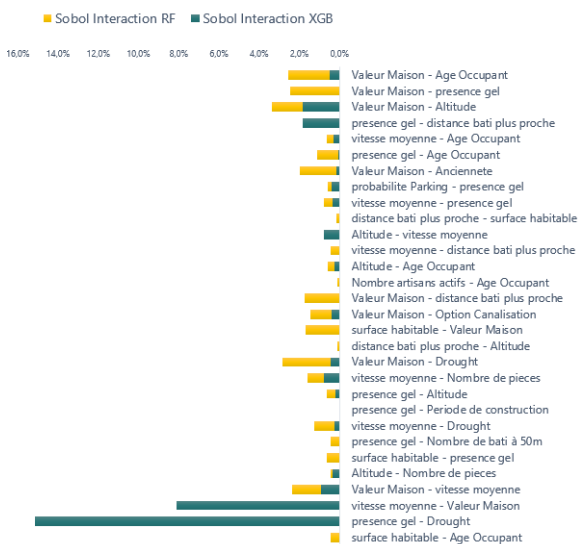


FIGURE 11 – Visualisation des interactions à l’ordre 2 pour les modèles de fréquence. Les interactions de la valeur de la maison sont les plus fortes, suivies par celles parmi les variables géographiques et climatiques (quantité de pluie, sécheresse, altitude, vitesse du vent).

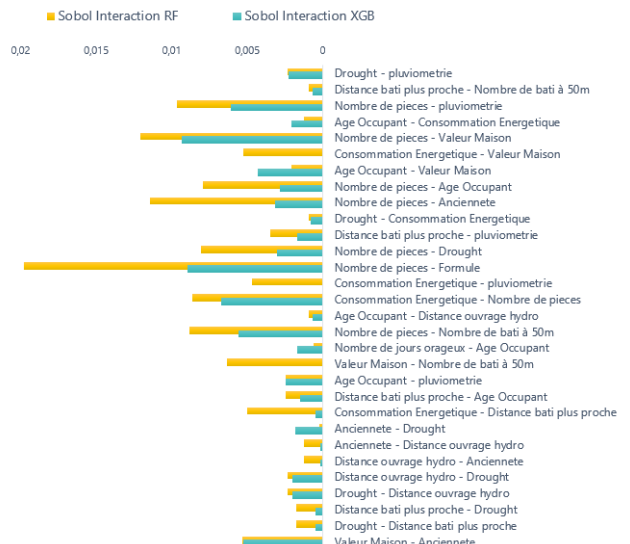


FIGURE 12 – Visualisation des interactions à l’ordre 2 pour les modèles de sévérité. Les interactions du nombre des pièces avec la valeur de la maison, la quantité de pluie, l’âge de l’occupant, la formule et la sécheresse sont les plus significatives, suivies par celles de la valeur de la maison.

À l'aide de la théorie des graphes, nous avons visualisé les interactions au-delà d'un seuil :
Visualisation des interactions selon l'analyse de Sobol
 Modèle: Fréquence-RF seuil: 0.015

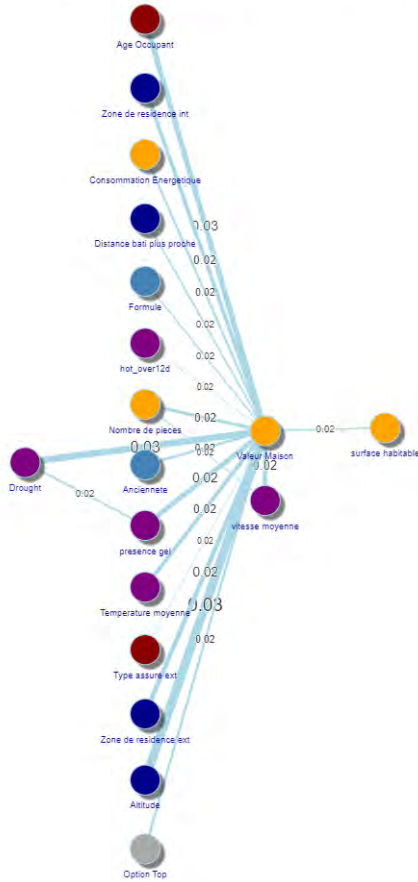


FIGURE 13 – Visualisation des interactions supérieures à 1.5% pour le modèle de fréquence de forêt aléatoire. La valeur de la maison est la variable qui interagit le plus avec les autres covariables.

Visualisation des interactions selon l'analyse de Sobol
 Modèle: Severite-RF seuil: 0.01

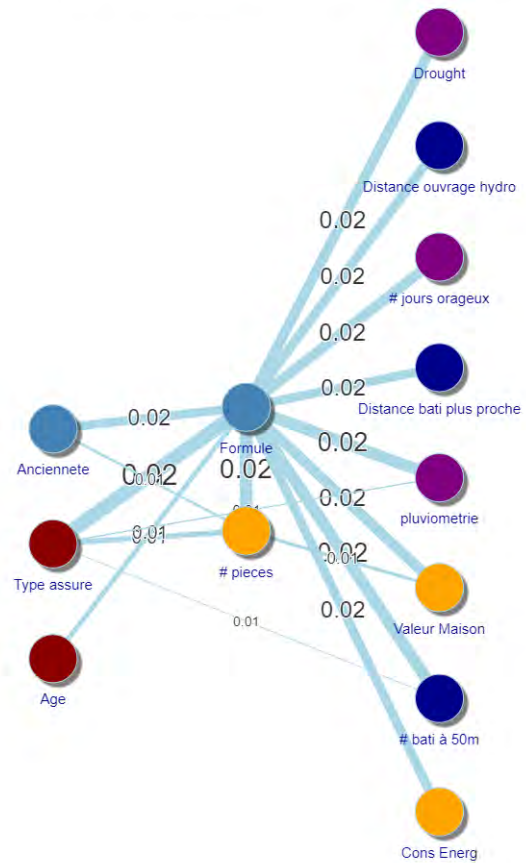


FIGURE 14 – Visualisation des interactions supérieures à 1% pour le modèle de sévérité de forêt aléatoire. La Formule est la variable qui interagit le plus avec les autres covariables.

Étant donné que l'interaction de Sobol est une mesure globale sur tout le domaine de définition des variables, nous avons fait le choix d'introduire dans le GLM simple des termes polynomiaux du deuxième degré.

Comparaisons des modèles

Comparaison entre les GLMs avec interactions et le GLM simple.

Enfin, nous avons comparé les indicateurs de performance des modèles composés d'interactions détectées dans le Benchmark initial..

| Fréquence | | | | | | | | | | Sévérité | | | | | | | | | | |
|-------------------------------|----------------|---------|---------|---------|--------|--------|--------|------------------|-------------------|-------------------------------|----------------|--------|--------|---------|--------|--------|--------|----------------|-------------------|--------|
| Base 100 | | | | | | | | | | Base 100 | | | | | | | | | | |
| Modele | Ecart Pred avg | MSE | RMSE | MAE | Q2 | GINI | AIC | Deviance Poisson | Residual Deviance | Modele | Ecart Pred avg | MSE | RMSE | MAE | Q2 | GINI | AIC | Deviance Gamma | Residual Deviance | |
| Benchmark | 100,00 | 100,000 | 100,000 | 100,000 | 100,00 | 100,00 | 100,00 | 100,000 | 100,00 | Benchmark | 100,00 | 100,00 | 100,00 | 100,000 | 100,00 | 100,00 | 100,00 | 100,00 | 100,00 | 100,00 |
| Clouton CART | 94,25 | 99,872 | 99,936 | 100,045 | 102,33 | 100,20 | 99,93 | 99,962 | 99,91 | Clouton CART | 98,07 | 98,994 | 98,997 | 98,998 | 100,55 | 102,54 | 99,98 | 100,58 | 99,52 | 99,91 |
| SHAP RF | 93,29 | 99,993 | 99,996 | 100,076 | 100,87 | 100,23 | 99,97 | 99,998 | 99,91 | SHAP RF | 97,07 | 99,983 | 99,992 | 100,252 | 102,06 | 100,25 | 99,90 | 99,982 | 99,83 | 99,83 |
| SHAP XGB | 79,07 | 99,983 | 99,992 | 100,252 | 102,06 | 100,25 | 99,90 | 99,982 | 99,83 | SHAP XGB | 98,89 | 98,998 | 98,999 | 100,131 | 100,93 | 110,73 | 99,98 | 100,98 | 99,30 | 99,30 |
| SHAP Toutes les interactions | 78,99 | 99,984 | 99,992 | 100,252 | 101,98 | 100,25 | 99,90 | 99,981 | 99,79 | SHAP Toutes les interactions | 98,89 | 98,998 | 98,999 | 100,131 | 100,93 | 110,73 | 99,98 | 100,98 | 99,30 | 99,30 |
| Sobol RF | 87,20 | 99,987 | 99,993 | 100,160 | 101,62 | 99,55 | 99,93 | 100,017 | 99,87 | Sobol RF | 97,54 | 98,999 | 98,999 | 100,041 | 101,05 | 103,25 | 100,00 | 100,32 | 99,70 | 99,70 |
| Sobol XGB | 98,22 | 99,989 | 99,995 | 100,021 | 101,30 | 99,97 | 99,93 | 99,999 | 99,91 | Sobol XGB | 99,79 | 98,999 | 98,999 | 100,172 | 101,05 | 101,72 | 100,00 | 100,51 | 99,70 | 99,70 |
| Sobol Toutes les interactions | 86,66 | 99,987 | 99,994 | 100,167 | 101,55 | 99,50 | 99,93 | 100,014 | 99,87 | Sobol Toutes les interactions | 97,50 | 98,999 | 99,000 | 100,029 | 101,07 | 103,62 | 100,00 | 100,30 | 99,70 | 99,70 |
| Toutes les Interactions | 79,37 | 99,987 | 99,994 | 100,251 | 101,52 | 100,20 | 99,93 | 100,026 | 99,74 | All | 96,21 | 99,990 | 98,995 | 100,050 | 100,15 | 117,06 | 99,95 | 102,00 | 98,94 | 98,94 |

FIGURE 15 – Métriques d'évaluation des modèles de fréquence et sévérité.

Parmi les modèles de fréquence construits, celui contenant toutes les interactions SHAP, sur l'ensemble de la base de test, se rapproche de plus à la moyenne observée et il est le plus discriminant (Gini plus élevé). Le

modèle dérivé du CART se rapproche de plus en moyenne aux observations (MSE plus faible) et il est le plus prédictif (Q2 plus élevé).

Les modèles de la sévérité avec les interactions sont moins performants que les modèles de fréquence : les interactions détectées pourraient avoir une nature plus complexe que les polynômes du deuxième degré ou ne pas être assez localisées. Le modèle avec toutes les interactions prédit une moyenne sur le portefeuille plus proche de celle observée que le Benchmark, il est le plus discriminant et réduit l'AIC.

Plus concrètement, l'ajout des interactions correspond à une modification de la surface de réponse du modèle GLM : nous montrons ici uniquement les interactions entre l'âge et la valeur de la maison, mais les résultats sont équivalents pour les autres interactions.

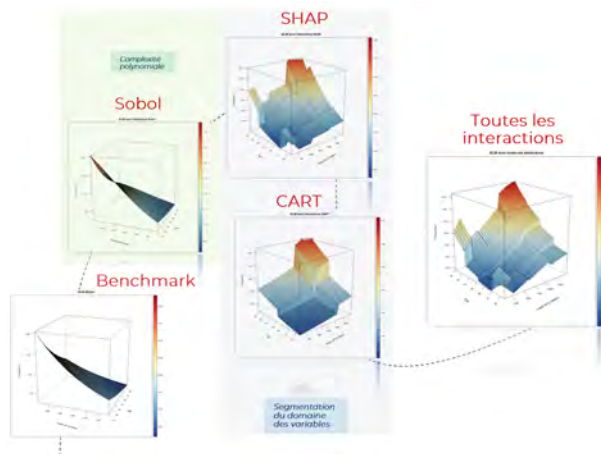


FIGURE 16 – Visualisation des surfaces de réponse des GLMs sans et avec interactions pour un modèle à deux variables (*Âge* et *valeur de la maison*) : la fonction de lien donne une structure exponentielle au modèle, ensuite les termes des interactions de Sobol confèrent une complexité avec l'ajout d'un polynôme du deuxième degré à la formule tarifaire et enfin les termes des interactions SHAP et CART partitionnent le domaine de définition des variables. Par rapport au modèle de Benchmark, le risque est beaucoup plus segmenté.

Conclusion

Dans le cadre de cette étude, une analyse de sensibilité a été menée sur des données à l'adresse pour améliorer le modèle de tarification de la garantie *Dégâts des eaux*. Elle a permis d'ajouter de la *complexité* tout en gardant une structure analytique, transparente et interprétable qui s'intègre parfaitement au processus de tarification traditionnel des organismes d'assurance.

Afin de respecter ces critères, nous avons reconduit le problème d'optimisation tarifaire au problème de détection et intégration des interactions parmi les variables, l'interaction étant une expression de la complexité du modèle.

Nous avons d'abord détecté les interactions des modèles plus sophistiqués, dits de type *boîte noire* à cause de leur structure prédictive non accessible, en nous appuyant sur des concepts de la théorie de jeux et de l'analyse de sensibilité selon Sobol.

Nous avons cherché un compromis entre la rapidité de calcul (parallélisation des tâches, passage par des métamodèles), la complexité (des modèles et des données), l'interprétabilité (des modèles de tarification) et l'agnosticité aux modèles. Les outils de l'analyse de sensibilité tels que les indices de Sobol et les indices de SHAP (une relecture plus récente selon la théorie des jeux) ont été choisis par les bonnes propriétés de convergence de ces estimateurs et par le type d'information apporté.

D'un côté, les indices de Sobol de l'ordre deux, nous informent de la part de la variance totale due à chacune des interactions, de l'autre les indices d'interaction SHAP déterminent dans quelle direction l'interaction a un impact sur la prédiction (si elle est à la hausse ou à la baisse à cause de l'interaction) et selon quelle intensité.

Ensuite nous avons intégré les interactions détectées localement et globalement en ajoutant des termes polynomiaux dans le modèle GLM de départ.

Ces termes améliorent le modèle GLM simple selon des métriques d'évaluation habituelles (MSE, RMSE, MAE, Q2, Gini, déviance, AIC) avec un gain entre 0.03% et 17%.

Plus généralement, cette méthodologie de détection ne se limite pas à l'optimisation tarifaire : son agnosticité aux modèles permet de l'appliquer en toute généralité à n'importe quel modèle complexe ou non. De plus, ce parcours nous a fait expérimenter une nouvelle approche portant sur la collaboration du Machine Learning et des modèles linéaires généralisés (GLM) : dans la littérature actuarielle, on a tendance à comparer la performance

des modèles plus complexes au GLM classique, ou à se servir de l'une ou de l'autre type de modèle pour deux études séparées.

Par ailleurs, la perception du risque géographique héritée par les modèles plus complexes et par les données à maille fine (à l'adresse et au bâtiment) est plus fine.

Ces données sont particulièrement adaptées à ce cadre d'application car la base de données est de grande dimension (300 variables) et l'analyse bivariée pour examiner toutes les interactions possibles peut être fastidieuse et lente.

De plus, l'inclusion des variables géographiques à maille très fine challenge les méthodes de zonage actuelles pour la prise en compte du risque spatial.

Limites de l'étude et ouvertures

Les **limites** principales de cette étude sont les suivantes :

- la **nature de la complexité des données et des modèles**. Les estimateurs de SHAP et des indices de Sobol ont permis de détecter les interactions parmi les variables et nous avons intégré ces relations à l'intérieur de la structure d'un GLM simple en ajoutant des termes polynomiale. Toutefois, la nature de ces liens peut être plus complexe (l'ajout des fonctions indicatrices n'est pas suffisant) ou très localisée (par opposition aux indices de Sobol qui sont globaux) : dans ce cas, l'ajout des termes peut se révéler non significatif selon le test du rapport de vraisemblance. Le lissage avec les techniques de *smoothing* ou *spline* pourrait améliorer la performance des modèles.
- Le **risque de propagation de l'erreur**. La construction des métamodèles de type CART, xgboost et forêt aléatoire peut propager son erreur de prédiction à l'estimation des indices de Sobol ou de SHAP et introduire des interactions artificielles, spécialement en présence de colinéarité. Pour limiter leur intégration dans le modèle de départ, il faut valider les interactions avant de les exploiter dans le modèle. De plus, la dimension des données peut ralentir la convergence de ces estimateurs et dans le cas où l'importance est très petite les estimations peuvent être négatives (pour les indices Sobol).

La théorie liée à la valeur SHAP qui se fonde sur l'explicabilité du modèle n'est pas nouvelle dans le secteur de l'assurance : c'est aussi un outil de pilotage tarifaire, de compréhension du portefeuille et de prévention au risque de sinistralité. Les références sous les logiciels **R** et **python** sont par ailleurs abondantes et sophistiquées. Son extension à la détection d'interactions est un des **éléments innovants** que l'on a voulu transmettre dans ce mémoire.

L'analyse de Sobol, quant à elle, a été un autre challenge pour ce type de problématique : à la différence des contextes où elle s'applique d'habitude (ingénierie aéronautique, finance,...), où l'on *pilote* une base, nous disposons des réalisations des variables sans connaître leur distributions et on a plutôt *subi* une base.

Nous nous sommes restreint à la sinistralité attritionnelle, mais dans la continuité des études d'analyse de sensibilité, il serait intéressant de calculer ces indices sur des quantiles différents et intégrer ainsi les interactions sur les **sinistres graves**.

Remerciements

Je tenais à remercier Guillaume ROSOLEK pour l'opportunité qu'il m'a offerte en intégrant son équipe et pour les insights actuariels partagés le long de mon stage.

J'exprime toute ma reconnaissance à mon directeur de mémoire, Nabil RACHDI, pour la proposition du sujet très passionnant, son encadrement et ses enseignements très enrichissants.

La réalisation de ce mémoire a été possible grâce au support de toute l'équipe P&C Pricing & Data de Addactis France, que je remercie pour les belles expériences dont je peux bénéficier chaque jour. Je suis sincèrement reconnaissante à Pierre CHATELAIN, François-Xavier CHAMOULAUD et Victoria DELAUDAUD pour les relectures et les conseils de rédaction.

Je voudrais remercier également mon tuteur pédagogique, Wissal SABBAGH, pour sa disponibilité et ses conseils.

Je tiens à remercier toutes les personnes qui ont contribué de loin à la réalisation de mon mémoire : Raphaëlle Lorenzo, Hassan El Hadraoui, Léa Nefussi et tous mes anciens collègues de CHUBB. Je remercie notamment Manhirath Amoussa, Nnana Yackson et Badreddine Ramchoun pour leur soutien.

J'adresse enfin mes sincères remerciements à tous les professeurs que j'ai pu croiser pendant mon parcours académique à l'ENSAE et à l'Université Gustave Eiffel. En particulier, Florence Merlevède, Thierry Jeantheau, Damien Lambertson, Romouald Elie et Caroline Hillairet dont le soutien et les enseignements ont été fondamentaux pour ma vie professionnelle.

Table des matières

| | | |
|------------|--|-----------|
| I | Contexte Assurantiel | 3 |
| 1 | Spécificités du marché de l'assurance MRH | 4 |
| 1.1 | Brève histoire de l'assurance habitation | 4 |
| 1.2 | Le contrat MRH Multirisque Habitation | 5 |
| 2 | Enjeux et contraintes en Assurance Habitation | 8 |
| 2.1 | Le parcours de souscription habitation | 8 |
| 2.2 | Contraintes techniques | 9 |
| 2.3 | Apport du mémoire | 10 |
| II | Principes de Tarification IARD | 12 |
| 3 | Rappel de Tarification en Assurance Non - Vie | 13 |
| 3.1 | Caractéristiques et principes de l'assurance Non-Vie | 13 |
| 3.2 | Modélisation de la prime pure | 17 |
| 3.3 | Le Modèle Linéaire Généralisé (GLM) | 18 |
| 3.4 | Critères de tarification en Habitation | 22 |
| 4 | Effet d'interaction en Tarification Non - Vie | 24 |
| 4.1 | Définition d'une interaction statistique | 24 |
| 4.2 | Comment on détecte et on intègre les interactions statistiques aujourd'hui | 25 |
| 5 | Des modèles sophistiqués pour apprendre les interactions entre les variables | 29 |
| 5.1 | CART : Arbres de classification et régression | 31 |
| 5.2 | Agrégation de modèles | 33 |
| 5.2.1 | Bagging pour bootstrap aggregating | 33 |
| 5.2.2 | Forêts aléatoires | 34 |
| 5.2.3 | Boosting | 34 |
| 5.2.4 | Gradient tree boosting | 35 |
| 5.2.5 | Extrem Gradient Boosting | 37 |
| III | Analyse de sensibilité et XAI | 39 |
| 6 | Analyse de sensibilité | 41 |
| 6.1 | Comment évaluer l'impact des entrées aléatoires sur la sortie d'un modèle mathématique | 41 |
| 6.2 | Modèles à entrées indépendantes | 45 |
| 6.2.1 | Indices de sensibilité pour modèles linéaires et/ou monotones | 45 |
| 6.2.2 | Méthodes basées sur des tests statistiques | 46 |
| 6.2.3 | Méthodes basées sur la Décomposition fonctionnelle de la variance | 47 |
| 6.2.4 | Estimation des indices de Sobol | 49 |
| 6.2.5 | Illustration d'un modèle à entrées indépendantes : le modèle Ishigami | 54 |
| 6.2.5.1 | Approximations des indices de Sobol | 55 |
| 6.2.5.2 | Approximation en passant par un métamodèle | 58 |
| 6.2.6 | Limites du glm vis-à-vis des interactions et pourquoi envisager une analyse de sensibilité . | 61 |
| 6.3 | Modèles à entrées dépendantes | 61 |

| | | |
|-----------|---|------------|
| 7 | Explainable artificial intelligence | 63 |
| 7.1 | Classe des méthodes <i>Additive Feature Attribution</i> | 63 |
| 7.2 | Valeur de Shapley dans la théorie de jeux | 66 |
| 7.3 | Valeur Shapley dans un modèle prédictif | 67 |
| 7.3.1 | SHAP zone : Sensibilité locale d'une donnée de sortie selon la valeur de Shapley | 67 |
| 7.3.2 | SHAP Feature Importance | 70 |
| 7.3.3 | Effets d'interaction SHAP | 70 |
| 7.4 | Estimation de la valeur Shapley | 72 |
| 7.5 | Illustration dans le modèle Ishigami | 81 |
| IV | Application à la tarification à l'adresse | 85 |
| 8 | Contexte d'étude : la tarification à l'adresse | 87 |
| 8.1 | Une nouvelle approche au risque | 88 |
| 8.2 | Description de la base des données | 92 |
| 8.3 | Impacts de la Tarification à l'adresse | 97 |
| 8.4 | Intérêt par rapport à un zonier | 97 |
| 9 | Présélection des variables | 98 |
| 9.1 | Traitement des données | 98 |
| 9.2 | Pre-sélection des variables | 100 |
| 9.2.1 | Présélection non supervisée des variables | 100 |
| 9.2.2 | Présélection supervisée des variables | 111 |
| 10 | Modélisation de la sinistralité | 113 |
| 10.1 | Modèle de fréquence | 113 |
| 10.2 | Modèle de Sévérité | 118 |
| 10.3 | Comparaison des modèles | 119 |
| 11 | Pratique de l'analyse de sensibilité et de l'XAI pour la détection des interactions dans la Tarification à l'adresse | 122 |
| 11.1 | Analyse de sensibilité | 122 |
| 11.1.1 | Spécificités du contexte assurantiel | 122 |
| 11.1.2 | Méthodologie | 123 |
| 11.2 | XAI | 128 |
| 11.2.1 | Méthodologie | 129 |
| 11.3 | Comparaisons | 138 |
| 12 | Intégration des interactions dans le modèle de prime pure : résultats et ouvertures | 141 |
| 12.1 | Ajout d'interactions | 141 |
| 12.2 | Comparaison entre le GLM simple et le GLM avec interactions | 143 |
| 12.2.1 | Comparaison des métriques | 143 |
| 12.2.1.1 | Surfaces de réponse des modèles utilisés | 145 |
| 12.2.1.2 | Comparaison du risque géographique | 146 |
| 12.2.2 | Arbre de décision | 147 |
| V | Conclusion | 149 |
| 13 | Conclusion | 150 |
| | Annexes | 163 |
| A | Spécificités du marché de l'assurance MRH | 164 |
| A.1 | Le marché français de l'assurance Habitation | 164 |
| A.2 | Réglementation du marché MRH | 169 |
| B | Pré-selection des variables | 172 |
| B.0.1 | Algorithme dirigé par la force | 172 |

| | | |
|----------|--|------------|
| C | Explainable artificial intelligence | 173 |
| C.0.1 | LIME | 173 |
| D | Tarification à l'adresse | 175 |
| D.1 | Une parenthèse sur la confiance des données | 175 |
| D.2 | Analyse descriptive | 178 |
| E | Méthodes de pénalisation pour la sélection finale | 185 |
| E.1 | Sélection via régression Lasso | 185 |
| F | Modèle de tarification par Machine Learning | 187 |
| F.1 | Optimisation des modèles de fréquence | 187 |
| F.2 | Optimisation des modèles de sévérité | 191 |
| G | Pratique de l'analyse de sensibilité | 194 |
| H | Résultats | 196 |

Introduction

Le marché de l'assurance a été bouleversé pendant les deux dernières décennies par des événements qui ont remis en question le système entier. D'un côté, la *digitalisation* a affecté le parcours de souscription de la plupart des assureurs en proposant des devis partiellement ou totalement en ligne. La *réglementation* en assurance, quant à elle, a renforcé le pouvoir du client (avec la loi Hamon par exemple), en l'autorisant à rompre son contrat d'assurance à n'importe quel moment une fois la première année écoulée. Enfin, l'exploitation du *Big Data* et des techniques d'apprentissage automatiques, connues pour réduire l'erreur de prédiction, permettrait à l'actuaire une connaissance plus fine du risque et de la clientèle.

Cependant, dans la littérature actuarielle, peu d'articles vont au-delà du modèle linéaire généralisé (GLM), et peu d'assureurs utilisent les méthodes de *machine learning* comme modèle de tarification pour leur sinistralité. En effet, trois contraintes majeures empêchent l'actuaire d'avoir recours à ces nouvelles méthodes :

- une contrainte *d'interprétabilité* due à l'effet boîte noire : la plupart des modèles de machine learning ne sont pas interprétables. Or, cette notion est obligatoire d'après la loi puisqu'un individu a le droit d'avoir une explication face à la décision du modèle ;
- une contrainte *opérationnelle* : avec la mise en place de ces modèles non linéaires, on s'éloigne de la structure multiplicative du modèle linéaire généralisé (GLM) qui est intuitive et pratique ;
- une contrainte *éthique* : l'utilisation du machine learning dans la tarification peut conduire à une "personnalisation du risque" extrême ou à une discrimination, au détriment de la mutualisation des risques par exemple, sous la forme de primes extrêmement élevées.

Ainsi, ces techniques ont aujourd'hui un rôle encore très marginal dans la tarification (dans la sélection des variables par exemple) et sont très rarement utilisées en tant que modèle de tarification principale.

L'alternative la plus utilisée pour améliorer les modèles linéaires généralisés est l'ajout manuel de termes d'interaction parmi les variables explicatives dans l'équation tarifaire, suivi par un test de significativité. Cette pratique est toutefois très limitée car les interactions à tester peuvent être nombreuses et demander un coût algorithmique non négligeable.

Pour pallier ceci, une méthodologie de détection d'interactions plus robuste que la méthode traditionnelle a été développée dans le cadre de ce mémoire, dans une vision plutôt inclusive et collaborative entre les modèles de machine learning et des GLMs. Ainsi, en supposant que l'interaction statistique est une manifestation de la complexité des modèles *black box*, l'optimisation d'un GLM simple à l'aide des interactions bénéficie des gains opérationnels des modèles d'apprentissage automatique.

En d'autres termes, les modèles de machine learning incluent déjà les interactions les plus fortes parmi les variables, mais elles ne sont pas lisibles pour l'effet boîte noire.

Par rapport à ce que la littérature propose sur la détection des interactions dans les modèles statistiques, notamment les travaux de Antoine Guillot[8] en assurance Non-Vie, le mémoire étend la recherche et l'exploitation des effets combinés des variables d'entrées sur une sortie à deux domaines : l'analyse de sensibilité et l'*Explainable Artificial Intelligence*.

Les outils de détection employés, à savoir les indices de Sobol et indices de SHAP, sont les clés de relecture des modèles prédictifs, puisqu'ils visualisent et quantifient les impacts des variables d'entrée sur la sortie selon un "juste" partage.

Le mémoire s'inscrit donc dans l'optimisation tarifaire d'un modèle GLM, pour garder une structure simple et intuitive, tout en bénéficiant du pouvoir prédictif des modèles *Black Box*.

Ce mémoire propose dans un premier temps une méthodologie de détection d'interactions et dans une deuxième

partie on évalue le gain opérationnel sur le modèle de prédiction traditionnel.

Le cadre d'application est le produit Multirisque habitation (MRH) de l'assurance du particulier *Smart Home Pricing* pour la garantie *Dégâts des eaux*. Sa particularité est que la tarification utilise des données météorologiques, économiques, climatiques, démographiques à mailles fines : jusqu'à l'adresse et même au bâtiment. Parmi les variables innovantes, nous utiliserons par exemple la présence de gel, le nombre de jours orageux ou le nombre d'artisans dans la commune.

Le mémoire s'articule en **cinq parties** :

- la *Partie I* présente le contexte assurantiel de l'étude : les spécificités de l'assurance : MRH en France, ses enjeux et ses contraintes, ainsi que l'apport du mémoire.
- la *Partie II* expose des rappels sur les principes de tarification IARD, le modèle linéaire généralisé en tant que modèle traditionnel, les limites de sa structure prédictive et la notion d'interaction statistique. On termine cette partie par la présentation des méthodes de tarification plus sophistiquées de type *Machine Learning*.
- la *Partie III* développe une méthodologie de détection des interactions statistiques, afin de rendre les modèles traditionnels plus performants. On décrit les principales techniques d'analyse de sensibilité et d'*Explainable Artificial Intelligence*. Pour introduire leur application, des exemples sur la fonction test *Ishigami* sont présentés.
- la *Partie IV* présente le cas d'application assurantiel, la tarification à l'adresse : dans un premier temps, nous avons construit un modèle de tarification simple, dit *Benchmark*, qui sera la base de comparaison, et des modèles plus avancés, dits *Métamodèles*. Ensuite, à partir de ces derniers, nous avons détecté les interactions des couples de variables. Enfin, nous avons montré le gain opérationnel dans un modèle de prime pure à l'aide de cette méthodologie.
- la *Partie V* conclut le mémoire et expose les limites et ouvertures.

Première partie

Contexte Assurantiel

Chapitre 1

Spécificités du marché de l'assurance MRH

Le chapitre suivant présente les caractéristiques de l'assurance multirisque habitation, son cadre juridique et les caractéristiques du marché français.

Il s'appuie principalement sur les documents publiés par la Fédération Française de l'Assurance (FFA) ainsi que sur les notes du cours de *Réglementation et Assurance* tenu par Franck Le Vallois.

1.1 Brève histoire de l'assurance habitation

Les premières notions d'assurance remontent loin dans le temps, sous le règne du roi Hammourabi de Babylone (1700 ans av JC), où le code Hammourabi exposait certains arrêts qui rappelleraient aujourd'hui l'assurance flotte et l'assurance emprunt : un prêt sur hypothèque pouvait être consenti pour fournir les fonds nécessaires à un voyage. Aucune prime n'était versée et si le bateau sombrait, les sommes avancées n'avaient pas à être remboursées.

Dans l'antiquité du V siècle avant JC, on retrouve une notion proche de l'assurance dans le droit maritime grec, le "prêt à la grosse aventure" : un négociant empruntait de l'argent (à un taux d'intérêt très fort, souvent jusqu'à 50 %) lorsqu'il avait besoin de fonds pour charger un navire et si le navire et sa marchandise disparaissaient en mer (tempête, attaques pirates, ...) l'"assuré" ne remboursait rien au prêteur, qui jouerait aujourd'hui le rôle de l'assureur qui s'engage à couvrir le risque.

Au Moyen-Âge, suite au développement des foires, les commerçants étaient exposés au risque d'être agressés ou volés lorsqu'ils traversaient les régions ; le "Conduit de foire", institué à ce moment, était ainsi un contrat (d'abord facultatif, et ensuite obligatoire avec le "Droit des marchés et des foires") dans lequel les Seigneurs s'engageaient à protéger les marchands, leur personne et leurs biens, en les escortant pendant la traversée de leurs terres.

À partir du XIVe siècle, c'est dans les ports italiens que l'assurance maritime va se développer pour accompagner les échanges commerciaux. Les marchands faisaient appel aux banquiers pour financer leurs expéditions maritimes et également la vie des hommes de l'équipage : en cas de naufrage, les marchands n'avaient rien à rembourser aux banques, mais si le bateau arrivait à bon port, le banquier était remboursé et pouvait recevoir une compensation financière. Une loi génoise vint en 1434 réglementer la profession des courtiers en assurance.

En 1652 le financier italien Lorenzo Tonti créa la tontine, une forme de contrat d'assurance proche de l'assurance vie, où des associations de personnes mettent en commun des fonds pendant une certaine durée. Au terme de cette période, les fonds étaient répartis entre les participants.

La succession d'incendies de la deuxième moitié du XVIIe siècle (Londres en 1666, Hambourg en 1676 et Bruxelles en 1695) a montré la nécessité de l'assurance et de la prévention : suite à l'incendie qui détruisit 13 200 bâtiments à Londres, Nicholas Barbon ouvre un bureau pour assurer les bâtiments.

Quelques années plus tard, sous le ministre Colbert, le marché de l'assurance incendie se développe en France, proposant une unique assurance couvrant incendie et vie. Ce n'est seulement qu'en 1788 que ces deux garanties furent séparées.

L'assurance évolue aussi techniquement grâce à l'utilisation du calcul des probabilités pour la construction des tables de mortalité, l'évaluation du risque de perte pour une compagnie d'assurance ou le calcul des rentes

viagères.

C'est ainsi que en 1816 avec la naissance de la première *Mutuelle d'assurance contre l'incendie* que l'assurance habitation s'ouvre au marché français. A sa création, cette branche garantissait uniquement la garantie Incendie, puis à partir des années 1970, d'autres garanties ont été ajoutées.

1.2 Le contrat MRH Multirisque Habitation

Aujourd'hui, il existe trois types des contrats d'assurance :

1. des personnes
2. l'assurance de biens
3. et l'assurance des responsabilités.

Le nouveau article 1101 du Code Civil définit le contrat comme *un accord de volontés entre deux ou plusieurs personnes destiné à créer, modifier, transmettre ou éteindre des obligations*.

Le **Contrat d'assurance**, en particulier, spécifie :

1. Les deux parties et leurs obligations :
 - L'assuré qui s'engage à verser une prime
 - L'assureur qui s'engage à verser les dédommagements ou prestations au titre de garanties
2. Le risque ou l'événement garanti

Le contrat d'assurance est *aléatoire*¹ car les flux (de prestation) ne sont pas stables et / ou les montants de prime peuvent être aléatoires aussi (quand l'aléa est la durée de vie, dans les contrats vie par exemple).

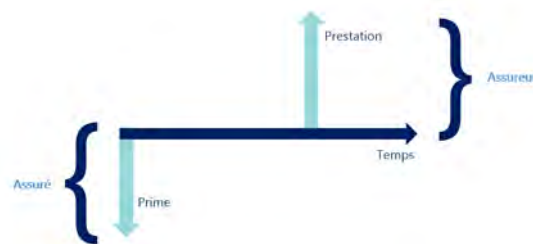


FIGURE 1.1 – Schéma général du contrat d'assurance

Définition 1. *Le contrat d'assurance multirisques habitation (MRH) est un contrat multi-garanties qui permet de protéger l'assuré, que l'on soit propriétaire ou locataire, le patrimoine familial (habitation et mobilier) contre les conséquences d'événements affectant son domicile ou mettant en cause sa responsabilité ou celle des membres de sa famille.*

La loi² impose de souscrire, au minimum, un contrat d'assurance couvrant les risques locatifs.

En principe, les organismes assureurs proposent toutes les garanties et acceptent éventuellement d'en retirer quelques unes, mais peu pour éviter des effets d'antisélection. La forme la plus courante d'un contrat MRH est celle d'un groupe de garanties, avec différents niveaux de protection qui peuvent être choisis.

Contrairement au Royaume-Uni, le renouvellement des polices en France est faible. Par conséquent, un portefeuille de MRH est généralement constitué de multiples couches de contrats plus ou moins anciens, désignant des garanties plus ou moins étendues. Il est important de réaliser que toute analyse statistique sur un tel portefeuille est effectuée sur une masse hétérogène, et que la comparaison des différents contrats constituant ce portefeuille n'a que peu de sens.

¹à l'exception des Contrats d'épargne (qui appartient aux contrats Vie) qui ne sont aléatoires car le risque est porté par l'assuré et l'assureur s'engage que à investir

²Loi n°89-462 du 6 juillet 1989 articles 7 et 8 visant à uniformiser et améliorer les rapports locatifs entre locataires et propriétaires.

Le contrat Multirisques Habitation



FIGURE 1.2 – Composition d'un contrat MRH

Les garanties

Aujourd'hui les **garanties** les plus courantes d'un contrat multirisque habitation sont les suivantes :

- les garanties de type **Dommages aux biens** :

- La garantie **Dégât des eaux**. Elle couvre des sinistres dont les faits générateurs sont des fuites d'eau, ruptures des conduites, débordements de canalisations, infiltrations. La couverture prend en charge les dégradations survenues chez l'assuré et s'étend aux logements voisins.

Remarque :

Dans le cas d'infiltration d'eaux, en règle générale, les eaux venant par le haut (vasistas, fenêtres...) relèvent de la police alors que celles qui viennent par le bas (remontée des nappes, dégâts des eaux par capillarisation) relèvent du régime Cat-Nat.

Dans le cas des fuites, la recherche de fuite est garantie, c'est-à-dire les dommages nécessaires pour trouver la fuite, mais la réparation n'est pas couverte. La garantie « recherche de fuites » illustre bien les liens entre garantie des dommages aux biens et prix de la construction. Cela souligne l'importance de l'indice Construction, à utiliser lors des études relatives aux dommages aux biens dans le cadre d'une MRH.

- La garantie **Incendie**. Elle couvre les dégâts causés par les événements : explosion, implosion, fumée accidentelle, chute de la foudre, chute d'appareil de navigation aérienne, choc de véhicule terrestre, dégâts consécutifs aux interventions des pompiers ; par exemple, la perte de valeur consécutive à l'introduction d'eau dans une saumure d'anchois suite à une intervention des pompiers chez un voisin rentre dans cette catégorie.
- La garantie **Bris de glace** prend en charge les dommages matériels (bris, fissures, etc.) aux éléments de séparation avec l'extérieur ou délimitant une pièce (portes, fenêtres, baies vitrées...).
- La garantie **Évènements Climatiques** : c'est une couverture (obligatoire des contrats qui garantissent les dommages d'incendie ou tous autres dommages à des biens situés en France) des effets du vent dus aux tempêtes, ouragans et cyclones. A cette garantie est associée une garantie grêle couvrant le choc de la grêle sur les toitures et le poids de la glace ou de la neige accumulée sur les toitures. L'assureur peut exclure de sa garantie certains bâtiments, éléments de bâtiments ou biens qui ne présenteraient pas une résistance suffisante à un vent violent, même si ces biens sont par ailleurs assurés contre l'incendie.

Cette garantie exclut ce qui est garanti en Cat-Nat.

- La garantie **Vol et Vandalisme** couvre la disparition, la destruction ou la détérioration des biens mobiliers résultant de vols, tentatives de vol et/ou d'actes de vandalisme commis dans les circonstances prévues au contrat et dont l'assuré doit en apporter la preuve. Certains biens peuvent être garantis selon un montant limité : les objets de valeur et les bijoux, les espèces et valeurs, le mobilier des dépendances, le mobilier de villégiature, etc.
- La garantie **Catastrophes Naturelles** est une garantie légale obligatoire depuis 1981 qui couvre les événements qui ont fait l'objet d'une déclaration par arrêté interministériel pour entraîner un dédommagement.

Des autres garanties proposées par les assureurs sont la garantie **Dommages Électriques** (qui parfois figure en annexe de la garantie incendie) et la garantie **Catastrophes Technologiques**

- les garanties de type **Responsabilité Civile** :

- Liées à l'habitation :
 - * la garantie **Risque locatif** : cette garantie peut être souscrite par le locataire, car il est présumé responsable en Incendie, et peut vouloir se prémunir contre le recours du propriétaire contre lui. Le locataire est contraint par la loi de s'assurer contre les risques locatifs.
 - * la garantie **Recours des locataires** : ceci concerne un sinistre survenu suite à une faute du propriétaire (par exemple : une cheminée qui n'est pas aux normes) ;
- La garantie **Responsabilité Civile Vie Privée** couvre les conséquences pécuniaires de la responsabilité civile encourue par l'assuré à la suite de dommages corporels, matériels ou immatériels consécutifs causés à des tiers au cours de la vie privée. Elle ne concerne que les sinistres survenus suite à des actions involontaires.

D'autres garanties, telles que les garanties **Protection juridique** ou **Assistance** à domicile, peuvent également être proposées.

Les biens assurables

- Les bâtiments : l'assureur désigne les bâtiments appartenant à l'assuré ainsi que leurs aménagements et installations qui ne peuvent être détachés sans être détériorés ou sans détériorer la construction (maison, appartement, greniers, cave, garages, abris de jardins, etc.).
- Le mobilier personnel : l'assureur garantit les meubles et objets personnels appartenant à l'assuré, aux membres de sa famille, à ses employés et ouvriers et à toute autre personne résidant où se trouvant momentanément dans les lieux assurés.
- Les biens à usage professionnel : il s'agit de tous les meubles, instruments, outillages et machines utilisés pour les besoins de la profession de l'assuré (la couverture de cette catégorie de biens est en général accordée de manière optionnelle dans le contrat MRH avec des limitations de capitaux).

Chapitre 2

Enjeux et contraintes en Assurance Habitation

Pendant la dernière décennie, le marché de l'assurance a été touché par plusieurs événements :

1. une concurrence toujours plus forte, d'autant plus que la réglementation avec la Loi Hamon protège le consommateur, aux dépens de l'assureur.
2. le progrès dans l'apprentissage statistique a amené en assurance des nouvelles méthodes d'exploitation d'un grand volume des données, qui performant souvent mieux que les méthodes traditionnelles.

Si d'un côté l'hyper-segmentation tarifaire semblerait être la piste d'évolution naturelle de la tarification actuelle, de l'autre des questions éthiques se posent, surtout vis-à-vis de l'exploitation des données personnelles. L'assureur doit donc trouver un compromis entre rentabilité et concurrence.

2.1 Le parcours de souscription habitation

Le parcours de souscription d'une assurance habitation, en ligne ou en agence, se compose d'une série de questions posées à l'assuré concernant son statut, le logement assuré, les occupants et les antécédents.

La plupart de ces questions représentent les "critères tarifaires", c'est-à-dire des variables utilisées par l'assureur pour modéliser la prime pure. Un détail plus exhaustif de ces questions sera traité par la partie dédiée aux principes de Tarification Non-Vie.

Aujourd'hui, le nombre de questions posées est en moyenne 20¹.

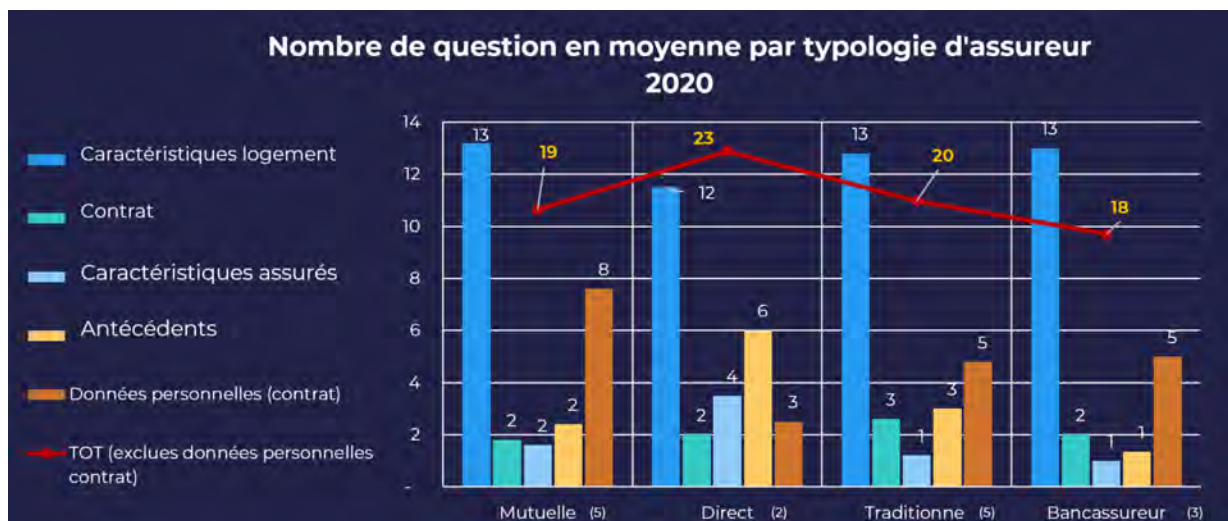


FIGURE 2.1 – Nombre de questions posées en ligne.
Source : étude interne au cabinet

¹exclues les données personnelles de rédaction de la police d'assurance

Mutualisation vs Hypersegmentation

Historiquement, les organismes d'assurance constituaient des paquets homogènes de risques afin de réaliser la mutualisation des risques : cela se traduisait par un nombre limité des questions à l'assuré ou le découpage des variables continues en classes. Aujourd'hui, les assureurs ont de plus en plus besoin d'affiner leur tarif à cause de la concurrence et de la réglementation qui défend le consommateur (Loi Hamon).

C'est la pratique connue comme segmentation en classes de risque, un découpage des variables tarifaires pour différencier les profils qui se ressemblent (en classe d'âge par exemple).

L'amélioration de la segmentation d'un portefeuille d'assurés constitue un enjeu économique et stratégique majeur pour ne pas se retrouver en difficulté financière. En effet, la sélection adverse peut se manifester auprès d'un assureur qui ne discrimine pas : en proposant la même prime aux mauvais et aux bons risques, ces derniers vont préférer s'assurer chez un autre assureur qui opère de la segmentation car il va leurs proposer une prime inférieure.

Si d'un côté on a l'impression de rentrer dans une *spirale de la segmentation*² car les questions de type "déclaratif" (la valeur de la surface habitable par exemple) et le nombre de modalités augmente, de l'autre on se demande à quel niveau il faut s'arrêter dans le découpage en classes pour qu'il y ait une solidarité minimale entre les assurés.

L'actuaire dans son nouveau rôle technique-marketing recherche l'équilibre parfait entre mutualisation, segmentation et personnalisation des tarifs. Toutefois, cela n'est pas toujours si simple, spécialement avec la diffusion des techniques d'intelligence artificielle puisqu'elles ont un pouvoir prédictif en moyenne supérieur aux méthodes linéaires et elles peuvent récupérer des données en ligne à granularité très fine.

2.2 Contraintes techniques

Limites des GLMs

Pour mesurer de façon adéquate leurs risques, les compagnies d'assurance s'appuient sur des modèles statistiques, qui quantifient les relations entre la valeur des contrats des risques assurés et les variables décrivant ce risque.

Le modèle linéaire généralisé (GLM) est l'outil principal pour modéliser la sinistralité. il s'agit d'un modèle paramétrique, qui en plus de la régression linéaire simple permet de modéliser des comportements non linéaires et des distributions de résidus non gaussiens.

Pendant les trente dernières années, l'effort technique en tarification Non-Vie a été concentré sur la maîtrise et l'amélioration des GLMs. Bien que performants, ces modèles imposent des contraintes sur la **structure** du risque modélisé et sur les **interactions** entre les variables explicatives du risque. Parmi les limites du GLM, nous avons focalisé notre attention sur la détection et la modélisation d'interactions entre les variables.

L'actuaire, en s'appuyant sur des connaissances métier, peut mesurer un certain nombre d'interactions et des corrélations parmi les variables explicatives à l'aide des tests statistiques. Cela peut avoir un coût non négligeable : si on voulait tester les interactions parmi N variables explicatives à n modalités, il existerait $\sum_{i=1}^N \binom{N}{i} n^i$ interactions possibles dont il faudrait tester la significativité.

La littérature propose des interactions à utiliser dans la tarification : par exemple, l'interaction entre "Maison isolée" et sélection d'une formule "sécurité"³, mais cela est limité à la disponibilité de ces informations dans les questionnaires ou/et dans les variables externes.

Contraintes dans la Tarification

À la différence de la statistique classique qui formule des hypothèses sur la structure et la distribution des données, les méthodes dits d'apprentissage statistique (ou Machine Learning) imposent uniquement que les données à prédire soient i.i.d. Un algorithme de prédiction est ensuite construit, évalué avec une fonction de risque et calibré à l'aide de techniques d'estimation de fiabilité (comme la validation croisée) et optimisé et validé pour réduire l'erreur de prédiction.

Malgré l'accrue popularité de l'apprentissage automatique et de l'exploitation du big data, dans la littérature de l'assurance peu d'articles vont au-delà du modèle traditionnel GLM, et peu d'assureurs utilisent le Machine

²Charpentier, Denuit, Elie : SEGMENTATION ET s LES DEUX FACES D'UNE MÊME PIÈCE ?

³Camille Desvilletes, *Modélisation du ratio combiné des affaires nouvelles d'un produit multirisque habitation*

Learning comme modèle prédictif de la sinistralité.

A la base de ce choix, il y a trois raisons principales :

1. Effet boîte noire : les modèles de Machine Learning, tels que Random Forest, xgboost, Réseaux de neurones ne sont pas interprétables et par la loi, les individus ont droit à une explication de la logique derrière la décision (RGPD art.22, RGPD Raison 71), ce qui signifie que les modèles de tarification doivent être transparents et faciles à communiquer à tous. Les *multiplieurs* (qui dérivent des coefficients de régression linéaire), dans le mécanisme de tarification, identifient l'effet que chacune des variables a indépendamment des autres sur la sinistralité. Or, l'identification de tous les impacts est impossible à cause du manque d'interprétabilité.
2. Mettre en place une tarification par un modèle non (forcément) linéaire signifierait changer radicalement la façon de travailler pour la plupart des départements d'actuariat.
3. L'assureur a le rôle social de créer une solidarité parmi les assurés. L'utilisation du Machine Learning dans la tarification ne doit en aucun cas conduire à une "personnalisation du risque" extrême ou à une discrimination, par exemple, sous la forme de primes extrêmement élevées.

2.3 Apport du mémoire

Aujourd'hui les techniques de Machine Learning ont un rôle encore très marginal dans la tarification (dans la sélection des variables par exemple) et très rarement elles sont utilisées en tant que modèle de tarification principal (fig. 2.2).

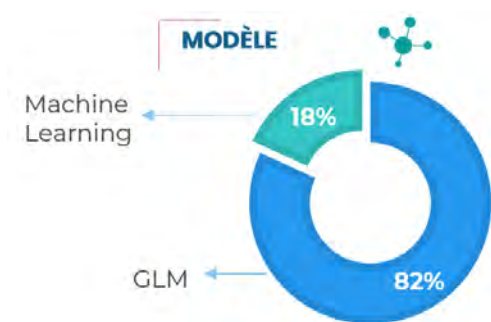


FIGURE 2.2 – Utilisation des méthodes en tarification dans le marché français (échantillon de 22 acteurs de l'assurance). Source : Enquête Fonction Pricing Data, Addactis France, Fev 2020.

Dans une vision plutôt inclusive et collaborative entre les modèles de Machine Learning et les GLMs, nous nous servirons des modèles de type boîte noire pour détecter des **interactions** parmi les variables, sans les spécifier au préalable, et puis nous testerons leur pertinence dans un GLM simple. Ainsi, les termes d'interaction, étant une expression de la complexité d'un modèle, permettent aux modèles GLM de bénéficier de bonnes propriétés des modèles plus sophistiqués.

Par rapport à ce que la littérature propose sur la détection des interactions dans les modèles statistiques, notamment aux travaux de Antoine Guillot[8] en Assurance Non-Vie, le mémoire étend la recherche et l'exploitation des effets combinés des variables d'entrées sur une sortie à deux domaines : l'analyse de sensibilité et l'*Explainable Artificial Intelligence* (XAI).

Les finalités ultimes qu'on souhaite atteindre sont deux :

- proposer un modèle de tarification simple qui prend en compte les interactions parmi les variables et les possibles contraintes opérationnelles lors de sa mise en place. On réalisera arbre de décision avec plusieurs trajectoires de tarifications ;
- comprendre les modèles, l'importance qu'ils attribuent aux variables globalement et localement, à l'aide de méthodes d'interprétation qui ne violent pas le Règlement général sur la protection des données (RGPD).

Le cadre d'application est le produit *SmartHome Pricing* pour la garantie Dégâts des eaux, une tarification habitation très segmentante qui utilise des données météorologiques, économiques, climatiques et démographiques hyperindividualisées associées à l'adresse et au bâtiment assuré.

Étapes de l'étude

La force de la méthodologie d'étude que nous proposons est l'indépendance du modèle, aussi appelé *agnosticité* au modèle, sur lequel les interactions sont détectées : peu d'hypothèses sont nécessaires pour estimer la valeur d'une interaction entre deux ou plusieurs variables.

En d'autres termes, elle peut s'appliquer à n'importe quel type de modèle.

Toutefois, les spécificités des données, à savoir la qualité ou le type de variable (catégorielle, continue), jouent un rôle fondamental dans la convergence ou dans la vitesse de convergence des estimateurs des interactions.

Dans un premier temps, nous modéliserons la fréquence et la sévérité, à partir des données de la tarification à l'adresse. Le modèle traditionnel GLM, appelé ici et par la suite *Benchmark* est la norme de l'industrie de l'assurance, et il sera celui auquel on ajoutera les interactions. Nous construirons trois modèles de Machine Learning, un arbre de régression, une forêt aléatoire et un Extreme gradient boosting (xgboost) : ils ne seront pas utilisés en tant que modèles de tarification, mais comme modèles complexes. On supposera que leur complexité est manifestée par l'introduction d'interactions dans leur structure prédictive.

En deuxième lieu, à l'aide de l'analyse de sensibilité (indices de Sobol) et de XAI (indices SHAP), nous détecterons les interactions des modèles complexes construits à l'étape précédente. Ces deux techniques sont complémentaires, car l'une intervient globalement, alors que l'autre représente des interactions localisées à des intervalles des domaines de définition spécifiques.

L'arbre de régression quant à lui est interprétable par construction.

Enfin, les interactions seront ajoutées au modèle de Benchmark : les interactions locales, étant des interactions différentes selon leur appartenance à un intervalle, seront ajoutées en tant que fonctions indicatrices ; celles globales, étendues à tout le domaine de définition, seront traduites par des termes polynomiaux du second degré.

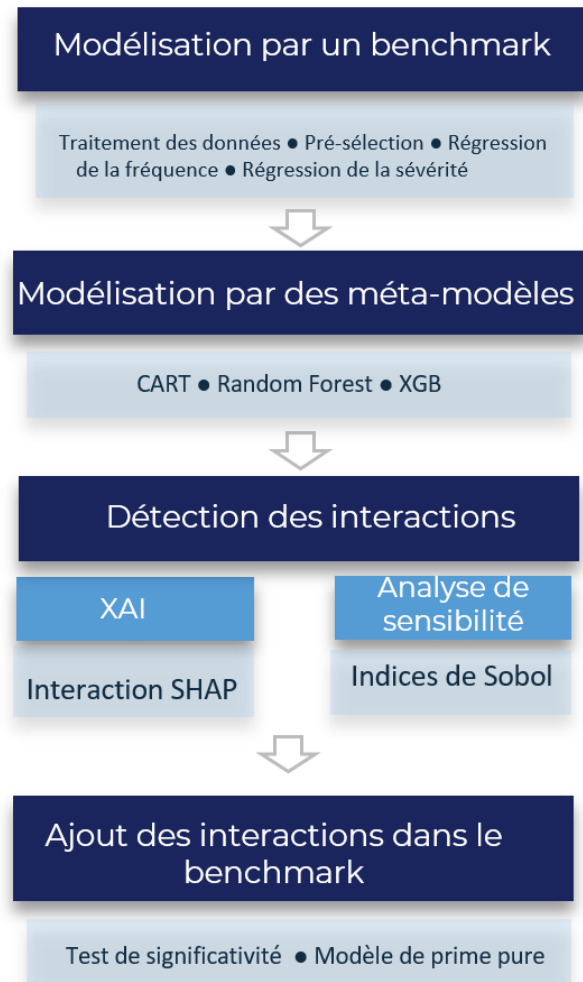


FIGURE 2.3 – Étapes de l'étude d'application de l'analyse de sensibilité pour l'optimisation tarifaire.

Deuxième partie

Principes de Tarification IARD

Chapitre 3

Rappel de Tarification en Assurance Non - Vie

Ce chapitre rappelle les notions principales de la tarification IARD et fait référence aux notes du cours *Actuariat Non Vie* tenu par Christophe Dutang [7].

3.1 Caractéristiques et principes de l'assurance Non-Vie

La définition la plus courante de l'opération d'assurance est celle de Joseph Hémard (1876-1932), juriste français et professeur en faculté de droit¹ :

“ L'assurance est une opération par laquelle une partie, l'assuré, se fait promettre moyennant une rémunération (la prime ou cotisation) pour lui ou pour un tiers, en cas de réalisation d'un risque, une prestation par une autre partie, l'assureur, qui prenant en charge un ensemble de risques, les compense conformément aux lois de la statistique. ”

L'assurance repose sur deux piliers :

1. Le caractère par nature aléatoire du risque sur lequel porte le contrat : au moment de la signature, ni l'assuré ni l'assureur ne doivent savoir si le sinistre se réalisera.
2. La mutualisation des risques : c'est le principe de verser la prime d'assurance pour régler les sinistres des autres assurés sachant que peut-être qu'un jour, ce sera pour soi-même. La mutualisation est fondée sur le partage du coût d'un sinistre sur un groupe de personnes : sans assurance, les personnes responsables d'un sinistre grave tel qu'un accident de la route, ou ayant subi un incendie ou encore souffrant d'une maladie chronique, ne pourraient pas régler leurs factures toutes seules.

L'assurance Non-Vie (ou IARD *Incendie, Accident et Risques Divers*) est une opération par laquelle l'assuré contracte, moyennant un paiement (la prime ou cotisation), une prestation par l'assureur en cas de réalisation d'un risque. À la différence de l'assurance vie, l'aléa réside à la fois sur le montant et sur la date de versement des flux. Le risque principal pour l'assureur est ainsi lié à la variabilité des sinistres que l'on qualifie de risque de variance.

La prime d'assurance

La prime d'assurance se décompose en plusieurs parties :

- la **prime pure**, notée $\mathbb{E}(S) := \int_0^{+\infty} x dF_S(x)$, qui doit permettre à l'assureur de faire face à la charge du sinistre S ; elle représente l'espérance mathématique du montant cumulé des sinistres S ;

¹Jean-Luc PÉTRICOUL, Guide pratique de l'assurance, 4ème édition, JLP Consultant, 2020

- la **prime technique**, notée $\Pi(S)$ qui ajoute des chargements techniques (entre 30% en assurance dommage et plus de 13% en assurance vie²);
- la **prime commerciale** qui est la prime proposée sur le marché, qui prend en compte des chargements fiscaux³, de la concurrence, de la stratégie commerciale,...

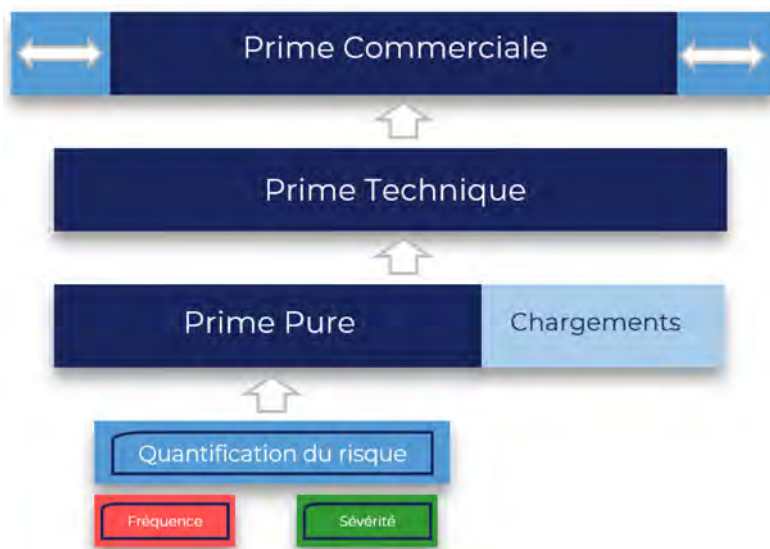


FIGURE 3.1 – Différentes primes d'assurance

À la différence des autres productions économiques, le cycle de production est inversé en assurance, c'est-à-dire que l'on vend un produit avant d'en connaître le coût de revient. C'est pour cela que la prime d'assurance est traitée comme une variable aléatoire et modélisée à l'aide d'outils statistiques et probabilistes. Elle prend en considération différents éléments :

1. la charge totale des sinistres,
2. la probabilité de survenance ou fréquence des sinistres,
3. le coût moyen ou sévérité
4. l'exposition du contrat au risque,
5. d'autres critères tarifaires (somme assurée, âge de l'assuré,...).

Le principe de calcul de prime est fondamental pour l'assureur car il détermine la loi de probabilité de son résultat futur ainsi que sa probabilité de ruine.

Définition 2. *Un principe de calcul de prime est une fonction qui associe à un risque S une prime technique "acceptable" pour l'assureur, qui quantifie la dangerosité de S :*

$$\begin{aligned} \Pi &: \{ \text{Espace des variables aléatoires} \} &\rightarrow & \mathbb{R} \\ &S &\mapsto & \Pi(S) \end{aligned}$$

La définition d'un principe de calcul repose sur plusieurs propriétés "souhaitables" :

1. des propriétés *de rationalité*, qui sont peu contestables, comme :

- *Pas de chargement inutile* : $\Pi(S) \leq \max(S) = \sup\{x : \mathbb{P}(S \leq x) < 1\}$

²les chargements techniques prennent en compte les fluctuations de la charge de sinistres autour de la moyenne (chargement de sécurité), des frais de gestion (17,6%), frais de rémunération du capital dans Sociétés par actions (2%), frais de production : rémunération des intermédiaires (12,2 %), frais d'encaissement des primes.

³9% pour risques divers, 18% pour risque automobile (a doublé en 1984) + 15% à Sécurité Sociale + 1,90 pour fonds de garantie + 0,10 au fonds de revalorisation, soit plus de 35%. Elle atteint 30% de la prime pour les risques incendie d'habitation des particuliers. Il faut ajouter diverses taxes parafiscales par contrat pour financer le fonds de garantie attentats, et le fonds d'indemnisation des victimes du SIDA.

- *Chargement positif* : une prime d'assurance doit être plus élevée que la prime pure $\mathbb{E}(S)$ sinon la ruine devient quasi-certaine lorsque la taille du portefeuille devient très grande
 - *Invariance par traslation* : pour toute constante $c \geq 0$, $\Pi(S + c) = \Pi(S) + c$
2. Propriétés d'Additivité et d'homogénéité, qui décrivent les comportements des risques lorsqu'on les agrège, comme :
- Propriété de *sous-additivité* pour laquelle la somme de deux risques réduit le risque global par diversification et la prime à demander est inférieure ou égales à la somme de deux primes pour chacun des risques individuellement : $\Pi(X + Y) \leq \Pi(X) + \Pi(Y)$.
Si X et Y sont indépendants, $\Pi(X + Y) = \Pi(X) + \Pi(Y)$.
3. des propriétés de comparaison des risques, qui expliquent comment comparer deux primes à partir d'une comparaison entre les risques sous-jacents, par exemple :
- Propriété de *monotonie* : si le montant de sinistres S est toujours inférieur au montant de sinistres Y (c'est-à-dire $\mathbb{P}(S \leq Y) = 1$ où S et Y sont définis sur le même espace de probabilité), alors $\Pi(S) \leq \Pi(Y)$
4. des propriétés mathématiques :
- Propriété de *convexité* : Pour tous risques X et Y définis sur le même espace de probabilité, pour toute constante $\alpha \in [0, 1]$,

$$\Pi(\alpha X + (1 - \alpha)Y) \leq \alpha\Pi(X) + (1 - \alpha)\Pi(Y)$$

- Propriété d'*itérativité* : Pour tous risques X et Y définis sur le même espace de probabilité,

$$\Pi(X) = \Pi(\Pi(X|Y))$$

- Propriété de *convergence en loi* : Si (X_n) converge en loi vers X et si $\max(X_n) \rightarrow \max(X)$,

$$\Pi(X_n) \rightarrow \Pi(X)$$

- Propriété de *stabilité par mélange* : Soient X' , X_1 et X_2 des risques et $p \in [0, 1]$. Si $\Pi(X_1) = \Pi(X_2)$, alors :

$$\Pi(pF_{X_1} + (1 - p)F_{X'}) = \Pi(pF_{X_2} + (1 - p)F_{X'})$$

Dans le cadre de ce mémoire nous avons retenu le principe suivant, appelé de l'espérance mathématique :

Définition 3. La compagnie assure le risque S au prix $\Pi(S)$ défini par

$$\Pi(S) = (1 + \beta)\mathbb{E}(S), \quad \beta > 0$$

où β est le taux de chargement.

Ce principe est simple, mais la dispersion de S autour de sa moyenne n'est pas prise en compte.

Le principe de la variance, défini par $\Pi(S) := \mathbb{E}(S) + \beta\text{Var}(S)$, permettrait de résoudre ce problème, mais la division du risque réduit la prime :

$$n\Pi\left(\frac{S}{n}\right) < \Pi(S)$$

Un principe qui prend en compte les fluctuations et la division du risque est celui de l'écart type ($\Pi(S) := \mathbb{E}(S) + \beta\sigma(S)$).

Principes de l'assurance Non-Vie

Cette section présente les contraintes encadrant la conception et la mise en oeuvre de modèles mathématiques servant à calculer les primes d'assurance.

Inversion du cycle de production

L'inversion du cycle de production est la principale caractéristique qui différencie le fonctionnement de l'assurance d'une quelconque production économique : les assureurs vendent l'engagement de verser des prestations avant de connaître le coût des sinistres.

Cette spécificité a deux conséquences :

1. la première est la nécessité de mettre en place des outils mathématiques sophistiqués afin d'évaluer le montant de la prime à demander à l'assuré pour le protéger du risque et éviter les pertes pour l'assureur. Ces outils sont principalement statistiques et probabilistes et utilisent les données historiques pour cerner la variabilité des risques.
2. la seconde est que l'assurance est extrêmement dépendante des données connues par l'assureur d'une part et l'assuré d'autre part sur le risque couvert par un contrat. Des asymétries d'informations entre les deux protagonistes sont régulièrement constatées et expliquent une partie des règles qui encadrent l'activité d'assurance.

La variabilité du risque est une notion importante en assurance puisque le bilan de l'assureur en dépend fortement. Qualifier et quantifier la variabilité du risque est une obligation pour l'assureur ainsi que instaurer des mécanismes pour s'en protéger (dont la mise en place d'une réserve de solvabilité et la réassurance).

Principe indemnitaire

il s'agit du principe en vertu duquel l'assurance ne doit pas permettre l'enrichissement de l'assuré. Il s'ensuit qu'au nom de ce principe, l'indemnité d'assurance ne peut pas dépasser la valeur du bien assuré dans les assurances de dommages.

Asymétrie d'information

L'asymétrie d'information décrit une situation dans laquelle tous les participants à un marché ne disposent pas de la même information. Elle se manifeste sous deux formes :

1. l'antisélection : Selon le dictionnaire de l'économie d'assurance, l'antisélection est *le mécanisme dû à l'asymétrie de l'information entre assureurs et assurés par lequel, dans une population hétérogène, les mauvais risques sont les plus demandeurs d'un contrat d'assurance donné. Si les assureurs proposent à tous les membres d'une population donnée un certain contrat d'assurance, et si l'on suppose que l'aversion pour le risque de tous les individus est la même, alors ce sont avant tout les « mauvais risques », c'est-à-dire les individus dont le risque est le plus élevé qui achèteront le contrat. La sinistralité effective se révélera donc plus élevée que la sinistralité moyenne estimée de l'ensemble de la population.* En d'autres termes, c'est le phénomène selon lequel les mauvais risques peuvent se faire passer pour de bons risques parce que l'assureur ne peut pas observer toutes les caractéristiques qui affectent leur probabilité de sinistralité.

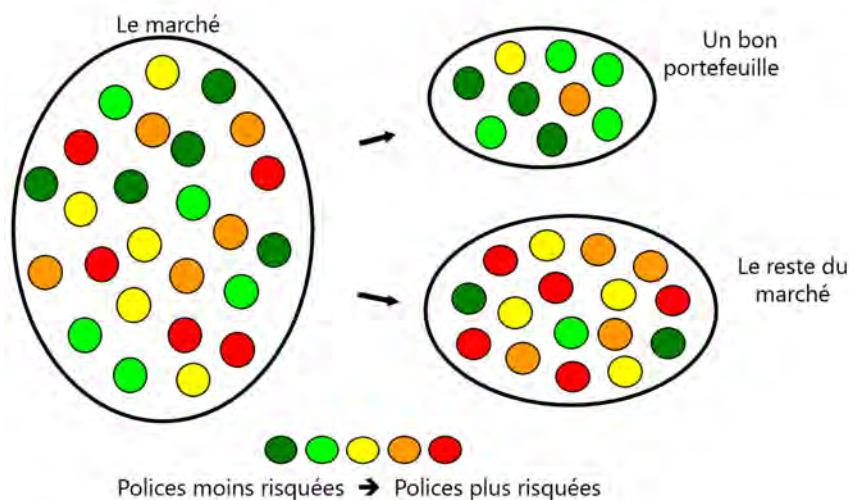


FIGURE 3.2 – Proposer une prime moyenne pour tous génère un effet d'antisélection ou de sélection adverse. Les bons profils partent du portefeuille parce que l'assureur surestime leur risques et les mauvais y restent.

2. et l'aléa moral : changement de comportement d'un individu (assuré) vis-à-vis du risque selon qu'il y soit exposé ou prémuni.

Mutualisation et Segmentation

En assurance, deux principes fondamentaux apparaissent antinomiques mais complémentaires : la mutualisation et la segmentation.

Le premier est le fondement historique de l'assurance, un mécanisme de partage des risques pour faire face à la variabilité de la réalisation des risques, de sorte qu'ils se compensent entre eux. C'est une relecture actuarielle de la loi des grands nombres, qui considère un grand nombre afin de réduire le risque moyen, n'ayant de sens qu'au sein d'une population de risques « homogènes ».

Pour faire face au risque d'antisélection, dont une des conséquences est que les bons profils partent du portefeuille suite à la sur-estimation de leur risque alors que les mauvais y demeurent, l'assureur applique le deuxième principe cité, la segmentation, pour avoir des groupes de risques homogènes et proposer une prime adaptée à un profil de risque particulier (ex. Prime MRH pour le propriétaire d'une maison, âgé de 40 ans, cadre, en ville).

La recherche de modèles probabilistes permettant une segmentation appropriée des risques est une activité marquante des mathématiques de l'assurance contemporaines. Il faut cependant noter que trop de segmentation conduit à une situation problématique puisque les primes de certains assurés peuvent devenir trop importantes et le caractère solidaire de l'assurance peut disparaître.

Les assurances dommages sont fondées sur la mutualisation des risques et modérées par la segmentation. Le suivi statistique permet de définir des classes de risque, i.e. les bons et les mauvais risques, et d'établir des tarifications personnalisées.

3.2 Modélisation de la prime pure

La modélisation des risques d'un portefeuille de polices d'assurance dommages suit l'approche du modèle de risque collectif (M.R.C), par opposition au modèle de risque individuel (M.R.I).

Le modèle de risque collectif a été proposé en 1903 par Filip Lundberg, qui dans sa thèse⁴ avait mis en avant l'intérêt du processus de Poisson dans la modélisation des risques en Assurance dommage.

Selon cette approche, le portefeuille de polices est considéré comme un groupe de risques homogènes ; l'élément majeur est la prise en considération du nombre de sinistres par police.

TABLE 3.1 – Différence entre Modèle individuel et collectif

| M.R.I. | M.R.C. |
|---|---|
| pour chaque police il ne peut y avoir qu'un unique sinistre | plusieurs sinistres peuvent survenir par police |
| les risques individuels sont indépendants et hétérogènes | les montants de sinistres individuels sont indépendants et identiquement distribués (i.i.d) |

Définition 4. *Le modèle collectif de risque est une suite infinie (X_k) $k \geq 1$ de variables aléatoires indépendantes (les montants de sinistres cumulés) et identiquement distribuées et une variable aléatoire N indépendante et à valeur entières (le nombre total de sinistres pour toutes les polices du portefeuille). Le montant cumulé des sinistres S est alors défini par*

$$S = X_1 + \dots + X_k = \sum_{k=1}^N X_k$$

Hypothèses du modèle

1. **(indépendance et homogénéité des sinistres)** les montants des sinistres X_i sont indépendants et identiquement distribués. Cette condition peut ne pas être valable pour certaines branches, telles que l'assurance tempête, lorsque les sinistres sont causés par le même fait générateur et n'est pas valable lorsque l'on souhaite tarifier une maison et une usine (risque non homogènes).

⁴Approximations of the Probability Function/Reinsurance of Collective Risks, 1903

Les sinistres sont *i.i.d.* à un facteur d'actualisation ou de mise as if près.

2. (**indépendance entre le montant et le nombre de sinistres**) le nombre de sinistres, N , est indépendant des X_i ;
3. Taille. Pour pouvoir opérer la mutualisation des risques, il faut par ailleurs qu'on dispose d'un portefeuille de taille assez grande ; si les risques ne sont pas nombreux, homogènes ou indépendants, l'assureur peut envisager la possibilité de se faire réassurer.

Proposition 1. Dans le modèle collectif de risque N , (X_k) $k \geq 1$ avec montant cumulé des sinistres $S = X_1 + \dots + X_N$ nous avons si $\mathbb{E}(X_1) < +\infty$ et $\mathbb{E}(N) < +\infty$:

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

Proof 1. On conditionne par rapport au nombre de sinistres :

$$\mathbb{E}(S) = \mathbb{E}[\mathbb{E}(S|N)]$$

Sachant $N = n$, S est la somme de n termes, chacun de moyenne $\mathbb{E}(X_1)$ on a :

$$\mathbb{E}(S|N = n) = \mathbb{E}[X_1 + \dots + X_n] = n\mathbb{E}(X_1) = n\mathbb{E}(X_1)$$

En intégrant par rapport à N on a enfin :

$$\begin{aligned} \mathbb{E}(S) &= \mathbb{E}[\mathbb{E}(S|N)] = \sum_{n=0}^{\infty} \mathbb{P}(N = n)\mathbb{E}(S|N = n) = \\ &= \sum_{n=0}^{\infty} \mathbb{P}(N = n)n\mathbb{E}(X_1) = \mathbb{E}(X_1) \sum_{n=0}^{\infty} \mathbb{P}(N = n)n = \\ &= \mathbb{E}(N)\mathbb{E}(X_1) \end{aligned}$$

Deux méthodes de tarification sont couramment utilisées par les assureurs :

- lorsqu'il s'agit de classer le risque à partir d'information disponible a priori (sur l'assuré, le bien assuré...) la tarification a une approche *à priori* : l'assureur essaie de prévoir, dès l'entrée d'un nouvel assuré, sa sinistralité future. L'idée de cette approche est de segmenter le portefeuille afin de constituer des sous-portefeuilles sur lesquels les risques peuvent être considérés comme équivalents. Les groupes de risques sont aussi appelés *classes de risques a priori*.
- si l'information sur l'historique des sinistres de l'assuré est prise en considération, la tarification est dite *à posteriori* : le tarif initial de l'assuré est adapté, au cours de la vie de son contrat, à sa sinistralité individuelle.

Nous allons présenter par la suite des méthodes de tarification à priori.

Le résultat de la proposition 1 nous permet ainsi de modéliser la fréquence des sinistres et le coût moyen ou sévérité séparément selon l'approche de modélisation *Fréquence - Sévérité*.

Il y a plusieurs méthodes de modélisation de la sinistralité, des modèles statistiques traditionnels au Machine Learning.

La section suivante sera consacrée aux GLM et le chapitre suivant traitera les modèles plus récents de Machine Learning.

3.3 Le Modèle Linéaire Généralisé (GLM)

Les Modèles Linéaires Généralisés (GLM) ont été introduits par Nelder et Wedderburn en 1972 et représentent aujourd'hui le benchmark de la tarification Non-Vie. Ces modèles doivent leur succès dans le domaine de l'assurance à leur interprétabilité et à leur facile application dans les plans tarifaires.

Le principe du GLM consiste à déterminer la loi de probabilité de la variable réponse Y , ici **le nombre des sinistres ou leur sévérité**, en fonction des variables explicatives X_1, \dots, X_n , en explicitant leur relation par trois composantes :

1. une **composante aléatoire**, la variable à expliquer $Y = (Y_1, \dots, Y_n)^T$, où l'on fait l'hypothèse d'appartenance à la famille exponentielle, c'est-à-dire que sa densité f s'écrit dans la forme suivante :

$$f(y, \theta, \phi) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)}\right) + c(y, \phi) \quad (3.1)$$

où

- y appartient au domaine de Y
- $(\theta, \phi) \in \mathbb{R}^2$ sont respectivement les paramètres canonique (ou paramètre naturel) et de dispersion de la loi de distribution ;
- b est une fonction connue, trois fois dérivable, telle que sa dérivée b' est inversible ;
- c et a sont des fonctions réelles connues et dérivables

Deux propriétés caractérisent la famille exponentielle :

$$(a) \quad \forall i = 1, \dots, p : \mathbb{E}(Y_i) = b'(\theta_i) \quad (3.2)$$

$$(b) \quad \text{Var}(Y_i) = b''(\theta)\phi$$

Cet élément étend ainsi les modèles de régression linéaire aux cas où Y ne suit plus une loi Normale mais une loi de famille exponentielle, dont la loi Normale fait partie.

En actuariat, pour prendre en compte l'exposition en années des contrats dans la modélisation de la fréquence, on introduit une pondération w_i de l'observation Y_i qui intervient dans la définition de $a(\theta) := \frac{\theta}{w_i}$.

2. une **composante déterministe**, les p réalisations de n variables explicatives : $\forall i = 1, \dots, p$ on dispose de $X_{1,i}, \dots, X_{n,i}$ décrivant Y_i .
3. une **fonction de lien** g déterministe, bijective, strictement monotone et définie sur \mathbb{R} telle que :

$$g(\mathbb{E}(Y)) = \beta_0 + \beta_1 X_1 + \dots + \beta_n X_n$$

Cet élément exprime une relation fonctionnelle entre la composante aléatoire et celle déterministe. La fonction lien qui associe la moyenne $\mathbb{E}(Y)$ au paramètre naturel θ est appelée fonction lien canonique. Dans ce cas, $g(\mathbb{E}(Y)) = \theta_i = \beta_0 + \beta_1 X_1 + \dots + \beta_n X_n$

Par exemple, la fonction de lien canonique d'une loi de Poisson est la fonction logarithmique $g = \ln$.

Ces modèles sont des extensions du modèle linéaire simple et permettent à la fois de modéliser des comportements non-linéaires (grâce aux fonctions de liens) et des distributions de résidus non-gaussiens.

Par leur flexibilité, les GLMs ont remplacé en assurance la régression linéaire simple gaussienne.

En effet, en assurance Non-Vie les coûts des sinistres, quand ils se concrétisent, suivent une densité très asymétrique et non gaussienne et l'utilisation des GLMs a permis d'améliorer la qualité des modèles de prédiction du risque.

Estimation des paramètres du GLM

L'estimation des coefficients β_i s'effectue par la méthode de maximum de vraisemblance.

Pour p observations supposées indépendantes et en tenant compte que pour la propriété 3.2(a) $\theta = b'(\mathbb{E}(Y))$ dépend de $\beta := (\beta_0, \beta_1, \dots, \beta_n)$, la log-vraisemblance s'écrit comme la somme de p contributions :

$$\mathcal{L}(\beta) = \ln \prod_{i=1}^p f(y_i, \theta_i, \phi) = \sum_{i=1}^p \ln f(y_i, \theta_i, \phi) := \sum_{i=1}^p l(\theta_i, \phi, y_i) \quad (3.3)$$

L'estimateur $\hat{\beta} \in \mathbb{R}^n$ est ainsi : $\hat{\beta} = \underset{\beta}{\operatorname{argmax}} \mathcal{L}(\beta)$

Qualité d'un modèle

On peut améliorer la qualité d'un modèle linéaire généralisé selon trois axes :

- en réduisant l'espace des variables explicatives aux **variables significatives** de la prédiction. On s'appuiera sur des tests statistiques de la significativité.
- en évaluant l'**ajustement du modèle** sur la base des différences entre observations et estimations.
- en évaluant la **robustesse du modèle**, à l'aide de l'étude graphique des résidus.

Test de significativité

Après l'estimation des coefficients de régression, afin d'évaluer la pertinence des variables explicatives dans la prédiction de Y , le test de déviance (type III) est utilisé afin de choisir les variables à exclure ou inclure dans le modèle.

Ce test évalue l'apport des variables explicatives supplémentaires dans l'ajustement du modèle. On considère deux modèles emboîtés : M_2 à q_2 variables et M_1 à $q_1 = q_2 - 1$ variables, avec pour contrainte que toutes les variables de M_1 se retrouvent dans M_2 .

On effectue le test de nullité de q coefficients libres, pour tester la significativité de la variable X_j à q modalités qui n'appartient pas au modèle M_1 , c'est à dire on teste l'hypothèse suivante :

$$H_0 : \beta_{j1} = \beta_{j2} = \dots = \beta_{jq} = 0 \quad \text{contre} \quad H_1 : \exists k \in \{1, \dots, q\} \text{ tel que } \beta_{jk} \neq 0 \quad (3.4)$$

Il existe plusieurs statistiques adaptées au test de significativité. Dans le cadre de cette étude nous avons utilisé le **Test de déviance ou du rapport de vraisemblance**⁵.

Ce test compare la déviance du modèle avec la variable X_j et la déviance du modèle M_1 sans cette variable.

Sous H_0 , les coefficients des variables supplémentaires que l'on retrouve dans M_2 sont tous nuls ; on notera la log-vraisemblance sous H_0 $\mathcal{L}(\beta_{H_0})$.

La statistique du test est ainsi la différence de ces deux déviations :

$$\xi^{RV} = \underbrace{D_1}_{\text{modele sans variable}} - \underbrace{D_2}_{\text{modele avec variable}} = 2[\mathcal{L}(\hat{\beta}) - \mathcal{L}(\beta_{H_0})] \quad (3.5)$$

Sous H_0 , ces deux statistiques convergent asymptotiquement en loi vers une loi du χ^2 à q degrés de liberté. La variable X_j est jugée significative si on rejette l'hypothèse nulle, c'est-à-dire lorsque ces statistiques dépassent le quantile d'ordre $1 - \alpha$ de la loi χ^2 à q degrés de liberté.

Si le paramètre θ est estimé, on compare la statistique à une loi de Fisher au lieu de celle du χ^2 . On considère qu'une variable n'est pas significative si sa p-value est supérieure à 0.5%.

Remarques :

- nous avons présenté le cas où on ajoute une seule variable, mais on peut considérer d'ajouter plusieurs variables à la fois.
- Ce test permet donc de tester la **significativité** de la diminution de la déviance par l'ajout de variables explicatives ou la prise en compte d'interactions, sachant que $D_1 - D_2 \geq 0$. Plus on rajoute de variables dans la régression, mêmes non pertinentes, plus faible sera la déviance.

Validation du modèle

Afin d'évaluer la qualité d'ajustement du modèle sur la base des différences entre observations et estimations, plusieurs critères sont proposés.

Déviance

Le modèle estimé est comparé avec le modèle dit *saturé*, c'est-à-dire que le modèle possédant autant de paramètres que d'observations et estimant donc exactement les données. Cette comparaison est basée sur l'expression de la déviance D des log-vraisemblances $\mathcal{L}(\text{modèle estimé})$ et \mathcal{L}_{sat} (modèle saturé) :

$$D = -2(\mathcal{L} - \mathcal{L}_{sat}) = 2\ln\left(\frac{L_{sat}}{L}\right)$$

⁵D'autres test communément utilisés sont le test Wald et le test du Score

Ce rapport remplace ou généralise l'usage des sommes de carrés propres au cas gaussien et donc à l'estimation des moindres carrés. Plus ce rapport est proche de 1, plus le modèle estimé est ajusté aux données observées.

On peut montrer que asymptotiquement D suit une loi du χ^2 à $n - p - 1$ degrés de liberté, p le nombre d'observations et n le nombre de variables explicatives. On peut construire un test de rejet ou d'acceptation du modèle selon que la déviance est jugée significativement ou non importante.

On utilise la règle suivante : si $\frac{D}{n-p-1}$ est proche de 1, alors le modèle est de bonne qualité.

Plus on ajoute de facteurs explicatifs au GLM, plus la déviance est petite et plus le rapport est proche de 1. Toutefois, elle ne prend pas en compte la complexité du modèle. C'est pour cela que l'AIC et le BIC sont des critères plus adaptés dans le cadre d'une tarification.

AIC et BIC

Ces deux critères sont utilisés pour comparer deux modèles non emboîtés et qui peuvent avoir des nombres de paramètres différents.

Le modèle privilégié est celui qui a un AIC ou un BIC plus faible.

AIC

Pour palier le défaut de la déviance, qui ne mesure pas le degré de complexité du modèle, le critère d'information d'Akaike **AIC** pénalise la déviance du modèle $-2\log(L)$ par 2 fois le nombre de paramètres libres. Il est défini par :

$$AIC = -2\ln(L) + 2k$$

où L est la vraisemblance maximisée et k le nombre de paramètres libres dans le modèle.

Pénaliser les modèles très fins, comprenant trop des variables, est aussi une façon d'éviter le sur apprentissage. Le critère AIC représente donc un compromis entre le **biais**, diminuant avec le nombre de paramètres libres, et la **parcimonie**, volonté de décrire les données avec le plus petit nombre de paramètres possibles.

BIC

Le critère d'information bayésien **BIC** pénalise la déviance par le nombre de paramètres et la taille de l'échantillon :

$$BIC = -2\ln(L) + \ln(n)k$$

L'utilisation de ces deux critères dépend du type d'analyse : le critère BIC peut devenir moins performant que l'AIC lorsque le modèle est complexe. Nous nous intéressons donc principalement à l'AIC.

Robustesse du modèle

L'analyse des résidus est appliquée pour indiquer si le modèle peut être amélioré. Les résidus sont obtenus en comparant les valeurs observées Y_i et les valeurs prédites \hat{Y}_i . Selon le type d'analyse, ils peuvent être pondérés par l'écart type (Résidus de Pearson) ou être transformés afin de construire des résidus suivant une loi normale (Résidus d'Anscombe).

Dans la modélisation, leur analyse permet de vérifier que l'erreur est aléatoire, de repérer les valeurs aberrantes ou trop influentes.

L'histogramme et le QQ-plot obtenus permettront de vérifier la normalité des résidus de déviance des modèles à composante aléatoire continue tel que le modèle Gamma.

Loi de fréquence et loi de sévérité

Le tableau ci-dessous présente les modèles utilisés le plus couramment en Assurance Non Vie.

| | Fréquence | Sévérité | Montant des sinistres |
|----------------------------------|-------------------------------|---------------------|-----------------------------------|
| Loi $Y X = x$ | Poisson ou Binomiale négative | Gamma | Loi mixte Poisson/Gamma (Tweedie) |
| Fonction de lien g | $\ln(x)$ | $-1/x$ | $\ln(x)$ |
| Poids w | Exposition | Nombre de sinistres | Exposition |
| Paramètre de dispersion θ | 1 | estimé | estimé |

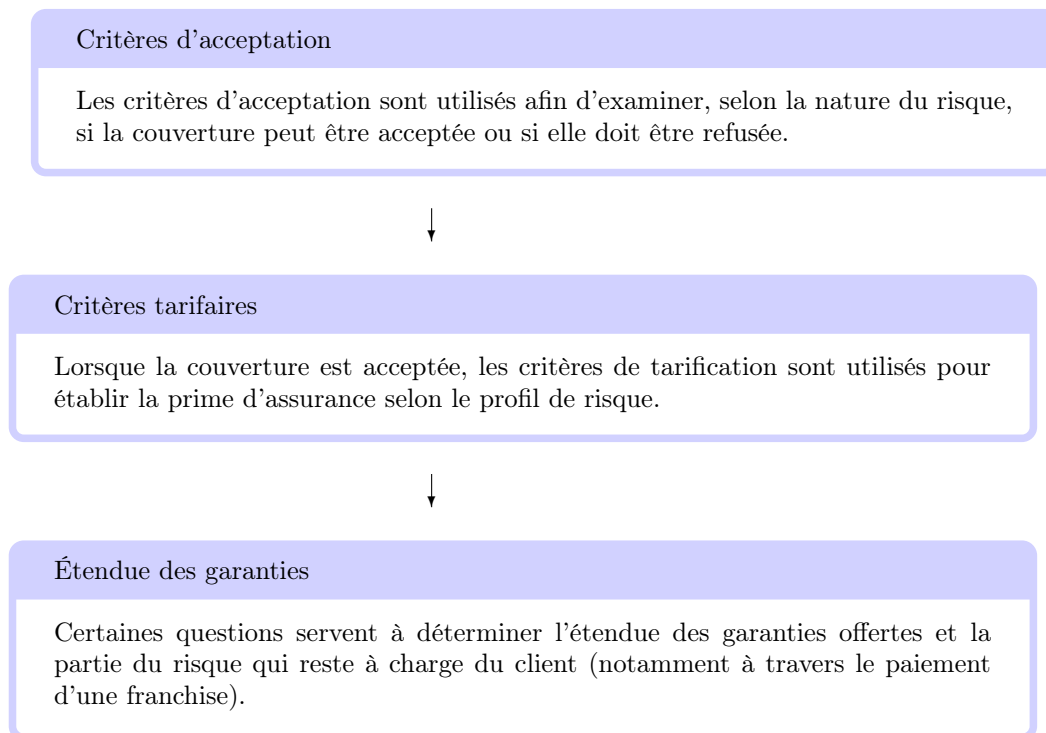
FIGURE 3.3 – Tableau récapitulatif des lois utilisées en tarification pour fréquence, coût moyen et prime pure

Dans la littérature IARD, le nombre de sinistres est très souvent modélisé par une loi de Poisson⁶, tandis que la loi Gamma est souvent calibrée pour le modèle de coût moyen.

En pratique la fonction de lien logarithmique est préférée parce qu'elle permet d'avoir un tarif multiplicatif.

3.4 Critères de tarification en Habitation

L'acceptation d'un risque, la tarification et l'étendue des garanties sont déterminées sur la base de critères objectifs qui permettent à un organisme d'assurance d'évaluer au mieux un risque déterminé.



Critères généraux

Caractéristiques des assurés

- la **Qualité** de l'assuré (locataire, propriétaire, propriétaire non occupant) induit des risques spécifiques et détermine l'étendue des garanties : le risque d'un locataire n'est pas par exemple le même que celui d'un propriétaire ;
- l'**Âge** de l'occupant peut avoir un effet sur le nombre de sinistres et sur leur coût.

⁶Frédéric PLANCHET, Guillaume SERDECZNY, " Modèles fréquence – coût : Quelles perspectives d'évolution ?", Mars 2014

Caractéristiques relatives au risque à assurer

Voici les caractéristiques relatives au risque à assurer :

- **Type de construction** : certains types de construction sont intrinsèquement plus sensibles aux sinistres que d'autres. Une construction en matériaux combustibles (telle qu'un chalet) est par exemple plus exposée au risque d'incendie. Une construction en matériaux légers (une caravane par exemple) présente un risque accru de dégâts provoqués par une tempête.
- **L'état d'une construction** (longtemps inoccupé, à démolir, en voie de construction ou de rénovation) peut avoir un impact significatif sur le risque.
- la **zone géographique** joue un rôle important pour le vol et les catastrophes naturelles.
- La **situation relative** d'un immeuble vis-à-vis d'autres immeubles influence aussi le risque de dégâts causés par le vol. La présence d'un immeuble attenant peut générer un contrôle social accru, réduisant le risque de vol. La situation relative jouera également un rôle dans le cadre d'autres garanties, telles que le vandalisme.
- **Type de résidence** : un immeuble qui ne sert pas de résidence principale sera plus exposé aux sinistres, du fait des périodes d'absence. Dès lors, un sinistre sera probablement moins vite constaté, aggravant ses conséquences (financières). Ces périodes d'absence rendent de surcroît l'immeuble plus « attrayant » pour le vol et les actes de vandalisme.
- **Mesures de prévention** : les études démontrent que le fait de prendre des mesures préventives (porte blindée, alarme) ont un effet significatif sur le risque de vol.

Chapitre 4

Effet d'interaction en Tarification Non - Vie

Afin d'améliorer le pouvoir prédictif des modèles linéaires généralisés, des termes peuvent être ajoutés dans sa structure linéaire.

Dans ce chapitre, les techniques les plus populaires d'intégration des interactions seront présentées. À partir de leurs limites, on posera une nouvelle problématique liée à la détection des interactions qui introduira les chapitres suivants de l'analyse de sensibilité et de l'*Explainable Machine Learning* .

4.1 Définition d'une interaction statistique

Définition 5. Le terme "interaction" désigne la situation où l'influence simultanée de deux variables sur une troisième n'est pas additive, c'est-à-dire quand l'impact d'une variable explicative sur la variable réponse dépend de la valeur de l'autre.

Plus formellement, considérant le modèle à deux entrées :

$$Y = f(X_1, X_2)$$

Il y a interaction entre X_1 et X_2 si l'effet marginal de X_1 sur Y , noté β_1 , dépend de X_2 :

$$\frac{\partial Y}{\partial X_1} = \beta_1(X_2)$$

La condition de Friedman et Popescue¹ généralise cette définition à plus de deux variables explicatives :

$$\mathbb{E}\left[\frac{\partial^2 \hat{Y}}{\partial X_1 \partial X_2}\right]^2 > 0$$

En d'autres termes, la relation entre les niveaux d'une variable n'est pas constante pour tous les niveaux d'une autre variable. Par exemple, en assurance automobile, la courbe d'âge de l'expérience de conduite n'a pas la même forme pour les hommes et les femmes.

Si $Y = X_1 + X_2$, avec X_1 indépendant de X_2 , alors il n'y a pas d'interaction dans le modèle.

Dans la structure du GLM, la seule interaction introduite implicitement par la fonction de lien est la suivante :

$$\frac{\partial Y}{\partial X_i} = \frac{\partial}{\partial X_i} g^{-1}(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n)$$

¹Friedman et Popescue : "PREDICTIVE LEARNING VIA RULE ENSEMBLES", 2008

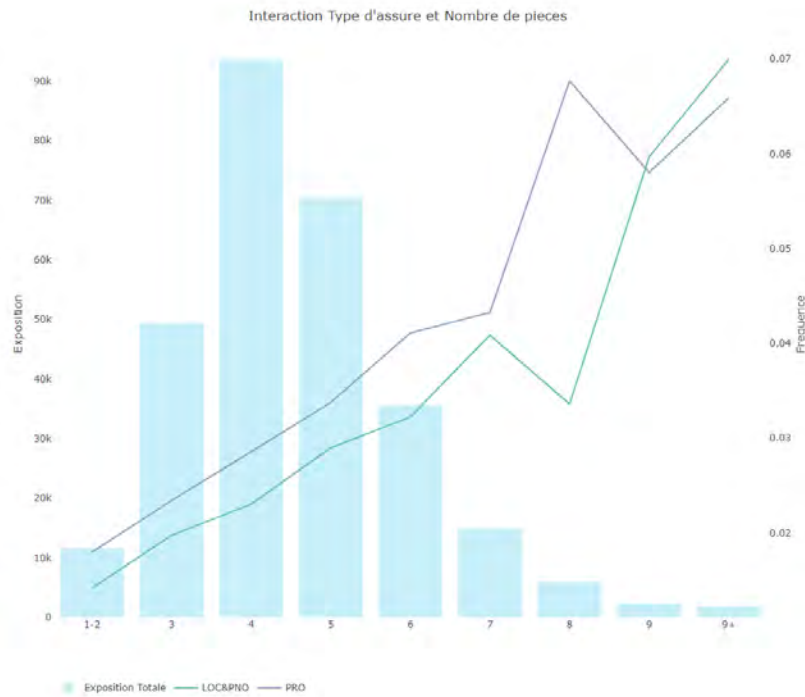


FIGURE 4.1 – Exemple d’interaction. Nous avons représenté la fréquence moyenne des sinistres par nombre de pièces et type d’assuré avec les segments violet et vert : l’écart entre les niveaux du propriétaire et de la classe "locataire et PNO" n’est pas constante, mais il varie selon le nombre de pièces. Cette tendance est particulièrement marquée lorsque l’exposition (histogramme en bleu) diminue et les estimations sont plus volatiles. Les valeurs ont été multipliées selon une constante.

4.2 Comment on détecte et on intègre les interactions statistiques aujourd’hui

Les GLMs ne permettent pas l’intégration d’interactions statistiques, au-delà de celle implicite de la fonction de lien, entre deux variables par défaut, et celles-ci doivent être spécifiées manuellement. En MRH, par exemple, une des principales interactions est entre le nombre de pièces et le type d’assuré (fig 4.1).

Méthode naïve

La méthode classique de détection d’interactions consiste à tester si le terme croisé $X_1 * X_2$ améliore la qualité du modèle. Le test de significativité est choisi parmi ceux décrits dans le chapitre 3 où l’on teste l’hypothèse nulle :

$$H_0 : \beta_i = 0$$

avec β_i est coefficient de régression du terme ajouté.

Si le modèle comporte un tel terme, on l’intègre dans l’équation tarifaire :

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 * X_2$$

L’ajout des interactions dans un GLM permet d’approcher une structure plus complexe des données d’apprentissage. En effet, le GLM simple utilise une structure linéaire pour modéliser une variable cible, alors qu’un GLM avec interactions inclut dans le modèle des polynômes.

| Modèle | Complexité | Intéprétabilité | Mise en place opérationnelle |
|-----------------------|------------|-----------------|------------------------------|
| GLM | + | +++ | +++ |
| CART | ++ | +++ | + |
| RF | +++ | + | + |
| XGB | +++ | + | + |
| GLM avec interactions | ++ | +++ | +++ |

FIGURE 4.2 – L'enjeu principal de la détection d'interactions est que l'on peut améliorer un modèle GLM traditionnel en lui conférant un degré de complexité plus élevé.

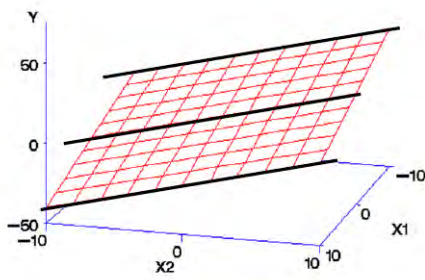


FIGURE 4.3 – Modèle linéaire sans prendre en compte des interactions : les effets simples (distances parmi les droites en noir) sont constants.

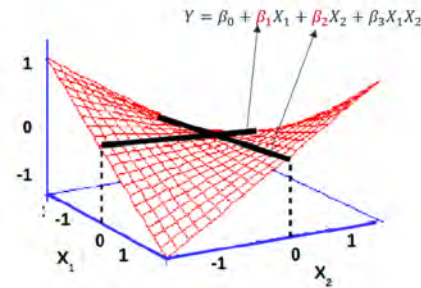


FIGURE 4.4 – Modèle linéaire avec prise en compte de l'interaction entre X_1 et X_2 : les droites ne sont plus parallèles. Les droites en noir représentent les effets moyens des variables.

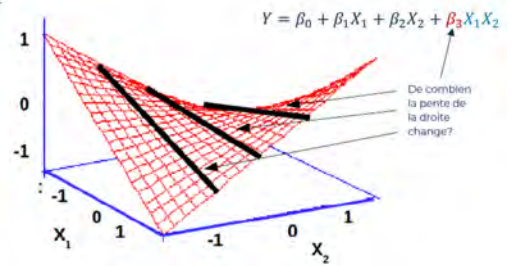


FIGURE 4.5 – Le coefficient du terme croisé $X_1 * X_2$ indique de combien l'effet de X_1 varie selon X_2 .

Analyse bivariée

Un effet d'interaction peut généralement être vu comme un ensemble de **lignes non parallèles**.

Pour examiner la relation entre deux variables catégorielles et une variable dépendante continue (la sortie), on regarde si l'effet de la première variable sur la sortie varie selon la catégorie de la deuxième.

Plus formellement, on considère *l'effet simple* de la première variable, défini comme la différence entre les moyennes des prédictions par chaque catégorie de la deuxième variable, et on vérifie s'il est constant. En d'autres termes, si les segments définis par les effets simples sont parallèles, il n'y a pas d'interaction.

| | PRO | LOC& PNO | Effets simples du type d'assuré |
|-----|-------|----------|---------------------------------------|
| | (A) | (B) | (B)-(A) |
| 1-2 | 1,80% | 1,40% | 0,40% |
| 3 | 2,30% | 2,00% | 0,40% |
| 4 | 2,90% | 2,30% | 0,60% |
| 5 | 3,40% | 2,90% | 0,50% |
| 6 | 4,10% | 3,20% | 0,90% |
| 7 | 4,30% | 4,10% | 0,20% |
| 8 | 6,80% | 3,40% | 3,40% |
| 9 | 5,80% | 6,00% | 0,20% |
| 9+ | 6,60% | 7,00% | 0,40% |

FIGURE 4.6 – Fréquence moyenne par niveaux de deux variables : les valeurs prises par les effets simples du type d'assuré ne sont pas constantes, c'est-à-dire que les droites par type d'assuré de la figure 4.1 ne sont pas parallèles. L'effet du nombre de pièces dépend ainsi du type d'assuré et il y a une interaction entre ces deux variables.

Dans le cas de variables qualitatives, celles-ci sont binarisées et sont donc représentées par plusieurs prédicteurs au sein du modèle.

Limites de ces méthodes

Les méthodes présentées ont toutefois des limites :

- la méthode naïve s'appuie sur les hypothèses statistiques sous-jacentes non vérifiées héritées du GLM ;
- les deux méthodes requièrent un coût algorithmique élevé puisqu'il faut tester un nouveau modèle par chaque interaction pour la méthode naïve et visualiser toutes les combinaisons des variables dans l'analyse bivariée. En particulier, dans notre étude à grande dimension, ces techniques ne semblent pas adaptées.

Afin de répondre à cet enjeu, nous allons proposer une nouvelle méthodologie de détection d'interactions, à l'aide des indice de Sobol et de SHAP qu'on présentera dans les chapitres suivants.

Corrélation et Interaction

Dans l'étape de pré-sélection nous avons éliminé la plupart des variables corrélées pour ne pas garder des informations redondantes.

Il peut toutefois s'avérer que des variables corrélées restent dans le modèle et soient détectées comme variables en interaction car en présence de collinéarité, des interactions artificielles peuvent apparaître.

Ces interactions, une fois détectées, devront être validées pour ne pas les confondre avec la corrélation.

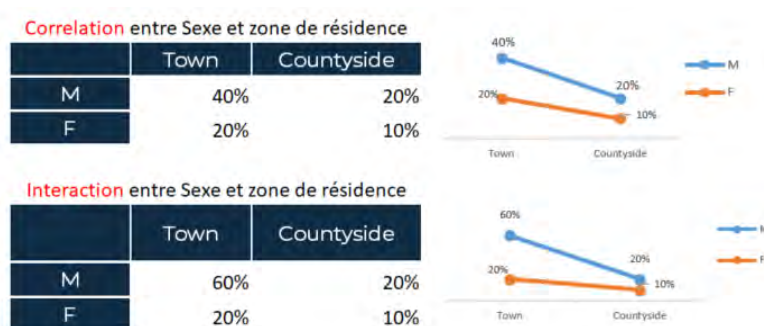


FIGURE 4.7 – Exemple de corrélation et interactions entre deux variables.

Conclusion

Dans les modèles linéaires généralisés présentés dans le chapitre précédent, nous n'avons inclus que les effets principaux.

Autrement dit, par définition de la structure linéaire des GLMs et par hypothèses d'indépendance des variables,

nous avons supposé qu'il n'y a pas d'autres interactions entre les facteurs tarifaires en plus de celle introduite par la fonction de lien.

Un premier axe d'amélioration est d'ajouter les effets de toutes les interactions bidirectionnelles entre les termes du modèle, et de tester la significativité de l'ajout de chacune des interactions dans le modèle initial. Dans la plupart des cas, les interactions testées sont pilotées par la connaissance métier.

Une alternative est d'utiliser des modèles non (forcément) linéaires tels que CART, forêt aléatoire et Xgboost, mais cela n'est pas encore très pratiqué en assurance. De plus, cela causerait des problèmes de mise en place d'outils. Par ailleurs, ces méthodes ne sont pas toutes facilement interprétables.

Par la suite nous allons présenter une nouvelle technique de détection d'interactions, à l'aide d'outils de l'analyse de sensibilité, qui permettrait de bénéficier de la structure non linéaire des modèle *black box*, tout en gardant un degré élevé d'interprétabilité et de mise en place opérationnelle.

Chapitre 5

Des modèles sophistiqués pour apprendre les interactions entre les variables

Les contraintes imposées par des modèles linéaires généralisés peuvent conduire à une estimation biaisée de la prime d'assurance. De nature non-paramétrique, les algorithmes d'apprentissage statistique qui seront présentés dans ce chapitre proposent une alternative à la structure linéaire des GLMs, avec un apport dans la qualité de la prédiction.

En particulier, ces méthodes introduisent dans leur structure prédictive les interactions statistiques. Toutefois, à l'exception de l'arbre de classification ou de régression, elles ne sont pas lisibles à cause de l'effet boîte noire. Cet effet dérive du fait que la prédiction d'un modèle est obtenue à partir de l'agrégation des plusieurs estimateurs.

Notions d'apprentissage statistique

À la différence de la statistique classique qui nécessite de formuler des hypothèses sur la structure et la distribution des données, la théorie de l'apprentissage statistique se base que sur une seule hypothèse : les données à prédire, notées Y sont générées de façons identiques et indépendantes par un processus de loi P à partir du vecteur des variables explicatives X .

L'étape d'apprentissage statistique ou machine learning consiste à construire un algorithme de prédiction, une fonction notée \hat{f} , qui approche l'association des variables en entrée à la variable cible, à partir d'une base de données d'apprentissage. Plusieurs critères de quantification de l'erreur commise dans la prédiction sont définis pour choisir le meilleur algorithme : ce choix dépend de la nature des données (variables continues, catégorielles,...) ainsi que de la tâche d'apprentissage (prédiction de la fréquence ou de la sévérité).

Prédiction d'une donnée de sortie

Supposons vouloir prédire une donnée de sortie Y , en observant $D_n = \{(X_{11}, \dots, X_{1m}, Y_1), \dots, (X_{n1}, \dots, X_{nm}, Y_n)\}$ iid de loi P .

- si Y est un ensemble discret de valeurs uniques ($Y = \{0, 1\}$ par exemple), ce problème est dit de classification ;
- si $Y = \mathbb{R}$, le problème est dit de régression.

Définition 6. On appelle **prédicteur** une fonction

$$f : \mathbb{R}^m \rightarrow Y \tag{5.1}$$

qui associe à une observation une valeur de Y
et on appelle prédiction :

$$\hat{Y} = f(X) \tag{5.2}$$

la réponse d'un nouvel individu de caractéristiques $X \in \mathbb{R}^m$,

Fonction de perte et risque de prédiction

Définition 7. On choisit une **fonction de perte** ou coût de la prédiction

$$l : Y \times Y \rightarrow \mathbb{R}_+ \quad (5.3)$$

qui s'applique à l'observation Y et à sa prédiction $\hat{Y} = f(X)$, pour juger la qualité de la prédiction.

Un bon prédicteur minimise la fonction de perte. Pour les données assurantielles, les fonctions de perte les plus utilisées sont :

| Fonction de perte | Définition | Application |
|--|--|---------------------|
| Erreur quadratique $SE(Y, \hat{Y})$ | $Y - \hat{Y}$ | Modèle de sévérité |
| Déviance $D(Y, \hat{Y})$ | $-2 \cdot \ln \left[\frac{\mathcal{L}(\hat{Y})}{\mathcal{L}(Y)} \right]$ | Tous les modèles |
| Déviance poissonnienne $D(Y, \hat{Y})$ | $2 \sum_{i=1}^n \left[Y_i \ln \frac{Y_i}{\hat{Y}_i} - (Y_i - \hat{Y}_i) \right]$ | Modèle de fréquence |
| Déviance gamma $D(Y, \hat{Y})$ | $2 \sum_{i=1}^n \alpha \left[\frac{Y_i - \hat{Y}_i}{\hat{Y}_i} - \ln \frac{Y_i}{\hat{Y}_i} \right]$ | Modèle de sévérité |

où

- $\frac{\mathcal{L}(f(X))}{\mathcal{L}(\hat{Y})}$ est le ratio de vraisemblance ;
- α est le paramètre d'échelle et peut être utilisé comme paramètre de poids.

L'erreur quadratique n'est pas un bon choix pour juger de la qualité d'un modèle de comptage ou d'un modèle de sévérité où la queue de distribution des données est étalée vers la droite. On préfère ainsi la déviance. Des autres métriques d'évaluation de la qualité d'un modèle sont les suivantes :

- le carré moyen des erreurs (MSE) : $\frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$
- l'erreur quadratique moyen (RMSE) : $\sqrt{\frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2}$
- l'erreur absolue moyenne (MAE) : $\frac{1}{n} \sum_{i=1}^n |Y_i - \hat{Y}_i|$

Définition 8. Un des enjeux du machine learning est de comparer différents prédicteurs et choisir celui qui minimise le **risque** de prédiction

$$R(f) = \mathbb{E}_{(X,Y) \sim P} [l(Y, f(X))], \quad (5.4)$$

avec l fonction de perte.

Un prédicteur est dit **optimal** ou oracle si son risque est minimal :

$$f^* \in \arg \min_{f \in F} R(f) \quad (5.5)$$

avec F : classe de tous les prédicteurs.

Règle de prédiction

Si le prédicteur f est estimé à partir des données D_n , on appelle règle de prédiction :

$$\hat{f} : \mathbb{R}^m \times \left(\bigcup_{n \geq 1} (\mathbb{R}^m \times \mathcal{Y}) \right) \rightarrow \mathcal{Y}$$

$$(X, D_n) \rightarrow \hat{f}(X, D_n) \stackrel{\text{notation}}{:=} \hat{f}(X)$$

où \mathcal{Y} est l'espace des prédictions.

Le risque de \hat{f} est ainsi défini comme :

$$R(\hat{f}) = \mathbb{E} [l(Y, \hat{f}(X) | D_n)],$$

Décomposition de l'excès de risque

Dans la pratique on cherchera à déterminer une règle de prédiction, notée $\hat{f}_{\mathcal{F}}(X)$, appartenante à un sous-ensemble \mathcal{F} (un modèle) de toutes les règles de prédiction f pour laquelle le risque de prédiction est proche à celui de l'oracle. En particulier, on voudra minimiser l'Excès de risque $R(\hat{f}_{\mathcal{F}}(X)) - R(f^*)$. On peut montrer que cette quantité se décompose en deux parties :

$$R(\hat{f}_{\mathcal{F}}) - R(f^*) = \underbrace{R(\hat{f}_{\mathcal{F}}) - \inf_{f \in \mathcal{F}} R(f)}_{\substack{\text{erreur stochastique} \\ \text{(Variance)}}} + \underbrace{\inf_{f \in \mathcal{F}} R(f) - R(f^*)}_{\substack{\text{erreur systématique} \\ \text{(Biais)}}$$

- l'erreur stochastique ou d'estimation est l'erreur causé par le choix de la règle de prédiction
- l'erreur systématique ou d'approximation est due au fait que l'on approche la réponse sur une classe limitée de fonctions

Par analogie avec la décomposition biais-variance du risque quadratique, l'erreur d'approximation est souvent appelée *biais* (du modèle \mathcal{F}), et l'erreur d'estimation est souvent appelée *variance*. La validation d'une règle de prédiction est un compromis entre le biais et la variance, pour éviter d'avoir un modèle "trop" adapté aux données d'apprentissage (sur-apprentissage) ou un modèle peu complexe (sous-apprentissage) :

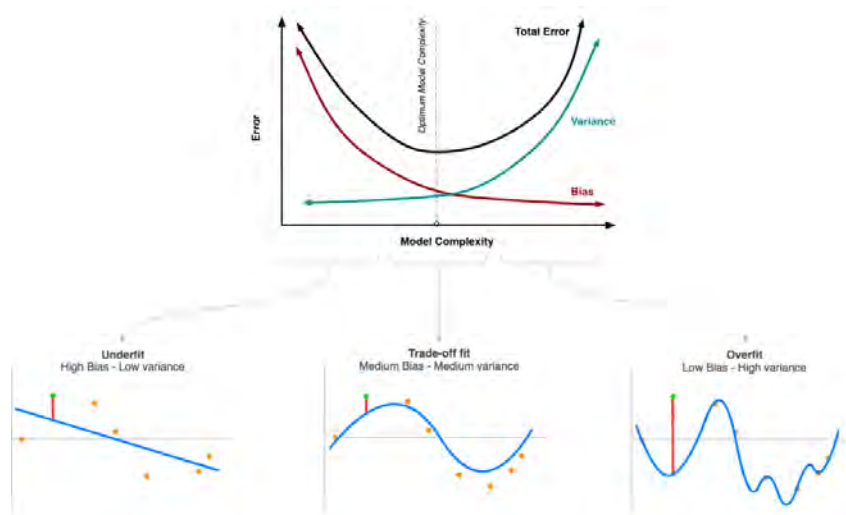


FIGURE 5.1 – Un bon modèle est le résultat d'un compromis entre le biais et la variance.

5.1 CART : Arbres de classification et régression

Définition 9. *Un arbre de décision est une suite de partitions de plus en plus fines de l'ensemble de tous les individus observés vis à vis d'une variable cible.*

La partition est obtenue par itération d'un algorithme. Nous utiliserons l'algorithme CART, introduit par Breiman et al. (1984).

À partir de la base d'apprentissage complète (appelée racine), l'algorithme CART construit un arbre en séparant la population en groupes. Chaque séparation donne naissance aux noeuds-fils jusqu'à qu'un critère d'arrêt soit vérifié et l'arbre aura atteint sa profondeur. La partition finale est ainsi constituée par des feuilles.

À chaque étape, on choisit une variable X_j et un seuil s et on sépare la population en deux groupes :

$$\text{groupe de gauche : si } X_j \leq s, \text{ groupe de droite : si } X_j > s$$

Le choix de X_j et de s se fait par l'algorithme *Glouton (greedy)* qui parcourt toutes les variables et tous les seuils de telle façon que la déviance (dans les problèmes de classification) ou de la somme des carrés (dans les problèmes de régression) de la nouvelle partition est minimale. Dans le cas d'étude, nous considérerons la variable sévérité ou fréquence des sinistres et nous séparerons la base d'apprentissage en minimisant la somme des carrés :

$$SCR_{\text{Racine}} = \sum_{i=1}^n (Y_i - \hat{\mu})^2 \geq$$

$$\geq SCR_1 = \sum_{j:x_j \in \text{Groupe de gauche}} (Y_j - \hat{\mu}_{\text{Groupe de gauche}})^2 + \sum_{j:x_j \in \text{Groupe de droite}} (Y_j - \hat{\mu}_{\text{Groupe de droite}})^2 \geq$$

$$\dots \geq SCR_i = \sum_{k=1}^K \sum_{j:x_j \in C_{ik}} (Y_j - \hat{\mu}_{ik})^2$$

où K est le nombre de feuilles de l'étape i , C_{ik} est le groupe k de l'étape i de l'algorithme et $\hat{\mu}_k$ est la moyenne des estimations du groupe k : $\hat{\mu}_k = \frac{1}{|C_{ik}|} \sum_{j:x_j \in C_{ik}} Y_j$.

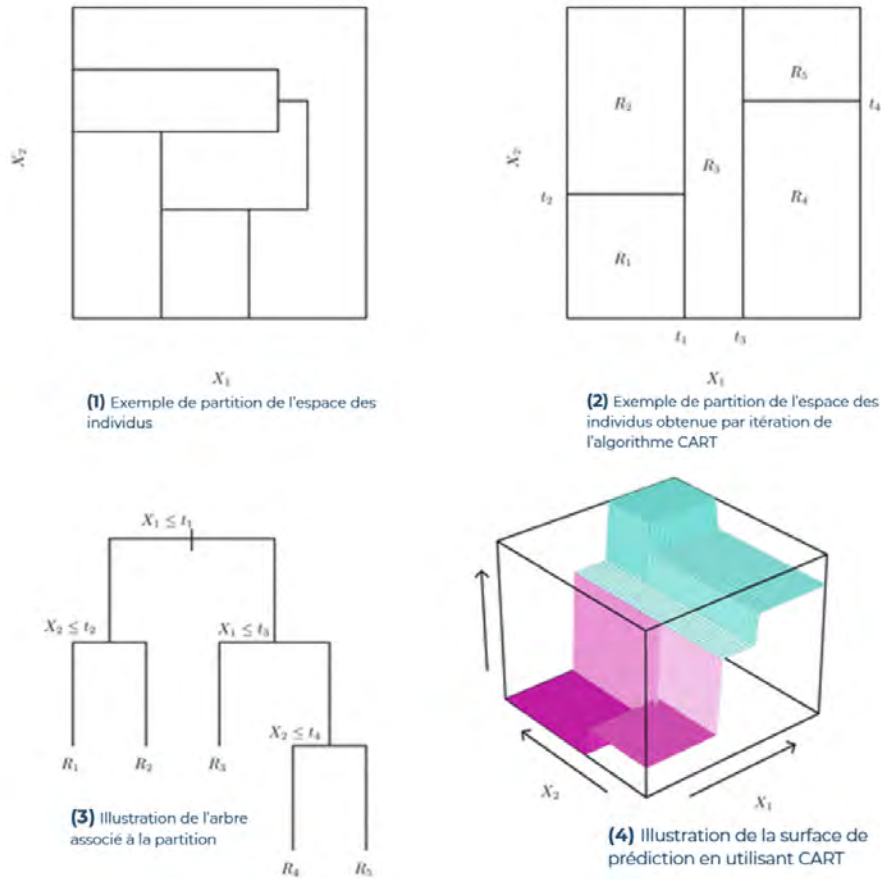


FIGURE 5.2 – L'algorithme CART sépare l'espace des individus pour obtenir une segmentation du risque (variable cible) plus fine (figure (4)).

Limite du CART

L'arbre peut être très complexe et avoir une grande profondeur (nombre d'étape maximale de la racine aux feuilles). Un risque non négligeable est qu'il sur-apprenne des données d'apprentissage. Pour éviter le sur-apprentissage on fait recours à une technique de pénalisation appelé *élagage* : on cherche l'arbre T tel que

$$T = \underset{t:\text{arbre}}{\operatorname{argmin}} \sum_{k=1}^K \sum_{i:x_i \in C_k} (Y_i - \hat{\mu}_k)^2 + \alpha |t|$$

avec $|t|$ nombre de noeuds-feuilles. α paramètre de complexité estimé par validation croisée.

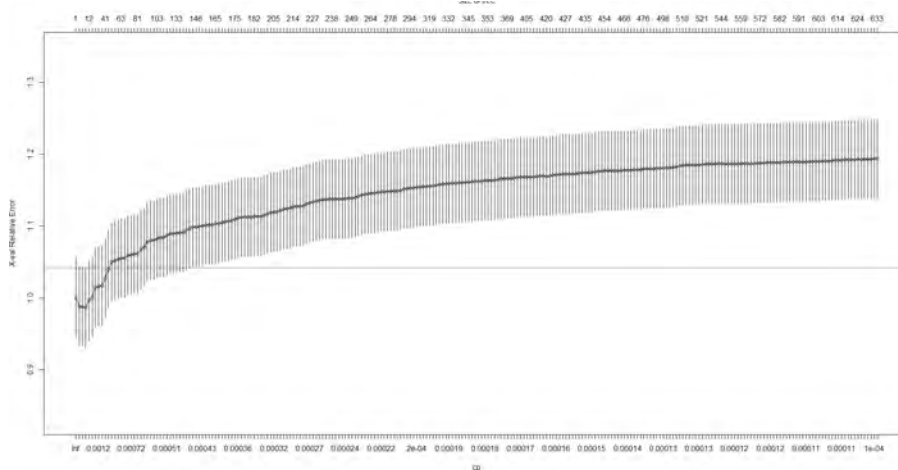


FIGURE 5.3 – Exemple de choix du paramètre de complexité pour le modèle de sévérité : le minimum a une valeur de 0.001702251

La facilité d'interprétation de ce modèle fait son grand succès, néanmoins, les arbres de régression sont relativement instables, très sensibles aux changements au sein du jeu de données. L'ajout des individus ou une permutation de deux observations de l'ensemble d'apprentissage peut engendrer un arbre très différent.

5.2 Agrégation de modèles

Les algorithmes décrits dans cette section sont basés sur des stratégies adaptatives (boosting, gradient boosting) ou aléatoires (bagging, random forest) permettant d'améliorer l'ajustement du modèle final par une combinaison ou agrégation d'un grand nombre de modèles (comme les arbres de classification ou régression) afin d'éviter le sur-apprentissage.

Il y a deux types d'algorithme :

- ceux qui reposent sur une construction aléatoire d'une famille de modèles :
 - la *bootstrap aggregating* qui combine le bootstrap et l'agrégation des arbres (Breiman 1996) ;
 - les forêts aléatoires (*Random forests*) de Breiman (2001) qui proposent une amélioration du bagging en ajoutant une perturbation aléatoire dans la construction des arbres, en jouant sur le mécanisme de sélection des variables de segmentation sur les nœuds. Cela permet de résoudre le problème de corrélation présent dans le bagging, de réduire davantage la variance et de stabiliser le modèle prédictif.
- et les algorithmes qui reposent sur une construction adaptative, déterministe ou aléatoire, d'une famille de modèles : la *boosting* (Freund et Shapiro, 1996).

Les principes du *bagging* ou du *boosting* s'appliquent à toute méthode de modélisation (régression, CART, réseaux de neurones). Dans le cas des arbres de régression en particulier, l'agrégation de modèles se propose comme solution à l'instabilité. Par contre, l'utilisation de ces algorithmes n'est pas aussi interprétable que les CARTs.

5.2.1 Bagging pour bootstrap aggregating

Le *Bagging* est un algorithme qui combine :

- l'agrégation des prédicteurs
- et le bootstrap.

D'abord on produit n échantillons bootstrap (avec remise) à partir de la base d'apprentissage : D_1, \dots, D_n et sur chacun d'eux on construit un arbre avec l'algorithme décrit précédemment.

Puis, la prédiction finale sera obtenue par la moyenne des n prédicteurs (ou vote à la majorité dans les problèmes de classification).

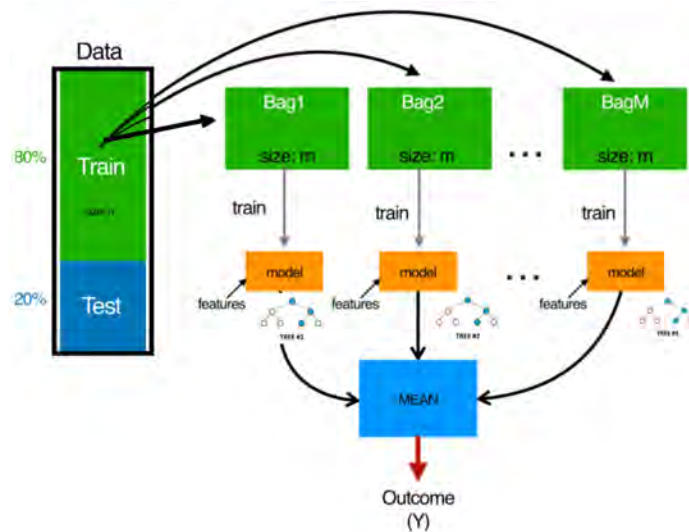


FIGURE 5.4 – Fonctionnement du Bagging par *bootstrap aggregating* : les arbres agrégés étant construits sur toutes les variables peuvent être corrélés, mais la prédiction est plus stable que l’algorithme CART.

Erreur out-of-bag

L’erreur *Out of Bag* est l’erreur de prédiction calculée sur les données qui ne sont pas dans l’échantillon bootstrap (près de 30% de la base¹). Ce principe permet de réduire la variance et donc de réduire l’erreur de prévision.

5.2.2 Forêts aléatoires

L’algorithme des forêts aléatoires (Random Forest) se base sur le *bagging* et il lui ajoute une composante aléatoire afin de décorréler les arbres : pour chaque échantillon bootstrap on tire au hasard le nombre de variables à traiter sur chaque noeud et on choisit la meilleure variable pour segmenter jusqu’aux feuilles et puis on agrège les prédictions par moyenne (ou vote à la majorité dans les problèmes de classification).

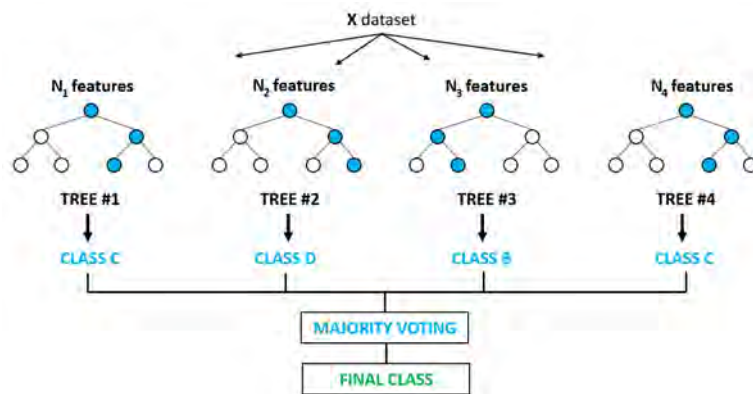


FIGURE 5.5 – Fonctionnement d’une forêt aléatoire.

5.2.3 Boosting

L’idée de base du boosting consiste à améliorer des prédicteurs de faible qualité moyenne (qui prédisent légèrement mieux que le hasard) et de les combiner entre eux pour faire un nouveau prédicteur meilleur sur l’erreur empirique.

Comme le *bagging*, le principe général est de construire une famille de modèles et puis d’agréger les prédictions. Néanmoins, il en diffère sur la façon de construire la famille qui est dans ce cas itérative : chaque modèle est construit sur un échantillon d’apprentissage non bootstrapé et donne plus de poids, lors de l’estimation suivante, aux observations mal ajustées ou mal prédites.

¹la probabilité que une des observations d’un échantillon bootstrapé n’appartient pas à l’échantillon précédent est $(1 - \frac{1}{m})^m \rightarrow e^{-1} \sim 30\%$, où m est la taille de l’échantillon.

L' algorithme corrige les poids des individus au fur et à mesure de l'apprentissage en concentrant ses efforts sur les observations les plus difficiles à ajuster.

Plusieurs algorithmes de *boosting* ont été développés, chacun reposant sur des spécificités :

- la façon de pondérer les observations mal estimées lors de l'itération précédente
- l'objectif, selon le type de la variable à prédire Y : binaire, qualitative, à k classes uniques, réelles ;
- la fonction perte, qui peut être choisie plus ou moins robuste aux valeurs atypiques, pour mesurer l'erreur d'ajustement ;
- et sur la façon de pondérer les modèles de base successifs.

ADABOOST est l'algorithme de base du *boosting* a été introduit dans le cadre d'une classification et puis adaptée à la régression :

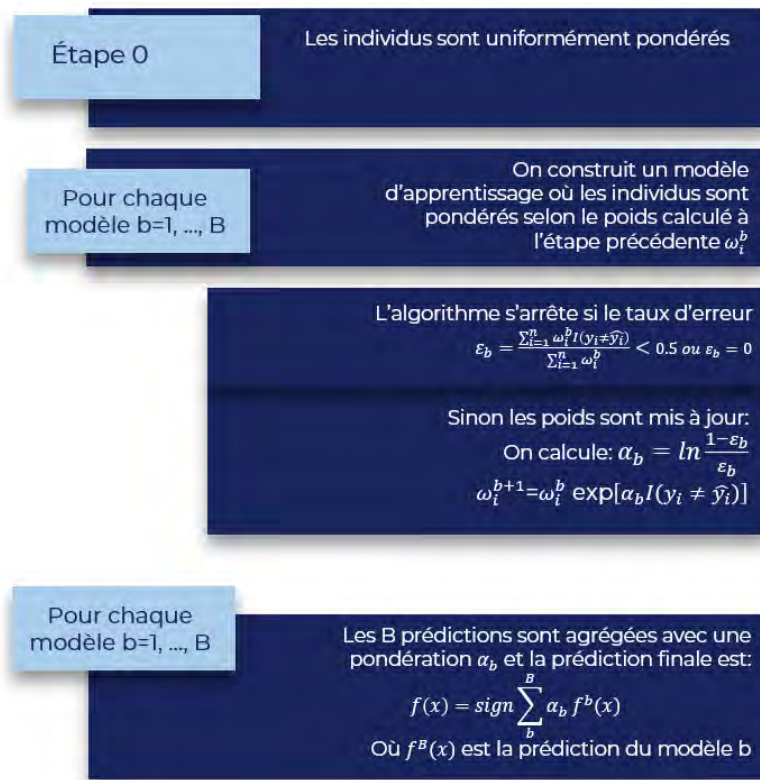


FIGURE 5.6 – Algorithme adaptatif du *boosting* dans une classification.

Dans les problèmes de régression, la fonction de perte binaire ($I(\hat{y} \neq y)$) n'est pas appropriée et on utilise communément la perte quadratique ($\frac{1}{2}(Y - f_i(X))^2$)

Le principe du *boosting* en lui-même ne précise pas comment choisir le modèle et cela dépend des objectifs de l'étude. Dans le cadre du mémoire, nous avons utilisé des techniques utilisant comme modèle de base un arbre de régression.

5.2.4 Gradient tree boosting

Le *Gradient Tree Boosting* est un cas particulier de la méthode du *boosting* dans la régression impliquant pour modèle un arbre de régression de type CART à chaque itération du processus de *Boosting*. La correction des poids s'effectue selon la performance des arbres. La fonction de perte est la somme au carré des résidus d'un arbre antérieur. Plus la valeur du poids affectée est importante, plus l'observation associée pose problème. Le modèle se concentrera ainsi sur cette observation lors de la prochaine itération. C'est la raison pour laquelle le *boosting* est une méthode de descente, où le modèle est corrigé à mesure que les itérations se succèdent, et devient particulièrement adaptée aux données d'apprentissage.

Formalisation du boosting pour la régression avec un modèle d'arbre

Dans un problème de régression partant d'un arbre de régression, on considère la réécriture naturelle de la variable à expliquer Y :

$$Y = A_1(X) + \epsilon_1$$

où ϵ_1 résume l'erreur de prédiction de l'arbre A_1 .

L'idée est de modéliser le résidu $r_1 = Y - A_1(X)$ avec un second modèle A_2 : $r_1 = Y - A_1(X) = A_2(X) + \epsilon_2$ et de l'associer au précédent pour une meilleure prédiction. À l'étape 2 on aura :

$$Y = A_1(X) + r_1 = A_1(X) + A_2(X) + \epsilon_2$$

Le rôle de A_2 est de compenser (additivement) les insuffisances de A_1 , puis on pourra avoir A_3 , etc.

En répétant l'étape B fois, nous obtenons la relation suivante :

$$Y = \sum_{i=1}^B A_i(X) + \epsilon_B$$

L'approximation finale à l'étape B sera ainsi : $\hat{Y} = f_B(X) = \sum_{i=1}^B A_i(X)$.

Descente de gradient

Comme dans le cas de la classification, on cherche un critère d'arrêt pour la définition des poids (selon une fonction de perte), afin d'améliorer les prédictions itérativement. Dans la régression il est coutume d'utiliser la fonction de coût quadratique ($l(Y, f_i(X)) = \frac{1}{2}(Y - f_i(X))^2$). Pour l'ensemble des n observations de la base d'apprentissage le coût sera pour chaque itération b $L(Y, f(X)) = \sum_{j=1}^n l(Y_j, f_i(X_j))$.

L'objectif est ainsi de minimiser cette fonction à chaque itération.

Cette problématique rentre dans le cadre d'application de la descente de gradient.

En apprentissage supervisé, la construction du modèle revient souvent à déterminer les paramètres (du modèle) permettant d'optimiser (maximisant ou minimisant) une fonction objectif.

\forall iteration $b = 1, \dots, B$: nous calculons la dérivée partielle de la fonction de coût l selon f_b , soit le gradient de la fonction de coût quadratique l :

$$\frac{\partial l(Y, f_b(X))}{\partial f_b(X)} = \frac{\partial [\frac{1}{2}(Y - f_b(X))^2]}{\partial f_b(X)} = f_b(X) - Y = \sum_{j=1}^b A_j(X) - Y = -\epsilon_b$$

Nous remarquons alors que les résidus correspondent au gradient négatif à chaque itération b . Par conséquent, en reprenant le modèle additif :

$$f_b(X) = f_{b-1}(X) + \gamma_b A_b(X) = f_{b-1}(X) - \gamma_b \frac{\partial l(Y, f_b(X))}{\partial f_b(X)} = f_{b-1}(X) - \gamma_b \nabla l(Y, f_b(X))$$

où $\gamma_b = \underset{\gamma}{\operatorname{argmin}} \sum_{j=1}^n l(Y_j, f_{b-1}(X_j) - \gamma A_b)$.

La méthode du *gradient tree boosting* procède comme suit :

1. l'arbre de régression initial utilisé est un arbre trivial réduit à la racine ;
2. à partir de ce modèle initial, on effectue pour chaque itération b de 1 à B :
 - le calcul du gradient négatif $\nabla l(Y, f)$ sur la base d'apprentissage, correspondant aux résidus avec une fonction de coût quadratique ;
 - la construction d'un nouvel arbre de régression A_b sur ces résidus ;
 - le calcul de l'estimation de notre modèle obtenu $f_b = f_{b-1}(X) + \gamma_b A_b$, étant égal à 1 dans ce cas précis.
3. On répète ces étapes B fois, jusqu'à convergence vers la solution optimale f_B , minimisant l .

5.2.5 Extrem Gradient Boosting

En 2016 Chen et Guestrin [6] ont proposé une amélioration du Gradient Boosting (qui par rapport au Tree Gradient boosting n'utilise pas forcément l'arbre comme modèle initial) avec l'*Extrem gradient boosting (xgboost)*. Cette version du boosting est très populaire dans les compétitions Kaggle pour les modèles prédictifs, étant très performante, mais en contrepartie, la complexité due au nombre des paramètres rend nécessaire l'optimisation de l'algorithme.

Les avantages majeurs que xgboost a par rapport aux modèles de gradient boosting sont :

- de nature calculatoire : il est particulièrement apprécié pour sa capacité de faire des calculs parallèles sur une seule machine.
- de nature technique : cet algorithme ajoute un terme de régularisation à la fonction objectif, pour éviter le sur-apprentissage :

$$\mathcal{L} = \underbrace{\sum_{i=1}^n l(\hat{y}_i, y_i)}_{\text{la fonction perte convexe différentiable}} + \underbrace{\sum_{m=1}^M l(\Omega(\delta_m))}_{\text{terme de régularisation}}$$

où $|\delta|$ est le nombre de feuilles de l'arbre de régression δ , w est le vecteur des valeurs attribuées à chacune de ses feuilles.

- sur le traitement des données : à la différence des autres méthodes de machine learning précédemment décrites, xgboost a la capacité de gérer des données manquantes en proposant à chaque division une direction par défaut si une donnée est manquante.

Une limite de l'algorithme de Boosting est toutefois sa sensibilité au bruit et aux valeurs extrêmes. Ceci est particulièrement dû à la concentration de l'algorithme sur des populations limitées à chaque itération et à la pondération des modèles par leurs qualités d'ajustements.

Optimisation des paramètres

Un bon paramétrage des modèles dans l'étape de validation du modèle peut améliorer la qualité du modèle, afin qu'il soit bien ajusté aux données, sans sur-apprendre. Il existe deux type de paramètres :

- des paramètres prioritaires (ex. nombre d'arbres, nombre de feuilles dans les algorithmes CART) dont le calibrage a la fonction de réduire le biais de prédiction ou le sur-apprentissage ;
- des hyper-paramètres, qui ont moins d'impact sur la précision du modèle, et plus d'importance sur sa complexité ou sa vitesse de convergence.

Les méthodes d'optimisation que nous avons exploré sont la *validation croisée V-fold* et *Random Search*.

Validation Croisée V-Fold

La validation croisée V-Fold est une méthode d'estimation de fiabilité d'un modèle fondée sur une technique d'échantillonnage.

Elle est utilisée pour estimer le risque de prédiction ou pour estimer les paramètres qui minimisent ce risque.

Algorithme

- On sépare les données en V sous-ensembles (dans la plupart des cas $V=5$)
- $\forall v = 1, \dots, V$ on réitère l'algorithme suivant :
 - on sélectionne l'échantillon v comme ensemble de validation et les autres $V - 1$ comme ensemble d'apprentissage
 - on construit respectivement l'estimateur sur la base d'apprentissage et on évalue le risque de prédiction sur l'échantillon v :

$$(\hat{f}^{(-v)}, \hat{R}^{(v)}(\hat{f}^{(-v)}))$$

- L'erreur de prédiction est la moyenne des V risques empiriques : $\hat{R}^{(CV)}(\hat{f}) = \frac{1}{V} \sum_{v=1}^V \hat{R}^{(v)}$

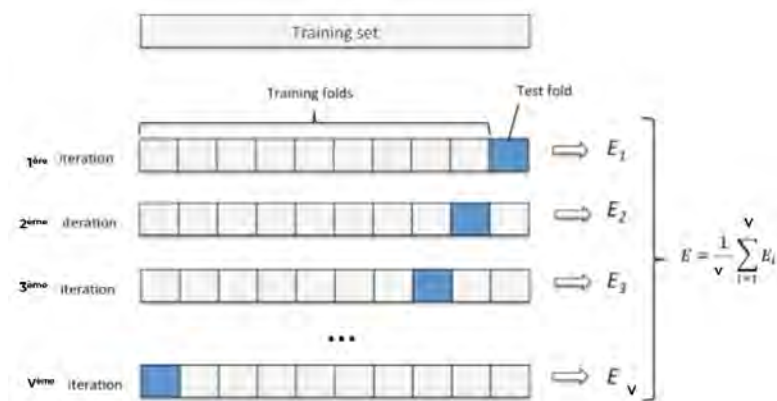


FIGURE 5.7 – Représentation du mécanisme de validation croisée 5-fold.

Si on veut choisir un paramètre α par validation croisée, on considère une grille de valeurs de $\alpha : \alpha_1, \dots, \alpha_n$ possibles et on applique pour chacun d'eux l'algorithme de validation croisée. Le α optimal est celui qui minimise le risque empirique :

$$\alpha = \underset{\hat{\alpha}}{\operatorname{argmin}} \{ \hat{R}_{\hat{\alpha}_1}^{(CV)}, \dots, \hat{R}_{\hat{\alpha}_n}^{(CV)} \}$$

Random Search

Random Search une méthode d'optimisation qui permet de tester une série de paramètres et de comparer les performances afin d'en déduire le meilleur paramétrage.

Elle croise chacune des hypothèses (intervalles pour chaque paramètre) et va créer un modèle pour chaque combinaison de paramètres.

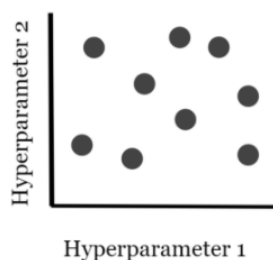


FIGURE 5.8 – Exemple simplifié de fonctionnement de *Random Search*.

Son utilisation est particulièrement adaptée aux modèles où il n'est pas possible de simuler toutes les combinaisons possibles des paramètres car ils ont un trop grand nombre de paramètres en entrée, comme xgboost, ou lorsque la taille de la base d'apprentissage rend l'optimisation très lente.

Impact d'un modèle de machine learning dans la tarification IARD

Les techniques présentées dans ce chapitre offrent plusieurs pistes d'amélioration du modèle linéaire généralisé :

- du point de vue du risque de prédiction, les prédicteurs ne se basent plus sur l'estimation par maximum de vraisemblance comme dans les GLMs, mais ils étendent leur construction à plusieurs fonctions de perte ;
- de plus, ces modèles n'imposent pas une structure linéaire des variables explicatives, en ajoutant implicitement des interactions parmi les variables.

Leur principal point de faiblesse est toutefois l'interprétabilité des résultats : à l'exception du CART, il est difficile de visualiser la structure tarifante sous-jacente le modèle de machine learning.

Pour pallier cette limite nous allons proposer dans la partie suivante des outils d'*ouverture* de ces modèles.

Troisième partie

Analyse de sensibilité et XAI

Introduction

Un état des lieux des modèles de tarification traditionnels les plus sophistiqués a été fait dans la *partie II* :

- le modèle GLM est caractérisé par une structure linéaire, interprétable et simple à implémenter ;
- le modèle CART partitionne les individus en groupes homogènes, il est interprétable, mais il n'est pas utilisé fréquemment comme modèle principal de pricing notamment à cause du risque de sur-apprentissage et de sa non robustesse ;
- la forêt aléatoire et le modèle xgboost sont les plus complexes parce qu'ils utilisent une agrégation des prédictions, mais ils sont difficilement interprétables et ils sont aussi difficilement implémentés.

La façon de lier les deux types de modèles (traditionnel et sophistiqué) que nous préconisons avec ces travaux est l'amélioration du premier en ajoutant les interactions statistiques introduites par le deuxième. En effet, ces interactions manifestent la complexité du modèle, en permettant ainsi de faire profiter des bonnes propriétés prédictives des GLMs. En d'autres termes, afin de pallier le manque de complexité des modèles linéaires généralisés, l'ajout des interactions permet d'approcher des structures plus complexes de sinistralité.

| Modèle | Complexité | Interprétabilité | Mise en place opérationnelle |
|--------|------------|------------------|------------------------------|
| GLM | + | +++ | +++ |
| CART | ++ | +++ | + |
| RF | +++ | + | + |
| XGB | +++ | + | + |

FIGURE 5.9 – Classification des méthodes de tarifications analysées par les critères de complexité, interprétabilité et mise en place opérationnelle.

De contraintes d'interprétation et d'explicabilité amènent notre recherche vers des techniques d'*ouverture* des modèles *black box*. Nous présenterons dans les chapitres suivants le cadre théorique permettant d'estimer des interactions parmi les variables, à partir de n'importe quel modèle et, en particulier, à partir des modèles complexes de forêt aléatoire et Xgboost.

La détection des interactions fait partie d'une problématique plus générale : la quantification de l'importance des variables, qui intéresse deux domaines :

- l'analyse de sensibilité, dont les outils principaux sont :
 1. les Indices Sobol, pour les modèles à entrées indépendantes,
 2. les Effets Shapley, pour les modèles à entrées dépendantes.
- l'Explainable AI (XAI), dont les outils principaux sont les *Shap Importances*.

Historiquement, les premiers indices de sensibilité, qui nécessitaient de peu d'hypothèses sur la régularité sur la fonction f associant des variables explicatives à une réponse, ont été introduits par Sobol en 1993. Ils sont basés sur la décomposition fonctionnelle de la variance.

En 2014, Owen introduit d'autres indices de sensibilité basés sur la décomposition de la variance, appelés *Shapley Effects* à partir des notions propres à la théorie des jeux. Contrairement aux *indices de Sobol*, ces nouveaux indices d'Owen sont bien adaptés lorsque les variables d'entrée sont dépendantes. Par ailleurs, ces indices sont exploités dans un cadre plus général que l'analyse de sensibilité, en couvrant aussi le domaine de l'interprétabilité des modèles d'intelligence artificielle (XAI).

Chapitre 6

Analyse de sensibilité

Dans ce chapitre nous introduisons le domaine de l'analyse de sensibilité : sa définition, son impact dans la compréhension d'un modèle statistique, ainsi que la classification des principales méthodes.

Introduction

L'analyse de sensibilité étudie comment des perturbations sur les variables explicatives du modèle engendrent des perturbations sur la variable expliquée.

Lors de la construction et de l'utilisation d'un modèle numérique de simulation, les méthodes d'analyse de sensibilité (AS) sont des outils précieux. Elles permettent notamment de déterminer quelles sont les variables d'entrée du modèle qui contribuent le plus à une quantité d'intérêt donnée en sortie du modèle. Elles expliquent aussi quelles sont celles qui n'ont pas d'influence et quelles sont celles qui interagissent au sein du modèle.

Il y a deux approches :

- les **méthodes d'analyse locale** [20] évaluent quantitativement l'impact d'une petite variation autour d'une valeur de départ des entrées, en associant une partie de variance de la sortie à une entrée ou à un ensemble d'entrée. Malgré leur avantage d'avoir un temps de calcul raisonnable et d'être très intuitives, elles sont limitées aux hypothèses de linéarité, de normalité et de variations locales.
- et les **méthodes d'analyse globale** [2] s'intéressent à la variabilité de la sortie du modèle dans l'intégralité de son domaine de variation (Saltelli et al.[1]). Ces méthodes ont été développées à partir de la fin des années 1980 (Bertrand Iooss et Lemaitre), pour pallier les limites des méthodes locales.

Dans le cadre de ce mémoire, l'approche retenue et présentée pour l'analyse de sensibilité est celle globale, puisque l'on a besoin du minimum d'hypothèses de régularité du modèle de prédiction, sans se limiter au cadre linéaire ou normal.

6.1 Comment évaluer l'impact des entrées aléatoires sur la sortie d'un modèle mathématique

L'analyse de sensibilité globale permet d'analyser un modèle mathématique en étudiant l'impact de la variabilité des facteurs d'entrée du modèle sur la variable de sortie.

Il est possible de grouper les méthodes d'analyse de sensibilité (globale) en trois classes :

1. *Screening ou Criblage*, qui consistent en une analyse qualitative de la sensibilité de la variable de sortie aux variables d'entrée [1], c'est un tri grossier des entrées les plus influentes parmi un grand nombre.
2. Définition des *mesures d'importance* : il s'agit d'indices quantitatifs donnant l'influence de chaque entrée.
3. Les outils d'*exploration du modèle*, qui mesurent les effets des entrées sur tout leur domaine de variation.

En déterminant les entrées responsables de cette variabilité à l'aide d'indices de sensibilité, l'AS permet de prendre les mesures nécessaires :

- pour diminuer la variance de la sortie si celle-ci est synonyme d'imprécision,
- ou d'alléger le modèle en fixant les entrées dont la variabilité n'influe pas la variable de sortie.

Des approches différentes sont proposées selon l'hypothèse d'indépendance des variables d'entrée ou non.

Apport d'une analyse de sensibilité

L'analyse de sensibilité étudie comment la réponse du modèle réagit aux variations de ses variables d'entrée. Cette méthode propose des nouveaux axes d'amélioration des modèles, en améliorant les éléments suivants :

1. **La qualité du modèle par rapport au phénomène/processus modélisé**

Si une variable d'entrée habituellement connue comme *non influente* ressort très influente, il sera nécessaire de remettre en cause la qualité du modèle et l'impact réel des variables d'entrée.

Par exemple, la caractéristique d'une maison pourrait influencer la sinistralité grave et non celle attritionnelle : dans ce cas, il faut calibrer les modèles attritionnel et grave avec des variables d'entrée différentes.

2. **La prédiction lorsque la variabilité est synonyme d'imprécision**

Si cette variabilité est synonyme d'imprécision sur la valeur prédite de la sortie, il sera alors possible d'améliorer la qualité de la réponse du modèle à moindre coût.

Il suffirait d'identifier les variables explicatives qui causent le plus de variabilité dans la réponse et réduire leur variabilité.

3. **La sélection des variables** On peut envisager de considérer les variables les moins influentes comme des paramètres *déterministes*, en les fixant à leur espérance.

4. **L'effet d'interaction** Extrapoler des nouvelles informations par le fait que des modalités de variables distinctes sont prises en même temps.

Ce dernier point sera en particulier l'usage qu'on fera de l'analyse de sensibilité dans le cas d'application assurantielle.

Choix de la méthodes d'AS globale

Il existe une multitude de méthodes globales qui se distinguent pour le type d'étude (qualitatif, quantitatif), le type de décomposition de la sortie, les données considérées et les objectifs.

Nous nous sommes restreints à présenter et appliquer la méthode basée sur la décomposition de la variance de la sortie, introduite par Sobol (1993).

L'autre méthode de référence est la méthode de Morris, qui est adaptée à un grand nombre de variables et qui appartient à la classe du criblage.

TABLE 6.1 – Synthèses des méthodes d'analyse de sensibilité globale

| Classe | Méthode |
|-------------------------------------|---|
| Criblage | <ul style="list-style-type: none">• Criblage à très grande dimension• Plans d'expériences usuels• Méthode de Morris |
| Définitions de mesures d'importance | <ul style="list-style-type: none">• Méthodes basées sur la régression linéaire• Méthodes basées sur des tests statistiques• Méthodes basées sur la décomposition de la variance fonctionnelle |
| Exploration du modèle | <ul style="list-style-type: none">• Méthodes de lissage• Métamodèles |

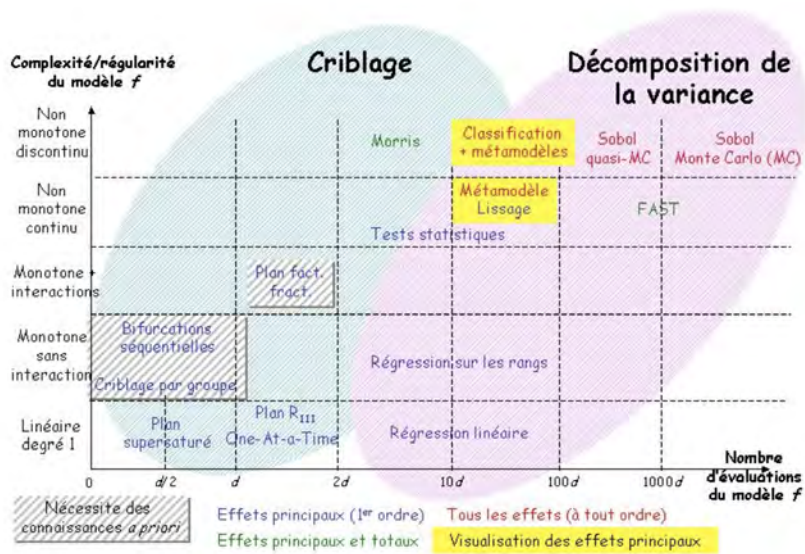


FIGURE 6.1 – Synthèse détaillée des méthodes d’AS. Source : Bertrand Iooss : Revue sur l’analyse de sensibilité globale de modèles numériques

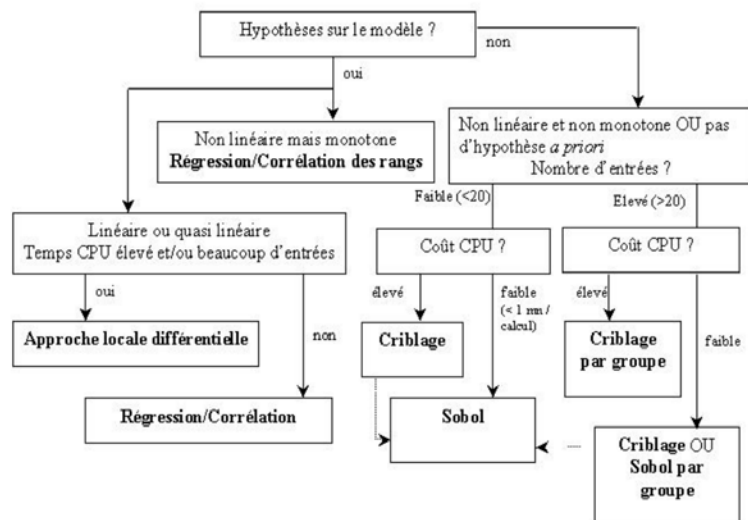


FIGURE 6.2 – Arbre de décision pour le choix de la méthode d’AS appropriée (d’après de Rocquigny et al.).

Approximation du modèle par un métamodèle

En dehors de fonctions analytiques simples, avec un très faible nombre d'entrées ou de codes de calcul demandant peu de ressources en temps, les coûts d'estimation des indices de Sobol sont souvent très élevés ou inatteignables.

Pour accélérer le calcul des indices de Sobol, les praticiens de l'analyse de sensibilité passent par des *Métamodèles* : des catégories de méthodes d'approximation du modèle numérique, qui réduisent les temps de calcul.

Une autre solution pour accélérer les calculs, en complément de ces modèles auxiliaires, est la réduction de la taille de l'échantillon avec des méthodes d'échantillonnage raffinées.

A la différence des domaines d'application historiques de l'analyse de sensibilité, comme l'ingénierie aérospatiale, nous ne possédons pas une fonction analytique représentant la sortie du modèle, mais une base des réalisations des variables aléatoires. Le passage par un métamodèle est ainsi naturel dans la cas où l'on "subit" une base d'observations. Nous traiterons les modèles de tarifications comme des métamodèles.

Le terme *métamodèle* fait référence à la "surface de réponse", outil connu depuis longtemps dans le domaine de la planification d'expériences. L'objectif de cette méthode est de construire une fonction pour simuler le comportement d'un phénomène physique ou chimique dans le domaine de variation des variables influentes, à partir d'un certain nombre d'expériences (Box & Draper).

Ces modèles simplifiés substituent l'exécution de codes de calcul nécessitant trop de temps d'exécution ou trop de ressources (Downing et al., Sacks et al., Fang et al.).

Tout comme dans les modèles de prédictions vus dans les chapitres précédents, le nombre de simulations nécessaires dépend :

- de la complexité du code et du scénario qu'il modélise,
- du nombre de variables d'entrée,
- et de la qualité d'approximation souhaitée.

Mesures d'importance et Indices de Sobol

Dans ce chapitre nous allons présenter les mesures d'importance d'une analyse de sensibilité.

On parle de mesures d'importance car ces techniques permettent une hiérarchisation précise de l'influence sur la sortie de toutes les entrées, contrairement aux techniques de criblage qui ont pour but de détecter les entrées non influentes.

L'importance des variables dépend de la distribution des variables, même si elles sont indépendantes. En voici les deux principales raisons :

- Poids d'une variable : la distribution d'une variable pourrait accroître ou atténuer son importance. Imaginons que nous étudions l'importance de l'âge et d'autres facteurs dans la fréquence de la sinistralité : sur une population jeune, nous nous attendons à ce que l'âge prédomine ; tandis qu'avec les adultes, un autre facteur pourrait être le plus important.
- Interactions : lorsqu'il y a de fortes interactions dans le modèle parmi deux ou plusieurs variables, une variation de la distribution d'une variable peut augmenter ou atténuer l'importance d'une autre.

Il y a trois méthodes de définition des mesures d'importance :

1. Les méthodes basées sur la régression linéaire, où l'on approxime un modèle par un modèle linéaire,
2. Les méthodes basées sur les tests statistiques,
3. Les méthodes basées sur la décomposition fonctionnelle.

En particulier, nous allons appliquer les méthodes basées sur la décomposition fonctionnelle de la variance.

6.2 Modèles à entrées indépendantes

Considérons un modèle mathématique avec :

- $\mathbf{X} \subset \mathbb{R}^p$ ensemble de variables d'entrée aléatoires,
- f , une fonction qui associe les variables explicatives à une réponse,
- Y variable de sortie aléatoire.

$$\begin{aligned} f : \mathbb{R}^p &\rightarrow \mathbb{R} \\ \mathbf{X} &\mapsto Y = f(\mathbf{X}) \end{aligned}$$

La forme analytique de f peut être très complexe ou ne pas être connue.

6.2.1 Indices de sensibilité pour modèles linéaires et/ou monotones

Nous étudierons tout d'abord l'hypothèse de linéarité du modèle, qui définit des indicateurs intuitifs de la sensibilité de la sortie du modèle aux variables d'entrée indépendantes. Puis nous généralisons ces indicateurs au cas moins restrictif des modèles monotones.

Exemple 1. Modèle linéaire La décomposition de la variance dans le cas où

$$Y = \beta_0 + \sum_{i=1}^{i=p} \beta_i X_i$$

est :

$$\text{Var}(Y) = \sum_{X_i \perp\!\!\!\perp}^{i=p} \beta_i^2 \text{Var}(X_i) \quad (6.1)$$

Définition 10. $\beta_i^2 \text{Var}(X_i)$ est la **part de variance** due à la variable X_i .

Définition 11. SRC (Standardized Regression Coefficient)

Une mesure de la **sensibilité de la variable de sortie Y à la covariable X_i** est le rapport de la part de variance due à X_i sur la variance totale :

$$SCR_i = \frac{\beta_i^2 \text{Var}(X_i)}{\text{Var}(Y)}$$

Cet indice SRC, toujours positif ($SCR \in [0, 1]$), est en outre le carré du coefficient de corrélation linéaire entre la réponse du modèle et ses variables d'entrée.

L'indice SRC est équivalent, au carré près, au coefficient de corrélation linéaire entre la réponse du modèle et ses variables d'entrée $\rho_{X_i, Y}$, qui est parfois utilisé comme indicateur de sensibilité. En effet, pour le modèle linéaire :

$$\text{Cov}(X_i, Y) = \beta_i \text{Var}(X_i)$$

d'où

$$\rho_{X_i, Y} = \frac{\text{Cov}(X_i, Y)}{\sqrt{\text{Var}(X_i) \text{Var}(Y)}} = \beta_i \sqrt{\frac{\text{Var}(X_i)}{\text{Var}(Y)}}$$

Donc $\rho_{X_i, Y}^2 = SCR_i$

On préférera utiliser l'indice SRC plutôt que celui de corrélation linéaire, puisque étant positif, il est plus approprié lorsque l'on quantifie des indicateurs de variabilité totale.

Définition 12. PCC (coefficient de corrélation partielle)

Une mesure de la **sensibilité de la variable de sortie Y à la covariable X_i** est le rapport de la part de variance due à X_i sur la variance totale :

$$PCC_i = PCC_i(X_i, Y) = \rho(Y - \hat{Y}, X_i - \hat{X}_i)$$

avec \hat{Y} la prévision du modèle linéaire dans lequel X_i n'est pas présent et \hat{X}_i la prévision du modèle linéaire qui exprime X_i en fonction des autres entrées.

Les PCCs permettent d'éliminer l'influence des autres variables et sont donc adaptés quand les variables d'entrée sont corrélées (Saltelli et al.).

Si on juge l'hypothèse de linéarité acceptable (par exemple si le coefficient de détermination $R^2 > 0.8$, ce qui signifie que plus de 80% de la variabilité de la sortie est expliquée par une relation linéaire), alors les indices de sensibilité Pearson (corrélation linéaire), SRC et PCC sont utilisables.

Dans le cas où la relation entre X et Y n'est pas linéaire mais monotone, les coefficients de corrélation et de régression basés sur les rangs (Spearman, SRRC, PRCC) peuvent être utilisés (Saltelli et al.).

Nous allons étendre cette définition dans le cas d'une fonction quelconque f , dont on ne connaît pas la forme analytique.

6.2.2 Méthodes basées sur des tests statistiques

A partir d'un échantillon *i.i.d.*, d'autres techniques d'AS peuvent être utilisées.

Par exemple, pour chaque entrée, un découpage en classes de valeurs équiprobables permet d'obtenir plusieurs échantillons de données. Des tests statistiques peuvent alors être appliqués pour mesurer l'homogénéité des populations entre les classes : moyennes communes (CMN) basées sur un test de Fisher ou médianes communes (CMD) basées sur un test du χ^2 .

Ces méthodes ne requièrent pas d'hypothèses sur la monotonie de la sortie en fonction des entrées mais présentent l'inconvénient d'être peu intuitives comparativement aux méthodes de régression.

6.2.3 Méthodes basées sur la Décomposition fonctionnelle de la variance

Pourquoi la variance ?

L'analyse de sensibilité est l'étude de la façon dont l'incertitude de la sortie d'un modèle peut être attribuée à l'incertitude dans ses entrées.

Ainsi, la variance, en tant que mesure de dispersion, émerge naturellement comme quantification de l'incertitude de la sortie d'un modèle. En effet, plutôt que d'étudier comment la moyenne d'une sortie varie par rapport à une perturbation des entrées, il est plus intéressant d'observer combien chaque variable contribue à la dispersion de la sortie, c'est-à-dire établir quelle variable est à améliorer pour réduire la dispersion de la sortie.

Décomposition de la variance :

L'importance d'une variable d'entrée X_i est mesurée par la partie de variance de Y dont elle est responsable. Nous dirons qu'elle est d'autant plus importante que la variance de la sortie Y en fixant X_i petite : en d'autres termes, nous étudions à combien la variance de Y décroît si on fixe la variable X_i à une valeur x_i^* : $\text{Var}(Y|X_i = x_i^*)$.

Pour ne pas biaiser cette estimation avec le choix de x_i^* , nous allons plutôt considérer son espérance :

$$\mathbb{E}(\text{Var}(Y|X_i))$$

Ainsi, **plus la variable X_i sera importante vis-à-vis de la variance de Y , plus cette quantité sera petite.**

De même, par la formule de la *variance totale* :

$$\begin{aligned} \text{Var}(Y) &\stackrel{\text{def}}{=} \mathbb{E}(Y^2) - \mathbb{E}^2(Y) = \\ &\stackrel{\text{espérance totale}}{=} \mathbb{E}(\mathbb{E}(Y^2|X_i)) - \mathbb{E}^2(\mathbb{E}(Y|X_i)) = \\ &\stackrel{\text{déf Variance}}{=} \mathbb{E}(\text{Var}(Y|X_i) + \mathbb{E}^2(Y|X_i)) - \mathbb{E}^2(\mathbb{E}(Y|X_i)) = \\ &\stackrel{\text{linearité espérance}}{=} \mathbb{E}(\text{Var}(Y|X_i)) + (\mathbb{E}(\mathbb{E}^2(Y|X_i)) - \mathbb{E}^2(\mathbb{E}(Y|X_i))) = \\ \text{Var}(Y) &= \mathbb{E}(\text{Var}(Y|X_i)) + \text{Var}(\mathbb{E}(Y|X_i)) \end{aligned} \tag{6.2}$$

la quantité $\text{Var}(\mathbb{E}(Y|X_i))$ sera d'autant plus grande que la variable X_i sera importante vis-à-vis de la variance de Y .

Afin d'utiliser un indicateur normalisé, nous définissons l'indice de sensibilité suivant.

Définition 13. Indice de sensibilité de premier ordre - importance measure

Sobol [17] a défini l'indice de sensibilité de premier ordre de Y à X_i :

$$S_i = \frac{\text{Var}(\mathbb{E}(Y|X_i))}{\text{Var}(Y)}$$

Cet indice appelé *indice de Sobol*, ou *indice de sensibilité de premier ordre par Sobol*, *correlation ratio par McKay*, ou encore *importance measure*, quantifie la part de variance de Y due à la variable X_i .

Remarque Dans le cas du modèle linéaire, cet indice de sensibilité est égal à l'indice SRC, puisque $\text{Var}(\mathbb{E}(Y|X_i)) = \text{Var}(\beta_i X_i) = \beta_i^2 \text{Var}(X_i)$

Décomposition fonctionnelle de Sobol :

Sobol a introduit les indices de sensibilité précédemment définis (Déf 13) en décomposant la fonction f du modèle par la somme de fonctions de dimension croissante. Cette décomposition est un cas particulier de la décomposition réalisée en analyse de variance, connue sous le nom anglais *ANOVA decomposition*, présentée notamment par Efron et Stein ¹.

¹B. Efron and C. Stein. The jackknife estimate of variance. The Annals of Statistics, 9(3) :586–596, 1981.

Théorème 1. Décomposition ANOVA Soit f une fonction définie de $\mathcal{U}[0, 1]^n$ dans R . La fonction f peut se décomposer en somme de 2^n fonctions de dimensions croissantes :

$$Y = f(X_1, \dots, X_p) = \mathbb{E}(Y|X_1, \dots, X_p) = f_0 + \sum_{i=1}^p f_i(X_i) + \sum_{1 \leq i < j \leq p} f_{ij}(X_i, X_j) + \dots + f_{1, \dots, p}(X_1, \dots, X_p)$$

où

$$\begin{aligned} V_i &= \text{Var}(\mathbb{E}(Y|X_i)) \\ f_0 &= \mathbb{E}(Y) \text{ (constante)} \\ f_i(X_i) &= \mathbb{E}(Y|X_i) - \mathbb{E}(Y) \\ f_{i,j}(X_i, X_j) &= \mathbb{E}(Y|X_i, X_j) - \mathbb{E}(Y|X_i) - \mathbb{E}(Y|X_j) + \mathbb{E}(Y) \\ f_{i,j,k}(X_i, X_j, X_k) &= \mathbb{E}(Y|X_i, X_j, X_k) - \mathbb{E}(Y|X_i, X_j) - \mathbb{E}(Y|X_i, X_k) - \mathbb{E}(Y|X_j, X_k) + \\ &\quad + \mathbb{E}(Y|X_i) + \mathbb{E}(Y|X_j) + \mathbb{E}(Y|X_k) - \mathbb{E}(Y) \\ &\dots \end{aligned}$$

Remarque L'hypothèse $X_i \sim \mathcal{U}[0, 1]$ n'est pas restrictive car $X_i \sim F_i$ (fonction de répartition), $F_i^{-1} \sim \mathcal{U}[0, 1]$, où F_i^{-1} est l'inverse généralisée de la fonction de répartition.

En ajoutant des contraintes d'orthogonalité parmi les termes de la somme de la décomposition ANOVA, cette décomposition est unique (Sobol, 1993) :

Théorème 2. Décomposition fonctionnelle de Sobol Sous les mêmes hypothèses que le théorème 1 et

- $\forall 1 \leq r \leq n$
- $\forall 1 \leq i_1 < i_2 < \dots < i_r \leq m,$
- $\forall 1 \leq s \leq r,$
- $\int_0^1 f_{i_1, \dots, i_r}(x_{i_1}, \dots, x_{i_r}) dx_{i_s} = 0$ (condition d'orthogonalité)

la décomposition ANOVA est unique.

La variance de Y , $\text{Var}(Y)$, peut alors se décomposer selon le théorème suivant.

Théorème 3. Décomposition de Sobol de la variance.

La variance du modèle à entrées indépendantes se décompose en :

$$V := \text{Var}(Y) = \sum_{i=1}^p V_i + \sum_{1 \leq i < j \leq p} V_{ij} + \dots + V_{1 \dots p}$$

où

$$\begin{aligned} V_i &= \text{Var}(\mathbb{E}(Y|X_i)) \\ V_{ij} &= \text{Var}(\mathbb{E}(Y|X_i, X_j)) - V_i - V_j \\ V_{ijk} &= \text{Var}(\mathbb{E}(Y|X_i, X_j, X_k)) - V_{ij} - V_{ik} - V_{jk} - V_i - V_j - V_k \\ &\dots \\ V_{1 \dots p} &= V - \sum_{i=1}^p V_i - \sum_{1 \leq i < j \leq p} V_{ij} - \dots - \sum_{1 \leq i_1 < i_{p-1} \leq p} V_{i_1 \dots i_{p-1}} \end{aligned}$$

Définition 14. À partir de la décomposition pour définir des **indices de sensibilité d'ordre k** .

$$S_{i_1 \dots i_k} = \frac{V_{i_1 \dots i_k}}{V}$$

Ces indices expriment la sensibilité de la variance de Y à l'interaction des variables X_{i_1}, \dots, X_{i_k} , c'est-à-dire **la sensibilité de Y aux variables X_{i_1}, \dots, X_{i_k} qui n'est pas prise en compte dans l'effet des variables seules.**

Interprétation des indices de Sobol :

L'interprétation des indices de Sobol est très intuitive, car grâce à la décomposition fonctionnelle de la variance, leur somme est égale à 1, et étant tous positifs, plus l'indice sera grand (proche de 1), plus la variable (ou le groupe des variables) aura d'importance.

Considérant la formule de la variance totale (Eq. 6.2), par exemple, si $\mathbb{E}(\text{Var}(Y|X_i))$ est faible, fixer X_i réduit la variabilité de Y donc X_i est influente (l'indice de Sobol de X_i est proche de 1).

L'indice du second ordre S_{ij} exprime la sensibilité du modèle à l'interaction entre les variables X_i et X_j , et ainsi de suite pour les ordres supérieurs.

Par ailleurs, si l'objectif de l'analyse de sensibilité est de savoir quelle variable ou quel groupe de variables conduit à la plus grande réduction de la variance de Y , Saltelli et Tarantola affirment que les indices de sensibilité d'ordre 1 sont toujours les indicateurs à utiliser en présence de corrélation : en effet, si en présence de corrélation l'indice d'ordre 1 S_i n'exprime plus uniquement la sensibilité à une variable X_i mais également une partie de sensibilité aux variables avec lesquelles elle est corrélée, fixer X_i conduit à jouer sur la distribution des variables avec lesquelles elle est corrélée, et donc conduit à réduire d'autant plus la variance de la réponse du modèle.

Les indices de Sobol ont toutefois un grand coût algorithmique car le nombre d'indices de sensibilité de l'ordre 1 à l'ordre n (n nombre de variables) est égale à $2^n - 1$: en grande dimension il risque d'exploser.

Pour contourner ce problème, Homma et Saltelli ont alors introduit des indices de sensibilité totaux, exprimant la sensibilité de la variance Y à une variable sous toutes ses formes (sensibilité à la variable seule et sensibilité aux interactions de cette variable avec d'autres variables). Cela permettra d'exclure le calcul des indices des interactions lorsque l'indice total est proche de celui du premier ordre, car l'effet d'une variable seule sera très proche à la somme de toutes ses interactions et de cet effet.

Définition 15 (L'indice de sensibilité total S_{T_i}). *L'indice de sensibilité total S_{T_i} à la variable X_i est défini comme la somme de tous les indices de sensibilité relatifs à la variable X_i :*

$$S_{T_i} = \sum_{k \in \#i} S_k$$

où $\#i$ représente tous les ensembles d'indices contenant l'indice i .

Par exemple, pour un modèle à trois variables d'entrée, nous avons : $S_{T_1} = S_1 + S_{12} + S_{13} + S_{123}$.

Ces indices de sensibilité ont l'avantage de ne faire aucune hypothèse sur la forme du modèle, mais nécessitent une hypothèse d'indépendance des variables d'entrée.

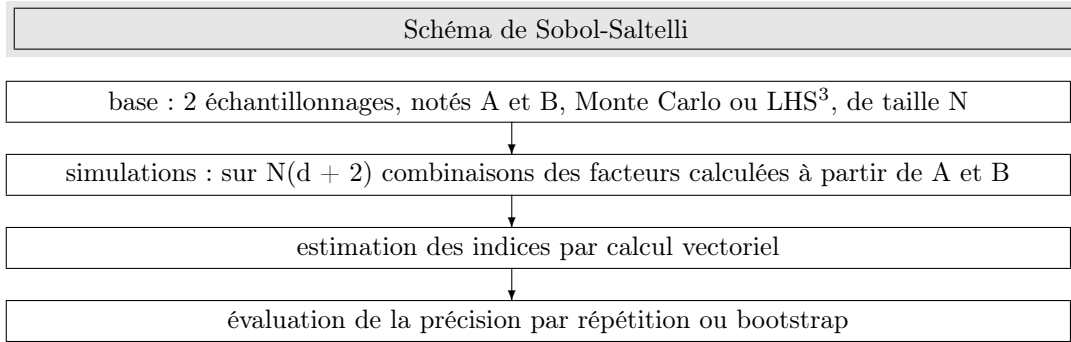
6.2.4 Estimation des indices de Sobol

En pratique il est souvent difficile d'obtenir analytiquement la valeur de ces indices.

Il existe plusieurs estimateurs des indices de Sobol.

Nous allons présenter l'estimateur de type Monte - Carlo (aussi appelé "pick-freeze"), qui repose sur la méthode de Sobol- Saltelli (2002), une amélioration de la précédente version de Sobol (1993), pour calculer les indices donnés par la décomposition de la variance jusqu'à un ordre spécifié.

Methode de Sobol - Saltelli :²



Description de l'algorithme

L'estimation des indices de sensibilité nécessite l'estimation de la variance de l'espérance conditionnelle. Supposons que l'on dispose d'un échantillon de taille N des d paramètres du modèle étudié, pris de façon aléatoire. On note X_i^k la k -ième valeur du paramètre X_i issue de l'échantillon.

Préliminairement, nous estimons l'espérance et la variance.

- L'estimée de l'espérance de y est donnée par :

$$\hat{\mathbb{E}}(y) = \frac{1}{N} \sum_{k=1}^N f(X_1^k, \dots, X_d^k)$$

- De même, l'estimée de la variance de y est donnée par :

$$\hat{\mathbb{V}}(y) = \frac{1}{N} \sum_{k=1}^N f^2(X_1^k, \dots, X_d^k) - \hat{\mathbb{E}}(y)^2$$

L'estimation des indices de sensibilité de premier ordre consiste à estimer la quantité :

$$V_i = \text{Var}(\mathbb{E}[y|X_i]) = \underbrace{\mathbb{E}[\mathbb{E}[y|X_i]^2]}_{:=U_i} - \mathbb{E}[\mathbb{E}[y|X_i]]^2 = U_i - (\mathbb{E}(y))^2$$

Sobol propose d'estimer la quantité U_i , c'est-à-dire l'espérance du carré de l'espérance de y conditionnellement à X_i , comme une espérance classique, mais en tenant compte du conditionnement par rapport à X_i en faisant varier entre les deux appels à la fonction f toutes les variables sauf la variable X_i .

Pour l'estimer, on a besoin de deux matrices des réalisations des paramètres, A et B , de dimension $N * d$ afin de faire varier tous les paramètres sauf X_i .

Il faut ensuite générer des matrices intermédiaires $C_i, \forall i = 1, \dots, d$, construites à l'aide de la colonne i de B et des colonnes $1, \dots, i-1, i+1, \dots, d$ de la matrice A .

²I. Sobol', Sensitivity estimates for nonlinear mathematical models, *Matematicheskoe Modelirovanie*, Vol. 2, p. 112-118, 1990 in Russian, translated in English in I. Sobol', Sensitivity analysis for non-linear mathematical models, *Mathematical Modeling Computational Experiment*, Vol. 1, p. 407-414, 1993.

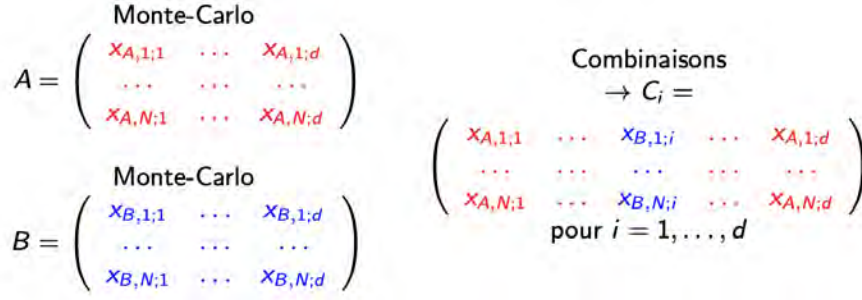


FIGURE 6.3 – 2 échantillonnages Monte Carlo A et B, de taille N

On calcule :

$$\hat{U}_i = \frac{1}{N} \sum_{k=1}^N f(X_{B;k;1}, \dots, X_{B;k;i-1}, X_{B;k;i}, X_{B;k;i+2}, \dots, X_{B;k;d}) * f(X_{A;k;1}, \dots, X_{A;k;i-1}, X_{B;k;i}, X_{A;k;i+1}, \dots, X_{A;k;d})$$

L'estimée de la variance de y conditionnellement à X_i , notée \hat{V}_i , est donnée par :

$$\hat{V}_i = \hat{U}_i - (\hat{\mathbb{E}}(y))^2$$

- L'estimateur de l'indice de sensibilité de premier ordre \hat{S}_i est ainsi :

$$\hat{S}_i = \frac{\hat{V}_i}{\hat{\mathbb{V}}(y)}$$

\hat{S}_i converge p.s. vers S_i quand $N \rightarrow \infty$

- L'estimateur de l'indice de sensibilité d'ordre 2 \hat{S}_{ij} est :

$$\hat{S}_{ij} = \frac{\hat{V}_{ij}}{\hat{\mathbb{V}}(y)}$$

avec

- \hat{V}_{ij} donné par :

$$\hat{V}_{ij} = \hat{U}_{ij} - (\hat{\mathbb{E}}(y))^2 - \hat{V}_i - \hat{V}_j$$

- et en estimant $U_{ij} = \mathbb{E}[y|X_i, X_j]^2$ de la même manière que U_i , en faisant varier entre les deux appels à la fonction toutes les variables sauf X_i et X_j :

$$\hat{U}_{ij} = \frac{1}{N} \sum_{k=1}^N f(X_{B;k;1}, \dots, X_{B;k;i-1}, X_{B;k;i}, X_{B;k;i+1}, \dots, X_{B;k;d}) * f(X_{A;k;1}, \dots, X_{A;k;i-1}, X_{B;k;i}, X_{A;k;i+1}, \dots, X_{A;k;j-1}, X_{B;k;j}, X_{A;k;j+1}, \dots, X_{A;k;d})$$

L'estimation des indices de sensibilité d'ordre i , avec $1 < i \leq d$, nécessite l'estimation des indices de sensibilité d'ordre 1 jusqu'à $i-1$. Par contre, les indices de sensibilité totaux peuvent être estimés directement. En effet, l'estimateur de l'indice de sensibilité d'ordre total nécessite l'estimation de la variance de l'espérance de Y conditionnellement à toutes les variables sauf X_i : $V_{\sim i} := \mathbb{E}[\mathbb{E}[Y|X_{\sim i}]^2] - \mathbb{E}[\mathbb{E}[Y|X_{\sim i}]] = U_{\sim i} - \mathbb{E}(y)^2$. On estime $U_{\sim i}$ par :

$$\hat{U}_{\sim i} = \frac{1}{N} \sum_{k=1}^N f(X_{B;k;1}, \dots, X_{B;k;i-1}, X_{B;k;i}, X_{B;k;i+2}, \dots, X_{B;k;d}) * f(X_{B;k;1}, \dots, X_{B;k;i-1}, X_{A;k;i}, X_{B;k;i+1}, \dots, X_{B;k;d})$$

- L'estimateur de l'indice de sensibilité d'ordre total est ainsi :

$$\hat{S}_{T_i} = 1 - \frac{\hat{V}_i}{\hat{\mathbb{V}}(y)}$$

avec $\hat{V}_{\sim i}$: estimation de la variance de y conditionnellement à tous les paramètres sauf X_i :

$$\hat{V}_{\sim i} = \hat{U}_{\sim i} - (\hat{\mathbb{E}}(y))^2$$

qui est estimée comme V_i , où varie uniquement X_i au lieu de faire varier toutes les variables sauf X_i .

Coût algorithmique et stratégie à adopter :

En utilisant une taille d'échantillon de Monte Carlo de N , le nombre de simulations des variables d'entrée nécessaire à l'estimation des indices de sensibilité est $2N$, puisque cette estimation nécessite deux jeux de simulations. Le nombre d'appels à la fonction du modèle est alors $N \times (k + 1)$, où k est le nombre d'indices estimés. Pour un modèle à d variables d'entrée, l'estimation de tous les indices de sensibilité nécessite $N \times (2^d)$ appels à la fonction.

Une stratégie à utiliser est d'estimer uniquement les indices de premier ordre et les indices totaux (cout de $N \times (2d+1)$ appels à la fonction) : s'il existe des écarts importants entre ces deux indices, la part des interactions est non négligeable et il peut être utile d'estimer les indices d'ordres intermédiaires. Dans le cas contraire, l'effet des variables d'entrée sera principalement de premier ordre et il ne sera pas utile de s'intéresser aux indices d'ordres intermédiaires.

En pratique, une taille d'échantillon de l'ordre de 10 000 sera suffisante pour estimer les indices de sensibilité d'un modèle comportant une dizaine de variables d'entrée.

Propriétés et problématiques :

La méthode de Sobol-Saltelli possède les caractéristiques suivantes :

- les hypothèses d'utilisation de la méthode sont peu contraignantes, il est seulement demandé que la fonction f soit de carré intégrable par rapport à la loi produit des entrées ($\mathbb{E}(y^2) < \infty$),
- elle permet d'estimer les indices d'ordre supérieur à 1 avec les effets dus aux interactions entre les paramètres d'entrée et les effets totaux,
- les estimateurs convergent et sont sans biais,
- la vitesse de convergence de la méthode est indépendante de la dimension d , toutefois elle est seulement d'ordre $\mathcal{O}(n^{-1/2})$ (vitesse du TLC) pour une taille N des plans d'expériences construits,
- le coût de calcul $N * (k + 1)$ avec k nombre des indices à estimer est élevé mais il peut être amélioré avec un échantillonnage pseudo probabiliste ou en passant par un métamodèle,
- la qualité de l'échantillon et la taille de l'échantillonnage peuvent causer une grande imprécision :
 - le signe des indices peut être négatif pour les facteurs les moins influents,
 - l'indice total peut être inférieur de celui d'ordre 1.

Pour pallier cette problématique, une technique de bootstrap est appliquée pour mesurer la robustesse.

Le choix de la méthode de l'analyse de sensibilité est généralement déterminé pour répondre aux contraintes que la fonction f ou la base de données impose. Les plus courantes sont :

- le coût de calcul : sauf dans de très rares cas, faire une analyse de sensibilité requiert un grand nombre d'évaluations. Cela devient une barrière quand une seule exécution du code prend beaucoup de temps et le code dispose d'un grand nombre d'entrées. L'espace à explorer devient trop grand ("Fléau de la dimension").
- Les entrées corrélées : l'hypothèse d'indépendance parmi les entrées du modèle simplifie beaucoup l'analyse de sensibilité, mais parfois on ne peut pas négliger que les entrées soient fortement corrélées. Les corrélations parmi les entrées doivent alors être prises en compte dans l'analyse.
- Les interactions : lorsque les entrées ont seulement des effets additifs, on dit qu'elle ne sont pas en interaction. Certaines techniques simples (nuage de point, OAT) sont alors applicables. Quand ce n'est pas le cas, d'autres techniques sont mieux indiquées (calcul des indices de Sobol simple et totaux).

Optimisation des méthodes décrites par le choix de l'échantillonnage :

Quand on souhaite réduire le temps d'exécution d'un algorithme d'approximation et l'utilisation de la mémoire vive des logiciels comme R , tout en gardant une bonne estimation des indices de Sobol, on recherche souvent à satisfaire les contraintes suivantes :

- en premier lieu, répartir les points dans l'espace de façon à capter les non-linéarités.
- En deuxième lieu, choisir des observations représentatives de l'espace.

- Ensuite, faire en sorte que ce remplissage de l'espace prenne en compte l'autocorrélation spatiale.

La qualité de la répartition spatiale est mesurée soit à l'aide de critères déterministes comme les distances minimax ou maximin (Johnson et al.,1990), soit à l'aide de critères probabilistes comme la discrédance.

Ainsi des méthodes de simulation quasi-aléatoires, comme l'échantillonnage stratifié ou par hypercube latin (LHS) peuvent être utilisés pour améliorer la vitesse de convergence et l'efficacité des estimateurs.

Monte Carlo vs Quasi Monte Carlo

En analyse numérique, la méthode de quasi-Monte-Carlo est une méthode d'intégration numérique et la résolution de problèmes numériques par l'utilisation de suites à discrédance faible. Elle s'oppose donc à la méthode de Monte-Carlo qui utilise des suites de nombres pseudo-aléatoires.

Les méthodes de Monte-Carlo et quasi-Monte-Carlo se basent sur le même problème : l'approximation de l'intégrale d'une fonction f par la moyenne des valeurs de la fonction évaluée en un ensemble de points x_1, \dots, x_N :

$$\int f(u)du \approx \frac{1}{N} \sum_{i=1}^N f(x_i)$$

avec x_i observation.

La différence entre les méthode de Monte-Carlo et quasi-Monte-Carlo tient dans le choix des valeurs x_i . Alors que la méthode de Monte-Carlo utilise une suite de nombres pseudo-aléatoires, la méthode de quasi-Monte-Carlo utilise la suite de Halton, la suite de Sobol ou la suite de Faure, connue pour leur discrédance faible. L'un des avantages est de permettre une convergence plus rapide de la méthode par l'utilisation de suites à discrédance faible (de l'ordre de $\mathcal{O}(\frac{1}{N})$), alors que la méthode de Monte-Carlo est en $\mathcal{O}(\frac{1}{\sqrt{N}})$.

Estimations d'erreurs d'approximation de la méthode de quasi-Monte-Carlo :

L'erreur d'approximation de la méthode de quasi-Monte-Carlo est bornée par un terme proportionnel à la discrédance de la suite des nœuds de calcul $x_1 \dots, x_N$. Plus précisément, l'inégalité de Koksma-Hlawka borne l'erreur par

$$\left| \int_X f(u) du - \frac{1}{N} \sum_{i=1}^N f(x_i) \right| \leq V(f)D_N \left| \int_X f(u) du - \frac{1}{N} \sum_{i=1}^N f(x_i) \right| \leq V(f)D_N$$

où $V(f)$ est la variation de Hardy-Krause de la fonction f^2 et D_N est la discrédance de l'ensemble (x_1, \dots, x_N) :

$$D_N = \sup_{Q \subset [0,1]^s} \left| \frac{\#Q}{N} - \text{volume}(Q) \right|, D_N = \sup_{Q \subset [0,1]^s} \left| \frac{\#Q}{N} - \text{volume}(Q) \right|,$$

où Q est un pavé rectangulaire inclus dans X dont chaque face est parallèle aux faces du cube unité et $\#Q$ représente le nombre de nœuds contenus dans Q .

Cette inégalité permet de montrer que l'erreur d'approximation par une méthode de quasi-Monte-Carlo est $\mathcal{O}\left(\frac{(\ln N)^s}{N}\right)$, alors qu'au mieux, la méthode de Monte-Carlo a une erreur en probabilités de $\mathcal{O}\left(\frac{1}{\sqrt{N}}\right)$.

Dans la pratique, pour couvrir de manière optimale l'espace variable d'entrée (à grande dimension par exemple) les séquences quasi-Monte Carlo sont préférées.

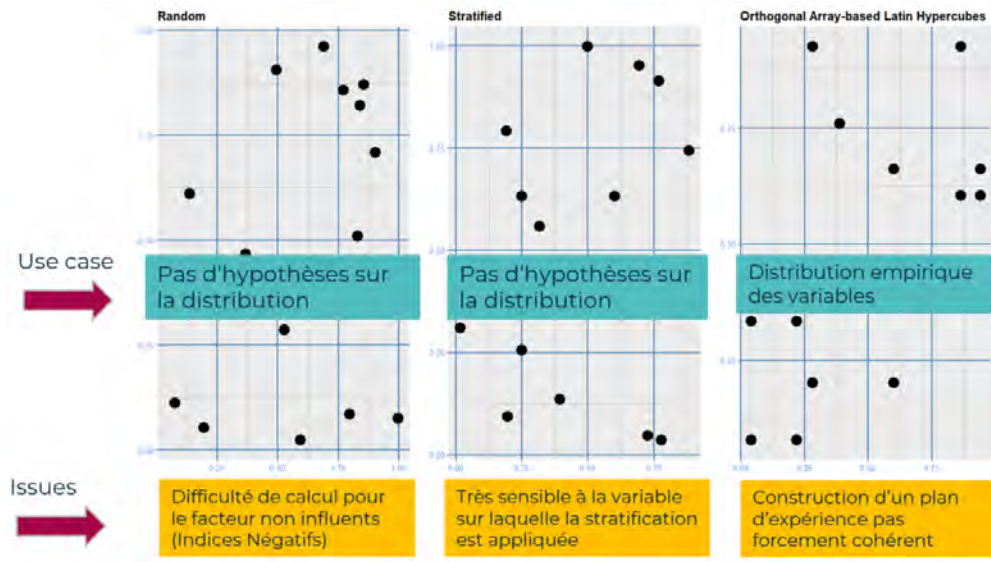


FIGURE 6.4 – Nous avons testé les échantillons suivants : aléatoire, stratifié par département ou commune (nous avons considéré une proportion fixée par catégorie) et par hypercube latin en utilisant la fonction de répartition empirique des variables. Malgré sa limite pour les facteurs non influents, l'échantillon aléatoire a donné des résultats plus robustes. Nous avons pallié sa limite en augmentant la taille de l'échantillon et parallélisé l'algorithme sous le logiciel R.

6.2.5 Illustration d'un modèle à entrées indépendantes : le modèle Ishigami

Dans cette section nous présentons un exemple classique de l'analyse de sensibilité globale, le modèle Ishigami, à 3 variables d'entrée indépendantes :

$$Y = f(X_1, X_2, X_3) = \sin(X_1) + a \sin(X_2)^2 + b X_3^4 \sin(X_1)$$

où $X_i \sim U[-\pi; \pi]$ pour $i = 1, 2, 3$ (les densités sont $f_i(X_i) = \frac{1}{2\pi}$) et a, b réels. Pour cette fonction, tous les indices de sensibilité Sobol ($S_1, S_2, S_3, S_{12}, S_{13}, S_{23}, S_{123}, S_{T_1}, S_{T_2}, S_{T_3}$) sont connus.

En effet, pour cette fonction analytique, il est facile d'obtenir l'espérance, la variance et les indices de sensibilité⁴.

Cet exemple est particulièrement intéressant car il permet de montrer deux résultats :

1. En premier lieu, la convergence des estimateurs des indices de Sobol et le temps d'exécution en fonction de la taille de l'échantillon, ce qui nous donne une idée de quel estimateur utiliser par la suite.
2. En deuxième lieu, en passant par un métamodèle de type *glm* nous voulons montrer que les modèles linéaires, les plus utilisés en assurance Non-Vie, ne prennent pas en compte l'interaction parmi les variables dans la modélisation.

On obtient, avec les notations comme dans 6.2.3 :

$$\begin{aligned} \mathbb{E}(Y) &= \frac{a}{2} \\ \mathbb{V}ar(Y) &= \frac{1}{2} + \frac{a^2}{8} + \frac{b^2 \pi^8}{18} + \frac{b \pi^4}{5} \\ f_0 &= \frac{a}{2} \\ f_1(X_1) &= \sin(X_1) \left(1 + b \frac{\pi^4}{5}\right) \\ f_2(X_2) &= -\frac{a}{2} + a (\sin(X_2))^2 \\ f_{1,3}(X_1, X_3) &= b \sin(X_1) \left(X_3^4 - \frac{X_3^4}{5}\right) \end{aligned}$$

⁴Pour la preuve voir : Baudin M., Martinez J.M. "Introduction to sensitivity analysis with NISP", 2013

$$f_3(X_3) = f_{1,2}(X_1, X_2) = f_{2,3}(X_2, X_3) = f_{1,2,3}(X_1, X_2, X_3) = 0$$

$$V_1 = \frac{1}{2} \left(1 + b \frac{\pi^4}{5}\right)^2$$

$$V_2 = \frac{a^2}{8}$$

$$V_{1,3} = b^2 \pi^8 \frac{8}{225}$$

$$V_3 = V_{1,2} = V_{2,3} = V_{1,2,3} = 0$$

Remarque : On déduit que $S_3 = 0$: cela signifie que X_3 n'a pas d'effet sur la sortie si prise toute seule, mais elle a un effet sur la sortie à cause de son interaction avec X_1 .

6.2.5.1 Approximations des indices de Sobol

Simulation de deux échantillons issus de la fonction Ishigami

Nous avons simulé les 3 entrées indépendantes (X_1, X_2, X_3) et calculé la fonction Ishigami avec $a = 7$ et $b = 0.1$. Deux échantillons de taille 1000 de (X_1, X_2, X_3) ont été simulés.

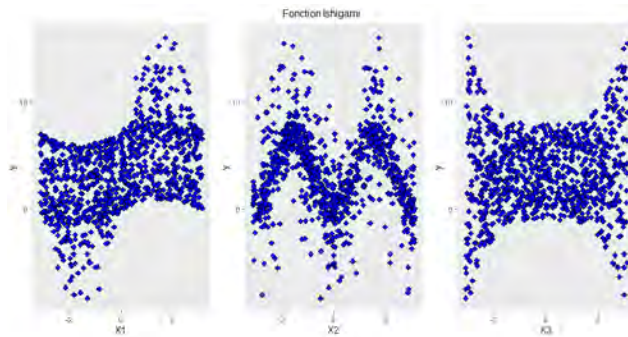


FIGURE 6.5 – Fonction Ishigami par rapport aux 3 entrées ($X_i \sim U[-\pi; \pi]$ pour $i = 1, 2, 3$)

Comparaison des trois méthodes Monte Carlo

L'algorithme de Liu-Owen, plus récent, (Liu et Owen, 2006) a été comparé à celui de Sobol et Sobol-Saltelli dans cet exemple, mais nous ne l'avons pas retenu dans l'application numérique de la tarification à l'adresse.

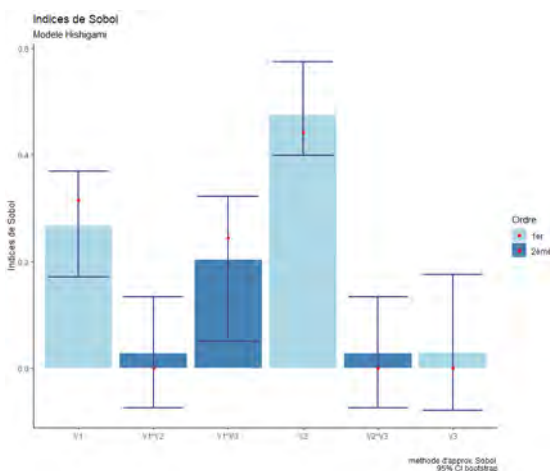


FIGURE 6.6 – Estimation des indices de Sobol avec la méthode de Sobol de type Monte-Carlo; en rouge la valeur exacte.

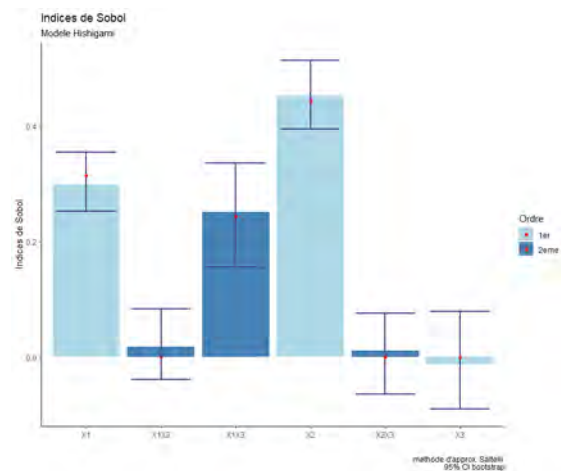


FIGURE 6.7 – Estimation des Indice de Sobol avec la méthode de Sobol-Saltelli de type Monte-Carlo; en rouge la valeur exacte.

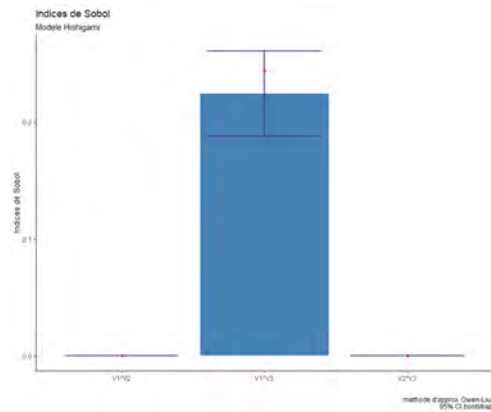


FIGURE 6.8 – Estimation des Indice de Sobol de l’ordre 2 avec la méthode de Owen-Liu (Liu et Owen, 2006) de type Monte-Carlo ; en rouge la valeur exacte.

TABLE 6.2 – Résumé des estimations des indice de Sobol (échantillon de taille 1000)

| <i>Methodes d’approximation des Indices de Sobol</i> | | | | | | | | | | |
|--|--------|--------|--------|-------------------|----------|-------|-------------------|----------|-------|-------------------|
| Indices | Exacte | Sobol | | | Saltelli | | | Owen-Liu | | |
| | | Approx | sd | IC _{95%} | Approx | sd | IC _{95%} | Approx | sd | IC _{95%} |
| S ₁ | 0,314 | 0,267 | 0,049 | [0,171 ; 0,368] | 0,3 | 0,024 | [0,253 ; 0,355] | | | |
| S ₂ | 0,442 | 0,474 | 0,041 | [0,399 ; 0,575] | 0,452 | 0,028 | [0,394 ; 0,514] | | | |
| S ₃ | 0 | 0,029 | 0,06 | [-0,079 ; 0,176] | -0,012 | 0,038 | [-0,089 ; 0,079] | | | |
| S ₁₂ | 0 | 0,028 | 0,055 | [-0,075 ; 0,134] | 0,019 | 0,031 | [-0,038 ; 0,084] | 0 | 0 | [0 ; 0] |
| S ₁₃ | 0,244 | 0,203 | 0,07 | [0,05 ; 0,322] | 0,251 | 0,043 | [0,156 ; 0,336] | 0,224 | 0,188 | [0,261 ; 0,244] |
| S ₂₃ | 0 | 0,028 | 0,055 | [-0,075 ; 0,134] | 0,012 | 0,033 | [-0,064 ; 0,076] | 0 | 0 | [0 ; 0] |
| S _{T₁} | 0,557 | 0,469 | 0,0125 | [0,520 ; 0,570] | | | | | | |
| S _{T₂} | 0,442 | 0,502 | 0,008 | [0,437 ; 0,467] | | | | | | |
| S _{T₃} | 0,244 | 0,231 | 0,004 | [0,230 ; 0,24] | | | | | | |

Lecture des résultats

Dans le modèle d'Ishigami :

- la variable qui a le plus d'influence sur la variance de la sortie (au sens de l'indice total) est la variable X_1 , avec un indice total de $0.31+0.24=0.55$ et près de 31% de la variance de Y est expliquée par cette variable toute seule.
- La variable X_2 n'intervient que seule (indice d'ordre un équivalent à indice total), en expliquant près de 44% de la variance de Y .
- La variable X_3 a une influence uniquement avec X_1 , avec un indice total d'environ 0.24.
- La variance de Y est donc due pour 44% à X_2 , 31% à X_1 et 25% à l'interaction entre X_3 et X_1 .

Dans cet exemple, l'interaction entre X_1 et X_3 est due à une relation non additive entre ces deux variables dans l'expression du modèle.

Pour observer l'effet qu'une variable sur la sortie, on fixe cette variable à son espérance :

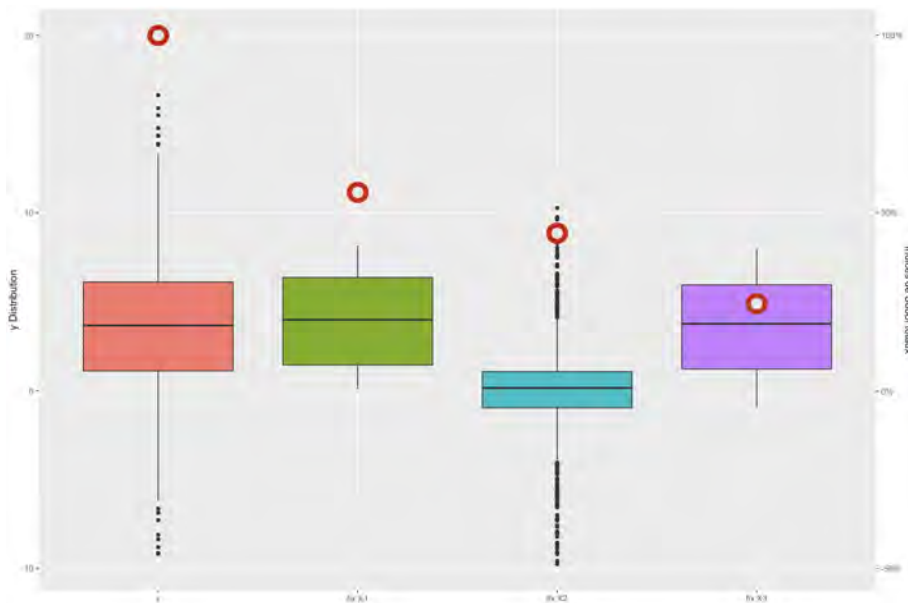


FIGURE 6.9 – Distributions de Y en fonction de la variable d'entrée fixée. La première boîte correspond à la distribution initiale de y . Les boîtes suivantes correspondent aux distributions de y en fixant les autres variables une à une, l'indice total est marqué en rouge.

La plus grande réduction de variance est obtenue en fixant la variable ayant l'indice de sensibilité total le plus important.

Visualisation des interactions à l'aide des graphes Network

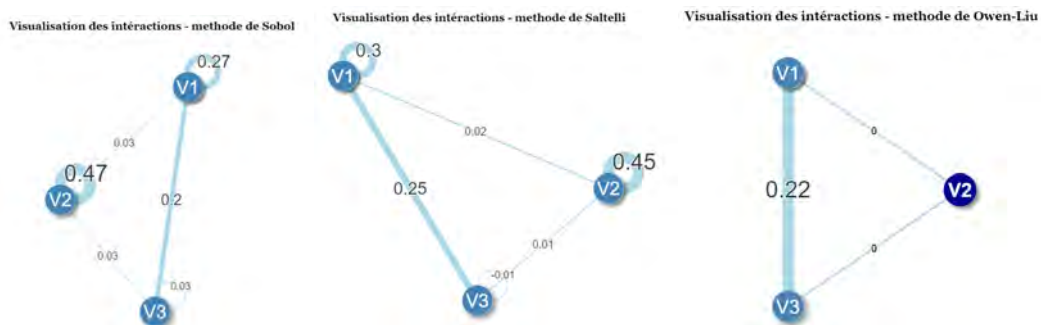


FIGURE 6.10 – Visualisations des indices de Sobol à l'ordre 1 et 2 (taille de l'échantillon : 1000). L'épaisseur des flèches est proportionnelles à la valeur de l'interaction (X_1).

Convergence à la valeur exacte

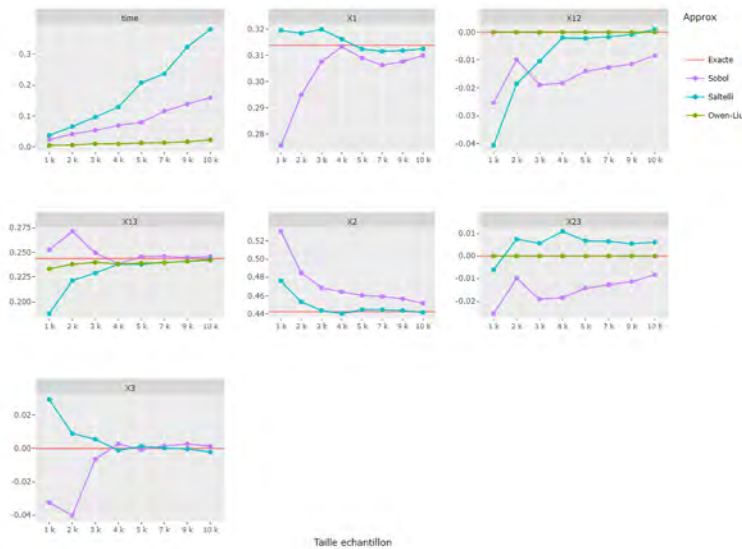


FIGURE 6.11 – Temps d’exécution des algorithmes d’approximation (secs) et convergence vers la valeur des indices de Sobol exacte en fonction du nombre de simulation Monte Carlo.

6.2.5.2 Approximation en passant par un métamodèle

En utilisant les échantillons simulés, nous voulons montrer qu’une modélisation de type *glm* simple ne prend pas en compte l’interaction existante entre les variables X_1 et X_3 alors que un modèle de type arbre de régression capte cette interaction. Enfin, on montre que l’ajout de l’interaction à la formule du GLM est significatif.

- Tout d’abord, nous montrons graphiquement que le modèle *glm* ne prend pas en compte l’interaction entre X_1 et X_3 :

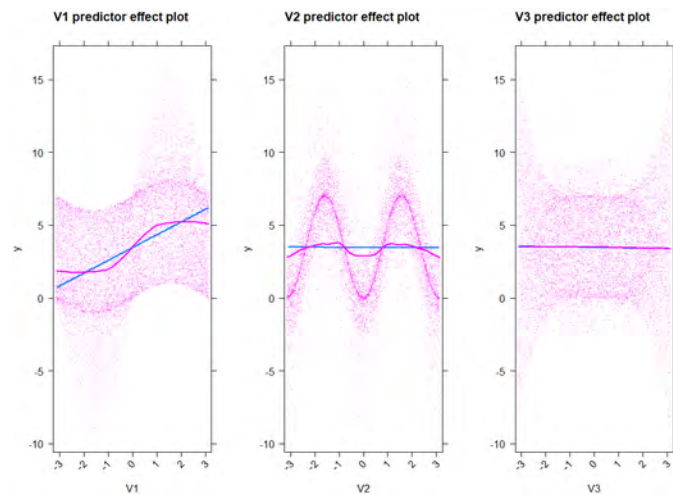


FIGURE 6.12 – En bleu, le modèle incorrect $y \sim X_1 + X_2 + X_3$ utilisé pour approcher la cible $E(Y|X) = \sin(X_1) + a * \sin(X_2)^2 + b * x_3^4 * \sin(X_1)$ (points en magenta). Nous remarquons que la structure linéaire pour X_2 n’est pas du tout adaptée et les résidus partiels par rapport à X_1 montrent qu’il y a un effet de X_1 qui n’est pas pris en compte.

- Nous avons construit un modèle CART en utilisant les données simulées précédemment et nous voudrions montrer que ce modèle prend en compte la non-linéarité parmi X_1 et X_3 . D’abord nous calibrons à l’aide de la validation croisée le paramètre de complexité :

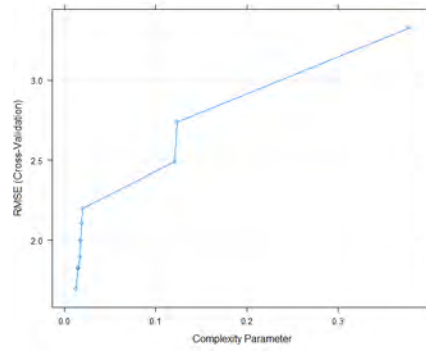


FIGURE 6.13 – Le modèle final sera celui avec le paramètre de complexité le plus petit : 0.01259719.

Au vue du coefficient de prédiction Q_2 qui est de 78%, nous pouvons valider le modèle.

- Ensuite nous estimons les indices de Sobol, comme vu dans les sections précédentes :

TABLE 6.3 – Indice de Sobol en utilisant un métamodèle de type CART

| Saltelli - Sobol | | | | | |
|------------------|--------|-------------|---------------|------------|----------------------------|
| | Exacte | Approx | Biais | sd | IC _{95%} |
| S_1 | 0.314 | 0.37069417 | 0.0015785102 | 0.04582366 | [0.27845707 ,0.45782311] |
| S_2 | 0.442 | 0.52228801 | 0.0015650898 | 0.04183838 | [0.44049503, 0.60338783] |
| S_3 | 0 | 0.06082380 | -0.0003343051 | 0.05831787 | [-0.04847389 ,0.17487190] |
| S_{12} | 0 | -0.08335067 | -0.0007779491 | 0.05427911 | [-0.19426770, 0.02245197] |
| S_{13} | 0.244 | 0.11046562 | -0.0024673732 | 0.06716990 | [-0.02228949 ,0.23520645] |
| S_{23} | 0 | -0.07161736 | -0.0005596565 | 0.05549022 | [-0.18159621 ,0.03381170] |
| S_{123} | 0 | 0.09173828 | 0.0010005124 | 0.05375803 | [-0.01148203, 0.20088404] |
| S_{T_1} | 0.557 | 0.4895474 | | | |
| S_{T_2} | 0.442 | 0.45905827 | | | |
| S_{T_3} | 0.243 | 0.19141036 | | | |

- Les indices de Sobol estimés à partir d'un métamodèle CART identifient une interaction entre les variables X_1 et X_3 , mais ils ne spécifient pas le type de fonction qui les lie.

Pour cela, on ajoute des termes quadratiques et cubiques de X_1 et X_3 à la structure linéaire du GLM :

```

              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.0223688    0.0178901  -1.250    0.2112
X2            6.9871399    0.0147314 473.660 < 2e-16 ***
SO(X1, X3)X1  0.1425924    0.0246852   5.776 7.86e-09 ***
SO(X1, X3)X3  0.0022736    0.0082248   0.276   0.7825
SO(X1, X3)X1:X3 -0.0121228    0.0137704  -0.769   0.4421
SO(X1, X3)X1^2  0.1345471    0.0626409   2.148   0.0317 *
SO(X1, X3)X3^2  0.0124403    0.0066818   1.862   0.0627 .
PQ(X1, X3)X1^2 NA NA NA NA
PQ(X1, X3)X3^2 NA NA NA NA
SO(X1, X3)X1:PQ(X1, X3)X1^2  0.0226430    0.0294473   0.769   0.4420
SO(X1, X3)X3:PQ(X1, X3)X1^2 -0.0090317    0.0081474  -1.109   0.2677
SO(X1, X3)X1:X3:PQ(X1, X3)X1^2  0.0456694    0.0162296   2.814   0.0049 **
SO(X1, X3)X1^2:PQ(X1, X3)X1^2 -0.1059837    0.0388260  -1.802   0.0716 .
SO(X1, X3)X3^2:PQ(X1, X3)X1^2 -0.0110202    0.0030701  -3.205   0.0025 **
SO(X1, X3)X1:PQ(X1, X3)X3^2  0.8442908    0.0025471 321.473 < 2e-16 ***
SO(X1, X3)X3:PQ(X1, X3)X3^2  0.0001956    0.0011146   0.176   0.8607
SO(X1, X3)X1:X3:PQ(X1, X3)X3^2 -0.0026088    0.0015813  -1.650   0.0990 .
SO(X1, X3)X1^2:PQ(X1, X3)X3^2 NA NA NA NA
SO(X1, X3)X3^2:PQ(X1, X3)X3^2 -0.0009698    0.0007053  -1.375   0.1693
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 0.2711473)

Null deviance: 136840.0 on 9999 degrees of freedom
Residual deviance: 2707.1 on 9984 degrees of freedom
AIC: 15346

```

FIGURE 6.14 – Les variables plus significatives sont X_2 , X_1 et $X * X_3^2$

On ajoute ainsi le terme " $X_1 * X_3^2$ " au modèle *glm* et on regarde sa significativité à l'aide d'un test statistique.

- Dans le cas d'un modèle linéaire :

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 * X_3,$$

on testerait l'hypothèse nulle " $H_0 : \beta_3 = 0$ ";

- Dans le cadre d'un *glm*, le test statistique multiple prend la forme d'un test du rapport de vraisemblance.

Comme la vraisemblance d'un modèle augmente presque toujours avec l'ajout de termes supplémentaires, le test statistique détermine si la différence de vraisemblance entre le modèle avec les termes croisés (H_1 : modèle non contraint) et celui sans ces termes (H_0 : modèle contraint) est statistiquement significative. S'il n'y a pas de différence significative, alors l'ajout des termes croisés n'améliore pas suffisamment la qualité du modèle et l'interaction envisagée est écartée. La statistique de test s'écrit :

$$S = -2 \ln \left(\frac{L_C}{L_{nc}} \right)$$

où L_C et L_{nc} sont les vraisemblances respectives des modèles contraint et non contraint. Cette statistique suit une distribution du χ^2 avec pour degrés de liberté le nombre de paramètres contraints.

- On considère deux modèles *glm* :

– simple

$$\mathbb{E}(Y) = g^{-1} \left(\sum_{i=0}^3 \beta_i X_i \right)$$

avec fonction de lien g

– avec interaction " $X_1 * X_3$ "

$$\mathbb{E}(Y) = g^{-1} \left(\sum_{i=0}^3 \beta_i X_i + X_1 * X_3 \right)$$

avec fonction de lien g

Le rapport de vraisemblance est donné par :

$$LR = Deviance_{\text{glm simple}} - Deviance_{\text{glm avec interaction}} = 29809.76$$

La différence est très importante et de plus, la probabilité critique (p-value)

(`pchisq(LR,df=1,lower.tail = FALSE)`) = 0 \ll 5% donc au risque de 5% le terme croisé contribue significativement dans le modèle.

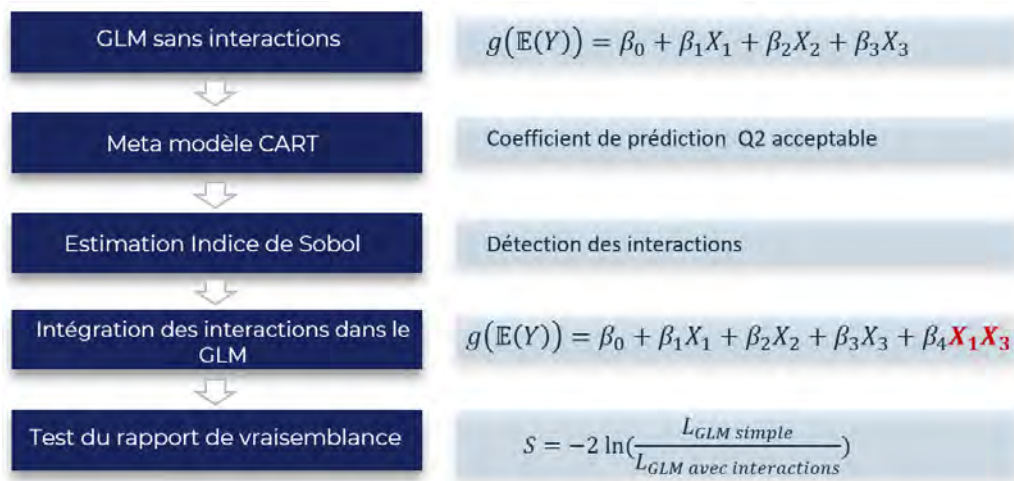


FIGURE 6.15 – Processus de détection des interactions.

6.2.6 Limites du glm vis-à-vis des interactions et pourquoi envisager une analyse de sensibilité

À l'aide de l'exemple du modèle Ishigami, nous avons montré les limites du *glm*, qui approche par une structure linéaire un phénomène plus complexe. L'analyse de sensibilité dans son but de quantification d'importance permet de pallier cette limite et de détecter les interactions parmi deux (ou plusieurs variables). Malgré le coût important que demande l'algorithme de Sobol-Saltelli, l'utilisateur peut bénéficier d'une méthode plus robuste et plus fiable que l'ajout manuel dans l'équation tarifaire du *glm*.

6.3 Modèles à entrées dépendantes

Dans le but de quantifier les contributions des entrées d'un modèle, l'analyse de sensibilité globale est limitée au cas où il n'y a pas de corrélations entre les variables explicatives.

En effet, dans le cas de colinéarité, un indice d'ordre un n'exprime plus la sensibilité à une unique variable si cette dernière est corrélée avec d'autres.

Indice multidimensionnel

Une solution pour intégrer l'analyse de sensibilité au cas de la colinéarité est de considérer des indices par groupe de variables dépendantes (les variables au sein d'un groupe sont dépendantes mais les variables de différents groupes sont indépendantes).

Ainsi, la sensibilité de la variance de Y à un groupe de facteurs dépendants est donnée par des indices de sensibilité multidimensionnels.

Définition 16. *Considérant un groupe de k variables $\{X_1, \dots, X_k\}$, dépendantes parmi elles, et indépendantes du reste des autres variables, la sensibilité à la variable multidimensionnelle $\{X_1, \dots, X_k\}$ est exprimée par l'indice multidimensionnel*

$$S_{\{1, \dots, k\}} = \frac{\text{Var}(\mathbb{E}[Y|X_1, \dots, X_k])}{\text{Var}(Y)}$$

Shapley Effects

Récemment introduits en analyse de sensibilité par la contribution de Owen (2014), les indices de Shapley (Shapley Effects) allouent une part de la variance de la sortie à chaque entrée.

Définition 17. *Soit d le nombre d'entrées du modèles. Les Shapley Effects sont des indices définis $\forall i = 1, \dots, d$:*

$$Sh_i = \sum_{u \subset \{1, \dots, d\} \setminus \{i\}} \frac{(d - |u| - 1)! |u|!}{d!} [c(u \cup \{i\}) - c(u)]$$

où $c(\cdot)$ est une fonction coût, souvent $c(u) = \text{Var}(\mathbb{E}[Y|X_u]) / \text{Var}(Y)$.

Remarque Les *Shapley Effects* avec $c(u)$ contiennent les effets des variables et des interactions. Ils sont les équivalents des indices totaux où l'on répartit les interactions équitablement. La somme des ces indices est 1, alors que la somme des indices de Sobol totaux n'est pas forcément 1.

Conclusion

Nous avons introduit les notions principales de l'analyse de sensibilité et les indices de sensibilité de Sobol.

Ensuite, nous avons montré la valeur ajoutée de ces outils dans le cadre d'une modélisation de type *glm*, avec l'exemple du modèle Ishigami. Les indices de Sobol quantifient l'effet pur (au premier ordre) de la variable, et les effets d'interaction (à partir de l'ordre 2).

Leur application dans les cas réels, dont la modélisation de la fréquence et de la sévérité des sinistres permet d'améliorer les modèles "traditionnels" de type *glm*, qui ne font pas intervenir aucune interaction parmi les variables.

Plus adaptés aux entrées dépendantes, les indices de Shapley (*Shapley Effects*) distribuent équitablement la variance de la sortie sur les entrées, afin de disposer d'indices totaux dont la somme est 1. Ils sont intéressants pour prioriser les entrées, par exemple pour la sélection des variables, mais ils n'ont pas une formule analytique pour isoler l'effet d'interaction comme les indices de Sobol du deuxième degré.

Dans le chapitre suivant, nous présenterons une extension des indices de Shapley pour la détection des interactions.

Chapitre 7

Explainable artificial intelligence

Les mesures d'importance des variables et des interactions présentées auparavant, les indices de Sobol, sont issus de l'analyse de sensibilité. Une autre approche de décomposition de l'effet d'une variable sur la sortie d'un modèle est la valeur SHAP, issue du domaine de l'*Explainable Artificial Intelligence*(XAI). Dans ce chapitre, nous présenterons ses mesures d'importances et d'interaction.

À la différence du chapitre précédent, la quantité décomposée sera la "valeur prédite".

Introduction

La tâche la plus courante de l'apprentissage statistique consiste à déterminer un modèle capable de prédire un résultat inconnu (variable de réponse) sur la base d'un ensemble de variables d'entrée connues. Dans les applications réelles, il est souvent crucial de comprendre pourquoi un certain ensemble de variables ou caractéristiques conduit exactement à cette prédiction.

Cependant, expliquer les prédictions à partir des modèles de machine learning complexes est une question pratique éthique et juridique : Puis-je faire confiance au modèle ? Est-ce biaisé ? Puis-je l'expliquer aux autres ? Dans ce chapitre nous expliquons les prédictions individuelles à partir d'un modèle d'apprentissage automatique complexe en apprenant des explications simples et interprétables.

En effet, deux critiques majeures sont soulevées par rapport à l'interprétabilité des modèles de tarification de la sinistralité individuelle :

- Dans le cas où on modélise par un modèle linéaire ou linéaire généralisé, l'interprétation est faite grâce aux coefficients de régression, mais cela est limité à un certain nombre de liens parmi les variables et n'explore pas toute les possibilités : si on voulait par exemple ajouter une interaction non linéaire parmi deux ou plusieurs variables à un modèle linéaire généralisé, cela demanderait un temps machine très élevé.
- Les modèles en Machine learning, malgré leur haute performance, sont limités par l'effet boîte noire¹.

Un autre façon de voir les choses est de s'intéresser au rôle ou au *poids* des variables dans la prédiction. Cela permet par ailleurs d'interpréter le modèle. Le premier outil dont on s'est intéressé est **l'importance des variables**, communément utilisé pour la sélection des variables et plus récemment (Owen, 2014) l'explication des modèles.

La notion de l'importance des variables se base sur la valeur de Shapley, un concept des jeux coopératifs, qui définit un partage d'une quantité (la valeur prédite de la fréquence ou de la sévérité des sinistres par exemple). Ce partage respecte l'apport de chaque variable dans une coalition (un sous-ensemble des variables) pour n'importe quel modèle prédictif.

7.1 Classe des méthodes *Additive Feature Attribution*

Définition 18. *Explainable Artificial Intelligence (XAI) fait référence au domaine où on étudie des méthodes et techniques d'intelligence artificielle de sorte que leurs prédictions puissent être comprises par les humains.*

Cela contraste avec le concept de *boîte noire* dans l'apprentissage automatique où même les développeurs du modèle ne peuvent pas expliquer pourquoi l'Intelligence artificielle (IA) est arrivée à une décision spécifique.

¹Un phénomène est appelé *boîte noire* » (ou *black box* en anglais), dans le sens où on peut juger des données qui entrent dans la boîte et des résultats qui en sortent, mais sans savoir ce qui se passe à l'intérieur

XAI peut être une mise en œuvre du droit social à l'explication².

La définition d'une classe de modèle d'explication vient de la nécessité de généraliser la notion d'explication à tous les modèles de prédiction : si pour des modèles intuitifs, comme ceux linéaires ou linéaires généralisés, la meilleure explication du modèle est le modèle lui-même (les coefficients d'une régression par exemple sont auto-explicatifs), lorsque les modèles sont de type *black box* (réseaux de neurones, forêt aléatoire, ...), leur complexité ne permet pas de les comprendre.

Afin d'unifier tous les modèles de prédiction dans l'explication, on souhaiterait associer une **structure additive** à n'importe quel modèle de prédiction, vérifiant de bonnes propriétés. Les termes de cette structure sont les *poids* ou importances de chaque variable et ils vérifient les **deux propriétés souhaitables** suivantes :

1. Importance additive (*Additive Importance*) : pour que l'on puisse exploiter dans la pratique, la somme des importances devrait conduire à une quantité qui a du sens (par exemple la prédiction $f(x)$).
2. Poids des variables (*Feature attribution*) : indique dans quelle mesure une quantité d'intérêt du modèle f dépend de chaque variable.

Définition 19 (Modèle d'explication locale pour la classe des méthodes *Additive feature attribution*). Soient

- f un modèle de prédiction,
- X_1, X_2, \dots, X_n ses variables explicatives,
- les "Coalitions", des sous-ensembles de l'espace engendré par les variables X_1, X_2, \dots, X_n ,
- $x = (x_1, \dots, x_n)$ une observation
- c' le vecteur de la coalition (dans la littérature est aussi noté *simplified input features*), un vecteur binaire dont les composantes $i = 1, \dots, n$ sont définies de la façon suivante :

$$c'_i = \begin{cases} 1, & \text{si la variable est présente dans la coalition} \\ 0, & \text{sinon} \end{cases}.$$

On notera x' le vecteur de coalition pour la grande coalition $\{X_1, \dots, X_n\}$ de l'observation x .

- et

$$\begin{aligned} h_x : \{0, 1\}^n &\longrightarrow X \\ c' &\longmapsto c \end{aligned}$$

la fonction qui associe à un vecteur de coalition c' , une entrée c de X , tel que :

$$c_i = \begin{cases} x_i & \text{si } X_i \text{ est dans la coalition} \\ \text{entrée manquantes (ou aléatoire)} & \text{si la variable } X_i \text{ n'est pas dans la coalition} \end{cases}$$

Un modèle d'explication g est la combinaison linéaire de variables binaires $c' \in \{0, 1\}^n$

$$g(c') = \varphi_0 + \sum_{j=1}^n \varphi_j c'_j = \varphi_0 + \sum_{j=1}^n \varphi_j \quad (7.1)$$

avec n le nombre des variables et φ_j l'effet de la variable X_j .

Les méthodes locales tel que LIME[14] assurent que

$$g(c') \sim f(h_x(c')) \text{ lorsque } c' \sim x' \text{ (hypothèse des méthodes locales).}$$

On peut remarquer que cette définition garantit la propriété d'additivité car la somme de tous les effets est une approximation de la sortie $f(x)$. De plus, elle attribue une quantité d'importance à chacune des variables.

Le modèle d'explication est ainsi défini comme un outil qui décompose la valeur prédite (locale) d'un modèle (pas forcément additif) à travers une structure additive, en fonction des contributions de chaque variable à la prédiction.

²Edwards, Lilian ; Veale, Michael (2017). "Slave to the Algorithm ? Why a 'Right to an Explanation' Is Probably Not the Remedy You Are Looking For". Duke Law and Technology Review.

Remarque : Si la valeur de Shapley seule garantit un partage équitable des poids des variables, l'hypothèse des méthodes locaux est fondamentale pour se reconduire la valeur prédite à une structure additive. La littérature appelle cette valeur de Shapley particulière *SHAP*.

Les indices de Shapley (*Shapley effects*), comme vous dans le chapitre précédent, dérivent des méthodes dits globaux qui décomposant la variance, au lieu de la valeur prédite.

Classification des mesures d'importance selon le but

Dans le cadre de l'explication des modèles³, il est possible de décomposer plusieurs quantités d'intérêt, selon le but de l'explication des modèles qu'on souhaite amener :

- Portée locale ou globale.
La portée locale est préférée pour analyser l'importance des variables le long d'une ou plusieurs dimensions. Alors que la portée globale est pertinente comme métrique récapitulative pour les décisions de haut niveau : la sélection de variables ou hiérarchisation des facteurs.
- Sensibilité ou pouvoir prédictif : les importances des variables devraient-elles être une mesure de la variation du modèle f ou de l'augmentation des performances prédictives? Selon la sensibilité, l'importance devrait se concentrer sur la façon dont le modèle f repose sur une variable. Alors que l'approche du pouvoir prédictif fixe les importances pour tenir compte de la contribution d'une variable à l'amélioration des performances prédictives (réduction de la fonction de perte).

Selon le type d'information donné et le type de décomposition, les mesures d'importance des variables définissent une matrice 2 x 2.

| | Sensitivity (model behaviour) | Predictive (model performance) |
|--------|--|--|
| Local | <p>SHAP zone</p> <p>Quantity of interest: $f(x), \text{ for } x \in X$</p> $f(x) = \Phi_0 + \sum \Phi_i(x)$ | <p>SHAPloss zone (SHAP form model loss)</p> <p>Quantity of interest: $-l(y, f(x)), \text{ for } (x, y) \in (X, Y)$</p> $-l(y, f(x)) = \Phi_0 + \sum \Phi_i(x)$ |
| Global | <p>SHAPLEY EFFECTS zone</p> <p>Quantity of interest: $\mathbb{V}(f(X)),$</p> $\mathbb{V}(f(X)) = \Phi_0 + \sum \Phi_i$ | <p>SAGE zone</p> <p>Quantity of interest: $\mathbb{E}(-l(Y, f(X)))$</p> $\mathbb{E}[-l(y, f(X))] = \Phi_0 + \sum \Phi_i$ |

FIGURE 7.1 – Matrice des importances des variables selon le but : la quantité ϕ_i indique l'importance de la variable par rapport à la quantité d'intérêt. Source : datajms

On montrera qu'il existe des quantités d'intérêt optimales pour chaque quadrant, dérivant de la Valeur de Shapley, une notion tirée de la théorie de jeux.

La **valeur de Shapley** est le seul cadre d'explication de prédiction avec une théorie solide qui garantit les deux propriétés souhaitables des modèles d'explication ainsi que d'autres propriétés qui seront présentée par la suite.

Dans un premier temps nous allons définir la valeur de Shapley dans la théorie de jeux et ensuite ses spécificités dans les modèles de prédiction.

³de manière équivalente aussi dans le cadre de l'analyse de sensibilité

Valeur SHAP

7.2 Valeur de Shapley dans la théorie de jeux

En théorie des jeux, plus précisément dans un jeu coopératif, la valeur de Shapley donne une répartition équitable des gains parmi les joueurs d'une coalition.

Shapley a proposé cette répartition « équitable » dans le cas où l'utilité est transférable.

Considérons un jeu caractérisé par le couplet (N, v) où

- $N = \{1, \dots, n\}$ est un ensemble de n joueurs, $n \in \mathbb{N}$,
- une *coalition* est un sous-ensemble de joueurs $Z \in \mathcal{P}(N)$
- et $v : \mathcal{P}(N) \rightarrow \mathbb{R}$ est la fonction caractéristique (aussi appelée jeu conditionnel), c'est-à-dire une fonction qui associe à chaque coalition C une valeur $v(C) \in \mathbb{R}$ telle que $v(\emptyset) = 0$ avec $\mathcal{P}(N)$ l'ensemble des parties de N .

La fonction caractéristique v décrit le gain ou importance de chaque coalition : l'objectif du jeu est alors de trouver un opérateur φ , qui assigne au jeu $(\{1, \dots, n\}, v)$, un vecteur $\varphi = (\varphi_1, \dots, \varphi_n)$, qui est la répartition la plus équitable du gain total (de la grande coalition $N = \{1, \dots, n\}$).

Définition 20. La valeur de Shapley pour le joueur i est notée $\varphi_i(v)$ et elle est définie par :

$$\varphi_i(v) = \sum_{Z \subseteq \mathcal{P}(N \setminus \{i\})} \frac{|Z|!(n - |Z| - 1)!}{n!} * [v(Z \cup \{i\}) - v(Z)] \quad (7.2)$$

où

- $|Z|$ est le nombre de joueurs de la coalition Z
- n le nombre total de joueurs
- v est la fonction caractéristique

Théorème 4. (Shapley (1953))[16] La valeur de Shapley φ est l'unique solution dans un jeu avec utilité transférable qui distribue le gain total $v(\{1, \dots, n\})$ équitablement parmi les joueurs et qui vérifie les quatre axiomes suivants :

1. *Efficacité* : $\sum_{i=1}^n \varphi_i(v) = v(\{1, \dots, n\})$
2. *Symétrie* : Pour tout couple de joueurs $(i, j) \in \{1, \dots, n\}^2$, si

$$\forall Z \in \mathcal{P}(\{1, \dots, n\} \setminus \{i, j\}) : \\ v(Z \cup \{i\}) = v(Z \cup \{j\}), \text{ alors } \varphi_i(v) = \varphi_j(v)$$

Par analogie, deux variables explicatives dans un modèle de prédiction, qui ont le même impact sur la prédiction auront des valeurs de contribution identiques.

3. *Facticité* : Soit $i \in \{1, \dots, n\}$ un joueur. Si $\forall Z \in \mathcal{P}(\{1, \dots, p\} \setminus \{i\}), v(S \cup \{i\})$, alors $\varphi_i(v) = 0$. Par analogie, une variable explicative dans un modèle de prédiction, qui a une contribution de 0, n'aura aucune influence sur la prédiction.

4. *Additivité* : Pour tous les jeux de coalitions (N, v) , (N, w) composés du même ensemble de joueurs, avec

$$\begin{aligned} v &: \mathcal{P}(N) \rightarrow \mathbb{R} \\ w &: \mathcal{P}(N) \rightarrow \mathbb{R} \end{aligned}$$

Alors

$$\phi(v + w) = \phi(v) + \phi(w),$$

avec :

$$\forall Z \in \mathcal{P}(\{1, \dots, n\}), (v + w)(Z) = v(Z) + w(Z)$$

Par analogie, si le modèle prédictif utilisé repose sur la moyenne de plusieurs modèles (comme les forêts aléatoires qui utilisent des arbres de décision) alors la contribution de ce modèle sera la moyenne des contributions de chaque modèle pris seul.

7.3 Valeur Shapley dans un modèle prédictif

Dans cette section, un jeu de prédiction de la sinistralité sera repensé selon la théorie des jeux coopératifs, en choisissant la quantité d'intérêt comme dans un des quadrants de la figure 7.1.

Prédiction d'une donnée de sortie

Supposons de vouloir prédire une donnée de sortie $Y \in \mathbb{R}$, en observant $\mathcal{D} = \{(d_{ij}, y_i) : j = 1, \dots, n; i = 1, \dots, p\}$, un jeu des données constituées de p individus d_1, \dots, d_p sur n variables explicatives X_1, \dots, X_n .

Définition 21. On appelle **prédicteur** une fonction

$$\hat{f} : \mathbb{R}^n \rightarrow Y \tag{7.3}$$

qui associe à une observation une valeur de Y , et **prédiction** :

$$\hat{y} = \hat{f}(X)$$

la réponse d'un nouvel individu de caractéristiques $X \in \mathbb{R}^n$,

Définition 22. On choisit une **fonction de perte** ou coût de la prédiction

$$l : Y \times Y \rightarrow \mathbb{R}_+ \tag{7.4}$$

qui s'applique à l'observation y et à sa prédiction \hat{y} .

7.3.1 SHAP zone : Sensibilité locale d'une donnée de sortie selon la valeur de Shapley

A l'aide de la valeur de Shapley on cherche à définir un partage équitable de la valeur prédite, pour tout modèle d'apprentissage, en respectant l'apport de chaque variable dans la coalition. La valeur prédite peut être expliquée en supposant que chaque variable d'une observation est un "joueur" dans un jeu où la prédiction est la sinistralité.

Afin d'attribuer une structure additive à la prédiction, Lundberg et Lee en 2017 [13] ont conçu et mis en œuvre SHAP (SHapley Additive exPlanations), une approche unifiée de l'importance des variables et de l'interprétation des prédictions.

Dans "*A unified approach to interpreting model predictions*" [13], ils introduisent trois notions principales :

1. Définition d'une classe des méthodes d'explication, appelée *Additive Feature Attribution Methods*. L'idée est que le fait de voir toute explication d'une prédiction est un modèle lui-même, noté modèle d'explication.
2. la quantité SHAP comme unique modèle d'explication qui appartient à la classe "Additive Feature Attribution Methods" et satisfait trois bonnes propriétés (précision locale, absence et cohérence, qui seront énoncées par la suite)

3. et des approximation de SHAP.

Cette théorie peut être appliquée à n'importe quel jeu de données, où l'on souhaite prédire une donnée de sortie.

Avec les notations comme dans la section 7.2, et l'ensemble des joueurs $\{X_1, \dots, X_n\}$, nous introduisons le jeu coopératif (\mathcal{D}, v_x) de prédiction de la sinistralité pour une observations $x = (x_1, \dots, x_n)$, où :

- le gain v_x est la prédiction de la sinistralité d'une coalition des variables, $\hat{f}(x)$, moins la prédiction moyenne de tous les individus du jeu de données ($\hat{f}(x) - \mathbb{E}(\hat{f}(X))$).

En d'autres termes, la fonction caractéristique v_x décrit le gain marginale d'une coalition selon les variables exclues de cette coalition :

$$v_x(Z) = \int \hat{f}(x_1, \dots, x_n) d\mathbb{P}_{X \notin Z} - \mathbb{E}_X(\hat{f}(X))$$

où x est une réalisation des variables X_1, \dots, X_n et la notation $X \notin Z$ est le vecteur aléatoire constitué des variables exclues dans la coalition Z .

Par exemple, si l'on considère un modèle à quatre entrées réelles X_1, X_2, X_3 et X_4 et on souhaite calculer le gain marginal de la coalition $S = \{X_1, X_3\}$, on a :

$$v_x(S) = v_x(\{x_1, x_3\}) = \int_{\mathbb{R}} \int_{\mathbb{R}} f(x_1, X_2, x_3, X_4) d\mathbb{P}_{X_2, X_4} - \mathbb{E}(f(X))$$

- les variables X_1, \dots, X_n sont des joueurs qui collaborent pour recevoir le gain.

L'objectif du jeu de prédiction est d'expliquer le gain d'un individu x selon les covariables du jeu de données. L'élément clé des explications est la contribution des caractéristiques d'entrée de chaque individu : une prédiction est expliquée en attribuant à chaque variable une quantité qui dénote son influence.

Valeur de Shapley : une méthode agnostique aux modèles

Une autre propriété souhaitable pour un modèle d'explication est l'agnosticité du modèle de prédiction, c'est à dire l'indépendance entre le modèle choisi et la définition de l'importance des variables.

Du point de vue opérationnel, il est plus facile de travailler avec des explications indépendantes du modèle lorsque plusieurs modèles d'apprentissages statistiques sont évalués et comparés.

Valeur de Shapley dans la régression linéaire

Dans une régression linéaire à n prédicteurs, par définition, la valeur prédite s'écrit comme la somme des multiplieurs :

$$\hat{f}(x) = \beta_0 + \sum_{i=1}^n \beta_i x_i$$

avec $\beta_i \in \mathbb{R}, \forall i = 1, \dots, n$.

La valeur de Shapley de la j -ème variable est ainsi :

$$\varphi_j(\hat{f}) = \beta_j x_j - \mathbb{E}(\beta_j X_j) = \beta_j x_j - \beta_j \mathbb{E}(X_j)$$

Cette valeur, appelée aussi **contribution** de la variable X_j à la prédiction $\hat{f}(x)$ est la différence entre l'effet de la variable pour cette observation en particulier et l'effet moyen de la variable.

Pour une observation x , la somme des contributions des variables est ainsi la différence entre la valeur prédite et la prédiction moyenne :

$$\begin{aligned} \sum_{j=1}^n \varphi_j(\hat{f}) &= \sum_{j=1}^n (\beta_j x_j - \mathbb{E}(\beta_j X_j)) \\ &= (\beta_0 + \sum_{j=1}^n \beta_j x_j) - (\beta_0 + \sum_{j=1}^n \mathbb{E}(\beta_j X_j)) \\ &= \hat{f}(x) - \mathbb{E}(\hat{f}(X)) \end{aligned}$$

Pour les modèles additifs, une estimation de la valeur de Shapley de la j -ème variable s'obtient à partir de l'estimateur de $\mathbb{E}(\hat{f}(x_1, \dots, X_j, \dots, x_n))$, en perturbant les valeurs de la j -ème variable, tandis que les valeurs des autres caractéristiques d'entrée restent fixes.

Toutefois, on ne connaît pas toujours la fonction de prédiction f , notamment dans les cas des modèles plus sophistiqués. Pour cela, il faut ainsi généraliser la valeur de Shapley à tous les modèles de prédiction, à l'aide de la prédiction conditionnelle $\mathbb{E}[\hat{f}(X_1, \dots, X_n) | X_i = x_i, \forall i \in Z]$.

Généralisation de la valeur de Shapley

On se place dans le cadre d'un modèle quelconque, avec \hat{f} la fonction de prédiction associée. Soit $x = (x_1, \dots, x_n)$ un individu dont on souhaite expliquer la prédiction.

Définition 23. Dans le cas général, nous allons utiliser comme fonction de gain (ou fonction caractéristique) la différence de prédiction. Il s'agit du changement dans la prédiction causé par l'observation des variables explicatives d'une coalition.

La **différence de prédiction** Δ^x associée au sous-ensemble de variables explicatives $Z = \{X_{i_1}, \dots, X_{i_s}\} \subset \{X_1, \dots, X_n\}$, où $\{i_1, \dots, i_s\} \subset \{1, \dots, n\}$ est un sous-ensemble d'indices, avec $s \in \{1, \dots, n\}$ est définie par :

$$\Delta^x(Z) = \underbrace{\mathbb{E}[\hat{f}(X_1, \dots, X_n) | X_{i_1} = x_{i_1}, \dots, X_{i_s} = x_{i_s}]}_{:=f_x(Z)} - \underbrace{\mathbb{E}[\hat{f}(X_1, \dots, X_n)]}_{:=f_x(\{\})} \quad (7.5)$$

Le couplet $(\{X_1, \dots, X_n\}, \Delta^x)$ forme un jeu coopératif.

Définition 24. La contribution de la variable explicative X_j , $j \in \{1, \dots, n\}$, est définie comme la valeur de Shapley du jeu de coopératif $(\{X_1, \dots, X_n\}, \Delta^x)$:

$$\varphi_i(\Delta^x) = \sum_{Z \in \mathcal{P}(N \setminus \{i\})} \frac{|Z|!(n - |Z| - 1)!}{n!} * [\Delta^x(Z \cup \{i\}) - \Delta^x(Z)] \quad (7.6)$$

La quantité $[\Delta^x(Z \cup \{i\}) - \Delta^x(Z)]$ représente la contribution de la variable i à la coalition Z . À la différence de la définition 7.2, ici la valeur de Shapley dépend de l'observation x .

Propriétés souhaitables des modèles d'explication

Les propriétés suivantes sont des propriétés souhaitables pour un modèle d'explication de la classe *Additive Feature Attribution*. Elles sont par ailleurs des propriétés vérifiées par la valeur de Shapley.

1. **Précision locale** : pour un modèle d'explication g (section 7.1),

$$f(x) = g(x') = \varphi_0 + \sum_{j=1}^n \varphi_j x'_j \quad (7.7)$$

En définissant $\varphi_0 = E_x(\hat{f}(x))$, avec le vecteur de coalition $x' = (1, \dots, 1)$, la 7.7 vient de la propriété d'efficacité de la valeur de Shapley : $f(x) = \varphi_0 + \sum_{j=1}^n \varphi_j x'_j = E_x(\hat{f}(x)) + \sum_{j=1}^n \varphi_j$.

2. **Absence** : si une variable est manquante sa contribution est nulle :

$$x'_j = 0 \Rightarrow \varphi_j = 0 \quad (7.8)$$

Cette propriété, appelée *minor book-keeping property*, est ajoutée aux propriétés classiques de la valeur de Shapley, pour indiquer que les variables constantes n'ont pas de contribution.

3. **Cohérence** Soit $f_x(z') = f(h_x(z'))$. Pour tout modèle f et f' tels que

$$f'_x(z') - f'_x(z'_j) \geq f_x(z') - f_x(z'_j)$$

pour tout $z' \in \{0, 1\}^n$, alors :

$$\varphi_j(f', x) \geq \varphi_j(f, x) \quad (7.9)$$

où la notation z'_j signifie que $z'_j = 0$. La propriété de cohérence indique que si un modèle change de sorte que la contribution marginale d'une variable augmente ou reste la même (quelles que soient les autres variables), la valeur Shapley augmente ou reste également la même. A partir de la propriété de cohérence on retrouve les propriétés de linéarité, facticité et symétrie de la valeur de Shapley.

Valeur SHAP

Lundberg et Lee ont montré que localement uniquement le modèle d'explication g appartenant à la classe *Additive Feature Attribution* et satisfaisant ces trois propriétés souhaitables est celui dont les importances de variables sont les valeurs de Shapley (déf.7.6).

Définition 25. SHAP est l'unique modèle d'explication qui appartient à la classe des modèles *Additive feature Attribution* et satisfait les trois propriétés souhaitables.

Réécriture de SHAP

Soient $Z = \{X_{i_1}, \dots, X_{i_s}\} \subset \{X_1, \dots, X_n\}$ une coalition où $\{i_1, \dots, i_n\} \subset \{1, \dots, n\}$ est un sous-ensemble d'indices, avec $s \in \{1, \dots, n\}$ et Z' son vecteur de coalition.

On rappelle que $h_x : \{0, 1\}^n \rightarrow X$ est la fonction qui associe à vecteur de coalition une entrée des composantes égales à celles de x lorsque les variables sont dans la coalition désigné par le vecteur de coalition, des composantes aléatoires autrement.

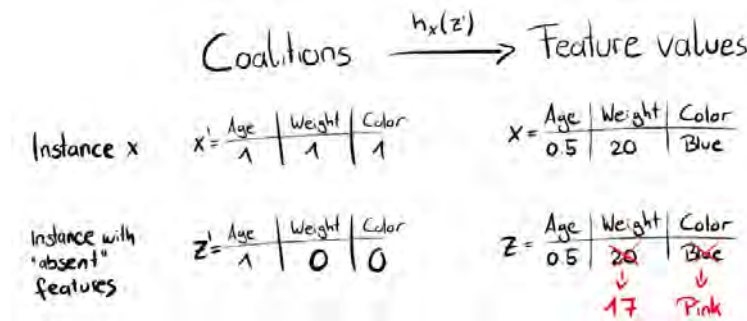


FIGURE 7.2 – Fonction h_x : lorsque les variables n'appartiennent pas à la coalition, la fonction associe des valeurs aléatoires. Elle peut associer la valeur manquante, quand le modèle de prédiction traite les valeurs manquantes (xgboost par exemple). Source : *Interpretable Machine Learning, A Guide for Making Black Box Models Explainable*, Christoph Molnar.

Avec les notations :

- $|Z'|$ est le nombre des entrées non nulles de Z' ,
- $\{Z' \subseteq x'\}$ indique tous les vecteurs de coalition Z' dont les entrées non nulles sont un sous-ensemble des entrées non nulles en x' , où x' est le vecteur de coalition de x
- $h_x(Z') = z_Z \in \mathbb{R}^n$, qui a des valeurs manquantes pour les variables qui ne sont pas dans Z ;
- $f_x(Z') = f(h_x(Z')) \approx \mathbb{E}(f(X)|z_Z)$

On peut simplifier la 7.6 :

$$\varphi_i(f, x) = \sum_{Z' \subseteq x'} \frac{|Z'|!(n - |Z'| - 1)!}{n!} [f_x(Z') - f_x(Z' \setminus i)] \quad (7.10)$$

Nous venons de présenter la valeur SHAP lorsque les variables sont indépendantes. Dans le cas de dépendance les méthodes présentées intègrent la notion d'ordre d'entrée des variables dans la prédiction.

7.3.2 SHAP Feature Importance

Malgré SHAP soit une méthode locale, il est possible construire un indicateur d'importance globale, noté I_j pour la variables X_j en agrégeant les valeurs SHAP de la base d'apprentissage :

$$I_j = \frac{1}{k} \sum_{i=1}^k |\varphi_j^{(i)}|$$

où k est la taille de la base d'apprentissage, et $\varphi_j^{(i)}$ est la valeur SHAP de la variable j associée à l'individu i .

7.3.3 Effects d'interaction SHAP

Les contributions des modalités sont généralement réparties entre les variables d'entrée, une pour chaque variable, mais nous pouvons obtenir des informations supplémentaires en séparant les effets d'interaction des effets principaux.

Ainsi, les effets d'interaction d'ordre deux seront donnés par l'impact de tous les couples de variables sur une prédiction du modèle. L'extension naturelle aux effets d'interaction, à partir des valeurs SHAP, est obtenue par la matrice symétrique des indices d'interaction de Shapley :

$$\Phi_{i,j} = \sum_{S \subseteq N \setminus \{i,j\}} \frac{|S|!(n - |S| - 2)!}{2(n-1)!} \nabla_{i,j}(S) \quad (7.11)$$

avec $i \neq j$ et

$$\begin{aligned}\nabla^{i,j}(S) &= f_x(S \cup \{i,j\}) - f_x(S \cup \{i\}) - f_x(S \cup \{j\}) + f_x(S) \\ &= f_x(S \cup \{i,j\}) - f_x(S \cup \{i\}) - [f_x(S \cup \{j\}) - f_x(S)]\end{aligned}$$

Effet principal d'une variable aléatoire

Tout comme l'indice de Sobol d'ordre 1 et l'indice de Sobol total, nous pouvons décomposer l'impact d'une variable entre la contribution de la variable seule et les contributions des interactions de la variable avec les autres variables.

Définition 26. Dans l'équation 7.11, la valeur d'interaction SHAP entre la variable X_i et la variable X_j est répartie également entre chaque variables donc $\Phi_{i,j} = \Phi_{j,i}$ et l'effet d'interaction total est $\Phi_{i,j} + \Phi_{j,i}$. L'effet principal pour une prédiction peut alors être défini comme la différence entre la valeur SHAP et les valeurs d'interaction SHAP d'ordre deux :

$$\Phi_{i,i} = \varphi_i - \sum_{j \neq i} \Phi_{i,j} \quad (7.12)$$

Ces valeurs d'interaction SHAP ont des propriétés similaires à SHAP et permettent de prendre en compte l'effet principal d'une variable et les effets d'interaction combinés pour une prédiction.

Interprétation

Coefficient de pondération de la valeur de Shapley

L'idée dans l'attribution de l'importance d'une variable selon la valeur de Shapley est de **moyenner l'impact** que la variable a pour toutes les combinaisons de variables possibles :

$$\varphi_i(\Delta^x) = \sum_{Z \in \mathcal{P}(N \setminus \{i\})} \underbrace{\overbrace{|Z|!(n - |Z| - 1)!}^{(C) \quad (D)}}_{\underbrace{n!}_{(B)}} * \underbrace{[\Delta^x(Z \cup \{i\}) - \Delta^x(Z)]}_{(A)}$$

Cette formule peut être interprétée :

- l'impact est mesuré par la contribution marginale de i , c'est-à-dire l'accroissement de gain que la coalition peut réaliser avec l'entrée de cette variable dans la coalition Z (A).

- Supposons qu'il est convenu que les joueurs vont rentrer successivement un par un dans une salle. Le joueur i , par exemple, va attendre son tour d'entrée et pour cela il va faire la queue, définie par un ordre (une permutation des joueurs).

Il y a $n!$ permutations possibles (B). Maintenant, pour tout coalition Z ne contenant pas i , il y a $|Z|! * (n - |Z| - 1)!$ permutations dans lesquelles les joueurs dans Z se trouvent tous devant i dans la queue (C), et après eux se trouve le joueur i .

Le terme (D) indique toutes les permutations possibles des joueurs derrière i dans la queue.

Si les permutations sont distribuées uniformément (aucun joueur ou coalition n'est prioritaire), le coefficient de pondération est ainsi la probabilité selon laquelle, quand le joueur i rentre dans la salle, il se trouve devant la coalition Z .

Puisque sa contribution marginale au moment où le joueur i rentre dans la salle est $[\Delta^x(Z \cup \{i\}) - \Delta^x(Z)]$, la valeur de Shapley pour i est tout simplement l'espérance des contributions marginales du joueur i sous l'hypothèse d'équiprobabilité de la formation des coalitions.

Intérprétation de la valeur SHAP

Nous transférons à présent cette interprétation au cas particulier de la valeur SHAP, où le gain de la coalition est défini par la différence de prédiction (7.5).

Lorsqu'un modèle attribue une prédiction pour une observation, toutes les variables ne jouent pas le même rôle : certaines d'entre elles peuvent avoir un grand impact sur la prédiction du modèle, tandis que d'autres

peuvent ne pas être pertinentes.

Par conséquent, on peut penser que l'effet de chaque caractéristique peut être mesuré en vérifiant quelle aurait été la prédiction si cette caractéristique était absente : plus le changement de la sortie du modèle est importante, plus la variable doit être importante.

Précisément, cette méthode est basée sur l'idée que les valeurs des variables d'une observation fonctionnent ensemble pour générer un changement dans la prédiction du modèle par rapport à la sortie attendue du modèle ($\mathbb{E}(f(X))$).

Ce changement total de prédiction est partagé parmi les variables de la manière *plus équitable* possible, prenant compte leurs contributions dans tous les sous-ensembles (coalitions) des variables.

La valeur de SHAP est donc la contribution marginale moyenne d'une variable (explicative) sur toutes les coalitions possibles.

L'ampleur et le signe de la valeur SHAP

L'ampleur et le signe des contributions sont importants.

Premièrement, si une variable a une contribution plus importante qu'une autre, elle a une plus grande influence sur la prédiction du modèle pour l'observation d'intérêt.

Deuxièmement, le signe de la contribution indique si la variable contribue à augmenter (si elle est positive) ou à diminuer (si elle est négative) la sortie du modèle pour une observation particulière.

Enfin, la somme des contributions des variables représente la différence entre la prédiction de sortie du modèle et la sortie attendue du modèle sans aucune information sur les réalisations des variables (propriété d'efficacité de la valeur de Shapley) :

$$f(x) = \varphi_0 + \sum_{j=1}^M \varphi_j x'_j = E_x(\hat{f}(x)) + \sum_{j=1}^M \varphi_j$$

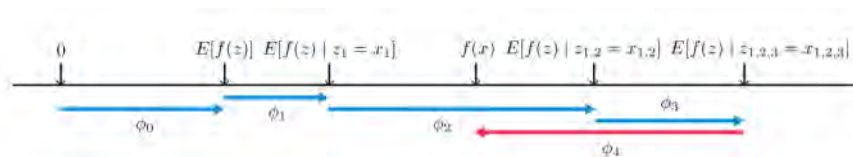


FIGURE 7.3 – Les valeurs SHAP (*SHapley Additive exPlanation*) attribuent à chaque variable le changement de prédiction attendue du modèle lors du conditionnement sur cette variable. Elles expliquent comment obtenir la prédiction à partir de la valeur de base $\mathbb{E}(f(X))$ qui serait prédite si nous ne connaissions aucune caractéristique de la prédiction actuelle. **Remarque** : lorsque le modèle est non linéaire ou que les entités en entrée ne sont pas indépendantes, l'ordre dans lequel les variables sont ajoutées compte, et les valeurs SHAP résultent de la moyenne des valeurs φ_i dans tous les ordres possibles.

7.4 Estimation de la valeur Shapley

Lorsque la distribution des variables du modèle et la forme analytique du modèle ne sont pas connues, les valeurs de Shapley sont estimées à partir de plusieurs algorithmes :

- **Shapley Sampling Values**, un estimateur de type Monte Carlo, développé par Štrumbelj et Kononenko (2014) ; nous noterons cet algorithme *SK* par la suite.
- **KernelSHAP**, qui se base sur les noyaux, inspirée aux modèles de substitution locaux (LIME), développé par Lundberg et Lee (2017) ;
- **TreeSHAP**[12], une extension de SHAP pour les modèles d'apprentissage basée sur les arbres tels que les arbres de décision, les forêts aléatoires et gradient boosting, développée par Lundberg, Erion, and Lee (2018). Nous présenterons par la suite ces trois estimateurs.
- Aas, Jullum, et Løland (2019) ont étendu la méthode KernelSHAP de Lundberg et Lee à des modèles à variables dépendantes.

De plus, il existe quatre méthodes spécifiques au modèle de prédiction : MaxSHAP, DeepSHAP, LinearSHAP et Low-OrderSHAP.

Ces approximations, à l'exception de celle introduite par Aas, Jullum et Løland, supposent que les variables sont indépendantes et que les individus sont choisis aléatoirement et indépendamment. Cette hypothèse est très forte puisque dans la pratique les variables ne sont pas forcément indépendantes.

Approximation par Monte-Carlo

Le temps de calcul des valeurs Shapley comme pour la définition (7.6) est de l'ordre exponentiel : en effet pour chaque variables, il faut calculer autant de quantités Δ^x que de sous-ensembles possibles (2^n) avec et sans cette variable.

Pour éviter cette contrainte calculatoire, des approximations ont été proposées. Tout d'abord, on passe par une écriture équivalente de (7.6),

$$\varphi_i(\Delta^x) = \frac{1}{n!} \sum_{O \in \mathcal{S}(N)} [\Delta^x(\text{Pre}^i(O) \cup \{i\}) - \Delta^x(\text{Pre}^i(O))] \quad (7.13)$$

où $\text{Pre}^i(O)$ est l'ensemble des joueurs qui sont prédécesseurs du joueur i dans la permutation $O \in \mathcal{S}(N)$ et $\mathcal{S}(N)$ est l'ensemble de toutes les permutations de N joueurs.

En simplifiant $\Delta^x(Z) = \sum_{y \in X} p(y)(f(\tau(x, y, Z)) - f(y))$ et en posant

- $\tau(x, y, Z) = (z_{(1)}, \dots, z_{(n)})$, avec $z_i = x_i$ si $i \in Z$, sinon $z_i = y_i$
- $p(\cdot)$ est la distribution de probabilité de X (individus du dataset), les variables sont supposées mutuellement indépendantes

l'équation (7.13) devient :

$$\varphi_i(\Delta^x) = \frac{1}{n!} \sum_{O \in \mathcal{S}(N)} \left[\sum_{y \in X} p(y) \underbrace{(f(\tau(x, y, \text{Pre}^i(O) \cup \{i\})) - f(\tau(x, y, \text{Pre}^i(O))))}_{:=V_{O,z}} \right]. \quad (7.14)$$

Strumbelj et Kononenko [21] ont proposé l'approximation suivante s'appuyant sur des méthodes de simulation par Monte Carlo :

$$\begin{aligned} \hat{\varphi}_j^M &= \frac{1}{M} \sum_{m=1}^M \left(\hat{f}(x_{+j}^m) - \hat{f}(x_{-j}^m) \right) \\ &= \frac{1}{M} \sum_{m=1}^M V_m \end{aligned}$$

où

- $j \in 1, \dots, n$ est l'indice de la variable qu'on souhaite expliquer ;
- $\hat{f}(x_{+j}^m)$ est la prédiction pour le vecteur $x = (x_1, \dots, x_n)$ de n variables explicatives, mais avec un nombre de caractéristiques aléatoires remplacées par un point z aléatoire, à l'exception de la valeur de la caractéristique.

Cet algorithme ne calcule pas les contributions marginales de toutes les coalitions : on fixe un nombre M de coalitions à échantillonner.

Puis, pour chaque coalition échantillonnée, les variables non sélectionnées dans la coalition sont remplacée par une simulation.

M est choisi suffisamment grand pour estimer avec précision les valeurs de Shapley, mais suffisamment petit pour que le temps de calcul soit raisonnable : un M petit réduit le temps de calcul, mais augmente la variance de la valeur de Shapley. En général, on établit un critère d'arrêt des itérations.

Algorithme d'approximation de la Valeur de Shapley de la j -ème variable, pour une observation $x \in X$ et le modèle f :

• **Inputs :**

1. modèle \hat{f} ,
2. observation x qu'on cherche à expliquer,
3. M : nombre d'itérations de l'algorithme
4. $\varphi_j = 0$

- **for** $m = 1$ **to** M **do**

1. choisir une permutation aléatoire $O \in S(n)$
2. choisir une observation $z = (z_1, \dots, z_n)$ du dataset initial
3. ré-ordonner x et z :
 - $x_O = (x_{(1)}, \dots, x_{(n)})$
 - $z_O = (z_{(1)}, \dots, z_{(n)})$
4. construire deux nouvelles instances :
 - avec la réalisation $x_j : \forall k \in \{1, \dots, n\}$,

$$x_k^{+,m} = \begin{cases} x_k & \text{si le joueur } k \in \text{Pre}^j(O) \cup \{j\} \\ z_k & \text{sinon} \end{cases}$$

- sans la réalisation $x_j : \forall k \in \{1, \dots, n\}$,

$$x_k^{-,m} = \begin{cases} x_k & \text{si le joueur } k \in \text{Pre}^j(O) \\ z_k & \text{sinon} \end{cases}$$

5. calculer la contribution marginale $\varphi_j^m = \varphi_j + \hat{f}(x_k^{+,m}) - \hat{f}(x_k^{-,m})$
à chaque itération $\hat{f}(x_k^{+,m})$ et $\hat{f}(x_k^{-,m})$ se basent sur des vecteurs identiques sauf pour la variable X_j
- calculer la valeur de Shapley comme moyenne des contributions marginales précédemment calculées :

$$\varphi_j^M(x) = \frac{1}{M} \sum_{m=1}^M \varphi_j^m$$

La moyenne prend en compte implicitement les échantillons par la distribution de probabilité de X .

On réitère l'algorithme n fois, pour obtenir les valeurs Shapley de toutes les variables.

Convergence de l'algorithme et Optimisation

Selon le Théorème Central Limite (TCL), l'estimateur $\hat{\varphi}_j$ suit asymptotiquement une loi normale de moyenne φ_j et une variance $\frac{\sigma_j^2}{M}$, avec σ_j^2 est la variance de la population pour la variable X_j .

$$\frac{\hat{\varphi}_j^M - \varphi_j}{\sigma_j / \sqrt{M}} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1)$$

$$\hat{\varphi}_j^M \xrightarrow{\mathcal{L}} \mathcal{N}\left(\varphi_j, \frac{\sigma_j^2}{M}\right)$$

Autrement dit, $\hat{\varphi}_j$ est un estimateur asymptotiquement sans biais et consistant de φ_j .

La vitesse de convergence dépend de plusieurs éléments :

- de la variance σ_j^2 , $j = 1, \dots, n$ des variables dans la base d'apprentissage
- du nombre de variables n
- du nombre d'itération Monte Carlo M

Si on voulait approcher la valeur de Shapley pour chacune des observations, le temps de calcul serait de l'ordre de $\mathcal{O}(2npM)$, avec n nombre de variables, p nombre des individus et M itérations.

Avec l'optimisation comme expliqué dans la figure 7.4, le nombre de calcul se réduit à $\mathcal{O}(2pn)$.

Nous pouvons également paralléliser l'algorithme sur n ou p , en fonction de ce qui est le plus avantageux.

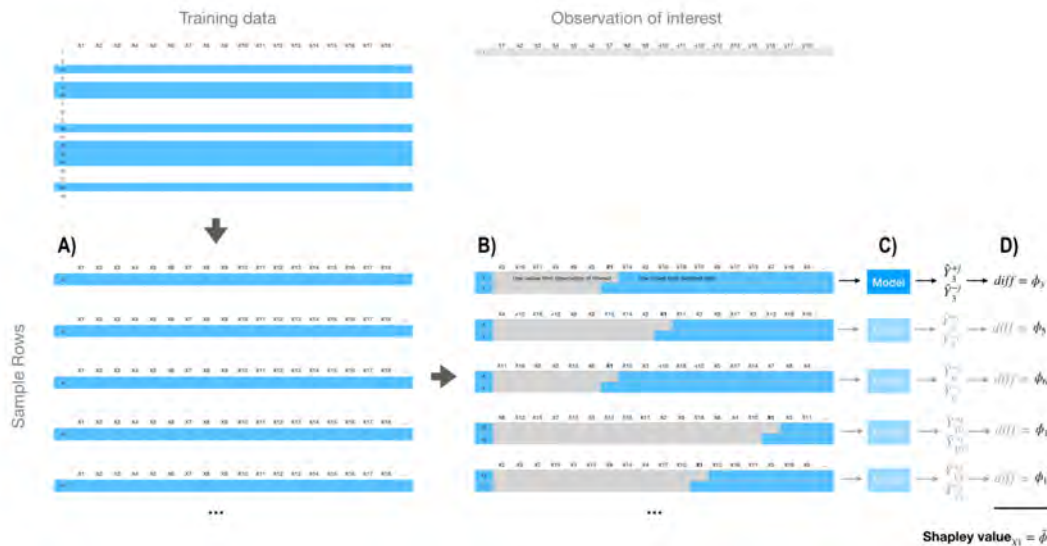


FIGURE 7.4 – Itération de l’algorithme *SK* pour l’estimation de φ_1 .

À l’étape (A), M observations sont sélectionnées. À l’étape (B), on crée deux copies d’une ligne échantillonnée individuellement et on choisit une observation z . Ensuite, dans une copie, nous incluons toutes les valeurs de l’observation d’intérêt pour les valeurs de la première colonne jusqu’à la j -ème variable incluse. Nous incluons ensuite les valeurs de l’observation z pour toutes les autres composantes. La deuxième copie est identique à l’exception de la variable X_j . Ensuite, à l’étape (C), on calcule les prédictions respectives et à l’étape (D) la différence entre les sorties prévues. Cette procédure est réitérée M fois et l’estimation de la valeur de Shapley pour la variable X_j est la différence moyenne sur toutes les lignes échantillonnées. Par rapport à l’algorithme de base, pour gagner du temps de calcul, plutôt que de calculer chaque itération de Monte Carlo, on sélectionne avant l’étape A) M observations à partir du *Training Data* et on applique la fonction de prédiction.

Intérêts et limites de la méthode de Monte Carlo

Le succès de cet algorithme dès lors sa première publication en 2013 est dû aux propriétés de la méthode de Monte-Carlo :

- Robustesse :
la méthode de Monte-Carlo est particulièrement bien adaptée au calcul d’intégrales a) Quand la dimension de l’espace d’intégration est grande (typiquement à partir de la dimension 4 ou 5) ; b) d’autant plus que la fonction à intégrer est peu régulière.
- Parallélisation :
on dit qu’un calcul peut être mené de façon parallélisable lorsqu’il peut être décomposé en une multitude de sous-calculs indépendants ou de façon équivalente, lorsque plusieurs centaines d’unités de calcul (intelligemment gérées) peuvent l’effectuer plus rapidement qu’un seul. La méthode de Monte-Carlo se parallélise très bien car les simulations sont indépendantes.

À l’inverse, la méthode de Monte-Carlo présente également des limites non négligeables.

La convergence vers la loi normale est souvent trop lente pour qu’on puisse supposer que l’approximation par la gaussienne soit valable. La convergence de l’estimateur est limitée à l’approximation des intégrales dont l’intégrande est de classe L^2 , ce qui n’est pas toujours vérifiable.

Une manière d’essayer de vérifier cette condition est de regarder si l’estimateur empirique de la variance reste bien stable quand le nombre de simulations augmente : on peut en effet montrer que, dès lors que l’intégrande n’appartient pas à L^2 , cet estimateur diverge presque-surement vers $+\infty$.

Des explications locales à la compréhension globale

La méthode de Shapley est locale, mais les valeurs estimées peuvent être agrégées pour créer des explications globales.

Cela, dans le cas d’un dataset à plusieurs milliers d’individus, n’est pas recommandable puisque le temps de calcul est trop élevé.

KernelSHAP

KernelSHAP est une autre estimation de SHAP, qui permet de résoudre le problème de la dimension lorsque les variables sont nombreuses.

Cette méthode utilise le lien parmi les équations des modèles de substitution locaux LIME⁴ et la définition de SHAP.

La méthode LIME interprète les prédictions individuelles du modèle en se basant sur une approximation locale du modèle autour d'une prédiction donnée. Pour déterminer les importances des variables comme dans la section 7.1, LIME minimise la fonction :

$$\eta = \underset{g \in G}{\operatorname{argmin}} L(f, g, \pi_{x'}) + \Omega(g)$$

où L est une fonction de perte, g est le modèle d'explication, $\pi_{x'}$ est un noyau local (utilisé pour pondérer), f le modèle de prédiction et Ω est une pénalité sur la complexité de g .

KernelSHAP est ainsi un cas particulier de LIME[14], où l'on choisit judicieusement la fonction de perte, les poids et la fonction de régularisation pour lui conférer les propriétés de précision local et de consistance.

Cette combinaison permet par ailleurs de fournir des explications plus parcimonieuses.

Théorème 5 (Shapley kernel[13]). *Un modèle d'explication de la classe Additive feature attribution qui minimise la fonction objectif de LIME et satisfaisant les trois propriétés 7.7, 7.8, 7.9 est tel que :*

$$\begin{aligned} \Omega(g) &= 0 \\ \pi_{x'}(z') &= \frac{(n-1)}{\binom{n}{|z'|} |z'| (n-|z'|)} \\ L(f, g, \pi_{x'}) &= \sum_{z' \in Z} [f(h_x(z')) - g(z')]^2 \pi_{x'}(z'), \end{aligned}$$

où z' et x' sont des vecteurs de coalition, $|z'|$ est le nombre des éléments non nuls de z' .

Remarque

- $\pi_{x'}(z') = \infty$, si $|z'| \in \{0, n\}$, qui implique que $\varphi_0 = f_x(\emptyset)$ et $f(x) = \sum_{i=0}^n \varphi_i$.
- Le modèle d'explication est ainsi $g(z') = \varphi_0 + \sum_{j=1}^n \varphi_j z'_j$, où g est un modèle linéaire entraîné en optimisant la fonction de perte : $L(f, g, \pi_{x'}) = \sum_{z' \in Z} [f(h_x(z')) - g(z')]^2 \pi_{x'}(z')$, la somme des erreurs quadratiques pondérées, avec Z ensemble des vecteurs de coalition de la base d'apprentissage.

Les coefficients φ_j sont les valeurs de Shapley.

Les valeurs de Shapley de la théorie des jeux peuvent donc être calculés en utilisant la régression linéaire pondérée. Puisque nous sommes dans un cadre de régression linéaire, nous pouvons également utiliser les outils standard de régression. Par exemple, nous pouvons ajouter des termes de régularisation pour rendre le modèle sparse. Si nous ajoutons une pénalité $L1$ à la perte L , nous pouvons créer des explications sparse, sous conditions que les coefficients soient des valeurs Shapley.

Avantages et limites

L'avantage majeur de cette technique est la parcimonie : le théorème 5 relie les valeurs de Shapley de la théorie des jeux à la régression linéaire pondérée. KernelSHAP utilise cette connexion pour calculer l'importance des variables. Cela conduit à des estimations plus **précises avec moins évaluations** du modèle original que les estimations précédentes fondées sur l'échantillonnage Monte Carlo.

Comme dans l'estimateur de Strumbelj et Kononenko [21], KernelSHAP est une méthode d'interprétation basées sur la permutation. Puisqu'elle simule les variables exclues par les coalitions, l'estimation risque de donner un poids importants aux cas les moins probables. On ne peut pas pallier cette limite, car si les variables absentes étaient échantillonnées à partir de la distribution conditionnelle les valeurs résultantes violeraient l'axiome Shapley d'absence (*Dummy*), qui dit qu'une caractéristique qui ne contribue pas au résultat a une valeur Shapley de zéro.

De plus, en grande dimension, le temps de calcul reste toujours très important.

Interprétation

KernelSHAP est une méthode qui donne plus de poids aux très petites coalitions et aux très grandes coalitions.

⁴Pour plus de détails sur la méthode LIME voir [14]

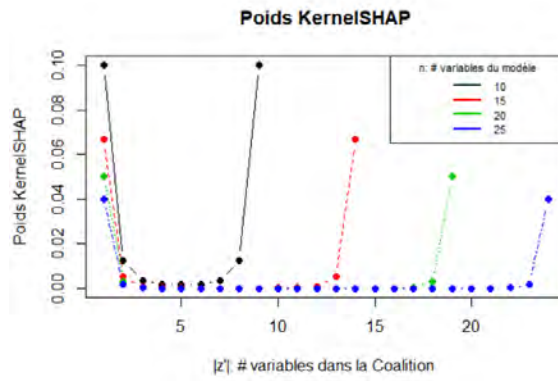


FIGURE 7.5 – Calcul des poids selon l’approximation KernelSHAP

L’intuition sous-jacente est la suivante : nous apprenons le plus sur les caractéristiques individuelles si nous pouvons étudier leurs effets individuels.

Si une coalition se compose d’une seule variable, nous pouvons en apprendre davantage sur l’effet principal isolé des caractéristiques sur la prédiction.

Si une coalition se compose de toutes les variables sauf une, nous pouvons en apprendre davantage sur l’effet total de ces variables (effet principal et interactions des variables).

Si une coalition comprend la moitié des variables, nous en apprenons peu sur la contribution d’une variable individuelle, car il existe de nombreuses coalitions possibles avec la moitié des variables.

LinearSHAP

Pour les modèles linéaires, dans le cas où les entrées sont indépendantes, les valeurs SHAP se calculent directement à partir des coefficients de régression linéaire du modèle. Cela a été montré dans la section 7.3.1.

TreeSHAP

Malgré leurs avantages respectifs, les estimateurs Strumbelj et Kononenko et KernelSHAP n’arrivent pas à estimer l’espérance conditionnelle $\mathbb{E}(f(x)|z_S)$ sans violer les propriétés souhaitables et la complexité exponentielle de la 7.10. Pour cela, ils simulent les valeurs manquantes des coalitions, ce qui augmente la variance des estimations.

TreeSHAP, une extension de SHAP spécifique aux modèles de type CART, Random Forest et Gradient Boosting permet de pallier cette limite. Nous allons détailler l’algorithme par un exemple.

Algorithme TreeSHAP dans un arbre de régression

Examinons les valeurs SHAP pour un arbre de régression simple afin d’illustrer l’algorithme.

Considérons un modèle de trois variables explicatives numériques indépendantes (à savoir : X, Y, Z) et une variable cible T .

Nous obtenons l’arbre ci-dessous :

Calculons les valeurs SHAP pour un individu i donnée par $x_i = 150, y_i = 75, z_i = 200$. La prédiction pour cette instance est $t_i = 20$.

Nous commençons par calculer les valeurs SHAP dans un modèle vide sans aucune variable indépendante, puis on calcule la contribution marginale moyenne lorsque chaque variable est ajoutée à ce modèle dans une permutation, puis on la moyenne sur toutes les permutations possibles.

Puisque nous avons ici 3 variables indépendantes, nous devons considérer $3! = 6$ permutations.

Calculons les contributions marginales pour chaque permutation.

La prédiction pour le modèle vide φ_0 (également appelée valeur de base) est la prédiction moyenne pour l’ensemble d’apprentissage :

$$\varphi_0 = (50 * 2 + 30 * 2 + 20 * 1 + 10 * 5) / 10 = 23$$



FIGURE 7.6 – n1, n2, n3, . . . , n7 représentent les nœuds de l’arbre. Les valeurs de s dans chaque nœud représentent le nombre d’observations de l’ensemble d’apprentissage qui appartiennent à chaque nœud.

On considère la permutation : $\sigma_1 = \{X, Y, Z\}$:

1. Tout d’abord, la variable X est ajoutée au modèle vide.
Pour l’observation sélectionnée i , nous pouvons calculer la prédiction exacte avec juste cette information car seule la variable X est utilisée dans les nœuds (n1 & n3) menant au nœud feuille n6.
Ainsi, la prédiction du modèle avec juste la caractéristique X est 20. Par conséquent, la contribution marginale de X dans cette permutation, $\varphi_{X_{\sigma_1}} = 20 - 23 = -3$.
2. Maintenant, ajoutons la variable Y au modèle. Puisque l’ajout de Y ne modifie pas la prédiction pour l’instance sélectionnée i , la contribution marginale pour y dans cette permutation, $\varphi_{Y_{\sigma_1}} = 23 - 23 = 0$.
3. De même, la contribution marginale de Z dans cette permutation, $\varphi_{Z_{\sigma_1}} = 0$.

On considère la permutation : $\sigma_2 = \{Y, Z, X\}$

1. Tout d’abord, la variable Y est ajoutée au modèle vide.
Le premier nœud n1 utilise X comme variable de division ; puisque X n’est pas encore disponible, nous calculons la prédiction comme

$$(4/10) * (\text{prédiction à partir des feuilles filles gauche n2}) + (6/10) * (\text{prédiction à partir des feuilles filles droit n3}),$$

100, 60 et 40 étant le nombre d’échantillons d’apprentissage tombant respectivement dans les nœuds n1, n2 et n3.

La prédiction pour le modèle avec juste la caractéristique Y est $(4/10) * 50 + (6/10) * (70/6) = 27$.
Par conséquent, la contribution marginale de Y dans cette permutation, $\varphi_{Y_{\sigma_2}} = 27 - 23 = 4$.

2. Ensuite, nous ajoutons la variable Z au modèle.
Puisque Z n’est pas utilisé comme variable de division dans aucun des nœuds internes de l’arbre, l’ajout de cette fonction ne modifie en aucune façon la prédiction.
Ainsi la contribution marginale de Z dans cette permutation est 0.
3. Enfin, nous ajoutons la variable X au modèle qui donne la prédiction 20. Par conséquent, la contribution marginale de X dans cette séquence est $\varphi_{X_{\sigma_2}} = 20 - 27 = -7$

On calcule la valeur SHAP pour les autres permutations et on calcule la moyenne :

$$\begin{aligned} \varphi_Y &= (\varphi_{Y_{\sigma_1}} + \varphi_{Y_{\sigma_2}} + \varphi_{Y_{\sigma_3}} + \dots + \varphi_{Y_{\sigma_6}}) / 6 = 2 \\ \varphi_X &= (\varphi_{X_{\sigma_1}} + \varphi_{X_{\sigma_2}} + \varphi_{X_{\sigma_3}} + \dots + \varphi_{X_{\sigma_6}}) / 6 = -5 \\ \varphi_Z &= (\varphi_{Z_{\sigma_1}} + \varphi_{Z_{\sigma_2}} + \varphi_{Z_{\sigma_3}} + \dots + \varphi_{Z_{\sigma_6}}) / 6 = 0 \end{aligned}$$

On vérifie que la somme des valeurs de Shapley et de l’effet moyen est bien la prédiction pour l’individu i : $\varphi_0 + \varphi_X + \varphi_Y + \varphi_Z = 23 + (-5) + 2 + 0 = 20$. On peut interpréter les valeurs de Shapley de la façon suivante :

“ La valeur de base de la prédiction en l’absence de toute information sur les variables indépendantes est de 23 ; sachant $x_i = 150$ la prédiction diminue de 5 et sachant $y_i = 75$ elle augmente de 2, donnant une prédiction finale de 20. Connaître $z_i = 300$ n’a aucun impact sur la prédiction du modèle. ”

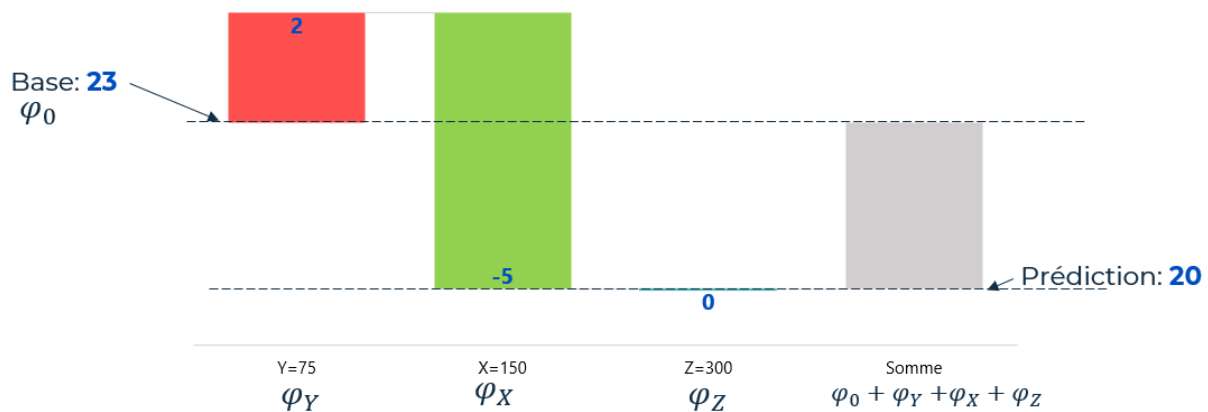


FIGURE 7.7 – Représentation de l’impact des variables dans la prédiction selon la valeur de Shapley : à partir de la prédiction moyenne, Y amène la prédiction vers le haut, alors que X la diminue de 5. Z n’a aucun rôle dans la prédiction.

Avantages et limites

Le grand avantage de TreeSHAP est qu’il permet de résoudre la limite des méthodes basées sur la permutation, sans simuler les variables lorsque elles sont exclues de la coalition.

En effet cette méthode estime l’espérance conditionnelle $\mathbb{E}_{X_S|X_{\bar{S}}}(f(X)|X_S)$ au lieu de l’espérance marginale, comme SK et KernelSHAP.

Le problème avec l’espérance conditionnelle est que les entités qui n’ont aucune influence sur la fonction de prédiction f peuvent obtenir une estimation TreeSHAP différente de zéro[10],[19]. L’estimation non nulle peut se produire lorsque l’entité est corrélée avec une autre caractéristique qui a en fait une influence sur la prédiction. La plus récente extension de TreeSHAP permet de résoudre ce problème.

Par ailleurs, c’est un algorithme récursif dont la complexité computationnelle est polynomiale $\mathcal{O}(TLD^2)$ (par opposition au coût exponentiel $\mathcal{O}(TL2^n)$ des méthodes précédentes), où T est le nombre d’arbres, L les nombre de feuilles et D le profondeur.

A différence des précédents, les implémentations de cet algorithme sous **R** et **python** intègrent aussi le calcul des interactions.

Synthèse des approximations de la valeur de Shapley

| Algorithme | Vitesse d'exécution | Agnosticité au modèle | Calcul des interactions intégré |
|------------|-----------------------|-----------------------|---------------------------------|
| SK | + ++ (parallélisé) | +++++ | Non |
| Exact | +++++ | + | Non |
| KernelSHAP | +++ | ++++ | Approximation |
| TreeSHAP | ++++ | ++ | Oui |

FIGURE 7.8 – Bien qu'il existe des techniques comme le clustering utilisant K-means pour réduire l'ensemble de données avant de calculer les valeurs Shapely, elles sont toujours lentes. KernelSHAP ignore la dépendance des fonctionnalités.

Pour réduire le temps d'exécution de KernelSHAP, l'algorithme de clustering k -means est utilisé pour sélectionner les observations.

| Algorithme | Avantages | Limites | Hypothèses |
|--------------------------|---|--|--|
| SK | <ul style="list-style-type: none"> Agnostique au modèle Robuste | <ul style="list-style-type: none"> Pas forcément convergent Remplacement aléatoire des valeurs manquantes Inadapté aux grand volume de données (problème de mémoire vive sous logiciel) | Indépendance variables |
| Exact Value (LinearSHAP) | Rapide | Spécifique au modèle linéaire | <ul style="list-style-type: none"> Indépendance variables Pas d'effet d'interaction Linéarité du modèle |
| KernelSHAP | Parcimonieux | Remplacement aléatoire des valeurs manquantes Inadapté aux grand volume de données | Indépendance variables |
| TreeSHAP | <ul style="list-style-type: none"> Rapide Calcul des interactions | Spécifique aux modèles basés sur les arbres | Modèle de type arbre |

FIGURE 7.9 – Récapitulatif des approximations de la valeur SHAP.

Remarques

Considérant la coalition des variables S , les algorithmes vus dans cette section proposent de calculer sa prédiction comme l'espérance conditionnelle empirique :

$$f(S) := \mathbb{E}[f(x_S, X_{\bar{S}} | X_S = x_S)] \quad (7.15)$$

où

- les variables de S sont fixes (observées)
- et \bar{S} est l'ensemble des variables exclues de la coalition, sur la base d'apprentissage

Afin d'avoir un estimateur robuste, il faudrait trop d'échantillons de $X_{\bar{S}}$ et l'algorithme serait trop coûteux. Les algorithmes se différencient parmi eux par la façon d'approcher cette quantité.

- La plupart des méthodes basées sur la permutation n'acceptent pas des modèles à variables manquantes, ici les variables exclues de S . Ces algorithmes, notamment **SK** et **KernelSHAP**, remplacent alors les valeurs manquantes par des valeurs aléatoires et approchent la prédiction d'une coalition (un sous-ensemble à variables manquantes) par l'espérance marginale :

$$f(S) := \mathbb{E}[f(x_S, X_{\bar{S}} | X_S = x_S)] \sim \mathbb{E}[f(x_S, X_{\bar{S}})]$$

Cette quantité est dite *interventionnelle*, pour souligner que l'intervention des variables S peut être considérée comme une intervention sur l'observation à expliquer.

Cependant, si deux variables appartenant une à S et l'autre à \bar{S} sont corrélées, l'intervention de X_S modifie la distribution de $X_{\bar{S}}$ et cela conduit à mettre trop de poids sur des données peu probables.

- **TreeSHAP**, calcule l'espérance conditionnelle :

$$\mathbb{E}_{X_{\bar{S}} | X_S}(f(X) | X_S)$$

Cette méthode peut toutefois introduire des effets inexistantes [19],[10].

7.5 Illustration dans le modèle Ishigami

Nous considérons le modèle de Ishigami, noté f , comme présenté dans la section 6.2.5.

On peut supposer que les joueurs sont les variables X_i et les quantités d'intérêt Q sont les fonctions caractéristiques et elles s'obtiennent de la section 7.3 en remplaçant le modèle f par l'espérance le long des variables manquantes du modèle f :

$$f_u = \mathbb{E}(f(X | X_u = x_u))$$

avec u coalition. Dans le modèle Ishigami a trois variables dans le modèle f , alors le nombre de coalitions possibles est $2^3 = 8$:

$$\begin{aligned} f_x(\{\emptyset\}) &= \mathbb{E}(f(X)) = \frac{a}{2} \\ f_x(\{x_1\}) &= \mathbb{E}(f(X | X_1 = x_1)) = \sin(x_1)(1 + b\frac{\pi^4}{5}) - \frac{a}{2} + f_x(\{\emptyset\}) \\ f_x(\{x_2\}) &= \mathbb{E}(f(X | X_2 = x_2)) = -\frac{a}{2} + a * ((\sin(X_2))^2) + f_x(\{\emptyset\}) \\ f_x(\{x_3\}) &= \mathbb{E}(f(X | X_3 = x_3)) = f_x(\{\emptyset\}) \\ f_x(\{x_1, x_2\}) &= \mathbb{E}(f(X | X_1 = x_1, X_2 = x_2)) = f_x(\{\emptyset\}) \\ f_x(\{x_1, x_3\}) &= \mathbb{E}(f(X | X_1 = x_1, X_3 = x_3)) = b * \sin(X_1) * (X_3^4 - \frac{X_3^4}{5}) + f_x(\{x_3\}) + f_x(\{x_1\}) - f_x(\{\emptyset\}) \\ f_x(\{x_2, x_3\}) &= \mathbb{E}(f(X | X_2 = x_2, X_3 = x_3)) = f(x_2, x_3) = f_x(\{\emptyset\}) \\ f_x(\{x_1, x_2, x_3\}) &= \mathbb{E}(f(X | X_1 = x_1, X_2 = x_2, X_3 = x_3)) = f_x(\{\emptyset\}) \end{aligned}$$

Les valeurs de Shapley du modèle de Ishigami sont :

$$\begin{aligned} \varphi_1(f, x) &= \sum_{Z' \subseteq x'} \frac{|Z'|!(n-|Z'|)!}{n!} [f_x(Z') - f_x(Z' \setminus X_1)] \stackrel{\text{equiprobabilité d'entrer dans la coalition}}{=} \\ &= \frac{1}{6} (\underbrace{[f_x(\{\emptyset\}) - f_x(\{\emptyset\})]}_{Z=\{\emptyset\}} + \underbrace{[f_x(\{x_1\}) - f_x(\{\emptyset\})]}_{Z=\{x_1\}} + \underbrace{[f_x(\{x_2\}) - f_x(\{\emptyset\})]}_{Z=\{x_2\}} + \underbrace{[f_x(\{x_3\}) - f_x(\{\emptyset\})]}_{Z=\{x_3\}}) + \\ &+ \underbrace{[f_x(\{x_1, x_2\}) - f_x(\{x_2\})]}_{Z=\{x_1, x_2\}} + \underbrace{[f_x(\{x_1, x_3\}) - f_x(\{x_3\})]}_{Z=\{x_1, x_3\}} + \underbrace{[f_x(\{x_2, x_3\}) - f_x(\{x_2, x_3\})]}_{Z=\{x_2, x_3\}} + \\ &+ \underbrace{[f_x(\{x_1, x_2, x_3\}) - f_x(\{x_2, x_3\})]}_{Z=\{x_1, x_2, x_3\}} = \\ &= \frac{1}{6} (\underbrace{[f_x(\{x_1\}) - f_x(\{\emptyset\})]}_{Z=\{x_1\}} + \underbrace{[f_x(\{x_1, x_2\}) - f_x(\{x_2\})]}_{Z=\{x_1, x_2\}} + \underbrace{[f_x(\{x_1, x_3\}) - f_x(\{x_3\})]}_{Z=\{x_1, x_3\}}) + \\ &+ \underbrace{[f_x(\{x_1, x_2, x_3\}) - f_x(\{x_2, x_3\})]}_{Z=\{x_1, x_2, x_3\}} \end{aligned}$$

$$\begin{aligned}
\varphi_2(f, x) &= \sum_{Z' \subseteq x'} \frac{|Z'|!(n-|Z'|)!}{n!} [f_x(Z') - f_x(Z' \setminus X_2)] \stackrel{\text{equiprobabilité d'entrer dans la coalition}}{=} \\
&= \frac{1}{6} \left(\underbrace{[f_x(\{x_2\}) - f_x(\{\emptyset\})]}_{Z=\{x_2\}} + \underbrace{[f_x(\{x_1, x_2\}) - f_x(\{x_1\})]}_{Z=\{x_1, x_2\}} + \underbrace{[f_x(\{x_2, x_3\}) - f_x(\{x_3\})]}_{Z=\{x_2, x_3\}} \right) + \\
&\quad + \underbrace{[f_x(\{x_1, x_2, x_3\}) - f_x(\{x_1, x_3\})]}_{Z=\{x_1, x_2, x_3\}} \\
\varphi_3(f, x) &= \sum_{Z' \subseteq x'} \frac{|Z'|!(n-|Z'|)!}{n!} [f_x(Z') - f_x(Z' \setminus X_3)] \stackrel{\text{equiprobabilité d'entrer dans la coalition}}{=} \\
&= \frac{1}{6} \left(\underbrace{[f_x(\{x_3\}) - f_x(\{\emptyset\})]}_{Z=\{x_3\}} + \underbrace{[f_x(\{x_1, x_3\}) - f_x(\{x_1\})]}_{Z=\{x_1, x_3\}} + \underbrace{[f_x(\{x_2, x_3\}) - f_x(\{x_2\})]}_{Z=\{x_2, x_3\}} \right) + \\
&\quad + \underbrace{[f_x(\{x_1, x_2, x_3\}) - f_x(\{x_1, x_2\})]}_{Z=\{x_1, x_2, x_3\}}
\end{aligned}$$

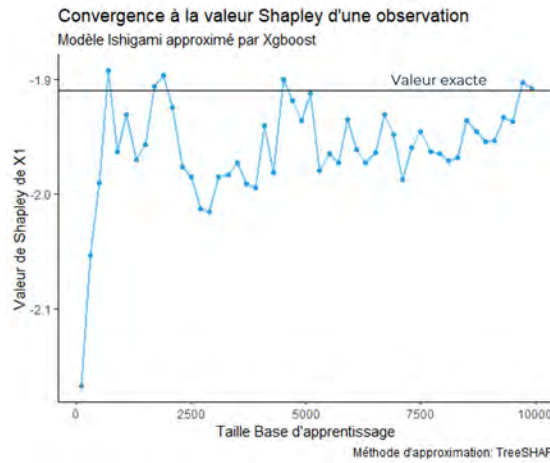


FIGURE 7.10 – Convergence de l’algorithme TreeSHAP en fonction de la taille de l’échantillon : nous représentons ici l’estimation de la valeur de Shapley de la variable X_1 du modèle d’Ishigami. Le métamodèle utilisé est le xgboost.

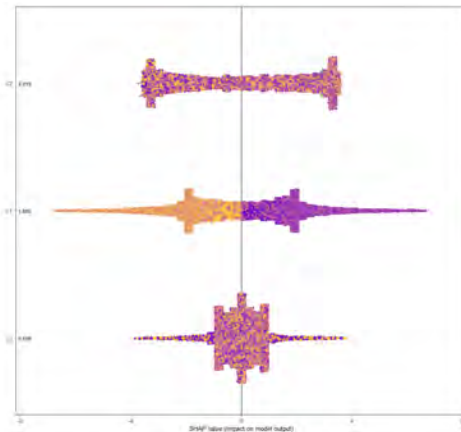


FIGURE 7.11 – Valeurs de Shapley des variables X_1, X_2, X_3 de toutes les observations, estimés par l’algorithme TreeSHAP, passant par un métamodèle xgboost. On remarque que comme pour les indices de Sobol totaux les valeurs de Shapley sont concentrés autour de zéro pour la variable X_3 .

Ces valeurs φ_j sont tels que :

$$f(x) = \varphi_0 + \sum_{j=1}^n \varphi_j = f_x(\{\emptyset\}) + \sum_{j=1}^n \varphi_j(f, x).$$

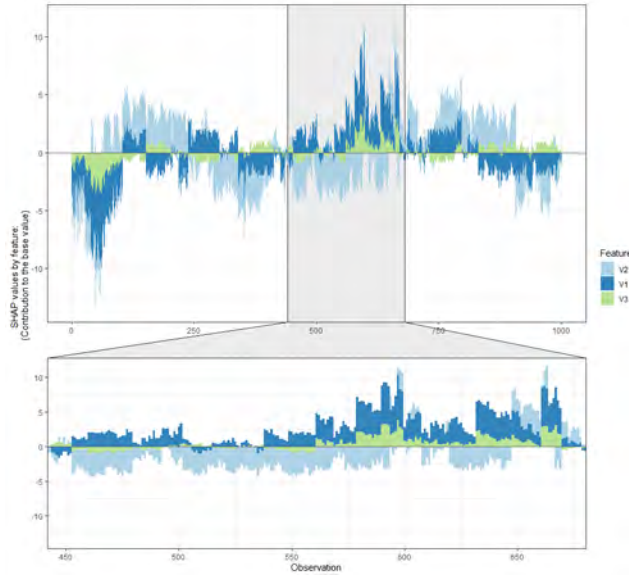


FIGURE 7.12 – Représentation des valeurs de Shapley : les valeurs de Shapley par variable (en ordonné) sont empilées par chaque observation (en abscisse). La somme de ces valeurs par observation mesurent l'écart entre la prédiction de l'observation et la valeur attendue. Les contributions positives amènent la prédiction vers le haut, alors que des valeurs négatives attirent la prédiction en dessous de la moyenne.

Les interactions SHAP du modèle de Ishigami sont :

$$\begin{aligned} \Phi_{1,2}(f, x) &= \sum_{Z' \subseteq x'} \frac{|Z'|!(n-|Z'|-2)!}{2[(n-1)!]} \nabla_{1,2}(Z') = \\ &= \frac{1}{4}(f_x(\{x_1, x_2, x_3\}) - f_x(\{x_1, x_3\}) - f_x(\{x_2, x_3\}) + f_x(\{x_1, x_2\})) \\ \Phi_{1,3}(f, x) &= \sum_{Z' \subseteq x'} \frac{|Z'|!(n-|Z'|-2)!}{2[(n-1)!]} \nabla_{1,3}(Z') = \\ &= \frac{1}{4}(f_x(\{x_1, x_2, x_3\}) - f_x(\{x_1, x_2\}) - f_x(\{x_2, x_3\}) + f_x(\{x_1, x_3\})) \\ \Phi_{2,3}(f, x) &= \sum_{Z' \subseteq x'} \frac{|Z'|!(n-|Z'|-2)!}{2[(n-1)!]} \nabla_{2,3}(Z') = \\ &= \frac{1}{4}(f_x(\{x_1, x_2, x_3\}) - f_x(\{x_1, x_2\}) - f_x(\{x_1, x_3\}) + f_x(\{x_2, x_3\})) \end{aligned}$$

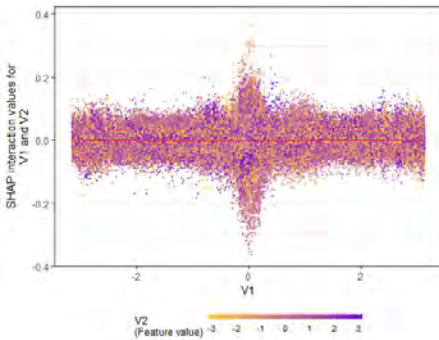


FIGURE 7.13 – Les interactions SHAP entre X_1 et X_2 estimées par l'algorithme TreeSHAP, avec un métamodèle xgboost (taille d'apprentissage de 100 000 observations) sont très faibles et dispersées autour de zéro.

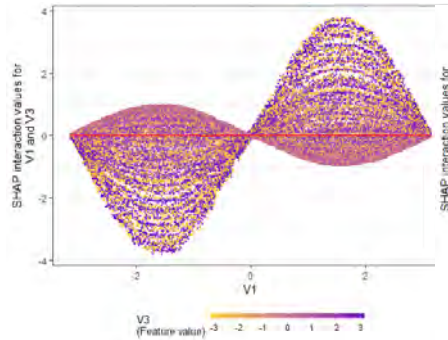


FIGURE 7.14 – Les interactions SHAP entre X_1 et X_3 estimées par l'algorithme TreeSHAP, avec un métamodèle xgboost (taille d'apprentissage de 100 000 observations) sont les plus fortes détectées par l'algorithme. On reconnaît une courbe sinusoïdale. Son amplitude dépend de X_3 : il est autant grand que la valeur absolue de X_3 .

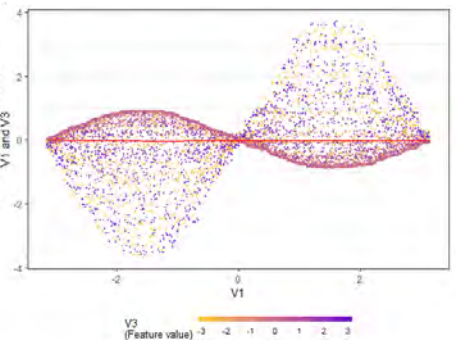


FIGURE 7.15 – Interaction SHAP entre X_1 et X_3 , avec un métamodèle xgboost, à taille d'échantillon d'apprentissage réduite (10 000 observations).

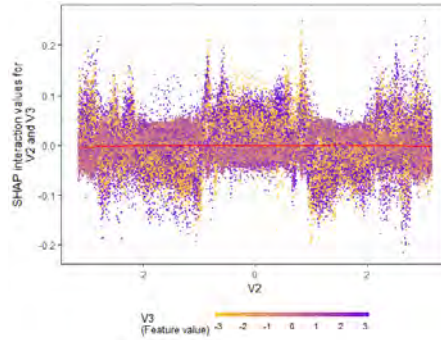


FIGURE 7.16 – Les interactions SHAP entre X_2 et X_3 estimées par l’algorithme TreeSHAP, avec un méta-modèle xgboost (taille d’apprentissage de 100 000 observations) sont faibles et dispersées autour de zéro. On remarque deux tendances : les individus avec une grande valeur absolue de X_3 et $X_1 \in [-2.8, -2.2] \cup [-2, -1] \cup [1, 2] \cup [2.2, 2.8]$ et les individus avec une faible valeur absolue de X_3 et $X_1 \in [-\pi, \pi] \setminus ([-2.8, -2.2] \cup [-2, -1] \cup [1, 2] \cup [2.2, 2.8])$ qui ont les interactions opposées. Il est possible en effet que des interactions artificielles soient introduites.

Conclusion

Les indices de Shapley (Shapley Effects), comme définis dans le chapitre précédent proposent une méthode globale de décomposition de la variance, mais ne permettent pas d’isoler les interactions des variables. Nous pouvons résoudre cette limite avec l’approche local de *SHAP*.

Nous avons proposé plusieurs algorithmes d’approximations de la valeur SHAP, mais nous avons retenu l’estimateur *TreeSHAP*, étant adapté à notre cadre d’étude et incluant le calcul des interactions.

Quatrième partie

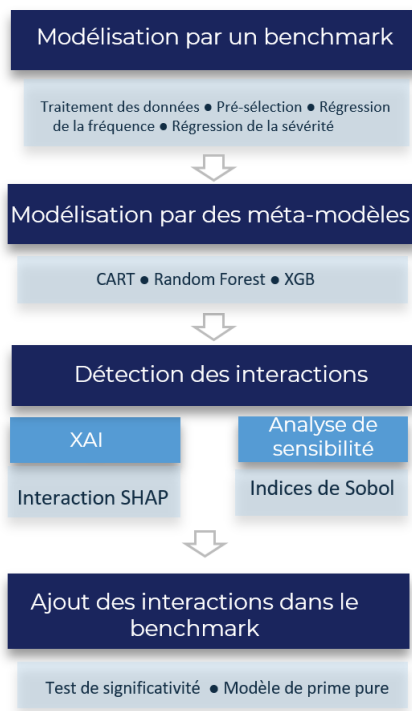
Application à la tarification à l'adresse

Rappel de la méthodologie

La *partie III* fournit les outils techniques pour interpréter les modèles, selon une approche globale ou locale, à partir de l'explication de la prédiction ou de la variance.

Bien que cette méthodologie soit applicable à n'importe quel modèle, des spécificités des données peuvent demander des étapes supplémentaires de retraitement.

Cette partie montrera l'application des techniques de l'analyse de sensibilité et XAI aux données assurantielles de la *tarification à l'adresse*. Vivant dans la grande dimension, les estimateurs de Sobol et SHAP pourraient avoir de problèmes de convergence. Pour pallier cela, une étape de pré-sélection a été effectuée. Les étapes qui résument la méthodologie de détection et d'intégration des interactions dans le modèle de *Benchmark* sont les suivantes :



Chapitre 8

Contexte d'étude : la tarification à l'adresse

Le chapitre suivant présente le contexte d'étude ainsi que la base de données utilisée. Il s'agit de la tarification du produit *Smarthome Pricing* multirisque habitation, où la modélisation de la sinistralité intègre les caractéristiques de l'assuré, du logement et des variables externes (climatiques, économiques, criminalité) au niveau de l'adresse du logement et du bâtiment même.

Le mémoire se situe dans la continuité des travaux menés au sein de l'équipe P&C Pricing and Data du groupe Addactis France par Pierre Chatelain et Victoria Delavaud. Les montants de la sinistralité pour des questions de confidentialité ont été multipliés par une constante.

L'étude a été limitée aux maisons individuelles sur le territoire métropolitain (hors Corse) de la **sinistralité attritionnelle** pour la garantie **Dégâts des eaux**.

Introduction

Comme vu dans le chapitre **2**, aujourd'hui le tarif d'un contrat multirisque habitation se détermine à partir d'environ 19 questions posées dans les devis en ligne ¹ :

- caractéristiques du logement : type de résidence (principale, secondaire, surface habitable, présence d'une piscine, type de chauffage, ...)
- de l'assuré : type d'assuré (propriétaire, locataire, propriétaire, non occupant)
- de l'occupant : âge, profession, ...
- du contrat : date de début, fréquence de paiement,..
- éligibilité des options : présence d'une cave, présence d'une piscine
- données personnelles
- les antécédents de sinistralité : nombre de sinistres dans les 2 dernières années, type de sinistres, ...

¹Pour des risques considérés "assurables" en ligne

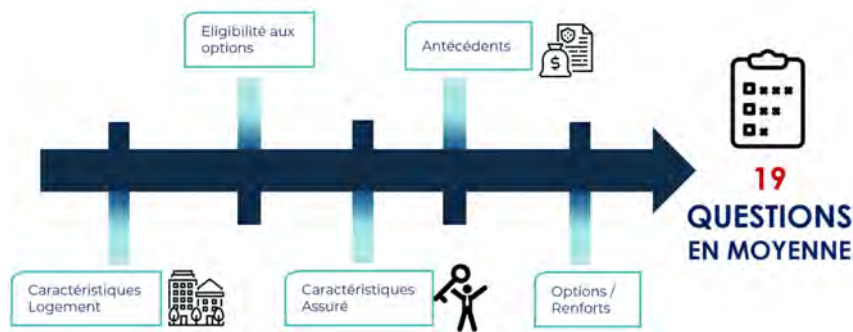


FIGURE 8.1 – Mécanisme simplifié du parcours de souscription d’aujourd’hui en assurance habitation.

À la différence de l’Auto, l’expérience client de la branche Multirisque habitation a peu évolué depuis les dernières décennies : les assureurs ont mis l’accent sur la segmentation et sur la maîtrise actuarielle, sans pourtant modifier leur parcours de souscription.

8.1 Une nouvelle approche au risque

La tarification à l’adresse inclut dans l’étude du risque des informations qui nécessitent la géolocalisation des biens assurés, liées à l’**adresse** et au **bâtiment**.



FIGURE 8.2 – Exemple de variables innovantes introduites par la tarification à l’adresse

Ces données sont fournies pour chaque adresse client par un Data Provider qui les récupère à l’aide de plusieurs techniques :

- Scraping (sources : IGN (Institut National de l’Information Géographique et Forestière), ADEME (Agence de l’environnement et de la maîtrise de l’énergie) , RPLS (Répertoire des logements locatifs des bailleurs sociaux), ...)
- Intelligence artificielle (ex. Computer vision pour l’analyse d’image qui utilise majoritairement des réseaux de neurones à convolution pour des variables comme : le type de toit, la présence de panneau solaire)
- Règle métier
- et par agrégation des trois, lorsque l’information est manquante

Des exemples de données seront présentés par la suite.



FIGURE 8.3 – Parcours de récupération des données intelligentes : tout d’abord les adresses clients sont géolocalisées. Puis, pour chaque adresse on choisit un bâtiment et on détermine différents attributs.

En plus de l’apport de nouvelles caractéristiques, la tarification à l’adresse propose des modèles de prime pure en fonction des objectifs et des besoins de l’assureur, s’il cherche un modèle plus performant ou plus rapide par exemple.

Cette approche remplacerait le modèle dit *basique* où la prime est calculée en fonction des réponses de l’assuré à un questionnaire, auxquelles on ajoute éventuellement des données externes. Nous distinguons 3 modèles de remplacement :

- le modèle dit *"0 questions"* a pour objectif la réduction des temps de souscription. En s’appuyant que sur les données intelligentes, ce modèle de prime permet de substituer les variables des questionnaires ;
- le modèle dit *Complet* ajoute au questionnaire de base les données géolocalisées par le fournisseur afin d’améliorer le modèle existant ;
- enfin le modèle dit *Hybride* se situe au milieu entre les modèles précédents et calcule une prime à partir d’un minimum de questions de base (ex. âge occupant et surface habitable) et les données intelligentes ;



FIGURE 8.4 – Mécanismes de Tarification : Modèle qui utilise les réponses du questionnaire et éventuellement des données externes comme réalisation des variables explicatives.

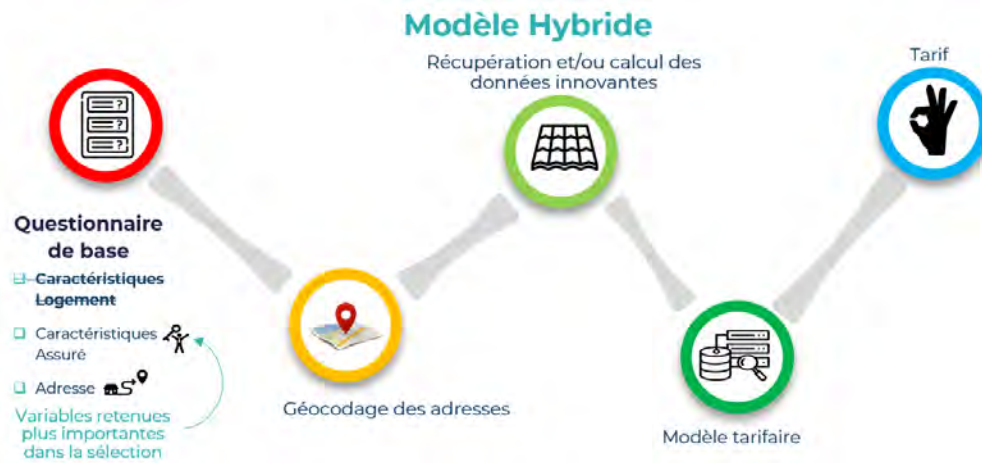


FIGURE 8.5 – Mécanismes de Tarification

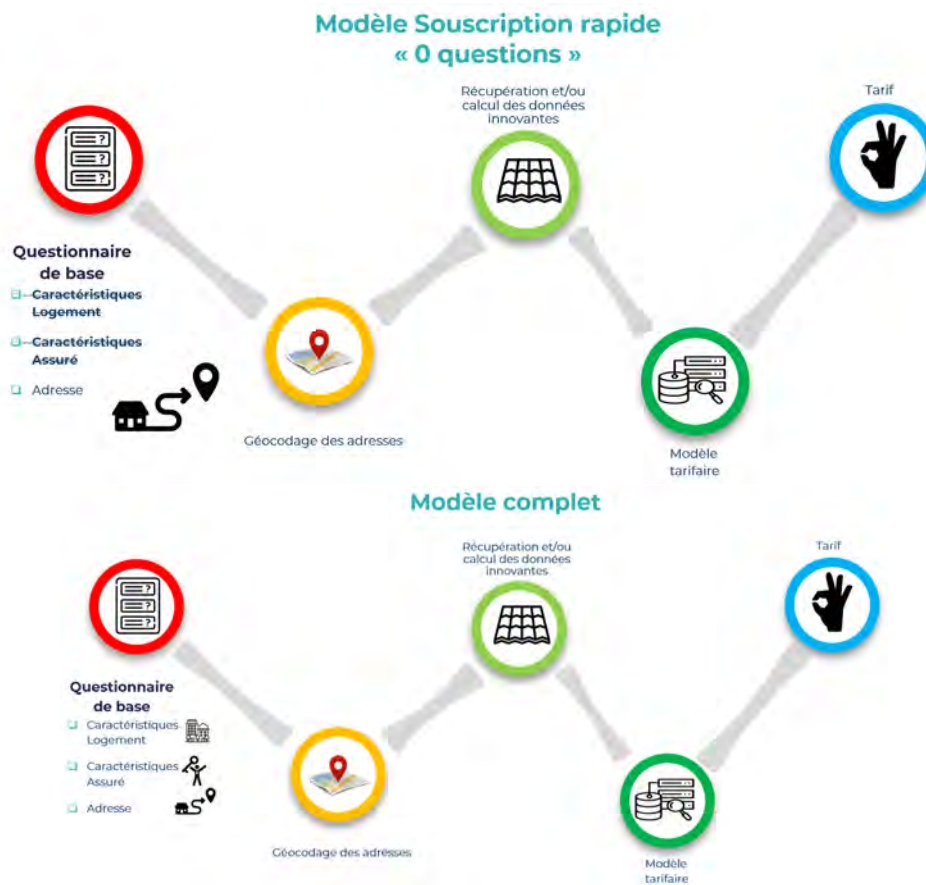


FIGURE 8.6 – Mécanismes de Tarification

Création des nouveaux attributs

Les données sont produites en agrégeant, par ordre de priorité, des Open Data ou données externes, des données issues d'un modèle Machine Learning et des données issues d'une règle métier.

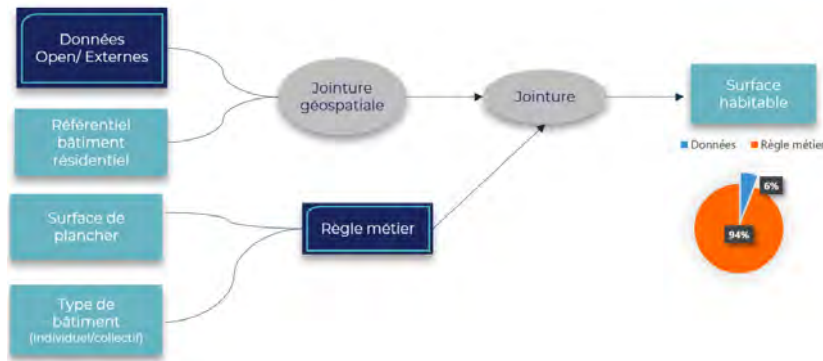


FIGURE 8.7 – Exemple de production de la variable surface habitable : 6% des données fournies provient des base des données sources (ADEME, DVF (valeur Foncière), IGN, ...), 94% provient des données produites par une règle métier.

Source : Nam.R



FIGURE 8.8 – Exemple de production de la variable période de construction : 8% des données fournies provient des base des données sources (IGN, Ministère de la transition écologique et solidaire,...), 92% provient des données produites avec un modèle de machine learning.

Source : Nam.R

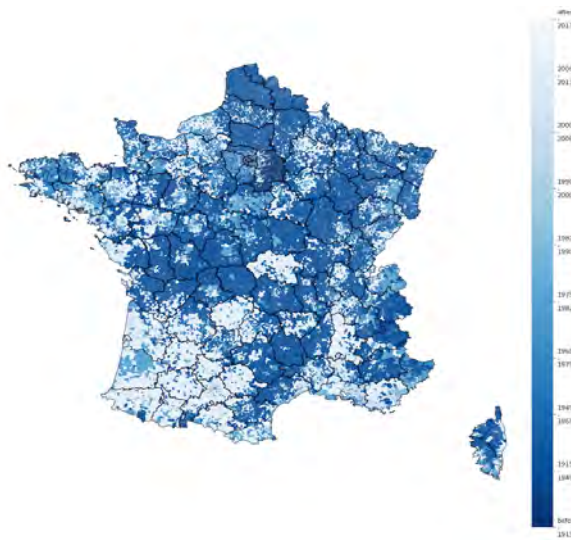


FIGURE 8.9 – Production de la variable période de construction : le modèle de machine learning a été validé avec 85% d'*accuracy*. Pour une meilleure lecture de la carte, les valeurs ont été agrégées à la maille IRIS.

8.2 Description de la base des données

La base utilisée pour l'étude de la sinistralité résulte de la jonction de deux bases :

- la base interne : il s'agit d'une base anonymisée provenant d'un assureur partenaire, contenant à la fois des informations sur la sinistralité et à la fois les caractéristiques de l'assuré ;
- la base externe : il s'agit d'une base provenant d'un *Data Provider* partenaire, qui fournit des caractéristiques climatiques, sociales, ... des adresses géographiques étudiées et dans le cas où la maille à l'adresse soit trop fine, des informations à la maille municipale, IRIS, départementale ou régionale.

La base interne contient les informations sur la sinistralité attritionnelle entre 2016 et 2018 des maisons individuelles, sur le territoire métropolitain (hors Corse).

L'identifiant de la base est le contrat avec son image, c'est-à-dire son état (en cours, avenant, résilié, annulé), et il est caractérisé par une date de début d'image et par une date de fin d'image.

La modélisation de la fréquence dépend de l'exposition au risque ou nombre d'années assurance, dont la définition prend en compte les dates de début et fin d'image. Elle représente la proportion de l'année de la situation du contrat.

Dans le cadre de la modélisation d'une loi de comptage, le nombre des sinistres sera normalisé par l'exposition de la situation du contrat sur une année.

Définition 27. *L'exposition du contrat A, d'image x, pour l'année N est défini :*

$$Expo(\text{Contrat } A; N; \text{Image}_x) := \frac{\min(31/12/N; \text{date fin d'image}) - \max(\text{date debut d'image}; 01/01/N)}{365.25}$$

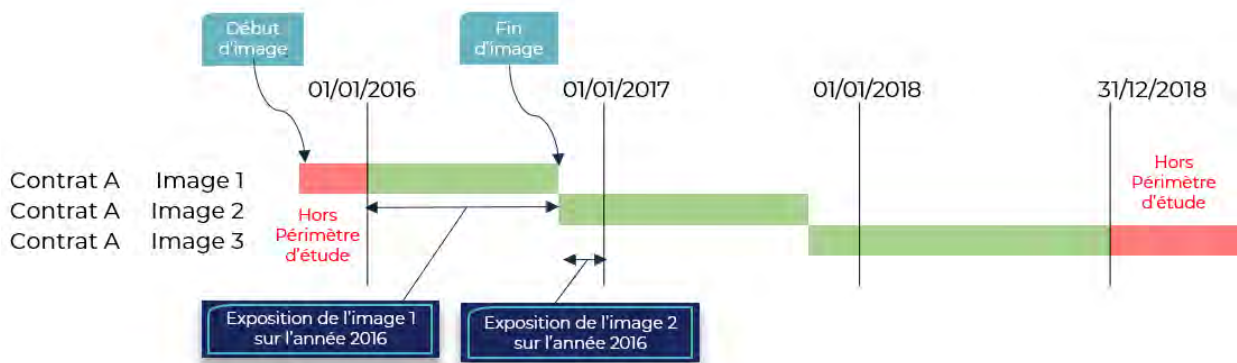


FIGURE 8.10 – Exposition en années pour chaque identifiant (Contrat & Image) de la base interne

Chaque sinistre sera par ailleurs affecté à la ligne du contrat où

la date de survenance \in [Date de début de l'image, Date de fin de l'image]

Afin d'utiliser une base fiable et un nombre de lignes suffisant pour la modélisation de la sinistralité, 8% de la base après jointure a été supprimée :

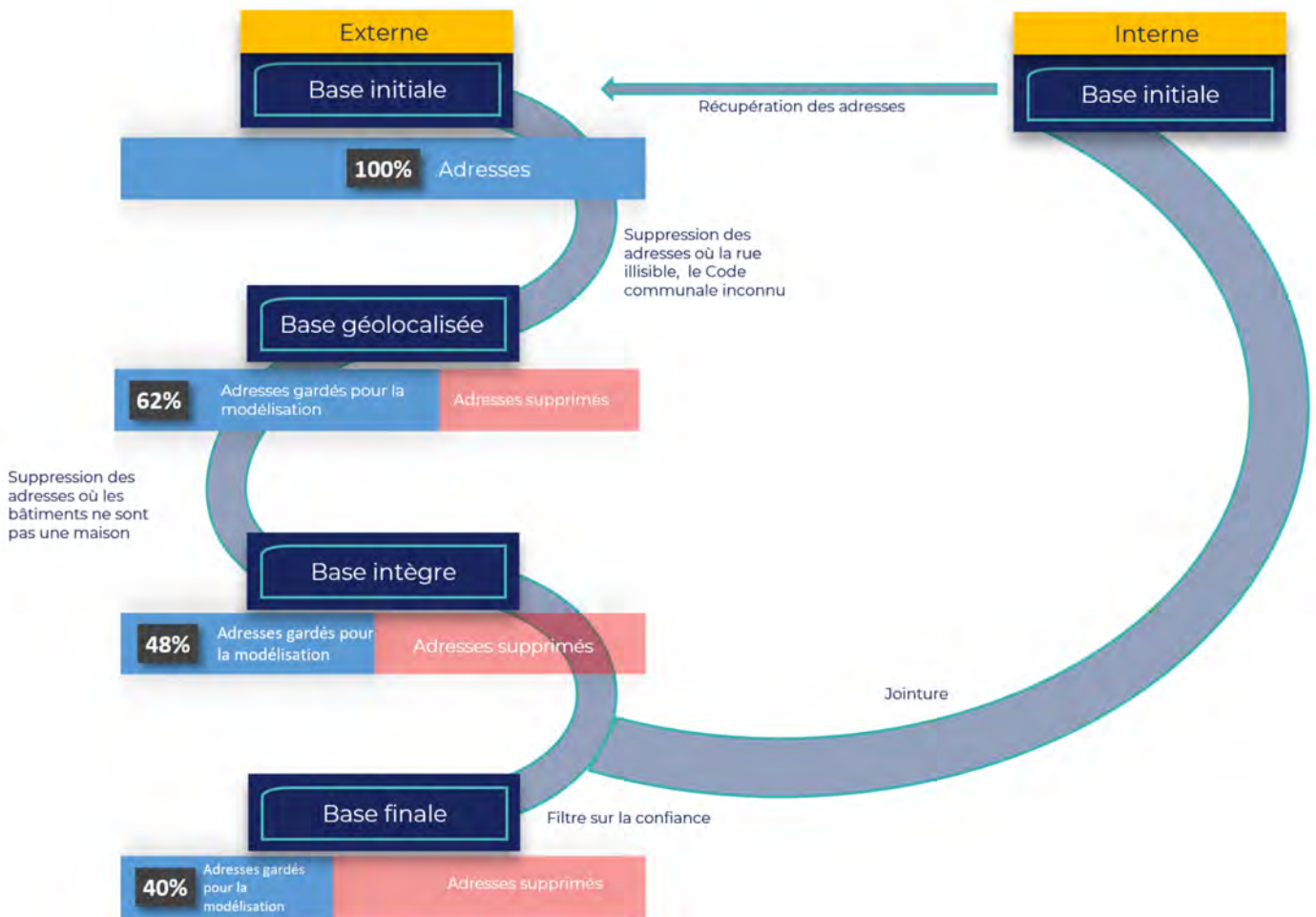


FIGURE 8.11 – Construction de la base d'étude pour la modélisation.

Pour la modélisation de la prime pure nous connaissons les informations suivantes :

| <i>Information</i> | Base | Nom des variables | Maille |
|--------------------|-------------|---|---------------|
| Contrat | Interne | Numéro de contrat, Année de souscription, numéro d'image Contrat, ... | Adresse |
| Assuré | Interne | Age occupant, ancienneté, situation professionnelle, type de personne (Propriétaire, Locataire,...) | Adresse |
| Antécédents | Interne | Exposition en années d'assurance, charge totale, nombre de sinistres, date de survenance | Adresse |
| Options | Interne | Options multimédia, chien, type de garantie , ... | Adresse |
| Bâtiment | Externe | Adresse, Surface de plancher du bâtiment, surface annexes, nombre d'étage, période de construction, type de sol, note DPE, type de toit, présence du parking, présence panneaux solaires, surface mitoyenneté, présence piscine, Emprise au sol en m2, Présence de cave dans le bâtiment, Présence de véranda ou loggia sur la parcelle du bâtiment, Mode de chauffage principal du bâtiment, Consommation énergétique estimée en kWh pour le bâtiment, ... | Bâtiment |
| | Interne | Informations sur l'annexe, Valeur des biens à assurer, nombre de pièces, type de résidence, ... | Bâtiment |
| Adresse | Externe | Code Département, Code INSEE, Code IRIS, Code Municipalité, probabilité d'être un résidence principale, Bâtiment inscrit dans une zone rurale ou urbaine, Altitude du bâtiment en mètres, Distance au bâtiment résidentiel le plus proche, Nombre de bâtiments à moins de 50 et à moins de 100m, ... | Adresse |
| | Interne | Code fractionnement, Code postal du lieu du risque, département, Code Département, Code INSEE, Code IRIS, Code Municipalité | Adresse |

FIGURE 8.12 – Variables utilisées pour la modélisation de la prime pure.

| <i>Information</i> | Base | Nom des variables | Maille |
|--------------------|-------------|--|---------------|
| Climatique | Externe | Nombre de jour de gel, nombre de jours où la température est supérieure à la moyenne saisonnière de 12 °C, de 8 °C, de 25%/50%/65%, nombre de jours de précipitations, Vitesse moyenne du vent, nombre de fois où la vitesse maximale dépasse d'au moins 10 nœuds sa vitesse moyenne (calculée sur 2 minutes), Quantité moyenne de précipitation en mm, Nombre de jours où les précipitations ont été supérieures à 25%/50% de la moyenne saisonnière, Nombre de jours de précipitation intense (50mm en 24h (région plaine) 100 mm en 24h (région montagneuse)), Nombre de jours où il a plu moins de 2mm par jour pendant 15 jours consécutifs, Nombre de jours où la température a été inférieure à 0°C/ -6,6°C/ -11,5°C/ -1°C, Nombre d'éclairs nuage-sol, nombre d'heures moyen d'ensoleillement, Nombre de jours moyen de neige, Nombre de jours de présence de neige au-dessus 2cm/heure, Nombre de jours de présence de neige au-dessus 4cm/heure et 8cm/heure, Coefficient moyen de la marée, Pression moyenne du niveau de la mer (millibars), Nombre de jour de présence de tornade, Nombre total de jours de tremblement, Nombre total de jours de tremblement multiplié à la magnitude de l'échelle de Richter, Nombre de jours où la visibilité a été inférieure aux normes de saison (Ou inférieure à 20km (région plaine) 60km (région montagneuse)),... | Municipalité |
| Criminalité | Externe | Achats et ventes sans factures, Aide à l'entrée, à la circulation et au séjour des Etrangers, Atteintes à la dignité et à la Personnalité, Atteintes à l'environnement, Atteintes aux intérêts, fondamentaux de la Nation, Atteintes sexuelles, Attentats à l'explosif contre des biens privés, Attentats à l'explosif contre des biens publics, Autres coups et blessures volontaires criminels, Autres délits , ... , Autres infractions à la police des étrangers, Autres vols à main armée, Cambriolages de locaux, d'habitations principales, Cambriolages de locaux industriels, commerciaux ou financiers, Cambriolages de résidences secondaires , ... | Département |
| Économique | Externe | Valeur immobilière de la maison, Nombre d'agriculteurs exploitants actifs occupés, Nombre d'artisans, commerçants, chefs d'entreprise actifs occupés, Nombre de cadres, professions intellectuelles supérieures actifs occupés, Nombre de profession intermédiaires actives occupés, Nombre de profession | IRIS |

FIGURE 8.13 – Variables utilisées pour la modélisation de la prime pure.

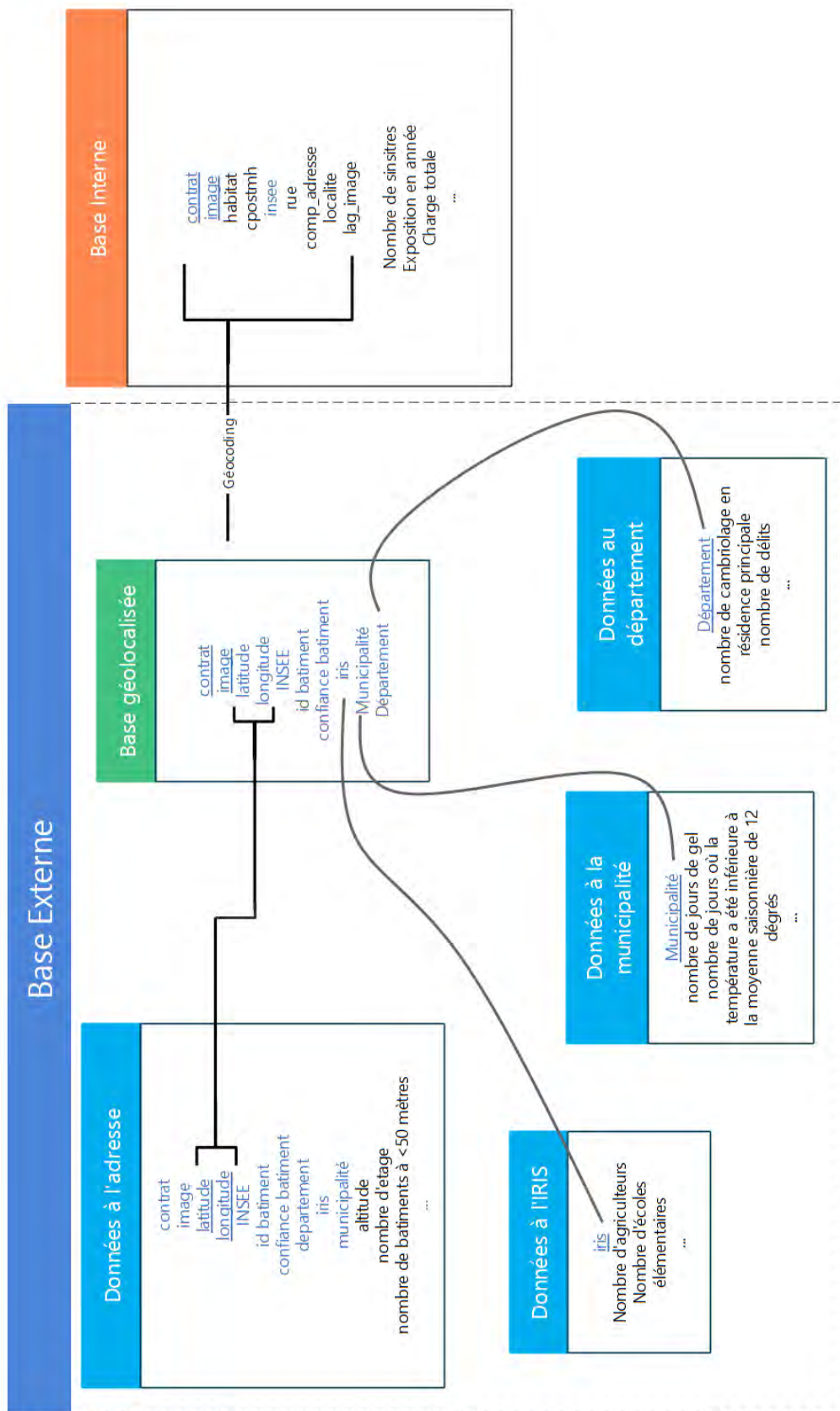


FIGURE 8.14 – Schéma relationnel simplifié pour la jointure des deux bases.

Des détails supplémentaires sur la confiance des données et sur l'analyse descriptive sont fournis dans l'annexe D.

8.3 Impacts de la Tarification à l'adresse

L'utilisation des données intelligentes pour l'assurance habitation a des impacts importants dans l'écosystème assurantiel car :

- la récupération des informations du bâtiment assuré, au lieu du questionnaire déclaratif, permet de diminuer le risque de fausse déclaration et d'avoir une connaissance plus fine du risque ;
- une connaissance plus approfondie du bien endommagé permet de réduire les coûts d'expertise et le temps pour clôturer un sinistre et par conséquent d'indemniser plus rapidement ;
- et l'assureur peut mettre en place des actions de prévention plus adaptées au profil du risque.

Du point de vue commercial, la tarification à l'adresse pourrait par ailleurs promouvoir la tarification multi couverture *Auto + MRH* aux clients seulement équipés de l'automobile en récupérant leur adresse.

8.4 Intérêt par rapport à un zonier

Pour intégrer les facteurs de risque spatial en assurance habitation, on fait recours à une variable appelée *zonier*. L'idée est de découper le territoire en différentes zones de risques selon une maille définie (communément la maille commune). Pour cela les variables géographiques sont écartées de la modélisation de la prime ; le résidu est alors composé du risque géographique. Ce risque est projeté sur la totalité du territoire par un lissage géospatial. Le zonier peut être à la maille département voire commune, mais il n'est que rarement poussé à une maille plus fine. Le modèle de tarification à l'adresse a déjà le niveau de granularité du zonier² et intègre en plus de ça la maille du bâtiment.

²Guillaume Beraud-Sudreau dans son mémoire "Construction d'un zonier en MRH à l'aide d'outils de data-science" utilise l'information spatiale jusqu'à la maille adresse

Chapitre 9

Présélection des variables

Dans ce chapitre, nous développons des modèles de régression de la fréquence et de la sévérité afin de calculer les primes d'assurance qui représentent un profil de risque. Pour cerner la variabilité du risque ce modèles seront ajustés à l'aide d'outils statistiques.

9.1 Traitement des données

La base de la sinistralité et des caractéristiques du portefeuille a été écrêtée en amont, afin d'éviter que les modélisations réalisées ne soient pas trop influencées par des sinistres atypiques. Les sinistres reçus ont été projetés à l'ultime, selon les coefficients propres au client partenaire.

Pour prendre en compte la situation économique de l'année à tarifier, les sinistres ont été mis *as-if*, étape dont on parlera dans la section de la modélisation de la sévérité.

Retraitement des variables

Des retraitements ont été effectués pour corriger des modalités des fautes de conversion d'un format à l'autre. De plus, afin de baisser la volatilité des coefficients de régressions et garantir la stabilité et la robustesse des autres modèles, nous avons regroupé les modalités des variables catégorielles pour qu'elles n'aient pas plus de 8 valeurs uniques. En effet, avec les approches qu'on présentera dans la partie 3, les estimateurs des indices de Sobol et SHAP recodent une variable catégorielle en plusieurs variables *dummy* et souffrent du fléau de la dimension et cela a plusieurs désavantages : du point de vue algorithmique et de la précision des estimateurs.

Nombre de pièces

Après une première analyse de la base jointe et filtrée sur l'indice de confiance, nous avons remarqué que le nombre de pièces, qui est une variable interne, couvre un intervalle assez ample. Pour éviter de considérer dans la base d'étude des observations erronées du nombre de pièces, nous avons ainsi isolé cette variable avec la surface habitable (variable externe) et à l'aide d'une carte thermique nous avons visualisé l'estimation de leur densité jointe.

La méthode d'estimation utilisée est celle par noyau (Parzen-Rosenblatt), une généralisation de la méthode d'estimation par un histogramme, qui est implémentée sous le logiciel R dans le package MASS[3].

Définition 28. *En statistique, l'estimation par noyau est une méthode non-paramétrique d'estimation de la densité de probabilité d'une variable aléatoire. Elle se base sur un échantillon d'une population statistique et permet d'estimer la densité en tout point du support.*

Si $x_1, x_2, \dots, x_N \sim f$ est un échantillon i.i.d. d'un vecteur aléatoire, alors l'estimateur non - paramétrique par la méthode du noyau de la densité jointe est :

$$\hat{f}_h(x) = \frac{1}{Nh} \sum_{i=1}^N K\left(\frac{x - x_i}{h}\right)$$

où

- K est un noyau, une fonction de pondération positive, intégrable, à valeurs réelles, qui satisfait deux conditions :

1. $\int_{-\infty}^{+\infty} K(u)du = 1$ (densité de probabilité)
2. $K(-u) = K(u) \quad \forall u$

- h est une matrice symétrique, définie positive appelé fenêtre, qui régit le degré de lissage de l'estimation.

Pour l'estimation de la densité jointe du nombre des pièces et la surface habitable, le noyau utilisé est celui gaussien :

$$K(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$$

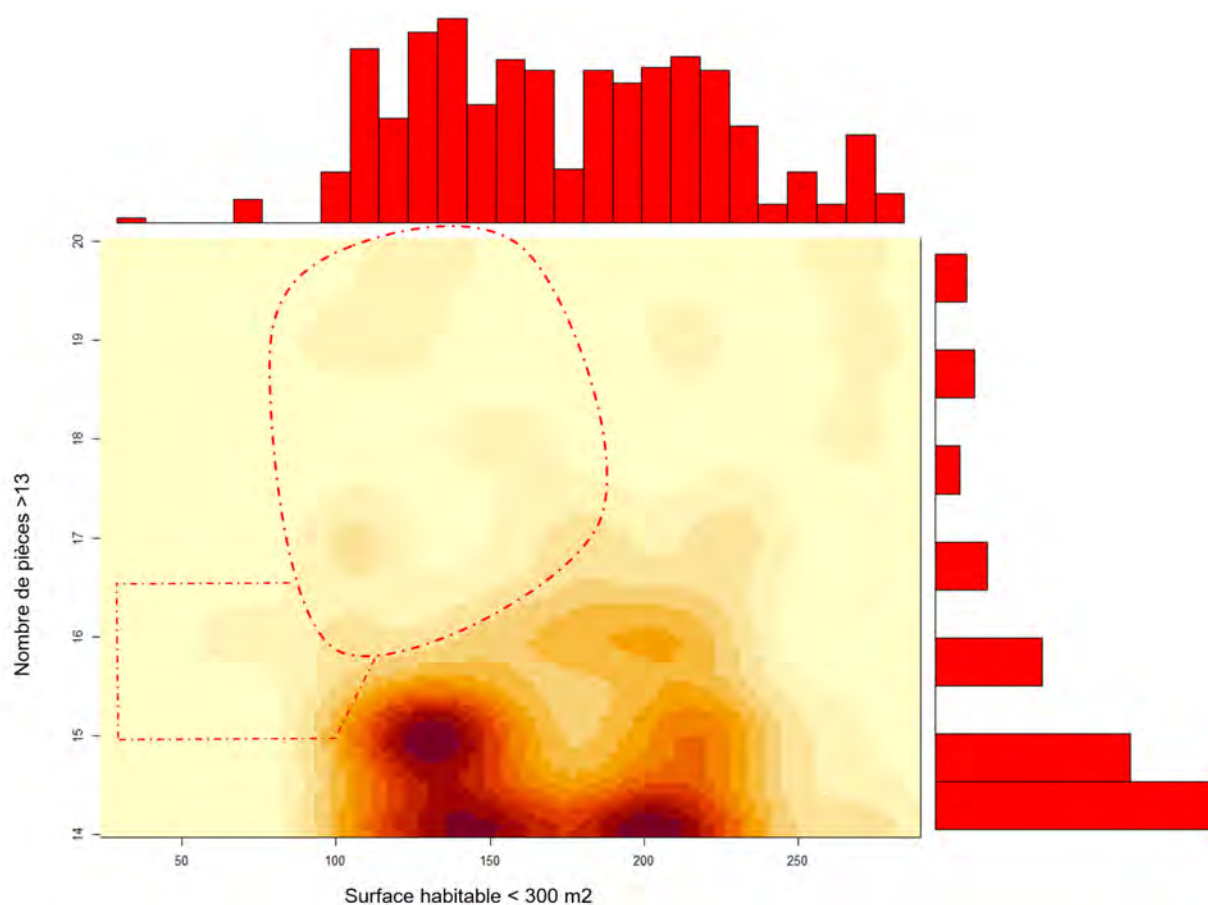


FIGURE 9.1 – Carte thermique de la densité jointe et densités marginales en rouge : la valeur de la densité jointe est d'autant élevée que la couleur des points de la carte thermique tend vers le bordeaux. Les observations ayant une surface inférieure à 100 m² et 15 pièces ont été supprimées, ainsi que les lignes où les pièces sont plus que 16 et la surface habitable inférieure à 200 m².

Nous nous sommes servis de cette carte pour détecter les modalités erronées : ces observations font souvent référence à des immeubles qui contiennent des maisons et des studios. Les modalités de 15 à 20 pièces, représentant une petite portion du portefeuille, ont été regroupées pour ne pas biaiser les résultats de la modélisation. Une répartition supplémentaire a été faite au moment des modèles de régression et validée avec les outils statistiques : l'ampleur de l'intervalle de confiance des coefficients (volatilité), la courbe Actual vs Expected qui compare les valeurs réelles et les estimations et la stabilité dans le temps.

9.2 Pre-sélection des variables

Malgré le fait qu'un grand nombre de variables puisse améliorer un modèle de tarification, en exploitant d'autres informations et des nouveaux liens parmi les variables, au moment de l'estimation de quantités comme la valeur de Shapley ou indice de Sobol, elles subissent inévitablement l'effet de la dimension : le temps d'exécution est $\mathcal{O}(n \times T(f))$, avec n : nombre de variables et $T(f)$: complexité du modèle.

Pour pouvoir exploiter des quantités empiriques "raisonnables" en temps de calcul, nous avons appliqué au préalable trois étapes de sélection de variables :

- afin d'éliminer les attributs redondants, nous avons calculé la matrice de corrélation des variables prises deux à deux. Étant une matrice à grande dimension, pour simplifier la visualisation et la lecture, nous avons fait recours à la visualisation statique de réseaux où nous avons proposé de garder seulement les variables représentatives par groupe des variables "corrélées" ;
- à l'aide de l'ACP, nous supprimons encore des variables corrélées en les projetant sur le plan engendré par les composantes principales ;
- à l'aide d'un Random Forest, car les deux approches précédentes s'appuient sur une mesure de corrélation linéaire parmi les variables et elles ne permettent d'éliminer qu'une partie de la redondance de façon non-supervisée.

9.2.1 Présélection non supervisée des variables

Matrice de corrélation

Afin de réduire le nombre de variables explicatives (plus de 150) et ne conserver que les variables importantes pour l'étude, nous allons effectuer une présélection de variables à l'aide de la matrice de corrélation.

Une matrice de corrélation est utilisée pour évaluer la dépendance entre plusieurs variables en même temps. Le résultat est une table contenant les coefficients de corrélation entre chaque variable et les autres.

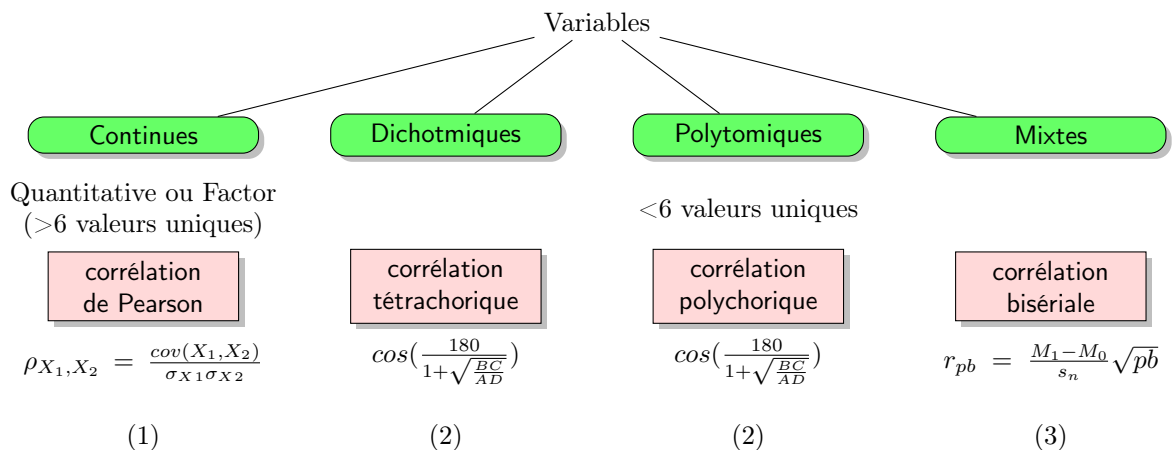
Il existe différentes méthodes de tests de corrélation selon les types de variables et le type de dépendance :

- le test de corrélation de Pearson sur des variables continues
- et des tests applicables par exemple lorsque une ou plusieurs variables en jeu sont réduites à deux ou plus valeurs possibles (la corrélation bisériale et la corrélation tétrachorique, le coefficient de concordance de Kendall, la corrélation point-bisériale, la corrélation "phi").

En statistique, la corrélation polychorique est une technique d'estimation de la corrélation entre deux variables à partir de deux variables ordinales. La corrélation tétrachorique est un cas particulier de la corrélation polychorique applicable lorsque les deux variables observées sont dichotomiques.

Le coefficient de corrélation bisériel ponctuel est un coefficient de corrélation utilisé lorsque au moins une des variables est dichotomique ou artificiellement dichotomisée.

Disposant de données catégorielles et quantitatives, nous avons calculé différentes mesures de dépendance selon le type de variable¹ :



avec en (1)

¹le package 'psych' de William Revelle calcule la matrice de corrélation pour données mixtes, mais au préalable nous avons converti les variables catégorielles en facteurs

| | Variable X Modalité 1 | Variable X Modalité 2 |
|--------------------------|--|---|
| Variable Y Modalité 1 | Nombre d'effectifs de Modalité 1 sur X et Modalité 1 sur Y A | Nombre d'effectifs de Modalité 2 sur X et Modalité 1 sur Y B |
| Variable Y Modalité 2 | Nombre d'effectifs de Modalité 1 sur X et Modalité 2 sur Y C | Nombre d'effectifs de Modalité 2 sur X et Modalité 2 sur Y D |

FIGURE 9.2 – Exemple simplifié de matrice de contingence

- $cov(X_1, X_2)$ covariance entre X_1 et X_2
- $\sigma_{X_1}, \sigma_{X_2}$ écarts types

En **(2)**, cas où les variables sont polytomiques (plusieurs valeurs uniques) ou dichotmiques (deux valeurs : 0 et 1), on fait référence au tableau de contingence, une matrice de dimension $s \times t$, où

- s est le nombre de modalité d'une variable X
- t est le nombre de modalité d'une variables Y
- et les éléments n_{ij} sont le nombre d'effectifs de modalités (i, j) sur (X, Y) .

En **(3)** on suppose que Y est une variable dichotomique et que l'ensemble de données est divisé en deux groupes (le groupe 1 qui a reçu la valeur "1" sur Y et le groupe 2 qui a reçu la valeur "0" sur Y), avec la notation suivante :

- M_1 est la valeur moyenne de la variable continue X pour tous les points de données du groupe 1
- M_0 la valeur moyenne de la variable continue X pour tous les points de données du groupe 2
- p est la proportion de points de données dans le groupe 1,
- b est la proportion de points de données dans le groupe 2
- s_n est l'écart type utilisé lorsque des données sont disponibles pour chaque membre de la population

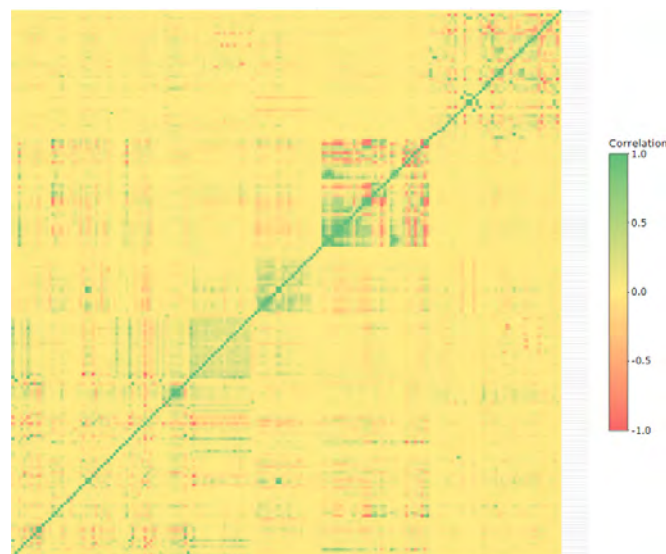


FIGURE 9.3 – Matrice de corrélation de 179 variables de la base Dégats des eaux

La complexité de la lecture d'une matrice de si grande dimension encourage à adopter un outil visuel ; c'est ainsi à ce moment qu'on se sert de la visualisation statique de réseaux².

Tracé de graphes

On peut réinterpréter le problème de visualisation de la matrice de corrélation au sens de la théorie des graphes. D'abord on présente des notions de base de la théorie des graphes.

En théorie des graphes, le tracé de graphes consiste à représenter des graphes dans le plan. Le tracé de graphes est utile à des applications telles que la conception de circuits VLSI, l'analyse de réseaux sociaux, la cartographie, et la bio-informatique.

La définition la plus simple de graphe est la suivante :

Définition 29. *Un graphe simple non orienté est un couple $G = (V, E)$ comprenant :*

- V un ensemble de sommets (ou noeuds ou points)
- $E \subset \{(x, y) | (x, y) \in V^2 \text{ et } x \neq y\}$ un ensemble d'arêtes (ou liens)

Un graphe peut être :

- *orienté* si ses arêtes ne peuvent être parcourues que dans un sens. L'orientation des arêtes est indiquée par des flèches sur les arêtes ;
- *étiqueté* si les liaisons entre les sommets (arêtes ou arcs) sont affectées d'étiquettes (mot, lettre, symbole, etc...)
- *pondéré* si toutes les étiquettes sont des nombres réels positifs ou nuls. Ces nombres sont les poids des liaisons (arêtes ou arcs) entre les sommets.

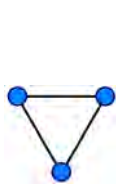


FIGURE 9.4 – Graphe non orienté et non pondéré

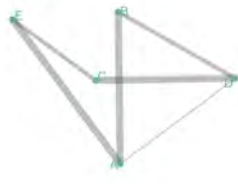


FIGURE 9.5 – Graphe non orienté et pondéré



FIGURE 9.6 – Graphe orienté et non pondéré



FIGURE 9.7 – Graphe orienté et pondéré

Par ailleurs, tout graphe $G = (V, E)$ peut être représenté par une matrice M , dite *matrice d'adjacence* dont chaque terme a_{ij} est égal au nombre d'arêtes orientées (d'arcs) allant du sommet i vers le sommet j :

$$M = (a_{ij})_{i,j=1,\dots,|V|} = \begin{pmatrix} 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}, \quad a_{ij} = \begin{cases} 1 & \text{si } (v_i, v_j) \in E \\ 0 & \text{sinon.} \end{cases}$$

Il est aussi possible d'utiliser une matrice d'adjacence pour implémenter un graphe pondéré : on remplace les 1 par les valeurs liées à chaque arc.

²*ggraph* est une extension de *ggplot2* adaptée aux visualisations graphiques et offre la même approche flexible pour construire des tracés couche par couche.

On peut maintenant définir le cadre d'application de la théorie des graphes à la problématique de visualisation de la matrice de corrélation.

On suppose que la matrice de corrélation est la matrice d'adjacence du graphe étiqueté, non orienté (la matrice est symétrique) et pondéré où les sommets sont les variables d'étude et les liens sont les corrélations importants parmi les variables. Étant intéressés par la suppression des variables à corrélations forte, nous enlevons les arêtes pour lesquelles la corrélation est inférieure à 0.4 en valeur absolue et la diagonale. La matrice d'adjacence est ainsi de la forme :

$$M = \begin{pmatrix} 0 & Cor(X_1, X_2) & & \\ Cor(X_1, X_2) & 0 & & \\ 0 & Cor(X_3, X_2) & \ddots & \\ & \dots & & \end{pmatrix}$$

Il y a plusieurs techniques de dessin d'un graphe :

- *Layout dirigé par les forces* : minimisation par descente de gradient d'une fonction d'énergie, inspiré par une métaphore physique ;
- *spectral layout* : layout basé sur une fonction énergie, qui est fondée sur des techniques d'algèbre linéaire ;
- *orthogonal layout* : layout avec des arcs courant horizontalement et verticalement, avec des approches pour réduire le nombre d'arcs s'entrecoupant ainsi que la superficie ;
- *symmetric layout* : en essayant de trouver les groupes de symétrie dans le graphe ;
- *dessins arborescents* : techniques spécialisées pour le tracé d'arbres ;
- *dessins hiérarchiques* : techniques qui essayent de trouver une source et un puits dans un graphe orienté et d'arranger les nœuds en strates en ayant le plus d'arcs commençant vers la source et suivant la direction du puits.

L'algorithme de dessin basé sur les forces (Force-based ou Force-directed algorithms) est particulièrement adapté à notre problème car il est simple à implémenter et permet de positionner les nœuds de manière optimale, organique et esthétique en utilisant un système de force appliqués entre les nœuds et les arcs. Il est couramment utilisé dans la visualisation de réseaux ou de grands graphes, représentation des connaissances, gestion de système et visualisation de maillage.

L'algorithme peut être décrit comme une analogie physique des composants du graphe :

- Les nœuds sont représentés par des particules dans un plan qui sont chargés électriquement et qui appliquent des forces de répulsion les uns contre les autres, respectant le principe des aimants. Plus les nœuds sont éloignés, moins ils se repoussent.
- Les arêtes relient ces points en simulant une force de ressort, attirant les nœuds adjacents
- Le modèle détermine de manière itérative les forces résultantes qui agissent sur les nœuds et tente de rapprocher les nœuds d'un équilibre où toutes les forces s'additionnent à zéro, et la position des nœuds reste stable. À chaque passe de l'algorithme, on applique la somme des forces sur chacun des nœuds. On déplace ces nœuds jusqu'à trouver un état stable.

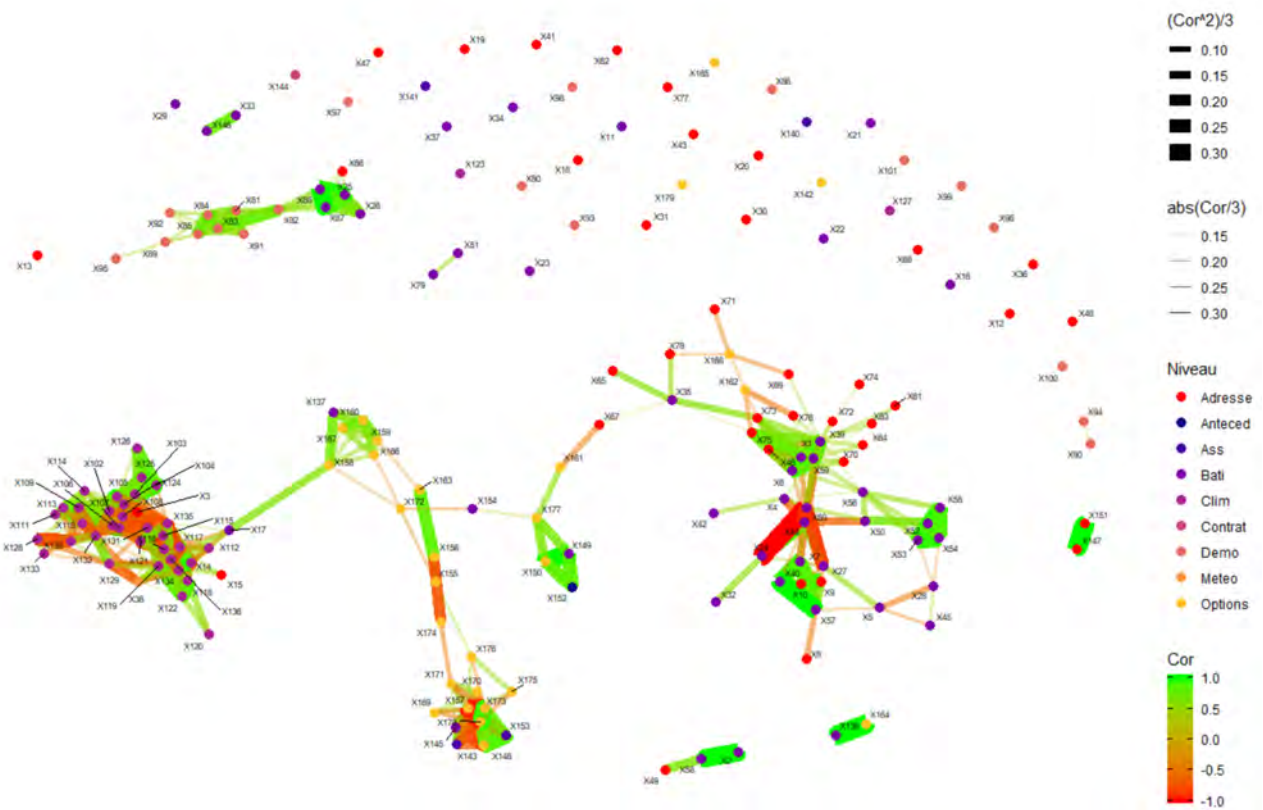


FIGURE 9.8 – Le diagramme de réseau basé sur l'énergie montre une distribution plus équilibrée que la matrice de corrélation, avec peu de croisements d'arêtes. Dans un même groupe de variables corrélées, les variables contenant plus d'informations (ex. surface des annexes contient plus d'informations que la variable dichotomique "bâtiment principal") et les variables "externes" ont été privilégiées dans la sélection.

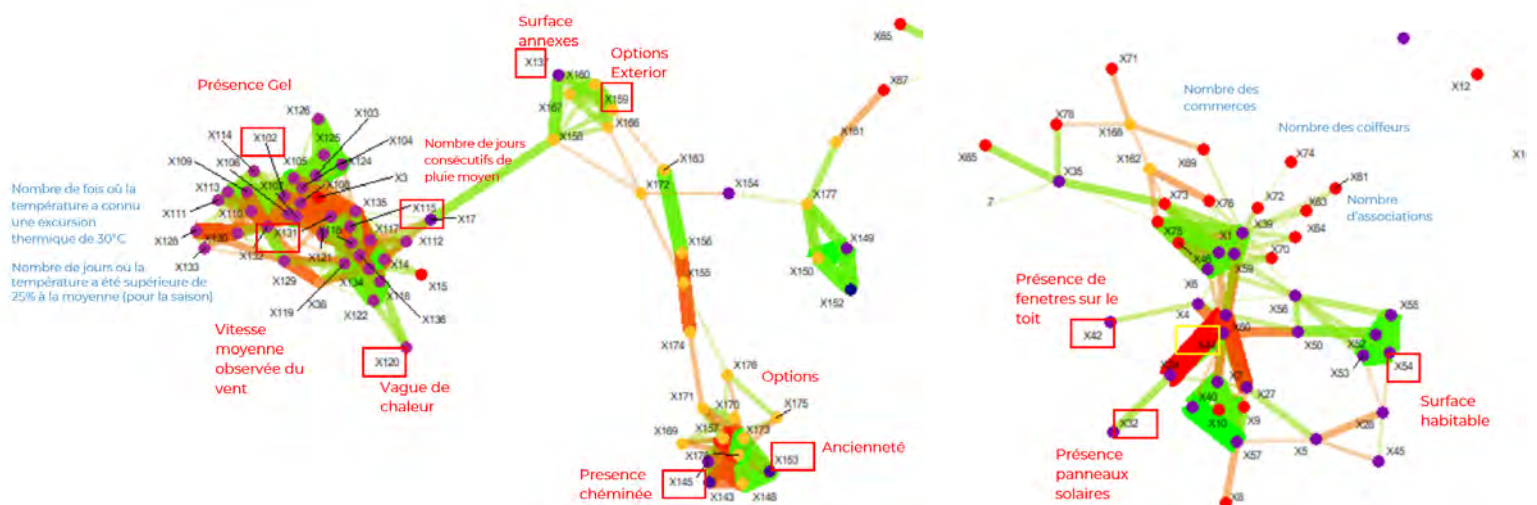


FIGURE 9.9 – Corrélation à l'aide de la théorie des graphes : en rouge les variables retenues : présence de gel, vague de chaleur, vitesse moyenne du vent, surface des annexes, surface habitable, présence panneaux solaires,...

À l'issue de cette procédure, on supprime environ 50 variables où il y a de la redondance d'information (ex. le nombre de jours où l'écart de la température par rapport à la moyenne saisonnière est significatif est capté par la variable présence de gel et l'ensoleillement est corrélé aux variables de pluviométrie).

Méthodes factorielles de représentation et de discrimination : ACP, ACM

L'Analyse en Composantes principales (ACP) et l'Analyse en Composantes multiples (ACM) font parties des méthodes descriptives multidimensionnelles appelées méthodes factorielles.

Apparues dans les années '30, ces méthodes ont été développées en France seulement dans les années '60, grâce à l'apport de Jean-Paul Benzécri qui en a exploité les aspects géométriques et les représentations graphiques.

On distingue généralement deux grandes approches :

- l'approche non-supervisée **descriptive** s'appuie sur un modèle géométrique. L'objectif est de proposer un nouveau système de représentation des variables latentes formées à partir de combinaisons linéaires des variables prédictives, qui permettent de discerner le plus possible les groupes d'individus. En ce sens, elle propose une représentation graphique dans un espace à dimension réduite, engendré par les *composantes principales* qui sont calculées sur les centres de gravité conditionnels des nuages de points avec une métrique particulière.
- l'approche supervisée **décisionnelle**, plus récent, cherche à construire une fonction de classement qui permet de prédire le groupe d'appartenance d'un individu à partir des valeurs prises par les variables prédictives. En ce sens, cette technique se rapproche des techniques supervisées en apprentissage automatique telles que les arbres de décision, les réseaux de neurones. Elle repose sur un cadre probabiliste. Elle est très séduisante dans la pratique car la fonction de classement s'exprime comme une combinaison linéaire des variables prédictives, facile à analyser et à interpréter. En assurance par exemple elle pourrait être exploitée avec la régression logistique, très utilisée dans le scoring, lorsque nous voulons par exemple caractériser l'appétence ou la propension à acheter d'un client face à un nouveau produit.

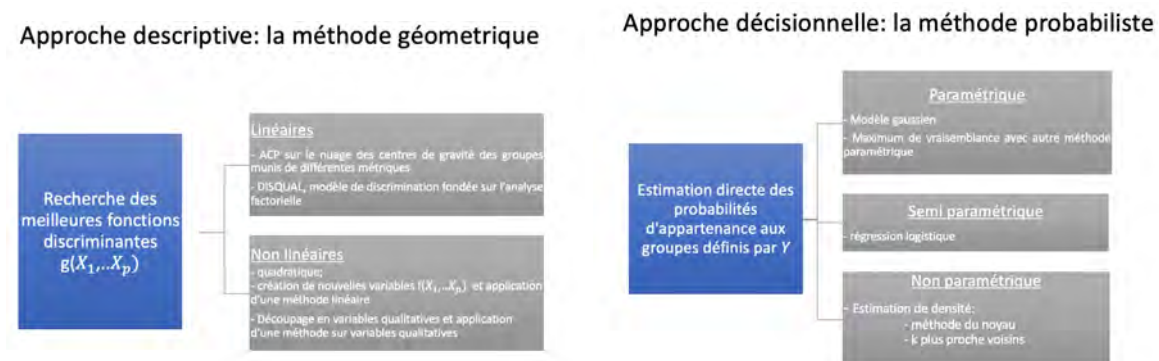


FIGURE 9.10

Nous avons limité l'utilisation de ces techniques au cadre non-supervisé, en complément de la matrice de corrélation avec les buts :

- de **visualiser les données**, en déterminant des éléments structurant les données (les axes de projection). D'un côté, les représentations des individus permettent de voir s'il existe une structure, non connue a priori, sur cet ensemble d'individus. D'un autre côté, les représentations des variables permettent d'étudier les structures de liaisons linéaires sur l'ensemble des variables considérées. Ainsi, on cherchera si l'on peut distinguer des groupes dans l'ensemble des unités en regardant quelles sont les unités qui se ressemblent, celles qui se distinguent des autres. Pour les variables, on cherchera quelles sont celles qui sont très corrélées entre elles, celles qui, au contraire ne sont pas corrélées aux autres.
- **Compression de données** en réduisant la dimension de l'espace de représentation des données (sélection des p premiers axes).
- **Extraction des caractéristiques** plus explicatives et réduction du nombre des variables prédictives pour un autre algorithme (typiquement les modèles explicatifs de fréquence et sévérité qui suivent). L'ACP et l'ACM serviront à avoir une connaissance plus profonde des données sur lesquelles on travaille, à détecter éventuellement des anomalies et éviter la redondance de l'information.

Selon le type de variable, ces techniques se basent sur :

- le tableau de contingence (pour l'étude des variables catégorielles)
- le tableau des données brutes normalisées, comportant les valeurs de q variables quantitatives pour n unités (appelées aussi individus). Ces données peuvent être issues d'une procédure d'échantillonnage ou bien de l'observation d'une population toute entière. Nous l'avons appliqué sur un échantillon représentatif de la base entière (20% de la base).

Les sorties de ces méthodes permettront de répondre aux questions suivantes :

- Quels sont les individus qui se ressemblent ?
- Quelles sont les variables importantes dans mon jeu de données (i.e. quelles sont celles qui apportent une information qui n'est pas redondante) ?
- Comment les variables sont-elles corrélées les unes aux autres ?

Bien qu'elles reposent sur des principes similaires, on appliquera :

- l'analyse des composantes principales (ACP), dans la manipulation des données où les variables sont quantitatives ;
- l'analyse des correspondances multiples (ACM), dans le cas des variables qualitatives, comme généralisation de l'analyse factorielle des correspondances (AFC) pour les cas de plus de 2 variables.

Quelques notions pour comprendre une analyse factorielle

Projection des individus sur la base orthonormale des composantes principales

L'idée à la base de cette technique est de projeter un individu (un contrat dans notre étude) dont on connaît D caractéristiques, noté $x \in \mathbb{R}^D$, sur un sous-espace linéaire de plus faible dimension engendré par une base orthonormale, tout en maximisant la variance des données, c'est-à-dire en captant le plus d'informations sur les données.

On appelle $X = (x_1, \dots, x_N)$ la matrice de dimension $N \times D$ de N contrats.

On commence par trouver un axe $\vec{u}_1 := X_1\alpha_1 + \dots + \alpha_D X_D$, issu d'une combinaison linéaire des variables de départ X_1, \dots, X_D , avec $\sum_{i=1}^D \alpha_i^2 = 1$ tel que la variance du nuage autour de cet axe soit maximale.

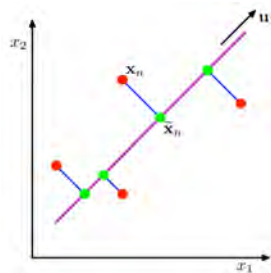


FIGURE 9.11 – En projetant les points en rouge sur l'axe \vec{u}_1 la variance des données est maximale.

On considère la moyenne empirique $\bar{x} = \frac{1}{N} \sum_{n=1}^N x_n$. La variance des points projetés dans la direction \vec{u}_1 est :

$$\frac{1}{N} \sum_{n=1}^N (\vec{u}_1^T x_n - \vec{u}_1^T \bar{x})^2 = \vec{u}_1^T S \vec{u}_1$$

où S est la matrice de covariance empirique.

Maximiser sous la contrainte $\|\vec{u}_1\|^2 = \vec{u}_1^T \vec{u}_1 = 1$ (multiplicateur de Lagrange) donne :

$$\begin{aligned} S \vec{u}_1 &= \lambda_1 \vec{u}_1 \\ \vec{u}_1^T S \vec{u}_1 &= \lambda_1 \end{aligned}$$

λ_1 correspond à la définition de la plus grande valeur propre associée au vecteur propre \vec{u}_1 . La première composante principale est ainsi le vecteur propre de la matrice variance covariance. Ensuite, on peut calculer incrémentalement la deuxième composante principale et ainsi de suite, en appliquant la contrainte d'orthogonalité

de la base : pour un vecteur aléatoire $x \in \mathbb{R}^D$

$$\vec{u}_K = \underset{\substack{a \in \mathbb{R}^D \\ \|a\|=1 \\ \langle a, \vec{u}_1 \rangle = 0, \dots, \langle a, \vec{u}_{K-1} \rangle = 0}}{\operatorname{argmax}} \operatorname{Var}(a^T x)$$

La k-ième composante principale est la combinaison linéaire de X_1, \dots, X_D , qui a la variance maximale de toutes les combinaisons linéaires qui ne sont pas corrélées avec $\vec{u}_1, \dots, \vec{u}_{K-1}$.

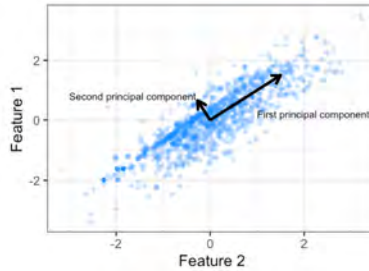


FIGURE 9.12 – Exemple : projection sur le plan engendré par les premières deux composantes principales de deux variables qui ont une corrélation de 0,56.

Cette idée repose en grande partie sur la notion d'*inertie* du nuage de points, qui est la généralisation du concept de variance dans le cas multivarié.

Inertie du nuage des points, Inertie portée par une CP

L'inertie est la quantité d'information contenue dans un tableau de données : si elle est nulle alors tous les individus sont presque identiques.

Considérant la représentation des individus dans l'espace engendré par les variables de départ X_1, \dots, X_D :

- L'inertie d'un nuage de points est définie comme la somme des distances au carré de chacun des points du nuage à leur centre :

$$I = \frac{1}{N} \sum_{i=1}^N (\bar{x} - x_i)^2.$$

Elle correspond à l'étalement du nuage de points.

- On peut montrer³ que la notion d'inertie est liée à celle de variance : $I = \sum_{i=1}^D \operatorname{Var}(X_i) = \operatorname{Tr}(\Sigma)$, avec $\operatorname{Var}(X_j) = \frac{1}{N} \sum_{i=1}^N (\bar{x}_j - x_{ij})^2$ et Σ matrice Variance-Covariance des variables X_1, \dots, X_D .
- Lorsque les variables sont centrées et réduites, l'inertie totale correspond au nombre de variables (colonnes) : $I_{red} = D$.

Considérant la représentation des individus dans l'espace engendré par les composantes principales $\vec{u}_1, \dots, \vec{u}_{D^*}$, où $D^* \leq D$ est le nombre de composantes principales qu'on retient :

- l'inertie de la k-ième CP est :

$$I_k = \operatorname{Var}(\vec{u}_k) = \lambda_k$$

Elle est aussi appelée variance expliquée par le k-ième axe principal.

- L'inertie du nuage des points réécrits dans la base des composantes principales est la somme des inerties des CPs :

$$I = \sum_{i=1}^D \lambda_i$$

$${}^3I = \frac{1}{N} \sum_{i=1}^N (\bar{x} - x_i)^2 = \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^D \left(\begin{array}{c} \frac{1}{N} \sum_{i=1}^N x_{i1} \\ \frac{1}{N} \sum_{i=1}^N x_{i2} \\ \dots \\ \frac{1}{N} \sum_{i=1}^N x_{iD} \end{array} \right)_j - x_{ij})^2 = \sum_{j=1}^D \frac{1}{N} \sum_{i=1}^N \left(\begin{array}{c} \frac{1}{N} \sum_{i=1}^N x_{i1} \\ \frac{1}{N} \sum_{i=1}^N x_{i2} \\ \dots \\ \frac{1}{N} \sum_{i=1}^N x_{iD} \end{array} \right)_j - x_{ij})^2 = \sum_{j=1}^D \operatorname{Var}(X_j)$$

Le tableau de Burt

Lorsque les variables sont qualitatives, on préfère appliquer l'analyse des composantes multiples (ACM) qui se base sur le tableau de Burt (par opposition à la matrice de variance-covariance de l'ACP).

Définition 30. La matrice de Burt B est une matrice symétrique carrée d'ordre p (avec p : nombre totale de modalités), composée de s^2 blocs (avec s : nombre des variables), où l'élément est B_{ij} est la fréquences des individus de modalité i et j .

La présentation de Burt comporte tous les tableaux de contingence des variables prises deux à deux qui sont autant de sous-matrices.

L'obtention de la matrice B du tableau de Burt s'obtient en multipliant la matrice du disjonctif complet par sa transposée.

Le tableau de Burt est un moyen de disposer des informations d'ordre qualitatif afin de les traiter par le calcul, tout comme le tableau disjonctif complet dont il est d'ailleurs issu.

Ce tableau est utilisé dans le cadre de l'analyse des correspondances multiples lorsque l'on cherche des liens entre plus de deux variables aléatoires, c'est-à-dire dans une situation où les unités statistiques se répartissent dans un espace à plus de deux dimensions.

| | Variable | Nombre de valeurs |
|--|--------------|-------------------|
| Dans notre cadre d'étude, la répartition des variables est équitable : | Numérique | 74 |
| | Catégorielle | 70 |

Application

L'hétérogénéité des variables pose des problèmes : comment considérer une combinaison linéaire de variables hétérogènes (la surface habitable a échelle et unité de mesure différentes de celles de la température moyenne par exemple) à la sensibilité des résultats de l'ACP et de l'ACM ?

Pour échapper à tous ce problème, on normalise les variables, c'est-à-dire qu'on travaille sur des variables centrées et réduites : chaque profil colonne x_i est donc remplacé par $(x_n - \bar{x}_n)/\sigma(x_n)$.

L'ACP sera conduite sur la matrice de corrélation. Dans cette procédure les variables ont été normalisées. Pour pouvoir prendre en compte les variables catégorielles dans l'ACP, nous avons transformé ces variables en variables *Dummy*, c'est-à-dire des variables indicatrices de la modalité.

Par la suite on considérera toujours la matrice de corrélation (plutôt que celle de covariance). On remarque que pour les finalités de l'analyse il n'est pas pénalisant d'utiliser une ou l'autre.

Valeurs aberrantes

Avant de retenir un repère orthonormé, il faut vérifier la qualité de la représentation des points, autrement dit il faut choisir des observations proches à la réalité, qui ne sont pas aberrantes : en effet, s'il y a des points aberrants, ils contribueraient sensiblement à l'inertie portée par l'axe et ils engendraient un angle dont le carré du cosinus est très petit.

Nous avons ainsi supprimé dans un premier temps les individus à faible cosinus et à forte contribution :

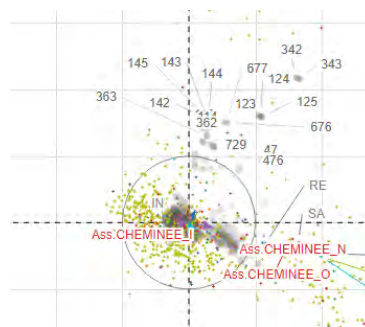


FIGURE 9.13 – Nous avons appliqué une première analyse factorielle sans tenir compte des individus extrêmes : nous les avons ainsi projetés sur le nouveau repère orthonormé. Les individus et les variables très bien représentés se trouvent sur la circonférence. Nous voulons à présent supprimer les valeurs aberrantes, qui ont une grande contribution (fixée ici à plus de 2) et faibles cosinus.

Choix du nombre des axes à retenir

Une fois que les valeurs aberrantes ont été supprimées, la première étape d'application d'une analyse factorielle est déterminer le nombre d'axes principales à retenir pour la projection sur le sous-espace engendré par la base orthonormale.

Il y a plusieurs critères pour choisir D^* le nombre d'axes :

- la variance expliquée : on retient les premiers axes qui expliquent au moins 85% de la variance totale ;
- le critère de Kaiser normée : on ne retiendra que les axes associés à des valeurs propre supérieures à 1 ;
- le critère du coude associé au *scree graph* : sur l'éboulis des valeurs propres, on observe un décrochement (coude) suivi d'une décroissance régulière. On sélectionne les axes avant le décrochement, donc les deux premières.

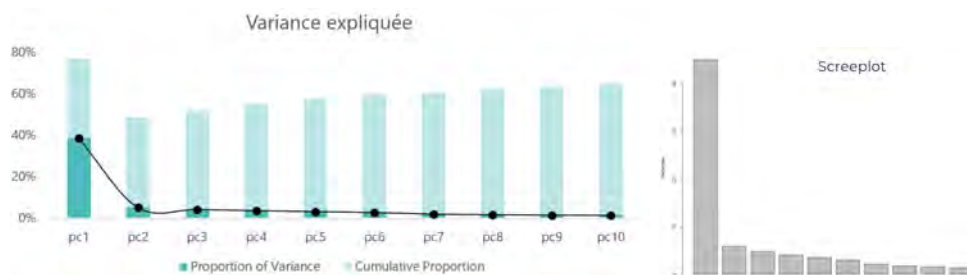


FIGURE 9.14 – ACP : Histogramme de la partie de variance totale expliquée par les premières 10 composantes principales ; dans la série cumulative la j -ème barre permet de visualiser la partie de variance expliquée jusqu'à la j -ième composante principale.

Jusqu'à présent, nous avons déterminé uniquement le plan factoriel où projeter les variables et les individus. Dans les sections suivantes nous chercherons à nous servir de cette nouvelle représentation pour regrouper des variables ayant un comportement vraisemblable au sens des composantes principales, ce qui nous permettra de retenir un nombre réduit de variables.

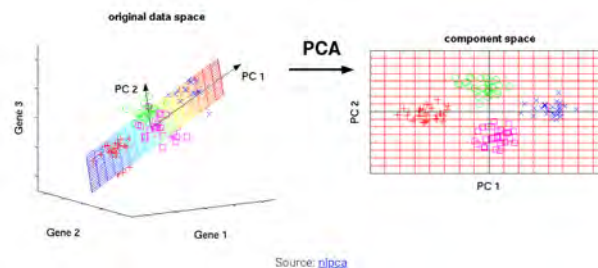


FIGURE 9.15 – Mécanisme de projection sur la base constituée des composantes principales

Projection des variables sur le plan factoriel

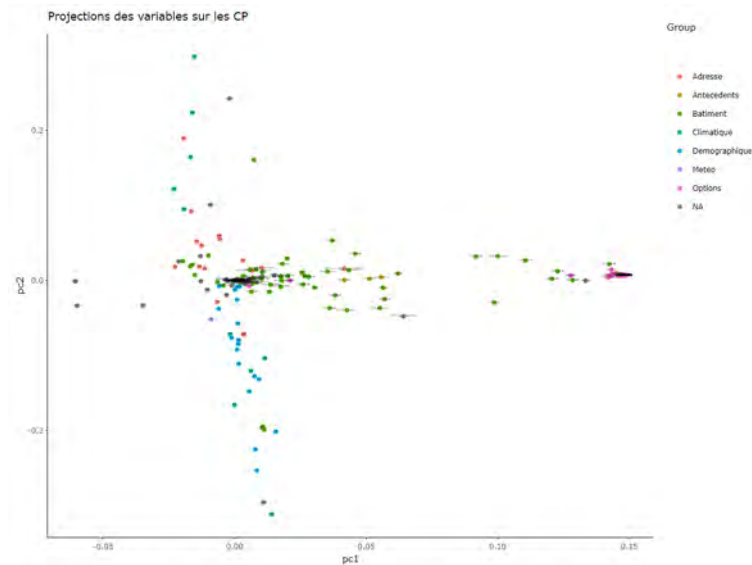


FIGURE 9.16 – La projections des variables sur le plan factoriel montre que le deuxième axe principal est porté par les variables climatiques, de l'adresse et démographiques, alors que le premier axe est porté par les variables relatives au bâtiment.

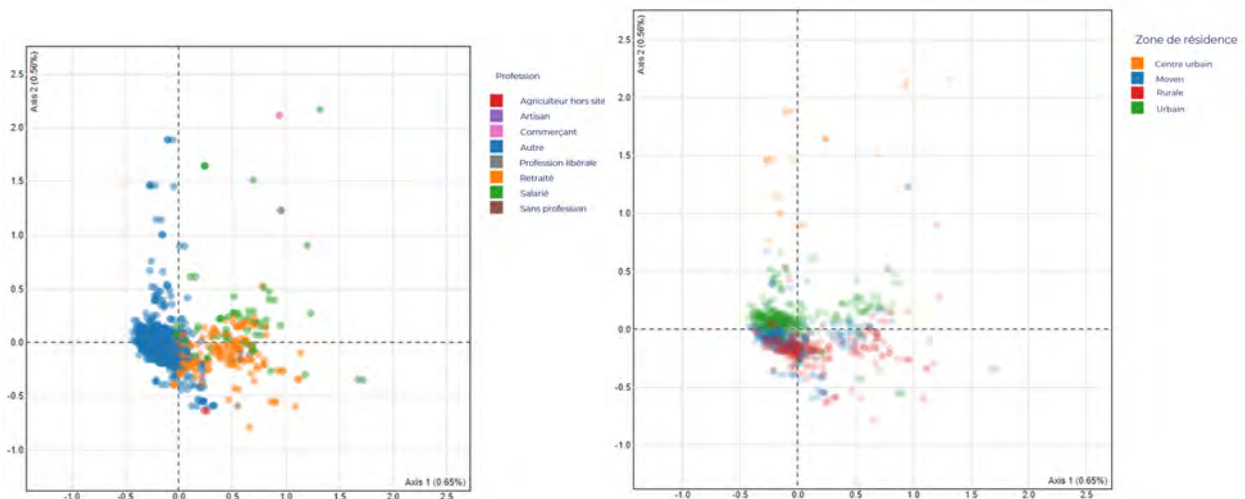


FIGURE 9.17 – ACM : nous avons projeté les individus sur le plan factoriel et coloré selon les variables zone de résidence et status professionnel : nous remarquons que les modalités "rurale " et "retraité" sont prises souvent ensemble. Les habitants du centre urbain ont quant à eux une situation professionnelle différente de ces proposées.

En étudiant les associations parmi les observations on peut définir des profils types, mais cela n'est pas l'objectif de notre analyse. Après ce descriptif visuel, nous allons réduire le nombre de variables. Comme dans les graphes des réseaux, le mécanisme de sélection est de retenir les variables représentatives d'un groupe. En particulier, nous projetons les variables sur les 3 composantes principales et virerons les variables ayant des coordonnées proches.

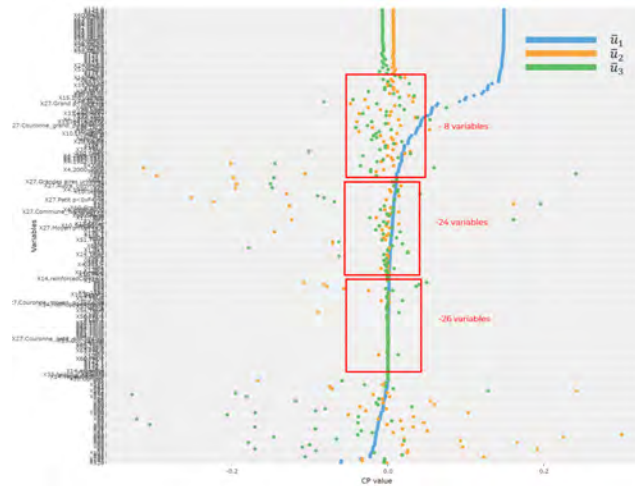


FIGURE 9.18 – Projection des variables sur une dimensions (avec l’ACP). Étant l’espace normé on peut facilement superposer les coordonnées des trois composantes principales par variable : nous avons identifié trois groupes de variables où les coordonnées ne sont pas trop dispersées et nous avons retenu les plus importantes.

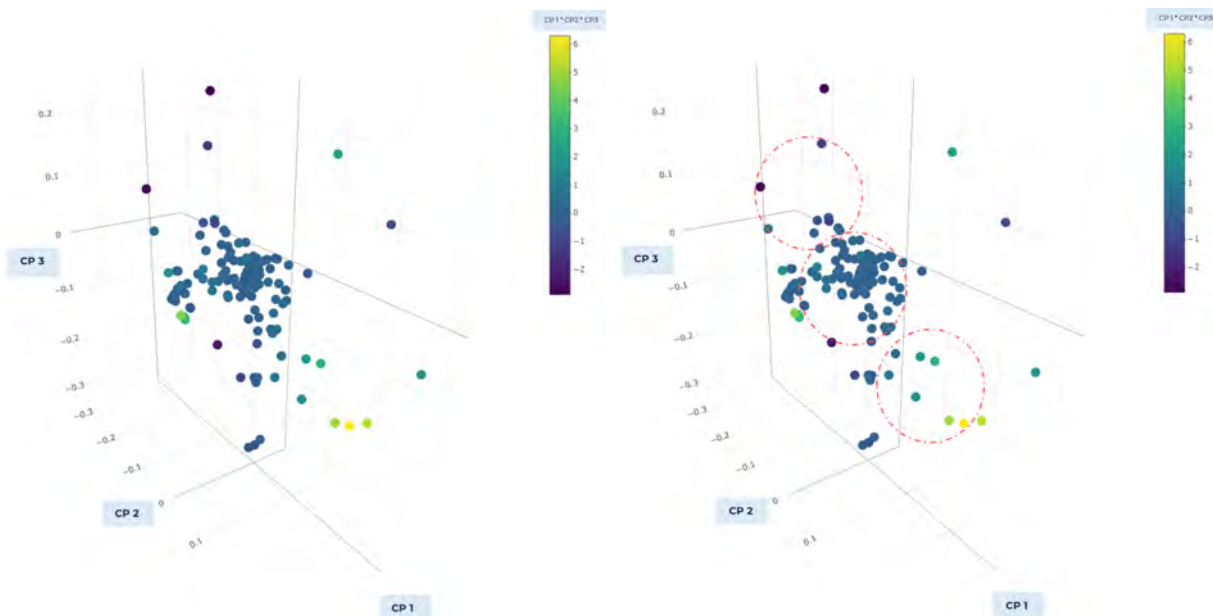


FIGURE 9.19 – Projection des variables sur un espace engendré par les trois composantes principales (avec l’ACP) : à gauche avant regroupement, à droite nous indiquons les trois groupes comme dans la figure 9.18.

9.2.2 Présélection supervisée des variables

Pourquoi autant d’étapes de présélection ?

Un des objectifs de la tarification à l’adresse est d’apporter des informations supplémentaires sur plusieurs axes : météorologiques, démographiques, bâti,... Si on gardait dans la modélisation toutes les variables, le risque est que deux (ou plusieurs) variables corrélées et significatives dans la modélisation saturer le nombre des variables retenues, en apportant que de l’information redondante.

Nous avons simulé un cas où on garde l’information *Age* deux fois, par mois et par années : les deux variables ont une corrélation de 0.9661595 et apparaissent dans la top 10 des variables importantes, donc on saurait naïvement amener à les garder, alors que le type d’information est le même :



FIGURE 9.20 – Exemple d’inclusion d’information redondante : supposons de vouloir modéliser la sévérité d’un sinistre avec un modèle de type Random Forest ⁴ ; nous chercherons à retenir qu’un nombre limité des variables pour garantir la robustesse des résultats. Si on suivait l’ordre d’importance suggéré par cette méthode sans présélectionner les variables, on utiliserait la même information deux fois, en renonçant à la possibilité d’inclure d’autres informations.

Random Forest non optimisé

La sélection des variables est complétée par un Random Forest : comme présenté dans le paragraphe précédent, pré-sélectionner des variables à l’aide des corrélations permet de ne pas retenir par exemple deux variables très corrélées dans la liste des premières 30 variables le plus importantes. Cette présélection est pilotée par la charge totale, par opposition à la sélection non-supervisé des corrélations et de l’analyse factorielle, selon un modèle de forêt aléatoire : l’optimisation des paramètres n’est pas effectuée à ce stade car on ne cherche pas la "meilleure modélisation", mais les variables les plus importantes.

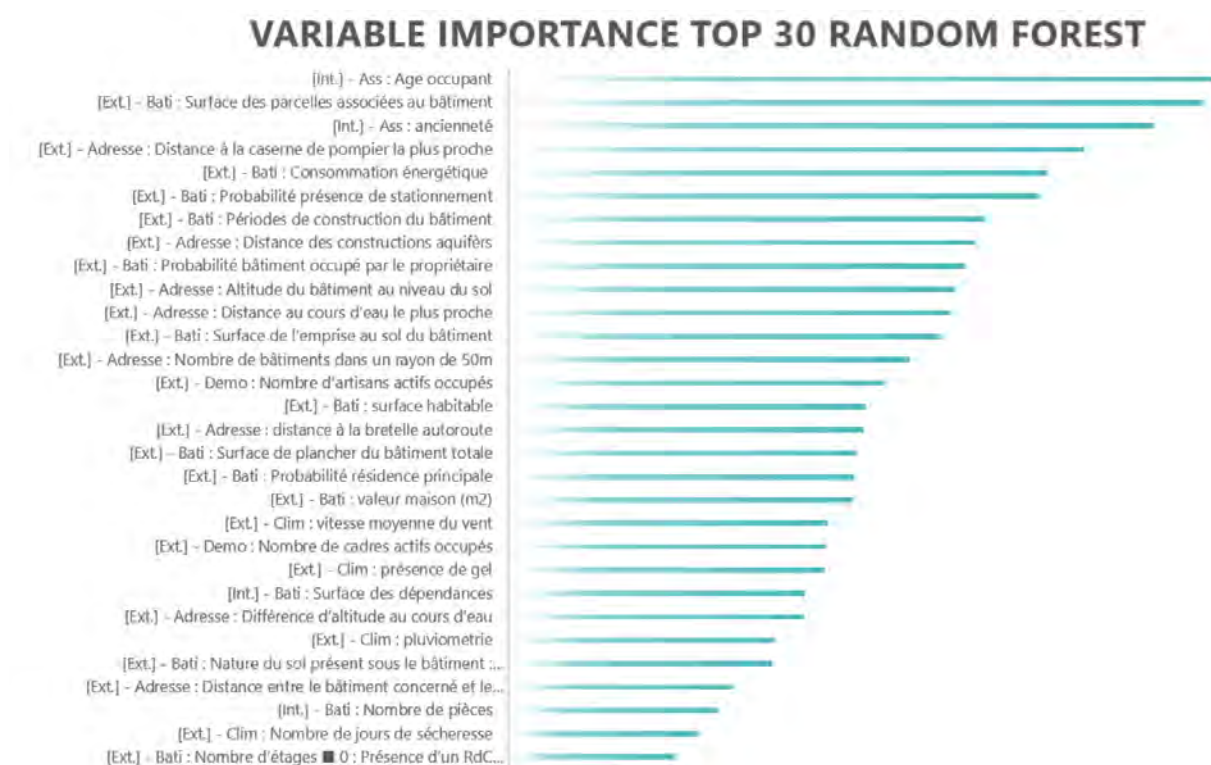


FIGURE 9.21 – Premières 30 variables par importance selon Random Forest : les données confidentielles ont été anonymisées et regroupées par type de variable.

En conclusion de l’étape de Présélection, nous avons retenu environ 40 variables : certaines variables internes sont des questions classiques dans le processus de souscriptions (ex. formule, âge occupant), donc nous les avons réintégrées pour la calibration du modèle *hybride*.

Chapitre 10

Modélisation de la sinistralité

Dans ce chapitre la fréquence et la sévérité des sinistres seront modélisées indépendamment, à l'aide des modèles de régression traditionnels en tarification Non-Vie, tel que le GLM, et de modèles plus sophistiqués de Machine Learning.

Commençons par diviser le jeu de données en plusieurs échantillons : les modèles seront construits dans l'échantillon d'apprentissage et calibrés dans l'échantillon de validation. L'échantillon de test a été utilisé pour comparer les modèles.

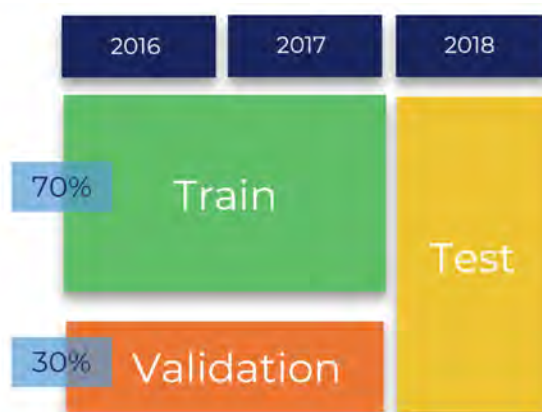


FIGURE 10.1 – Séparation en bases d'apprentissage, validation et test. Pour avoir des résultats robustes dans le temps, les données historiques ont été "mises as if".

L'objectif final étant l'amélioration d'un modèle GLM à l'aide des interactions, nous allons introduire quatre types de modèle :

1. le modèle linéaire généralisé, qui sera le modèle base des comparaison (*Modèle Benchmark*) ;
2. le modèle CART, qui est le première modèle de type machine learning ;
3. le modèle Random Forest (RF) d'agrégation des arbres, en tant que modèle "black box" ;
4. le modèle extreme gradient boosting (XGB), en tant que deuxième modèle "black box".

A l'exception du premier modèle, les autres ne reposant pas nécessairement sur une structure linéaire, peuvent inclure dans leur tâches d'apprentissage des interactions parmi les variables. Ils seront utilisés dans la partie d'analyse de sensibilité comme métamodèles afin de détecter les interactions.

10.1 Modèle de fréquence

Nous allons présenter dans cette section quatre modèles de fréquence développés sous les logiciels R et Python : GLM du package `stats`, CART su package `rpart`, random forest des packages `randomForest` et `sklearn` et extreme gradient boosting du package `xgboost`.

| Coefficients: | | | | |
|--|------------|------------|---------|--------------|
| | Estimate | Std. Error | z value | Pr(> z) |
| (Intercept) | -4.098e+00 | 6.552e-01 | -6.256 | 3.96e-10 *** |
| Nam.roof_materialglass | 3.901e-01 | 5.940e-01 | 0.657 | 0.511417 |
| Nam.roof_materialhotMoppedAsphalt | 2.258e-01 | 4.801e-01 | 0.470 | 0.638069 |
| Nam.roof_materialmetal | -2.565e-01 | 4.534e-01 | -0.566 | 0.571543 |
| Nam.roof_materialreinforcedConcrete | 8.453e-03 | 4.792e-01 | 0.018 | 0.985925 |
| Nam.roof_materialslate | -3.040e-01 | 4.530e-01 | -0.671 | 0.502242 |
| Nam.roof_materialtile | -3.022e-01 | 4.529e-01 | -0.667 | 0.504532 |
| Nam.roof_materialvegetatedRoof | 6.042e-01 | 6.916e-01 | 0.874 | 0.382329 |
| Nam.air_duct_presenceTRUE | 4.101e-02 | 1.635e-02 | 2.508 | 0.012132 * |
| Nam.hot_over12d | -4.265e-01 | 6.699e-02 | -6.366 | 1.94e-10 *** |
| Nam.urban_area_zoneCouronne_grand_pôle | 1.707e-01 | 3.063e-02 | 5.573 | 2.50e-08 *** |

FIGURE 10.2 – Sélection des variables avec le test de Wald : on ne garde que les variables significative au risque de 5%.

GLM

Les lois de fréquence les plus utilisées en assurance sont la loi de Poisson et la binomiale négative. Si les données sont sur-dispersées ($\mathbb{E}(N) < \text{Var}(N)$) la binomiale négative est préférée. En assurance dommages, comme Lundberg avait décrit dans sa thèse, la loi de Poisson est particulièrement adaptée à la modélisation de la fréquence.

Dans notre cas, l'espérance est 0.01363264 et la variance est 0.0133257 et nous avons ainsi retenu le modèle poissonien, avec une fonction de lien logarithmique pour avoir un modèle multiplicatif et éviter de construire un modèle qui renvoie des valeurs négatives.

Avant de tester la significativité des variables et l'ajustement du modèle nous appliquons une dernière étape de sélection des variables, à l'aide d'une méthode de pénalisation de type Lasso (voir annexe E).

Nous avons enfin supprimé la variable *distance de l'insertion sur l'autoroute*.

Significativité des variables

À l'aide du test de Wald, nous avons supprimé les variables : matériel du toit, surface annexe, nombre de jours orageux, type de toit, nombre d'étages, type de sol, surface habitable, surface plancher, surface parcelle, distance d'un ouvrage hydraulique, différence d'altitude par rapport à une autoroute.

Remarque : limites de la méthode Stepwise

Une technique pour sélectionner des variables dans le contexte de modèles linéaires généralisés aurait pu être la procédure *stepwise*, qui appartient aux procédures de sélection de variables "pas à pas".

Elle consiste à considérer un modèle de départ faisant intervenir un certain nombre de variables explicatives, puis élimine (méthode descendante) ou ajoute successivement des variables (méthode ascendante). Pour choisir le modèle meilleur à chaque étape, elle se sert des critères R^2 ou AIC.

C'est une technique aussi naturelle que controversée ; d'un côté, en grande dimension la mémoire vive des logiciels peut être contraignante pour ce type de sélection, car il faut calculer toutes les régressions possibles impliquant un sous-ensemble des k variables explicatives, et d'un autre côté elle peut donner des valeurs R^2 fortement biaisées pour être élevées. D'autres arguments qui défavorisent la technique *stepwise* sont détaillés par Frank Harrell¹.

Dans un premier temps, nous n'avons pas appliqué la procédure Stepwise, puisque le nombre de variables est élevé, et on utilisera l'AIC pour évaluer l'ajustement du modèle.

Qualité du modèle

Regroupements et stabilité dans le temps

Nous avons effectué des retraitements supplémentaires pour améliorer la qualité du modèle, au sens de l'AIC ou de la valeur prédite.

Par ailleurs, pour que le modèle ait un bon pouvoir prédictif, il faut que les variables soient stables dans le temps. Pour les variables où certaines modalités ou valeurs n'ont pas assez d'exposition pour que les coefficients soient robustes, nous avons créé des regroupements.

¹dans l'article : What are some of the problems with stepwise regression ?

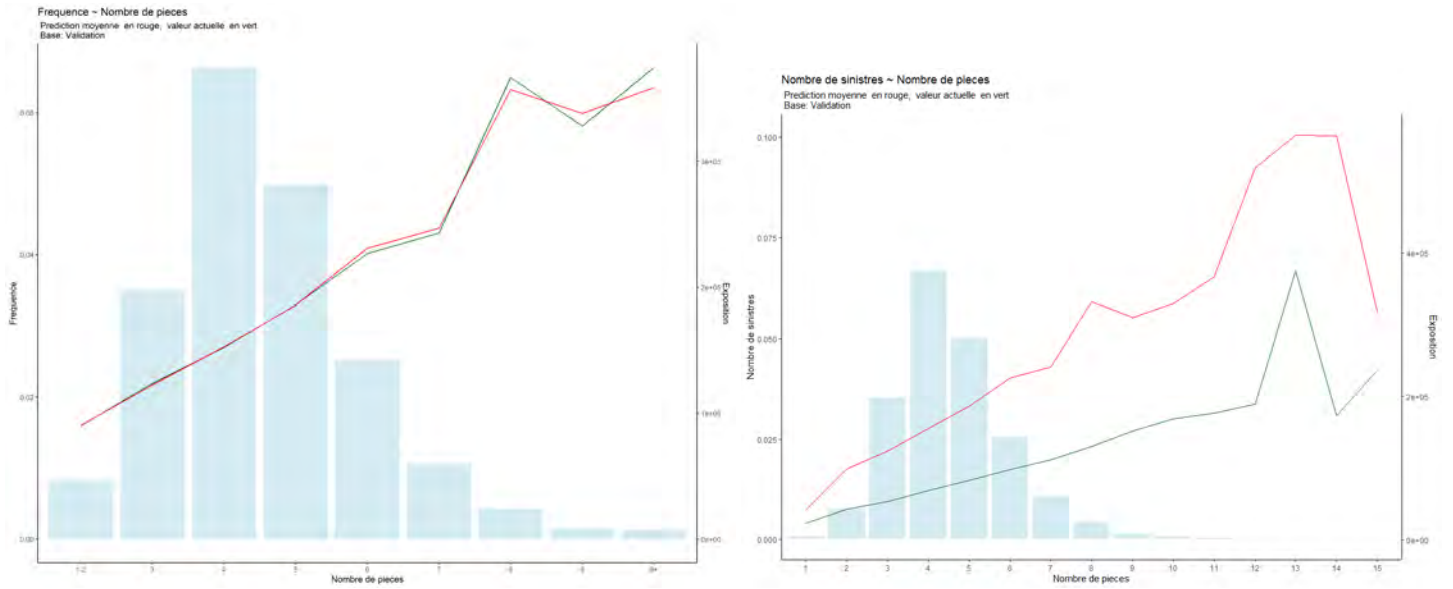


FIGURE 10.3 – Regrouper les modalités réduit la volatilité des coefficients : à droite, avec toute les catégories, les prédictions sont très éloignées de la valeur actuelle, alors que à gauche, le regroupement "1-2" et "9+" réduit l'écart entre les valeurs observées et les prédictions.

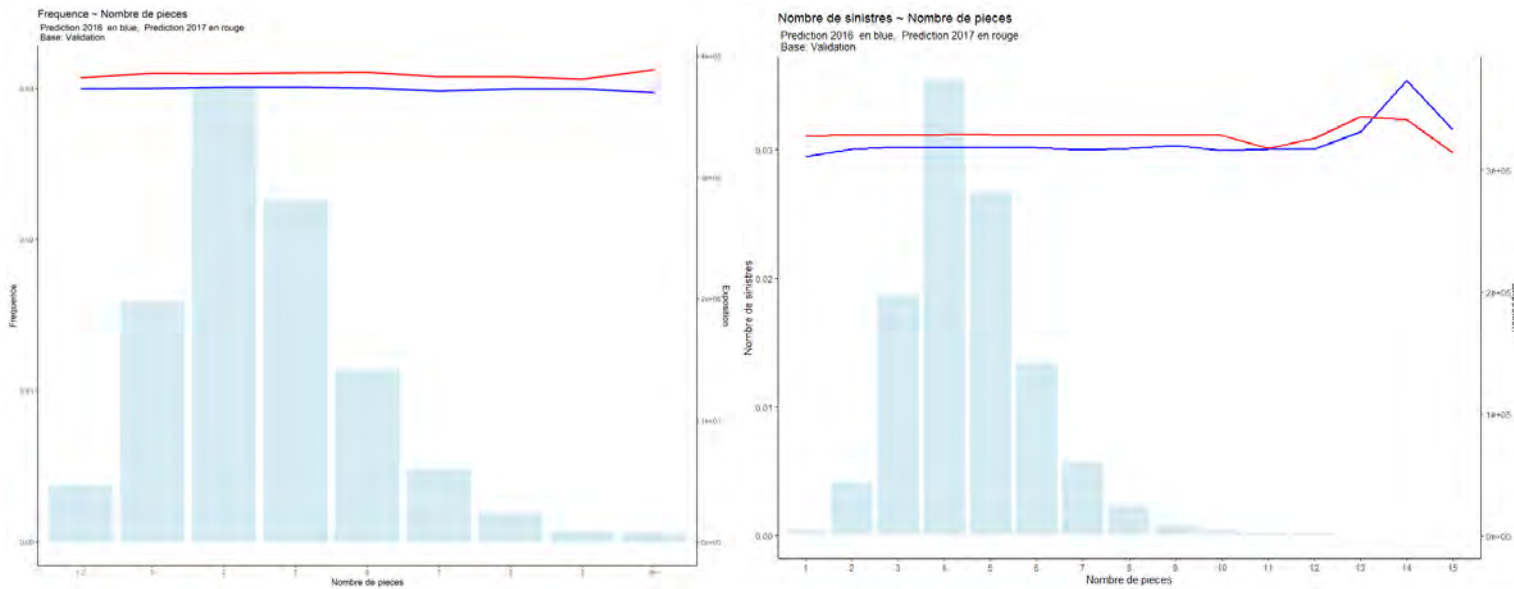


FIGURE 10.4 – Nous remarquons que le regroupement utilisé rend la variable matériel du toit plus stable dans le temps.

Résidus

Pour évaluer les risques potentiels et s'assurer de la robustesse du modèle, il faut garantir le caractère aléatoire des résidus du modèle.

L'une des méthodes consiste à analyser l'adéquation de la distribution des résidus à une loi normale en utilisation QQ-PLOT.

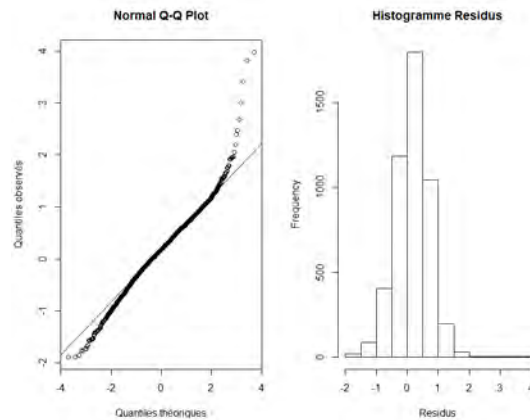


FIGURE 10.5 – Les résidus sont centrés en zéro et les quantiles observés sont proche de ceux d'une loi normale. Nous pouvons conclure que l'ajustement de modèle est pertinent.

Les trois modèles qui suivent sont de modèles de *Machine Learning*.

Du fait de la non-linéarité de leur structure, ces modèles captent les interactions parmi les variables. En contrepartie, ces interactions ne sont pas lisibles à cause de l'effet "boîte noire".

CART

L'arbre de régression sur la fréquence est le deuxième modèle que nous allons construire. Pour éviter le sur-apprentissage nous calibrons le paramètre de complexité de l'arbre avec l'erreur de validation croisée et puis avec la technique de pénalisation d'élagage de l'arbre :

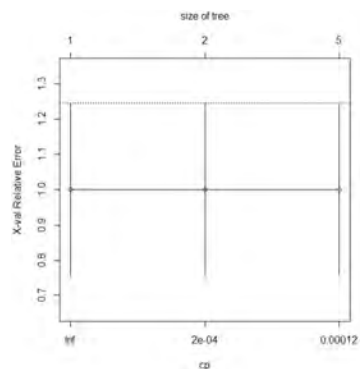


FIGURE 10.6 – Le paramètre de complexité choisi est celui qui minimise l'erreur de validation croisée : 0.0001538328.

L'arbre après élagage est facile à interpréter et on remarque que la valeur de la maison, l'âge, le nombre de pièces et la période de construction sont les variables utilisées par l'algorithme *greedy* pour créer une partition de la base d'apprentissage. Les feuilles filles de l'arbre indiquent

- le nombre d'observations dans la feuille fille : pour la feuille à droite $n = 616$;
- la fréquence moyenne pour les observations de la feuille : 0.69 pour la feuille à droite ;
- la portion de la base initiale présente dans la feuille : celle de droite représente moins de 1% de la base d'apprentissage.

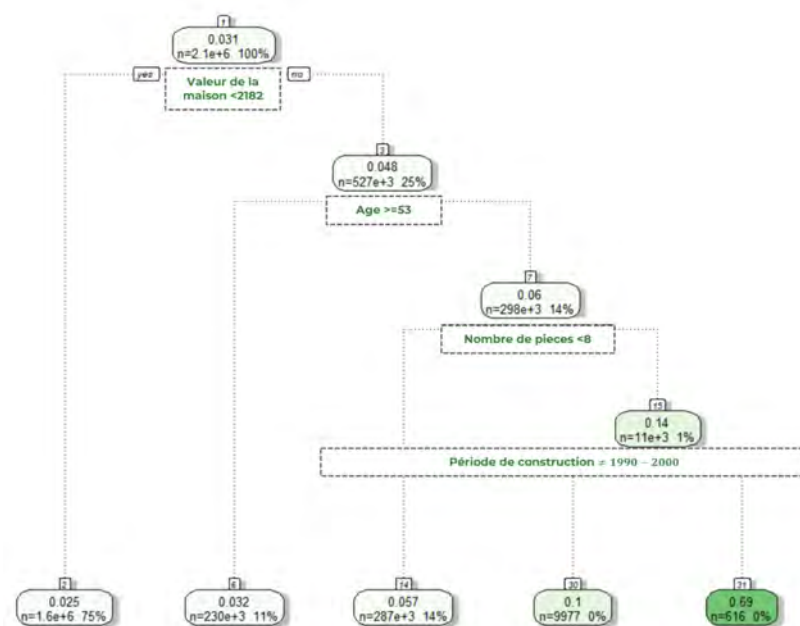


FIGURE 10.7 – À la différence du GLM, le CART ajoute des interactions : en effet la valeur de la maison a un impact différent sur la fréquence selon la valeur des autres variables. Nous remarquons que la période de construction entre 1990 et 2000 est discriminante : la fréquence prédite pour cette classe de bâtiment est de 0.69 sinistres (par an) lorsque l’occupant est âgé de plus de 53 ans, le nombre de pièces est supérieur à 8 et la maison vaut en moyenne plus que 2 182 EUR/m².

Random Forest

La forêt aléatoire est une agrégation d’arbres de régression où l’on sélectionne aléatoirement les variables à chaque étape de partition de la base. Afin de retenir le meilleur modèle nous avons calibré les paramètres de ce modèle. Pour réduire le temps de calcul nous avons réduit la base d’apprentissage à un échantillon de taille 50 000 observations.

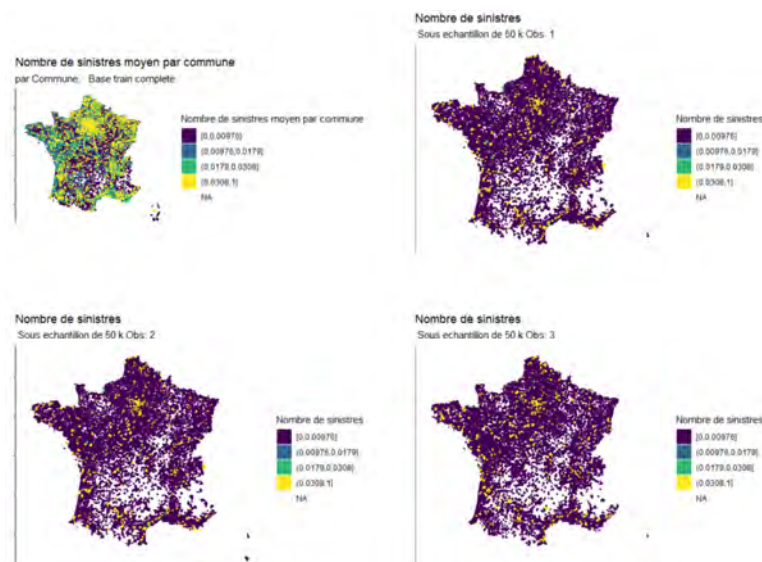


FIGURE 10.8 – Nous avons comparé trois échantillons à la maille communale et choisi celui qui se rapproche le plus à la moyenne par commune (en haut à gauche).

XGBoost

Pour le modèle Extreme Gradient Boosting (XGBoost) nous avons choisi comme fonction objectif la *régression poissonnienne* et comme métrique d’évaluation du modèle la déviance poissonnienne : dans la littérature actua-

rielle, des paramètres initiaux sont proposés², mais les variables explicatives utilisées dans l'exemple pourraient ne pas être adaptées à la base de données.

C'est pour cela que nous avons optimisé les paramètres (annexe F). Cette méthode, comme toutes celles qui dérivent du boosting est sensible aux données catégorielles. Nous les avons donc au préalable encodées en variables binaires (one-hot encoding).

10.2 Modèle de Sévérité

Les coûts des sinistres ont été mis en "as -if" en utilisant l'indice IPEA résidentiel et l'inflation :

$$Sinistre_{2016, corrigé} = \left(\frac{IPEA_{2018(T4)}}{IPEA_{2016(T4)}} - 1 + (1 + Inflation_{2018}) * (1 + Inflation_{2017}) \right) * Sinistre_{2016} \quad (10.1)$$

$$Sinistre_{2017, corrigé} = \left(\frac{IPEA_{2018(T4)}}{IPEA_{2017(T4)}} - 1 + (1 + Inflation_{2018}) \right) * Sinistre_{2017} \quad (10.2)$$

GLM

Les lois gamma et log-normale sont les plus utilisées en assurance pour modéliser le coût moyen³. Nous avons retenu que la loi Gamma était la plus adaptée à la base d'apprentissage, vis-à-vis du critère AIC.

Comme pour la fréquence, les variables ont été sélectionnées par pénalisation Lasso et test de significativité. Le nombre de variables supprimées est supérieur à celui de la fréquence. Nous gardons 15 variables dans le modèle finale.

La qualité du modèle a été jugée comme dans la section de la fréquence.

CART

L'arbre de régression sur la sévérité est le deuxième modèle de sévérité que nous allons construire. Comme dans la section de la fréquence, nous choisissons le meilleur paramètre de complexité pour que l'arbre ne sur apprenne pas de sa base d'apprentissage :

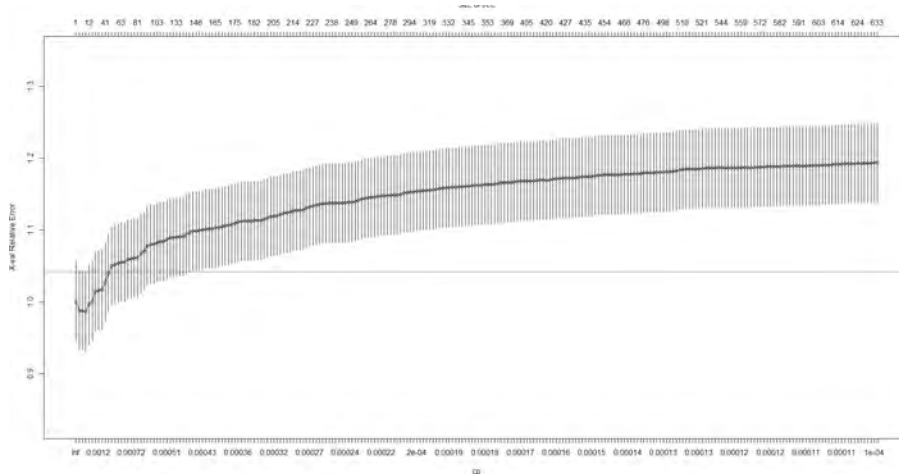


FIGURE 10.9 – Le paramètre de complexité choisi est celui qui minimise l'erreur de validation croisée : 0.001702251.

L'arbre après élagage est facile à interpréter et on remarque que la valeur de la maison, l'âge, le nombre de pièces et la période de construction sont les variables utilisé par l'algorithme *greedy* pour créer une partition de la base d'apprentissage. Les feuilles filles de l'arbre indiquent

- le nombre d'observations dans la feuille fille : pour la feuille de gauche $n = 2879$;
- le coût moyen pour les observations de la feuille : 796 pour la feuille à gauche ;

²Simon Coulombe : "<https://www.simoncoulombe.com/2019/01/bayesian/>"

³Frédéric PLANCHET, Guillaume SERDECZNY : "Modèles fréquence – coût : Quelles perspectives d'évolution ?"

- la portion de la base initiale présente dans la feuille : celle de gauche représente environ le 12% de la base d'apprentissage.

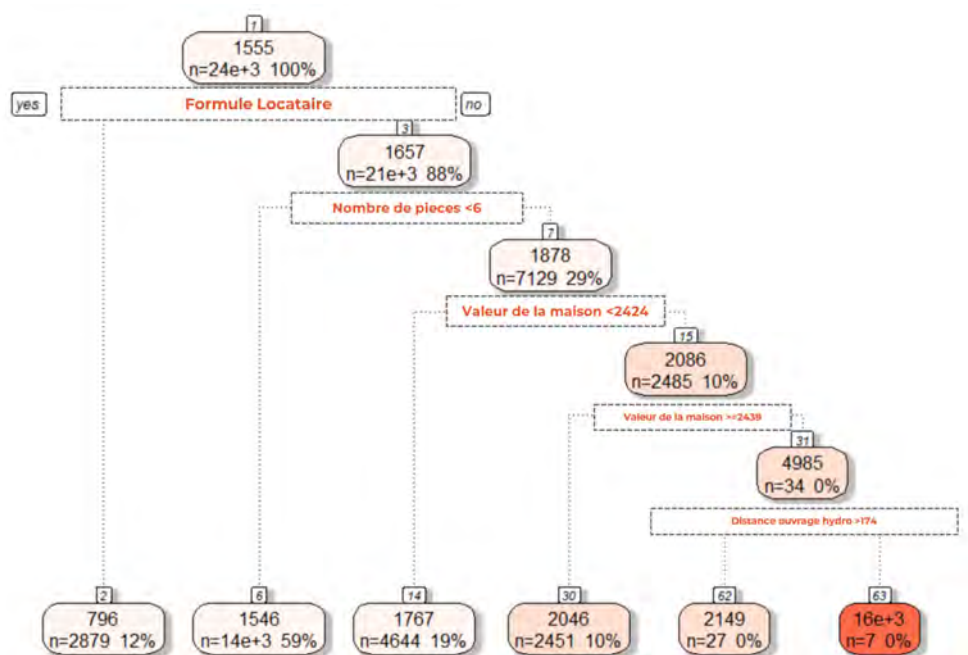


FIGURE 10.10 – À la différence du GLM, le CART ajoute des interactions : en effet le nombre de pièces a un impact différent sur la sévérité selon la valeur des autres variables : la valeur de la maison, la distance d'un ouvrage hydraulique et du type de garanties choisies dans la formule.

Les modèles de forêt aléatoire et xgboost ont été calibrés et optimisés comme pour la fréquence (voir annexe F).

10.3 Comparaison des modèles

Pour évaluer la fiabilité d'un modèle, nous avons fait appel à des indicateurs de performance et à des graphes "prédictions vs valeurs observées".

| Model | Fréquence | | | |
|-------|-----------|----------|----------|----------|
| | Pred Avg | MSE | RMSE | MAE |
| Glm | 0,0321 | 100 | 100 | 100 |
| CART | 0,0310 | 100,0588 | 100,0294 | 98,6923 |
| RF | 0,0358 | 100,5704 | 100,2848 | 105,7835 |
| XGB | 0,0296 | 100,1106 | 100,0553 | 96,4163 |

FIGURE 10.11 – Performance des modèles de fréquence (GLM en base 100) : les modèles sont à peu près équivalents au niveau des métriques d'erreur. La moyenne des prédictions est plus prudente avec Random Forest et sous-estimée avec XGB.

| Model | Severité | | | |
|-------|----------|-------|-------|-------|
| | Pred Avg | MSE | RMSE | MAE |
| Glm | 1555,2 | 100,0 | 100,0 | 100,0 |
| CART | 1554,9 | 100,8 | 100,4 | 100,2 |
| RF | 1572,7 | 100,5 | 100,2 | 100,9 |
| XGB | 1401,7 | 99,7 | 99,9 | 93,9 |

FIGURE 10.12 – Performance des modèles de sévérité (GLM en base 100) : les modèles sont à peu près équivalents au niveau des métriques d'erreur. XGB performe mieux individu par individu, mais la sévérité moyenne est sous-estimée.

Nous avons comparé les prédictions des modèles dans la base de test avec la valeur réelle : les comportements dépendent de la variable choisie. Les modèles de machine Learning souffrent lorsque les catégories ont une faible exposition. Pour un problème de confidentialité, nous présentons uniquement un nombre de variables réduit.

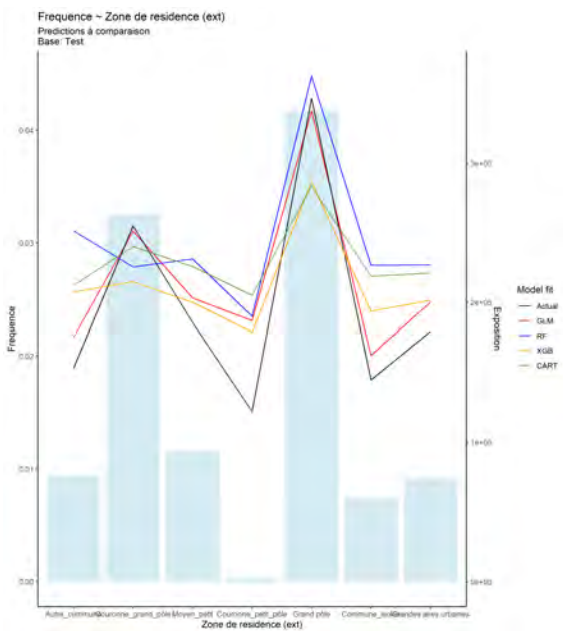


FIGURE 10.13 – Modèles de fréquences à comparaison : la zone de résidence est bien prédite par le GLM, à l'exception des modalités à très faible exposition.

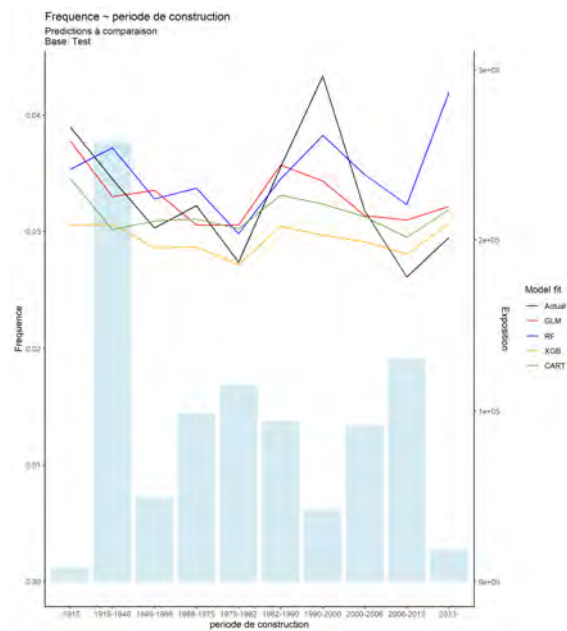


FIGURE 10.14 – Modèles de fréquences à comparaison : les prédictions de la période de construction sont très volatiles sur les quatre modèles.

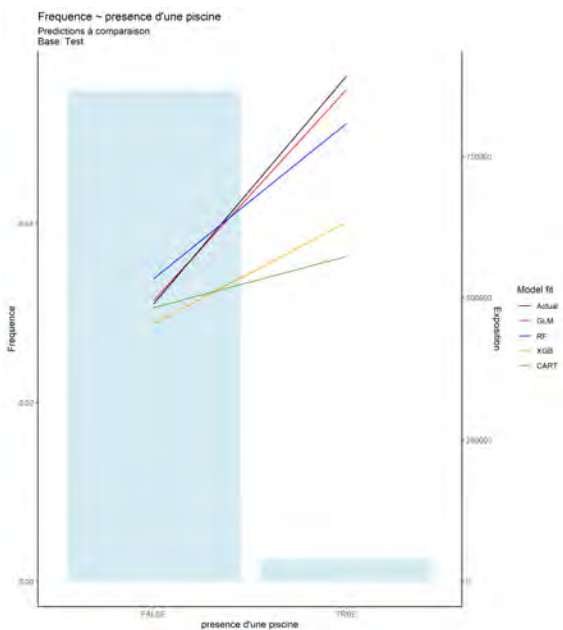


FIGURE 10.15 – Modèles de fréquences à comparaison : la présence d'une piscine non déclarée est bien captée par le GLM et Random Forest.

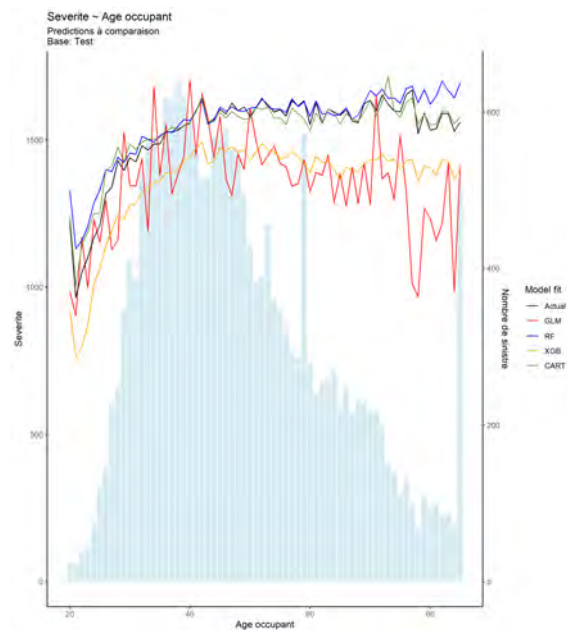


FIGURE 10.16 – Modèles des sévérité à comparaison pour l'âge : CART et Random Forest sont le plus proche à la valeur observée.

Comparaison de l'effet spatial

On observe que pour le GLM (deuxième graphe de la fig. 10.17), le nombre moyen de sinistres est élevé dans les grandes villes (la variable de zone de résidence a été pris en compte). CART partitionne "brusquement" les grandes zones urbaines (la valeur de la maison est la variable plus importante), tandis que le Random Forest et le Xgboost semblent prendre en compte d'autres facteurs (climatiques et météorologique) dans leur prédiction.

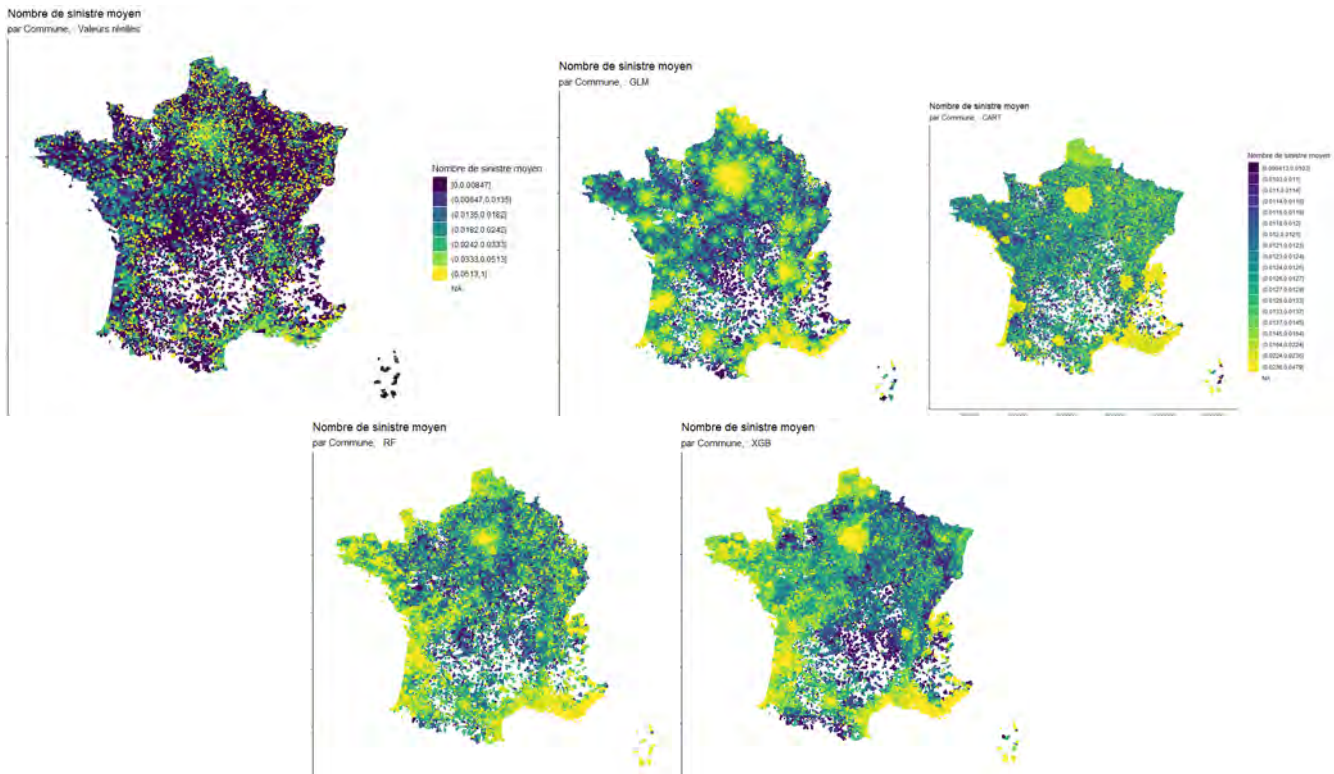


FIGURE 10.17 – Nombre de sinistres moyen prédits dans la base de Test (les échelles des couleurs ne correspondent aux quantiles pour chacun des résultats).

Conclusion

Les modèles construits dans cette section ne sont pas ceux retenus pour le calcul de prime pure. Nous pouvons constater que, même si les quatre modèles ont le même ordre d'erreur, ils aperçoivent le risque de façon différente : la comparaison sur la carte de France montre que les modèles de machine learning prennent en compte le risque géographique de façon plus fine.

L'objectif étant d'intégrer des améliorations sur le modèle linéaire généralisé, nous allons chercher dans les modèles subsidiaires CART, Random Forest et Xgboost des interactions significatives. La partie suivante traitera la détection de ces quantités.

Chapitre 11

Pratique de l'analyse de sensibilité et de l'XAI pour la détection des interactions dans la Tarification à l'adresse

11.1 Analyse de sensibilité

Comme vu dans les chapitres dédiés à l'analyse de sensibilité, le type et la technique d'analyse de sensibilité dépendent des objectifs que l'on souhaite atteindre.

Afin d'améliorer la qualité du modèle initial, nous souhaitons calculer les mesures d'importance à l'aide des indices de Sobol.

L'indice du premier ordre nous informe de l'effet principal de la variable d'entrée sur la sortie et correspond à la partie de la variance totale induite par la variable seule. C'est l'indice le plus facilement interprétable et calculable. Les indices totaux indiquent l'effet total d'une variable d'entrée et correspondent à la somme de l'indice du premier ordre et de la partie de la variance totale due aux interactions que la variable a avec les autres entrées. Nous utiliserons ces indices pour exclure de l'étude les interactions des variables où l'indice total est très proche de celui du premier ordre. Enfin, les indices du deuxième ordre sont ceux qui prennent le plus de temps de calcul et qui ont le plus de difficulté à converger.

Problèmes rencontrés

Lorsque les effets sont très faibles, les indices calculés peuvent être supérieurs à l'indice total ou ils peuvent être négatifs. En augmentant la taille de l'échantillon ces problèmes sont résolus et la qualité d'estimation s'améliore.

De plus, la modélisation par métamodèles, dans notre cas représentée par les modèles de tarification, peut propager l'erreur de prédiction à l'estimation des indices de Sobol. Si le coefficient de prédiction est très faible par exemple, l'erreur se propage et on risque d'introduire des interactions qui n'existent pas. Pour cela, nous nous sommes assurés que le coefficient de prédiction (Q2) était suffisamment élevé (> 70%) et nous avons retenu uniquement les interactions les plus fortes (> 2% au deuxième ordre).

Enfin, les implementations actuelles de méthodes d'analyse de sensibilité sur logiciel ne prenant pas en compte correctement les données mixtes, nous avons retraité au préalable les données pour ne pas avoir plus de 8 modalités uniques.

11.1.1 Spécificités du contexte assurantiel

Dans la plupart des analyses de sensibilité, un plan d'expérience est construit en "pilotant" les entrées, afin d'estimer les indices de Sobol : on fait varier les variables, en imposant éventuellement des contraintes, afin de calculer (ou estimer dans le cas des métamodèles) les sorties et pour en réaliser l'analyse de sensibilité souhaitée.

À la différence de ces applications où le modèle a une forme analytique et la distributions des entrées est connue, nous disposons d'une base historique de données sinistres, c'est-à-dire de réalisations de variables aléatoires, ce qui fait "subir une base" plutôt que de la "piloter".

La convergence des estimateurs peut ainsi être plus lente que dans les cas usuels.

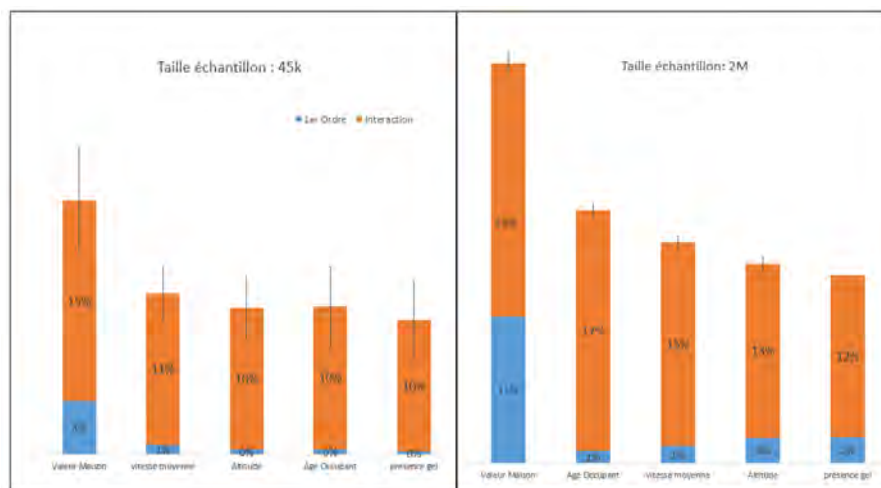


FIGURE 11.1 – Estimation des indices du premier ordre et d’ordre total (somme du premier ordre et des Interactions), détails des premières cinq variables. A gauche, la taille de l’échantillon utilisée pour l’estimation avec la méthode de Sobol-Saltelli est de 45 000 observations : les intervalles de confiance sont amples et les effets du premier ordre presque inexistantes. A droite, la taille de l’échantillon est de 2 millions : les intervalles de confiance sont beaucoup plus réduits.

11.1.2 Méthodologie

L’objectif de cette section est d’effectuer une analyse de sensibilité globale qui permettra ensuite de comparer le modèle simple à d’autres modèles qui prennent en compte les interactions parmi les variables.

La méthode utilisée est celle de Sobol-Saltelli pour l’estimation des indices de Sobol.

- D’abord nous avons commencé par construire quatre métamodèles (deux forêts aléatoires et deux xgboost, pour la modélisation de la fréquence et de la sévérité).
- Ensuite nous avons choisi deux échantillons aléatoires avec remise dans la base initiale, de taille 2 millions pour la fréquence et 50 000 pour la sévérité.
- Nous avons estimé les indices de Sobol du premier ordre, d’ordre total et du deuxième ordre à l’aide de la fonction *sobol* du package *sensitivity* de R.
- Enfin, nous avons réitéré le processus, afin de mesurer la robustesse des estimateurs avec les intervalles de confiance par bootstrap.

Fréquence

Indice du premier ordre et d’ordre total

Dans les deux modèles de fréquence, les variables qui ont un impact significatif sur la variance de la variable cible sont : la valeur de la maison, l’âge de l’occupant, la vitesse moyenne du vent, l’altitude et la présence du gel.

Nous remarquons que la valeur de la maison dans le modèle de forêt aléatoire (RF) participe à elle seule au 11% de la variance totale, alors que dans le modèle de xgboost (XGB) son impact est surtout dû aux interactions.

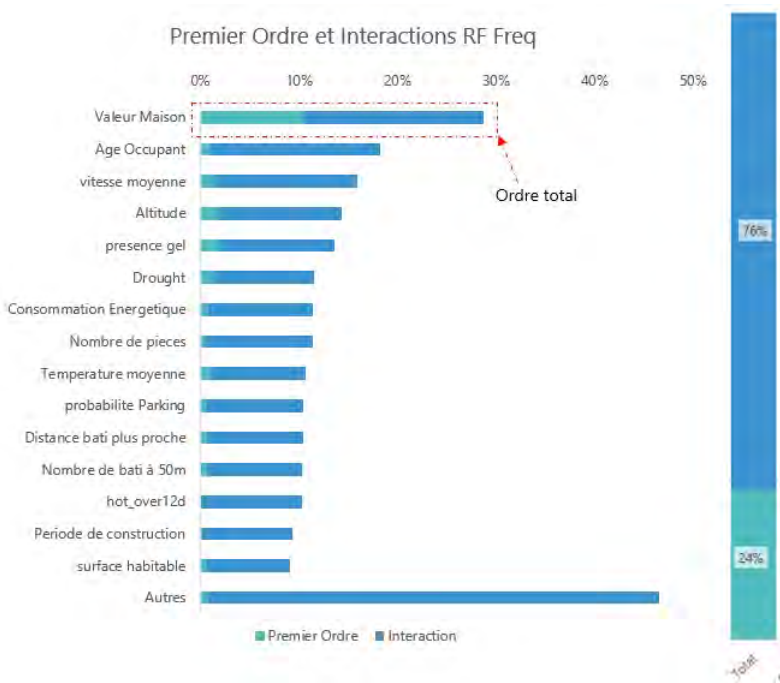


FIGURE 11.2 – Indice de Sobol du premier ordre et impacts des interactions dans le modèle de forêt aléatoire pour la fréquence : les effets principaux expliquent 24% de la variance totale de la sortie du modèle, alors que les interactions contribuent à 76%. La variable la plus importante au sens de la décomposition fonctionnelle de Sobol est la valeur de la maison.

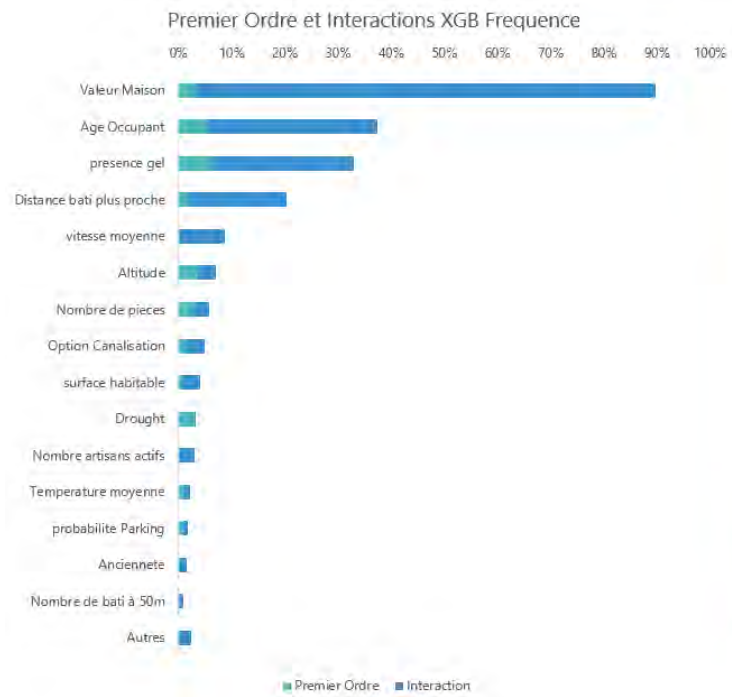


FIGURE 11.3 – Indice de Sobol du premier ordre et impacts des interactions dans le modèle xgboost pour la fréquence : les effets principaux expliquent le 31% de la variance totale de la sortie du modèle, tandis que les interactions contribuent à la hauteur de 69%. La valeur de la maison contribue significativement à la variance totale avec les interactions avec les autres variables.

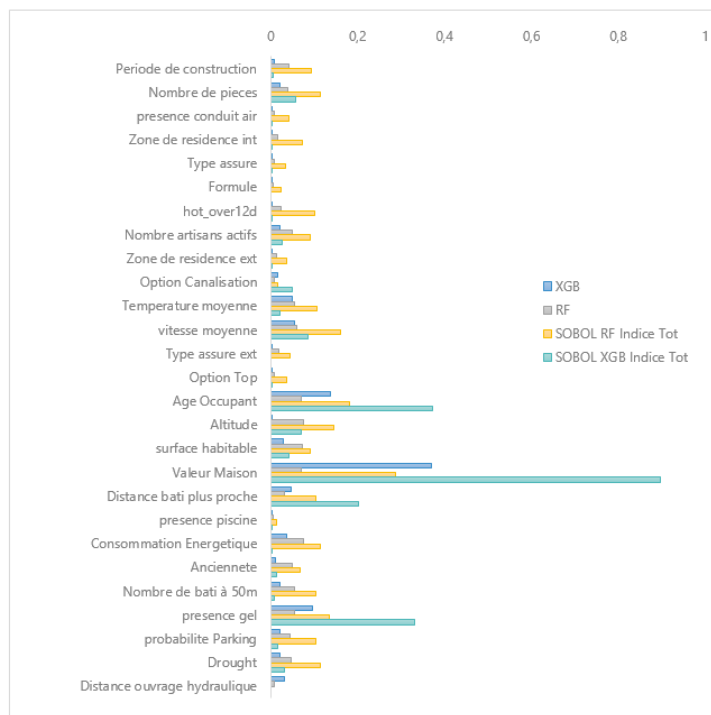


FIGURE 11.4 – Comparaison des importances des variables grâce à la méthode de Sobol et celles de Random Forest et xgboost basées sur le gain d'une fonction de coût : l'importance Xgboost et Sobol avec un métamodèle xgboost sont plus proche au niveau de la priorisation des variables que la forêt aléatoire et Sobol avec un métamodèle de forêt aléatoire.

Interactions par les indice du deuxième ordre

Les indices du deuxième ordre nous permettent de définir les interactions captées par les modèles. Nous n'avons gardé que les interactions supérieures à 1% pour éviter d'intégrer des interactions artificielles : les interactions de la valeur de la maison sont les plus fortes, suivies par celles parmi les variables géographiques et climatique (quantité de pluie, sécheresse, altitude, vitesse du vent).

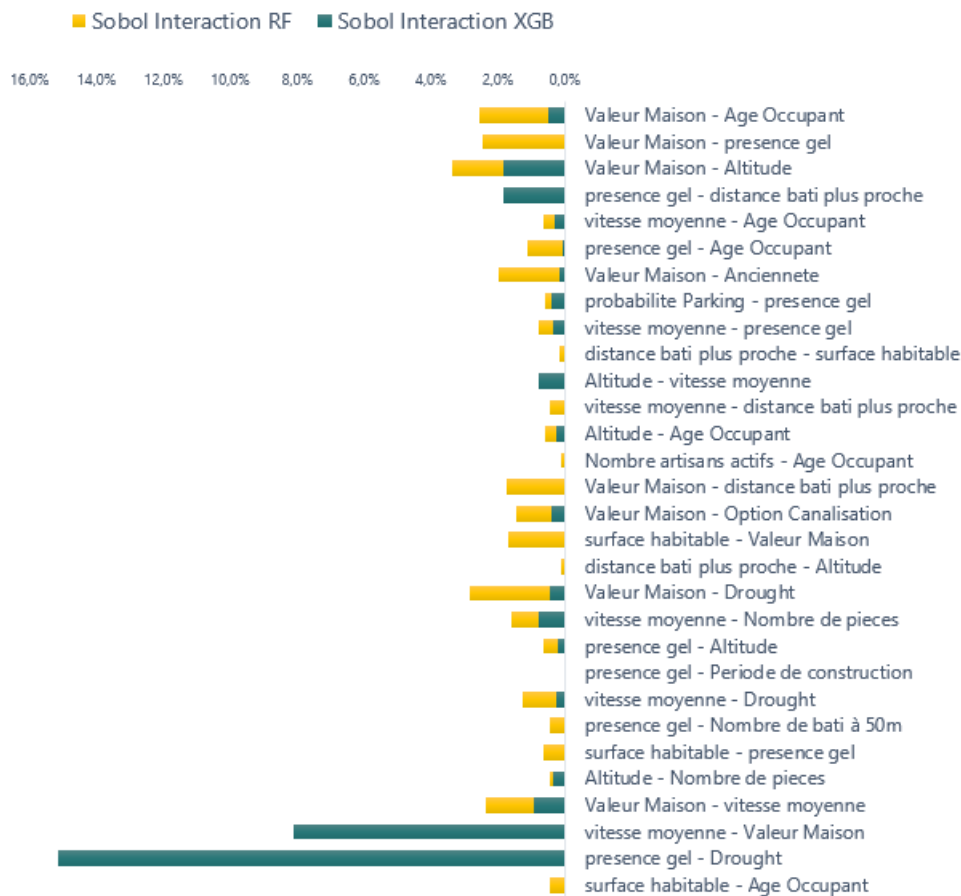


FIGURE 11.5 – Visualisation des interactions les plus significatives à l'ordre 2 pour les modèles de fréquence. Les interactions les plus fortes sont celles de la valeur de la maison avec l'âge, la présence de gel, l'altitude, l'ancienneté, la sécheresse, la vitesse du vent. L'interaction "présence de gel - sécheresse" contribue à 15% de la variance totale.

Sévérité

Indice du premier ordre et d'ordre total

Les modèles de sévérité sont en général bien expliqués par des modèles linéaires généralisés. La contribution des interactions à la variance totale est ainsi limitée. Comme dans la fréquence, les modèles de forêt aléatoire et xgboost captent des interactions différemment. Xgboost attribue aux interactions d'ancienneté, nombre de bâtiments dans le rayon de 50 mètres, la quantité de pluie annuelle et la consommation énergétique 4 % de la variance totale. La variance de la sortie du modèle de forêt aléatoire est quant à elle surtout causée par les interactions des toutes les variables, à l'exception de la Formule et du type d'assuré.

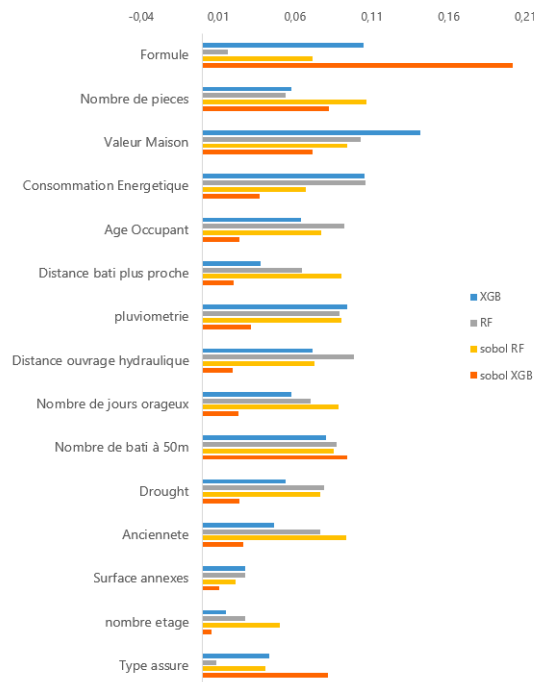


FIGURE 11.8 – Comparaison parmi les importance selon Sobol et celles de la forêt aléatoire et xgboost.

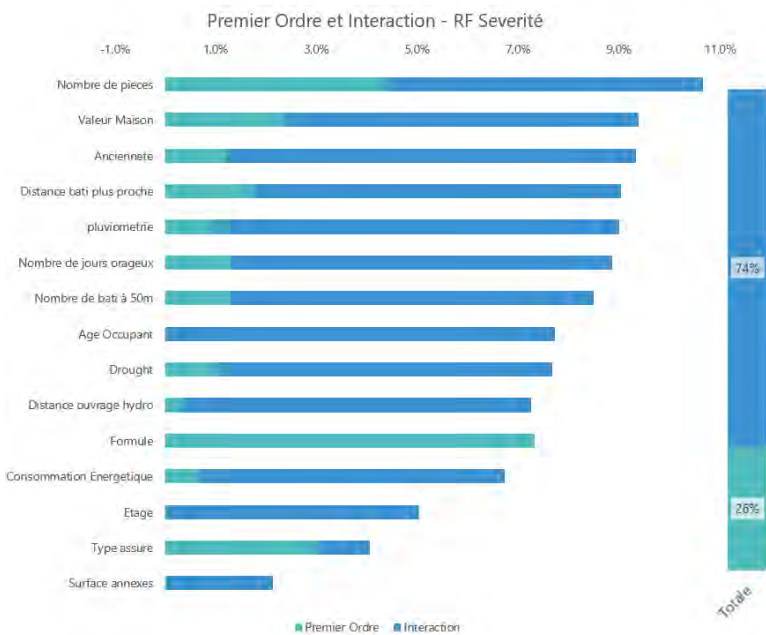


FIGURE 11.6 – Indice de Sobol du premier ordre et impacts des interactions dans le modèle forêt aléatoire pour la sévérité : les effets principaux expliquent le 26% de la variance totale de la sortie du modèle, alors que les interactions contribuent au 74%. Le nombre de pièces du logement contribue significativement à la variance totale et plus de la moitié de son impact est dû aux interactions avec les autres variables.

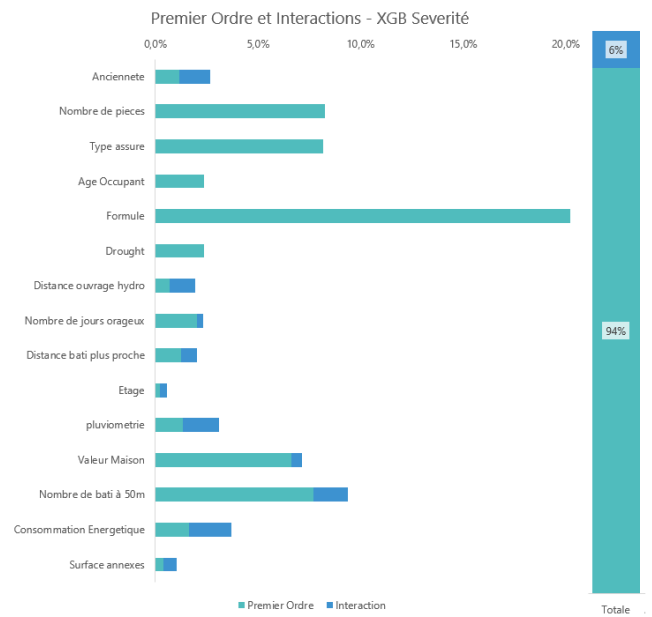


FIGURE 11.7 – Indice de Sobol du premier ordre et impacts des interactions dans le modèle xgboost pour la sévérité : les effets principaux expliquent le 94% de la variance totale de la sortie du modèle, alors que les interactions contribuent au 6%. La Formule du contrat contribue significativement à la variance totale, suivie par le nombre de bâtiments à distance de 50m, la valeur de la maison, le nombre de pièces et le type d'assuré. Les interactions du modèles ont un impact moindre que dans la forêt aléatoire.

Interactions par les indice du deuxième ordre

Par rapport au modèle de fréquence, les interactions sont beaucoup moins fortes : les interactions du nombre des pièces avec la valeur de la maison, la quantité de pluie, l'âge de l'occupant, la formule et la sécheresse sont les plus significatives, suivies par celles de la valeur de la maison.

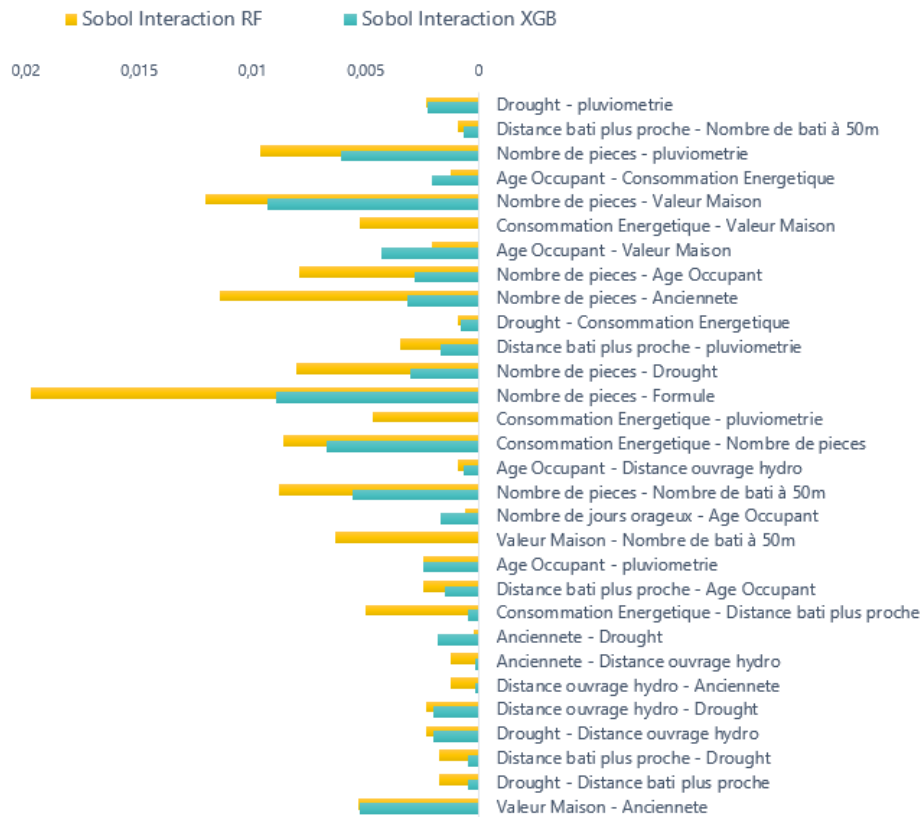


FIGURE 11.9 – Visualisation des interactions plus significatives à l'ordre 2 pour les modèles de sévérité. Les interactions plus fortes sont celles du nombre de pièces avec la pluviométrie, la valeur de la maison, l'âge, l'ancienneté, la formule, la sécheresse, la consommation énergétique, le nombre de bâtiments à 50m.

Visualisation à l'aide de la théorie des graphes

Nous représentons les indices d'ordre 2 au-delà d'un seuil sur un graphe non orienté, où les arcs qui relient deux variables sont autant épais que l'interaction est importante. Un code de couleurs des niveaux d'information



des variables a été utilisé pour faciliter la lecture des graphes :

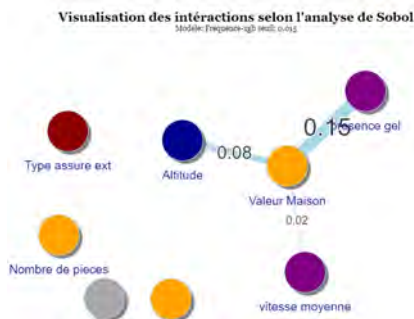


FIGURE 11.10 – Visualisation des interactions supérieures à 1.5% pour le modèle de fréquence xgboost. Peu d'interactions sont visualisées : la valeur de la maison est la variable qui interagit le plus.

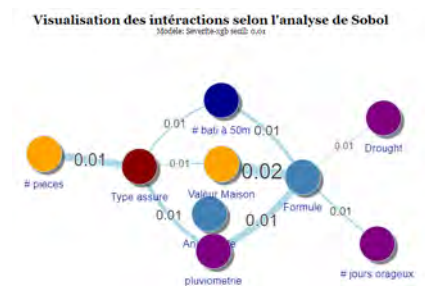


FIGURE 11.11 – Visualisation des interactions supérieures à 1% pour le modèle de sévérité xgboost. La Formule et la valeur de la maison sont les variables qui interagissent de plus.

Visualisation des interactions selon l'analyse de Sobol

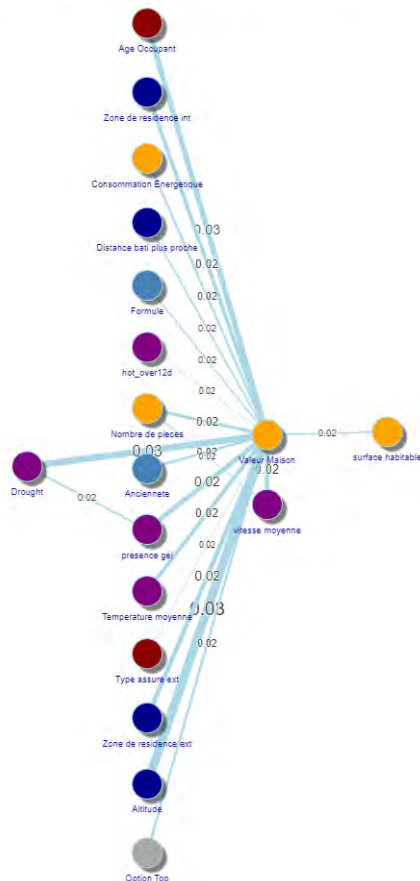


FIGURE 11.12 – Visualisation des interactions supérieures à 1.5% pour le modèle de fréquence de forêt aléatoire. La valeur de la maison est la variable qui interagit le plus avec les autres covariables.

Visualisation des interactions selon l'analyse de Sobol

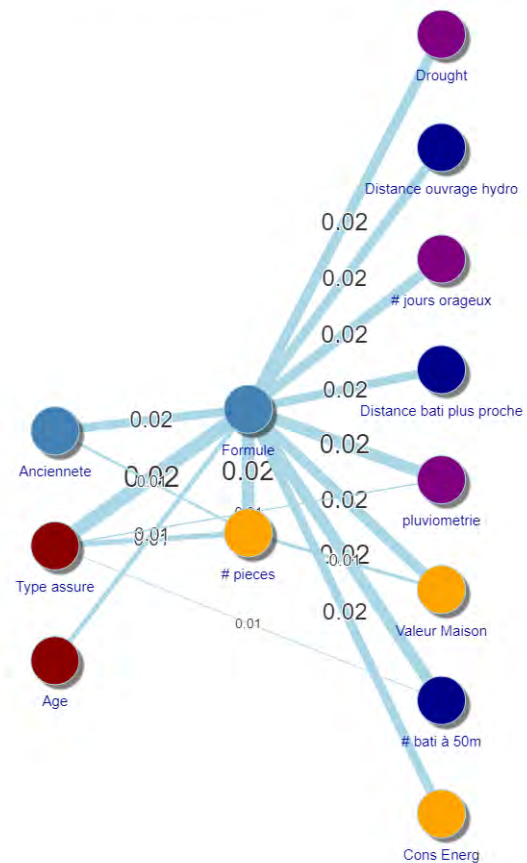


FIGURE 11.13 – Visualisation des interactions supérieures à 1% pour le modèle de sévérité de forêt aléatoire. La Formule est la variable qui interagit le plus avec les autres covariables.

11.2 XAI

L'analyse de sensibilité de Sobol a été utilisée dans son approche globale : nous avons détecté des variables qui, par leur façon de se croiser, déterminent une variation non constante sur la sortie. Cette analyse est très utile pour fournir une quantification du poids que chaque interaction a sur la variance de la sortie, elle est toutefois réalisée sur tout le domaine de variation des variables explicatives.

En effet, il peut s'avérer que deux variables aient une forte interaction sur un très petit intervalle du domaine de définition des variables.

Dans ce cas même si l'interaction globale est assez significative, étant la variance sensible aux valeurs extrêmes, au moment de l'intégration du terme d'interaction dans le GLM de départ, les régions où l'effet d'interaction est absent peuvent cacher sa significativité.

Nous avons ainsi fait recours à une analyse locale, au complément de celle globale de Sobol.

Les modèles d'explication de la classe *Additive Feature Attribution* ont été choisis en tant qu'outils locaux, dans le cas où ils vérifient les propriétés de la valeur de Shapley. En particulier, nous avons calculé le poids des variables avec les valeur de SHAP.

L'analyse de Sobol locale est quant à elle limitée aux hypothèses de linéarité ou normalité, alors que nous souhaitons exploiter n'importe quel modèle ou hypothèse sous-jacente¹. Afin de comparer les deux méthodes, des indices globaux ont été introduits comme dans la section 7.3.2

Par ailleurs, réaliser une analyse locale permettrait de détecter des profils comme "jeune locataire dans le centre ville" ou "étudiant dans un centre urbain", etc. qui peuvent se concrétiser dans la proposition de formules MRH particulières.

¹La régularité des modèles (classe L2 par exemple) est quand même souhaitée pour avoir des estimateurs convergents, mais dans la pratique il est compliqué de le vérifier.

Problèmes rencontrés

La taille de l'échantillon d'apprentissage est cruciale pour avoir des estimateurs de qualité.

Pour les modèles de sévérité, nous avons utilisé toute la base d'apprentissage, environ 50 000 observations, afin d'appliquer l'algorithme Strumbelj - Kononenko ou celui TreeSHAP.

Pour la fréquence, où la base d'apprentissage a près de 2 millions de lignes, nous avons eu des problèmes d'utilisation de mémoire vive sous le logiciel.

Nous avons ainsi réduit la taille de la base d'apprentissage à un sous-échantillon représentatif, satisfaisant ces critères :

1. La distribution des variables doit se rapprocher à celle de la base d'apprentissage entière.

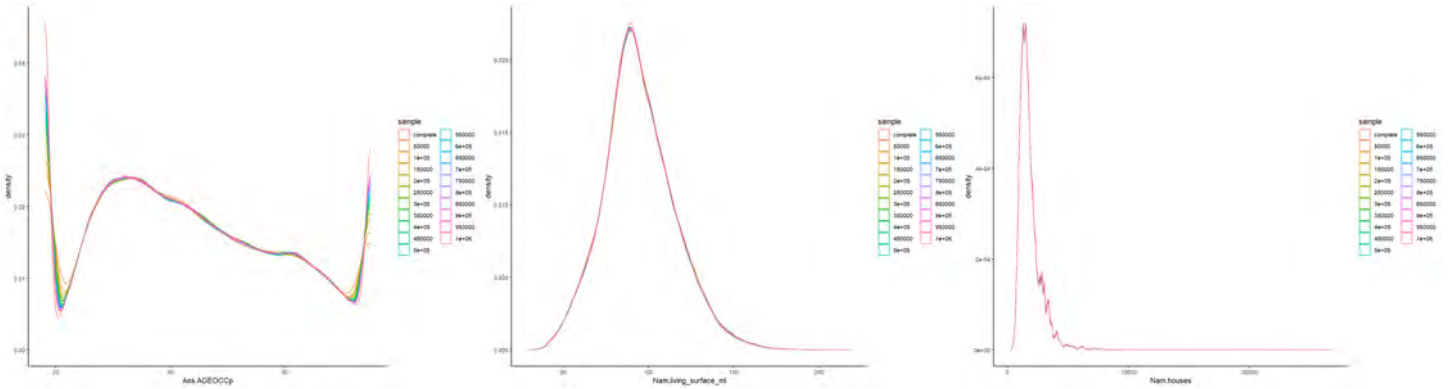


FIGURE 11.14 – Distribution de la variable *Surface habitable*, *Âge occupant* et *Valeur de la maison* selon la taille de l'échantillon de la base d'apprentissage : nous avons choisi 50 000 comme taille pour pouvoir appliquer les algorithmes Strumbelj - Kononenko ou TreeSHAP.

2. La taille de l'échantillon et le nombre des itérations doivent garantir la convergence de l'algorithme

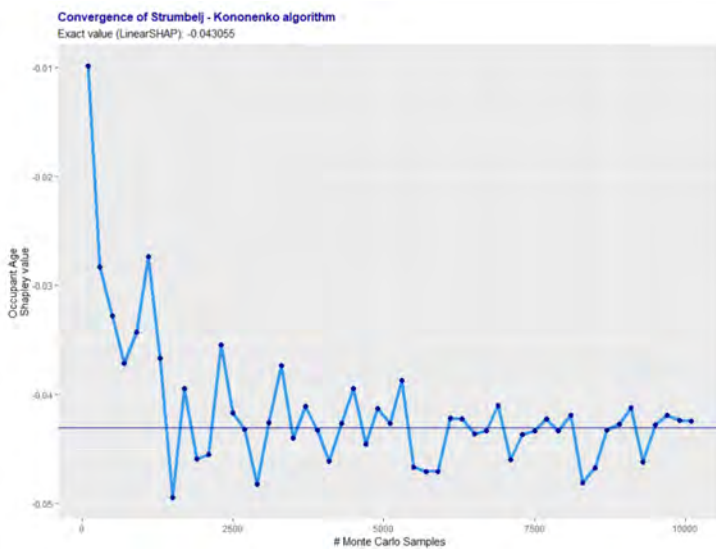


FIGURE 11.15 – Pour avoir une idée de l'ordre de grandeur de la taille requise, nous avons appliqué l'algorithme de Strumbelj - Kononenko pour le calcul de la valeur de Shapley de l'âge de l'occupant avec un metamodèle linéaire parce que on connaît la vraie valeur (l'estimateur LinearSHAP est exacte).

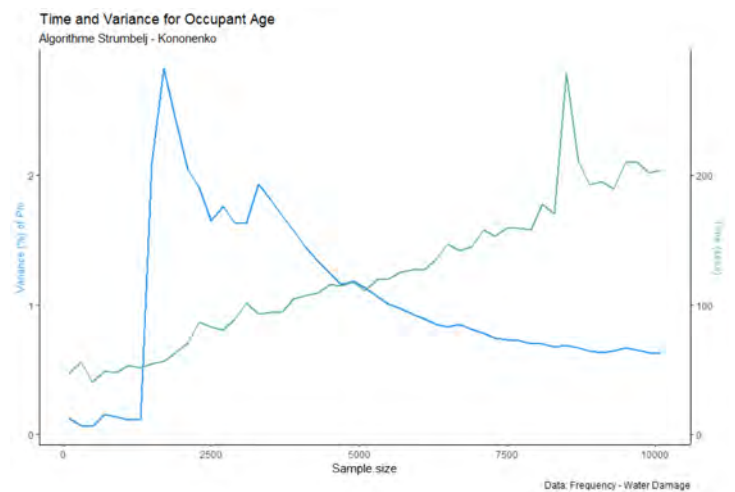


FIGURE 11.16 – Avec le même metamodèle nous avons mesuré la variance des estimations de la valeur de Shapley en bleu et le temps d'exécution de l'algorithme en vert qui est linéaire avec la taille.

11.2.1 Méthodologie

Les modèles d'explication font partie des récents développements de l'*Explainable Artificial Intelligence*.

Leur objectif est d'associer une structure additive à une prédiction où les termes qui la composent sont les

importances des variables. Une partie de l'importance d'une variable (que l'on peut comparer à l'indice total de Sobol) est due à l'interaction de la variable avec les autres covariables.

Afin de détecter ces interactions :

- nous avons construit quatre métamodèles (deux forêts aléatoires et deux xgboost, pour la modélisation de la fréquence et de la sévérité) ;
- ensuite nous avons choisi l'échantillon d'apprentissage si la réduction de la base était nécessaire ;
- et puis nous avons estimé les valeurs de SHAP à l'aide des packages *xgboost*, *SHAPforxgboost* de R et du package *shap* de python, en utilisant les algorithmes Strumbelj - Kononenko, KernelSHAP et TreeSHAP. Pour réduire le temps d'exécution avec KernelSHAP, il est possible de sous-échantillonner ultérieurement avec la méthode de *k*-means ; pour réduire le temps de TreeSHAP, on peut limiter la profondeur de l'arbre. Les estimateurs auront toutefois plus de variance.
- Les interactions sont calculées avec l'algorithme TreeSHAP.
- Afin de mesurer la robustesse de l'estimateur nous avons réitéré le processus pour obtenir un intervalle de confiance par bootstrap.

SHAP Summary Plot

Le graphique synthétique des valeurs de SHAP

$$\{\varphi_i(x) : i = 1, \dots, p \text{ variables explicatives ; } x \in \mathcal{V} \text{ (Base de validation)}\}$$

sur toute (ou une partie) de la base de validation nous donne un aperçu de l'importance des variables considérées par les modèles de prédiction. L'importance est définie comme la contribution marginale moyenne d'une variable sur toutes les coalitions possibles auxquelles elle peut participer. Dans ces graphes, l'effet d'interaction, ainsi, n'est pas isolé de l'effet principale.

Le signe de la valeur indique si la variable participe à augmenter la prédiction (positif) ou à la baisser (négatif), vérifiant la propriété d'efficacité de la valeur de Shapley :

$$f(x) = \varphi_0 + \sum_{j=1}^p \varphi_j x'_j = E_x(\hat{f}(x)) + \sum_{j=1}^p \varphi_j$$

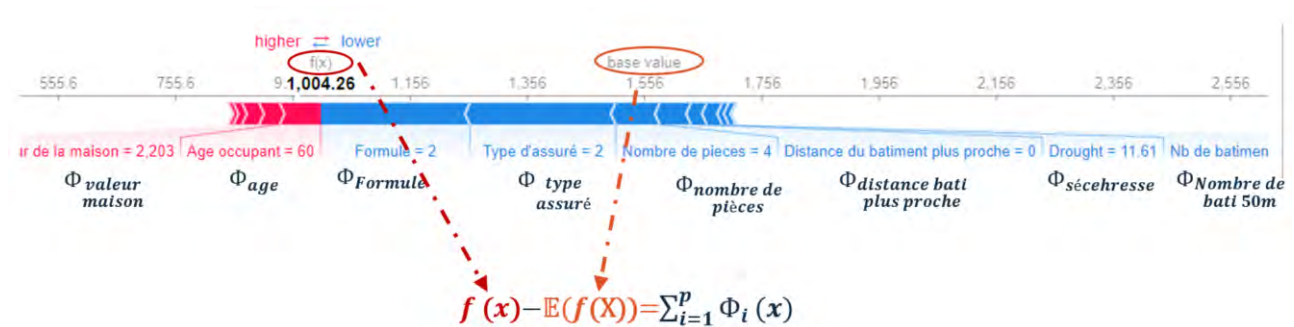


FIGURE 11.17 – Contributions à la prédiction du modèle de sévérité de forêt aléatoire pour une observation particulière.

Lecture : la prédiction pour cet individu est en dessus de la moyenne, les variables de valeur de Shapley négative (en bleu) amènent la sévérité vers le bas.

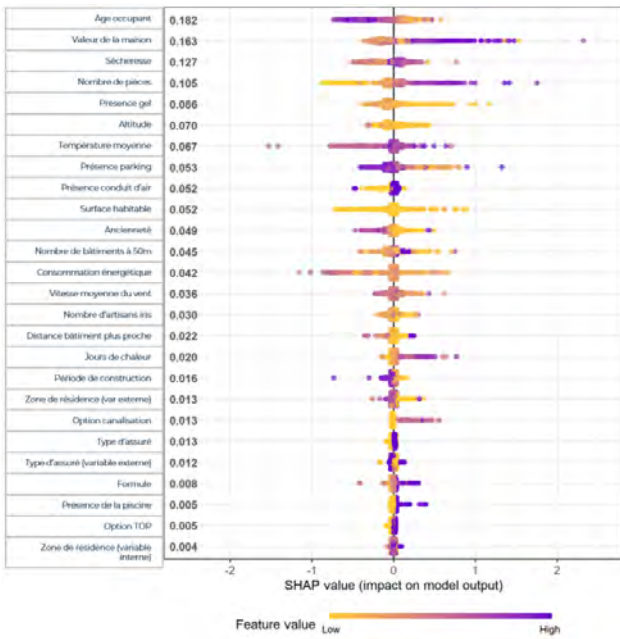


FIGURE 11.18 – Le SHAP *Summary Plot* du modèle de fréquence xgboost associe à chaque variable son impact total (valeur SHAP) sur chaque prédiction (points du graphe). Considérant la variable *Age de l'occupant*, les profils senior (couleur violet) ont en moyenne une valeur négative, ce qui détermine une baisse du coût du sinistre. A l'inverse, les profils jeunes (en jaune) colloquent la prédiction en moyenne au dessous de la prime moyenne.

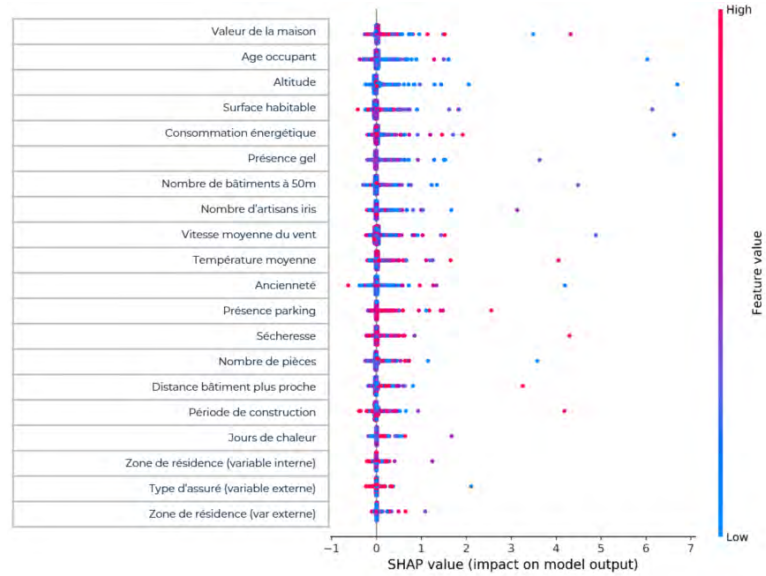


FIGURE 11.19 – Le SHAP *Summary Plot* du modèle de fréquence de la forêt aléatoire associe à chaque variable son impact total (valeur SHAP) sur chaque prédiction (points du graphe). Il est plus difficile de déduire de comportements globaux, car il y a plus de valeurs aberrantes (impact absolu >2) qui rendent le graphique peu lisible. Les profils junior sont en moyenne responsables de la hausse de la fréquence des sinistres, toutefois on remarque que son impact sur la sortie entre 1.5 et 2 est en moyenne du à des valeurs élevés. Cela peut être expliqué par les interactions avec une autre variable localisées à un intervalle particulier.

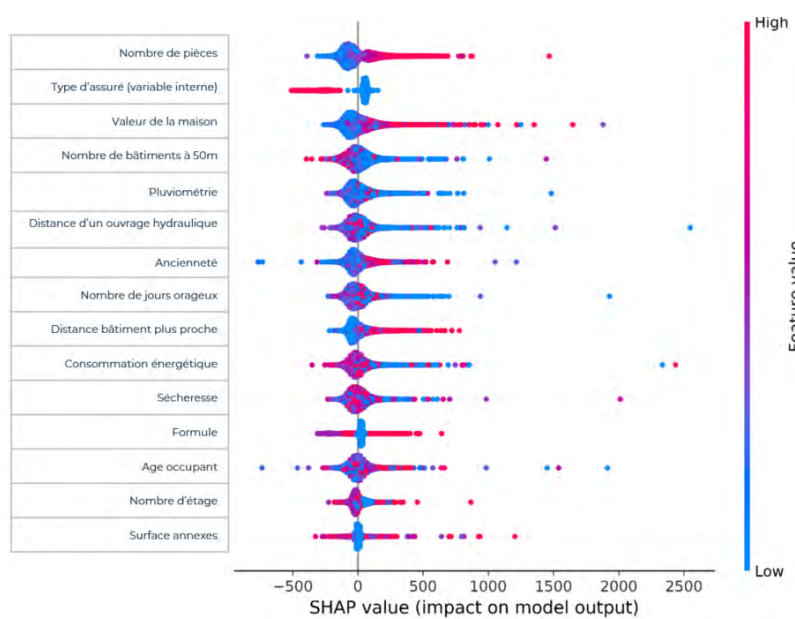
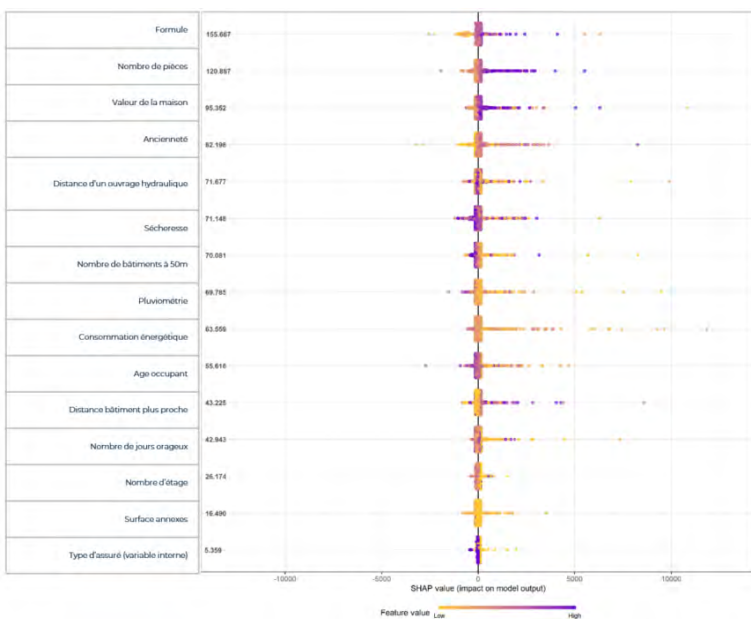


FIGURE 11.20 – Les SHAP *Summary Plot* des modèles de sévérité xgboost (à gauche) et de la forêt aléatoire (à droite) associe à chaque variable son impact total (valeur SHAP) sur chaque prédiction (points du graphe). Considérant la variable *Nombre de pièces*, il détermine une hausse du coût du sinistre en moyenne pour des valeurs élevés (couleur violet)

Nous venons de représenter les valeurs de SHAP sur toute ou une grande partie de la base de validation : à la différence de l'approche globale de Sobol, nous avons pu visualiser intuitivement la présence des interactions en regardant la variation de la valeur de SHAP selon la valeur de la variable. L'âge par exemple semblerait avoir un comportement non linéaire et non constant vis à vis de la fréquence. Nous chercherons ainsi ses interactions qui pourraient justifier ce comportement pour l'intégrer dans le GLM de départ.

SHAP Interaction Values

Les interactions sont les contributions des couples de variables à la prédiction. Elles représentent la partie de la valeur de SHAP qui exclut l'effet individuel (de la variable toute seule) :

$$\underbrace{\Phi_{i,i}}_{\text{effet individuel de la variable } X_i} = \underbrace{\varphi_i}_{\text{valeur SHAP de } X_i} - \underbrace{\sum_{j \neq i} \Phi_{i,j}}_{\text{Interactions de } X_i}$$

La différence par rapport aux indices de Sobol d'ordre deux est que les interactions SHAP ne sont pas des mesures relatives à une quantité (comme la part de variance totale).

Une solution pour normaliser les interactions ou les valeurs SHAP est de considérer l'indice :

$$\frac{|\Phi_{i,j}|}{|f(x) - \mathbb{E}(f(X))|}$$

qui indique la partie de l'écart entre la prédiction et la valeur attendue due à l'interaction.

Dans le cadre de cette étude, notre analyse s'est limitée à la visualisation des interactions dans la base de validation.

Fréquence

Détection d'interaction à partir du modèle xgboost

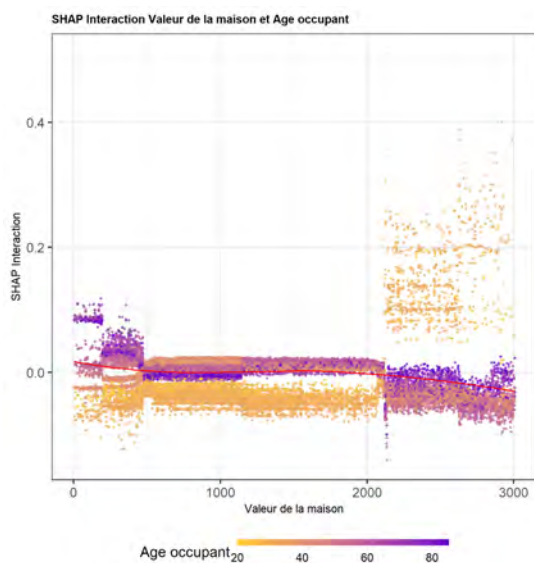


FIGURE 11.21 – Interactions SHAP entre la valeur de la maison et l'âge de l'occupant captées par le modèle de fréquence Xgboost : visualisation des interactions par individu de la base de validation. L'intensité moyenne est de 2,1%, la plus forte du portefeuille. La portée de cette interaction est à la fois locale et à la fois globale. Nous remarquons que la contribution des profils senior habitant dans les zones décentrées (valeur de la maison inférieure) est positive, alors que les jeunes tendent à déclarer plus de sinistres dans des logements qui valent plus (des maisons en centre ville par exemple).

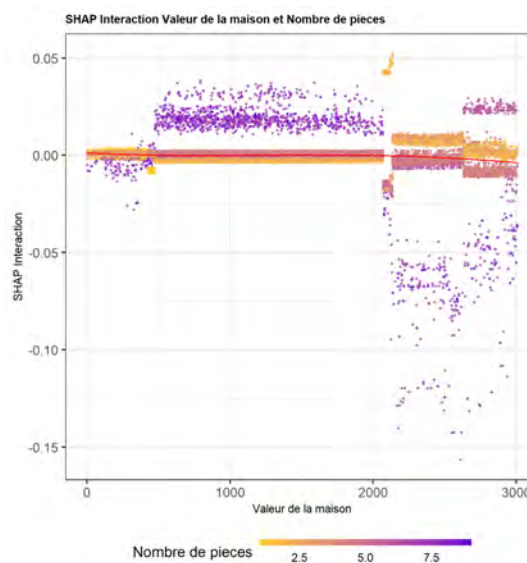


FIGURE 11.22 – Interactions SHAP entre la valeur de la maison et le nombre de pièces captées par le modèle de fréquence Xgboost : visualisation des interactions par individu de la base de validation. L'intensité moyenne est de 0,35%. L'effet du nombre de pièces dans prédiction n'est pas constante selon la valeur de la maison : pour des valeurs jusqu'à 500 Euro/m2 des maisons très grandes font que le nombre de sinistres diminue.

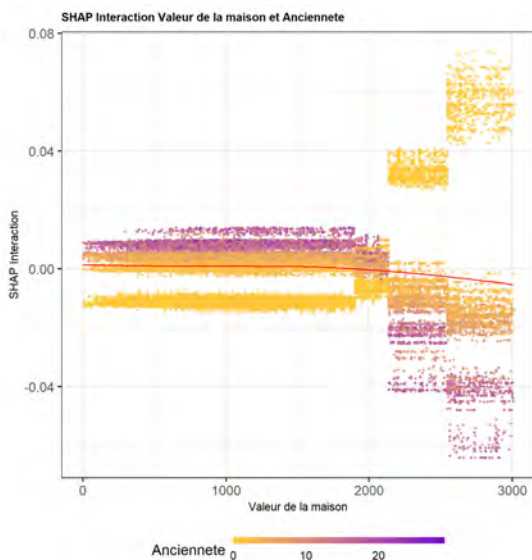


FIGURE 11.23 – Interactions SHAP entre la valeur de la maison et l’ancienneté de l’assuré dans le portefeuille captées par le modèle de fréquence Xgboost : visualisation des interactions par individu de la base de validation. L’intensité moyenne est de 0,89%. Les clients dans le portefeuille depuis plus que 10 ans, ont la tendance à déclarer des sinistres plus fréquemment lorsque la maison vaut jusqu’à 2000 Euro/m². Les nouveaux clients, au contraire, en déclarent plus lorsqu’ils ont de maison de valeur plus élevée.

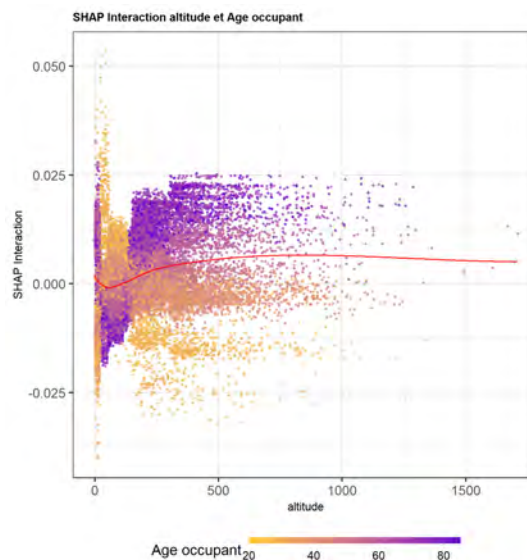


FIGURE 11.24 – Interactions SHAP entre l’âge de l’occupant et l’altitude captées par le modèle de fréquence Xgboost : visualisation des interactions par individu de la base de validation. L’intensité moyenne est 0,67%. Les interactions varient de façon non constante dans l’âge et dans l’altitude de la maison : des profils junior selon le modèle déclareraient moins en altitude que les seniors.

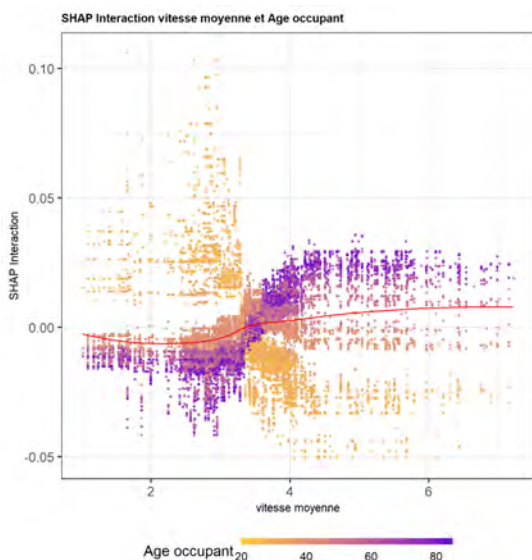


FIGURE 11.25 – Interactions SHAP entre l’âge de l’occupant et la vitesse moyenne du vent captées par le modèle de fréquence Xgboost : visualisation des interactions par individu de la base de validation. L’intensité moyenne est 1,07%. Lorsque la maison est située dans des zone où le vent est fort, les jeunes seraient plus prudents et déclareraient moins de sinistres des profils senior.

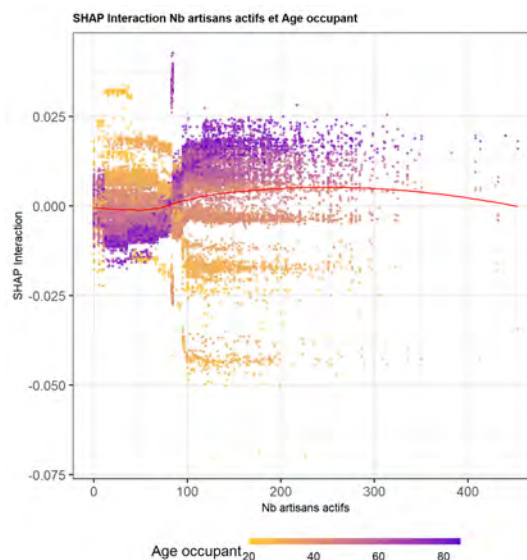


FIGURE 11.26 – Interactions SHAP entre le nombre d’artisans et l’âge de l’occupant captées par le modèle de fréquence Xgboost : visualisation des interactions par individu de la base de validation. L’intensité moyenne est 0,66%. La présence d’artisans actifs de résidence a un impact différent selon l’âge de l’occupant : les profils senior déclarent plus lorsqu’il y a plus de 100 artisans dans l’iris, au contraire des jeunes.

Les interactions suivantes ont une intensité moyenne (moyenne des valeurs absolues dans la base de validation) très faible. Il peut s'agir d'interactions très localisées ou artificielles. Le test de significativité après ajout dans le GLM nous guidera si retenir ou non ces interactions.

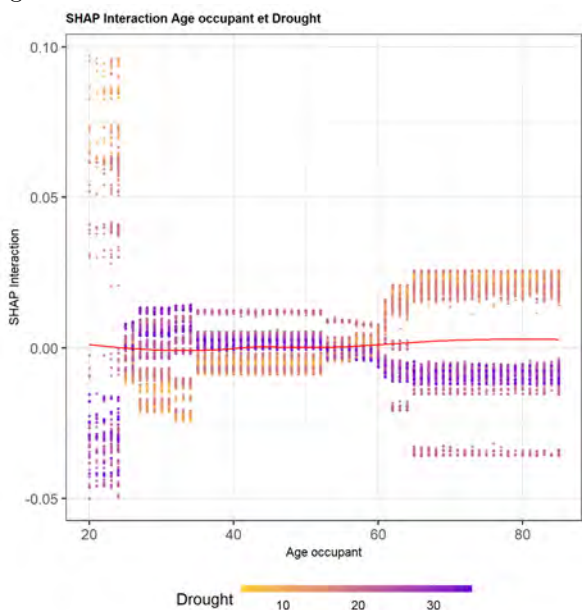


FIGURE 11.27 – Interactions SHAP entre l'âge de l'occupant et la variable météo de sécheresse captées par le modèle de fréquence Xgboost. On reconnaît des comportements différents des interactions selon l'âge et le nombre de jours où il a plu moins de 2mm par jour pendant 15 jours consécutifs : dans des territoires plutôt secs, la population entre 25 ans et 55 ans ont tendance à déclarer plus de sinistres.

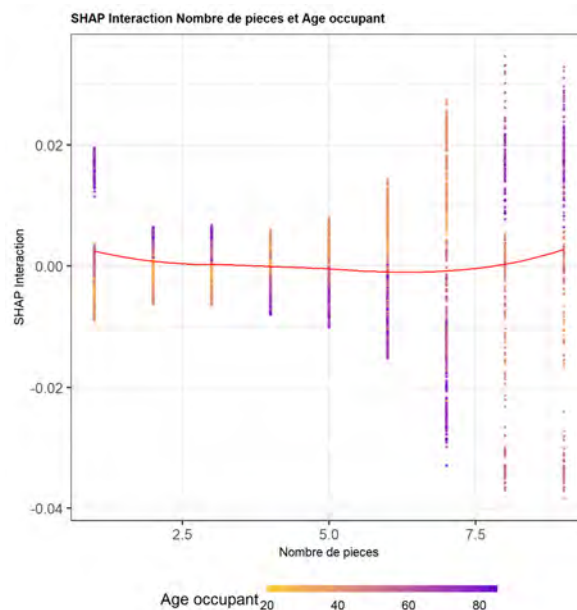


FIGURE 11.28 – Interactions SHAP entre le nombre de pièces et l'âge captées par le modèle de fréquence Xgboost. Jusqu'à trois pièces et à partir de 8 pièces, la population âgée tend à déclarer plus de sinistres, alors que entre 4 et 7 pièces, la sinistralité diminue avec l'âge.

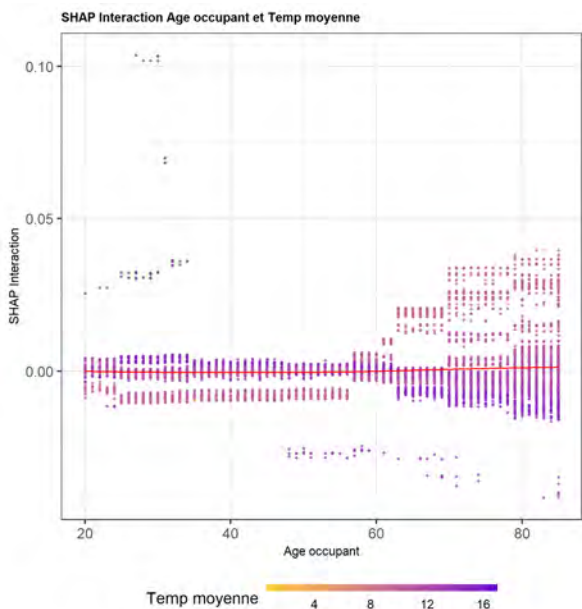


FIGURE 11.29 – Interactions SHAP entre l'âge de l'occupant et la température moyenne captées par le modèle de fréquence Xgboost. Il s'agit d'une interaction très localisée ou d'une interaction artificielle. On remarque dans des territoires plus chauds les jeunes causeraient plus de sinistres que les personnes plus âgées.

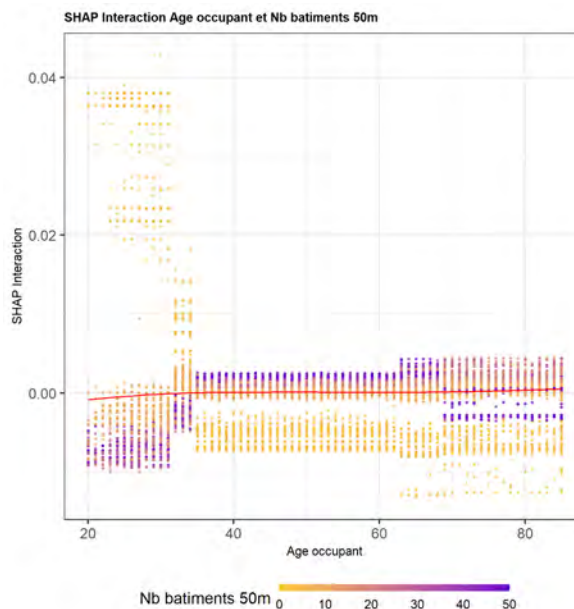


FIGURE 11.30 – Interactions SHAP entre le nombre de bâtiments dans le rayon de 50m et l'âge de l'occupant captées par le modèle de fréquence Xgboost. Dans des territoires très denses (plus de 40 bâtiments dans le rayon de 50m) les profils junior et senior déclarent moins de sinistres.

Détection d'interaction à partir du modèle de forêt aléatoire

Les interactions détectées par la forêt aléatoire (intensité moyenne de 0,001%) sont beaucoup moins fortes que celles du xgboost (intensité moyenne de 0,2%).

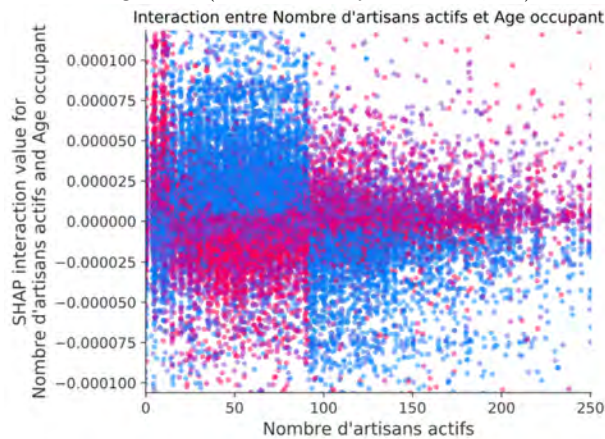


FIGURE 11.31 – Interactions SHAP entre l'âge de l'occupant et le nombre d'artisan par IRIS, captées par le modèle de fréquence forêt aléatoire. Pour des profils senior, le nombre d'artisans a une interaction positive lorsque il y en a peu ou beaucoup, ce qui pourrait être lié à la taille de la ville. La situation est inverse pour des profils junior.

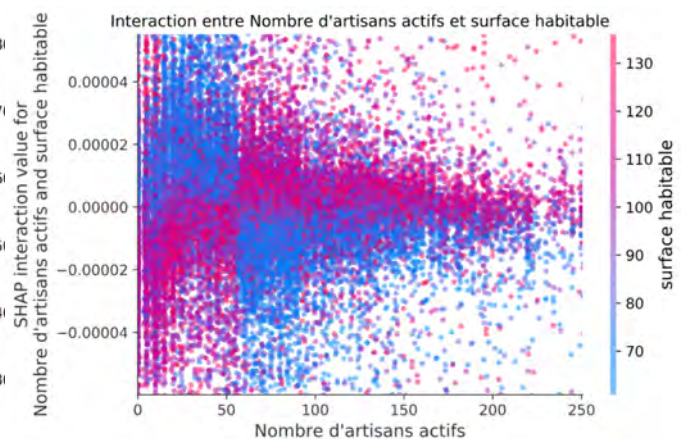


FIGURE 11.32 – Interactions SHAP entre le nombre d'artisans dans l'IRIS et la surface habitable captées par le modèle de fréquence forêt aléatoire. Pour des surfaces habitables entre 100 et 120 m², un nombre élevé d'artisans amènerait la sinistralité à augmenter, qui pourrait être expliqué par le fait que le logement nécessite plusieurs expertises, inversement aux petites surfaces.

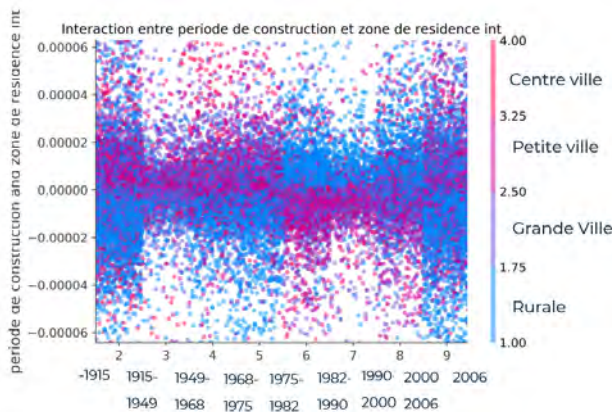


FIGURE 11.33 – Interactions SHAP entre la zone de résidence (variable interne) et la période de construction captées par le modèle de fréquence forêt aléatoire. Les logements dans le centre ville et dans les grandes ou petites villes très récents et précédents à la période 1975-1982 amènent le nombre de sinistres vers le haut. Habiter dans une zone rurale dans un logement construit entre 1975 et 2006 est une autre interaction qui contribue positivement à la déclaration des sinistres.

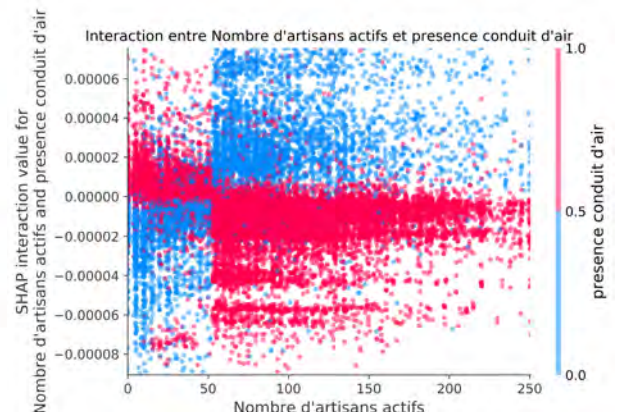


FIGURE 11.34 – Interactions SHAP entre le nombre d'artisans dans l'IRIS et la présence d'un conduit d'air captées par le modèle de fréquence forêt aléatoire. La présence d'un conduit d'air contribuerait positivement dans la prédiction du nombre de sinistres lorsque le nombre d'artisans est faible (petits centres ou zones rurales).

Sévérité

Dans les modèles de sévérité, les interactions SHAP ont un impact très limité dans la prédiction. Nous n'avons retenu que peu d'interactions sur le modèle de xgboost. Dans le modèle de forêt aléatoire, les interactions étaient très dispersées et nous n'avons pas détecté des tendances particulières dans les graphes.

Détection d'interaction à partir du modèle xgboost

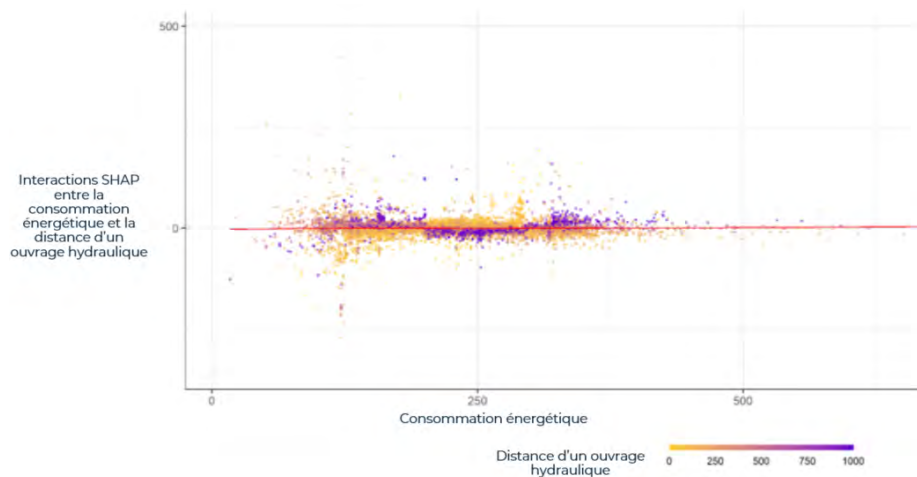


FIGURE 11.35 – Interactions SHAP entre la distance d'un ouvrage hydraulique et la consommation énergétique captées par le modèle de sévérité xgboost : visualisation des interactions par individu de la base de validation. Ils se définissent des profils de sinistralité selon les niveaux de consommation énergétique (abscisse) et la distance d'un ouvrage hydraulique (couleur).

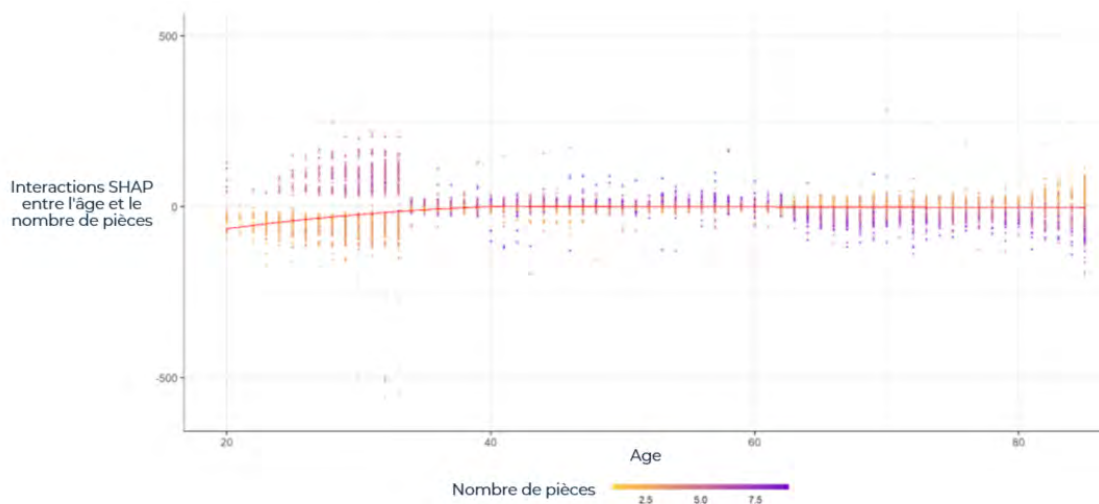


FIGURE 11.36 – Interactions SHAP entre le nombre de pièces (couleur) et l'âge de l'occupant (abscisse) captées par le modèle de sévérité xgboost : visualisation des interactions par individu de la base de validation. L'intensité moyenne est de 14,46. Les profils senior déclareraient des sinistres plus élevés dans des petites maisons, alors que les jeunes causeraient des sinistres plus faibles dans des maisons avec moins de 4 pièces.

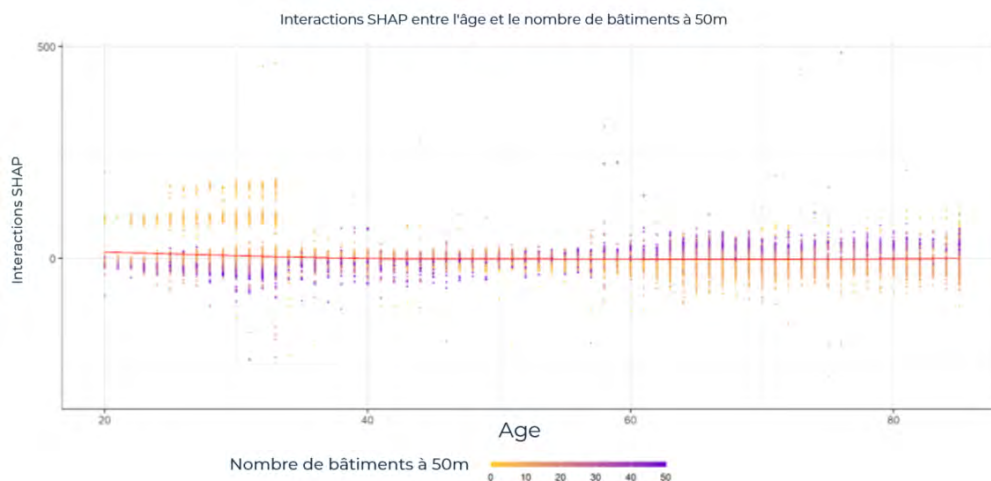


FIGURE 11.37 – Interactions SHAP entre le nombre de bâtiments dans le rayon de 50m et l'âge de l'occupant captées par le modèle de sévérité xgboost : visualisation des interactions par individu de la base de validation. L'intensité moyenne est de 10,21. Les sinistres déclarés par les jeunes dans des zones très denses seraient plus faibles de la moyenne.

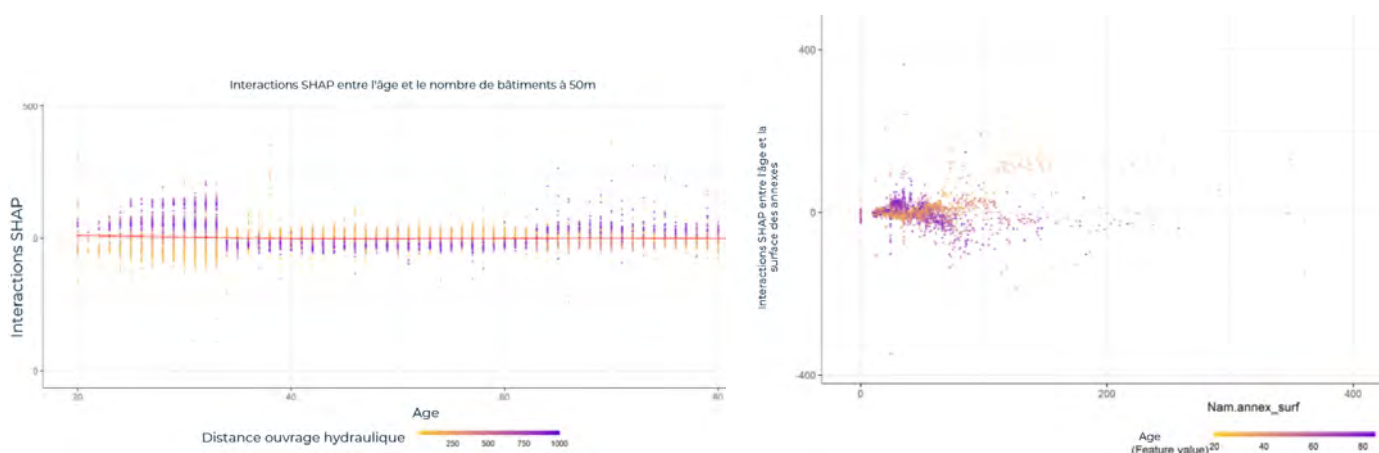


FIGURE 11.38 – Interactions SHAP entre la distance d'un ouvrage hydraulique et l'âge de l'occupant captées par le modèle de sévérité xgboost : visualisation des interactions par individu de la base de validation. Pour des profils de risques caractérisés par un logement proche d'un ouvrage hydraulique et par l'âge de l'occupant entre 35 et 65 ans, le coût moyen du sinistres augmente (interactions de signe positif).

FIGURE 11.39 – Interactions SHAP entre la surface des annexes et l'âge de l'occupant captées par le modèle de sévérité xgboost : visualisation des interactions par individu de la base de validation. On reconnaît deux comportements différents selon l'âge et la surface de la surface : les profils seniors ont un poids important sur la sévérité pour des surfaces jusqu'à 60 m², alors que les juniors sont responsables de sinistres plus élevés pour des surfaces plus grandes.

Avantages et limites de la valeur de Shapley

La valeur Shapley est plus appréciée que la méthode d'interprétation LIME puisque elle explique la différence entre la prédiction et la prédiction moyenne globale (dans notre étude la prédiction moyenne de la base de validation), tandis que LIME explique la différence entre la prédiction et une prédiction moyenne locale. Dans le cadre du RGPD et du "droit à l'explication", la valeur de Shapley est la seule méthode juridiquement conforme, car elle est basée sur une théorie solide et justifie la différence entre la prime d'un individu et la prime moyenne de manière "juste" entre les différentes variables utilisées par le modèle.

La valeur Shapley est une bonne solution pour un nombre de variable compris entre 10 et 15, toutefois elle nécessite beaucoup de temps de calcul, en grande dimension. Il n'existe pas encore une technique qui réduit la dimension à des explications sparses (explications qui contiennent peu de variables), car SHAP renvoie autant de coefficients que de variables explicatives rendant parfois l'interprétation complexe. C'est pour cette raison que nous avons envisagé de pré-sélectionner les variables.

11.3 Comparaisons

Dans cette section nous allons comparer les méthodes SHAP et Sobol : étant représentées par des quantités différentes, part de la variance totale pour Sobol, et contribution à la prédiction pour l'approche SHAP, nous nous limiterons la comparaison à la priorisation des interactions et des variables dans les modèles.

Pour obtenir un indice global des valeurs de SHAP et des interactions SHAP, nous avons considéré les SHAP Feature Importance comme définit dans la section 7.3.2 :

$$I_j = \frac{1}{k} \sum_{i=1}^k |indice^{(i)}|,$$

où $indice \in \{\text{Valeur SHAP}; \text{Interaction SHAP}\}$

Indices Principaux

Les indices principaux sont ceux qui résument la contribution d'une variable aux respectives quantités d'intérêt, sans isoler la partie due à l'interaction.

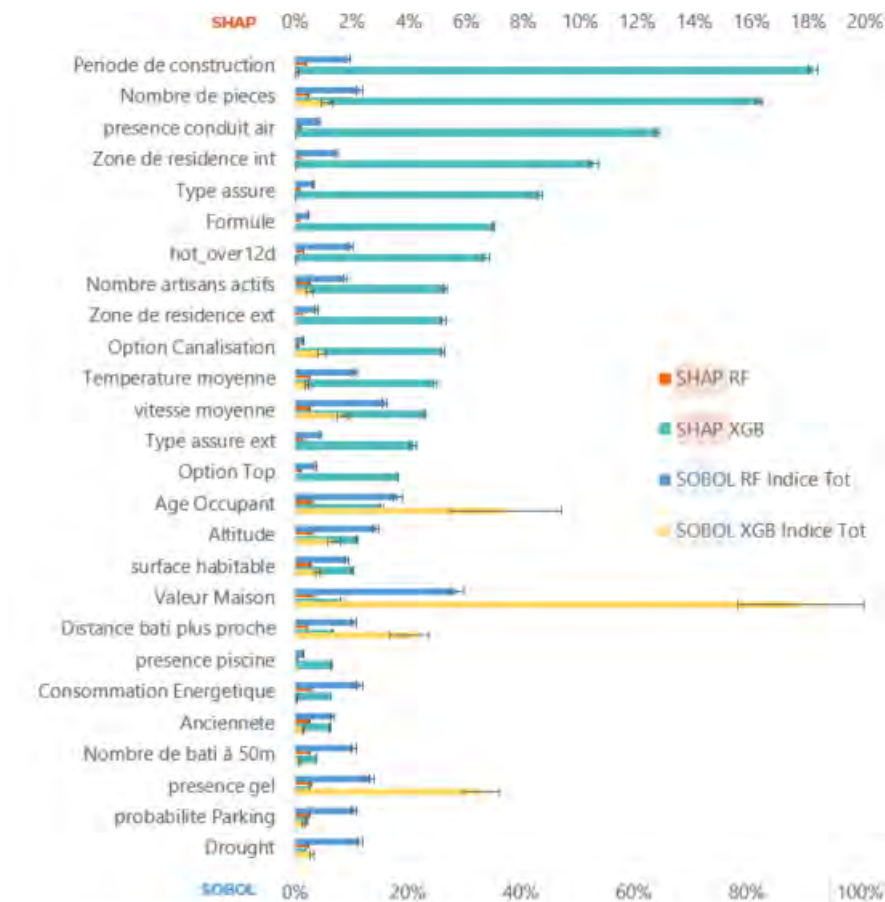


FIGURE 11.40 – Comparaison des effets principaux Sobol et SHAP pour la **fréquence**. Les intervalles de confiance au niveau 95% ont été calculés par bootstrap.

Selon l'approche SHAP, la période de construction, le nombre de pièces, la présence d'un conduit d'air, la zone de résidence et le type d'assuré sont les variables qui contribuent le plus à la prédiction du modèle xgboost ; la valeur de la maison, l'âge de l'occupant, l'altitude, la surface habitable et la consommation énergétique sont les plus importantes du modèle de forêt aléatoire (RF).

Selon l'approche de Sobol, le nombre de pièces, la valeur de la maison, l'ancienneté, la distance du bâtiment plus proche et la pluviométrie sont les variables qui contribuent le plus à la variance totale pour le modèle de forêt aléatoire, alors que la Formule, le nombre de bâtiments à 50m, le nombre de pièces, le type d'assuré et la valeur de la maison représentent environ 53% de la variance totale du modèle xgboost.



FIGURE 11.41 – Comparaison des effets principaux Sobol et SHAP pour la **sévérité**. Les intervalles de confiance au niveau 95% ont été calculés par bootstrap.

Selon l’approche **SHAP**, la Formule, le nombre de pièces, la valeur de la maison, la consommation énergétique et l’âge de l’occupant sont les variables qui contribuent le plus à la prédiction du modèle xgboost ; le nombre de pièces, valeur de la maison, le nombre de bâtiments à 50m et le type d’assuré sont les plus importantes du modèle de forêt aléatoire (RF). Selon l’approche de **Sobol**, la valeur de la maison, l’âge de l’occupant, la présence de gel et la vitesse moyenne du vent sont les variables qui contribuent le plus à la variance totale pour les deux modèles.

Indices d’interactions

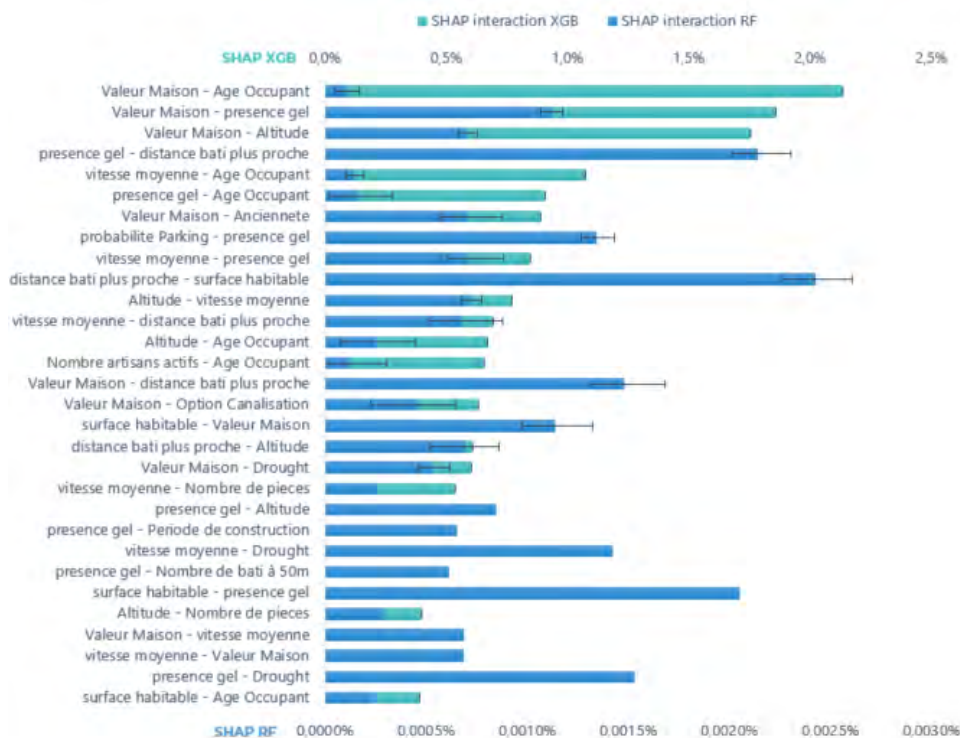


FIGURE 11.42 – Comparaison des effets d’interactions SHAP pour la fréquence : le modèle xgboost a des valeurs beaucoup plus importantes que le modèle de forêt aléatoire et dans la prédiction il prend en compte les interactions de la valeur de la maison avec l’âge, la présence du gel, l’altitude, ainsi que des interactions moins fortes. Le modèle *random forest* quant à lui prend en compte dans la prédiction les interaction de la distance du bâtiment plus proche et de la surface habitable.

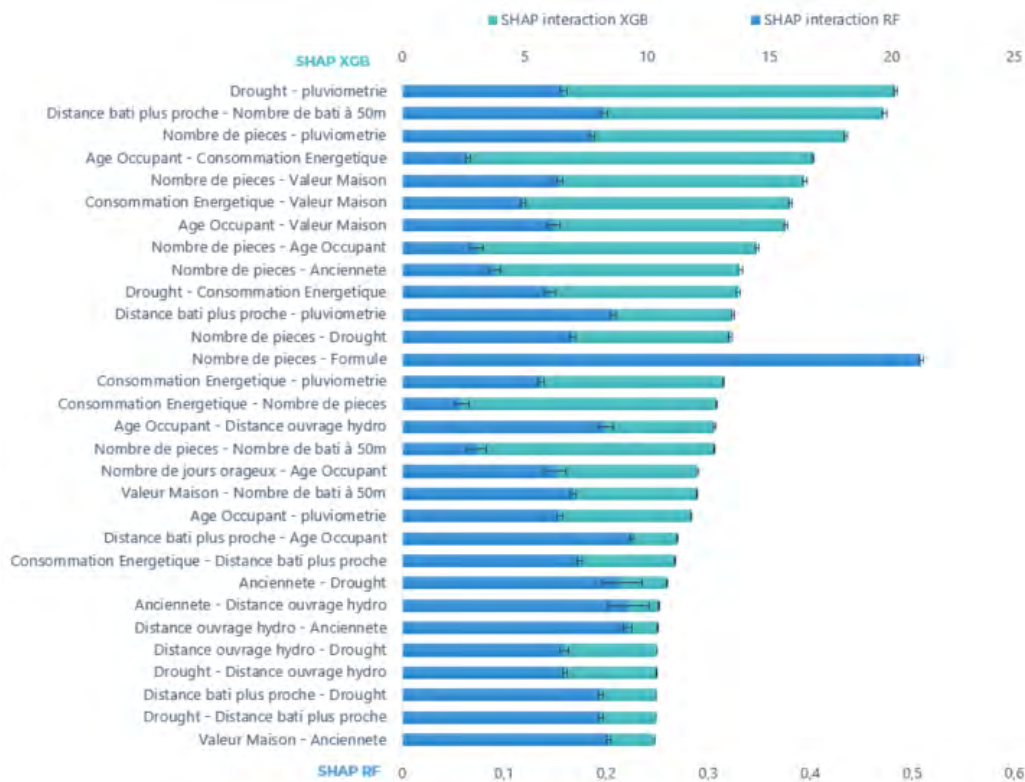


FIGURE 11.43 – Comparaison des effets d’interactions SHAP pour la sévérité : les interactions de forêt aléatoire sont très faibles (moyenne 0,17) par rapport à celles intégrées par xgboost (moyenne à 5,9). La prédiction xgboost prend en compte les interactions de la valeur de la maison avec l’âge, le nombre de pièces, la consommation énergétique ainsi que des interactions entre variables climatiques (pluviométrie et sécheresse).

Nous regardons enfin le top 15 des interactions ordonnées selon l’indicateur :

$$(\text{Interactions SHAP}_{RF} + \text{Interactions SHAP}_{XGB}) * (\text{Interactions Sobol}_{RF} + \text{Interactions Sobol}_{XGB})$$

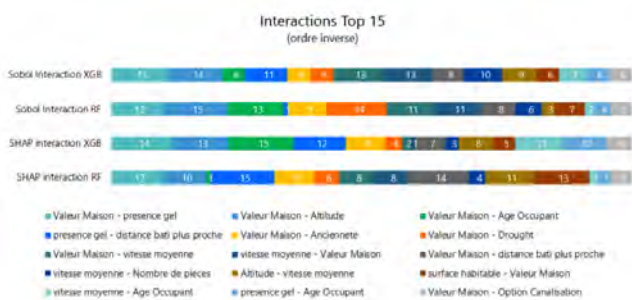


FIGURE 11.44 – TOP 15 des interactions des modèles de fréquence : l’ampleur de l’intervalle est proportionnel à l’interaction. Les interactions de la valeur de la maison sont les plus importants.

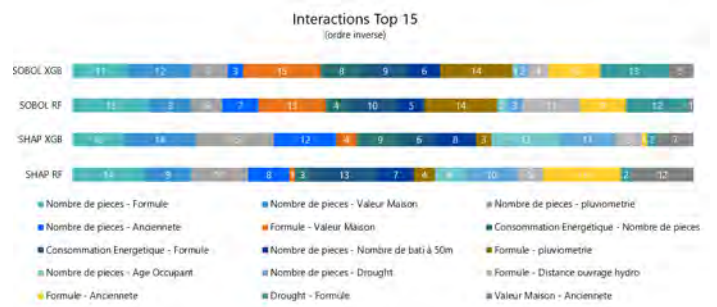


FIGURE 11.45 – TOP 15 des interactions des modèles de sévérité : l’ampleur de l’intervalle est proportionnel à l’interaction. Les interactions de la formule sont les plus importants.

Chapitre 12

Intégration des interactions dans le modèle de prime pure : résultats et ouvertures

Le chapitre suivant est consacré à la dernière étape de notre méthodologie d'optimisation tarifaire : l'intégration des interactions dans le modèle de prime pure "Benchmark", à savoir le GLM simple.

Plus généralement, cette méthodologie de détection ne se limite pas à l'optimisation tarifaire : son agnosticité aux modèles permet de l'appliquer en toute généralité à n'importe quel modèle, complexe ou non. Ce parcours de plus a fait expérimenter une nouvelle approche qui voit le Machine Learning et les modèles linéaires généralisés en collaboration : dans la littérature actuarielle, on a tendance à comparer la performance des modèles plus complexes au GLM classique, ou à se servir du machine learning ou du GLM de manière séparée.

Ainsi, les modèles plus sophistiqués recouvrent un rôle principal dans la tarification en assurance Non-Vie, étant les termes d'interaction une expression de la complexité de ces modèles.

12.1 Ajout d'interactions

Les interactions détectées avec l'analyse de sensibilité selon Sobol et selon SHAP devront être ajoutées au GLM de départ, ici appelé "Benchmark". Nous ajouterons aussi celles introduites par les partitions CART.

La détection ne fournit pas de détails précis sur la nature de la relation entre deux variables : les interactions SHAP suggèrent par exemple des fonctions linéaires par morceaux (fig. 11.28), sinusoïdales (fig. 11.38) ou racine n -ième avec n entier (fig. 11.25), mais en général il est assez compliqué réaliser une analyse rigoureuse en grande dimension.

Les interactions locales, faisant référence à l'appartenance à des intervalles spécifiques des domaines de définition de deux variables, conduisent naturellement à une segmentation de l'espace engendré par ces deux variables.

Les interactions globales, au contraire, représentent un comportement sur tout le domaine de définition des variables. Afin de restreindre les possibilités de complexité à une classe de fonctions, un choix naturel serait d'utiliser par exemple les fonctions polynomiales, très simples à implémenter.

Nous avons ainsi ajouté les interactions de la façon suivante :

- pour chaque interaction locale (SHAP et CART), nous avons établi plusieurs seuils sur les variables en interaction X_i et X_j , et nous avons ajouté à la formule du GLM les produits de deux fonctions indicatrices $I_{X_i \in I_k} * I_{X_j \in I_l}$, où I_k et I_l sont les intervalles des domaines des variables définis par les seuils.

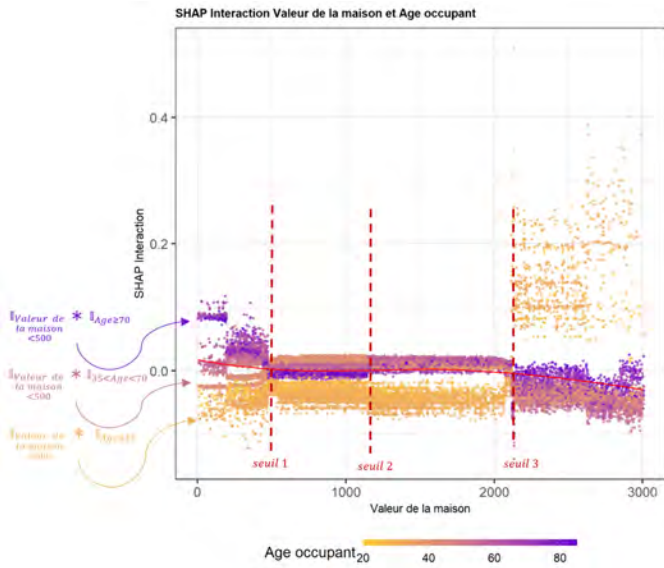


FIGURE 12.1 – Création des termes d'interactions pour le GLM à partir des graphes d'interactions SHAP.

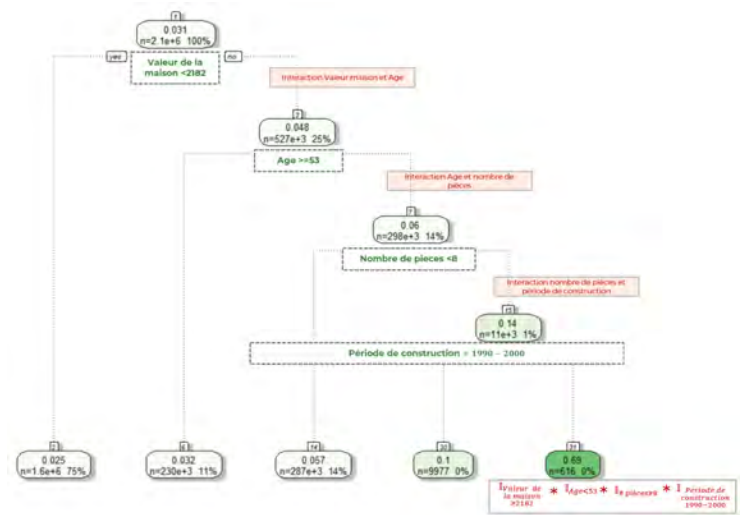


FIGURE 12.2 – Création des termes d'interactions pour le GLM à partir de la partition CART.

- pour chaque interaction globale (Sobol) entre les variables X_i et X_j , nous avons ajouté le polynôme du deuxième ordre $X_i * X_j + X_i^2 + X_j^2$ à la formule de départ.

Avec des outils plus sophistiqués (*smoothing* ou fonction spline), il est possible de lisser les fonctions de prédiction.

Significativité des ajouts

À l'aide du test du rapport de vraisemblance, nous montrons la significativité de l'ajout des termes d'interaction.

En particulier, le terme ajouté est jugé significatif si on rejette l'hypothèse H_0 "nullité du coefficient de régression", c'est-à-dire lorsque la statistique de test (ici la différence de déviance du Benchmark et le nouveau modèle) dépasse le quantile d'ordre $1-\alpha$ de la loi de Fisher. La probabilité de dépassement de ce seuil (nous avons utilisé $\alpha = 5\%$), est dite probabilité critique.

Les interactions ajoutées au GLM où la probabilité critique selon le test du rapport de vraisemblance était supérieure au seuil 5% n'ont pas été retenues. Nous montrons ci dessous les interactions retenues les plus importantes au sens du gain dans la réduction de la déviance : l'apport d'un terme dans un modèle est mesuré par la réduction que ce terme induit dans la déviance résiduelle.

| Modele | Analyse de sensibilité | Metamodelle | Interaction ajoutée dans le GLM | Rapport de vraisemblance | | Gain dans la réduction de la déviance | Probabilité critique | intensité globale |
|--------|------------------------|-------------|---|--------------------------|--------------|---------------------------------------|----------------------|-------------------|
| | | | | | Deviance res | | | |
| Sev | SHAP | XGB | Toutes les interactions | 150,0809 | 28 590 | 0,5% | 0,000000% | |
| Sev | GLOUTON | CART | Toutes les interactions | 125,17542 | 28 615 | 0,4% | 0,000000% | |
| Sev | GLOUTON | CART | "nombre de pieces >=6" & "2424<Valeur de la maison <=2439" & "distance ouvrage hydro >=174" | 119,994235 | 28 620 | 0,4% | 0,000000% | |
| Sev | Sobol | XGB | Toutes les interactions | 81,7346 | 28 658 | 0,3% | 0,000000% | |
| Sev | Sobol | RF | Toutes les interactions | 76,88228 | 28 663 | 0,3% | 0,000000% | |
| Sev | SHAP | XGB | Nb de batiments à 50m - Pluviometrie | 47,349845 | 28 693 | 0,2% | 0,000000% | 8,59714747 |
| Sev | Sobol | RF | Nb de jours orageux - Nombre de pieces | 32,063196 | 28 708 | 0,1% | 0,000001% | 0,16296066 |
| Sev | Sobol | XGB | Type d'assurance - Anciennete | 29,10527 | 28 711 | 0,1% | 0,000007% | 0,24480651 |
| Sev | SHAP | XGB | Drought - Anciennete | 27,907026 | 28 712 | 0,1% | 0,000013% | 10,7860801 |
| Sev | Sobol | XGB | Type d'assurance - Nombre de pieces | 26,25267 | 28 714 | 0,1% | 0,000030% | 0,67809648 |
| Sev | Sobol | RF | Nombre de pieces - Type d'assurance | 26,252666 | 28 714 | 0,1% | 0,000030% | 0,18717685 |

FIGURE 12.3 – Interactions ajoutées au GLM de la sévérité par ordre de réduction de la déviance résiduelle : uniquement les interactions où la probabilité critique $< 5\%$ ont été retenues. La notation "Toutes les interactions" prend en compte toutes les interactions selon le modèle et l'approche utilisés retenues significatives. L'intensité globale est soit l'indice de Sobol du deuxième ordre soit la moyenne des interactions SHAP en valeur absolue.

Ici, l'ajout de toutes les interactions permet de gagner entre 0.1% et 0.5% de la déviance résiduelle. En particulier l'interaction entre le *nombre de bâtiments dans un rayon de 50m* et la *pluviométrie* et celle entre le *nombre de jours orageux* et le *nombre de pièces* sont les interactions qui font le plus baisser la déviance résiduelle du modèle.

| Modele | Analyse de sensibilité | Metamodelle | Interaction ajoutée dans le GLM | Gain dans la réduction de la déviance | Probabilité critique | intensité globale |
|-----------|------------------------|-------------|--|---------------------------------------|----------------------|-------------------|
| Frequence | SHAP | XGB | Toutes les interactions | 0,5% | 0,000000% | |
| Frequence | GLOUTON | CART | Toutes les interactions | 0,3% | 0,000000% | |
| Frequence | GLOUTON | CART | "Valeur de la maison>=2182"&"Age occupant<53"&"nombre de pieces >=8"&"periode de construction <= "1990-2000" | 0,3% | 0,000000% | |
| Frequence | Sobol | RF | Toutes les interactions | 0,1% | 0,000000% | |
| Frequence | SHAP | RF | Toutes les interactions | 0,1% | 0,000000% | |
| Frequence | GLOUTON | CART | "Valeur de la maison>=2182"&"Age occupant<53"&"nombre de pieces <8" | 0,1% | 0,000000% | |
| Frequence | GLOUTON | CART | (Valeur de la maison>=2182)*!(Age occupant>=53) | 0,1% | 0,000000% | |
| Frequence | SHAP | XGB | Age occupant - Valeur Maison | 0,1% | 0,000000% | 2,13% |
| Frequence | Sobol | XGB | Toutes les interactions | 0,1% | 0,000000% | |
| Frequence | Sobol | RF | Valeur Maison - Anciennete | 0,1% | 0,000000% | 1,98% |
| Frequence | Sobol | RF | Valeur Maison - Nb artisans actifs | 0,1% | 0,000000% | 1,47% |
| Frequence | Sobol | RF | Valeur Maison - Temperature moyenne | 0,1% | 0,000000% | 2,25% |
| Frequence | GLOUTON | CART | Valeur de la maison<2182 | 0,1% | 0,000000% | |

FIGURE 12.4 – Interactions ajoutées au GLM de la fréquence par ordre de réduction de la déviance résiduelle (détails en annexe : uniquement les interactions où la probabilité critique < 5% ont été retenues. La notation "Toutes les interactions" prend en compte toutes les interactions selon le modèle et l'approche utilisés retenues significatives). L'intensité globale est soit l'indice de Sobol du deuxième ordre soit la moyenne des interactions SHAP en valeur absolue. Ici, l'ajout de toutes les interactions permet de gagner entre 0.1% et 0.5% de la déviance résiduelle. En particulier l'interaction entre les variables *Valeur de la maison*, *l'âge de l'occupant*, *le nombre de pièces* et *la période de construction* du modèle CART est celle qui fait baisser le plus la déviance résiduelle du modèle.

12.2 Comparaison entre le GLM simple et le GLM avec interactions

12.2.1 Comparaison des métriques

Nous avons comparé plusieurs métriques de performance de modèles sans et avec l'ajout des interactions :

- le rapport des écarts à la moyenne observé : $\frac{|\bar{Y} - \hat{Y}_m|}{|\bar{Y} - \hat{Y}_{Benchmark}|}$, où \hat{Y}_m est la prédiction moyenne du modèle m et \bar{Y} est la moyenne des observations ;
- les AIC, MSE, RMSE et MAE ;
- le coefficient de prédictivité Q2 du modèle : $Q_2 = 1 - \frac{\sum_{i=1}^n [Y_i - \hat{Y}_i]^2}{\sum_{i=1}^n (Y - Y_i)^2}$ où $(Y_i)_i$ sont les observations de la base de test de taille n , \bar{Y} est la moyenne des Y_i , \hat{Y}_i est la valeur prédite. Q2 correspond au R2 calculé sur une base de test ;
- le coefficient de Gini est une métrique utilisée pour évaluer le pouvoir discriminatoire d'un modèle, à savoir entre les clients à haut risque de sinistralité et les clients à faible risque de sinistralité. Cet indice est souvent utilisé pour comparer la qualité de plusieurs modèles et évaluer leur pouvoir de prédiction. Dans un modèle à pouvoir discriminant très élevé, le coefficient de Gini s'approche de 100%.
- la déviance poissonnienne ou gamma et la déviance résiduelle.

| Fréquence | | | | | | | | | |
|-------------------------------|----------------|---------|---------|---------|--------|--------|--------|------------------|-------------------|
| Base 100 | | | | | | | | | |
| Modele | Ecart Pred avg | MSE | RMSE | MAE | Q2 | GINI | AIC | Deviance Poisson | Residual Deviance |
| Benchmark | 100,00 | 100,000 | 100,000 | 100,000 | 100,00 | 100,00 | 100,00 | 100,000 | 100,00 |
| Clouton CART | 94,25 | 99,978 | 99,989 | 100,065 | 102,74 | 100,20 | 99,93 | 99,962 | 99,91 |
| SHAP RF | 93,29 | 99,993 | 99,996 | 100,076 | 100,87 | 100,23 | 99,97 | 99,988 | 99,91 |
| SHAP XGB | 79,07 | 99,983 | 99,992 | 100,252 | 102,06 | 100,25 | 99,90 | 99,982 | 99,83 |
| SHAP Toutes les interactions | 78,89 | 99,984 | 99,992 | 100,252 | 101,98 | 100,29 | 99,90 | 99,981 | 99,79 |
| Sobol RF | 87,20 | 99,987 | 99,993 | 100,160 | 101,62 | 99,55 | 99,93 | 100,017 | 99,87 |
| Sobol XGB | 98,22 | 99,989 | 99,995 | 100,021 | 101,30 | 99,97 | 99,93 | 99,999 | 99,91 |
| Sobol Toutes les interactions | 86,66 | 99,987 | 99,994 | 100,167 | 101,55 | 99,50 | 99,93 | 100,014 | 99,87 |
| Toutes les interactions | 79,37 | 99,987 | 99,994 | 100,251 | 101,52 | 100,20 | 99,9 | 100,026 | 99,74 |

FIGURE 12.5 – Parmi les modèles de fréquence construits, celui contenant toutes les interactions SHAP sur l'ensemble de la base de test se rapproche le plus à la moyenne observée et il est le plus discriminant (Gini plus élevé). Le modèle dérivé du CART se rapproche de plus en moyenne aux observations (MSE plus faible) et il est le plus prédictif (Q2 plus élevé).

| Sévérité | | | | | | | | | |
|------------------|----------------|--------|--------|---------|--------|--------|--------|----------------|-------------------|
| Base 100 | | | | | | | | | |
| Modele | Ecart Pred avg | MSE | RMSE | MAE | Q2 | GINI | AIC | Deviance Gamma | Residual Deviance |
| Benchmark | 100,00 | 100,00 | 100,00 | 100,000 | 100,00 | 100,00 | 100,00 | 100,00 | 100,00 |
| Glouton CART | 98,07 | 98,994 | 98,997 | 98,953 | 100,55 | 102,54 | 99,98 | 100,58 | 99,52 |
| SHAP RF | | | | | | | | | |
| SHAP XGB | 98,89 | 98,998 | 98,999 | 100,131 | 100,93 | 110,73 | 99,98 | 100,98 | 99,30 |
| SHAP Toutes les | 98,89 | 98,998 | 98,999 | 100,131 | 100,93 | 110,73 | 99,98 | 100,98 | 99,30 |
| Sobol RF | 97,54 | 98,999 | 98,999 | 100,041 | 101,05 | 103,25 | 100,00 | 100,32 | 99,70 |
| Sobol XGB | 99,79 | 98,999 | 98,999 | 100,172 | 101,05 | 101,72 | 100,00 | 100,51 | 99,70 |
| Sobol Toutes les | 97,50 | 98,999 | 99,000 | 100,029 | 101,07 | 103,82 | 100,00 | 100,30 | 99,70 |
| All | 96,21 | 98,990 | 98,993 | 100,050 | 100,15 | 117,08 | 99,95 | 102,00 | 98,56 |

FIGURE 12.6 – Les modèles de sévérité avec les interactions sont moins performants que ceux de fréquence : les interactions détectées pourraient avoir une nature plus complexe que les polynômes du deuxième degré ou ne pas être assez localisées. Le modèle avec toutes les interactions prédit une moyenne sur le portefeuille plus proche à celle réelle que le Benchmark, il est le plus discriminant et réduit l'AIC.

Nous avons combiné les neuf modèles de fréquence et les huit modèles de sévérité pour le calcul de la prime pure :

| Prime pure | | | | | | |
|---------------------------------------|----------------|----------|----------|----------|----------|----------|
| Base 100 | | | | | | |
| Modele | Ecart Pred avg | MSE | RMSE | MAE | Q2 | GINI |
| NB Benchmark_Sev Benchmark | 100,00 | 100,0000 | 100,00 | 100,00 | 100,000 | 100,000 |
| NB glm_shap_rf_Sev ALL Interactions | 101,35 | 99,9812 | 99,9906 | 99,9834 | 99,9305 | 102,5057 |
| NB glm_shap_xgb_Sev ALL Interactions | 101,57 | 99,9796 | 99,9898 | 99,9782 | 99,9246 | 101,8951 |
| NB Benchmark_Sev glm_glouton_CART | 99,63 | 100,0019 | 100,0010 | 100,0043 | 100,0070 | 100,2264 |
| NB glm_shap_xgb_Sev glm_sobol_xgb | 101,50 | 99,9869 | 99,9934 | 99,9780 | 99,9513 | 100,0699 |
| NB glm_sobol_all_Sev glm_glouton_CART | 100,06 | 100,0070 | 100,0035 | 99,9990 | 100,0259 | 98,8819 |

FIGURE 12.7 – Modèles de prime pure à comparaison (les six premiers) en base 100. Le modèle le plus discriminant est celui ayant l'indice de Gini le plus fort, à savoir le modèle avec les interactions de fréquence SHAP de la forêt aléatoire et toutes les interactions de la sévérité. Le modèle avec les interactions SHAP xgboost pour la fréquence et toutes les interactions de sévérité ou celle de Sobol de la forêt aléatoire est celui qui réduit l'erreur. L'ajout de toutes les interactions de fréquence et les interactions CART rend le modèle final le plus prédictif.

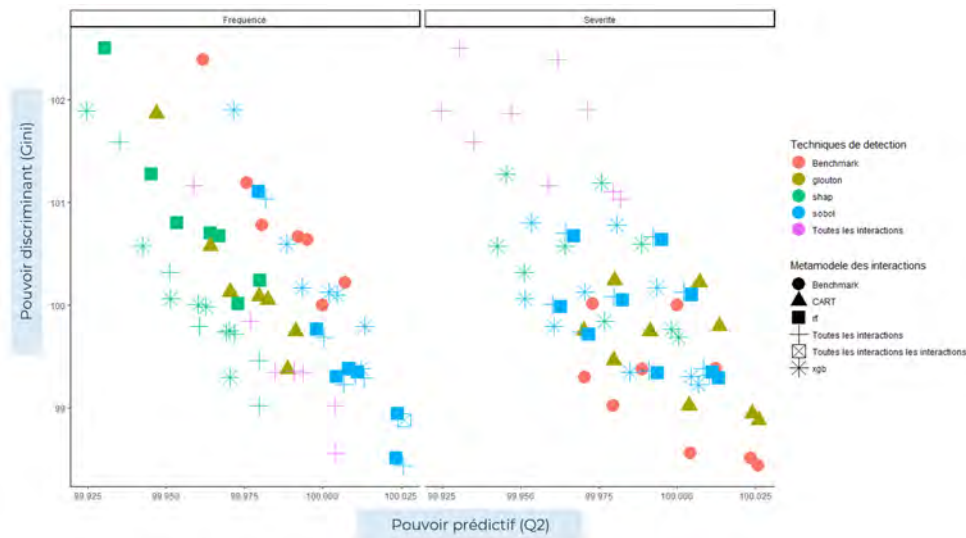


FIGURE 12.8 – A l’aide d’un nuage de points nous visualisons les modèles de prime pure selon leurs caractéristiques sur les modèles de fréquence et sévérité et leurs performances : les modèles souhaitables se positionnent sur la première bissectrice du plan engendré par le pouvoir prédictif (Q2) et le pouvoir discriminant (Indice de Gini). On remarque que pour avoir une première amélioration du modèle de prime pure, il suffirait d’ajouter uniquement toutes les interactions sur la sévérité (point en rouge dans le graphe de gauche, ou point en violet dans celui de droite).

12.2.1.1 Surfaces de réponse des modèles utilisés

Afin de visualiser l’ajout de la complexité au modèle de départ, nous avons utilisé la méthode des surfaces de réponses (MSR) qui a pour but d’explorer les relations entre les variables dépendantes et indépendantes impliquées dans une expérience¹. Il s’agit de représenter les prédictions avec un plan d’expérience. Pour mieux visualiser l’apport des interactions au modèle, nous avons simulé un plan à deux variables, l’âge et la valeur de la maison car ils interagissent, et représenté la surface de réponse à partir des modèles de fréquence.

Les modèles qui dérivent des interactions SHAP et CART segmentent le domaine produit des variables, alors que les interactions de Sobol ont été ajoutées comme des polynômes du deuxième degré.

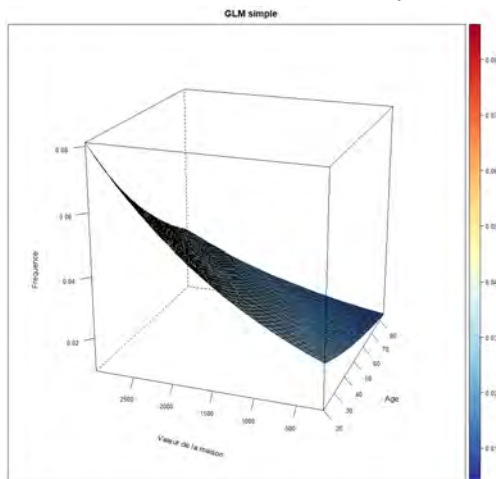


FIGURE 12.9 – Visualisation de la surface de réponse du GLM simple (Benchmark). La variation de la prédiction Y en fonction de l’âge est la fonction $\frac{\partial Y}{\partial X_{Age}} = \beta_{Age} \exp(\beta_{Age} X_{Age}) * \exp(\beta_{Valeur\ de\ la\ maison} X_{Valeur\ de\ la\ maison})$.

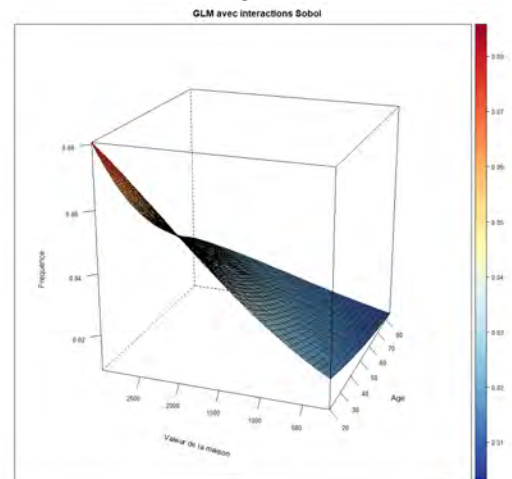


FIGURE 12.10 – Visualisation de la surface de réponse du GLM avec les interactions Sobol : nous avons ajouté le polynôme du deuxième degré $X_{Age} * X_{Valeur\ de\ la\ maison} + X_{Age}^2 + X_{Valeur\ de\ la\ maison}^2$

¹Elle est due aux travaux de 1951 de George Box et K. B. Wilson.

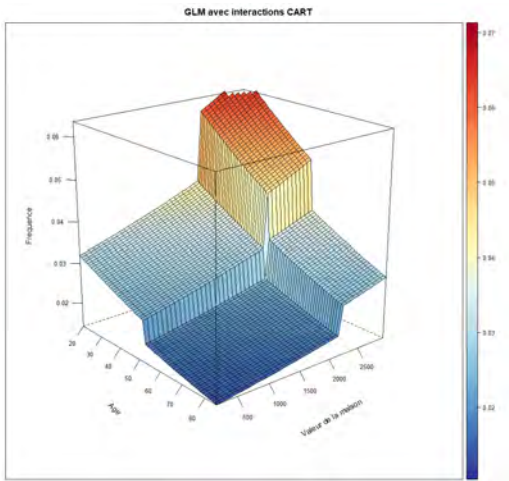


FIGURE 12.11 – Visualisation de la surface de réponse du GLM avec les interactions CART. Si l'on fixe la variable de la valeur de la maison l'effet de l'âge sur la prédiction n'est pas constant, mais il dépend de la valeur de la maison. Cet ajout crée une partition du domaine de l'âge et de la valeur de la maison.

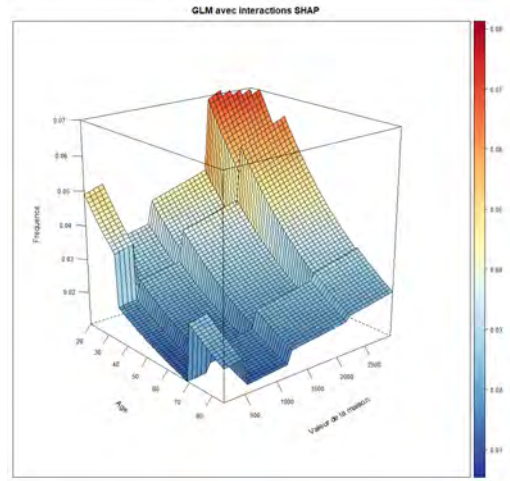


FIGURE 12.12 – Visualisation de la surface de réponse du GLM avec les interactions SHAP. Si l'on fixe la variable de la valeur de la maison l'effet de l'âge sur la prédiction n'est pas constant, mais il dépend de la valeur de la maison. Cet ajout crée une partition du domaine de l'âge et de la valeur de la maison plus finement que les ajouts du CART.

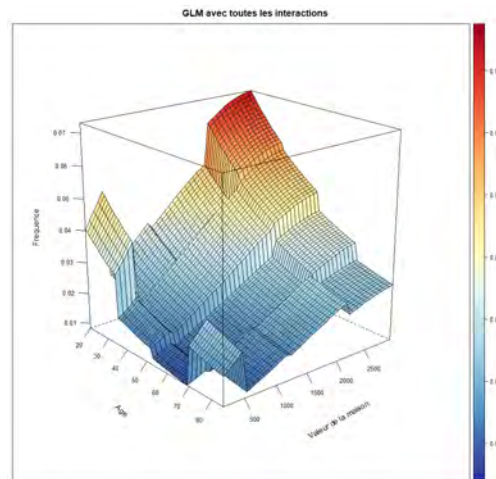


FIGURE 12.13 – Visualisation de la surface de réponse du GLM avec toutes les interactions : la fonction de lien donne une structure exponentielle au modèle. Les termes des interactions de Sobol confèrent une complexité avec l'ajout d'un polynôme du deuxième degré à la formule tarifaire. Enfin les termes des interactions SHAP et CART partitionnent "justement" (par rapport au partage équitable des importances de Shapley) le domaine de définition des variables. Par rapport au modèle de Benchmark, le risque est plus segmenté.

12.2.1.2 Comparaison du risque géographique

Avec les termes ajoutés, nous voulons, en dernière analyse, visualiser l'effet spatial : en particulier nous visualiserons où la prédiction change comparativement au GLM simple.

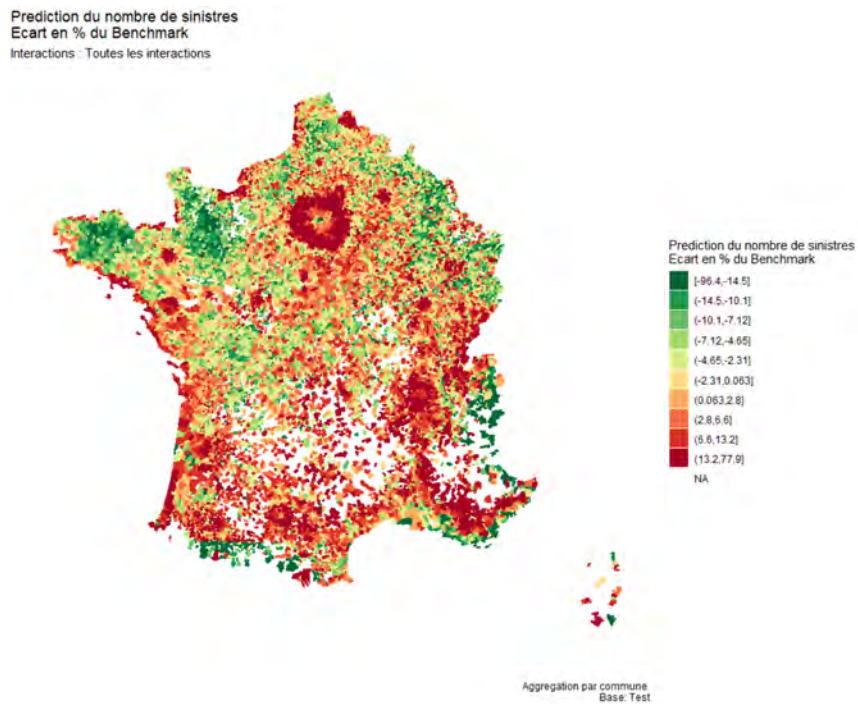


FIGURE 12.14 – Écart entre les prédictions moyennes par commune du GLM simple et celle du GLM avec interactions : les communes plus impactées par l’ajout des interactions sont Saint-Tropez (96%) et les communes de la couronne parisienne.

12.2.2 Arbre de décision

Le choix d’un modèle de tarification dépend de plusieurs critères : *techniques* sur la qualité du modèle, *opérationnels* selon la mise en place dans son industrialisation, et enfin d’*interprétabilité*.

L’étude a cherché à unifier ces critères à l’aide des interactions statistiques, qui ajoutent de la complexité et de l’interprétabilité aux modèles, tout en gardant une structure simple.

Nous avons classifié les modèles selon leurs trajectoires de tarification, répondant à ces critères.

Notations :

- RF : modèle de forêt aléatoire
- XGB : modèle xgboost
- SHAP RF : GLM avec ajout des interactions SHAP du modèle de forêt aléatoire
- SHAP XGB : GLM avec ajout des interactions SHAP du modèle xgboost
- Sobol RF : GLM avec ajout des interactions Sobol du modèle de forêt aléatoire
- Sobol XGB : GLM avec ajout des interactions Sobol du modèle xgboost

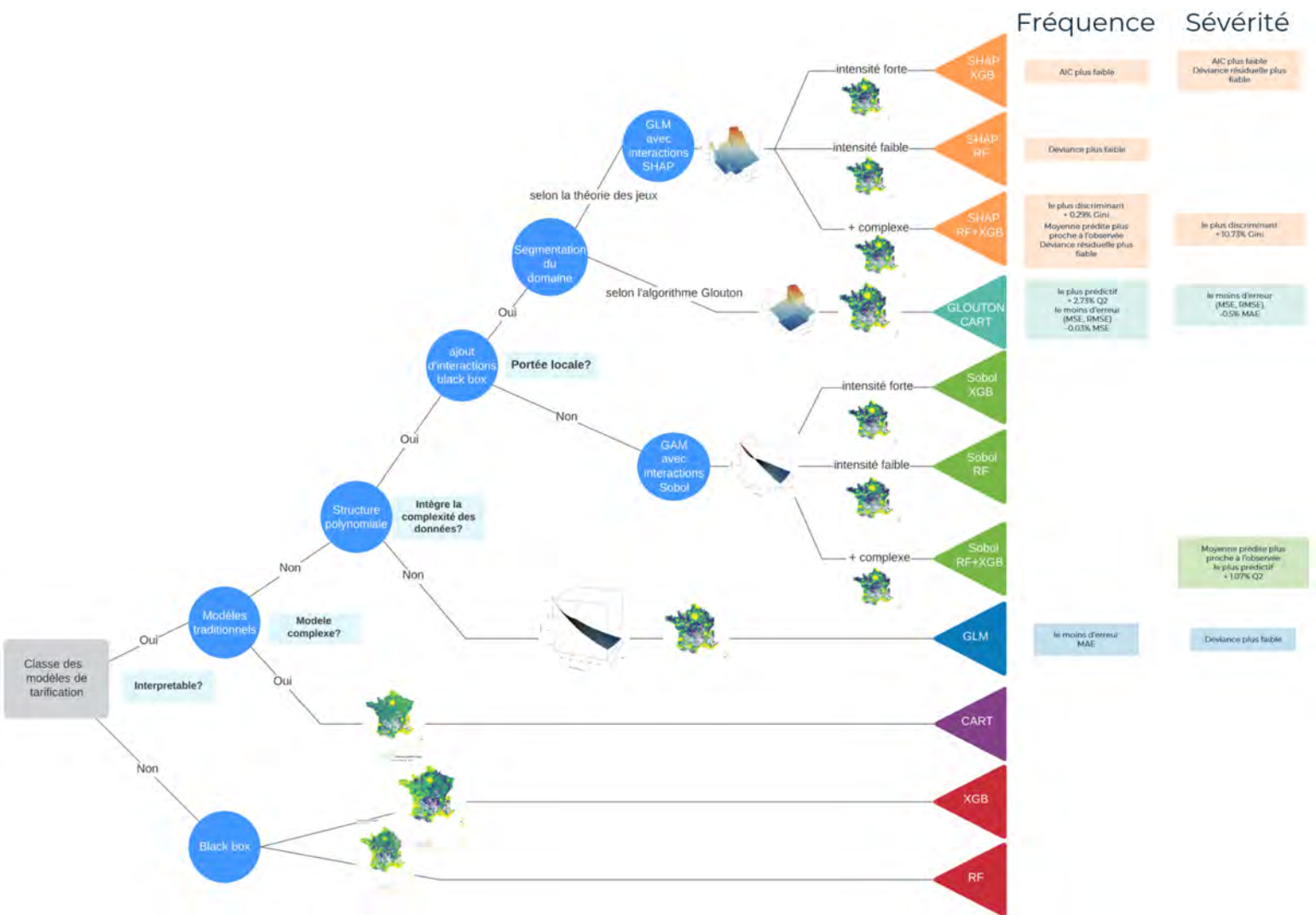


FIGURE 12.15 – Trajectoires des modèles de fréquence et sévérité selon les critères d'interprétabilité, complexité du modèle, prise en compte de la complexité des données, prise en compte de la mise en place opérationnelle et des métriques de performance.

Cinquième partie

Conclusion

Chapitre 13

Conclusion

Dans le cadre de cette étude, une analyse de sensibilité a été menée sur des données à l'adresse pour améliorer le modèle de tarification de la garantie *Dégâts des eaux*.

La base des données utilisée contient environ 300 variables selon les informations suivantes :

- Contrat : numéro de contrat, année de souscription, numéro d'image, date d'effet, ...
- Caractéristiques de l'assuré : âge de l'occupant, ancienneté, status professionnel, type de personne (propriétaire ou locataire ou PNO), ...
- Antécédents : nombre de sinistres, exposition en années d'assurances, charge totale ;
- Caractéristiques bâtiment : surface habitable, numéro d'étages, présence panneaux solaires, consommation énergétique, ...
- Caractéristiques de l'adresse : Code IRIS, Code INSEE, nombre de bâtiments dans le rayon de 50 mètres, altitude, distance du bâtiment le plus proche, ...
- Données climatiques : nombre de jours de gel, nombre de jours orageux, quantité de pluie moyenne par an, vitesse du vent, ...
- Données économiques : valeur immobilière de la maison, nombre d'agriculteurs par IRIS, nombre d'artisans par IRIS, ...

Afin de manipuler des estimateurs robustes par la suite, le nombre de variables a été réduit à environ quarante, à l'aide des techniques de présélection non supervisée (Matrice de corrélation, ACP, ACM) et supervisée (forêt aléatoire).

L'intérêt principal des techniques d'analyse de sensibilité dans la modélisation de la prime pure est de pouvoir ajouter de la *complexité* au modèle initiale tout en gardant une structure analytique, transparente et interprétable qui s'intègre parfaitement au processus de tarification traditionnel des organismes d'assurance.

Afin de respecter ces critères, nous avons reconduit le problème d'optimisation tarifaire au problème de détection et intégration des interactions parmi les variables, l'interaction étant une expression de la complexité du modèle.

On a d'abord détecté les interactions des modèles les plus sophistiqués, dits de type *boîte noire* à cause de leur structure prédictive non accessible, en nous appuyant sur des concepts de la théorie de jeux et de l'analyse de sensibilité selon Sobol. Ces domaines ont fourni un compromis entre la rapidité de calcul (parallélisation des tâches, passage par des métamodèles), la complexité (des modèles et des données), l'interprétabilité (des modèles de tarification) et l'agnosticité aux modèles. Les outils de l'analyse de sensibilité tels que les indices de Sobol et les indices de SHAP (une relecture plus récente de la théorie des jeux) ont été choisis par les bonnes propriétés de convergence de leurs estimateurs et par le type d'information apportée.

D'un côté, les indices de Sobol de l'ordre deux informent de la part de la variance totale due à chacune des interactions, de l'autre les indices d'interaction SHAP déterminent dans quelle direction l'interaction a un impact sur la prédiction (si elle est à la hausse ou à la baisse à cause de l'interaction) et selon quelle intensité.

Ensuite nous avons intégré les interactions détectées localement et globalement en ajoutant des termes polynomiaux dans le modèle GLM de départ.

Ces termes améliorent le modèle GLM simple selon des métriques d'évaluation habituelles (MSE, RMSE, MAE, Q2, Gini, déviance, AIC) avec un gain entre 0.03% et 17%.

Par ailleurs, la perception du risque géographique héritée par les modèles plus complexes et par les données à

maille adresse ou bâtiment est plus fine.

En effet, vivant la base de données dans la grande dimension, l'analyse bivariée pour examiner toutes les interactions possibles est fastidieuse et lente, car il faudrait visualiser et ajouter manuellement chaque terme croisé.

De plus, l'inclusion des variables géographiques à maille très fine challenge les méthodes de zonage actuelles pour la prise en compte du risque spatial.

Limites de l'étude

Les **limites** principales de cette étude sont les suivantes :

- la **nature de la complexité des données et des modèles**. Les estimateurs de SHAP et des indices de Sobol ont permis de détecter les interactions parmi les variables et nous avons intégré ces relations à l'intérieur de la structure d'un GLM simple en ajoutant des termes polynomiaux. Toutefois, la nature de ces liens peut être plus complexe (l'ajout des fonctions indicatrices n'est pas suffisant) ou très localisée (par opposition aux indices de Sobol qui sont globaux) : dans ce cas, l'ajout des termes peut se révéler non significatif selon le test du rapport de vraisemblance. Le lissage avec les techniques de *smoothing* ou *spline* pourrait améliorer la performance des modèles.
- Le **risque de propagation de l'erreur**. La construction des métamodèles de type CART, xgboost et forêt aléatoire peut propager son erreur de prédiction à l'estimation des indices de Sobol ou de SHAP et introduire des interactions artificielles, spécialement en présence de colinéarité. Pour limiter leur intégration dans le modèle de départ, il faut valider les interactions avant de les exploiter dans le modèle. De plus, la dimension des données peut ralentir la convergence de ces estimateurs et dans le cas où l'importance est très petite les estimations peuvent être négatives (pour les indices Sobol).

La théorie liée à la valeur SHAP qui se fonde sur l'explicabilité du modèle n'est pas nouvelle dans secteur de l'assurance : c'est aussi un outil de pilotage tarifaire, de compréhension du portefeuille et de prévention au risque de sinistralité. Les références sous les logiciels **R** et **python** sont par ailleurs abondantes et sophistiquées.

Son extension à la détection d'interactions est un des **éléments innovants** que l'on a voulu transmettre dans ce mémoire.

L'analyse de Sobol, quant à elle, a été un autre challenge pour ce type de problématique : à la différence des contextes où elle s'applique d'habitude (ingenierie aeronautique, finance,...), où l'on *pilote* une base, nous disposons des réalisations des variables sans connaître leur distributions et on a plutôt *subit* une base.

Nous nous sommes restreint à la sinistralité attritionnelle, mais dans la continuité des études d'analyse de sensibilité, il serait intéressant de calculer ces indices sur des quantiles différents et intégrer ainsi les interactions sur les **sinistres graves**.

Executive Summary

Executive Summary

Introduction

The popularity of machine learning and big data mining has changed predictive modeling for many business applications, nevertheless in the actuarial science, few dissertations go beyond the traditional Generalized Linear Models (GLM) model, and few insurers use Machine Learning methods as a pricing model for the claims experience.

Three major reasons lead the actuary not to use these new methods :

1. the *Black box* effect, referring to **lack of interpretability** of machine learning models, such as Random Forest, Xgboost, Neural Networks. In many countries, as in France, law makes mandatory to provide an explanation of a model output, that is to say in insurance, the price of the policy. Furthermore, pricing models should be transparent and easy to communicate ;
2. by changing the principle of the "calculator" based on a **multiplicative structure**, the way most actuarial and IT departments work will be strongly affected ;
3. The use of machine learning in pricing can lead to extreme **personalization of risk**, to the detriment of risks mutualization, in the form of extremely high premiums.

Nowadays, Machine Learning techniques still have very marginal role in pricing (in the selection of variables for example) and they are rarely used as main pricing model.

In a more inclusive and collaborative vision between Machine Learning models and GLM, we will use black box models to detect **interactions** among the variables, without specifying it beforehand, and then we will test their statistical significance, in a simple GLM. The interaction terms, thus, being an expression of the complexity of a model, allow GLM models to take advantage of the good properties of more sophisticated models.

Data

The scope of application is the product *Smart Home Pricing* of personal property insurance, for the *Water damage* coverage. Its particularity is that the inputs of models are hyper-individualized meteorological, economic, climatic and demographic data (at the address and even at the building) which requires the geocoding of the insured home and contents. Among the innovative variables, we will use for example the presence of frost, the number of stormy days or the number of craftsmen in the municipality.

Interactions and importance of variables

Compared to what the literature proposes on the detection of interactions in statistical models, in particular to the work of Antoine Guillot [8] in non-life insurance, this work extends the research and the exploitation of the combined effects of variables inputs to the model output to two fields : sensitivity analysis and explainable artificial intelligence (XAI).

GLM : an interpretable, but not very complex model

The detection of these interactions by innovative methods was motivated by a limitation of the GLM : this model, despite its high degree of interpretability, does not include terms of a higher degree (multivariate polynomials), i.e., it lacks the complexity generated by the crossed effect of two (or more) variables.

However, as a sophistication of linear models from which they inherit a linear structure, they lead to predictions through a tabular form, easily translatable into tariff plans. That's why nowadays the Generalized Linear

Models, introduced by Nelder and Wedderburn (1972), are the insurance industry standard for developing analytical pricing models.

Indeed, it is possible to decompose the prediction of a GLM as the sum of the effects of each of the variables of the model :

$$g(\mathbb{E}(Y)) = \underbrace{\beta_0}_{\text{average effet}} + \underbrace{\beta_1 X_1}_{\text{effet variable } X_1} + \underbrace{\beta_2 X_2}_{\text{effet variable } X_2} + \dots + \underbrace{\beta_n X_n}_{\text{effet variable } X_n}$$

where g is the link function.

According to the model assumptions, the effect of each independent variable is constant regardless of the value taken by the other independent variables. However, the effect of X_1 , or X_2 , ... or X_n may vary depending on the values taken by one of the other independent variables introduced into the model. We define this behavior an *interaction* between these two variables.

Let a model $Y = f(X_1, X_2, \dots, X_n)$, we say that there is an interaction between X_1 and X_2 if the marginal effect of X_1 , noted β , depends on X_2 :

$$\frac{\partial Y}{\partial X_1} = \beta(X_2), \text{ here, } \beta \text{ is a function.}$$

Naive method of adding interactions

The GLM model does not include any statistical interactions, except the natural one brought by the link function.

However, we can add the cross term $X_1 * X_2$ manually as a model input :

$$g(\mathbb{E}(Y)) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_{n+1} X_1 * X_2$$

and then test its significance using the likelihood ratio test. In particular, we test the null hypothesis : $H_0 : \beta_{n+1} = 0$

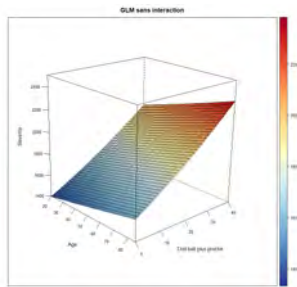


FIGURE 13.1 – Response surface of a GLM **without interactions** whose explanatory variables are the age of the occupant and the distance from the nearest building.

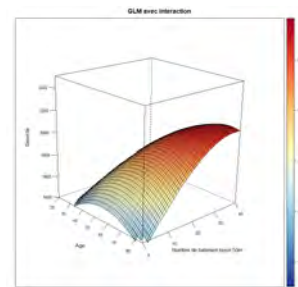


FIGURE 13.2 – Response surface of a GLM **with the interaction** between the age of the occupant and the distance from the nearest building.

This procedure, used by the variable selection methods *Forward/ Backward*, has two main limitations : computational, because it tests all possible combinations of the variables and this exhausts the memory of the software in high dimensional data ; and in terms of interaction model, because it is GLM-dependent. For example, if the link structure used or the random component is not appropriate, relevant interactions may not be meaningful.

Innovative method of interaction detection

Instead of testing the significance of the addition of each term, we introduce another approach, introducing models that intrinsically use interactions. In particular, *Tree based* models that use regression or classification trees as a basic model are known for their ability to model the effects of interactions among variables (Buchner et al., 2017 ; Schiltz et al., 2018). These models assume only the i.i.d hypothesis about the distribution of the data output.

In this study, interactions will be detected from the **decomposition of a quantity of interest** :

1. the **variance** : interactions will be quantified by Sobol indices. We will rely on the theory of sensitivity analysis and the main contributions of Ilya Sobol [17]. The variance of the model is decomposed by the

sum of functions of increasing dimensions :

$$V := \mathbb{V}(Y) = \sum_{i=1}^n V_i + \sum_{1 \leq i < j \leq n} V_{ij} + \dots + V_{1\dots n}$$

with $\forall i, j = 1, \dots, n$

$$V_i = \mathbb{V}ar(\mathbb{E}(Y|X_i))$$

$$V_{ij} = \mathbb{V}ar(\mathbb{E}(Y|X_i, X_j)) - V_i - V_j$$

$$V_{1\dots n} = V - \sum_{i=1}^n V_i - \sum_{1 \leq i < j \leq n} V_{ij} - \dots - \sum_{1 \leq i_1 < i_2 < \dots < i_{p-1} \leq n} V_{i_1 \dots i_{p-1}}$$

Each term of this sum, normalized by the total variance, is a Sobol index, said *of order k* if k is the number of variables with respect to which we condition the model expectation. The strength of these indices lies in the fact that their sum is 1, so they are very intuitive.

The **interactions are Sobol's indices of order 2** :

$$S_{i,j} = \frac{V_{i,j}}{V}, \forall i = 1, \dots, n \text{ with } i \neq j \text{ and } n \text{ number of explanatory variables}$$

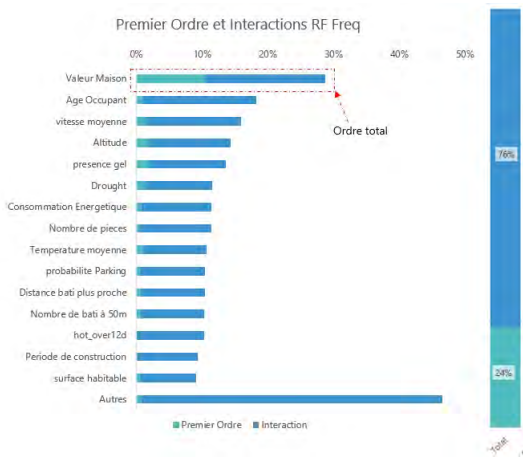


FIGURE 13.3 – First-order Sobol Index and impacts of interactions in the random forest model for frequency : the main effects explain 24% of the total variance of the model output, while interactions contribute 76%. The most important variable according to the Sobol functional decomposition is the value of the house.



FIGURE 13.4 – Interactions (Sobol Indexes of order 2) of House Value : stronger interactions in the random frequency forest model.

- of the **predicted output** : the interactions will be quantified by the SHAP interaction indices. This notion is linked as much to game theory as to the most recent field of Explainable Artificial Intelligence (XAI). The idea is to decompose the prediction by the sum of the importance of the n explanatory variables in order to be able to interpret any model. For an observation $x = (x_1, \dots, x_n)$ the predicted value $f(x)$ is decomposed as follows :

$$f(x) = \sum_{j=1}^n \underbrace{\varphi_j(\Delta^x)}_{\text{contribution of explanatory variable } X_j \text{ to the prediction } f(x)} + \underbrace{\mathbb{E}(f(X))}_{\text{prediction expected on Train dataset}}$$

$\varphi_j(\Delta^x)$ is the SHAP value for the variable X_j of the cooperative game $(N = \{X_1, \dots, X_n\}, \Delta^x)$:

$$\varphi_i(\Delta^x) = \sum_{S \in \mathcal{P}(N \setminus \{X_i\})} \frac{|S|!(p - |S| - 1)!}{p!} * [\Delta^x(S \cup \{X_i\}) - \Delta^x(S)]$$

where the quantity $[\Delta^x(S \cup \{X_i\}) - \Delta^x(S)]$ represents the contribution of the variable X_i to the coalition S (an element of N , the collection of all subsets).

In this framework, the **interactions are the SHAP interaction indices** :

$$\Phi_{i,j} = \sum_{S \subseteq N \setminus \{X_i, X_j\}} \frac{|S|!(p - |S| - 2)!}{2(p - 1)!} \nabla_{i,j}(S)$$

with $i \neq j$, n number of input variables

$$\begin{aligned} \nabla_{i,j}(S) &= f_x(S \cup \{X_i, X_j\}) - f_x(S \cup \{X_i\}) - f_x(S \cup \{X_j\}) + f_x(S) \\ &= f_x(S \cup \{X_i, X_j\}) - f_x(S \cup \{X_i\}) - [f_x(S \cup \{X_j\}) - f_x(S)] \\ f_x : \{X_{i_1}, \dots, X_{i_s}\} &\subset \{X_1, \dots, X_n\} \mapsto \mathbb{E}[f(X_1, \dots, X_n) | X_{i_1} = x_{i_1}, \dots, X_{i_s} = x_{i_s}] \\ \text{avec } \{i_1, \dots, i_n\} &\subset \{1, \dots, n\} \text{ et } s \in \{1, \dots, n\} \end{aligned}$$

As First order and total order Sobol indices, one can separate the *individual* effect of a variable in a SHAP prediction that represents the effect of that variable in all coalitions in which it participates. The *interaction effects* are thus the part of the SHAP value where the effect of the variable alone is excluded :

$$\underbrace{\Phi_{i,i}}_{\text{individual effect variable } X_i} = \underbrace{\varphi_i}_{\text{SHAP value of } X_i} - \underbrace{\sum_{j \neq i} \Phi_{i,j}}_{\text{Interactions with } X_i}$$

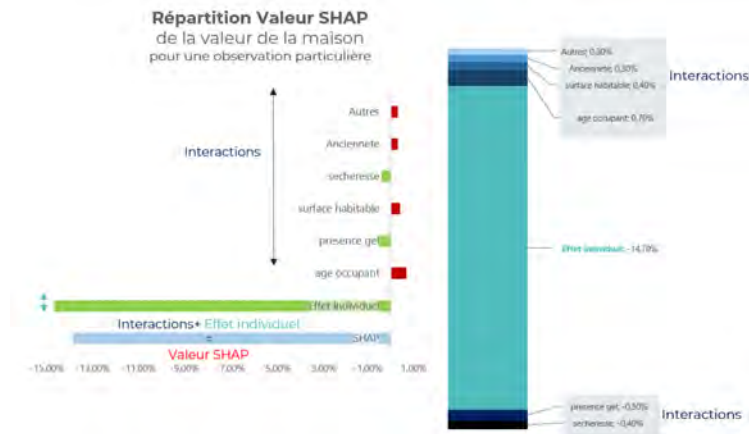


FIGURE 13.5 – Distribution of the SHAP value according to the individual effect of the variable "value of the house" and its interaction effects with the other explanatory variables. This is the xgboost model for the claim frequency fit. The SHAP value of the house for this particular observation is -13.8%, i.e. its contribution brings down (negative sign) the prediction. The client's age and seniority in the portfolio are the only interactions with the value of the house responsible for the increase in the number of claims.

The lack of explicability due to the black box effect of Machine Learning leads to open these models with these tools : if linear models or logistic regression have a limited number of parameters and are interpreted through an additive *natural* structure, this is not the case for black box models such as Random Forest or xgboost.

The main advantage of going through the domains of sensitivity analysis and Explainable Artificial Intelligence is their agnosticity to model, which allows to generalize the additive representation of a quantity of interest for any model. Furthermore, these two techniques are complementary : Sobol indices are *global*, i.e. they quantify the importance on the whole set of outputs, while the SHAP value is *local*, i.e. it quantifies the contribution of coalitions of variables for a specific observation.

Application to the Smart Home Pricing

Working Steps

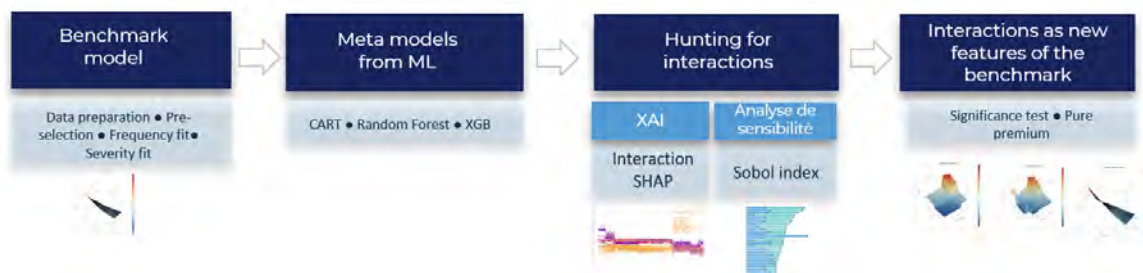
In order to measure the gain generated by interaction terms, we compared the performance of GLM models with and without the addition of interaction terms.

In particular, in a first step, we have built a simple model, called *Benchmark*, which will be the basis of comparison, and more advanced models, called *Meta-models*.

Then, from the latter, we detected the interactions of the pairs of variables.

Finally, we added these terms to the *Benchmark* and quantified their gain.

This methodology can concern any model *Benchmark* :

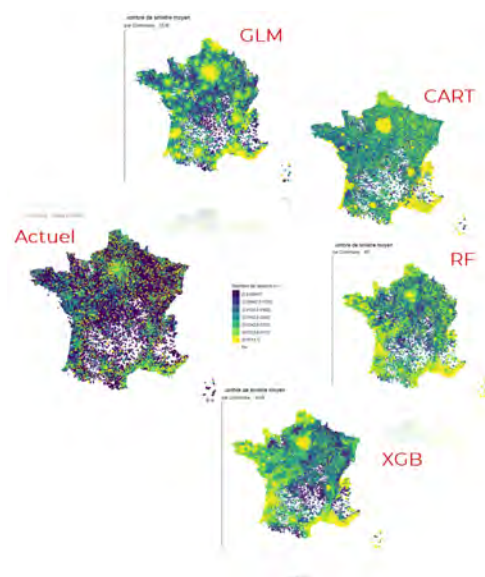


Effect of High Dimension and Variable Pre-selection

The Monte Carlo estimators of the Sobol and SHAP indices (Strumbelj and Kononenko [21]) can suffer from the curse of dimensionality or type of variable (categorical, quantitative, continuous, ...).

Indeed, the execution time of a non-optimized algorithm is $\mathcal{O}(n \times T(f) \times M)$, with n : number of variables, $T(f)$: complexity of the model f and M : number of Monte Carlo iterations. In order to use robust estimators without exhausting the software's RAM (R or python), we have pre-selected about 40 explanatory variables, reducing the base used to a representative subsample and/or estimators that converge faster, such as **TreeShap**.

Frequency and Severity fits before adding interactions



We want to propose to the insurer an advanced GLM, including interactions, in order to add complexity while keeping the rate structure simple to implement.

First, we calibrated the GLMs for frequency and severity.

Then, we set up three machine learning models : a regression tree, a Random Forest and an Extreme Gradient Bosting. These models will be the models from which interactions will be detected.

We observed that the four frequency models segment the risk in different ways : the comparison on the map of France of the average number of predicted claims, on the left, shows that the machine learning models take into account the geographical risk in a finer and smoother way, especially in the countryside.

Detection of interactions

CART : a first sophisticated and interpretable model

The *Greedy* algorithm used by the CART model (*Classification and Regression Tree*) is, by construction, an interpretable model. It builds up a tree in several steps by dividing the population into two groups at each step, maximizing the inter-class variance

(the groups are subsets whose outputs are as dispersed as possible). Each division of the population describes an interaction introduced by the model.

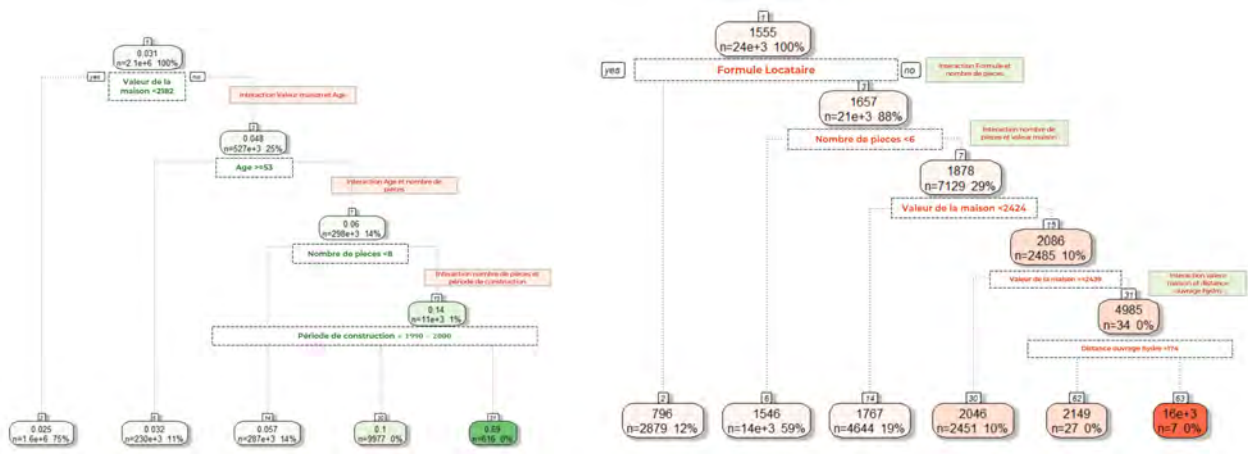


FIGURE 13.6 – CART frequency (left) and on severity (right) : interactions are detected by the partition steps of the Train dataset. For example, in the second step of the frequency model, the model introduces the interaction between the value of the house (root division) when it is greater than 2182 and age.

SHAP Interactions

Now, we computed SHAP interactions for the random forest and xgboost models to detect their interactions. In order not to introduce artificial interactions or very small effects in the final model, the interactions will be visualized and validated graphically.

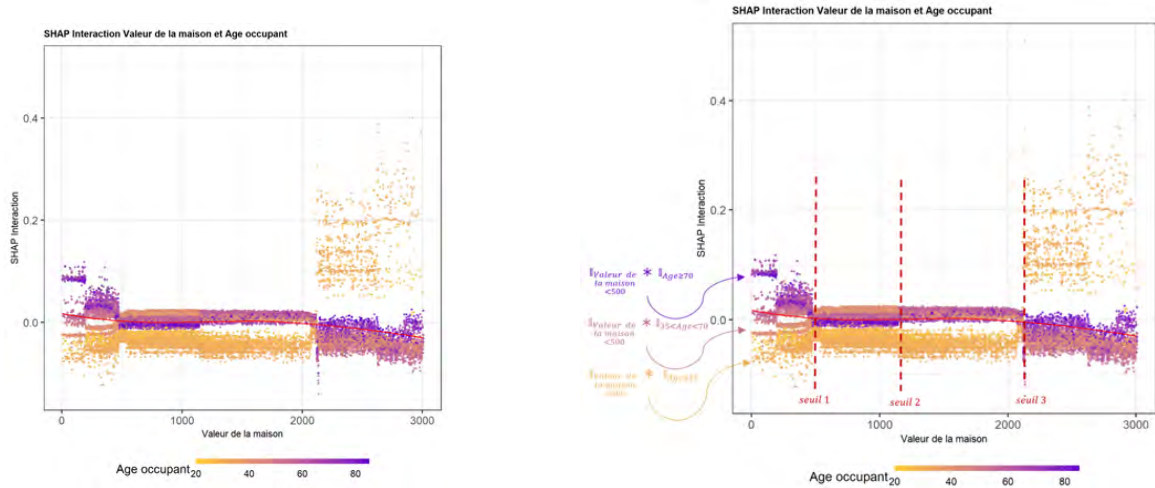


FIGURE 13.7 – Visualization of interactions by individual of the validation dataset, on the X-axis the value of the house and the color according to the age of the occupant. The average intensity (in absolute value) is 2.1%, the highest in the portfolio. The scope of this interaction is both local and global. We note that the contribution of senior profiles living in outside the central areas (lower house value) is positive, whereas young people tend to report more claims in most expensive homes (houses in the city center for example).

FIGURE 13.8 – Intervals for interaction terms for GLM from SHAP interaction graphs.

A natural choice to integrate local interactions into GLM is to cut the definition domain of the interacting variables into intervals. For CART these intervals are defined by construction, whereas for SHAP interactions, we defined thresholds according to the interaction graphs (fig. 13.8).

We noticed that the interactions detected with the *random forest* model are lower compared to those introduced by *xgboost* :

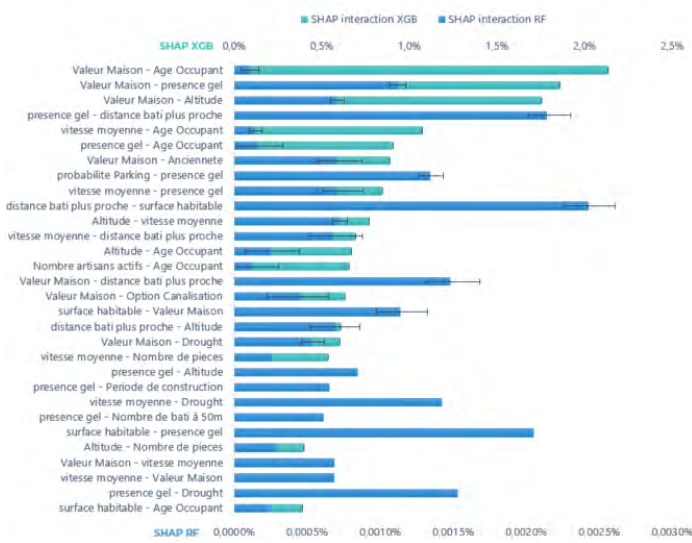


FIGURE 13.9 – Comparison among SHAP interactions values for frequency : *xgboost* has much higher values than the *random forest* model. It considers for the prediction the interactions of house value with age, frost presence, altitude and less strong interactions as well. The *random forest* model, on the other hand, takes into account in the prediction the interactions of the distance from the nearest building and the living area.

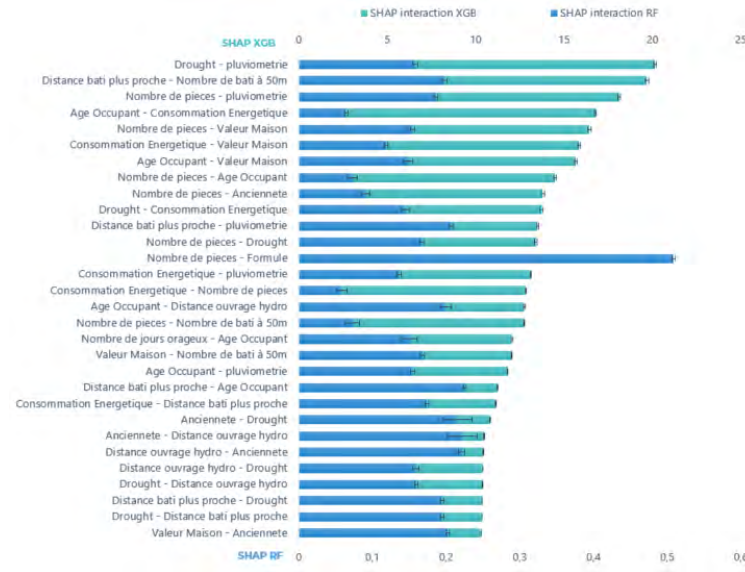


FIGURE 13.10 – Comparison among SHAP interactions for severity models : *random forest* interactions are very small (mean 0.17) compared to those integrated by *xgboost* (mean 5.9). In the prediction, *xgboost* takes into account the interactions of house value with age, number of rooms, energy consumption and interactions between climatic variables (rainfall and drought).

Interactions according to Sobol analysis

The Sobol indices of order 2 identify global interactions. According to these indices, the interactions of frequency are stronger than those of severity ; the frequency models thus are more complex in the sense of the Sobol sensitivity analysis.

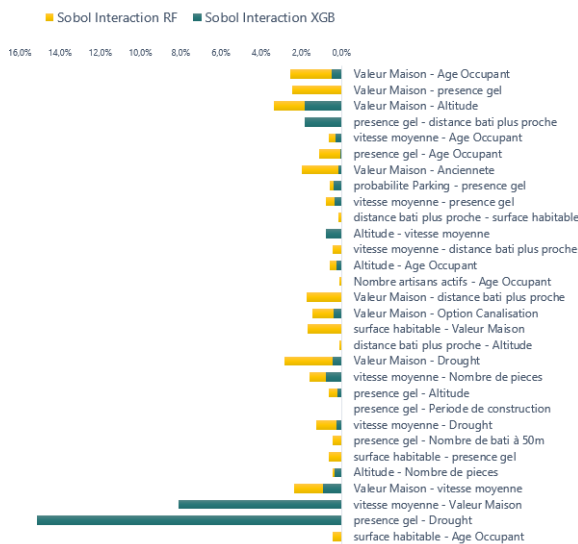


FIGURE 13.11 – Visualization of interactions at order 2 for frequency models. The interactions of the house value are the strongest, followed by those among the geographic and climatic variables (rainfall quantity, drought, altitude, wind speed).

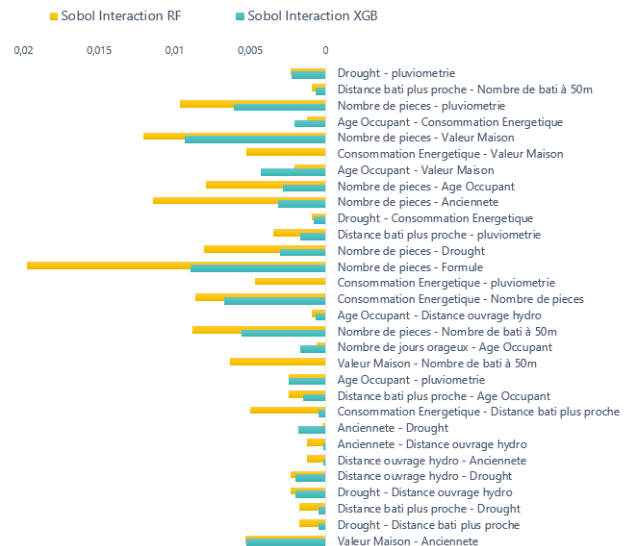


FIGURE 13.12 – Visualization of interactions at order 2 for severity models. The interactions of the number of rooms with the value of the house, the amount of rainfall, the age of the occupant, the formula and the drought are the most significant, followed by those of the value of the house.

By using Graph theory, we visualized the interactions beyond a threshold :

Visualisation des interactions selon l'analyse de Sobol
 Modèle: Frequence-RF seuil: 0.015

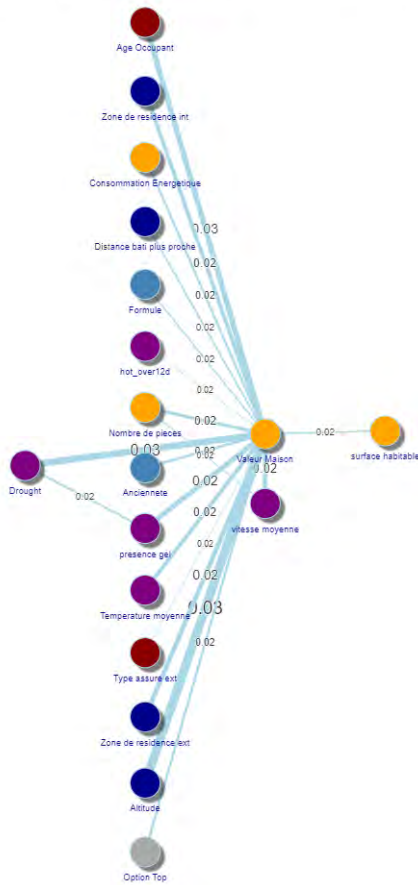


FIGURE 13.13 – Visualization of interactions greater than 1.5% for the random forest frequency model. The value of the house is the variable that interacts the most with the other covariates.

Visualisation des interactions selon l'analyse de Sobol
 Modèle: Severite-RF seuil: 0.01

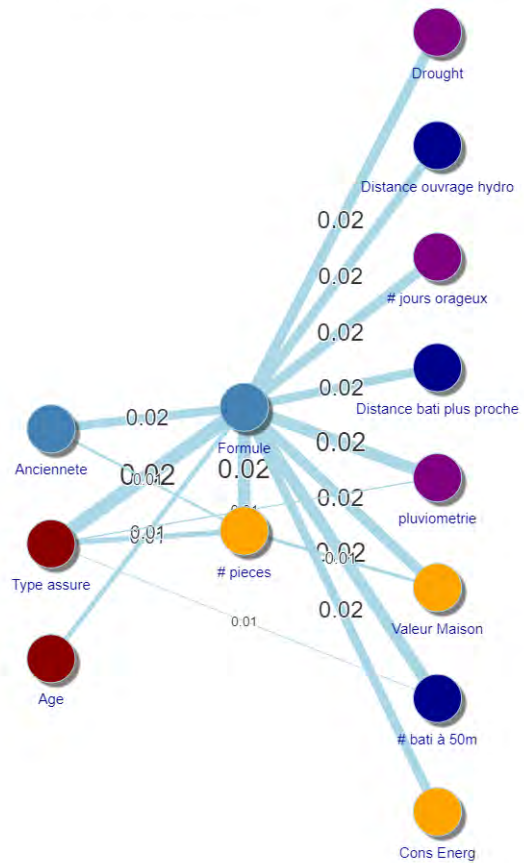


FIGURE 13.14 – Visualization of interactions greater than 1% for the random forest severity model. The "Formula" is the variable that interacts the most with the other covariates.

Sobol interactions are a global measure over the whole variable definition domain. As a simple way to introduce them in GLM *Benchmark*, then we use polynomial terms of the second degree.

Model comparisons

Comparison between GLMs with interactions and simple GLM.

As last analysis, we compared on the Test dataset several performance indicators of the GLM models with the interactions detected.

| Modele | Fréquence | | | | | | | | | Sévérité | | | | | | | | | |
|-------------------------------|----------------|---------|---------|---------|--------|--------|--------|------------------|-------------------|----------------|--------|--------|---------|--------|--------|--------|----------------|-------------------|--------|
| | ecart Pred avg | MSE | RMSE | MAE | Q2 | GINI | AIC | Deviance Poisson | Residual Deviance | ecart Pred avg | MSE | RMSE | MAE | Q2 | GINI | AIC | Deviance Gamma | Residual Deviance | |
| Benchmark | 100,00 | 100,000 | 100,000 | 100,000 | 100,00 | 100,00 | 100,00 | 100,000 | 100,00 | 100,00 | 100,00 | 100,00 | 100,000 | 100,00 | 100,00 | 100,00 | 100,00 | 100,00 | 100,00 |
| Clouton CART | 94,25 | 99,971 | 99,986 | 100,065 | 102,33 | 100,20 | 99,93 | 99,962 | 99,91 | 98,07 | 98,994 | 98,997 | 98,353 | 100,55 | 102,54 | 99,98 | 100,58 | 99,52 | |
| SHAP RF | 93,29 | 99,993 | 99,998 | 100,076 | 100,87 | 100,23 | 99,97 | 99,963 | 99,91 | 98,07 | 98,993 | 98,992 | 100,252 | 102,06 | 100,25 | 99,90 | 99,982 | 99,83 | |
| SHAP XGB | 78,07 | 99,983 | 99,992 | 100,252 | 102,06 | 100,25 | 99,90 | 99,982 | 99,83 | 98,89 | 98,998 | 98,999 | 100,131 | 100,93 | 110,73 | 99,98 | 100,98 | 99,30 | |
| SHAP Toutes les interactions | 78,59 | 99,984 | 99,992 | 100,252 | 101,98 | 100,25 | 99,90 | 99,981 | 99,79 | 98,89 | 98,998 | 98,999 | 100,131 | 100,93 | 110,73 | 99,98 | 100,98 | 99,30 | |
| Sobol RF | 87,20 | 99,987 | 99,993 | 100,160 | 101,62 | 99,55 | 99,93 | 100,017 | 99,87 | 97,54 | 98,999 | 98,999 | 100,041 | 101,05 | 103,25 | 100,00 | 100,32 | 99,70 | |
| Sobol XGB | 98,22 | 99,989 | 99,995 | 100,021 | 101,30 | 99,97 | 99,93 | 99,999 | 99,91 | 99,79 | 98,999 | 98,999 | 100,172 | 101,05 | 101,72 | 100,00 | 100,51 | 99,70 | |
| Sobol Toutes les Interactions | 96,66 | 99,987 | 99,994 | 100,167 | 101,55 | 99,50 | 99,93 | 100,014 | 99,87 | 97,50 | 98,999 | 99,000 | 100,029 | 101,07 | 103,82 | 100,00 | 100,30 | 99,70 | |
| Toutes les Interactions | 79,37 | 99,987 | 99,994 | 100,251 | 101,52 | 100,20 | 99,93 | 100,026 | 99,74 | 96,21 | 99,990 | 98,991 | 100,050 | 100,15 | 117,08 | 99,95 | 102,00 | 98,48 | |

FIGURE 13.15 – Metrics for evaluation of frequency and severity models

Among the frequency models built, the one containing all SHAP interactions over the entire Test dataset is closer to the observed mean and is the most discriminant (higher Gini). The model derived from CART comes closer on average to the observed values (lower MSE) and is the most predictive (higher Q2).

Severity models with interactions are less efficient than frequency models : the interactions detected could be more complex in nature than second degree polynomials or not localized enough. The model with all interactions predicts a portfolio mean closer to the real one than the Benchmark, it is the most discriminating and reduces the AIC.

More concretely, the addition of the interactions corresponds to a modification of the response surface of the GLM model : we show here only the interactions between age and house value, but the results are equivalent for the other interactions.

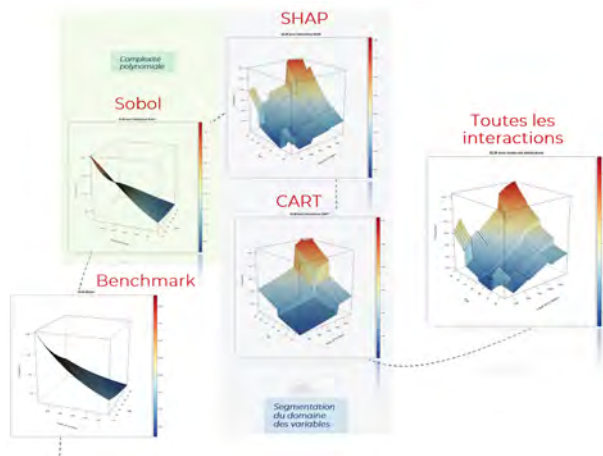


FIGURE 13.16 – Visualization of GLM response surfaces with and without interactions for a 2-dimension model (*age* and *house value*) : the link function gives an exponential structure to the model, then the terms of the Sobol interactions confer complexity with the addition of a second degree polynomial to the tariff formula, and finally the terms of the SHAP and CART interactions partition the variable definition domain. Compared to the Benchmark model, the risk is much more segmented.

Conclusion

In this work, we carried out a sensitivity analysis on the *Smart Home Pricing* data to improve the *Water Damage* rating model, adding *complexity* while maintaining an analytical, transparent and interpretable structure that joined perfectly the traditional insurance company rating process.

In order to satisfy these criteria, we have extended the problem of rate optimization to the problem of detecting and integrating interactions among variables, because the interaction is an expression of the complexity of the model.

We first detected the interactions introduced by more sophisticated models, known as *black box* models because of their inaccessible predictive structure, using concepts from Game Theory and sensitivity analysis according to Sobol approach.

We found a compromise among computational speed (parallelisation of tasks, metamodels), complexity (of models), interpretability (of pricing models) and model agnosticity. Sensitivity analysis tools such as Sobol indices and SHAP indices (a more recent review according to the Game Theory) were chosen because of the good convergence properties of their estimators and the type of information provided.

Not only the Sobol indices of order two inform us of the share of the total variance due to each interaction, furthermore the SHAP interaction indices determine in which direction the interaction has an impact on the prediction (whether it is increasing or decreasing because of the interaction) and with which intensity.

Next we integrated the interactions detected locally and globally by adding polynomial terms in the initial GLM model.

These terms improve the usual evaluation metrics of the simple GLM model (MSE, RMSE, MAE, Q2, Gini, deviance, AIC) with a gain estimated between 0.03% and 17%.

More generally, this detection methodology is not limited to pricing optimization : its model agnosticity property allows it to be applied to any model, complex or not. In the actuarial literature, there is a tendency to compare the performance of more complex models with the classical GLM, or to use them separately. Instead, through this new approach, Machine Learning and Generalized Linear Models work in collaboration.

Thus, the more sophisticated models play a major role in Non-Life insurance pricing, because interaction terms are an expression of the complexity of these models.

In addition, the perception of geographic risk inherited by more complex models and deeper granularity data (address and building data) is much more segmented.

These data are particularly adapted to this application framework because the database lives in high dimension (300 variables) and bivariate analysis to examine all possible interactions can be tedious and slow. In addition, the inclusion of deeper granularity geographic variables challenges current zoning methods for taking into account spatial risk.

Study limitations and research perspectives

The main **limitations** of this study are the following :

- the **complexity of data and models**. SHAP estimators and Sobol indices helped us to detect interactions among variables and we have integrated these functions within the structure of a simple GLM by adding polynomial terms. However, the nature of these relationships can be more complex (the addition of the indicator function is not sufficient) or very localized (as opposed to Sobol indices which are global) : in this case, the addition of the terms may not be significant according to the likelihood ratio test. Smoothing with *spline* techniques could improve the performance of the models.
- The **propagation of error**. The calibration of CART, *xgboost* and *random forest* metamodels can propagate the prediction error to the estimation of Sobol or SHAP indices and introduce artificial interactions, especially in the presence of collinearities.
For bounding their integration in the initial model, the interactions must be validated before being exploited in the *Benchmark* model.
Furthermore, the size of the data can slow down the convergence of these estimators and in the case where the importance is very small the estimates can be negative (for Sobol indices).

The SHAP value theory, based on the explicability of the model, is not new to the world of insurance : it is also a monitoring tool, portfolio understanding and risk selection. References under the software programs R and python are also abundant and sophisticated.

Its extension to the detection of interactions is one of the **innovative elements** that we wanted to convey with this dissertation.

Sobol's analysis has been another challenge for this type of problem : unlike the contexts where it is usually applied (aeronautical engineering, finance,...), where we know the distributions of variables, in insurance use cases we have the realizations of the variables without knowing their distributions.

We restricted the scope to attritional claims, but in the continuity of sensitivity analysis studies, it would be interesting to calculate these indices on different quantiles and thus integrate the interactions on **large claims**.

Bibliographie

- [1] K. Chan A. Saltelli and E.M. Scott. Sensitivity analysis, 2000.
- [2] T. Andres F. Campolongo J. Cariboni D. Gatelli M. Saisana S. Tarantola A. Saltelli, M.Ratto. Global sensitivity analysis., 2007.
- [3] Douglas M. Bates Kurt Hornik Albrecht Gebhardt David Firth Brian Ripley, Bill Venables.
- [4] economie.gouv.fr. *Assurance multirisque habitation*. 2018.
- [5] Dirk Eddelbuettel et Brandon Greenwell. *Fast Approximate Shapley Values*. 2017.
- [6] Tianqi Chen et Carlos Guestrin. *XGBoost : A Scalable Tree Boosting System, Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016.
- [7] Arthur Charpentier et Michel Denuit.
- [8] Antoine Guillot. *DETECTION ET EXPLOITATION DES INTERACTIONS EN TARIFICATION NON-VIE*.
- [9] Antonio Verbelen Henckaerts, Côté.
- [10] Lenon Minorics et Patrick Blöbaum. Janzing, Dominik. *Quantification de la pertinence des fonctionnalités dans l'IA explicable : un problème de causalité*. 2019.
- [11] JACQUES Julien. Pratique de l'analyse de sensibilité : comment évaluer l'impact des entrées aléatoires sur la sortie d'un modèle mathématique, 2011. [PUB. IRMA, LILLE 2011 Vol. 71, No III].
- [12] Erion G. Lundberg, Scott M. and Su-In Lee. *Consistent individualized feature attribution for tree ensembles*. 2018.
- [13] Scott M. Lundberg and Su-In Lee. *A unified approach to interpreting model predictions*. 2017.
- [14] Christoph Molnar. *Interpretable Machine Learning A Guide for Making Black Box Models Explainable*. 2020.
- [15] Sandra Maria Nawar. *Machine Learning Techniques for Detecting Hierarchical Interactions in Insurance Claims Models*. 2016.
- [16] Shapley Lyod S. *value for n-person games. Contributions to the Theory of Games*. 1953.
- [17] I.M. Sobol. Sensitivity estimates for nonlinear mathematical models. mathematical modelling and computational experiments,, 1993.
- [18] Nelson B. L. Staum J. Song, E. *Shapley effects for global sensitivity analysis : Theory and computation. SIAM/ASA Journal on Uncertainty Quantification, 4(1), 1060-1083*. 2016.
- [19] Mukund et Amir Najmi. Sundararajan. *Les nombreuses valeurs de Shapley pour l'explication du modèle*. 2019.
- [20] T. Turanyi. Sensitivity analysis of complex kinetic system, tools and applications., 1990.
- [21] Štrumbelj E. et Kononenko I. *Explaining prediction models and individual predictions with feature contributions*. 2014.

Annexe A

Spécificités du marché de l'assurance MRH

A.1 Le marché français de l'assurance Habitation

Cette section a pour but de donner un aperçu du marché de l'assurance habitation en France. En 2019 le marché de l'assurance habitation a constitué 18% du marché IARD en terme de cotisations.

Sinistralité Multirisques habitation (MRH) en 2019

En Habitation, le coût d'un sinistre le plus élevé en moyenne est en cas d'incendie, alors que les dégâts des eaux sont la première cause de sinistre.

| Branche | Fréquence | Sévérité | | |
|-----------------------|-----------|----------|----------------------|---------------|
| | | | 2019 | 19/18 |
| Incendie | 5,1 ‰ | 8 272€ | | |
| Tempête, grêle, neige | 11,4 ‰ | 2 123€ | | |
| Vol | 9,0 ‰ | 1 893€ | | |
| Dégâts des eaux | 33,3 ‰ | 1 082€ | Cotisations (M €) | 10 930 + 4.1% |
| Responsabilité civile | 10,1 ‰ | 1 103€ | Prime moyenne HT (€) | 255 + 2.0% |
| Bris de glaces | 7,2 ‰ | 449€ | | |

$$\text{Fréquence des sinistres (en‰)} := \frac{\text{Nombre de sinistres}}{\text{Nombre de garanties souscrites}} * 1000$$

Source : FFA

La météo explique en grande partie cette tendance : les fortes pluviométries et les inondations caractérisent les années où la fréquence des sinistres en Dégâts des eaux est élevée; l'activité orageuse a un impact sur la sévérité en incendie.

La garantie Vol, quant à elle, tend à décroître de pair avec l'installation des équipements de sécurité dans les habitations : depuis 2010, le nombre de caméras de sécurité vendues augmente de +11% chaque année, alors que les alarmes, les portes blindées et les digicodes entre 1.8% et 2.3% par an.

Évolution des résultats techniques en MRH

Depuis 2014 l'indicateur de rentabilité *Ratio Combiné* montre que le marché de l'assurance habitation n'est plus déficitaire : les assureurs ont en effet décidé de réorienter leur politique pour faire face à la croissance des événements climatiques pénalisant les résultats techniques. Ils ont en particulier actionné quatre leviers pour maîtriser la charge climatique¹ :

¹Sources : Argus de l'assurance "Quatre leviers pour maîtriser la charge climatique", 30/01/2015

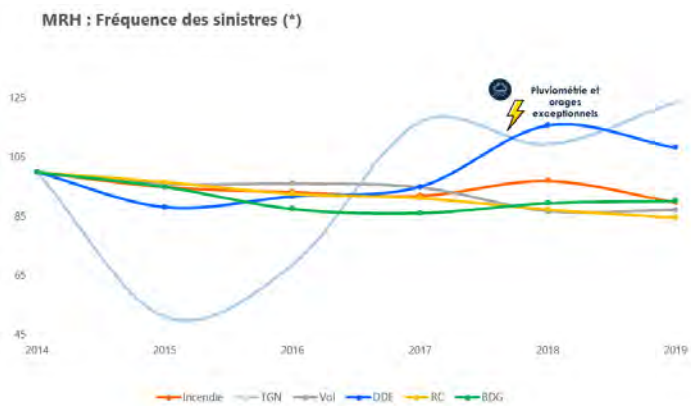


FIGURE A.1 – (*) : Indice base 100 année 2014.
Source : FFA

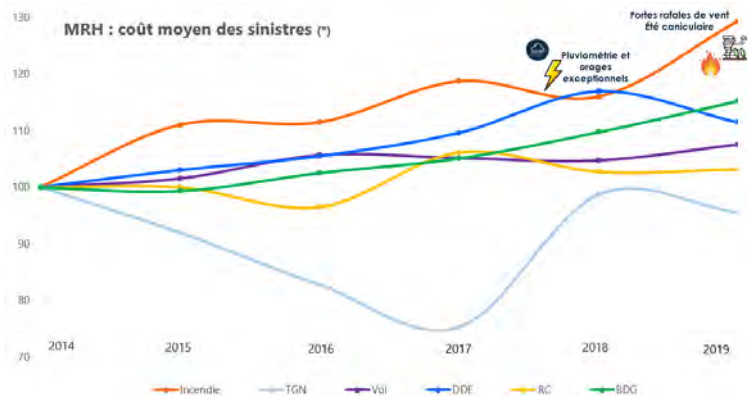


FIGURE A.2 – (*) : Indice base 100 année 2014.
Source : FFA

- Utiliser la réassurance pour se prémunir du cumul de sinistres, en réponse à l'augmentation du risque de fréquence de sinistres liés à la tempête
- Sophistiquer la tarification du risque tempête : s'agissant d'événements exceptionnels, les assureurs ne disposaient pas d'historique pertinent pour tarifier finement cette garantie. Avec la multiplication des événements climatiques de type tempête : Quinten (2009), Klaus (2009), Joaquim (2011), Dirk (2013), la modélisation repose sur une segmentation qui tient compte des critères du risque sous-jacent, tout en conservant une dimension de mutualisation importante.
- Activer des politiques de prévention.
- Utiliser les données de géocodage pour cartographier finement les zones à risques et gagner en visibilité sur leurs expositions.

L'année 2019, ainsi que les cinq précédentes années, a été particulièrement touchées par les événements climatiques et les catastrophes naturelles (sécheresse, inondations du Sud-Est et dans le Languedoc, séisme du Teil, Orages,...). La branche catastrophe naturelle de l'assurance habitation est encore déficitaire, mais la compensation des garanties fait que les ratios combinés en assurance Habitation sont depuis 2014 en dessous de 100%.

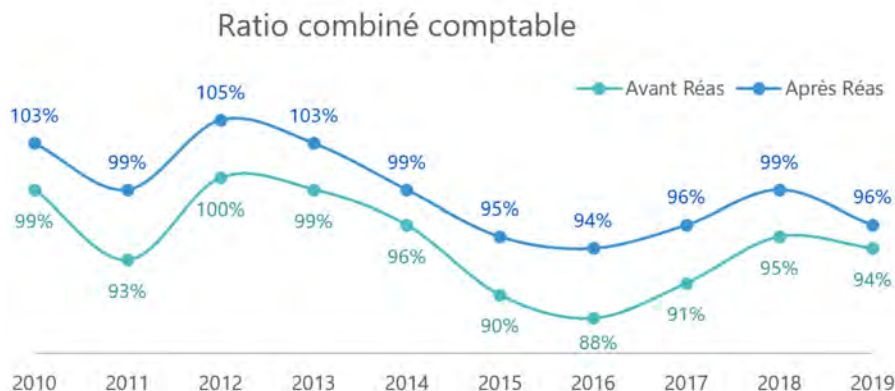


FIGURE A.3 – Ratio avant réassurance :
$$:= \frac{\text{Charge des prestations (brutes de réassurance)} + \text{frais généraux (bruts de réassurance)}}{\text{Primes émises (brutes de réassurance)}}$$

Ratio après réassurance :
$$:= \frac{\text{Charge des prestations (nettes de réassurance)} + \text{frais généraux (nettes de réassurance)}}{\text{Primes émises (nettes de réassurance)}}$$

Source : FFA

Parc des logements en 2019

Le bilan de l'INSEE révèle l'existence d'environ 36,6 millions de logements en France au début de l'année 2019 : le parc des logements est dominé par les résidences principales (82%), auxquelles s'ajoutent 10 % de résidences

secondaires ou de logements occasionnels, et de 8 % de biens inoccupés.

21 % de ces résidences principales sont localisées dans les communes rurales, tandis que l'agglomération parisienne en concentre 16 %.

TABLE A.1 – Parc de logements au 1er janvier (France hors Mayotte)

| | 2017 | 2018 | 2019 | 19/18 |
|--------------------------|---------------|---------------|---------------|--------------|
| en <i>k</i> | | | | |
| Résidence principale | 29 499 | 29 688 | 29 916 | +0.8% |
| dont logement individuel | 16 763 | 16 844 | 16 960 | +0.7% |
| dont logement collectif | 12 736 | 12 844 | 12 956 | +0.9% |
| Résidence secondaire | 3 461 | 3 528 | 3 590 | +1.8% |
| Logement vacant | 2 984 | 3 031 | 3 103 | +2.4% |
| Ensemble du parc | 35 944 | 36 247 | 36 609 | +1.0% |

Source : INSEE

Indexation en Assurance habitation

La hausse des primes d'assurance (entre 2% et 3% par an) est expliquée par plusieurs facteurs : la météo (surtout dans la garantie Dégâts des eaux) , des événements économiques tel qu'une crise (pour la garantie Vol), l'entrée en vigueur d'une norme de sécurité (pour la garantie Incendie), un changement de politique de souscription, et aussi des indices comme l'inflation, FFB, IRL, SMIC, IPEA,...

L'indice ICC FFB

L'indice ICC FFB est l'Indice du Coût de Construction émis par la Fédération Française du Bâtiment par rapport au coût de la construction d'un immeuble en France.

Il quantifie les évolutions du prix de tous les éléments qui rentrent dans la composition de la construction du bâtiment, en excluant la valeur du terrain d'origine.

Dans le cadre de l'estimation du prix de l'assurance habitation, c'est une référence pour toutes les compagnies d'assurance et il est utilisé pour indexer le prix des **franchises** d'assurance habitation, **cotisations**, **primes** d'assurance et les **capitaux garantis** au contrat. Plus cet indice est élevé, plus les assureurs vont répercuter sa valeur sur le coût de l'assurance habitation.

L'indice IPEA

L'indice des prix des travaux d'entretien-amélioration des bâtiments (IPEA) est un indice trimestriel qui mesure l'évolution des prix hors TVA pratiqués par les entreprises de la construction (y compris les entreprises artisanales) pour leurs travaux d'entretien et d'amélioration des bâtiments résidentiels et non résidentiels réalisés au cours du trimestre estimé.

Pour pouvoir être comparables d'une année à l'autre les sinistres sont corrigés à l'aide d'un ou plusieurs indicateurs ; ce processus est aussi appelé mise en *as if* : le nombre de contrats ou l'inflation ou des indices comme l'IPEA ou FFB sont ainsi utilisés pour transformer le coût d'un sinistre survenu dans le passé par le coût du sinistre s'il était survenu aujourd'hui (ou à la date de vision). Un calcul simplifié de ce processus est le

suivant :

$$Sinistre_{N, \text{corrigé}} = \frac{Indice_{N_{max}}}{Indice_N} * Sinistre_N$$

Principaux organismes d'assurance habitation

La bancassurance continue d'engranger des parts de marché significatives sur le marché de l'assurance dommages des particuliers, aux dépens des réseaux traditionnels : mutuelles d'assurance, agents généraux et courtage.

Ils détiennent environ 25% des contrats en MRH . Entre 2018 et 2019 leur part de marché (en chiffre d'affaires) est passée de 10,7 à 14,9%. Les mutuelles sans intermédiaire ainsi que les agents généraux ont reculé de un peu moins que 1 point de pourcentage, à différence des courtiers où la baisse est plus significative (de 25,6% à 23,7%).

La vente directe detient, quant à elle, une petite du marché (1.3%).

Principaux acteurs du marché français

Par chiffre d'affaire 2019, top 20

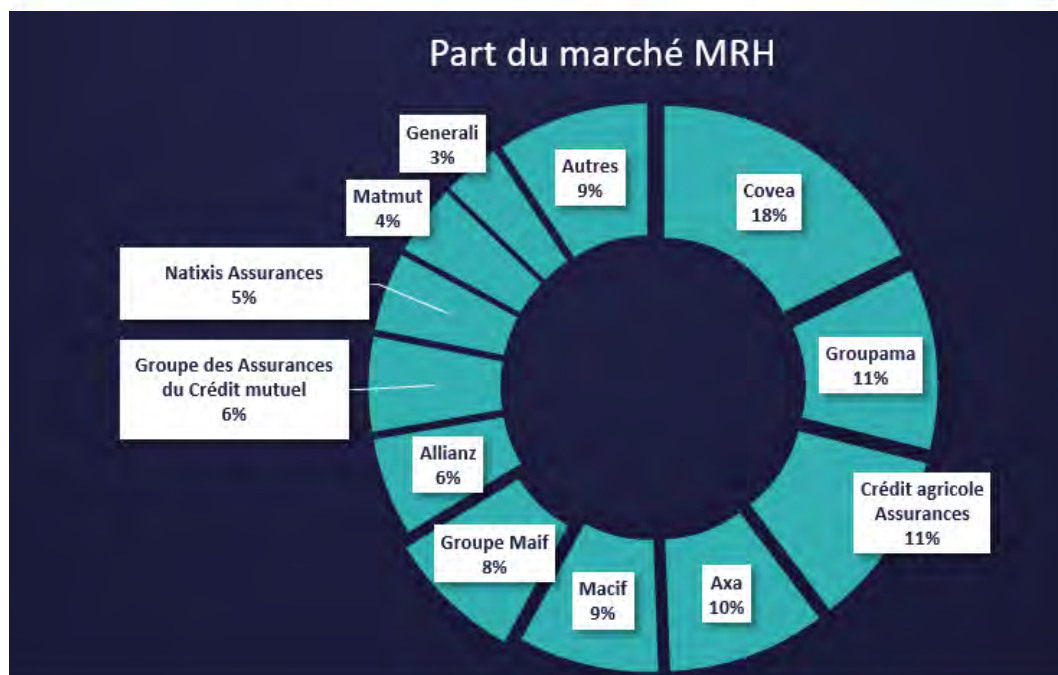


FIGURE A.4 – Le groupe d'assurance mutuelle française Covéa, réunissant notamment les marques MAAF, MMA et GMF détient la plus grande part de marché suivi par l'assureur traditionnel Groupama et le bancassureur Crédit Agricole Assurances.

Source : Argus Auto-MRH 2020 : coup d'arrêt sur les marchés !

A.2 Réglementation du marché MRH

Le règlement des dommages

Déclaration du sinistre par l'assuré

Le délai légal pour que l'assuré déclare un sinistre est de cinq jours à partir du jour où il en a connaissance (deux jours en cas de vol, dix jours qui suivent la parution de l'arrêté interministériel en cas de catastrophe naturelle). Le contrat peut prévoir des obligations conventionnelles, comme le dépôt de plainte au commissariat en cas de vol ou de vandalisme, et les pièces utiles pour évaluer les pertes, par exemple l'état estimatif des biens endommagés, détruits ou volés. Dans la mesure du possible, l'assuré doit parer au plus pressé pour limiter l'importance des dégâts et sauvegarder ses biens.

L'assuré qui manque ses obligations peut être privé de la garantie du sinistre.



Expertise

C'est à l'initiative de l'assureur ou de l'assuré. Elle a pour objet de déterminer les circonstances du sinistre, d'évaluer l'étendue et le montant du préjudice, de préconiser les modalités de remise en état.



Calcul de l'indemnité

Le montant de l'indemnité est évalué par rapport au jour du sinistre (on n'actualise pas le règlement en fonction de la date de règlement) selon le principe indemnitaire : l'assurance ne peut pas s'enrichir sans cause et la prestation ne peut pas être supérieure au dommage.

Lorsque le bien assuré a été totalement détruit,

- s'il était destiné à la vente, le montant versé correspond à la valeur vénale
- autrement, le montant correspond à la valeur d'usage.

Lorsque le bien est partiellement endommagé, le montant du sinistre correspond au \min (coût de réparation, valeur vénale ou d'usage)



Éventuelle Contre-expertise

L'assuré peut demander une contre-expertise dont les frais restent à sa charge.



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. Par exemple, l'assureur n'aura plus à indemniser un sinistre, et l'assuré n'aura plus 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 ou de la date à laquelle l'assuré en a eu connaissance.

Résiliation du contrat d'assurance multirisque habitation

Le contrat d'assurance habitation peut être résilié par l'assuré :

- à la première échéance du contrat
- après la première année de contrat
- lors d'un changement de situation (déménagement, aggravation du risque, modification des clauses du contrat)
- suite au décès de l'assuré

Le contrat d'assurance habitation peut être résilié par l'assureur :

- à l'échéance du contrat
- en cas de non-paiement de la cotisation
- pour fausse déclaration ou omission
- pour aggravation du risque
- après un sinistre
- suite au décès de l'assuré

Loi Hamon

La loi n.2014-344 du 17 mars 2014 relative à la consommation, connue comme loi Hamon du nom du ministre délégué à l'Économie sociale et solidaire et à la consommation, est un texte de loi français ayant pour objet de renforcer les droits des consommateurs.

Entrée en vigueur le 1er janvier 2015, la loi Hamon est une loi sur la libre consommation dont le but est de mettre en place des outils économiques pour rééquilibrer les pouvoirs entre consommateurs et professionnels.

Cette loi donne aux consommateurs le pouvoir de résilier leur contrat d'assurance Auto et Habitation à la date de leur choix, passé un an de contrat.

Lorsque l'on change de couverture, le nouvel assureur prend en charge les démarches de résiliation dans un délai d'un mois et veille à la continuité de la couverture de l'assuré entre les deux contrats, à l'exception des contrats multirisque habitation pour les propriétaires (couverture non obligatoire) qui souhaitent mettre terme au contrat.

La convention CIDRE et la convention IRSI

La convention **CIDRE** est la Convention d'Indemnisation Directe et de Renonciation à Recours en dégâts des Eaux. Signée par la grande majorité des sociétés d'assurance établies en France, elle a pour objet de faciliter le règlement des sinistres dégâts des eaux dans les immeubles.

Elle a été élaborée pour simplifier et accélérer la procédure d'indemnisation. Cependant, son application est parfois sujette à controverse, notamment sur le coût important d'assurance qu'elle fait supporter aux copropriétés.

Cette convention concerne les dégâts d'eaux exclusivement. Elle pose deux principes importants :

1. Le paiement est effectué par l'assureur de la victime ;
2. L'expertise initiale est incontestable ;

La convention Cidre s'applique si le sinistre réunit les conditions suivantes :

- Intervention d'au moins deux assureurs adhérents à la Convention : celui du lésé (subissant le dommage) et celui du responsable, et même si certaines des parties ne sont pas assurées ou sont assurées auprès d'une société non adhérente. Si le lésé et le responsable sont une seule et même entité, il n'y a qu'un assureur et donc pas d'application de la convention. Ex : dans le cas où le dégât des eaux a été provoqué par l'occupant et n'a endommagé que son logement, hors partie immobilière.
- Le montant des dommages est évalué entre 250 € HT et 1600 € HT pour les dommages matériels, et en dessous de 800 € HT pour les dommages immatériels.
- et si le fait générateur du sinistres est un des suivants :
 - fuite, rupture, engorgement, débordement ou renversement
 - * d'une conduite non enterrée d'adduction et de distribution d'eau froide ou chaude, d'évacuation des eaux pluviales, ménagères ou de vidange, de chéneau et gouttière (par «conduite non enterrée», on exclut celle dont l'accès nécessite des travaux de terrassement) ;
 - * d'une installation de chauffage central à eau ou à vapeur, sauf en ce qui concerne les canalisations enterrées ;
 - * d'un appareil à effet d'eau ;
 - * d'un récipient.
 - Infiltration à travers la toiture
 - Infiltration par les joints d'étanchéité aux pourtours d'une installation sanitaire et au travers du carrelage

Depuis le 1er juin 2018, la convention CIDRE est remplacée par la convention **IRSI** (Convention d'Indemnisation et de Recours des Sinistres Immeuble).

L'une des dispositions de la convention IRSI par rapport à la convention Cidre est de fixer des règles claires de désignation des assureurs. L'assureur gestionnaire du sinistre est celui de l'occupant du logement où le sinistre a eu lieu. Il sera seul en charge de la gestion du dossier d'indemnisation.

Par rapport à la CIDRE, elle étend la simplification de la gestion aux sinistres de type incendie et augmente le montant maximum pris en charge à 5 000 € HT.

Par ailleurs, la convention IRSI définit plus précisément que la convention Cidre l'organisation des modalités de recherche de fuite. En cas de dégâts des eaux, les frais de recherche de fuite exposés avant ou au moment de la déclaration du sinistre doivent être supportés par l'assureur de celui qui a effectué les démarches. Il pourra se retourner contre l'assureur responsable de l'indemnisation si les frais sont supérieurs à 1 600 €.

Annexe B

Pré-selection des variables

B.0.1 Algorithme dirigé par la force

La première implementation de l'algorithme basée sur la force est celle publiée par Fruchtermann et Reingold. L'avantage de l'algorithme réside dans la simplicité de la mise en œuvre. De plus, cela fonctionne plutôt bien pour la plupart des graphes où le nombre d'arêtes est similaire au nombre de nœuds. Les graphes denses avec trop d'arêtes ou les graphes très clairsemés qui n'ont pratiquement aucune structure ont tendance à ne pas bien fonctionner. C'est pour ça que nous avons choisi un seuil pour enlever certaines arêtes.

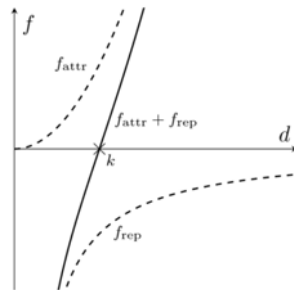
Algorithme Fruchtermann - Reingold

On suppose qu'il y a deux forces : une force attractive f_{attr} qui tire les nœuds connectés l'un vers l'autre et une force répulsive f_{rep} qui disperse les nœuds en les repoussant l'un de l'autre. La valeur absolue des forces peut être calculée comme suit :

$$f_{attr}(u, v) = \frac{k^2}{distance(u, v)}$$

$$f_{rep}(u, v) = \frac{distance^2(u, v)}{k}$$

Les directions des forces sont déterminées à partir des positions des sommets donnés sous forme de vecteurs bidimensionnels ; pour deux sommets, la direction de répulsion et d'attraction est inverse. La force complète affectant un sommet v_i est calculée en ajoutant les forces répulsives pour tous les autres sommets et les forces d'attraction pour tous les sommets connectés ensemble. Comme le montre la figure suivante, k décrit la distance entre deux sommets connectés dont les forces attractives et répulsives sont en équilibre.



Le facteur k est une constante et généralement choisie en fonction de la surface du dessin. Si la distance entre deux sommets diminue jusqu'à zéro, la force répulsive augmente indéfiniment. De même, pour deux sommets connectés, la force d'attraction croît avec la distance qui les sépare.

Cette approche fonctionne bien pour les sommets en forme de point, cependant, elle ne peut pas traiter les sommets à deux dimensions : elle ne peut pas garantir que les sommets ne se chevauchent pas, mais empêche seulement leurs centres de se toucher.

Annexe C

Explainable artificial intelligence

C.0.1 LIME

Les modèles de substitution locaux sont des modèles interprétables qui sont utilisés pour expliquer les prédictions individuelles des modèles d'apprentissage de type boîte noire. Cette méthode consiste à utiliser un modèle de substitution local qui approche au mieux le modèle de machine learning sous-jacent.

D'abord, LIME va générer aléatoirement de nouveaux individus fictifs \tilde{X} proches de l'individu sélectionné et va pondérer ces individus en fonction de leur proximité avec lui. Ensuite, on applique le modèle sous-jacent afin de reconstruire la variable à expliquer correspondante $\hat{y} = f(\tilde{X})$ et calcule un modèle linéaire sur la base de ces nouvelles données, en pondérant chaque observation de \tilde{X} en fonction de sa proximité avec les données initiales. Le modèle linéaire que LIME a créé est facile à interpréter et nous permet donc d'avoir une explication pour cet individu.

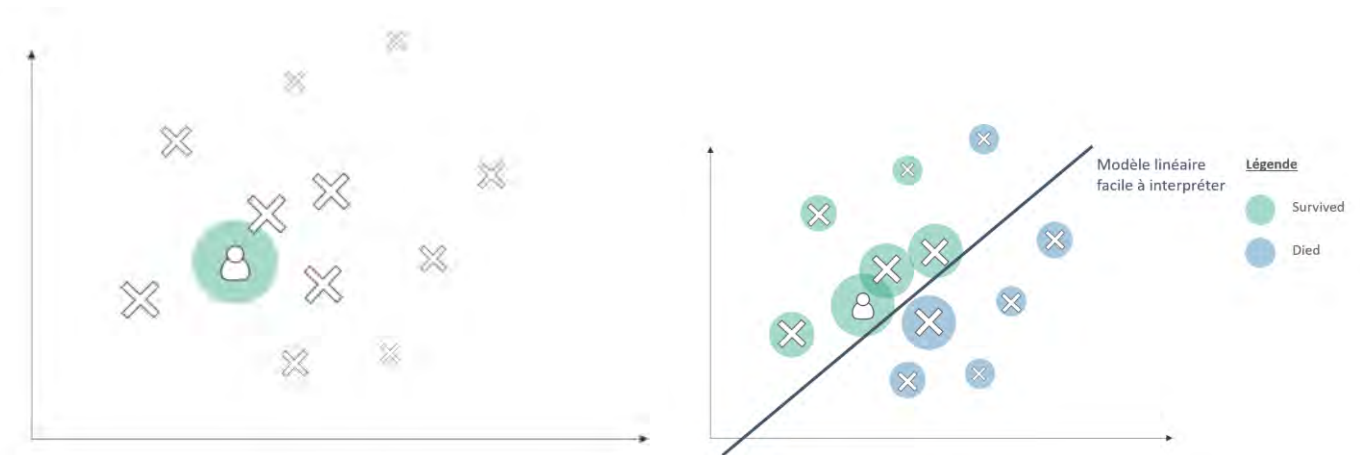


FIGURE C.1 – Procédure d'explication LIME sur un modèle de classification

Le modèle d'explication associé à cette méthode est solution du problème d'optimisation suivant :

$$\hat{g} = \underset{g \in \mathcal{G}}{\operatorname{argmin}} [L(f, g, \pi_{x'}) + \Omega(g)]$$

où :

- L est la fonction de cout
- f le modèle sous-jacent
- g la fonction associé au modèle d'explication qu'on souhaite optimiser
- \mathcal{G} classe des modèle d'explication
- $\pi_{x'}$ une mesure de proximité définissant la taille du voisinage autour de l'individu initiale, qui représente le poids donnée à chaque observation x' de la base de données perturbée
- Ω une fonction traduisant la complexité d'un modèle

LIME vs KernelSHAP

La grande différence entre LIME et KernelSHAP est la pondération des observations dans le modèle de régression. LIME pondère les observations en fonction de leur proximité avec l'individu d'origine. Plus il y a de 0 dans le vecteur de coalition, plus le poids donné par LIME est petit. SHAP pondère les observations échantillonnées en fonction du poids que la coalition obtiendrait dans l'estimation de la valeur de Shapley. Les petites coalitions (quelques 1) et les grandes coalitions (c'est-à-dire plusieurs 1) obtiennent les pondérations les plus importantes.

L'intuition sous-jacente est la suivante : nous en apprenons le plus sur les caractéristiques individuelles si nous pouvons étudier leurs effets individuels. Si une coalition se compose d'une seule caractéristique, nous pouvons en apprendre davantage sur l'effet principal isolé des caractéristiques sur la prédiction. Si une coalition se compose de toutes les variables sauf une, nous pouvons en apprendre davantage sur l'effet total de ces variables (effet principal et interactions des fonctionnalités). Si une coalition comprend la moitié des variables, nous en apprenons peu sur la contribution d'une variable individuelle, car il existe de nombreuses coalitions possibles avec la moitié des fonctionnalités.

Annexe D

Tarification à l'adresse

D.1 Une parenthèse sur la confiance des données

Pour chaque contrat de la base interne est associé un bâtiment et des attributs, mais des erreurs peuvent se réaliser pendant cette association.

Des mesures de confiance ont été ainsi attribuées pour chaque contrat.

Il y a deux types de variables de confiance indépendants entre eux :

1. la confiance du géocodage, c'est-à-dire pendant les étapes d'association d'une adresse à un bâtiment.
2. la confiance sur la qualité des attributs (ex. surface parcelle, type de toit,...) qui dépend de la source ou de la méthode qui a permis de le générer.

Confiance du Géocodage

La confiance du géocodage considère deux confidences :

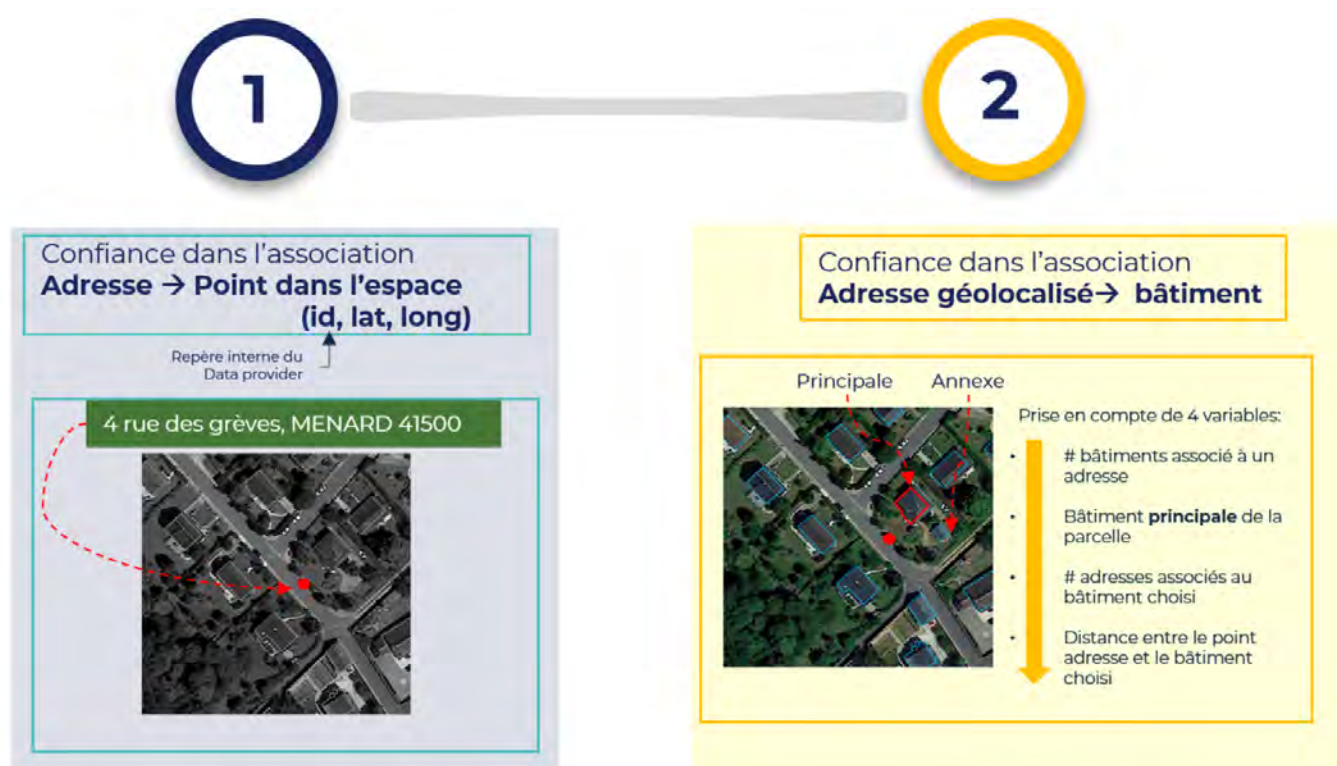


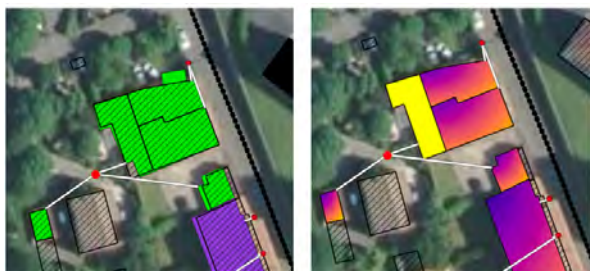
FIGURE D.1 – Confiance du Géocodage : la première confiance est calculée à partir d'un score de match, entre 0 à 1, entre l'adresse en entrée et un référentiel interne; et la deuxième est un score entre 0 et 1 été calculée avec plusieurs critères d'association (type et vecteur de relation)

Les scores de confiance sont restitués par le fournisseur des données et convertis en variable catégorielles à

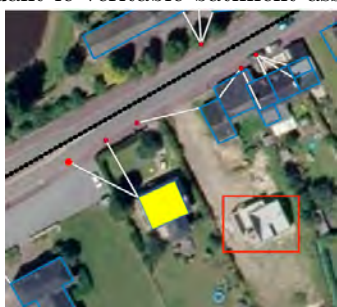
modalité : *Very Low, Low, Medium, High et Very High*¹

Utiliser des bons indices de confiance permet d'éviter les situations suivantes :

- lorsque plusieurs bâtiments sont reliés à la même adresse (bâtiments principaux ou annexes), mais rien ne permet de déterminer si c'est la maison ou l'annexe qui a été assurée



- lorsque l'adresse a bien été géo-localisée mais le bâtiment associé est le bâtiment jaune et non le rouge qui est cependant le véritable bâtiment assuré. Dans ce cas le lien entre l'adresse géo-localisée et le bâtiment



est faible.

Nombre d'adresse, nombre de batiments, batiments principal



FIGURE D.2 – Il est possible que un bâtiment présente deux maison, dont un studio par exemple. Dans ce cas, les adresses seront bien géolocalisés, mais affectés au même bâtiment. Les caractéristiques du bâtiment seront ainsi les mêmes pour les deux maisons car on ne prend pas en compte la dissociation des deux surfaces habitables. il en résulte que la surface habitable sera sur-estimée.

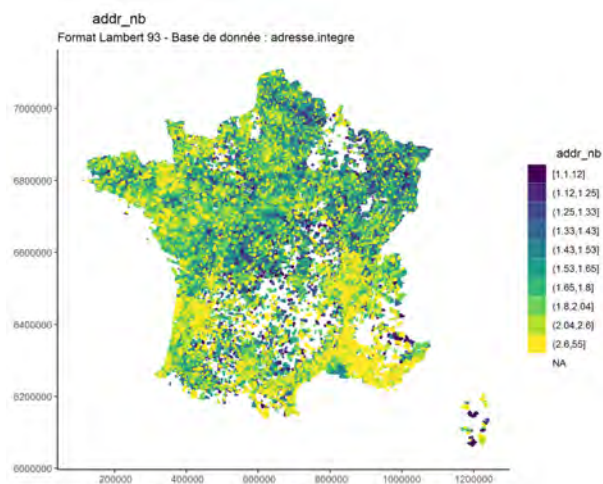


FIGURE D.3 – Nombre d'adresses associées au même bâtiment (moyenne à la maille IRIS) : plus il y a d'adresses associées à un bâtiment, plus cela sous-entend la présence de studio, de maison partagées, de construction nouvelles ou d'entreprises sur la même parcelle. Environ 25% des bâtiments ont plusieurs adresses.

¹On ne peut pas comparer les niveaux de confiance de plusieurs variables, que ce soit la confiance du géocodage ou celle relative aux autres variables ajoutées, car chaque confiance repose sur des méthodes différents. Par exemple, il est possible que une variable à confiance *Medium* soit plus fiable que un autre variable *High*.

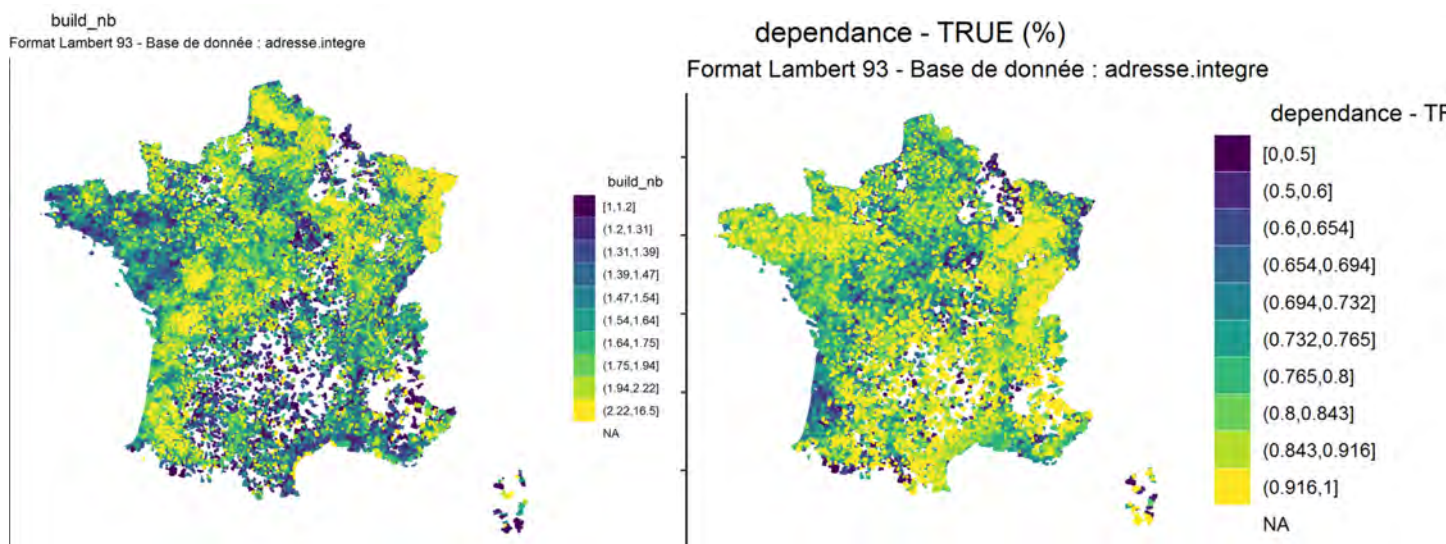


FIGURE D.4 – A gauche : nombre de bâtiments associées à la même adresse (moyenne à la maille IRIS) ; à droite : présence des dépendance (moyenne à la maille IRIS). Le nombre de bâtiments associés avec une même adresse est très fortement corrélé à la présence de dépendance ainsi qu'à la ruralité (en particulier au Sud de la France)

Visualisation de la confiance du Géocodage

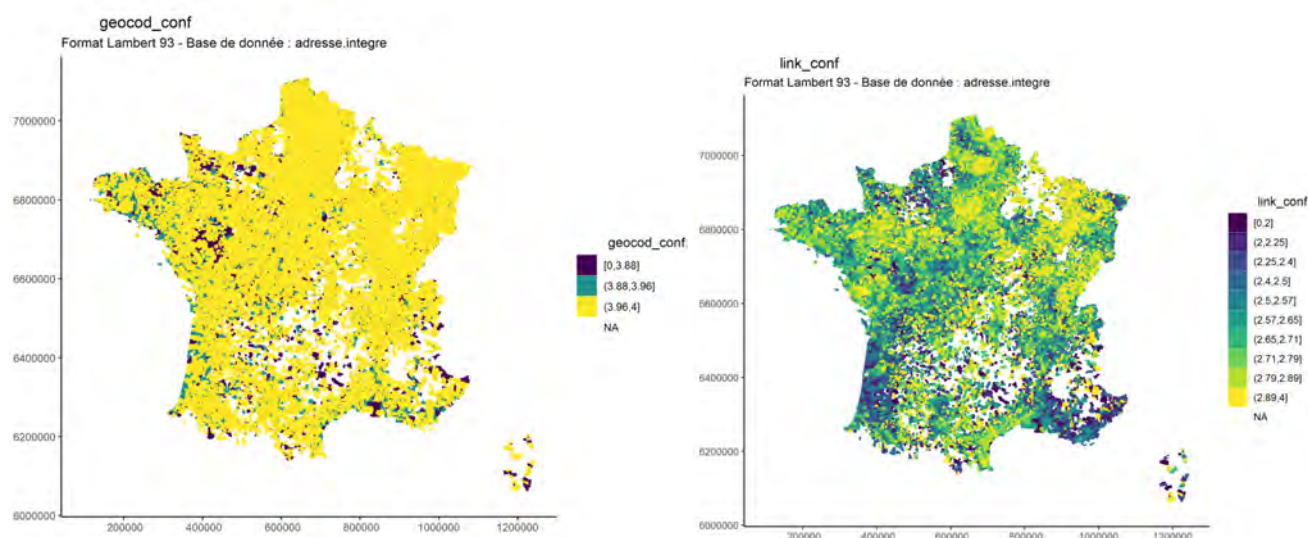


FIGURE D.5 – A gauche, la confiance du lien entre l'adresse et le géo-codeur : la confiance est très élevée (99% des adresses est bien géolocalisé) ; à droite, la confiance du lien entre l'adresse et le bâtiment : plus de 75 % des bâtiments sont liés avec une qualité moyenne (3) et plus 50% avec une qualité élevée (4).

Pour ne pas biaiser nos analyses par la qualité des variables, nous avons sélectionné dans le cadre de cet étude environ 80% des données, où les observations ont :

- la confiance dans la qualité du géo-coding supérieure à *High*
- la confiance du lien entre l'adresse et le géo-codeur supérieure à *Medium*
- le nombre d'adresses associé au même bâtiment < 5
- le nombre de bâtiments associé au même adresse < 5

Confiance des attributs

La confiance des attributs est définie comme le minimum des confiances de chaque base de données utilisée (Externe (payante ou open), provenant d'un modèle prédictif), du géocodage éventuel et des jointures effectuées :

$$\text{Confiance attribut} = \min(\text{Confiance}_{\text{Dataset}_1}, \text{Confiance}_{\text{Dataset}_2}, \dots)$$

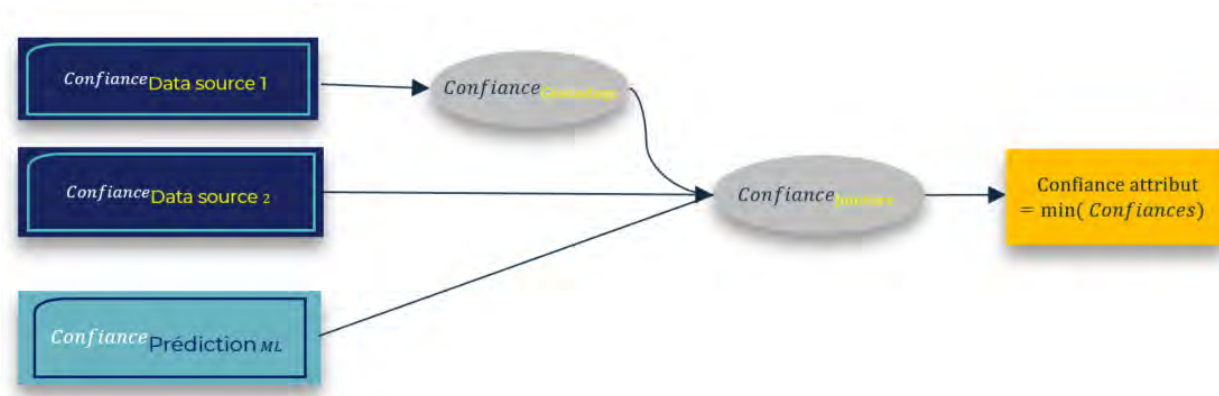


FIGURE D.6 – La confiance de l'attribut hérite la confiance des données sources, du géocodage, des jointures et de la règle métier et de prédiction. La confiance des prédictions dépend des métriques du modèle (MAE, RMSE,...)

Source : Nam.R

D.2 Analyse descriptive

Après la jonction de deux bases et le filtre sur la confiance, nous avons sorti des statistiques des bases sur les trois années.

Pour une question de confidentialité, les chiffres sensibles (charge totale des sinistres du portefeuille) ne seront pas affichés.

Granularité des données selon le découpage géographique

les données reçues ont plusieurs mailles de référence :

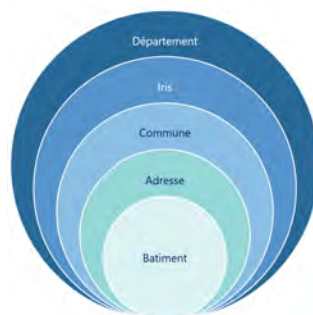


FIGURE D.7 – Niveaux d'agrégation des données utilisés dans l'étude : du plus grand (département) à la répartition plus fine (bâtiment)

IRIS

Définition 31. Le découpage en **IRIS** « Îlots Regroupés pour l'Information Statistique » est une répartition du territoire français en mailles de taille homogène, introduit par l'INSEE en 1999 pour préparer la diffusion du recensement de la population. Depuis, l'IRIS constitue la brique de base en matière de diffusion de données infra-communales.

Quelques variables

Après le filtre sur la confiance, nous avons affiché des statistiques de base.

Nombre de pièces

Il s'agit d'une variable interne, posé par la plupart des questionnaire MRH en 2020.

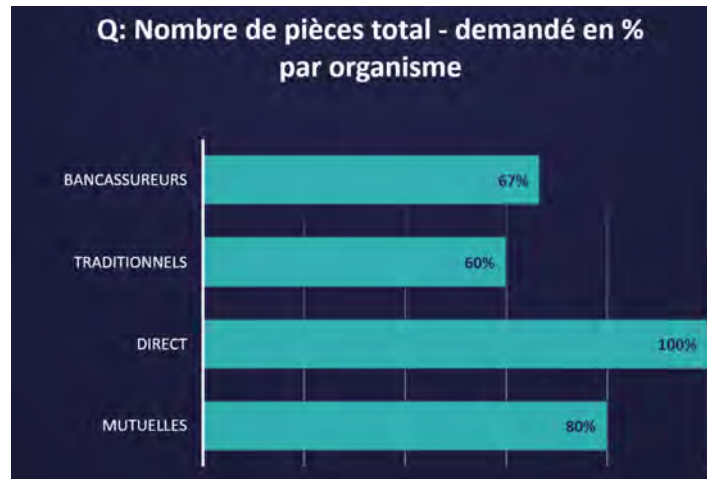


FIGURE D.8 – Organismes qui utilisent cette variable comme critère tarifaire
Source : Benchmark interne, Mai 2020

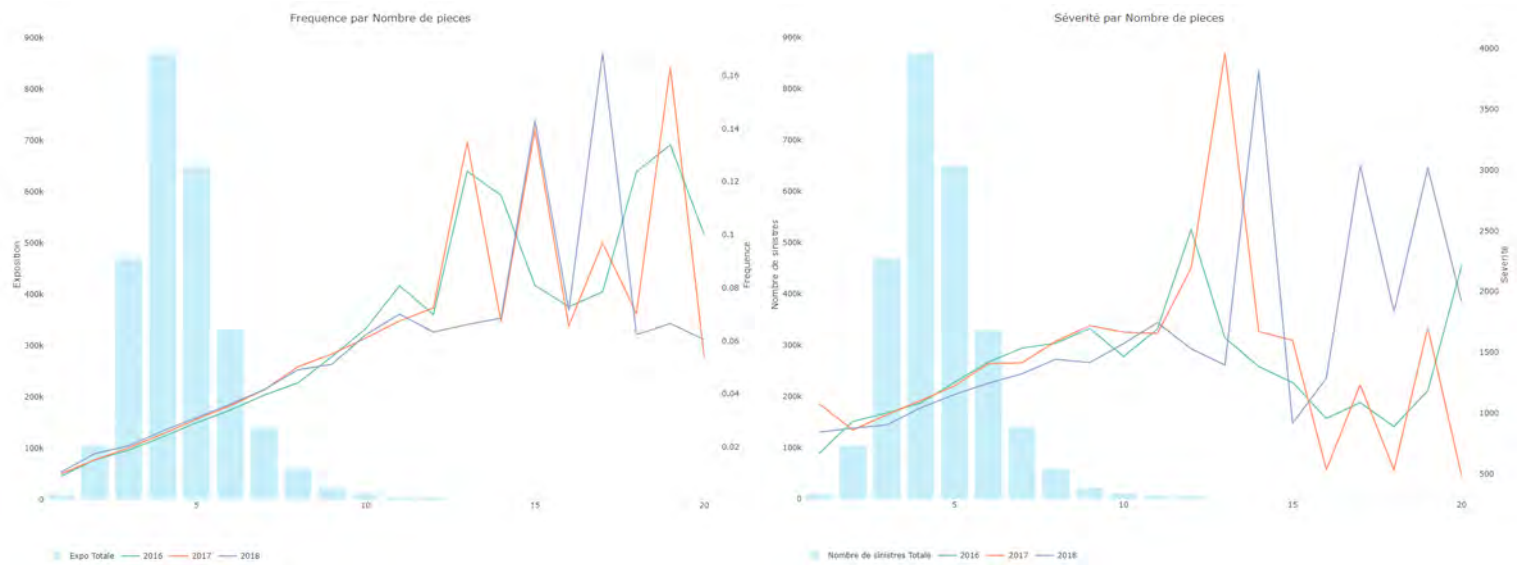


FIGURE D.9 – Sinistralité par Nombre de pièces : on remarque une forte volatilité pour des nombres de pièces supérieurs

Type d'assuré

Il s'agit d'une variable interne.

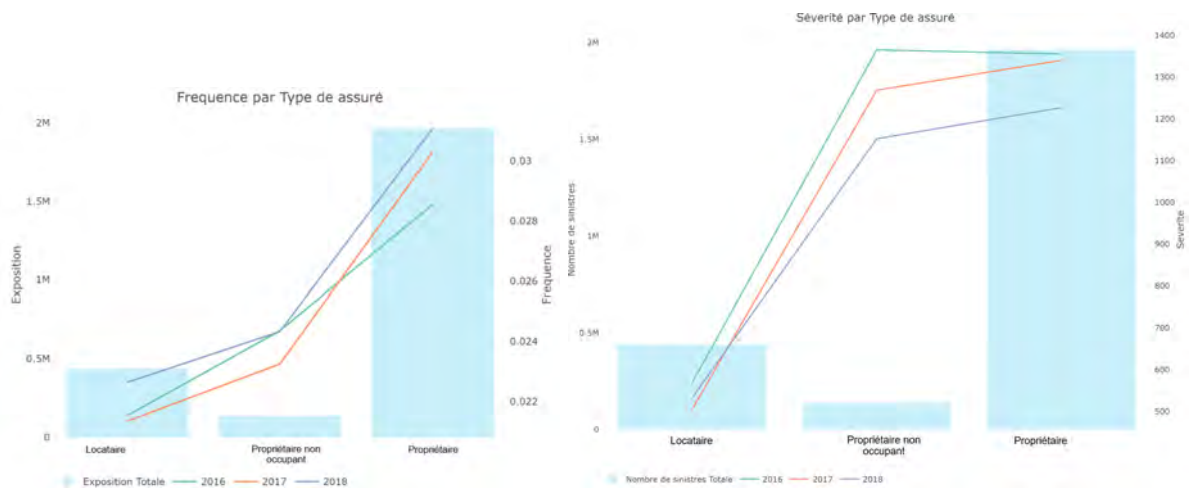


FIGURE D.10 – Sinistralité par type d'assuré : la plupart du risque est porté par les propriétaires

Âge

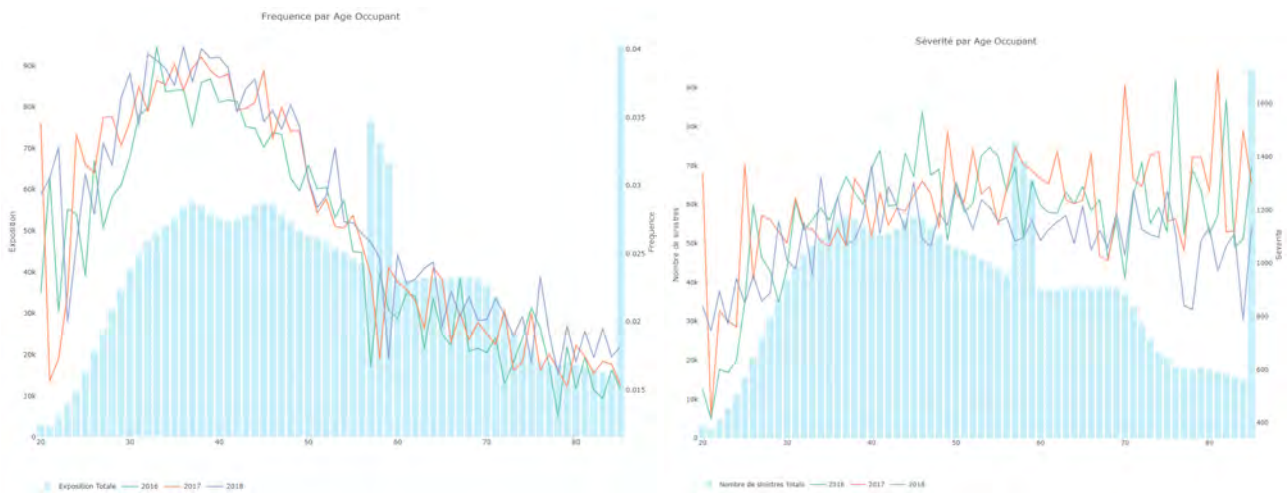


FIGURE D.11 – Sinistralité par Âge Occupant : la classe "20" contient tous les âges inférieurs à 20 et la classe "85" contient tous les âges supérieurs ou égaux à 85 ans

Surface habitable

Le calcul de la surface habitable se fait à l'aide de l'analyse d'image après que l'adresse a été renseigné, la parcelle a été détectée, le bâtiment a été choisi et le nombre d'étages (variable de qualité élevée) a été déterminé. C'est une variable souvent traitée de façon catégorielle sur les questionnaires en ligne.

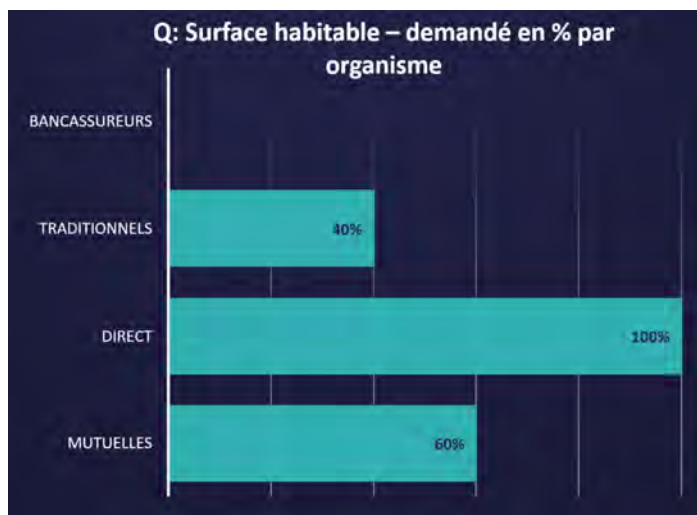


FIGURE D.12 – Source : Benchmark interne, Mai 2020

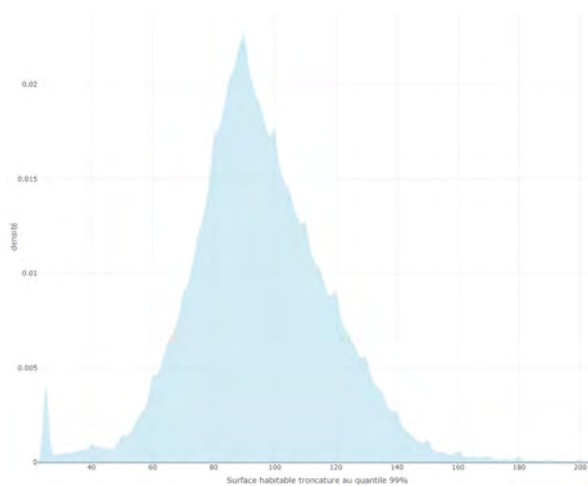


FIGURE D.13 – Distribution de la surface habitable : les observations extrêmes correspondent à des maisons singulières ou à des cas où la variable a une confiance faible.

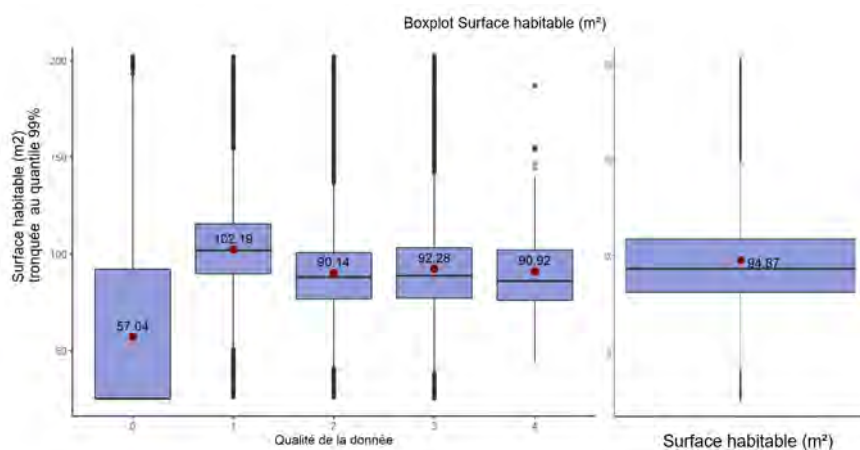


FIGURE D.14 – Boxplot de la surface habitable : certaines valeurs extrêmes correspondent à des maisons singulières. On remarque que la confiance de cette variable est faible pour des surfaces de moyenne 57m², ce qui fait penser au cas où on associe le même bâtiment à plusieurs adresse car les studios se trouvent dans un immeuble plus grand.

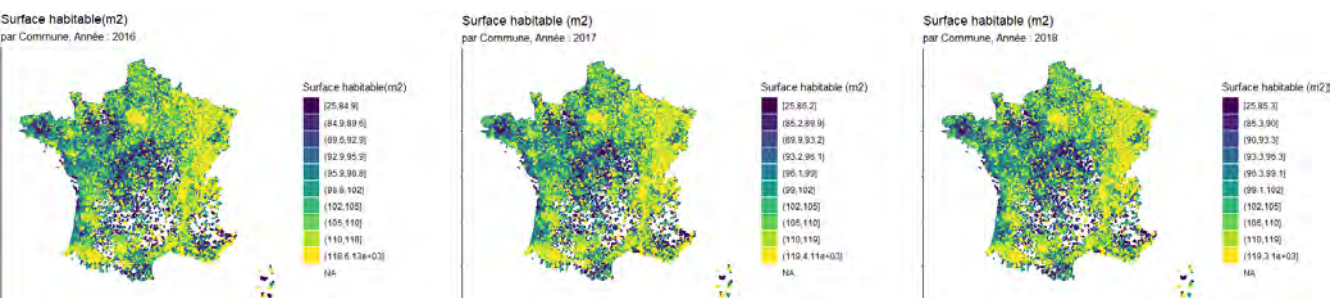


FIGURE D.15 – Surface habitable en m² moyenne par commune

Période de Construction

La période de construction est une variable qui permet de prendre en compte la vétusté du bâtiment et permet d'intégrer dans les modèles de fréquence des sinistres l'effet des réglementations.

Cette variable est produite au 92% par des méthodes de machine learning, à partir des sources : IGN, ADAME et MTES.

Elle a été découpé de façon à considérer la réglementation thermique RT2012² en 10 classes :

1. Avant 1915
2. 1915-1949
3. 1950-1968
4. 1969-1975
5. 1976-1982
6. 1983-1990
7. 1991-2000
8. 2001-2006
9. 2007-2013
10. Après 2013

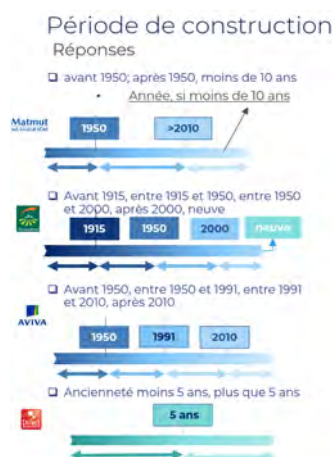


FIGURE D.16 – Prise en compte de la période de construction dans le marché habitation
Source : Benchmark interne, Mai 2020

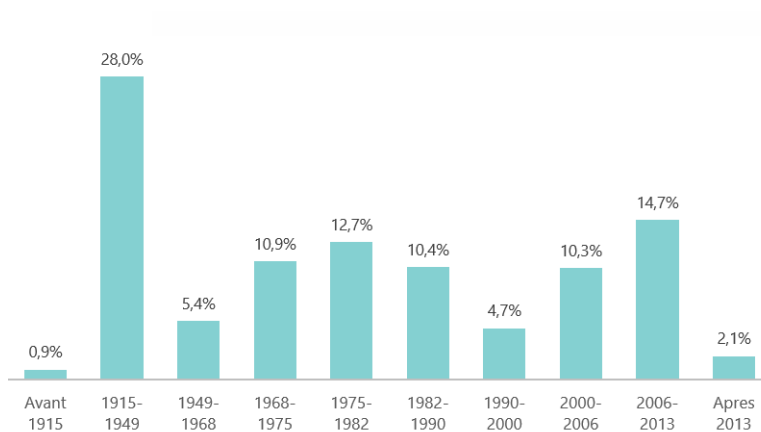


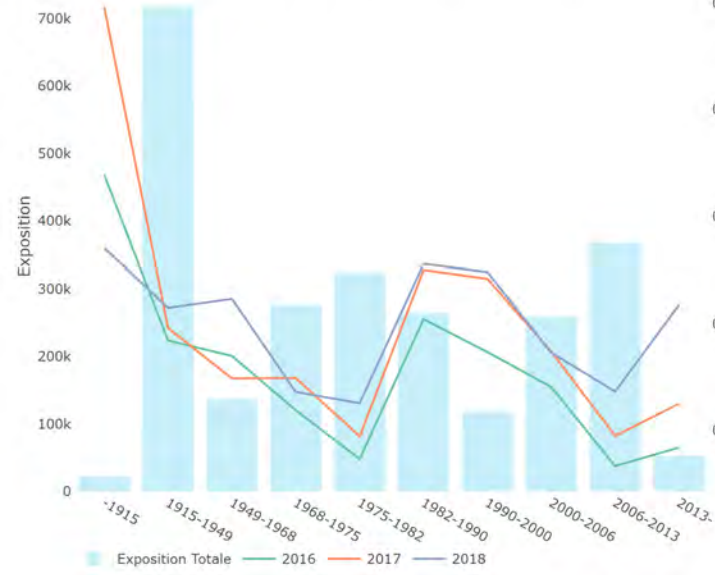
FIGURE D.17 – Répartition des bâtiments dans le portefeuille selon leur période de construction

²La réglementation thermique RT 2012 (*) Sur la consommation d'énergie. La RT 2012 a pour objectif de limiter la consommation d'énergie primaire des bâtiments neufs à un maximum de 50 kWhEP/(m².an) en moyenne. La RT 2012 est applicable à tous les permis de construire :

- déposés depuis le 28 octobre 2011 pour certains bâtiments neufs du secteur tertiaire (bureaux, bâtiments d'enseignement primaire et secondaire, établissements d'accueil de la petite enfance) et les bâtiments à usage d'habitation construits en zone ANRU ;
- déposés depuis le 1er janvier 2013 pour tous les autres bâtiments neufs

Source : <https://www.ecologique-solidaire.gouv.fr/exigences-reglementaires-construction-des-batiments>

Frequence par Periode de construction



Severité par Periode de construction



FIGURE D.18 – Période de construction

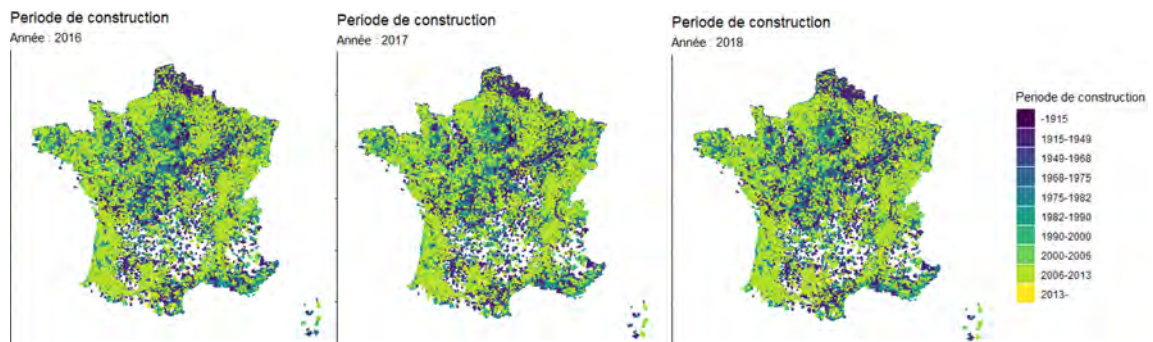


FIGURE D.19 – Période de construction plus fréquent pas commune

Valeur de la maison

La valeur de la maison est une variable calculée à la maille IRIS à l'aide de la base DVF. Cette variable permet de déterminer la richesse de l'aire urbaine associée (prix au mètre carré).

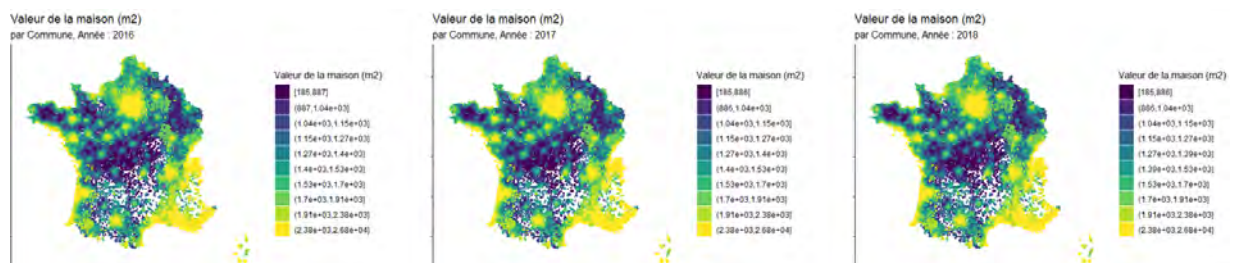


FIGURE D.20 – Valeur de la maison en €/ m² moyenne par commune

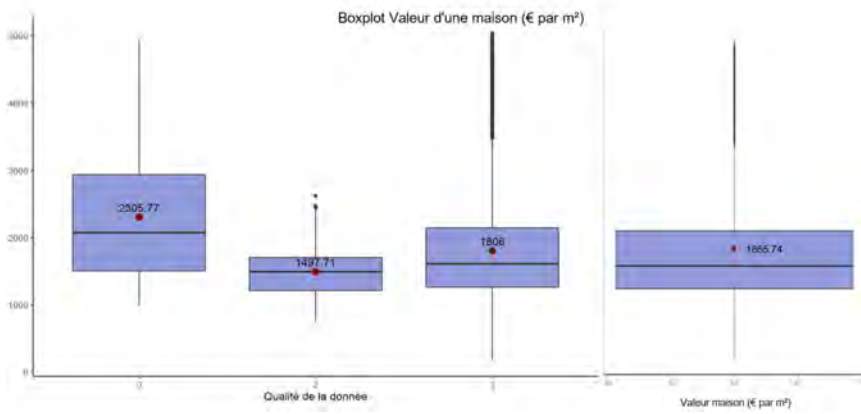


FIGURE D.21 – Boxplot du prix au mètre carré des maisons

Nombre de bâtiment dans un rayon de 50 mètres

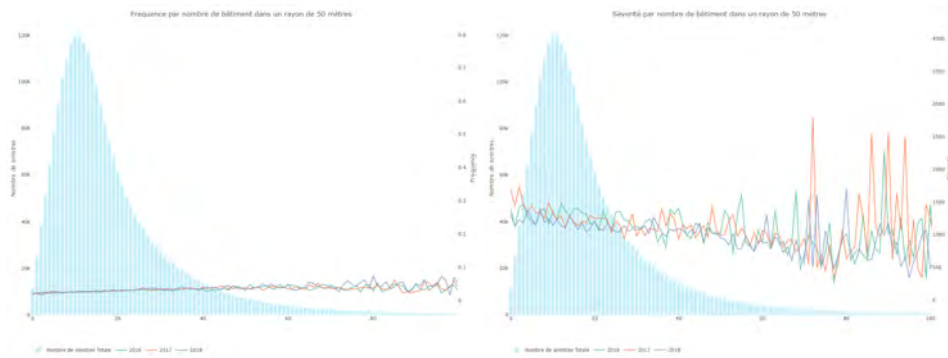


FIGURE D.22 – Nombre de bâtiment dans un rayon de 50 mètres : pour éviter trop de volatilité sur les valeurs extremes, cette variable sera retraitée.

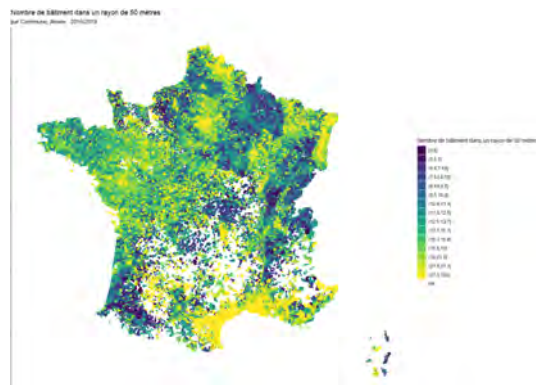


FIGURE D.23 – Nombre de bâtiment dans un rayon de 50 mètres

Annexe E

Méthodes de pénalisation pour la sélection finale

E.1 Sélection via régression Lasso

Dans le cadre de la régression linéaire, où on souhaite estimer le vecteur $\beta \in \mathbb{R}^D$ qui vérifie $Y = X \beta + \epsilon$, où D est le nombre de variables, on se sert souvent de méthodes dites pénalisées : à partir de l'estimateur des moindres carrés

$$\hat{\beta}^{MC} = \underset{\beta \in \mathbb{R}^D}{\operatorname{argmin}} \|Y - X \beta\|_2^2$$

on ajoute une pénalité envers la complexité du modèle, c'est-à-dire qu'on ajoute de l'information à un problème, s'il est mal posé ou pour éviter le sur-apprentissage.

La méthode Lasso est une méthode de pénalisation qui a l'avantage de permettre de sélectionner les variables plus importantes : β des coefficients estimé peut être "sparse" (il contient beaucoup de zéros). Comme le nombre de variables explicatives est encore élevé, nous avons retenu intéressant d'utiliser cette méthode pour contraindre les coefficients.

Robert Tibshirani a développé la méthode du lasso, une méthode de contraction des coefficients, qui consiste à introduire dans la fonction de vraisemblance une pénalisation proportionnelle à la norme $L1$ du vecteur β des coefficients :

$$\hat{\beta}^{Lasso} = \underset{\beta \in \mathbb{R}^D}{\operatorname{argmin}} \frac{1}{2N} \|Y - X \beta\|_2^2 + \lambda \|\beta\|_1$$

avec $\|\beta\|_1 := \sum_{j=1}^D |\beta_j|$ norme $L1$ du vecteur β des coefficients, N taille de l'échantillon d'apprentissage. Le paramètre λ contrôle la complexité du modèle ou le degré de généralisation du modèle : des faibles valeurs de λ contractent moins les coefficients et permettent au modèle de s'ajuster plus finement aux données, alors qu'un λ élevé conduit à un modèle plus parcimonieux et moins ajusté.

L'estimation de λ est souvent issue d'une validation croisée, pour éviter le sur-apprentissage ou d'utiliser un modèle trop simple.

- D'abord nous déterminons le λ optimale par validation croisée.

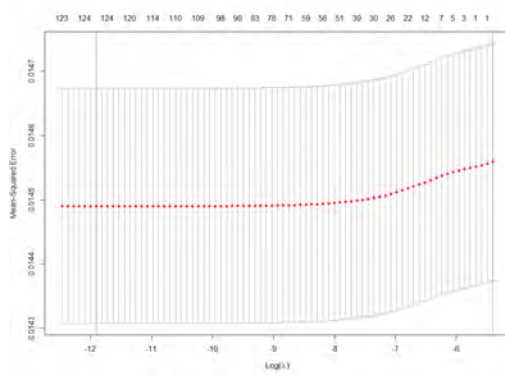


FIGURE E.1 – Estimation du paramètre de complexité λ par validation croisée 5-folds basées sur le MSE : comme λ_{opt} optimal est proche de zéro ($6.718791e-06$), nous remarquons que l'estimateur par Lasso est proche de celui des moindres carrés.

- Une fois que la valeur optimale de λ est déterminée, nous estimons le vecteur des coefficients β pour sélectionner les variables à coefficients significatifs (non nuls) que nous retiendrons in fine pour le modèle GLM.

Annexe F

Modèle de tarification par Machine Learning

F.1 Optimisation des modèles de fréquence

Optimisation de la forêt aléatoire

Dans un algorithme de forêt aléatoire il n'y a que quelques paramètres à ajuster sont :

- *ntrees* : le nombre d'arbre ;
- *mtry* : le nombre de variables testées à chaque division. La valeur par défaut étant le nombre de variables divisé par 3 (régression). C'est le principal paramètre à modifier, avec le nombre d'arbre ;
- *nodesize* : le nombre minimale d'observations dans les nœuds terminaux. (plus elle est élevée, plus les arbres seront courts) ;
- *samplesize* : qui fixe la taille de l'échantillon "in bag" (par défaut est 63,25% de l'ensemble d'apprentissage)
- *maxnodes* : le nombre maximum de nœuds terminaux.

Pour paramétrer le Random Forest, nous avons construit une grille de 120 modèles (Random Search) en combinant les paramètres principaux en laissant les autres paramètres par défaut :

- pour chaque couple de paramètres (# Arbres, # variables par division) nous avons construit un modèle sur la base d'apprentissage ;
- pour choisir les paramètres, en évitant le sur-apprentissage, nous avons testé chaque modèle sur la base de validation et retenu les paramètres appropriés

En dehors de l'optimisation du code et des algorithmes, une autre façon d'obtenir un code performant, nous avons parallélisé les algorithmes afin de faire des opérations simultanées sur des parties distinctes d'un même problème, en utilisant différents cœurs de calcul. On ne réduit pas le temps de calcul total nécessaire, mais l'ensemble des opérations s'exécute plus rapidement.

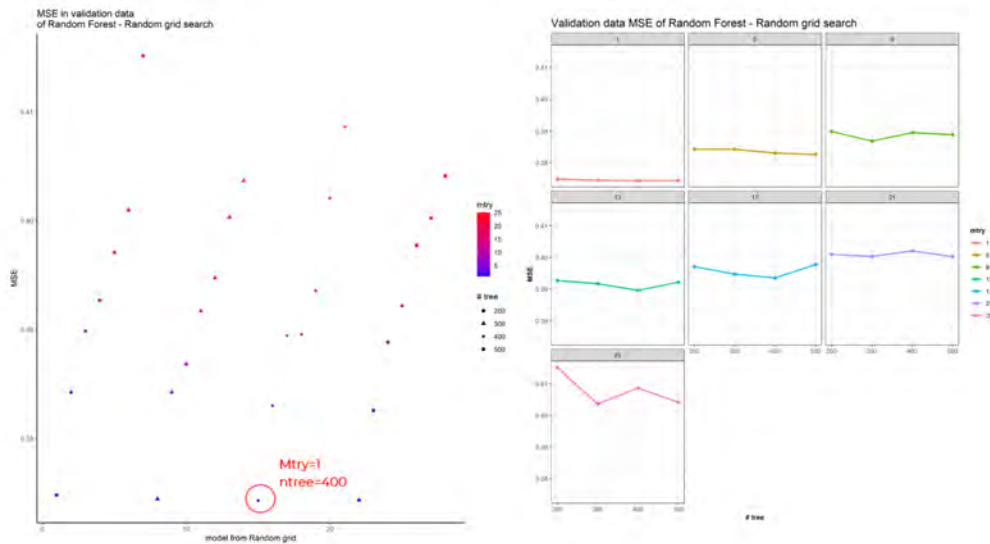


FIGURE F.1 – Recherche des paramètres $mtry$ et $ntrees$: pour éviter le sur-apprentissage nous avons déterminé les meilleurs paramètres au sens du MSE dans la base de validation.

Nous avons ainsi choisi :

| Paramètre | Valeur |
|-----------|--------|
| $ntree$ | 400 |
| $mtry$ | 1 |

Optimisation de xgboost

Avant d'exécuter XGBoost, nous devons définir trois types de paramètres :

- les **paramètres généraux** concernent le booster que nous utilisons pour faire du boosting, généralement un modèle de type arbre ou linéaire ;
- les paramètres de la **tâche d'apprentissage** décident du scénario d'apprentissage ;
- et les paramètres **de ligne** de commande concernent le comportement du modèle XGBoost.

Sous le logiciel R l'algorithme `xgboost` a plusieurs bonnes propriétés : la possibilité de paralléliser les tâches (donc économiser le temps machine), une fonction de validation croisée interne, la capacité à gérer les valeurs manquantes et la capacité à élaguer l'arbre jusqu'à ce que l'amélioration de la fonction de perte soit inférieure à un seuil.

Pour cela nous avons pu ajuster plus de paramètres que pour Random Forest :

- $nrounds$: le nombre d'itérations boosting
- η : le paramètre d'élagage. C'est un paramètre de pénalisation qui permet de limiter la taille des arbres. Il est équivalent au coefficient γ ;
- $subsample$ est un paramètre qui détermine le pourcentage d'individus de l'échantillon à utiliser à chaque itération pour construire une règle de prédiction ;
- γ : gain minimale pour splitter un noeud de l'arbre
- $colsample_bytree$, le pourcentage de variables à utiliser à chaque itération pour construire un prédicteur. Il est conseillé d'utiliser de valeurs entre 0.3 et 0.8 en grande dimension.
- max_depth , la profondeur maximale des arbres utilisés et aide à éviter le sur-apprentissage ;

- et *min child weight* : le nombre minimale d'observation dans une partition. Plus il est grand, plus l'algorithme sera conservateur.

Puisque un choix arbitraire des hyperparamètres peut causer le sur-apprentissage du modèle, nous les avons optimisé à l'aide de validation-croisée 5-fold.

Choix des paramètres

Étant la base de données très volumineuse et un grand nombre de paramètres nous avons réduire la taille de la base à un échantillon représentatif et puis appliquer la technique de Random Search, en étudiant les paramètres en deux blocs :

1. dans le premier bloc, nous avons étudié l'effet combiné de la profondeur maximale des arbres (*max depth*), *nrounds* et *eta*, car ils sont liés
2. et dans le deuxième bloc les autres paramètres.

Dans chaque blocs les paramètres qui ne seront pas analysés seront fixés à des valeurs trouvés dans la littérature actuarielle (modèle notée "benchmark").

Dans notre exemple nous avons crée une grille de 50 modèles par bloc de paramètres et les paramètres sont choisis en utilisant comme fonction de perte la déviance poissonienne.

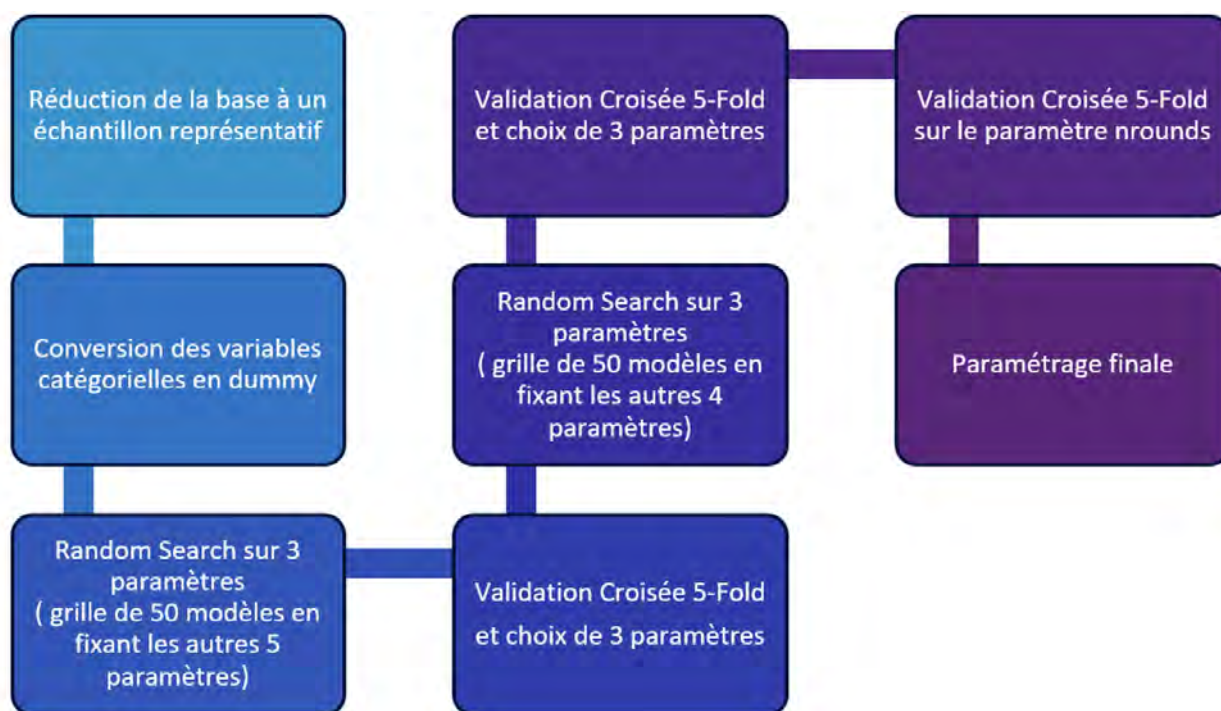


FIGURE F.2 – Processus d'optimisation

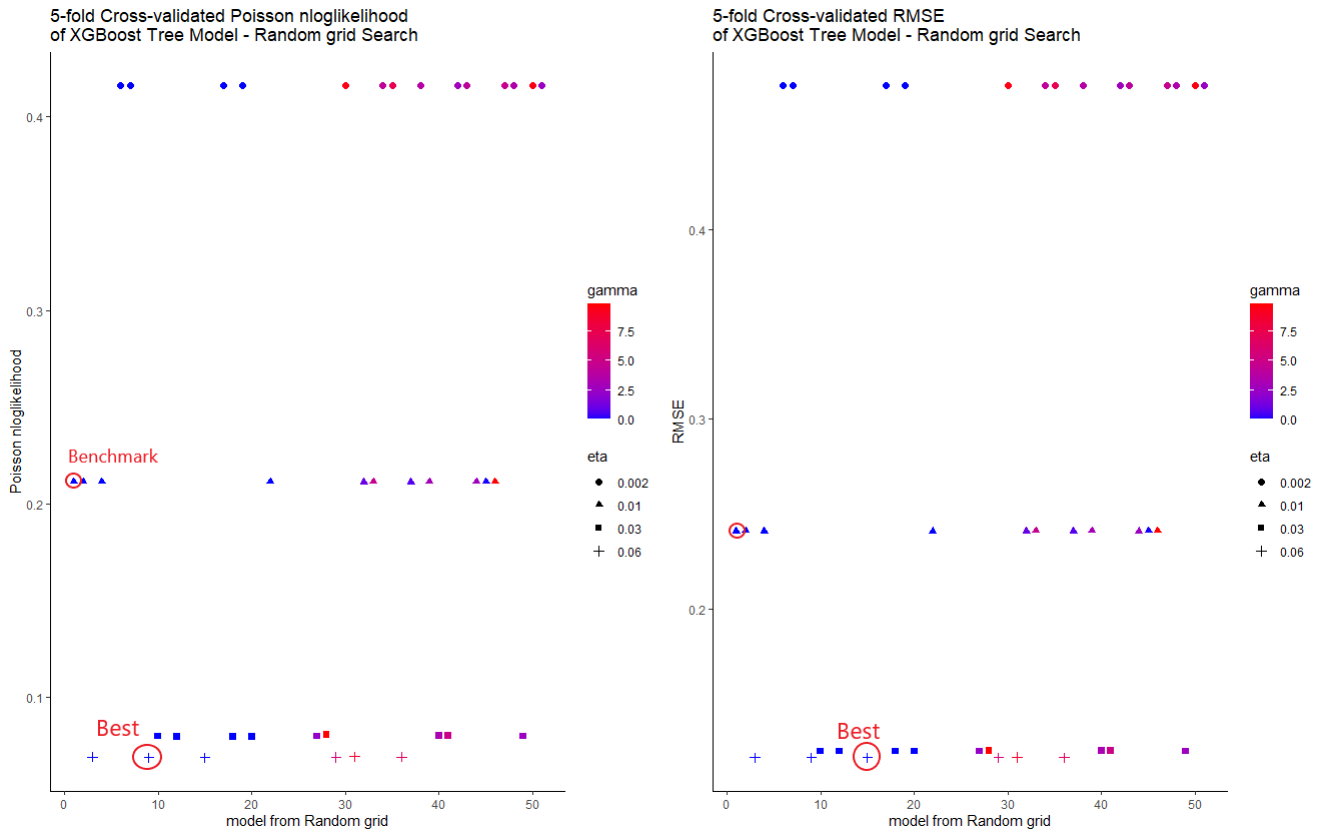


FIGURE F.3 – Choix des paramètres η , γ , Profondeur des arbres.

Les paramètres retenus après optimisation sont :

| Paramètre | Valeur |
|------------------|--------|
| η | 0.06 |
| γ | 0 |
| Max depth | 3 |
| subsample | 0.8 |
| colsample bytree | 0.8 |
| nrounds | 200 |
| min child weight | 7 |

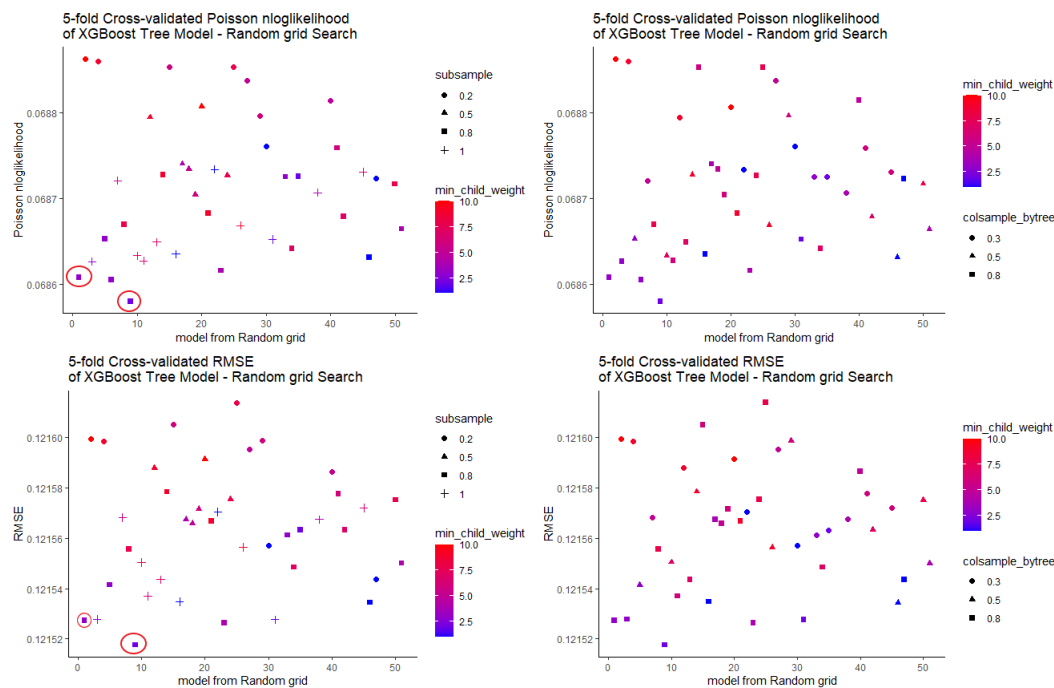


FIGURE F.4 – Choix des paramètres *subsample*, *colsample_bytree* et la somme minimale des poids par noeud fils (après le choix des paramètres η , γ , profondeur des arbres)

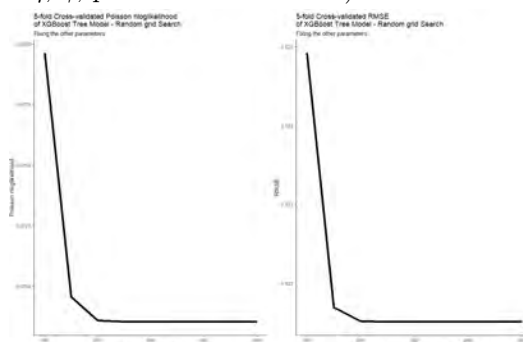


FIGURE F.5 – Choix du nombre d'itération : à gauche en utilisant la déviance poissonnienne comme fonction de perte, à droite le RMSE.

F.2 Optimisation des modèles de sévérité

Optimisation de la forêt aléatoire

Nous avons procédé comme pour la fréquence et appliqué la procédure de calibrage de paramètre en utilisant comme valeur cible la *Sévérité as if*. Nous avons construit une grille 120 modèles sur la base d'apprentissage complète où :

- $n_{tree} \in \{150, 200, 250, 300, 350, 400, 450, 500\}$
- $m_{try} \in \{1, \dots, 15\}$

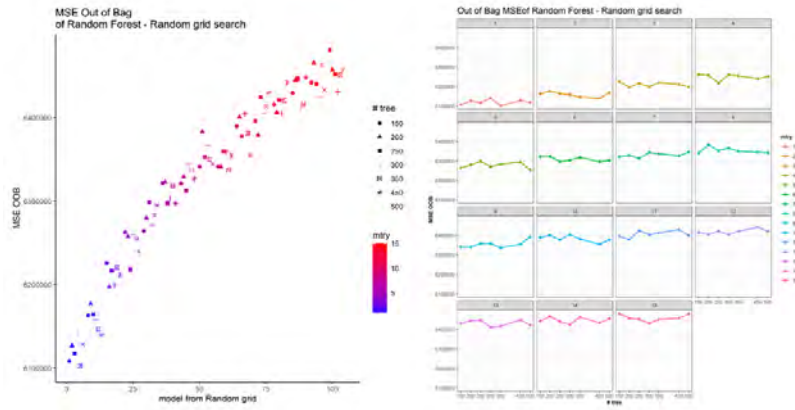


FIGURE F.6 – Procédure Random Search basée sur l’erreur quadratique moyen la base d’apprentissage.

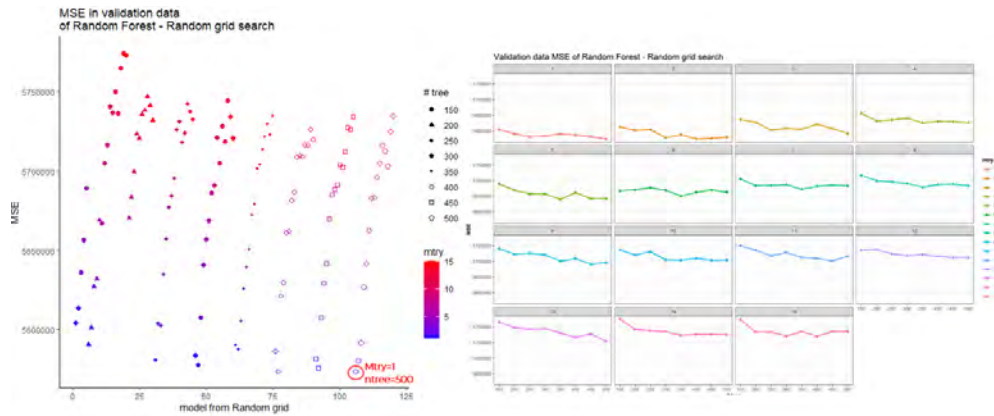


FIGURE F.7 – Procédure Random Search basée sur l’erreur quadratique moyen sur la base de validation.

Nous avons ainsi choisi :

| Paramètre | Valeur |
|-----------|--------|
| ntree | 500 |
| mtry | 1 |

Optimisation de xgboost

Le modèle Extreme Gradient Boosting (XGBoost) plus adapté à la prédiction de la sévérité utilise comme fonction objectif la *régression gamma* et comme métriques d’évaluation du modèle le RMSE et la déviance gamma.

L’optimisation des paramètres suit les étapes effectuées pour la fréquence : nous avons testé 50 modèle et choisit la combinaison de paramètres qui minimise la déviance gamma et le RMSE.

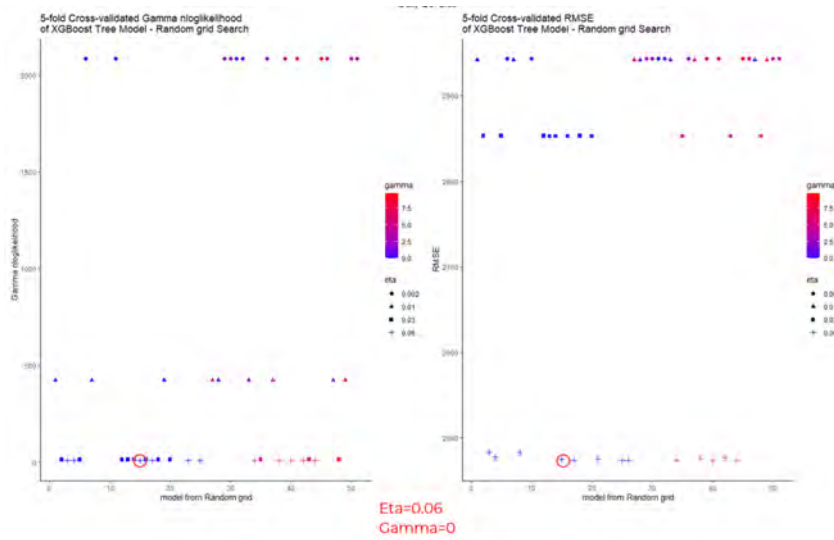


FIGURE F.8 – Choix des paramètres de calibrage de Xgboost pour la sévérité.

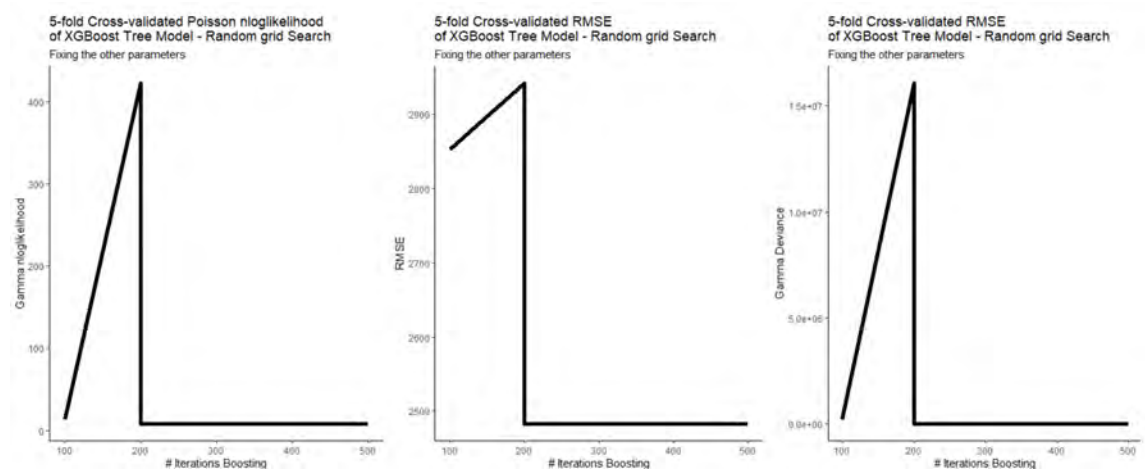


FIGURE F.9 – Choix du nombre d’itérations selon plusieurs métriques d’évaluation du modèle Xgboost de sévérité.

Les paramètres retenus après optimisation sont :

| Paramètre | Valeur |
|------------------|--------|
| η | 0.06 |
| γ | 0 |
| Max depth | 7 |
| subsample | 0.8 |
| colsample bytree | 0.3 |
| nrounds | 200 |
| min child weight | 10 |

Annexe G

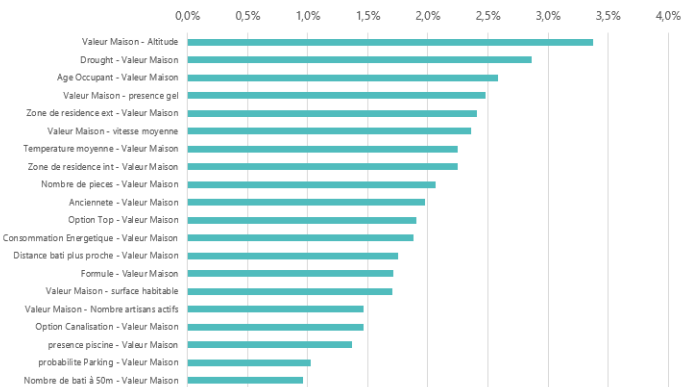
Pratique de l'analyse de sensibilité

Indices de Sobol d'ordre 2

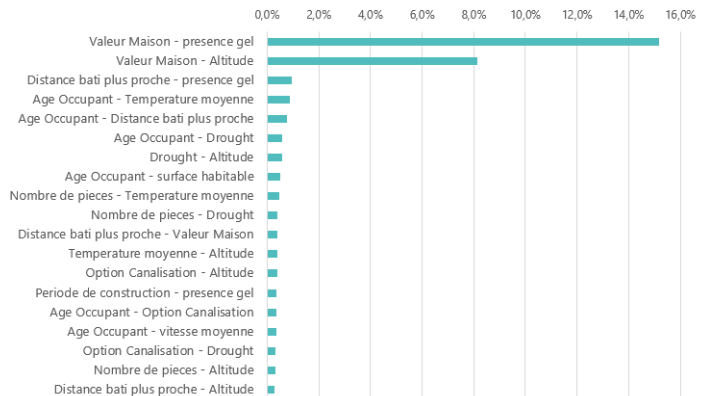
Nous visualisons les indices de Sobol d'ordre 2 plus significatifs par modèle.

Fréquence

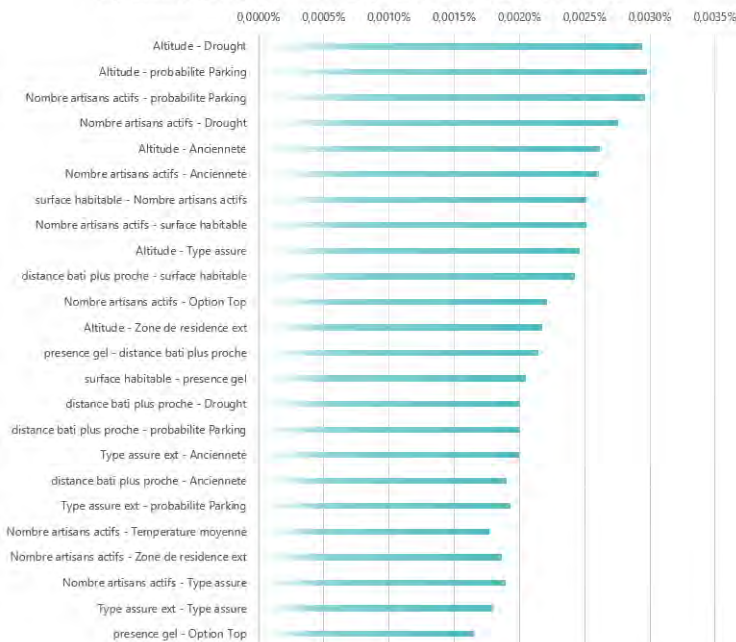
INTERACTION SOBOLE FREQUENCE RANDOM FOREST



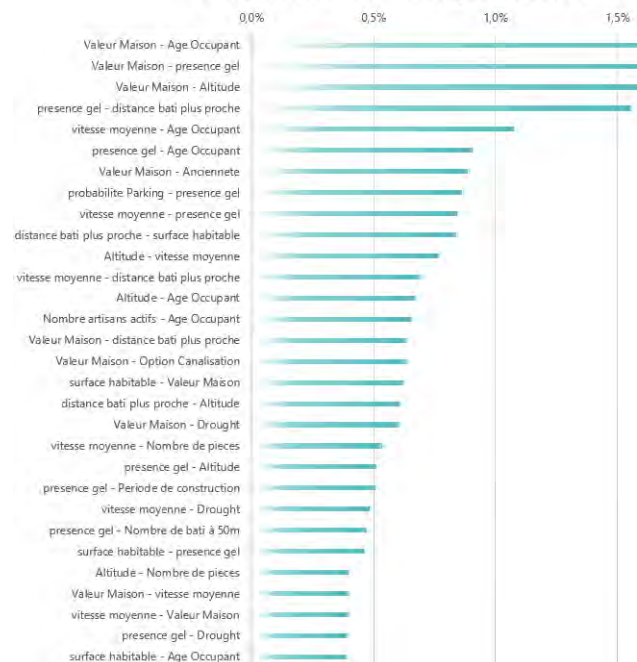
INTERACTION SOBOLE FREQUENCE XGB



INTERACTION SHAP FREQUENCE RANDOM FOREST



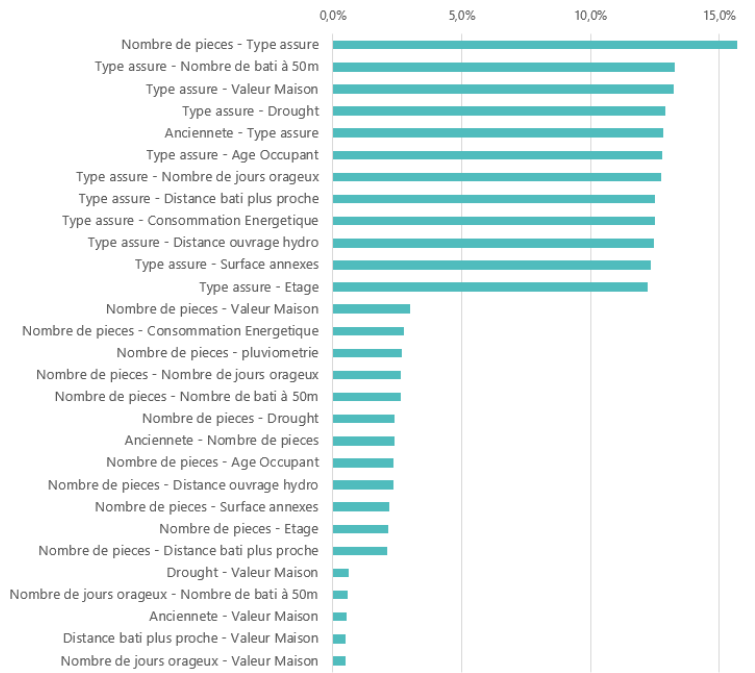
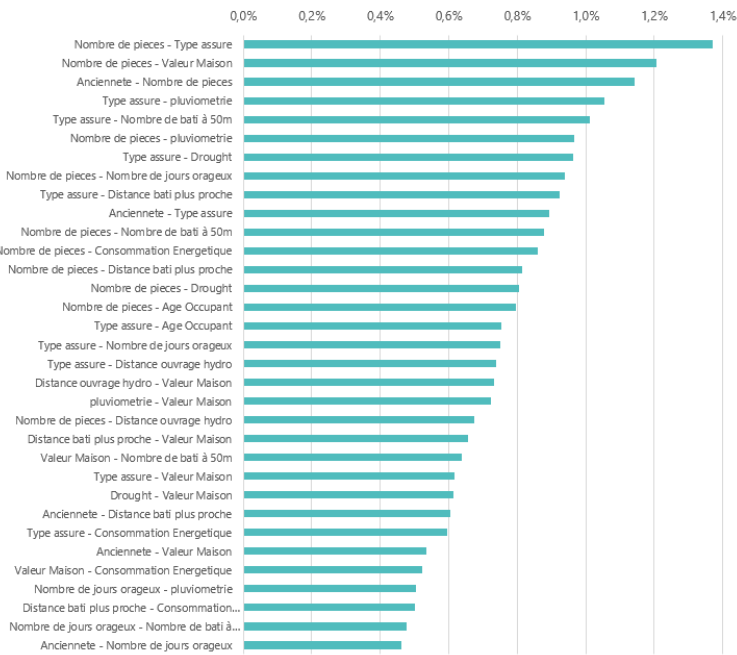
INTERACTION SHAP FREQUENCE XGB



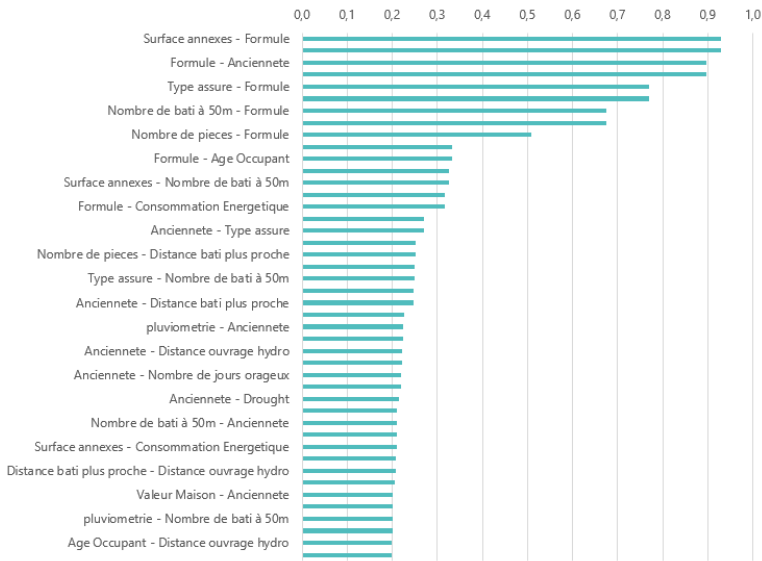
Sévérité

INTERACTION SOBOLE SEVERITE XGB

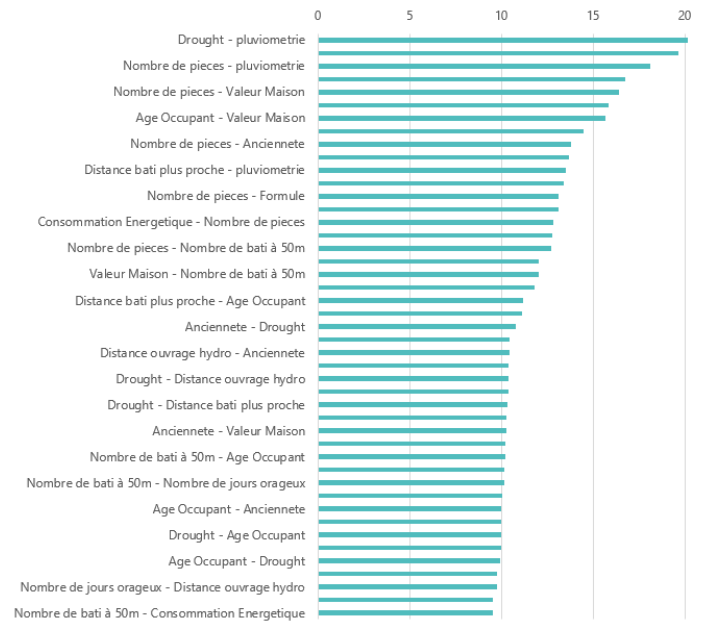
INTERACTION SOBOLE SEVERITE FREQUENCE



SHAP INTERACTION SEVERITE RANDOM FOREST



SHAP INTERACTION SEVERITE XGB



Annexe H

Résultats

Gain opérationnel : compléments

| Modele | Analyse de sensibilité | Metamodelle | Interaction ajoutée dans le GLM | Rapport de vraisemblance | Gain dans la réduction de la déviance | Probabilité critique | intensité globale |
|--------|------------------------|-------------|---|--------------------------|---------------------------------------|----------------------|-------------------|
| Sev | SHAP | XGB | Dist ouvrage hydro - Distance bâtiment plus proche | 21,591789 | 0,1% | 0,000337% | 8,08538527 |
| Sev | SHAP | XGB | Nb pieces - Pluviometrie | 20,38092 | 0,1% | 0,000635% | 18,0957828 |
| Sev | SHAP | XGB | Drought - Dist ouvrage hydro | 20,101697 | 0,1% | 0,000734% | 10,4584381 |
| Sev | Sobol | XGB | pluviometrie - Nombre de pieces | 16,17908 | 0,1% | 0,005763% | 18,0957828 |
| Sev | SHAP | XGB | Drought - Pluviometrie | 15,89276 | 0,1% | 0,006703% | 20,1386586 |
| Sev | SHAP | XGB | Nb de jours orageux - Distance bâtiment plus proche | 14,982425 | 0,1% | 0,010852% | 4,43075339 |
| Sev | SHAP | XGB | Age occupant - Nb pieces | 11,439256 | 0,0% | 0,071909% | 14,4683454 |
| Sev | SHAP | XGB | Age occupant - Drought | 10,099283 | 0,0% | 0,148327% | 12,7498156 |
| Sev | Sobol | RF | Valeur de la maison - Nombre de pieces | 8,902417 | 0,0% | 0,284793% | 0,15442444 |
| Sev | Sobol | XGB | Valeur de la maison - Nombre de pieces | 8,902417 | 0,0% | 0,284793% | 16,4298946 |
| Sev | Sobol | RF | Distance bâtiment plus proche - Anciennete | 8,581982 | 0,0% | 0,339506% | 6,71823687 |
| Sev | GLOUTON | CART | "nombre de pieces >=6" & "Valeur de la maison <2424" | 8,249437 | 0,0% | 0,407647% | |
| Sev | Sobol | RF | Anciennete - Nombre de pieces | 6,584102 | 0,0% | 1,028935% | 0,09076484 |
| Sev | SHAP | XGB | Age occupant - Nombre d'étages | 6,416546 | 0,0% | 1,130619% | 6,00177566 |
| Sev | Sobol | RF | Valeur de la maison - Drought | 6,167275 | 0,0% | 1,301349% | 8,20699313 |
| Sev | Sobol | XGB | Type d'assure - Valeur de la maison | 6,150768 | 0,0% | 1,313550% | 1,30749747 |
| Sev | SHAP | XGB | Nb de jours orageux - Nb de batiments à 50m | 5,063284 | 0,0% | 2,443789% | 10,17107 |
| Sev | Sobol | XGB | Type d'assure - Cons energ | 4,525928 | 0,0% | 3,338496% | 0,72498719 |
| Sev | Sobol | RF | Cons energ - Nombre de pieces | 3,814925 | 0,0% | 5,079787% | 0,05775011 |
| Sev | Sobol | XGB | Type d'assure - Surface annexes | 3,646472 | 0,0% | 5,618812% | 0,08260201 |
| Sev | SHAP | XGB | Nombre d'étages - Nb de batiments à 50m | 3,567806 | 0,0% | 5,891010% | |
| Sev | GLOUTON | CART | "nombre de pieces >=6" & "Valeur de la maison >=2439" | 3,238094 | 0,0% | 7,194431% | |
| Sev | Sobol | RF | Drought - Nombre de pieces | 1,79425 | 0,0% | 18,040920% | 0,1669473 |
| Sev | SHAP | XGB | Nb pieces - Nb de batiments à 50m | 1,171156 | 0,0% | 27,916390% | 12,7326175 |
| Sev | Sobol | RF | Nb de batiments à 50m - Nombre de pieces | 0,443575 | 0,0% | 50,540200% | 0,07350975 |
| Sev | Sobol | XGB | Nb de batiments à 50m - Nombre de pieces | 0,443575 | 0,0% | 50,540200% | 12,7326175 |
| Sev | SHAP | XGB | Age occupant - Cons Energ | 0,4387615 | 0,0% | 50,772080% | 16,7709803 |
| Sev | Sobol | XGB | Type d'assure - Nb de jours orageux | 0,000303983 | 0,0% | 98,608950% | 0,07446191 |
| Sev | GLOUTON | CART | nombre de pieces <6 | 0 | 0,0% | 100,000000% | |

FIGURE H.1 – Compléments au modèle de sévérité

| Modele | Analyse de sensibilité | Metamodelle | Interaction ajoutée dans le GLM | Gain dans la réduction de la déviance | Probabilité critique | intensité globale |
|-----------|------------------------|-------------|--|---------------------------------------|----------------------|-------------------|
| Frequence | GLOUTON | CART | Valeur de la maison<2182 | 0,1% | 0,000000% | |
| Frequence | SHAP | XGB | Nb pieces - Valeur Maison | 0,1% | 0,000000% | 0,35% |
| Frequence | Sobol | RF | Valeur Maison - zone de residence ext | 0,1% | 0,000000% | 2,41% |
| Frequence | Sobol | RF | Valeur Maison - surface habitable | 0,0% | 0,000000% | 0,96% |
| Frequence | Sobol | RF | Valeur Maison - Nombre de bati 50 m | 0,0% | 0,000000% | 0,96% |
| Frequence | SHAP | XGB | Anciennete - Valeur Maison | 0,0% | 0,000000% | 0,89% |
| Frequence | SHAP | XGB | Distance batiment plus proche - Valeur Maison | 0,0% | 0,000000% | 0,63% |
| Frequence | Sobol | RF | Age occupant - Valeur Maison | 0,0% | 0,000000% | 2,58% |
| Frequence | Sobol | RF | Valeur Maison - Distance batiment plus proche | 0,0% | 0,000000% | 1,75% |
| Frequence | Sobol | RF | Valeur Maison - Cons energ | 0,0% | 0,000000% | 1,88% |
| Frequence | Sobol | RF | Vitesse moyenne vent - Valeur Maison | 0,0% | 0,000000% | 2,36% |
| Frequence | Sobol | RF | Valeur Maison - option canalisation | 0,0% | 0,000000% | 1,46% |
| Frequence | SHAP | RF | Age occupant - Couverture des Canalisation | 0,0% | 0,000000% | 0,00% |
| Frequence | Sobol | XGB | Distance batiment plus proche - Age occupant | 0,0% | 0,000000% | 0,76% |
| Frequence | Sobol | RF | Altitude - Valeur Maison | 0,0% | 0,000000% | 3,37% |
| Frequence | SHAP | XGB | Altitude - Valeur Maison | 0,0% | 0,000000% | 1,75% |
| Frequence | Sobol | RF | Drought - Valeur Maison | 0,0% | 0,000000% | 2,86% |
| Frequence | SHAP | XGB | Age occupant - Drought | 0,0% | 0,000000% | 0,00% |
| Frequence | Sobol | XGB | presence gel - Valeur Maison | 0,0% | 0,000000% | 15,16% |
| Frequence | Sobol | RF | Valeur Maison - presence gel | 0,0% | 0,000000% | 2,48% |
| Frequence | SHAP | XGB | Valeur Maison - presence gel | 0,0% | 0,000000% | 0,00% |
| Frequence | SHAP | XGB | Age occupant - Anciennete | 0,0% | 0,000000% | 0,00% |
| Frequence | SHAP | RF | Age occupant - Nb pieces | 0,0% | 0,000000% | 0,00% |
| Frequence | SHAP | XGB | Age occupant - Nb pieces | 0,0% | 0,000000% | 0,00% |
| Frequence | SHAP | XGB | Vitesse moyenne vent - Type d'assuré | 0,0% | 0,000000% | 0,00% |
| Frequence | SHAP | RF | Age occupant - Nb artisans | 0,0% | 0,000000% | 0,12% |
| Frequence | SHAP | XGB | Age occupant - Temperature moyenne | 0,0% | 0,000000% | 0,00% |
| Frequence | SHAP | RF | Nb artisans - surface habitable | 0,0% | 0,000000% | 0,00% |
| Frequence | SHAP | RF | Age occupant - periode de construction | 0,0% | 0,000000% | 0,00% |
| Frequence | SHAP | XGB | Age occupant - Nb de batiments à 50m | 0,0% | 0,000000% | 0,00% |
| Frequence | SHAP | RF | Age occupant - Zone de residence (int) | 0,0% | 0,000000% | 0,00% |
| Frequence | Sobol | RF | Altitude - Temperature moyenne | 0,0% | 0,000000% | 0,56% |
| Frequence | SHAP | RF | Nb de batiments à 50m - Nb artisans | 0,0% | 0,000000% | 0,00% |
| Frequence | SHAP | RF | Zone de residence (int) - periode de construction | 0,0% | 0,000000% | 0,00% |
| Frequence | Sobol | XGB | Age occupant - Drought | 0,0% | 0,000000% | 0,58% |
| Frequence | SHAP | RF | Nb artisans - presence conduit d'air | 0,0% | 0,000000% | 0,00% |
| Frequence | Sobol | RF | Vitesse moyenne vent - presence gel | 0,0% | 0,000000% | 0,80% |
| Frequence | SHAP | XGB | Age occupant - Altitude | 0,0% | 0,000000% | 0,67% |
| Frequence | Sobol | XGB | Valeur Maison - Altitude | 0,0% | 0,000001% | 8,13% |
| Frequence | SHAP | XGB | Age occupant - Nb artisans | 0,0% | 0,000037% | 0,66% |
| Frequence | Sobol | XGB | Age occupant - surface habitable | 0,0% | 0,000073% | 0,53% |
| Frequence | SHAP | RF | Distance batiment plus proche - presence conduit d'air | 0,0% | 0,000100% | 0,00% |
| Frequence | SHAP | RF | presence conduit d'air - presence gel | 0,0% | 0,000897% | 0,00% |
| Frequence | Sobol | XGB | Age occupant - Temperature moyenne | 0,0% | 0,002481% | 0,87% |
| Frequence | Sobol | RF | Age occupant - presence gel | 0,0% | 0,002657% | 1,15% |
| Frequence | SHAP | XGB | Age occupant - Vitesse moyenne vent | 0,0% | 0,039586% | 1,07% |
| Frequence | SHAP | RF | Type d'assuré (variable externe) - surface habitable | 0,0% | 0,077200% | 0,00% |
| Frequence | SHAP | XGB | presence gel - Type d'assuré | 0,0% | 0,104022% | 0,13% |
| Frequence | Sobol | XGB | presence gel - Distance batiment plus proche | 0,0% | 0,171907% | 0,94% |
| Frequence | SHAP | XGB | Distance batiment plus proche - presence gel | 0,0% | 0,171907% | 1,56% |
| Frequence | SHAP | RF | Age occupant - Drought | 0,0% | 0,358021% | 0,00% |
| Frequence | SHAP | XGB | Drought - presence gel | 0,0% | 0,799410% | 0,40% |
| Frequence | SHAP | RF | periode de construction - Vitesse moyenne vent | 0,0% | 0,837507% | 0,00% |
| Frequence | SHAP | RF | Drought - presence conduit d'air | 0,0% | 1,067243% | 0,00% |
| Frequence | SHAP | RF | Type d'assuré (variable externe) - presence gel | 0,0% | 9,744852% | 0,00% |
| Frequence | SHAP | RF | periode de construction - Zone de residence (ext) | 0,0% | 100,000000% | 0,00% |

FIGURE H.2 – Compléments au modèle de fréquence