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

le 24 Septembre 2021

Par : Sophie NAVARRO
Titre : L'*Open Data* au service de la tarification à l'adresse

Confidentialité : Non

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

Membre présent du jury de l'Institut des Actuaires :
Anaëlle LE BERRE
Guillaume BIESSY
Signature :

Entreprise :
Sia Partners
Signature :

Membres présents du jury de l'EURIA :
Franck VERMET

Directeur de mémoire en entreprise :
Romain LAILY
Claire NICOLLE
Signature :

Invité :

Signature :

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

Signature du responsable entreprise :

Signature du candidat :

EURIA
EURo Institut d'Actuariat

6, avenue le Gorgeu
CS 93837
29238 Brest Cedex 3

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

Résumé

Dans un environnement toujours plus concurrentiel, l'expérience et le bien-être client sont au cœur des préoccupations de nombreux assureurs. Suivi personnalisé des clients, déclarations facilitées des sinistres en ligne, remboursement rapide des sinistres ou encore digitalisation des démarches, de nombreux moyens existent pour améliorer l'expérience client. Cependant, certains axes restent encore à améliorer, en particulier, le processus de souscription, un processus long, fastidieux, imposant au client de répondre à de multiples questions ce qui le conduit très souvent à abandonner le processus avant sa fin.

Par ailleurs, les données ouvertes (*Open Data*) connaissent aujourd'hui un développement important : elles sont de plus en plus nombreuses qu'elles soient mises à disposition par l'Etat ou qu'elles proviennent de la contribution de particuliers.

Dans ce contexte, il semble alors possible d'utiliser ces données nouvellement disponibles afin de simplifier le processus de souscription client en mettant en place un processus de tarification basé sur une seule variable : l'adresse. Côté client, le gain de temps serait non négligeable. Côté assureur, la simplification du processus de souscription deviendrait, de fait, un argument marketing permettant d'attirer une nouvelle clientèle. Elle permettrait également d'augmenter le taux de conversion des clients potentiels en clients réels.

L'objectif de ce mémoire est donc d'examiner s'il est possible, aujourd'hui, de concurrencer la méthode de tarification traditionnelle par une méthode de tarification à l'adresse reposant essentiellement sur de l'*Open Data*. Une réflexion plus générale sur la qualité des données et la faisabilité d'une telle pratique sera également menée.

Mots clefs: expérience client, assurance MRH, *Open Data*, tarification à l'adresse, GLM, forêt aléatoire, marché concurrentiel.

Abstract

In today's increasingly competitive environment, the customer experience and its well-being are at the heart of many insurers' concerns. Personalized customer follow-up, easy online claims declarations, rapid claims reimbursement and digitalization of procedures are all ways to improve the customer experience. However, some process still need to be improved, in particular the underwriting process, which is long, tedious and requiring the customer to answer multiple questions. Very often, it is leading the customer to abandon the process before it is completed.

In addition, Open Data is currently undergoing significant development : there are more and more of them, whether they are made available by the State or that they come from the contribution of individuals.

In this context, it seems possible to use this newly available data to simplify the customer underwriting process by implementing a pricing process based on a single variable : the address. On the customer side, the time savings would be significant. For the insurer, the simplification of the underwriting process would become a marketing argument to attract new customers. It would also increase the conversion rate of potential customers into real customers.

The objective of this thesis is therefore to study whether it is possible to compete with the traditional pricing method by an address-based pricing method (essentially based on Open Data). A more general reflection on the quality of the data and the feasibility of such a practice will also be conducted.

Keywords: customer experience, multi-risk home insurance, Open Data, address-based pricing, GLM, random forest, competitive market

Note de synthèse

Dans un marché toujours plus concurrentiel, les assureurs doivent faire face à un défi de taille : améliorer l'expérience de leurs clients dans le but de se démarquer. Une manière d'y parvenir est de simplifier le processus de souscription client, un processus long et fastidieux qui impose au client de répondre à de multiples questions ce qui le conduit très souvent à abandonner le processus avant sa fin.

Pour simplifier la démarche, une solution possible consiste en la mise en place d'un processus de tarification à l'adresse, c'est-à-dire une tarification à l'aide d'une seule variable : l'adresse. Ce mémoire propose donc une approche dans laquelle une seule question sera posée au client ce qui lui demandera moins de 30 secondes d'attention.

FIGURE 1 – Enjeu du mémoire

Ce gain de temps, porté à la connaissance des clients potentiels, serait très probablement de nature à les influencer dans le choix de leur assureur. Quant à l'assureur, il pourrait alors augmenter ses parts de marché en :

- utilisant le processus de tarification à l'adresse comme un **argument marketing** permettant alors d'**attirer une nouvelle clientèle** (en mettant en avant l'économie de temps) ;
- **augmentant le taux de conversion** des prospects, ceux-ci étant plus nombreux à aller jusqu'au bout du processus de souscription compte tenu de la simplification de la démarche.

Un processus de tarification à l'adresse semble donc comporter de nombreux avantages. C'est pourquoi, aujourd'hui, de nombreux acteurs du secteur s'intéressent à la question et tentent de mettre en place de manière opérationnelle une telle méthode de tarification. L'objectif de ce mémoire est donc d'examiner s'il est possible, aujourd'hui, de concurrencer la méthode de tarification traditionnelle par une méthode de tarification à l'adresse reposant essentiellement sur de l'*Open Data*.

« La tarification à l'adresse peut-elle concurrencer la tarification usuelle ? »

Après un examen de la base de données mise à disposition pour réaliser cette étude (base dite « assureur » dans le mémoire et contenant uniquement les variables déclaratives de la souscription), un rappel théorique des différentes méthodes utilisées a été effectué. La constitution de

la base de données dite « à l'adresse » a ensuite fait l'objet d'une présentation détaillée. Cette dernière a été construite à partir de l'adresse des assurés et a été complétée par des données en *Open Data* et, à la marge, des données internes à Sia Partners : aucune variable déclarative n'est donc présente dans cette base. Enfin, les modèles traditionnel et à l'adresse mis en place ont été explicités avant de mesurer l'impact de la tarification à l'adresse par rapport à la tarification traditionnelle dans un marché concurrentiel.

Présentation de la base de données « assureur »

La base de données de référence utilisée dans le cadre de cette étude correspond à un portefeuille multirisque habitation d'un assureur français sur la période 2010-2014.

En conservant l'objectif du mémoire à l'esprit, à savoir mettre en place un processus de tarification à l'adresse, il a été décidé de se focaliser sur un périmètre restreint du portefeuille : les maisons. Ce choix a été fait pour plusieurs raisons, la principale étant qu'il est beaucoup plus aisé d'obtenir des informations en *Open Data* sur les maisons plutôt que sur les appartements. Enfin, seules les lignes pour lesquelles des données en *Open Data* ont pu être récupérées ont été conservées ce qui représente un total de 255 000 lignes.

Pour la suite du mémoire, il a également été considéré que le produit d'assurance MRH tarifié ne comportait qu'une garantie dégâts des eaux (DDE) et une garantie vol, ces dernières étant intuitivement les plus susceptibles d'être expliquées par des variables externes issues de l'*Open Data*.

Enfin, une analyse des données a été entreprise afin de mieux cerner le portefeuille et de développer une première intuition quant aux variables potentiellement tarifaires tant pour la garantie DDE que pour la garantie vol. La corrélation entre les différentes variables a également fait l'objet d'une étude détaillée.

Constitution de la base de données « à l'adresse »

Afin de constituer la base de données « à l'adresse », les adresses de la base « assureur » ont fait l'objet d'une extraction. Puis, à partir de ces seules adresses, différentes informations ont été récupérées en *Open Data*.

Ce mémoire se focalisant sur la tarification à l'adresse, une attention toute particulière a été consacrée à la récupération de données à la maille adresse. Cependant, afin de conserver une certaine exhaustivité dans cette étude, des données à la maille commune/département et des données météorologiques ont également été utilisées. Les variables recueillies sont :

- des données **géographiques préliminaires** ;
- des données **altimétriques** ;
- des données de type « **point d'intérêt** » ;
- des données de **distance** ;
- des données issues de la **reconnaissance d'images** ;
- des données concernant les **risques naturels et technologiques** ;
- des données provenant de la base **« Demandes de valeurs foncières »** ;
- des données à la **maille commune/département** ;
- des données **météorologiques**.

Une réflexion sur la qualité des données a par ailleurs été menée en parallèle, cette notion impactant directement la qualité des modèles proposés par la suite.

De la même manière que pour la base « assureur », une analyse des données a ensuite été effectuée. Cette dernière a validé la pertinence des variables nouvellement recueillies dans le cadre de la modélisation des garanties dégâts des eaux et vol, ce qui a permis ensuite de procéder à la tarification à l'adresse de ces garanties.

Mise en place des tarifications

Une fois la base de données constituée, différents modèles ont été mis en place afin de tarifier les garanties DDE et vol d'un contrat d'assurance MRH. Pour chaque garantie, trois modèles ont ainsi été construits. Le premier modèle a repris la méthode de tarification traditionnelle : un GLM fréquence-coût a été calibré sur la base « assureur » ne contenant aucune variable externe. C'est ce modèle qui a, par la suite, constitué le modèle de référence auquel les deux autres modèles, à l'adresse cette fois-ci, ont été comparés. Le deuxième modèle, qui est donc à l'adresse, repose, lui aussi, sur une tarification GLM fréquence-coût tandis que le troisième, une forêt aléatoire à l'adresse, modélise séparément la fréquence et le coût moyen. Ces deux derniers modèles ont été calibrés uniquement à l'aide des données à l'adresse.

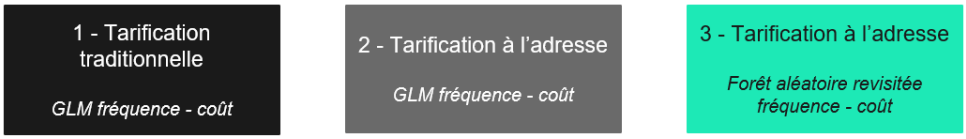


FIGURE 2 – Les différentes modélisations effectuées

Chaque modèle a ensuite fait l'objet d'une optimisation et d'une analyse de ses coefficients. Pour les modèles à l'adresse, une attention toute particulière a également été portée sur l'interprétation des coefficients. Avant de présenter les résultats des comparaisons des modèles entre eux, quelques remarques d'ordre général ont pu être formulées sur les modèles précédemment construits.

Ainsi, concernant les modèles traditionnels calibrés, l'étude des résultats a pu soulever des questions quant à la modélisation de la fréquence de la garantie DDE : une forte différence de prédiction a en effet été observée sur la base de test. Ce phénomène peut s'expliquer par la faible volumétrie des données ou encore par la potentielle absence d'une variable très tarifaire dans la modélisation de la garantie DDE. Il convient d'avoir conscience de ce phénomène même s'il ne peut lui être trouvé une solution dans le cadre de ce mémoire. Quant au modèle de coût moyen DDE, il est apparu plutôt correct. Concernant la garantie vol, aucune remarque particulière n'a été formulée, les modélisations de la fréquence et du coût moyen paraissant raisonnables sur la base de test.

Pour les modèles à l'adresse, basés essentiellement sur de l'*Open Data*, deux effets majeurs se sont distingués. Ainsi, certaines variables tendaient à approcher voire à répliquer des variables déjà présentes dans la tarification traditionnelle apportant ainsi des informations cruciales à la modélisation des différents phénomènes et permettant au tarif à l'adresse d'approcher la tarification usuelle. C'était le cas, par exemple, de la surface plane de la maison (issue de la reconnaissance d'images) qui tente de remplacer la variable déclarative correspondant au nombre de pièces. C'était également le cas pour la part des maisons à la commune construites entre 1991 et 2005 qui s'assimile à l'année de construction, une variable usuellement présente dans la tarification traditionnelle. D'autres variables (la distance à une autre maison ou encore la vitesse annuelle maximale du vent) apportaient, quant à elles, des informations complémentaires qu'une tarification traditionnelle n'aurait pu capter.

Par ailleurs, il convient de noter que la pertinence de la tarification à l'adresse a différé en fonction de la garantie considérée. Ainsi, pour la garantie DDE, peu de variables en *Open Data* ont eu un réel impact sur les modélisations. En revanche, pour la garantie vol, l'*Open Data* a permis d'ajouter de nombreuses informations qui sont venues compléter l'explication de la fréquence des vols ou de leur coût moyen. Ce constat a pu amener à penser que la tarification à l'adresse pourrait peut-être, dans le cas de la garantie vol, surpasser la tarification traditionnelle. Pour la garantie DDE, cela semble plus difficile.

		Fréquence		Coût moyen		Total	
		RMSE app	*RMSE test*	*RMSE app*	*RMSE test*	*RMSE app*	*RMSE test*
DDE	*Tarification traditionnelle*	0,11792	0,11017	2 876	2 500	390,18	339,97
	GLM à l'adresse	0,11801	0,11023	2 872	2 553	390,36	340,28
	Forêt aléatoire à l'adresse	0,11759	0,11020	2 642	2 572	389,42	340,26
Vol	*Tarification traditionnelle*	0,07699	0,07651	4 185	4 987	435,14	454,39
	GLM à l'adresse	0,07698	0,07648	4 184	5 195	435,19	454,35
	Forêt aléatoire à l'adresse	0,07690	0,07649	3 636	5 098	434,91	454,13

FIGURE 3 – RMSE sur les bases réduites pour tous les modèles

Concernant la comparaison des RMSE, aucun modèle de tarification n'a semblé clairement se démarquer à la lecture des résultats. Cependant, les valeurs très similaires des RMSE des différents modèles se sont révélées encourageantes et ont démontré que les modèles à l'adresse, basés uniquement sur des données en *Open Data*, pouvaient très fortement approcher les modèles traditionnels. Toutefois, pour répondre à la question centrale du mémoire, à savoir « La tarification à l'adresse peut-elle concurrencer la tarification traditionnelle ? », il n'est pas possible de se baser uniquement sur les RMSE des modèles, ces dernières n'intégrant pas la notion même de concurrence. Une étude plus approfondie a donc du être menée.

Marché concurrentiel

L'objectif du marché concurrentiel est de simuler un environnement se rapprochant au plus près de la réalité afin de mettre en concurrence plusieurs assureurs proposant des tarifs différents pour un même produit.

Au sein de ce marché, les clients choisissent leur assureur en fonction de la prime proposée. A ce stade, l'hypothèse la plus couramment émise, quant au comportement attendu du client, consiste à dire qu'il choisira l'assureur le moins cher. Cependant, dans le cadre de cette étude, cette hypothèse a dû être nuancée. En effet, ainsi qu'évoqué précédemment, l'objectif des méthodes de tarification à l'adresse est de permettre au client de gagner du temps en n'ayant à renseigner qu'un seul élément : son adresse. Nul doute donc, qu'outre le niveau du tarif, le paramètre "gain de temps" sera également pris en compte au moment du choix de l'assureur. Pour tenter de modéliser ce phénomène, une certaine élasticité au prix a donc été calibrée.

Cette élasticité a permis de mesurer la différence de prix qu'un client est prêt à accepter en contrepartie du gain de temps offert par la tarification à l'adresse. Elle est exprimée en pourcentage de la prime et dépend de deux paramètres : l'âge et le capital assuré.

➥ **La tarification traditionnelle versus le GLM à l'adresse**

		Groupe d'apprentissage			Groupe de test		
		Seul	**En concurrence**		**Seul**	**En concurrence**	
		Résultat	Résultat	Part de marché	Résultat	Résultat	Part de marché
DDE	*Tarification traditionnelle*	622K	-12K	52%	412K	23K	52%
	GLM à l'adresse	533K	-354K	48%	356K	-48K	48%
VOL	*Tarification traditionnelle*	527K	-50K	52%	239K	-53K	52%
	GLM à l'adresse	671K	-243K	48%	300K	-63K	48%

FIGURE 4 – Résultats du marché concurrentiel (GLM traditionnel versus GLM à l'adresse)

Les premiers résultats obtenus ont montré une déficience du modèle à l'adresse par rapport à la tarification traditionnelle. Par ailleurs, l'observation de la répartition des différentes parts de marché par variable a montré des phénomènes d'anti-sélection impliquant que des

éléments-clés avaient certainement été omis dans la modélisation. Les variables formule, sinistralité passée et capitaux assurés étaient notamment concernées. Quelques exemples sur la base de test sont présentés ci-après :

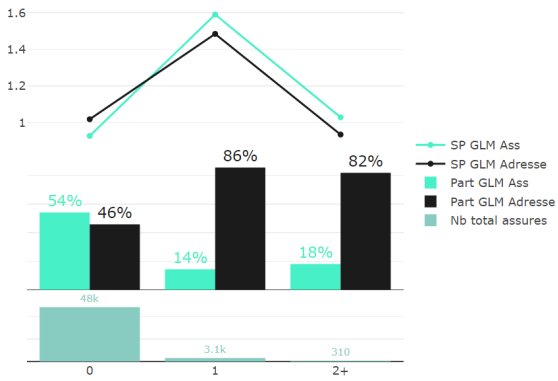


FIGURE 5 – Parts de marché de la sinistralité passée dégâts des eaux

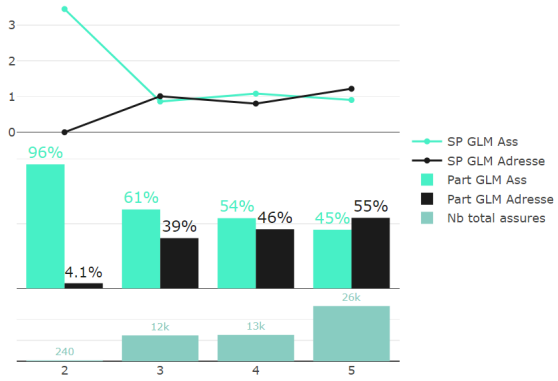


FIGURE 6 – Parts de marché de la formule pour la garantie dégâts des eaux

FIGURE 7 – Parts de marché du capital bijou pour la garantie vol

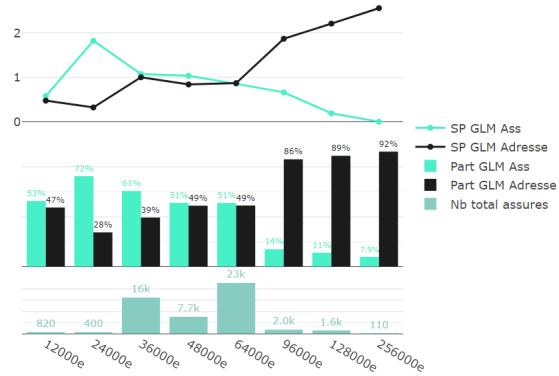


FIGURE 8 – Parts de marché du capital mobilier pour la garantie dégâts des eaux

Or, il semble logique qu'un assureur dispose de certaines de ces informations qui relèvent d'ailleurs plus d'un choix du client que du risque en lui-même (la formule et les capitaux assurés). L'assureur ayant également le droit de refuser certains assurés dont la sinistralité passée est trop importante, il doit nécessairement avoir connaissance de cette variable. Un nouveau modèle à l'adresse a donc été considéré : le GLM à l'adresse de base auquel ont été ajoutées les variables formule, sinistralité passé et capitaux assurés (tant mobilier que bijou et dépendance). Les résultats ont ensuite été mis à jour :

		Groupe d'apprentissage			Groupe de test		
		Seul	En concurrence		Seul	En concurrence	
		Résultat	Résultat	Part de marché	Résultat	Résultat	Part de marché
DDE	*Tarification traditionnelle*	622K	-109K	46%	412K	39K	46%
	GLM à l'adresse	547K	-161K	54%	376K	-18K	54%
VOL	*Tarification traditionnelle*	527K	-186K	47%	239K	-75K	47%
	GLM à l'adresse	640K	11K	53%	290K	10K	53%

FIGURE 9 – Résultats du marché concurrentiel avec ajout de la formule, de la sinistralité passée et des capitaux assurés dans le GLM à l'adresse (GLM traditionnel versus GLM à l'adresse)

Ces derniers résultats, obtenus sur la base de test, se sont alors révélés plutôt concluants pour la garantie vol. Cela est certainement lié aux ajouts importants d'informations provenant de l'utilisation de l'*Open Data* pour cette garantie (distance à une autre maison, à la station de police la plus proche, etc). En revanche, pour la garantie DDE, le GLM à l'adresse n'est pas parvenu à concurrencer de manière viable la tarification traditionnelle et ce, même si ce dernier enregistre plus de parts de marché que la tarification traditionnelle. Cela s'explique notamment par la nature des données externes qui ne donnent pas d'informations sur le bien en lui-même (outre la surface). Or, une des causes principales de dégâts des eaux, à savoir les infiltrations, est fortement liée à la vétusté de la maison. La faible performance du modèle à l'adresse pour le DDE peut ainsi s'expliquer par l'absence de variables de ce type.

Une analyse des différentes parts de marché a de nouveau été menée : il est apparu, cette fois-ci, que la variable relative au nombre de pièces était assez déséquilibrée, entraînant un phénomène d'anti-sélection et ce, que ce soit pour la garantie DDE ou pour la garantie vol.

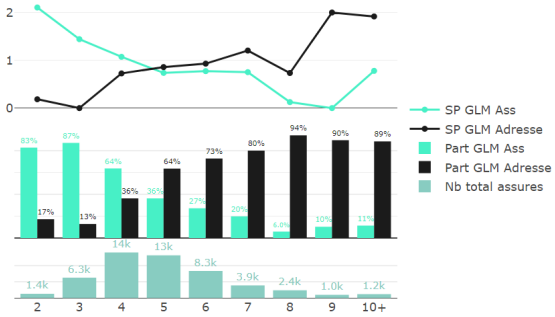


FIGURE 10 – Parts de marché du nombre de pièces pour la garantie dégâts des eaux

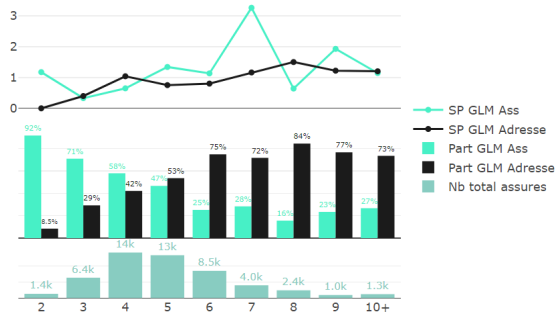


FIGURE 11 – Parts de marché du nombre de pièces pour la garantie vol

Ainsi, après ajout de la variable nombre de pièces au modèle à l'adresse, le GLM comportait maintenant, en plus des données externes, les informations suivantes : formule, sinistralité passée, capitaux et nombre de pièces. Les résultats ont alors été de nouveau mis à jour :

		Groupe d'apprentissage			Groupe de test		
		Seul	En concurrence		Seul	En concurrence	
		Résultat	Résultat	Part de marché	Résultat	Résultat	Part de marché
DDE	*GLM Ass*	622K	-229K	34%	412K	103K	34%
	GLM Adresse	557K	184K	66%	386K	18K	66%
VOL	*GLM Ass*	527K	-289K	44%	239K	-85K	44%
	GLM Adresse	653K	215K	56%	294K	64K	56%

FIGURE 12 – Résultats du marché concurrentiel avec ajout de la formule, de la sinistralité passée, des capitaux assurés et du nombre de pièces dans le GLM à l'adresse (GLM traditionnel versus GLM à l'adresse)

Ils s'en sont trouvés améliorés que ce soit au niveau des parts de marché ou des résultats. Malheureusement, pour la garantie DDE, ce dernier ajout n'a toujours pas été suffisant pour que la tarification à l'adresse surpasse la tarification traditionnelle.

➥ **La tarification traditionnelle versus la forêt aléatoire à l'adresse**

		Groupe d'apprentissage			Groupe de test		
		Seul	**En concurrence**		**Seul**	**En concurrence**	
		Résultat	Résultat	Part de marché	Résultat	Résultat	Part de marché
DDE	*GLM Ass*	622K	-391K	58%	412K	103K	58%
	RF Adresse	468K	42K	42%	328K	-125K	42%
VOL	*GLM Ass*	527K	-297K	58%	239K	-34K	58%
	RF Adresse	352K	-52K	42%	155K	-111K	42%

FIGURE 13 – Résultats du marché concurrentiel (GLM traditionnel versus forêt aléatoire à l'adresse)

L'ensemble des résultats obtenus a fait ressortir un constat : un phénomène de surapprentissage a été observé pour les forêts aléatoires et ce, malgré les optimisations effectuées. En effet, quel que soit les scénarios (présentés en section 6.3), il a été possible de noter que les résultats en concurrence du modèle à l'adresse sur la base d'apprentissage étaient bien meilleurs que ceux constatés sur la base de test. Une explication possible à ce phénomène pourrait être liée à la faible volumétrie de la base de données à disposition qui ne permet pas d'apprendre correctement et sans surapprentissage le modèle.

Conclusion

Les résultats de ces travaux, bien qu'encourageants, ont ainsi montré quelques limites.

En effet, il est apparu, qu'à ce jour, il semblait encore difficile de mettre en place une tarification basée uniquement sur l'adresse de l'assuré, certaines informations relatives aux choix de l'assuré ou encore à sa sinistralité passée devant absolument être fournies pour que la tarification à l'adresse soit viable. Par ailleurs, le nombre de pièces, autre élément essentiel à la tarification, a, dans le cadre de ce mémoire, été difficilement capté par les données à l'adresse (reconnaissance d'images). Ce faisant, un phénomène d'anti-sélection est apparu au sein du marché concurrentiel. Ce phénomène lié à l'absence d'information sur le nombre de pièces pourrait cependant être supprimé par l'intégration du *Street View* dans la reconnaissance d'images. Enfin, d'autres ajouts de variables aux modèles à l'adresse pourraient également participer à l'amélioration de cette tarification à l'adresse notamment pour la garantie DDE qui reste tout de même peu satisfaisante : il pourrait, par exemple, être possible d'intégrer à l'étude la base des permis de construire, la base des diagnostics de performance énergétique (DPE) ou encore la variable année de construction, trois éléments pouvant donner des informations quant à la vétusté des maisons.

Toutefois, bien que comportant des limites, ces travaux ont également démontré que la tarification à l'adresse pourrait être une solution d'avenir. En effet, il est, par exemple, apparu que les données à l'adresse permettaient tant d'approcher les variables tarifaires usuelles que de compléter la modélisation du risque par l'apport de nouvelles informations. Cela s'est révélé particulièrement vrai pour la garantie vol avec les données issues de la reconnaissance d'images (isolement de la maison, distance à la station de police la plus proche, etc).

Pour terminer, compte tenu de la taille relativement faible de la base de données utilisée dans le cadre de ce mémoire, il est important de souligner qu'il ne peut être tiré de conclusion générale quant à la performance de la tarification à l'adresse par rapport à la tarification traditionnelle. Il paraît en effet opportun de poursuivre et compléter ces travaux en utilisant une base de données plus conséquente pour confirmer ces premières tendances.

Executive summary

In an increasingly competitive market, insurers are facing a major challenge in order to stand out : improving their customers' experience. One way to do this is to simplify the customer underwriting process, a long and tedious process that requires the customer to answer multiple questions and often leads him to abandoning the process before it is complete.

To simplify the process, a possible solution is the implementation of an address-based pricing process, i.e., pricing using a single variable : the address. This paper proposes an approach in which the customer is asked a single question that requires less than 30 seconds of attention.

FIGURE 14 – Issue of the thesis

This time saving, made known to potential clients, would most likely influence their choice of insurer. As for the insurer, it could then increase its market share by :

- using the address-based pricing process as a marketing argument to attract new customers (by highlighting the time savings) ;
- increasing the conversion rate of potential customers to actual customers, as more potential customers will complete the application process due to the simplification of the process.

An address-based pricing process seems therefore to have many advantages. This is why, today, many players in the sector are interested in the issue and are trying to implement such a pricing method in an operational manner. The objective of this paper is therefore to examine whether it is possible, today, to compete with the traditional pricing method by an address-based pricing method essentially based on *Open Data*.

"Can address-based pricing compete with traditional pricing ?"

After an examination of the database made available to carry out this study (database known as "insurer database" and containing only the declarative variables of the subscription), a theoretical reminder of the various methods used was carried out. The constitution of the database called "at the address" was then the subject of a detailed presentation. This database was built from

the addresses of the insured and was completed by Open Data and, on the margin, internal data from Sia Partners : no declarative variable is therefore present in this database. Finally, the traditional and address-based models were implemented before measuring the impact of address-based pricing compared to traditional pricing in a competitive market.

Presentation of the "insurer database"

The reference database used in this study corresponds to a French insurer's multi-risk home portfolio over the period 2010-2014.

Keeping in mind the objective of the thesis, namely to implement an address-based pricing process, it was decided to focus on a limited perimeter of the portfolio : houses. This choice was made for several reasons, the main one being that it is much easier to obtain information in Open Data on houses rather than on apartments. Finally, only the lines for which Open Data could be recovered were kept, which represents a total of 255,000 lines.

For the remainder of the paper, it was also considered that the priced multi-risk home insurance product only includes a water damage guarantee and a theft coverage, both being intuitively the most likely to be explained by external variables from Open Data.

Finally, an analysis of the data was undertaken in order to better define the portfolio and to develop an initial intuition as to the potential pricing variables for both the water damage guarantee and theft coverage. The correlation between the different variables was also studied.

Creation of the database "at the address"

In order to constitute the database "at the address", the addresses of the "insurer database" were extracted. Then, from these addresses alone, different informations were retrieved in Open Data.

Since this thesis focuses on address-based pricing, special attention was paid to the recovery of data at the address level. However, in order to keep a certain exhaustiveness, data at the commune/department level and meteorological data were also used.

The collected variables are :

- **preliminary geographical** data ;
- **altimetric** data ;
- "**point of interest**" data ;
- **distance** data ;
- data from **image recognition** ;
- **natural and technological risk** data ;
- data from the database "**Requests for property values**" ;
- data at the commune/department level ;
- **meteorological** data.

A reflection on the quality of the data was also carried out, as this notion has a direct impact on the quality of the models proposed later.

In the same way as for the insurer's database, an analysis of the data was then carried out. This analysis validated the relevance of the newly collected variables within the framework of the modeling of the theft and water damage coverage.

Setting up of tariffs

Once the database was built, different models were set up in order to price the water damage guarantee and the theft coverage of a multi-risk home insurance product. For each coverage, three models were built. The first model used the traditional pricing method : a frequency-cost GLM was calibrated on the basis of the insurer's data, without any external variables. This model was then used as the reference model against which the other two models, at the address, were compared. The second model, which is therefore at the address, is also based on GLM frequency-cost pricing, while the third, a random forest at the address, models frequency and average cost separately. The latter two models were calibrated using only the address data.

FIGURE 15 – Implemented models

Each model has been optimized and its coefficients analyzed. For the address-based models, special attention was also paid to the interpretation of the coefficients. Before presenting the results of the comparisons between the models, some general remarks could be made about the models previously built.

Concerning the traditional models calibrated, the study of the results raised questions about the modeling of the frequency of the water damage guarantee : a strong difference in prediction was in fact observed on the test basis. This phenomenon can be explained by the low volume of data or by the potential absence of a highly priced variable in the modeling of the water damage guarantee. It is important to be aware of this phenomenon, even if a solution cannot be found in the context of this paper. As for the average cost model of the water damage guarantee, it appeared to be rather correct. Concerning the theft guarantee, no particular remark was made, the frequency and average cost models appearing reasonable on the basis of the test.

For the address-based models, based essentially on Open Data, two major effects stood out. Some variables tended to approximate or even replicate variables already present in traditional pricing, thus providing crucial information for modeling the various phenomena and allowing the address-based tariff to approximate the usual pricing. This was the case, for example, for the flat surface of the house (derived from image recognition) which attempts to replace the declarative variable corresponding to the number of pieces. This was also the case for the share of houses in the municipality built between 1991 and 2005, which is assimilated to the year of construction, a variable usually present in traditional pricing. Other variables (the distance to another house or the maximum annual wind speed) provided additional information that traditional pricing could not capture.

In addition, it should be noted that the relevance of address-based pricing differed depending on the coverage considered. Thus, for the water damage guarantee, few variables in Open Data had a real impact on the models. On the other hand, for the theft guarantee, the Open Data enabled the addition of a great deal of information that completed the explanation of the frequency of thefts or their average cost. This observation suggested that address-based pricing might be able to outperform traditional pricing in the case of theft coverage. For the water damage coverage, this seems more difficult.

		Frequency		Average cost		Total	
		RMSE learning	*RMSE test*	*RMSE learning*	*RMSE test*	*RMSE learning*	*RMSE test*
Water damage	*Traditional pricing*	0.11792	0.11017	2 876	2 500	390.18	339.97
	Address-based GLM	0.11801	0.11023	2 872	2 553	390.36	340.28
	Address-based random forest	0.11759	0.11020	2 642	2 572	389.42	340.26
Theft	*Traditional pricing*	0.07699	0.07651	4 185	4 987	435.14	454.39
	Address-based GLM	0.07698	0.07648	4 184	5 195	435.19	454.35
	Address-based random forest	0.07690	0.07649	3 636	5 098	434.91	454.13

FIGURE 16 – RMSE for all models

Regarding the RMSE comparison, no pricing model seemed to clearly stand out. However, the very similar RMSE values of the different models were encouraging and demonstrated that address-based models, based solely on Open Data, could very strongly approach the traditional models. However, to answer the central question of the thesis, namely, "can address-based pricing compete with traditional pricing ?" it is not possible to rely solely on the RMSE of the models, as the latter do not integrate the very notion of competition. A more in-depth study had to be conducted.

Competitive market

The objective of the competitive market is to simulate an environment that is as close as possible to reality in order to put in competition several insurers offering different premiums for the same product.

In this market, customers choose their insurer based on the premium offered. At this stage, the most common assumption about the customer's expected behavior is that he will choose the cheapest insurer. However, in the context of this study, this hypothesis had to be qualified. Indeed, as previously mentioned, the objective of address-based pricing methods is to allow the client to save time by having to fill only one element : his address. Thus, there is no doubt that, in addition to the level of the premium, the "time saving" parameter will also be taken into account when choosing the insurer. To try to model this phenomenon, a price elasticity has been calibrated.

This elasticity has been used to measure the price difference that a client is ready to accept in exchange for the time saving offered by address-based pricing. It is expressed as a percentage of the premium and depends on two parameters : age and sum insured.

➥ **Traditional pricing versus address-based GLM**

		Learning			Test		
		Alone	**In competition**		**Alone**	**In competition**	
		Result	Result	Market share	Result	Result	Market share
Water damage	*Traditional pricing*	622K	-12K	52%	412K	23K	52%
	Address-based GLM	533K	-354K	48%	356K	-48K	48%
Theft	*Traditional pricing*	527K	-50K	52%	239K	-53K	52%
	Address-based GLM	671K	-243K	48%	300K	-63K	48%

FIGURE 17 – Competitive market results (traditional pricing versus address-based GLM)

The first results obtained showed a deficiency of the address-based model compared to traditional pricing. Moreover, the observation of the distribution of the market shares showed anti-selection phenomena implying that some key elements had certainly been omitted in the modeling. The variables formula, past claims experience and insured capital were particularly concerned. Some examples (with the test basis) are presented below :

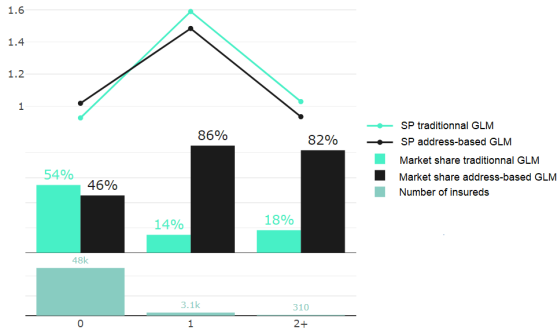


FIGURE 18 – Market share of past water damage claims

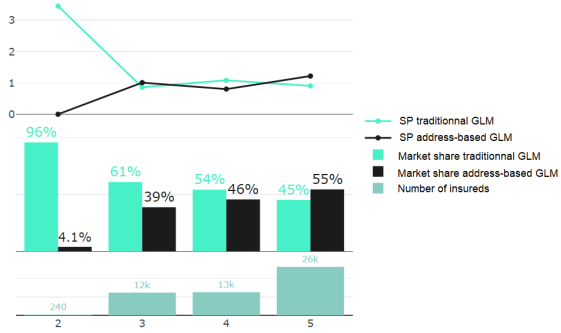


FIGURE 19 – Formula market share for water damage coverage

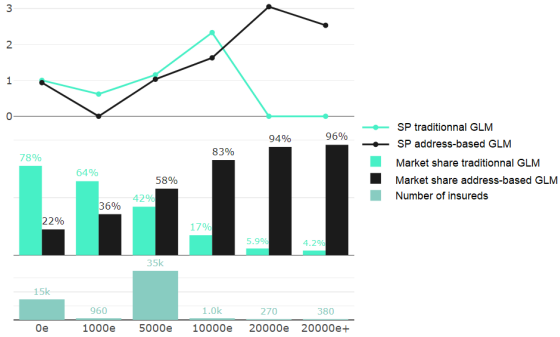


FIGURE 20 – Market share of jewelry capital for theft coverage

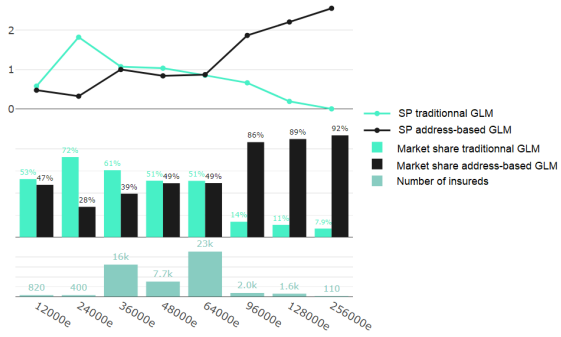


FIGURE 21 – Market share of personal property capital for water damage coverage

Yet, it seems logical that an insurer should have some of this information, which is more a matter of customer choice than of the risk itself (the formula and the insured amounts). Since the insurer also has the right to refuse certain policyholders whose past claims experience is too high, he must necessarily be aware of this variable. A new model at the address was therefore considered : the basic GLM at the address to which were added the variables formula, past claims experience and insured amounts (both personal property, jewelry and outbuilding). The results were then updated :

		Learning			Test		
		Alone	In competition		Alone	In competition	
		Result	Result	Market share	Result	Result	Market share
Water damage	*Traditional pricing*	622K	-109K	46%	412K	39K	46%
	Address-based GLM	547K	-161K	54%	376K	-18K	54%
Theft	*Traditional pricing*	527K	-186K	47%	239K	-75K	47%
	Address-based GLM	640K	11K	53%	290K	10K	53%

FIGURE 22 – Competitive market results with the addition of the formula, the past loss experience and the insured amounts in the address-based GLM (traditional pricing versus address-based GLM)

These last results, obtained on the basis of the test, proved to be rather conclusive for the theft guarantee. This is certainly related to the important additions of information coming from the use of Open Data for this coverage (distance to another house, distance to the nearest police station, etc). In contrast, for the water damage guarantee, address-based GLM has not been able to viably compete with traditional pricing, even though it has a larger market share than traditional pricing. One reason for this is the nature of the

external data, which does not provide information about the property itself (other than area). However, one of the main causes of water damage, namely seepage, is strongly linked to the age of the house, hence the poor performance of the address-based model for water damage.

An analysis of the different market shares was again carried out : this time it appeared that the variable number of pieces was rather unbalanced, leading to an anti-selection phenomenon, for both the water damage guarantee or the theft coverage.

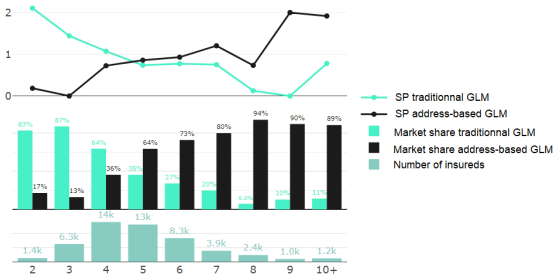


FIGURE 23 – Market share of the number of pieces for water damage coverage

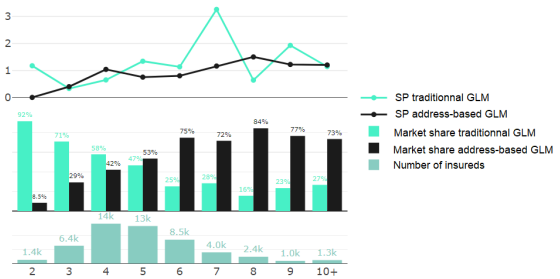


FIGURE 24 – Market share of the number of pieces for theft coverage

After adding the number of pieces to the address-based model, the GLM now included, in addition to the external data, the following information : formula, past claims, insured amounts and number of pieces. The results were then updated again :

		Learning			Test		
		Alone	In competition		Alone	In competition	
		Result	Result	Market share	Result	Result	Market share
Water damage	*Traditional pricing*	622K	-229K	34%	412K	103K	34%
	Address-based GLM	557K	184K	66%	386K	18K	66%
Theft	*Traditional pricing*	527K	-289K	44%	239K	-85K	44%
	Address-based GLM	653K	215K	56%	294K	64K	56%

FIGURE 25 – Competitive market results with the addition of the formula, the past loss experience, the insured amounts and the number of pieces in the address-based GLM (traditional pricing versus address-based GLM)

This has improved both market share and results. Unfortunately, for the water damage guarantee, this last addition was still not enough to make address-based pricing outperform traditional pricing.

Traditional pricing versus address-based random forest

		Learning			Test		
		Alone	In competition		Alone	In competition	
		Result	Result	Market share	Result	Result	Market share
Water damage	*Traditional pricing*	622K	-391K	58%	412K	103K	58%
	Address-based RF	468K	42K	42%	328K	-125K	42%
Theft	*Traditional pricing*	527K	-297K	58%	239K	-34K	58%
	Address-based RF	352K	-52K	42%	155K	-111K	42%

FIGURE 26 – Competitive market results (traditional pricing versus address-based random forest)

All the results obtained highlighted one observation : an overlearning phenomenon was observed for the random forests and this, despite the optimizations performed. Indeed, regardless of the scenarios (presented in section 6.3), the results in competition of the address-based model on the learning base were much better than those observed on the test base. A possible explanation for this phenomenon could be linked to the low volume of the database which does not allow the model to learn correctly and without overlearning.

Conclusion

The results of this work, although encouraging, have shown some limitations.

Indeed, it appeared that, to date, it still seemed difficult to set up a pricing system based solely on the insured's address, as certain information relating to the insured's choices or to his or her past claims history had to be provided in order for the address-based pricing system to be viable. Moreover, the number of pieces, another essential element for pricing, was, in the context of this paper, difficult to capture by address-based data (image recognition). In doing so, an anti-selection phenomenon appeared in the competitive market. This phenomenon related to the lack of information on the number of pieces could, however, be eliminated by integrating *Street View* into the image recognition. Finally, other additions of variables to the address-based models could also contribute to improve the address-based pricing, particularly for the DDE guarantee, which is still unsatisfactory : it could, for example, be possible to integrate into the study the building permit database, the energy performance diagnosis database or the year of construction variable, three elements that could provide information about the age of the houses.

However, although it has its limitations, this work has also shown that address-based pricing could be a solution for the future. For example, it appeared that address-based data could be used to approximate the usual rate variables and to complete the risk modeling by adding new information. This was particularly true for theft coverage with data from image recognition (isolation of the house, distance to the nearest police station, etc.).

Finally, given the relatively small size of the database used in this paper, it is important to emphasize that no general conclusion can be drawn about the performance of address-based pricing compared to traditional pricing. It seems appropriate to continue and complete this work using a larger database to confirm these initial trends.

Remerciements

Je tiens tout d'abord à remercier l'entreprise Sia Partners et notamment Michaël DONIO et Ronan DAVIT, les deux directeurs de l'unité de compétences (UC) Actuariat qui ont accepté de m'accueillir en stage au sein de leur service. Ce stage m'a permis, outre la réalisation de mon mémoire, de développer de réelles compétences dans le domaine du conseil et, pour cela, je leur en suis très reconnaissante.

Je souhaite ensuite remercier chaleureusement mes deux encadrants, Romain LAILY, manager au sein de l'UC Actuariat et Claire NICOLLE, consultante senior au sein de l'UC Actuariat. Tous deux ont su m'encadrer, me conseiller et me guider tout au long de mon mémoire. Pour tout cela, je les en remercie.

Un grand merci également à l'ensemble des membres de l'UC Actuariat qui ont pris le temps de répondre aux questions que je me posais lorsque le besoin s'en faisait ressentir.

Je n'oublie pas bien sûr les actuaires externes à Sia Partners qui ont contribué, eux aussi, au bon déroulement de mon stage et de mon mémoire de par leurs conseils avisés. Je pense notamment à Jordan MARIE-ROSE, mon tuteur académique, et Sarah DIDO qui ont, tous deux, su me rassurer lorsque le besoin s'en faisait ressentir. J'aimerais aussi les remercier pour leurs relectures qui m'ont permis, à chaque fois, d'améliorer mon mémoire. Toute ma gratitude également à Léonie LE BASTARD et Bryan GAUTIER pour leur soutien depuis le début.

Enfin, j'aimerais terminer en remerciant ma famille et mes amis qui ont su m'écouter et me guider tout au long de la réalisation de ce mémoire. Mille mercis à vous, je n'aurai pas pu le faire sans votre soutien.

Table des matières

Introduction **1**

1 Contexte **2**
1.1 L'assurance multirisque habitation en France . . . 2
1.1.1 Les biens assurables . . . 2
1.1.2 Les dommages assurables aux biens . . . 2
1.1.3 Les responsabilités assurables . . . 3
1.1.4 La protection juridique et l'assistance à domicile . . . 3
1.1.5 Le caractère obligatoire de l'assurance multirisque habitation . . . 4
1.1.6 Le marché de l'assurance multirisque habitation aujourd'hui . . . 4
1.2 L'*Open Data* en France . . . 6
1.2.1 Définition . . . 6
1.2.2 Un peu d'histoire . . . 6
1.2.3 Les utilisations possibles . . . 8
1.3 Les enjeux du mémoire . . . 9
1.3.1 L'expérience client en assurance multirisque habitation . . . 9
1.3.2 Le processus de devis et souscription client . . . 10
1.3.3 Les enjeux de la tarification à l'adresse . . . 12

2 Présentation de la base de données « assureur » **14**
2.1 La base de données . . . 14
2.2 La stabilité du portefeuille . . . 17
2.3 Les analyses univariées et bivariées . . . 18
2.3.1 L'analyse de la garantie dégâts des eaux . . . 18
2.3.2 L'analyse de la garantie vol . . . 22
2.4 Les corrélations entre les différentes variables . . . 26
2.4.1 Le test d'indépendance du χ^2 . . . 26
2.4.2 Le V de Cramer . . . 27
2.4.3 Le calcul des corrélations entre les variables . . . 27

3 Théorie **29**
3.1 La reconnaissance d'images . . . 29
3.1.1 La représentation informatique d'une image . . . 30
3.1.2 Quelques opérations utiles . . . 32
3.2 Les méthodes d'apprentissage statistique . . . 34
3.2.1 Les modèles collectifs . . . 34
3.2.2 Quelques généralités sur l'apprentissage statistique . . . 35
3.2.3 Les principes majeurs de l'apprentissage statistique supervisé . . . 36
3.2.4 Le modèle linéaire généralisé . . . 40
3.2.5 Les forêts aléatoires . . . 43
3.2.6 La comparaison des modèles . . . 46

4 Constitution de la base de données « à l'adresse » 47
4.1 La base de données . . . 47
4.1.1 Quelques données géographiques préliminaires . . . 48
4.1.2 Les données à la maille adresse . . . 53
4.1.3 Les données à la maille commune/département . . . 62
4.1.4 Les données météorologiques . . . 64
4.1.5 Résumé de l'ensemble des données . . . 65
4.2 Les analyses univariées et bivariées . . . 67
4.2.1 L'analyse univariée de la garantie dégâts des eaux . . . 67
4.2.2 L'analyse univariée de la garantie vol . . . 69
4.2.3 Les analyses bivariées . . . 72
4.3 Les corrélations entre les différentes variables . . . 72

5 Mise en place des méthodes de tarification 73
5.1 Le modèle de tarification traditionnelle . . . 73
5.1.1 La modélisation de la garantie dégâts des eaux . . . 74
5.1.2 La modélisation de la garantie vol . . . 78
5.2 Le GLM à l'adresse . . . 82
5.2.1 Problématiques liées à l'utilisation de l'*Open Data* . . . 82
5.2.2 La modélisation de la garantie dégâts des eaux . . . 84
5.2.3 La modélisation de la garantie vol . . . 90
5.3 La forêt aléatoire à l'adresse . . . 94
5.3.1 La modélisation de la garantie dégâts des eaux . . . 95
5.3.2 La modélisation de la garantie vol . . . 97
5.4 Comparaison des différents modèles . . . 99

6 Marché concurrentiel 101
6.1 Les hypothèses du marché concurrentiel . . . 101
6.1.1 Les clients . . . 101
6.1.2 Les assureurs . . . 102
6.2 La tarification traditionnelle versus le GLM à l'adresse . . . 103
6.3 La tarification traditionnelle versus la forêt aléatoire à l'adresse . . . 105
6.4 Conclusion et limites . . . 107

Conclusion 108

Bibliographie 109

Introduction

L'assurance multirisque habitation (MRH), bien qu'étant un secteur stable, est aujourd'hui saturée et de nombreux acteurs se disputent les différentes parts du marché. Les diverses lois promulguées ces dernières années (lois Hamon et Chatel) ont d'ailleurs accru cette concurrence, déjà bien installée, en facilitant la résiliation des contrats.

Dans ce contexte, les assureurs doivent donc se démarquer que ce soit pour fidéliser leurs clients ou en attirer de nouveaux. Ainsi, pour se différencier, beaucoup se sont tournés vers un argument de vente majeur : la satisfaction client. De nombreuses actions ont en effet été menées ces dernières années afin d'améliorer la satisfaction des utilisateurs : suivi personnalisé des clients, déclarations facilitées des sinistres en ligne, remboursement rapide des sinistres ou encore digitalisation des démarches. Si ces mesures ont contribué à améliorer l'expérience client dans certains cas, de nombreux travaux restent encore à entreprendre dans ce domaine, notamment au niveau du processus de souscription client.

En effet, ce processus apparaît comme étant long et fastidieux pour le client et lui impose de répondre à de multiples questions pouvant très souvent le conduire à abandonner le processus avant sa fin et ce, malgré la digitalisation des démarches.

Un enjeu actuel consiste donc à faciliter davantage ce processus de souscription en diminuant, par exemple, le nombre de questions posées dans le formulaire. Pour ce faire, une solution possible consiste en la mise en place d'un processus de tarification à l'adresse, c'est-à-dire, un processus de tarification basé sur une seule variable : l'adresse. Ce mémoire propose donc une approche dans laquelle une seule question sera posée au client. Ainsi, moins de 30 secondes d'attention lui seront nécessaires pour souscrire son assurance MRH. Ce faisant, le gain de temps serait non négligeable pour le client. Côté assureur, la simplification du processus de souscription deviendrait, de fait, un argument marketing, à valoriser, permettant d'attirer une nouvelle clientèle. Elle permettrait également d'augmenter le taux de conversion des clients potentiels en clients réels.

Si la mise en place opérationnelle d'une telle procédure s'avérait encore impossible il y a quelques années, la croissance du *Big Data* et notamment le développement de l'*Open Data* permettent aujourd'hui d'élargir le champ des possibles. Dans le souci d'améliorer l'expérience client, ce mémoire va donc s'intéresser à la possibilité de mettre en place une telle méthode et examiner la qualité de la démarche par rapport à la méthode de tarification traditionnelle.

Après un examen de la base de données mise à disposition pour réaliser cette étude (base dite « assureur » dans la suite du mémoire et contenant uniquement les variables déclaratives de la souscription), un rappel théorique des différentes méthodes utilisées sera effectué. La constitution de la base de données dite « à l'adresse » fera ensuite l'objet d'une présentation détaillée. Cette dernière sera construite à partir de l'adresse des assurés et sera complétée par des données en *Open Data* et, à la marge, des données internes à Sia Partners : aucune variable déclarative ne sera donc présente. Enfin, les modèles traditionnel et à l'adresse mis en place seront explicités et interprétés avant de mesurer l'impact de la tarification à l'adresse par rapport à la tarification traditionnelle dans un marché concurrentiel.

Chapitre 1

Contexte

Composé de trois parties, ce chapitre a vocation à apporter des éléments de contexte afin de mieux appréhender les problématiques traitées dans ce mémoire. Le marché de l'assurance multirisque habitation et l'histoire de l'*Open Data* vont ainsi être présentés avant d'expliquer l'enjeu du mémoire visant à lier ces deux thématiques.

1.1 L'assurance multirisque habitation en France

Le contrat d'assurance multirisque habitation (MRH) est un contrat ayant pour double objectif de couvrir les dégâts subis par le patrimoine de l'assuré lorsque ce dernier est victime d'un sinistre et de couvrir les dégâts causés par l'assuré lorsque ce dernier est responsable d'un sinistre. Ce contrat comporte également, parfois, des garanties de protection juridique et d'assistance à domicile.

1.1.1 Les biens assurables

Dans le cadre d'un tel contrat, trois types de biens sont assurables et constituent le patrimoine de l'assuré. Parmi ces biens assurables, sont présents :

- Les **bâtiments** (et leurs installations) qui peuvent être des maisons, des appartements, des greniers, des caves, des garages ou encore des abris de jardins ;
- Le **mobilier personnel** qui représente l'ensemble des meubles et objets personnels appartenant aux personnes résidant ou se trouvant momentanément dans les lieux assurés ;
- Les **biens à usage professionnel** qui regroupent l'ensemble des objets utilisés pour les besoins de la profession de l'assuré. La couverture de cette catégorie de biens est en général accordée de manière optionnelle.

1.1.2 Les dommages assurables aux biens

Ces biens peuvent être couverts contre différents dommages en fonction des garanties souscrites. Les garanties pouvant figurer dans un contrat d'assurance MRH sont les suivantes :

- La garantie **incendie-explosion** qui indemnise les dommages matériels résultant d'un incendie d'origine accidentelle, d'une explosion, d'une implosion, de la chute de la foudre, ou encore des dommages matériels provoqués en éteignant un feu (lors de l'intervention des secours) ;
- La garantie **dommages électriques** qui complète la garantie incendie et couvre les appareils électriques endommagés par la foudre, une surtension, sous-tension ou encore un court-circuit ;

- La garantie **dégâts des eaux** qui prend en charge les dégâts causés lors d'une fuite d'eau, d'une rupture des conduites, d'un débordement des canalisations d'eau, d'infiltrations accidentelles ou encore d'inondations. Il est à noter que cette garantie couvre les conséquences d'un dégât des eaux mais ne prend en aucun cas en charge les réparations à l'origine du dommage ;
- La garantie **tempête, grêle et neige** qui couvre les dommages causés au domicile par les tempêtes, les pluies de grêle et les chutes de neige ;
- La garantie **vol** qui couvre la disparition, la destruction ou la détérioration des biens mobiliers résultant de vols et de tentatives de vol ;
- La garantie **vandalisme** qui indemnise les dommages matériels résultant d'actes de vandalisme ;
- La garantie **bris de glace** qui prend en charge les dommages matériels (bris, fissures, etc.) subis par les vitres, les fenêtres, les baies vitrées, les vélux, les garde-corps, les parois séparatrices de balcons, ainsi que les verres et glaces du mobilier ;
- La garantie **catastrophes technologiques** qui prend en charge de manière rapide et totale les particuliers victimes d'une catastrophe technologique ;
- La garantie **catastrophes naturelles** qui complète les garanties décrites ci-dessus qui ne s'appliquent pas en cas de catastrophe naturelle. Pour que la garantie s'active un arrêté d'état de catastrophe naturelle doit être publié au Journal Officiel ;
- La garantie en cas d'**actes de terrorisme/ d'attentats** ;
- La garantie en cas d'**émeutes**.

Cette liste est non exhaustive et de nombreuses autres garanties optionnelles peuvent figurer dans un contrat d'assurance MRH.

1.1.3 Les responsabilités assurables

En outre, une garantie responsabilité civile permet de couvrir les dégâts causés par l'assuré lorsque ce dernier est responsable d'un sinistre. Cette garantie peut être scindée en deux catégories :

- la garantie **responsabilité civile occupant**, qui couvre, pour un propriétaire, les dommages subis par les locataires ou des tiers à cause de son habitation, et qui couvre, pour un locataire, les dommages causés à l'habitation ou à des tiers ;
- la garantie **responsabilité civile vie privée**, qui couvre les dommages non-intentionnels causés par l'assuré à des tiers au cours de sa vie privée.

1.1.4 La protection juridique et l'assistance à domicile

Deux autres garanties peuvent également être présentes dans un contrat d'assurance MRH :

- La garantie **protection juridique** qui permet à l'assuré, dans le cadre d'un litige qui l'oppose à un tiers, d'avoir une partie de ses frais de procédure prise en charge par l'assureur et d'être assisté par un juriste afin de défendre ses droits ;

➥ La garantie **assistance à domicile** qui couvre :

— les transports à l'hôpital, gardes d'enfants, gardes d'animaux, etc., suite aux accidents subis par l'assuré à son domicile ;

— les frais d'hébergement, de gardiennage, de déménagement, d'aide-ménagère etc., suite aux conséquences d'un sinistre frappant le domicile de l'assuré ;

— la transmission de messages urgents à la famille ou à l'employeur de l'assuré etc., suite aux problèmes de la vie quotidienne, comme par exemple, en cas de vol ou perte de clefs du domicile assuré.

1.1.5 Le caractère obligatoire de l'assurance multirisque habitation

L'article 7 de la loi n°89-462 du 6 juillet 1989 relative aux rapports locatifs impose à l'ensemble des locataires de logement non meublé la détention d'une assurance habitation couvrant les risques locatifs. Ces risques locatifs comprennent les dommages causés au logement par un incendie, une explosion, ou un dégât des eaux. Cependant, ce minimum légal ne protège ni les biens personnels de l'assuré, ni sa responsabilité civile en cas de dommages corporels ou matériels causés à un tiers (des voisins par exemple). Le locataire étant responsable des dégâts qu'il pourrait causer tant à ses biens mobiliers qu'à ses voisins, il est donc fortement recommandé de souscrire à des garanties facultatives. Le contrat d'assurance MRH comprend donc l'ensemble de ces garanties, obligatoires et facultatives.

Par application de cette loi, le locataire d'un logement non meublé doit donc fournir chaque année à son bailleur une attestation d'assurance. En cas de défaut, le bailleur peut se servir de ce motif pour rompre le contrat. Le locataire reste cependant libre quant au choix de l'assureur.

Le 24 mars 2014, la loi Alur a étendu l'ensemble de ces directives aux locataires de logement meublé. Elle évoque également le cas des copropriétaires pour lesquels une garantie responsabilité civile est a minima obligatoire.

Enfin, pour ce qui est des propriétaires (occupants des lieux ou non), l'assurance habitation n'est pas obligatoire mais reste vivement conseillée.

1.1.6 Le marché de l'assurance multirisque habitation aujourd'hui

De nombreuses études telles que celles menées par la Fédération Française de l'Assurance (FFA) [9] [10] montrent que l'assurance MRH est un marché relativement stable. En 2019, ce marché représentait, selon les données de la FFA, un montant total de cotisations de 10 930 milliards d'euros soit environ **4.8% des cotisations reçues** par le marché de l'assurance française.

A la lecture du rapport annuel de la FFA de 2019, une certaine augmentation du montant total des cotisations est observable ces dernières années. Cette évolution est à mettre en parallèle tant avec l'augmentation du nombre de contrats qu'avec la hausse des primes moyennes hors-taxe, hausse pouvant elle-même être expliquée par l'inflation annuelle. Entre 2018 et 2019, la charge des sinistres, quant à elle, progresse sensiblement (notamment en ce qui concerne la garantie Tempête, Grêle et Neige). Cependant, cette augmentation reste moindre comparée à celle des cotisations. Le ratio combiné s'en trouve donc amélioré.

Concernant les garanties principales d'un contrat d'assurance MRH, une grande diversité de coût et de fréquence existe. Ainsi, certaines garanties sont peu coûteuses mais très fréquentes à l'instar de la garantie dégâts des eaux dont la fréquence moyenne est de 33,3% et le coût moyen de 1 082 euros. A contrario, d'autres garanties sont très coûteuses et peu fréquentes comme la garantie incendie avec une fréquence moyenne de 5,1% et un coût moyen de 8 272 euros.

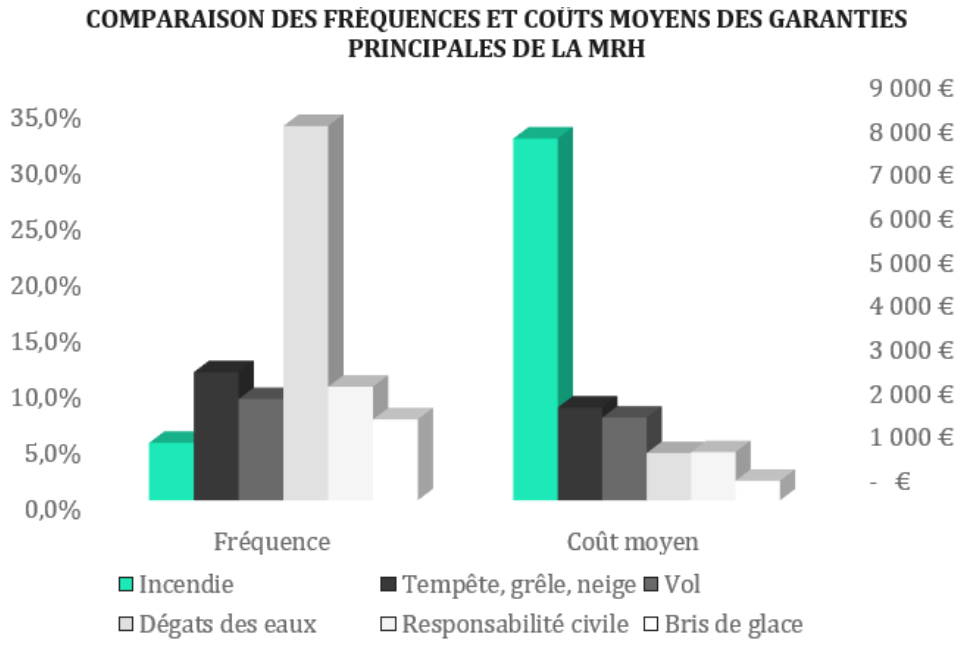


FIGURE 1.1 – Comparaison des fréquences et coûts moyens des garanties principales de l'assurance MRH, source : [9]

Une autre différence notable de comportement entre les garanties peut également être soulignée à travers le type de sinistres, les sinistres corporels de la responsabilité civile nécessitant généralement une prise en charge plus longue que les sinistres matériels des autres garanties.

L'assurance MRH, bien qu'étant un secteur stable, est aujourd'hui saturée et de nombreux acteurs se disputent les différentes parts du marché. Parmi eux, deux grands types de structure s'opposent : **les assureurs et les mutuelles** d'une part et **les bancassureurs** d'autre part. En effet, bien que représentant le mode de distribution historique, les assureurs et mutuelles sont aujourd'hui concurrencés par les bancassureurs, proposant les mêmes produits, mais jouissant d'un réseau de distribution conséquent et bien implanté ainsi que d'un contact étroit avec la clientèle. Ainsi, d'après Les Echos, c'est plus d'un contrat d'assurance MRH sur quatre qui est aujourd'hui commercialisé par un bancassureur et ce, très souvent, à l'occasion d'un prêt immobilier.

Les différentes lois promulguées ces dernières années ont d'ailleurs accru cette concurrence. En effet, auparavant, pour résilier un contrat d'assurance MRH (qui est un contrat à tacite reconduction), il était nécessaire de transmettre à l'assureur un courrier recommandé au minimum deux mois avant la date d'échéance annuelle du contrat. Depuis 2005, une nouvelle loi a été votée, la loi Chatel, dont l'objectif principal tend à conforter la confiance et la protection du consommateur. Cette loi a ainsi obligé les assureurs à rappeler au souscripteur leur date limite de résiliation avant reconduction tacite, facilitant alors la résiliation par les assurés. Par la suite, la loi Hamon, promulguée en 2015, a permis la résiliation à tout moment (à partir d'un an d'engagement) ce qui a eu pour effet d'exacerber encore plus la concurrence déjà bien présente sur le marché.

Aujourd'hui, dans le top 20 des assureurs habitation 2019 de la FFA [7], les bancassureurs sont donc de plus en plus représentés. C'est d'ailleurs ces derniers qui connaissent les plus fortes progressions de chiffres d'affaires. Ci-contre, en vert sont représentés les bancassureurs et en noir les assureurs et mutuelles.

Rang	Assureur	CA 2018 (M€)	CA 2017 (M€)	Variation 2018/2017
1	Covéa	1 808,0	1 753,0	3,1%
2	Groupama	1 140,5	1 117,0	2,1%
3	Crédit agricole Assurances	1 039,6	957,9	8,5%
4	Axa	1 021,0	1 017,0	0,4%
5	Groupe Maif	838,0	824,0	1,7%
6	Macif	790,2	780,9	1,2%
7	Allianz	627,0	614,0	2,1%
8	Groupe des assurances du Crédit mutuel	575,0	545,0	5,5%
9	Natixis Assurances	473,1	444,0	6,6%
10	Matmut	434,4	431,7	0,6%
11	Generali	351,0	340,0	3,2%
12	Aviva	194,4	194,8	-0,2%
13	La Banque postale Assurances IARD	164,2	150,7	9,0%
14	Société générale Assurances	156,0	148,0	5,4%
15	BNP Paribas Cardif	101,0	100,0	1,0%
16	Suravenir Assurances	98,5	96,3	2,3%
17	Mutuelle de Poitiers Assurances	85,4	82,2	3,9%
18	Thélem Assurances	76,0	73,2	3,8%
19	Groupe MACSF	66,4	62,8	5,7%
20	Sada Assurances	7,0	7,0	0,0%

FIGURE 1.2 – Top 20 des assureurs habitation 2019 de la FFA (chiffres France hors taxes 2018), source : [7]

1.2 L'*Open Data* en France

1.2.1 Définition

L'*Open Data* (ou données ouvertes en français) désigne l'ensemble des données numériques en libre accès, c'est-à-dire les données accessibles et utilisables par toute personne tierce et pouvant être partagées librement. Ces données peuvent provenir de services publics comme privés. Cependant, la plupart du temps, ce sont les services publics qui en sont à l'origine. Dans le cadre de ce mémoire, une attention particulière sera portée aux données issues de ces services publics.

Pour que les données soient considérées comme ouvertes, il ne doit y avoir aucune restriction qu'elle soit d'ordre technique, juridique ou financière. Il ne doit donc y avoir aucune condition de réutilisation : n'importe quel utilisateur doit être libre d'utiliser, de modifier, de combiner ou même de partager la donnée, y compris à des fins commerciales. Les licences ouvertes s'assurent du respect de ce principe. En droit français, les données doivent également faire l'objet d'une autorisation préalable de publication et, très souvent, d'une anonymisation de manière à ne pas contenir d'informations sensibles (par exemple il ne doit pas être possible d'identifier une personne).

Il est également important de ne pas faire d'amalgames lorsque l'on parle d'*Open Data*. Le format, la structure ou encore la lisibilité des données par la machine n'influent pas sur le caractère ouvert de la donnée. L'ouverture des données est relative aux possibilités d'utilisation et non aux modes de mise à disposition. Ainsi, un fichier CSV est certes plus accessible à un utilisateur tiers qu'une base SQL, mais cela ne rend pas pour autant la donnée plus ouverte.

1.2.2 Un peu d'histoire

Historiquement, ce n'est que depuis quelques années qu'il existe un réel déploiement de l'*Open Data*. Ce développement résulte notamment d'une volonté de transparence accrue des différentes administrations envers les citoyens.

Par le passé, différentes réglementations ont édicté des principes précurseurs à l'*Open Data*. Par exemple, dès 1789, l'article 15 de la déclaration des droits de l'homme et du citoyen évoquait le fait que tout citoyen a le droit de demander des comptes aux agents publics de son administration. La loi CADA du 17 juillet 1978 complète cela et institue le principe de libre accès aux documents administratifs. Grâce à cette loi, toute personne peut, en en faisant la demande préalable, obtenir la communication de documents administratifs. Ces deux textes donnent donc un premier cadre juridique au libre accès aux informations publiques. Cependant, l'*Open Data* est encore loin et aucune donnée n'est encore publiée sans demande expresse, de manière pro-active et massive.

A l'échelle européenne, il faut attendre encore un peu pour que davantage de mesures soient prises en faveur de l'*Open Data*. Ainsi, c'est seulement en 2003 que l'Union Européenne (UE) encourage le développement de l'*Open Data* avec la directive « Informations du Secteur Public » ou ISP. D'autres directives européennes, telles que la directive INSPIRE de 2008 destinée à favoriser l'échange de données à l'échelle européenne dans le domaine de l'environnement, seront ensuite mises en place pour favoriser le libre accès aux données.

En France, l'*Open Data* se développe d'abord au niveau des collectivités territoriales avec l'exemple de Rennes et de Paris en 2010, puis au niveau national, avec la mission Etalab en 2011. Cette mission, visant à développer et coordonner la politique publique de données ouvertes, a permis le lancement du portail data.gouv.fr et a marqué l'arrivée de l'*Open Data* dans les mœurs françaises. Différents textes ont par la suite été votés avec notamment la loi d'octobre 2016 pour une république numérique.

Aujourd'hui, la France fait partie des leaders européens et mondiaux en matière d'*Open Data* et différents classements le prouvent.

Ainsi, au niveau européen, l'*Open Data Maturity Report* de la Commission européenne [20] évalue les progrès réalisés par les pays européens selon quatre critères :

- la maturité du cadre politique ;
- la maturité du portail national d'*Open Data* ;
- la qualité de la mesure de l'impact de l'*Open Data* ;
- la qualité des données ouvertes.

Avec un score global de 89% la France occupe, en 2019, la 3ème place derrière l'Irlande (91%) et l'Espagne (90%).

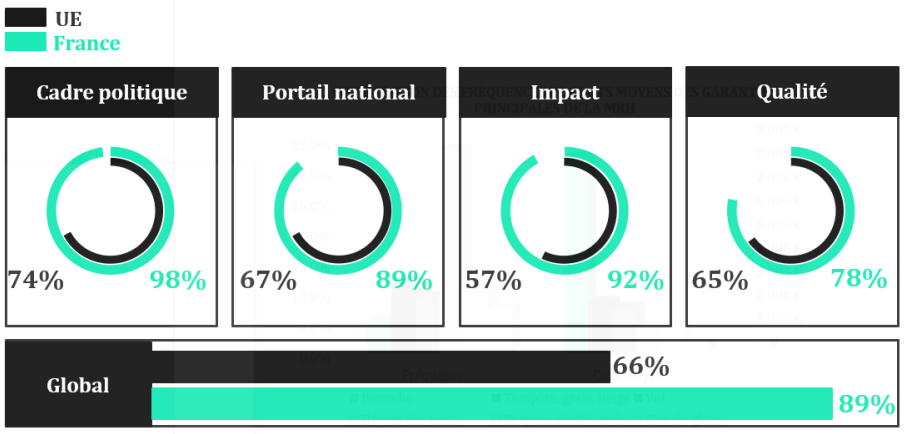


FIGURE 1.3 – Score de la France selon l'*Open Data Maturity Report*, source : [20]

L'*OURdata Index* est un autre indice évaluant l'efficacité des politiques d'ouverture des données gouvernementales. Ce dernier est mondial et repose sur trois indicateurs essentiels :

- la disponibilité des données (notamment le nombre de jeux de données disponibles) ;
- l'accessibilité des données (notamment la qualité et la complétude des données) ;
- le soutien gouvernemental à la réutilisation des données (la promotion de l'*Open Data*, la mise en place de formations ou encore la mesure de l'impact de l'ouverture).

En 2019 [19], au sein de ce classement, la France s'est vue attribuer la 2ème place se plaçant alors devant de nombreux pays de l'Organisation de coopération et de développement économiques (l'OCDE) comme le Brésil, la Chine, l'Inde, le Japon, etc.

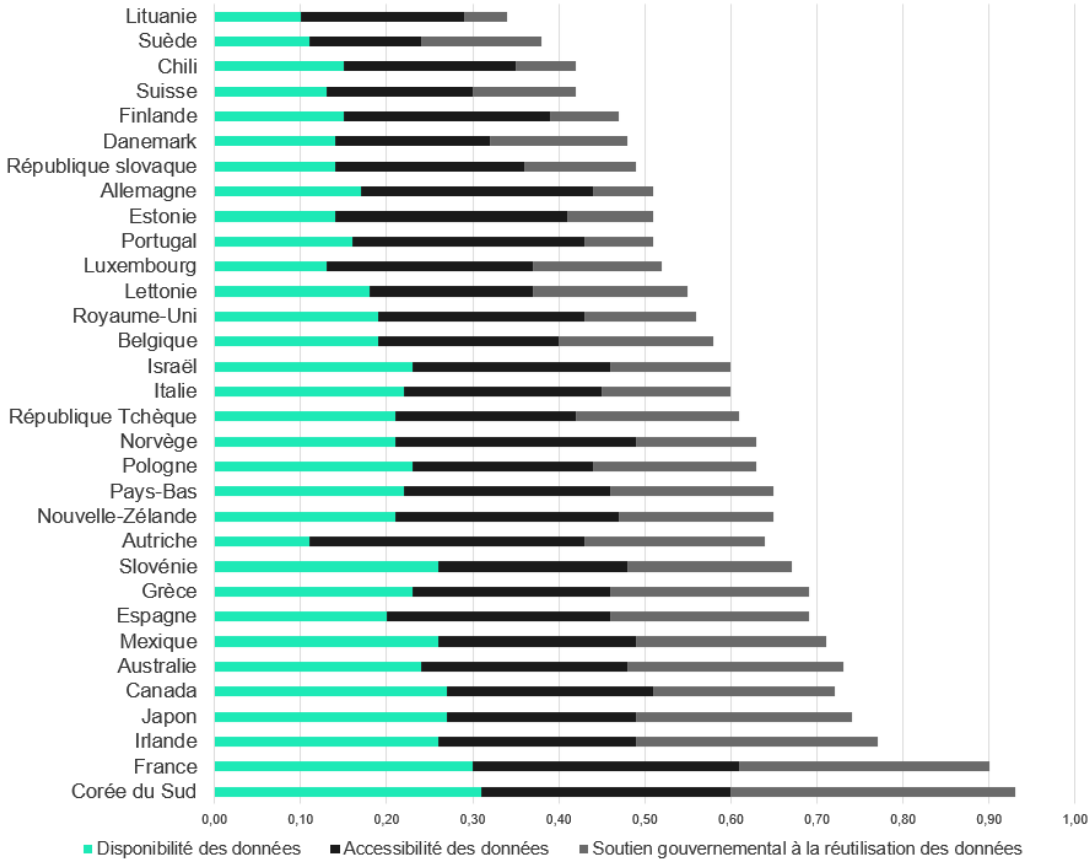


FIGURE 1.4 – Classement *OURdata Index*, source : [19]

Parallèlement au développement de l'*Open Data* public, de nombreuses entreprises privées telles que la SNCF, la RATP ou encore ENGIE ont perçu l'importance de l'*Open Data* et ont commencé à partager certaines de leurs données en libre accès.

1.2.3 Les utilisations possibles

Ces données nouvellement ouvertes permettent de répondre à de nombreux besoins.

Tout d'abord, elles participent à l'élaboration d'une transformation gouvernementale visant à améliorer la transparence des administrations envers les citoyens. Cela peut s'illustrer notamment dans le cadre de l'utilisation des fonds publics : l'*Open Data* aide ainsi à prouver que les fonds publics sont dépensés à bon escient et que les politiques sont bien implémentées.

L'*Open Data* offre également de nouvelles opportunités aux entreprises en se positionnant au cœur d'un processus de création de valeur visant à transformer les données en information puis en savoir. De nombreuses applications sont ainsi apparues suite à l'ouverture des données. Pour n'en citer que quelques exemples :

- le site « Où recycler » qui recense les sites de recyclage du verre, des bouchons, du plastique, des ampoules, des téléphones, des vêtements, des chaussures, des cartouches d'encre, etc. dans la France entière ;
- l'application « Où sont les toilettes » qui répertorie les toilettes publiques en France ;
- le site « Géovélo » proposant des itinéraires adaptés aux cyclistes en France notamment.

L'*Open Data* constitue donc un levier de création de valeur pour de nombreuses entreprises évoluant sur des marchés et secteurs variés et il est certain, qu'avec l'accélération de la digitalisation de la société, davantage de réutilisations de données verront le jour. Néanmoins, ce mémoire se focalisera exclusivement sur les possibilités offertes par l'*Open Data* dans le cadre de la tarification non-vie.

1.3 Les enjeux du mémoire

1.3.1 L'expérience client en assurance multirisque habitation

Ces dernières années, il a été démontré que :

- « Les entreprises qui offrent une **meilleure expérience client** obtiennent des **revenus entre** 4% **et** 8% **supérieurs** à leur marché. » Source : Bain & Company ;
- « 89% des consommateurs ont commencé **à faire affaire avec un concurrent** suite à une **expérience client médiocre**. » Source : Harris Interactive ;
- « Une **augmentation de** 5% **de la fidélisation** de la clientèle peut générer 25% **de bénéfices supplémentaires**. » Source : Bain & Company ;
- « Lors d'un achat, 64% des gens considèrent que l'**expérience client est plus importante que le prix**. » Source : Gartner ;
- « 52% des consommateurs déclarent avoir effectué un **achat supplémentaire** auprès d'une entreprise **après une expérience de service client positive**. » Source : Dimensional Research.

L'ensemble de ces chiffres montre donc l'importance de la satisfaction client à l'heure actuelle que ce soit pour fidéliser des clients ou en attirer de nouveaux. Ce constat est observable dans tous les secteurs et en particulier celui de l'assurance MRH. En effet, ce dernier marché étant saturé et ultra-concurrentiel, il est devenu difficile aujourd'hui pour les assureurs de se démarquer les uns des autres avec les tarifs proposés. Il a donc fallu, pour se départager, que ces derniers mettent l'accent sur un autre argument de vente : la satisfaction client. Ainsi, par exemple, selon une récente étude de la CGI [1] et de l'EBG [2], plus de 78% des professionnels de l'assurance estiment que l'amélioration de l'expérience client est une priorité *business* cruciale.

De nombreux assureurs et bancassureurs tentent donc d'améliorer l'expérience de leurs clients en plaçant la satisfaction de ces derniers au cœur de leurs préoccupations. Pour ce faire, il est important de maîtriser l'ensemble du parcours client, de l'écriture du devis à la résiliation du contrat.

1. Une entreprises de services numériques et de conseil.
2. Le principal think-tank français sur l'innovation digitale.

FIGURE 1.5 – Parcours type d'un client en assurance MRH

L'étude de ce parcours type fait ressortir de nombreuses pistes d'amélioration :

- **Devis & Souscription** : processus lourd et long avec beaucoup de formalités administratives et une certaine difficulté de compréhension du périmètre des garanties ;
- **Premier remboursement** : processus lourd et complexe pour la demande de remboursement (i.e. la déclaration du sinistre) et un manque d'informations claires sur le suivi du remboursement ;
- **Évolution du contrat** : difficulté pour entrée en contact avec l'assureur et manque de suivi ;
- **Reconduction** : pas de valorisation de la fidélité.

Face à ce constat, différentes actions ont été apportées pour améliorer ces points et l'un des leviers majeurs reste aujourd'hui la digitalisation. Quatre néo-assureurs, les « 4L » (Luko, Leocare, Lovys, Lemonade), l'ont bien compris et proposent aujourd'hui des contrats personnalisables avec des garanties lisibles, une déclaration de sinistres facilitée à l'aide d'une simple photo ou vidéo prise avec son smartphone, et plus généralement des parcours clients simplifiés et digitalisés. Le gain de temps pour le souscripteur est alors indéniable : à l'inscription, lors de la déclaration et même au moment de l'indemnisation (Leocare promet par exemple de diviser le temps de traitement des dossiers par trois lors de l'indemnisation).

Toutes ces mesures contribuent donc à l'amélioration de l'expérience client. Cependant, de nombreux travaux sont encore à entreprendre, notamment au niveau du processus de souscription client. C'est d'ailleurs cette étape du parcours client qui sera d'intérêt dans le cadre de ce mémoire.

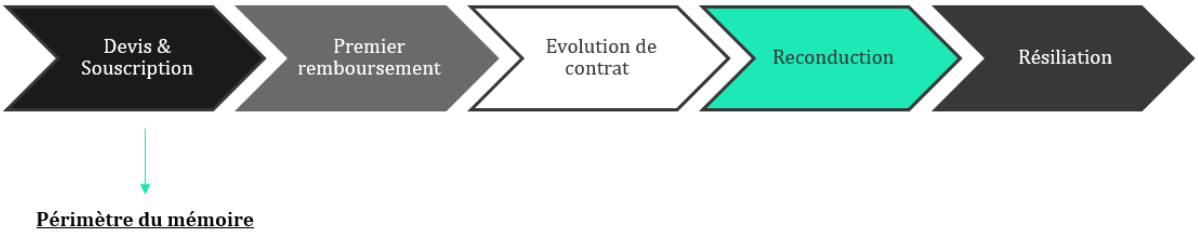


FIGURE 1.6 – Périmètre du mémoire

1.3.2 Le processus de devis et souscription client

Lors de la réalisation d'un devis pour un contrat d'assurance MRH, le client potentiel doit répondre à un ensemble de questions afin que l'assureur puisse cerner au mieux son risque et trouver le tarif approprié à la couverture de ce risque. Une fois ce devis réalisé, le client peut alors décider de souscrire à l'offre si cette dernière l'intéresse. Il devra alors remplir de nombreuses formalités administratives.

Côté assureur, un tel processus entraîne forcément des charges importantes liées à la gestion administrative qui se répercutent sur le prix de l'assurance et empêchent l'assureur de proposer des tarifs avantageux. Côté client, ce parcours est long, fastidieux et ce d'autant plus en mettant

en perspective le fait qu'une personne réalisera probablement plusieurs devis avant de trouver l'assurance lui correspondant (sans utilisation de comparateur).

Afin d'optimiser ce processus tant pour les clients potentiels que pour les assureurs, certains assureurs ont décidé d'y apporter des modifications en profondeur ces dernières années. Ainsi, plusieurs acteurs du marché ont décidé de digitaliser leur processus de souscription facilitant ainsi la démarche. D'autres se sont tournés vers l'intelligence artificielle afin d'automatiser certaines tâches : reconnaissance instantanée des documents de l'assuré, extraction automatique des informations clés, relances automatiques en cas de pièces manquantes... Tout cela a permis de diminuer les charges administratives des assureurs ainsi que le temps humain que représente un processus de souscription.

Malgré toutes ces améliorations, le processus de devis et souscription reste très souvent, pour les clients potentiels, long et fastidieux de par le nombre de questions posées. Une enquête statistique[3] menée dans le cadre de ce mémoire montre en effet qu'en moyenne, les clients potentiels répondent à environ **27 questions** (ce nombre pouvant varier de 23 à 31) et mettent en moyenne pour cela **4 minutes et 38 secondes**.

En tenant compte du nombre de processus de souscription entamé par l'assuré afin de pouvoir choisir son assurance, ces chiffres prennent toute leur importance et expliquent pourquoi un grand nombre de clients potentiels abandonne très souvent le processus avant sa fin (entraînant ainsi une perte du chiffre d'affaires côté assureur).

Un enjeu actuel consiste donc à faciliter encore plus ce processus de souscription côté client en diminuant par exemple le nombre de questions posées dans le formulaire de souscription. Pour ce faire, une solution possible consiste en la mise en place d'un processus de tarification à l'adresse, c'est-à-dire une tarification à l'aide d'une seule variable : l'adresse. Ce mémoire propose donc une approche dans laquelle **une seule question** sera posée à l'utilisateur et demandera **moins de 30 secondes** d'attention à ce dernier.

FIGURE 1.7 – Enjeu du mémoire

La mise en place d'une telle solution, utopique il y a encore quelques temps, pourrait peut-être aujourd'hui s'avérer possible avec la croissance du *Big Data* et notamment le développement de l'*Open Data*. Dans un objectif d'amélioration de l'expérience client, ce mémoire va donc s'intéresser à la possibilité de mettre en place une telle méthode et examiner la qualité de la démarche par rapport à la méthode de tarification traditionnelle.

La tarification à l'adresse peut-elle concurrencer la tarification usuelle ?

3. Enquête statistique basée sur un échantillon d'une cinquantaine de personnes non averties ayant réalisé le devis d'un même assureur dont le processus de souscription est représentatif de ceux du marché actuel.

1.3.3 Les enjeux de la tarification à l'adresse

La mise en place d'un processus de tarification à l'adresse nécessite de récolter des données à des mailles très fines. Ces données peuvent être achetées ou bien récupérées en *Open Data*. Dans le cadre de ce mémoire, une attention toute particulière sera portée à l'*Open Data* qui constituera la source principale des données à l'adresse. Cependant, quelques informations déjà détenues en interne par Sia Partners seront utilisées à la marge.

Jusqu'à maintenant, de nombreuses études ont illustré la pertinence de l'utilisation de l'*Open Data* dans le but d'enrichir les modèles utilisés en assurance MRH. L'approche la plus fréquente consiste à intégrer l'*Open Data* et les nouvelles informations géographiques ainsi obtenues dans la création d'un zonier ou micro-zonier qui viendra compléter les variables usuellement tarifaires. L'ensemble des travaux s'étant intéressé à la question ont obtenu des résultats plutôt concluants que ce soit en assurance automobile ou en assurance MRH. L'apport de l'*Open Data* semble donc indéniable et permet d'améliorer les connaissances des assureurs. Cependant, aujourd'hui, de nombreux assureurs aimeraient aller encore plus loin dans le cadre de l'assurance MRH et mettre en place un processus de tarification à l'adresse c'est-à-dire un processus de tarification basé sur une seule variable : l'adresse de l'assuré. L'approche serait alors différente de celle des micro-zoniers et consisterait à utiliser directement et uniquement les données en *Open Data* dans la construction du tarif. Des questions quant à la faisabilité d'une telle pratique vont immanquablement être soulevées et ce mémoire a pour objet d'y répondre tout en tenant compte des diverses contraintes pesant sur les données utilisées.

En effet, afin de mettre en place un processus de tarification à l'adresse fiable, les données doivent être :

- **précises** : idéalement, il est souhaitable de disposer de données à la maille la plus fine possible c'est-à-dire l'adresse. Plus la maille est précise, plus la donnée sera juste et représentative de la zone géographique considérée. Cependant, plus l'information est précise, plus elle est difficile à obtenir. En outre, il se peut qu'à un moment, le fait d'affiner la précision de la donnée ne procure pas un effet significatif sur la tarification par rapport à la donnée précédente de maille plus grossière. Un compromis entre accessibilité de la donnée, précision de la maille et impact doit donc être effectué ;
- **de bonne qualité** : le processus de tarification à l'adresse relevant uniquement de données externes (essentiellement de l'*Open Data*), le tarif sera de pauvre qualité si les données ne sont pas fiables. Une étude de la qualité des données devra donc être réalisée avant toute interprétation des résultats ;
- **stables dans le temps** : il faut s'assurer que les données utilisées dans le cadre du processus de tarification à l'adresse puissent être accessibles à tout moment afin de pouvoir mettre à jour la tarification.

La donnée est donc au centre de ce mémoire. Elle fait l'objet de nombreux enjeux qui impacteront directement la qualité du tarif à l'adresse.

En supposant qu'une telle méthode de tarification à l'adresse puisse concurrencer la méthode de tarification traditionnelle, il conviendrait alors de s'intéresser aux impacts qu'elle pourrait avoir vis-à-vis des assureurs et des clients potentiels.

Ainsi, au lieu de répondre à une trentaine de questions, le client potentiel renseignerait uniquement son adresse économisant alors un temps considérable. Ce gain de temps porté à la connaissance des clients potentiels (campagnes de communication, alerte en début de procédure, ...), pourrait très certainement se révéler crucial dans la captation de nouvelles parts de marché. Cet atout concurrentiel serait en effet très probablement de nature à influencer les clients potentiels dans le choix de leur assureur. Quant à l'assureur, il pourrait alors augmenter ses parts de marché en se démarquant par rapport à la concurrence. Il pourrait ainsi :

- utiliser le processus de tarification à l'adresse comme un **argument marketing** permettant alors d'**attirer une nouvelle clientèle** (en mettant en avant l'économie de temps) ;
- **augmenter le taux de conversion** des prospects, ceux-ci étant plus nombreux à aller jusqu'au bout du processus de souscription compte tenu de la simplification de la démarche.

En outre, les informations recueillies en *Open Data* pourraient également participer à la vérification des informations déclarées par les assurés dans le cadre d'une souscription usuelle. Cela constituerait donc un outil supplémentaire pour lutter contre la fraude.

Un processus de tarification à l'adresse semble donc comporter de nombreux avantages. C'est pourquoi, aujourd'hui, de nombreux acteurs du secteur s'intéressent à la question et tentent de mettre en place de manière opérationnelle une telle méthode de tarification. Les acteurs qui seront les premiers à y parvenir disposeront d'un avantage considérable qui leur permettra de se démarquer.

Ce mémoire a donc pour objectif principal d'examiner s'il est possible, avec l'*Open Data* disponible à ce jour, de remplacer la méthode de tarification usuelle par une méthode de tarification à l'adresse. Il évoquera également les différents enjeux évoqués précédemment.

Après un examen de la base de données mise à disposition pour réaliser cette étude (base dite « assureur » dans la suite du mémoire et contenant uniquement les variables déclaratives de la souscription), un rappel théorique des différentes méthodes utilisées sera effectué. La constitution de la base de données dite « à l'adresse » fera ensuite l'objet d'une présentation détaillée. Cette dernière sera construite à partir de l'adresse des assurés et sera complétée par des données en *Open Data* et, à la marge, des données internes à Sia Partners : aucune variable déclarative ne sera donc présente. Enfin, les modèles traditionnel et à l'adresse mis en place seront explicités avant de mesurer l'impact de la tarification à l'adresse par rapport à la tarification traditionnelle dans un marché concurrentiel.

Chapitre 2

Présentation de la base de données « assureur »

Ce chapitre a pour objet de présenter la base de données « assureur » utilisée dans le cadre du mémoire. Après délimitation du périmètre du portefeuille, quelques chiffres seront donnés afin de le cerner dans sa globalité. Des analyses univariées et bivariées seront également menées afin de développer une première intuition quant au caractère tarifaire ou non de certaines variables. La corrélation des différentes variables fera également l'objet d'une attention particulière.

2.1 La base de données

La base de données de référence utilisée dans le cadre de cette étude correspond à un portefeuille multirisque habitation d'un assureur français sur la période 2010-2014.

En conservant l'objectif du mémoire en tête, à savoir mettre en place un processus de tarification à l'adresse, il a été décidé de se focaliser sur un périmètre restreint du portefeuille : les maisons. Ce choix a été fait pour plusieurs raisons, la principale étant qu'il est beaucoup plus aisé d'avoir des informations en *Open Data* sur les maisons plutôt que sur les appartements. Suite à cette restriction du périmètre, 800 000 lignes ont été conservées sur les 1 200 000 lignes initialement présentes dans la base. En outre, pour des raisons évidentes de temps, seules les lignes pour lesquelles des données en *Open Data* ont pu être récupérées ont été conservées : cela a eu pour effet de diminuer le nombre de lignes de la base à 255 000 lignes (suppression de 70% des 800 000 lignes initiales). Plus de détails sur ces suppressions seront donnés dans le chapitre 4 dédié à la constitution de la base de données « à l'adresse ».

Cette base réduite comporte donc 255 000 lignes et une quarantaine de variables. Quelques chiffres clés permettent de cerner la sinistralité de ce portefeuille dans son ensemble :

Année	Exposition	Coût total	Fréquence	Coût moyen par sinistre
2010	24 843	3 550 530 €	5,40%	2 645,7 €
2011	26 137	3 540 202 €	5,18%	2 614,6 €
2012	25 357	4 164 331 €	5,89%	2 787,4 €
2013	23 784	3 485 069 €	5,26%	2 788,1 €
2014	22 229	2 750 508 €	5,17%	2 393,8 €

FIGURE 2.1 – Quelques chiffres clés sur la base globale

Sinistralité

Cette sinistralité peut se distinguer en fonction des garanties présentes dans le contrat d'assurance MRH. Dans le cadre de ce portefeuille, quatre garanties sont présentes, garanties pour lesquelles tous les contrats de la base de données sont couverts :

- la garantie **dégâts des eaux** (DDE) ;
- la garantie **vol** ;
- la garantie **incendie** ;
- la garantie **responsabilité civile** (RC).

La répartition des sinistres selon ces garanties est représentée ci-dessous selon la fréquence et le coût des sinistres :

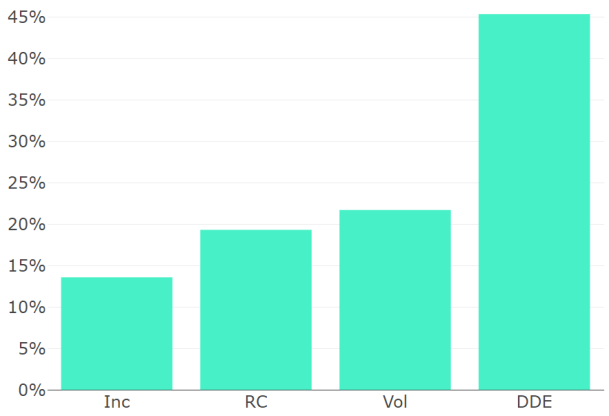


FIGURE 2.2 – Répartition en fréquence des sinistres par garantie

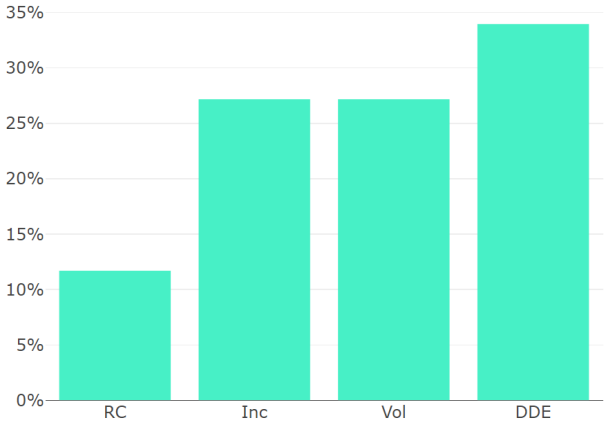


FIGURE 2.3 – Répartition en coût des sinistres par garantie

Sur ce portefeuille, il apparaît ainsi que les dégâts des eaux sont les sinistres les plus représentés que ce soit en coût ou en fréquence. Les sinistres incendie et vol représentent, pour leur part, environ 25% du coût total du portefeuille même si l'on observe une fréquence moindre des incendies par rapport aux vols. Cela s'explique par le fait que les incendies bien que plus rares, causent des dommages assez importants. Enfin, en ce qui concerne la RC, l'inverse s'observe et la sinistralité, bien qu'élevée, n'entraîne que des coûts relativement faibles.

Pour la suite du mémoire, seules les garanties DDE et vol seront traitées, ces dernières étant intuitivement les plus susceptibles d'être expliquées par des variables externes issues de l'*Open Data*. Il sera donc considéré par la suite que le produit d'assurance MRH tarifié ne comporte qu'une garantie DDE et une garantie vol.

Variables

Concernant les variables présentes dans la base, ces dernières peuvent être divisées en deux catégories : d'une part, les variables qui ne sont pas tarifaires liées à la police et au risque et d'autre part les variables potentiellement tarifaires.

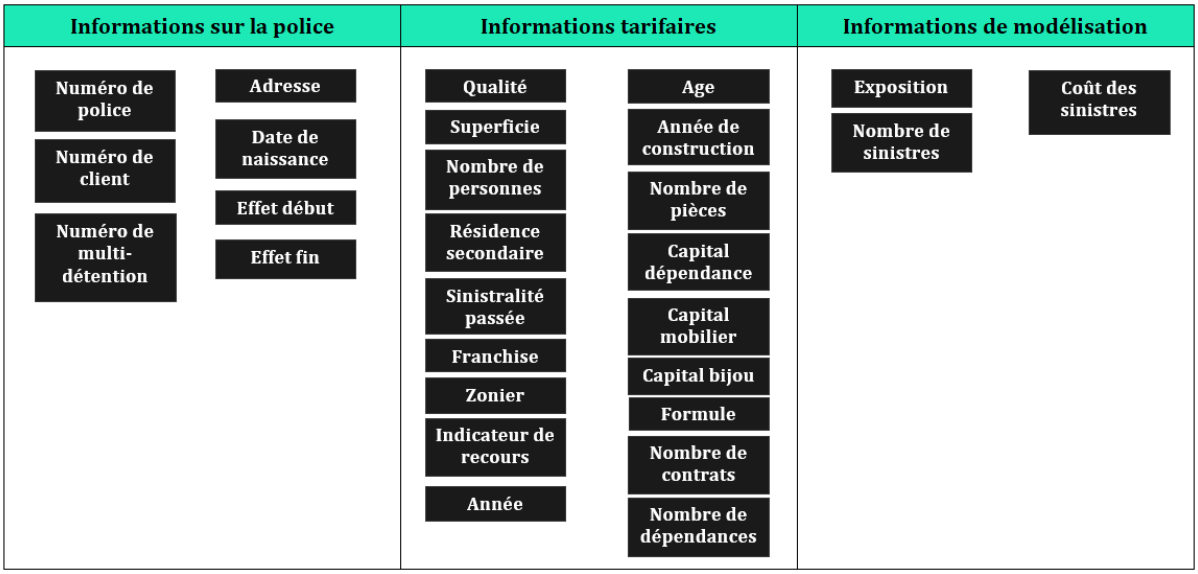


FIGURE 2.4 – Variables de la base « assureur »

Retraitements

Cette base de données ayant été utilisée lors de travaux internes, de nombreux retraitements ont déjà été effectués et seules quelques vérifications ont été réalisées concernant :

- **Les valeurs aberrantes et les informations manquantes** :

  La qualité des données étant primordiale lors de la mise en place d'un processus de tarification, il est nécessaire de s'assurer de l'absence de valeurs aberrantes et de la complétude des données.

  Concernant les valeurs aberrantes, des tests de cohérence ont été menés. Il en est ressorti qu'aucune valeur aberrante n'était présente. Après vérification, ces dernières avaient déjà été traitées lors de travaux internes à l'aide d'approximations simples (date de naissance retrouvée par le biais de l'âge, etc.).

  Concernant la complétude des données, il est apparu que l'ensemble des variables était globalement complète.

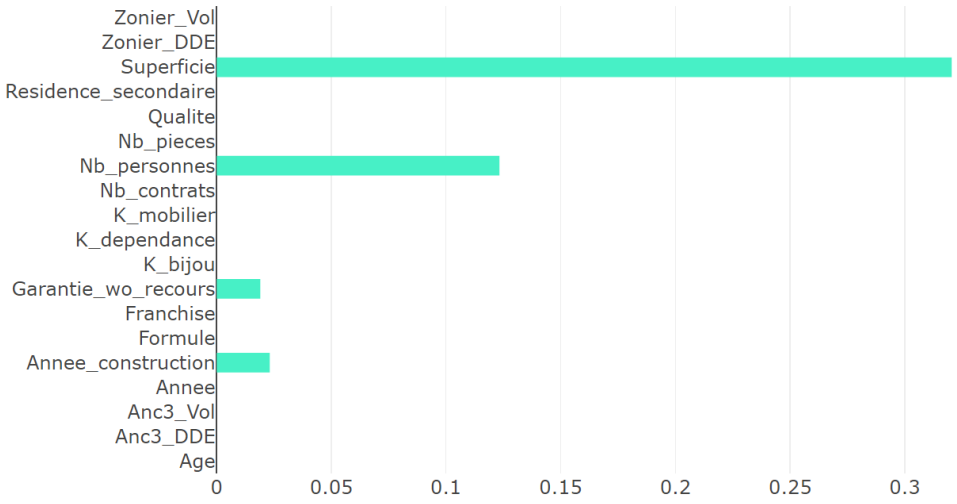


FIGURE 2.5 – Taux de *Not avalaible* (NA) par variable

Pour les quelques variables présentant de nombreuses informations manquantes, une imputation des données a été effectuée selon des règles simples (imputation par la moyenne, par la médiane, etc).

➥ **La prise en compte de l'inflation annuelle**

L'inflation annuelle constitue un biais important sur les coûts des sinistres de la base de données. En effet, le coût d'un même sinistre peut fortement évoluer en fonction de son année de survenance, du fait de l'augmentation régulière des coûts de construction et de réparation qui impactent d'autant le montant des indemnités. Cette évolution imputable à l'inflation est donc indépendante des caractéristiques du sinistre. Afin de corriger ce biais, il convient de retraiter les sinistres de l'inflation. Pour ce faire, plusieurs méthodes existent :

— intégrer l'année de survenance du sinistre au sein du modèle en tant que variable qualitative ;

— calculer le coût simulé du sinistre en considérant l'impact de l'inflation à l'aide de différents indices (BT01, FFB, IPC ...).

Dans le cadre de ce mémoire, il a été décidé de retenir la première méthode, à savoir intégrer au modèle l'année de survenance du sinistre en tant que variable qualitative.

➥ **Les sinistres graves**

Concernant les sinistres graves, il est nécessaire de les traiter en amont de la modélisation, ces derniers pouvant biaiser fortement les modèles. En cas de sinistres graves, différentes actions peuvent être menées, la plus courante étant l'écrêtement des sinistres. Aucun sinistre grave ne figurant dans la base de données, aucune de ces actions n'a été nécessaire.

Une fois l'ensemble de ces retraitements effectués, la base finale « assureur » est obtenue. Il convient ensuite de s'assurer de la stabilité du portefeuille dans le temps et de mener une analyse univariée/bivariée.

2.2 La stabilité du portefeuille

Pour s'assurer de la stabilité du portefeuille dans le temps, il a été réalisé des graphes permettant d'observer l'exposition de chaque modalité d'une variable au cours du temps. Tous ces graphes ne sont pas reportés dans ce mémoire mais en voici un exemple :

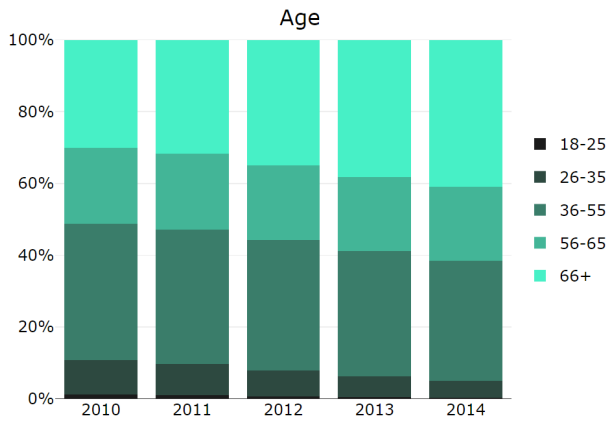


FIGURE 2.6 – Stabilité dans le temps de la variable âge

Après vérification de l'ensemble des variables de la base de données, il ressort que le portefeuille est globalement stable dans le temps. Il est donc maintenant possible de procéder aux analyses univariées et bivariées pour chaque garantie.

2.3 Les analyses univariées et bivariées

Suite à la présentation de la base de données, une analyse des données a été entreprise afin de mieux cerner le portefeuille et de développer une première intuition quant aux variables potentiellement tarifaires tant pour la garantie DDE que pour la garantie vol. Cette analyse permettra également de voir les regroupements de modalités pouvant avoir du sens. Dans la suite du mémoire, seuls quelques graphiques seront présentés pour chaque garantie et type d'analyse.

2.3.1 L'analyse de la garantie dégâts des eaux

Les analyses univariées en fréquence et en coût moyen présentent respectivement, pour chaque modalité d'une variable d'intérêt, la fréquence moyenne annuelle des sinistres et leur coût moyen.

Analyse univariée en fréquence

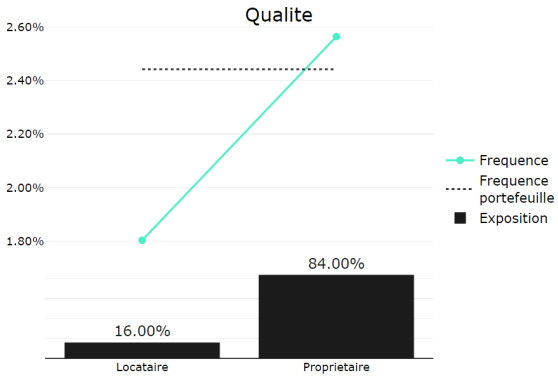


FIGURE 2.7 – Fréquence par qualité

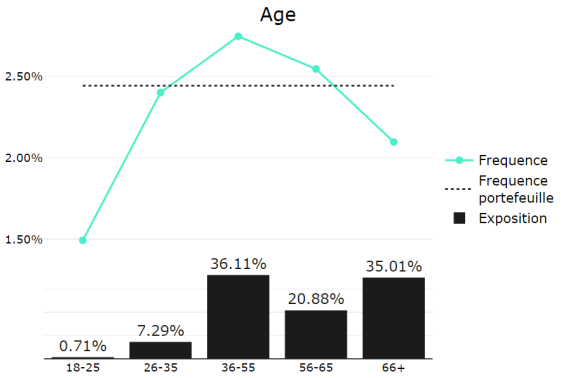


FIGURE 2.8 – Fréquence par classe d'âge

De ces graphiques, il ressort que les variables âge et qualité sont potentiellement tarifaires et impactent de manière non négligeable la fréquence des dégâts des eaux. Pour la variable âge, une courbe en cloche se dessine, ce qui reste cohérent avec la plupart des phénomènes observés en assurance MRH.

Il est à noter cependant la faible exposition de la modalité locataire pour la variable qualité et des modalités 18-25 et 26-35 pour la variable âge. Ceci provient du fait que, lors du traitement de la base de données, un périmètre réduit a été utilisé et seuls les assurés vivant dans une maison ont été conservés. Or, la plupart des assurés vivant dans une maison sont généralement propriétaires et ont plus de 35 ans : un déséquilibre d'exposition des modalités est donc créé.

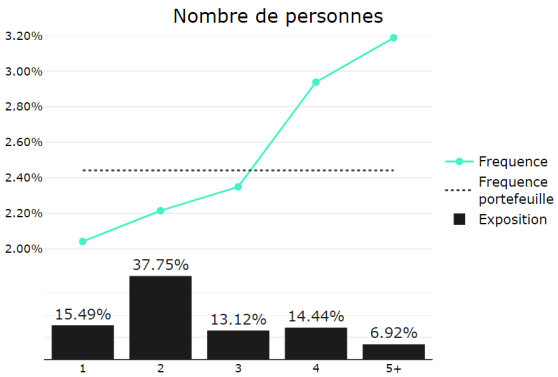


FIGURE 2.9 – Fréquence par nombre de personnes

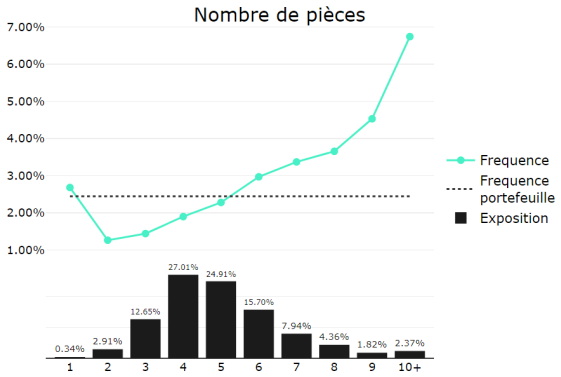


FIGURE 2.10 – Fréquence par nombre de pièces

Il apparaît que les deux variables nombre de personnes et de pièces impactent fortement la fréquence des sinistres. Ainsi, plus il y a de personnes ou de pièces dans le logement, plus la probabilité d'avoir un dégât des eaux semble forte ce qui reste cohérent avec la réalité du terrain.

Du fait du périmètre réduit, un biais potentiel est de nouveau observable pour le nombre de pièces avec très peu de maisons d'une ou deux pièces rendant ainsi l'interprétation de ces modalités plus délicate.

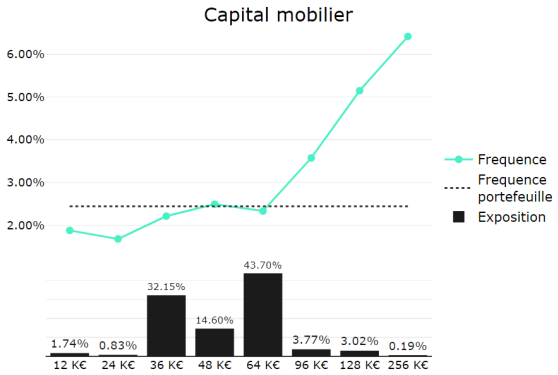


FIGURE 2.11 – Fréquence par montant mobilier assuré

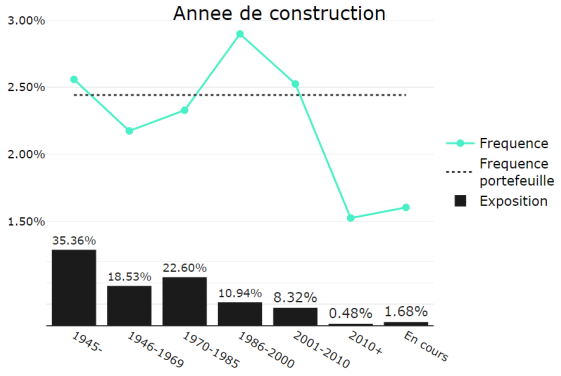


FIGURE 2.12 – Fréquence par année de construction

Pour le capital mobilier, la fréquence des sinistres semble augmenter drastiquement au fur et à mesure que le capital mobilier assuré augmente. Cependant, cette interprétation est à mettre en perspective avec la faible exposition de certaines modalités (respectivement les capitaux faibles et élevés).

A l'inverse pour la variable année de construction, il est plus difficile de se prononcer quant à un réel impact de la variable sur la fréquence des sinistres. Dans le doute, cette variable sera tout de même intégrée à la modélisation de la fréquence des dégâts des eaux.

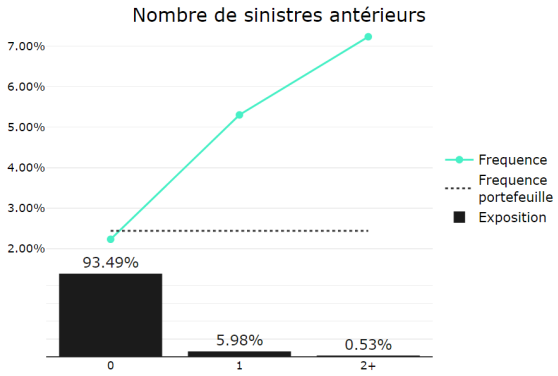


FIGURE 2.13 – Fréquence par sinistralité antérieure

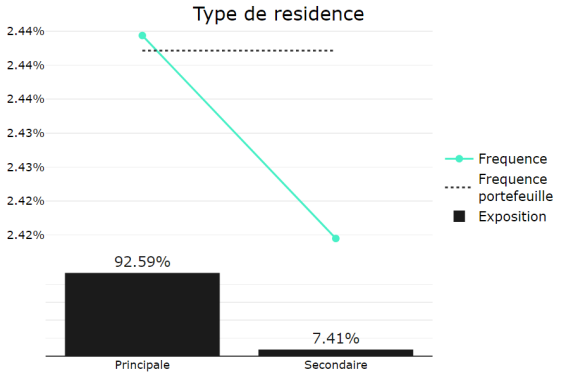


FIGURE 2.14 – Fréquence par type de résidence

Enfin, il apparaît que les assurés ayant déjà fait l'objet d'un dégât des eaux ont plus de risques d'en subir un de nouveau. Cependant, encore une fois, cette interprétation est à prendre avec précaution de par la faible exposition des modalités 1 et 2+.

Le type de résidence, quant à lui, semble plus faiblement impacter la fréquence des sinistres.

Pour l'ensemble des modalités ayant une faible exposition, il est possible d'envisager par la suite des regroupements de risques cohérents. Cela sera fait lors de la modélisation du risque de fréquence.

Analyse univariée en coût moyen

Le même raisonnement est effectué concernant le coût moyen d'un dégât des eaux.

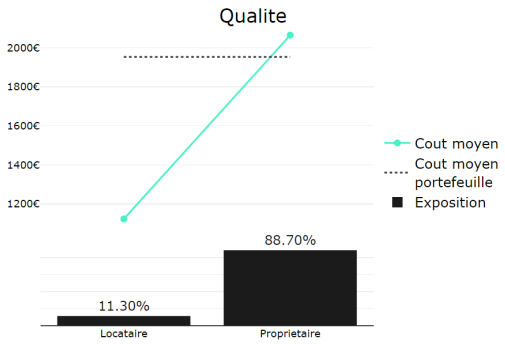


FIGURE 2.15 – Coût moyen par qualité

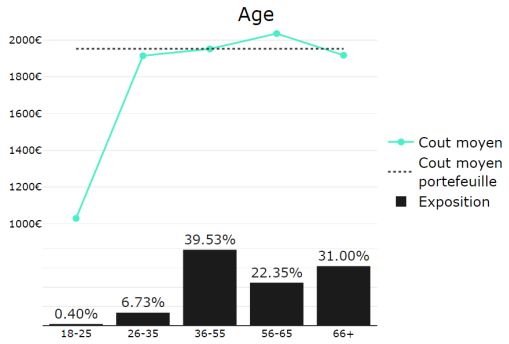


FIGURE 2.16 – Coût moyen par classe d'âge

Ainsi, la qualité de l'assuré impacte fortement le coût moyen d'un dégât des eaux tandis que l'âge de l'assuré n'influe que très peu sur ce coût (en omettant dans l'interprétation la modalité 18-25 qui est trop faiblement exposée).

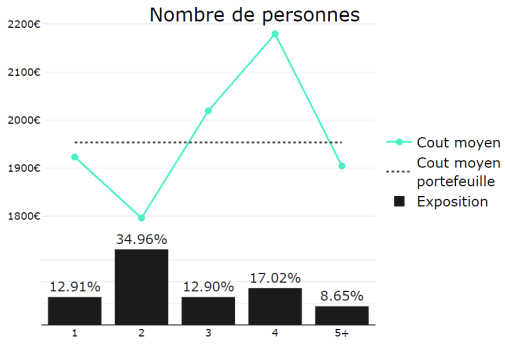


FIGURE 2.17 – Coût moyen par nombre de personnes

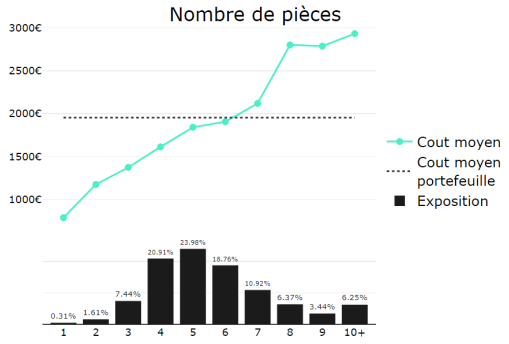


FIGURE 2.18 – Coût moyen par nombre de pièces

Concernant le nombre de personnes du logement assuré, il est délicat de conclure quant à un impact ou non de la variable. En revanche, il est clair que le nombre de pièces du logement est très tarifaire pour la modélisation du coût moyen des sinistres.

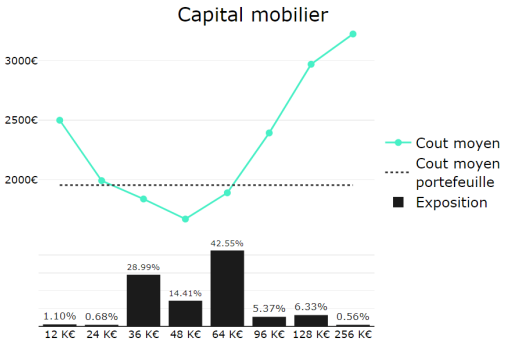


FIGURE 2.19 – Coût moyen par montant de mobilier assuré

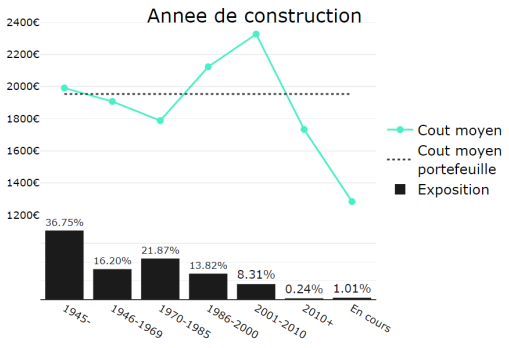


FIGURE 2.20 – Coût moyen par année de construction

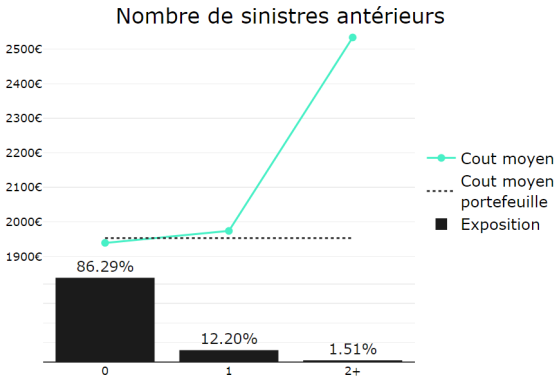


FIGURE 2.21 – Coût moyen par sinistralité antérieure

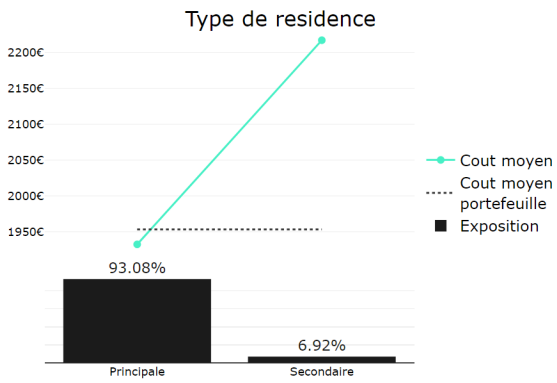


FIGURE 2.22 – Coût moyen par type de résidence

Le capital mobilier, quant à lui, semble également tarifaire. En revanche, eu égard aux variables année de construction, sinistralité antérieure et type de résidence, l'impact reste faible (pour l'année de construction et le type de résidence) voire quasi inexistant (pour la sinistralité antérieure).

Analyse bivariée

Il est également possible que l'influence simultanée de deux variables ait un impact non additif sur la variable à expliquer. Il est alors question d'impact bivarié et cela peut fortement influencer le modèle si ces interactions ne sont pas prises en compte. Dans le cadre de cette étude, seules les deux interactions les plus courantes en assurance MRH seront vérifiées, à savoir :

- l'interaction entre l'**âge** de l'occupant et sa **qualité** (locataire ou propriétaire) ;
- l'interaction entre le **nombre de pièces** du logement et la **qualité** de l'occupant.

Afin d'étudier la présence ou non d'un impact bivarié, il est nécessaire de représenter la variable cible en fonction d'une des variables prédictives, pour toutes les modalités de l'autre variable. Il est ensuite possible de conclure quant à l'interaction des deux variables en fonction de la forme des courbes. Ainsi, des courbes parallèles démontrent qu'aucun impact bivarié n'existe alors que des courbes se croisant indiquent la présence d'un impact bivarié.

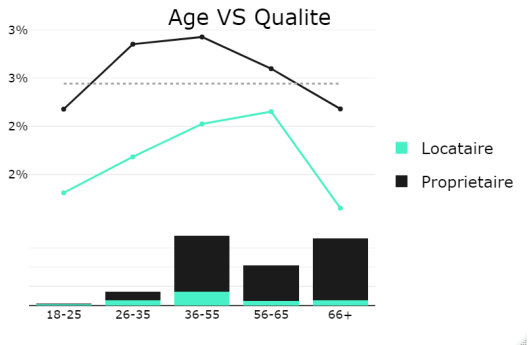


FIGURE 2.23 – Fréquence par classe d'âge et qualité

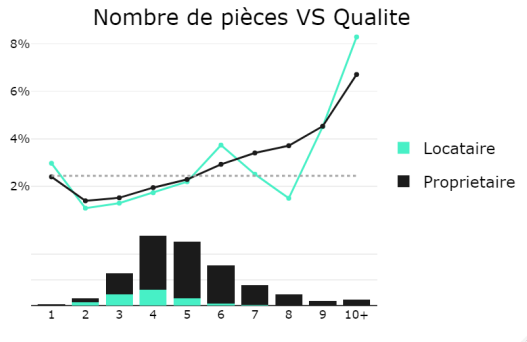


FIGURE 2.24 – Fréquence par nombre de pièces et qualité

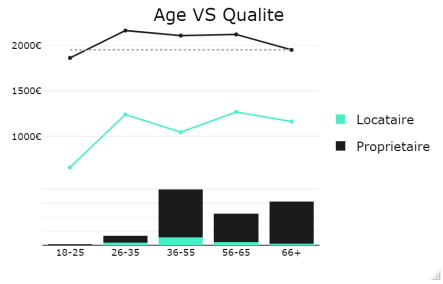


FIGURE 2.25 – Coût moyen par classe d'âge et qualité

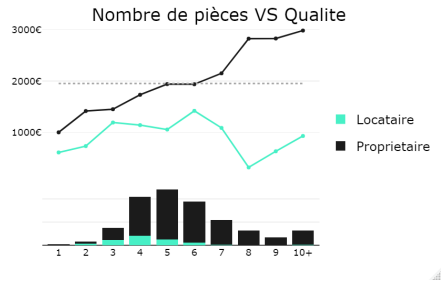


FIGURE 2.26 – Coût moyen par nombre de pièces et qualité

Toujours en ne considérant que les couples de modalités assez exposés pour être sujets à interprétation, et outre les fluctuations usuelles, il ressort qu'il n'existe, a priori, pas d'impacts bivariés dans ces analyses. Un doute peut cependant subsister entre l'âge et la qualité : les droites ne sont pas très parallèles autour de la modalité 56-65 lors de l'étude de la fréquence. Toutefois, il est possible d'attribuer cela à la faible exposition du couple Locataire/56-65. Concernant l'interaction nombre de pièces/qualité, les droites de fréquence sont bien parallèles pour les modalités 2, 3, 4 et 5 pièces (i.e. lorsque l'exposition est suffisante pour les couples nombre de pièces/qualité). En revanche, pour le coût, l'interprétation est plus difficile, les droites n'étant pas strictement parallèles pour ces mêmes couples. Il sera tout de même considéré par la suite que cela provient de légères fluctuations liées à la base de données et donc qu'aucun impact bivarié significatif n'est présent. Enfin, pour l'interaction âge/qualité en coût moyen, aucune interaction ne semble être présente.

En l'absence d'impacts bivariés significatifs, des modèles sans interaction vont donc être construits que ce soit pour la fréquence ou pour le coût.

Zonier

Enfin, il ne serait pas complet de ne pas terminer cette analyse sans mentionner le zonier à disposition, les zoniers étant généralement les seules variables représentant le risque géographique des assurés. Dans le cadre de cette étude, le zonier avait déjà fait l'objet de travaux précédents et était ainsi déjà disponible dans la base. Ce dernier a donc été utilisé en l'état.

2.3.2 L'analyse de la garantie vol

Une analyse similaire à celle menée pour la garantie DDE est maintenant réalisée pour la garantie vol.

Analyse univariée en fréquence

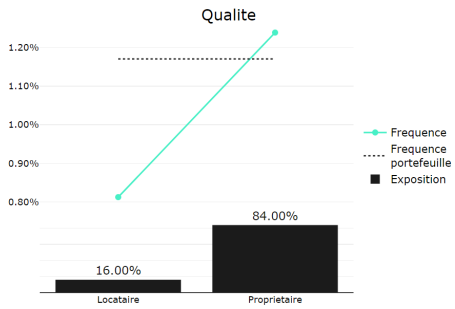


FIGURE 2.27 – Fréquence par qualité

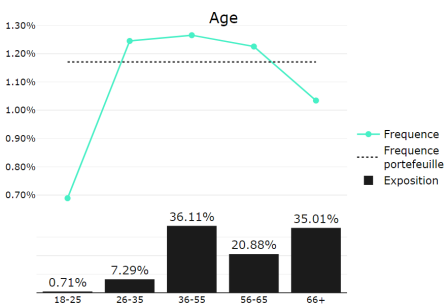


FIGURE 2.28 – Fréquence par classe d'âge

La qualité de l'assuré semble ainsi avoir un impact conséquent sur la fréquence des vols contrairement à l'âge qui semble être moins influent.

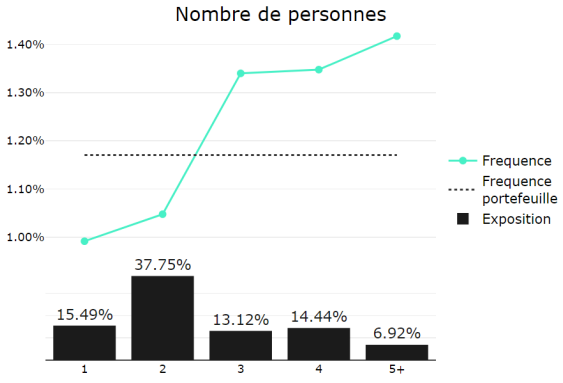


FIGURE 2.29 – Fréquence par nombre de personnes

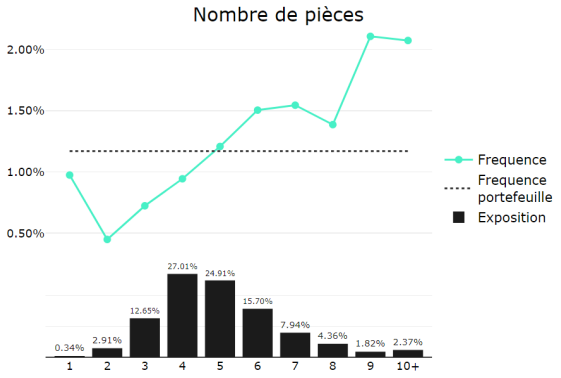


FIGURE 2.30 – Fréquence par nombre de pièces

Tant le nombre de personnes que de pièces impacte la fréquence des sinistres bien que ce soit dans une moindre mesure en ce qui concerne le nombre de personnes.

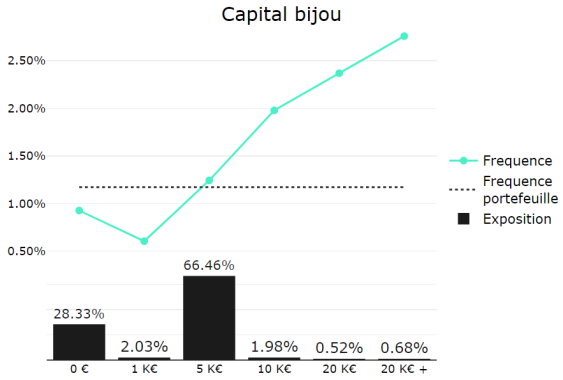


FIGURE 2.31 – Fréquence par montant bijou assuré

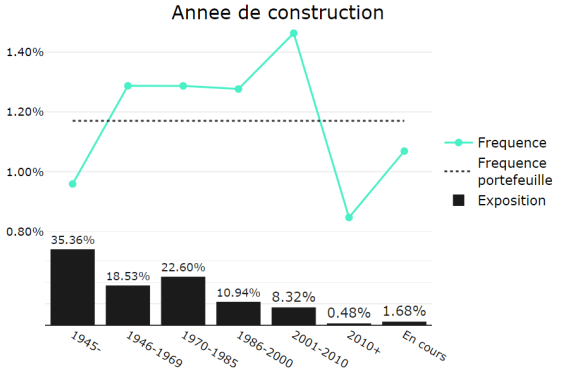


FIGURE 2.32 – Fréquence par année de construction

Concernant le capital bijou, il est difficile d'interpréter le graphique au vu de la faible exposition de nombreuses modalités, mais il semble apparaître que plus ce dernier augmente, plus la fréquence des vols est importante. Pour l'année de construction, en revanche, un impact « par morceaux » se dessine et sera à vérifier lors de la modélisation.

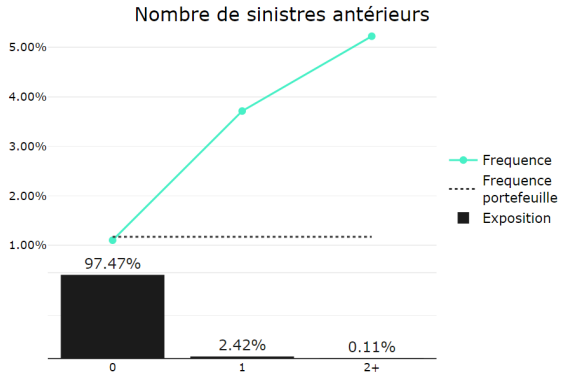


FIGURE 2.33 – Fréquence par sinistralité antérieure

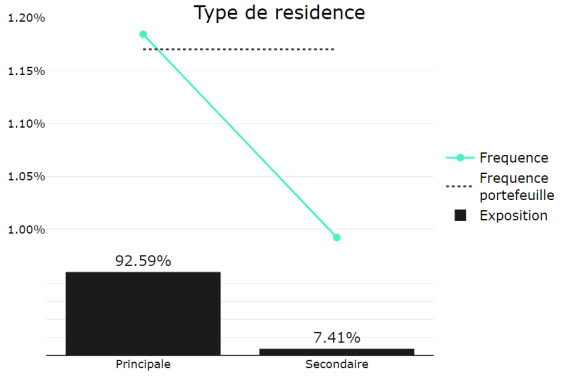


FIGURE 2.34 – Fréquence par type de résidence

Enfin, une maison ayant déjà fait l'objet d'un ou plusieurs cambriolages présente davantage de risques d'en subir un nouveau. Par ailleurs, une résidence principale est plus susceptible d'être cambriolée qu'une résidence secondaire.

Analyse univariée en coût moyen

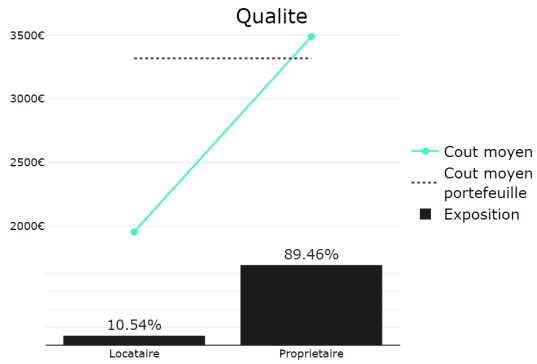


FIGURE 2.35 – Coût moyen par qualité

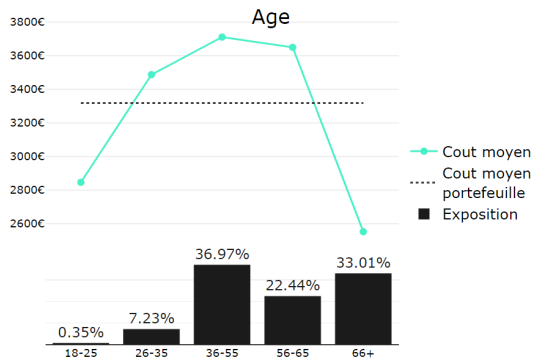


FIGURE 2.36 – Coût moyen par classe d'âge

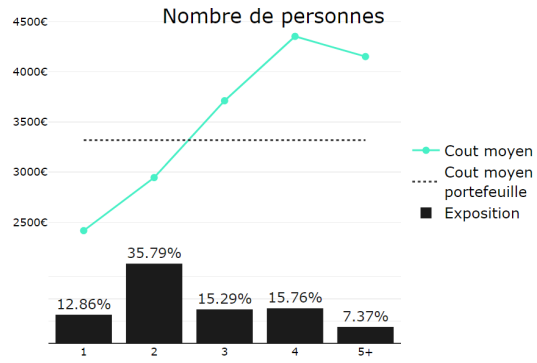


FIGURE 2.37 – Coût moyen par nombre de personnes

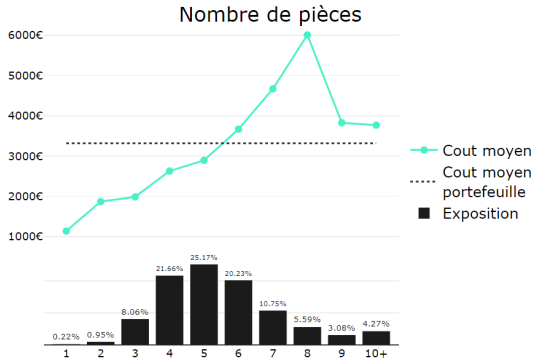


FIGURE 2.38 – Coût moyen par nombre de pièces

Que ce soit la qualité, l'âge, le nombre de personnes ou le nombre de pièces, toutes ces variables semblent une nouvelle fois impacter assez fortement le coût moyen de cette garantie. Les remarques, formulées précédemment, concernant l'exposition de certaines modalités restent par ailleurs toujours valables.

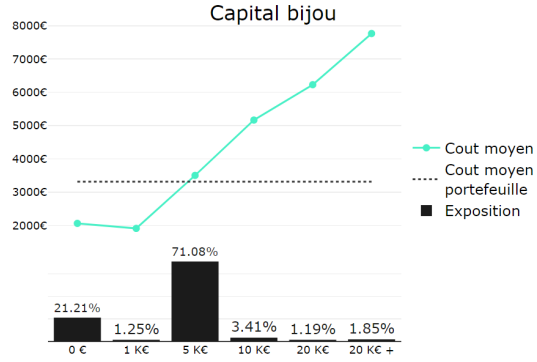


FIGURE 2.39 – Coût moyen par montant bijou assuré

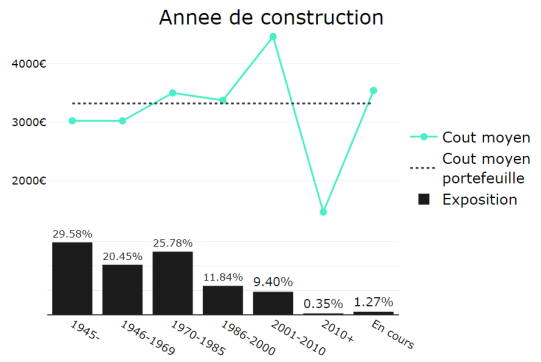


FIGURE 2.40 – Coût moyen par année de construction

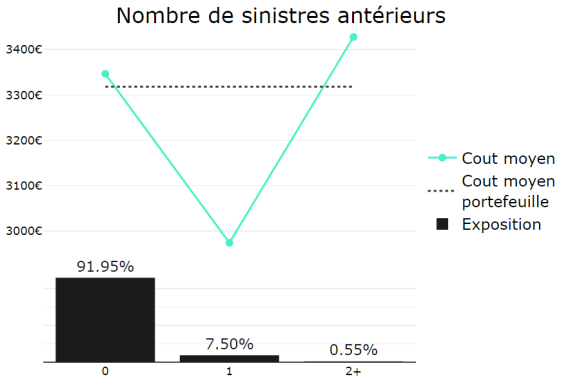


FIGURE 2.41 – Coût moyen par sinistralité antérieure

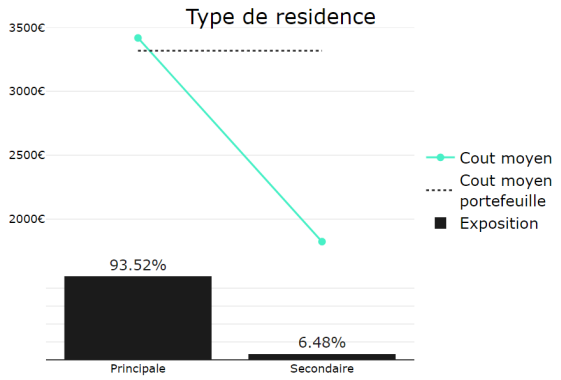


FIGURE 2.42 – Coût moyen par type de résidence

Enfin, les variables capital bijou et type de résidence apparaissent toutes les deux assez tarifaires concernant le coût moyen des vols. En revanche, les variables sinistralité antérieure et année de construction semblent, quant à elles, moins influer sur le coût moyen. Elles seront tout de même intégrées au modèle de coût.

Analyse bivariée

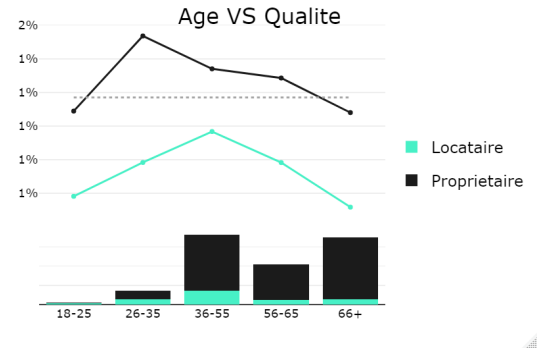


FIGURE 2.43 – Fréquence par classe d'âge et qualité

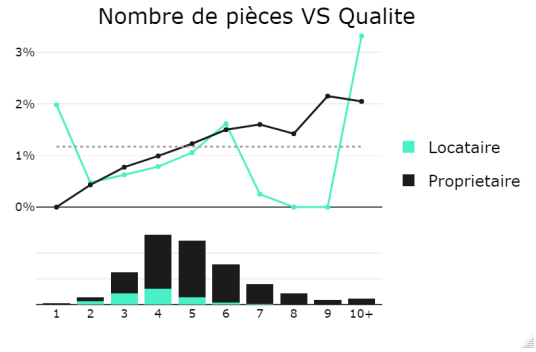


FIGURE 2.44 – Fréquence par nombre de pièces et qualité

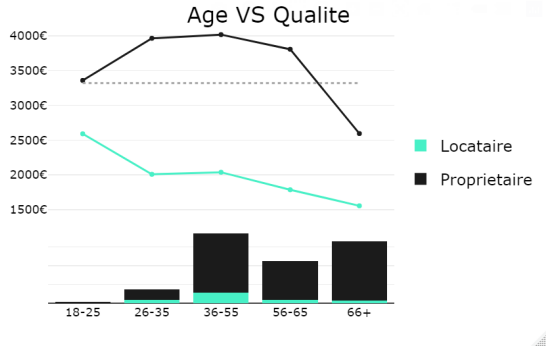


FIGURE 2.45 – Coût moyen par classe d'âge et qualité

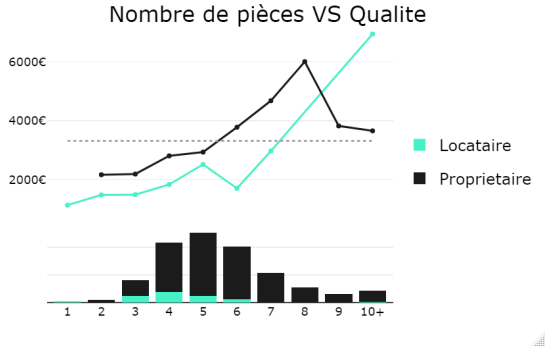


FIGURE 2.46 – Coût moyen par nombre de pièces et qualité

Outre les légères fluctuations et en considérant uniquement les couples de modalités suffisamment exposés, il ne semble pas y avoir d'impacts bivariés conséquents entre les variables qualité et âge et les variables nombre de pièces et qualité, que ce soit pour la fréquence ou le coût. Les modèles de fréquence et de coût pour la garantie vol seront donc par la suite sans interaction.

Zonier

Comme pour l'analyse de la garantie DDE, la variable zonier pour le vol a fait l'objet d'études précédentes et était donc déjà disponible. Cette variable a donc été reprise en l'état.

2.4 Les corrélations entre les différentes variables

Une fois les analyses univariées et bivariées effectuées, il est important de s'intéresser aux corrélations entre les variables, ces dernières pouvant fortement impacter les modèles GLM (variance erronée entraînant des problèmes de significativité des tests de Student, résultats instables, etc).

Afin de mesurer ces corrélations, différentes mesures de dépendance existent : le coefficient de corrélation linéaire de Pearson, les corrélations de rang (le ρ de Spearmann et le τ de Kendall) ou encore le test du χ^2 ou le V de Cramer. Chacune de ces mesures de dépendance présente des inconvénients et des avantages qui ne seront pas détaillés dans ce mémoire mais qui amènent à retenir le V de Cramer (qui présente l'avantage de prendre en compte des variables catégorielles) comme mesure de dépendance pour la suite du mémoire.

Avant de présenter cette mesure de dépendance, il est tout de même nécessaire de s'intéresser au test du χ^2 qui établit la dépendance ou l'indépendance de deux variables sans la quantifier.

2.4.1 Le test d'indépendance du χ^2

Plusieurs tests du χ^2 existent : le test du χ^2 d'adéquation, d'homogénéité ou encore d'indépendance. Au cas particulier, c'est le test du χ^2 d'indépendance qui est d'intérêt. Ce dernier test permet en effet de vérifier à partir d'un échantillon l'absence de lien statistique entre deux variables X et Y.

Plus formellement, soit X et Y deux variables aléatoires qualitatives à valeurs dans $x_1, ..., x_n$ pour X et $y_1, ..., y_p$ pour Y avec $n, p \geq 2$.

Disposant d'un échantillon de valeurs, il est possible de construire le tableau de contingence suivant :

	y_1	y_2	...	y_p	**Total**
x_1	$n_{1,1}$	$n_{1,2}$	...	$n_{1,p}$	$n_{1,.}$
x_2	$n_{2,1}$	$n_{2,2}$	...	$n_{2,p}$	$n_{2,.}$
...	...	...	...	...	...
x_n	$n_{n,1}$	$n_{n,2}$	...	$n_{n,p}$	$n_{n,.}$
Total	$n_{.,1}$	$n_{.,2}$	...	$n_{.,p}$	n

où $n_{i,j}$ est l'effectif observé des données avec $X = x_i$ et $Y = y_j$

Le test s'intéressant à l'indépendance des variables X et Y, il est possible de poser les hypothèses suivantes :

- H0 : X et Y sont indépendantes ;
- H1 : X et Y ne sont pas indépendantes.

Pour répondre à ce test, une statistique de test doit être introduite. Cette dernière se base sur l'écart entre les effectifs du tableau de contingence observé et les effectifs sous l'hypothèse d'indépendance des variables, c'est-à-dire les effectifs donnés par :

$$n^*_{i,j} = \frac{n_{i,.}n_{.,j}}{n}$$

Ceci amène à considérer la statistique de test suivante : $T = \sum_{i,j} \frac{n_{i,j}-n^*_{i,j}}{n^*_{i,j}}$.
Sous H0, il s'avère que $T \sim \chi^2_{(n-1)(p-1)}$. Ainsi, en comparant la statistique de test T au quantile de la loi du χ^2, $\chi^2_{1-\alpha,(n-1)(p-1)}$, il est possible de conclure quant à l'indépendance des deux variables. Si $T > \chi^2_{1-\alpha,(n-1)(p-1)}$, alors les deux variables ne sont pas indépendantes (H0 est rejetée) sinon, elles le sont (H0 est acceptée).

Ce test, bien que permettant de conclure quant à l'indépendance de deux variables, ne quantifie pas l'intensité de cette relation. C'est pourquoi le V de Cramer a été introduit.

2.4.2 Le V de Cramer

Le V de Cramer mesure donc l'intensité de la relation entre deux variables en utilisant le test du χ^2 précédemment présenté. Cette mesure est comprise entre 0 et 1. Plus la mesure est proche de 1, plus les variables considérées sont dépendantes entre elles et, à l'inverse, plus la mesure est proche de 0, moins il y a de lien.

De manière plus mathématique, le V de Cramer est donné par la formule suivante :

$$V = \sqrt{\frac{T}{n.min(n-1, p-1)}}$$

Cette formule correspond à la racine du rapport entre le χ^2 et le χ^2_{max} égal à l'effectif multiplié par le plus petit côté du tableau (nombre de lignes ou de colonnes) moins 1.

2.4.3 Le calcul des corrélations entre les variables

A partir de la matrice de corrélation présentée ci-après, il semblerait donc qu'il n'y ait que des corrélations de faible ou moyenne importance entre les variables. Aucune variable n'est donc éliminée par ce biais. Toutefois, pour la suite du mémoire, il est important de garder à l'esprit les corrélations d'importance modérée.

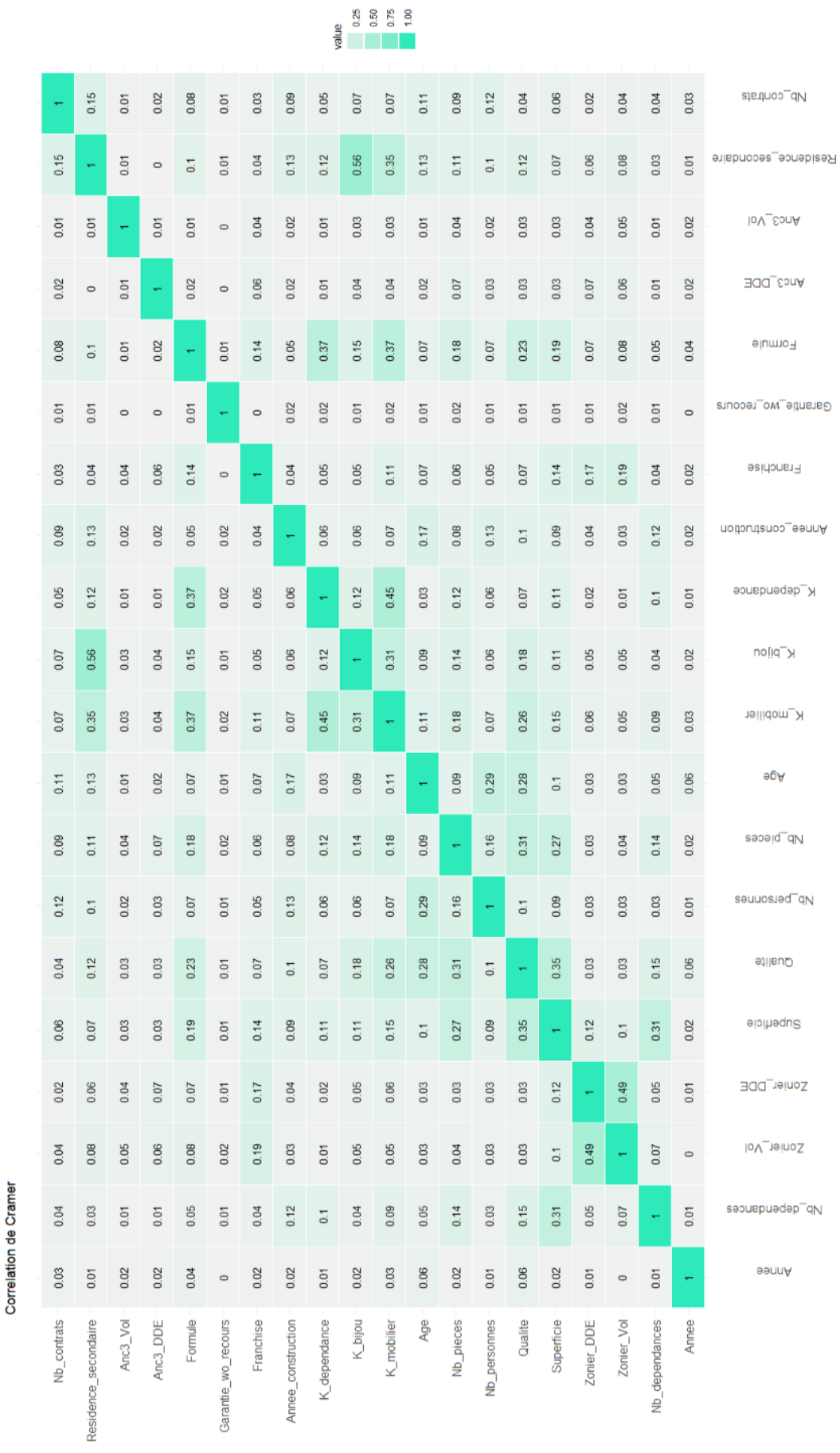


FIGURE 2.47 – Corrélation des variables

Chapitre 3

Théorie

L'objet de cette partie est d'exposer la théorie se cachant derrière chacune des méthodes utilisées dans le cadre de ce mémoire. Cette compréhension plus fine des méthodes permettra par la suite de prendre du recul quant à l'interprétation des résultats obtenus.

3.1 La reconnaissance d'images

Une grande partie des données à l'adresse présentées dans ce mémoire repose sur la reconnaissance d'images : calcul des surfaces des maisons, des piscines, du jardin, etc. Cette reconnaissance d'images peut être réalisée de deux manières différentes :

- soit des **règles** sont définies afin d'extraire des informations à partir d'une image à l'aide des formes et des nuances de couleurs présentes dans l'image. Cela peut être fait par exemple avec la librairie **OpenCV** de Python. Dans ce cas, l'ordinateur ne fait donc qu'appliquer ces règles ;
- soit l'ordinateur apprend lui-même des règles à partir d'une base d'apprentissage labellisée en passant par des méthodes plus complexes de ***deep learning*** par exemple.

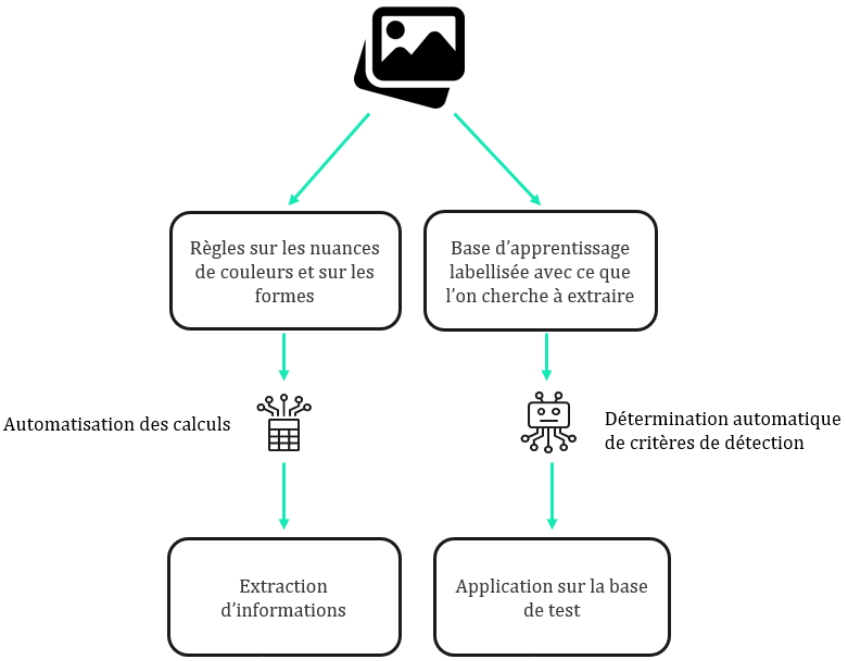


FIGURE 3.1 – Méthodes de reconnaissance d'images

Dans le cadre de cette étude, ces deux approches ont été envisagées. Pour des considérations de temps, de complexité et d'apport dans le mémoire, c'est la première méthode avec **OpenCV** qui a été retenue. Cependant, la deuxième méthode reste valable et d'intérêt pour de futures études.

Les opérations principales utilisées dans le cadre de cette reconnaissance d'images vont maintenant être expliquées de manière théorique. La reconnaissance d'images étant un domaine complet et vaste, l'objectif n'est pas de présenter ce domaine de manière exhaustive mais seulement de donner quelques éléments utiles à la compréhension du mémoire.

3.1.1 La représentation informatique d'une image

Avant toute chose, il est important de comprendre comment une image est représentée informatiquement.

Après acquisition de la donnée, une image peut être perçue comme un signal analogique, c'est-à-dire un signal mesuré en continu. Par exemple, pour chaque point de l'espace, une valeur du signal est associée. Cela peut par exemple être une couleur. Les supports physiques stockant ces informations étant limités en matière de capacité et de mémoire, il convient de numériser ce signal.

Cette numérisation comporte deux étapes majeures :

- une étape d'**échantillonnage** qui discrétise l'espace ;
- une étape de **quantification** qui discrétise, quant à elle, la valeur du signal. Cette étape permet pour chaque point de l'espace discrétisé d'associer une valeur du signal (par exemple une couleur). L'ensemble des valeurs possibles pour le signal devient donc fini.

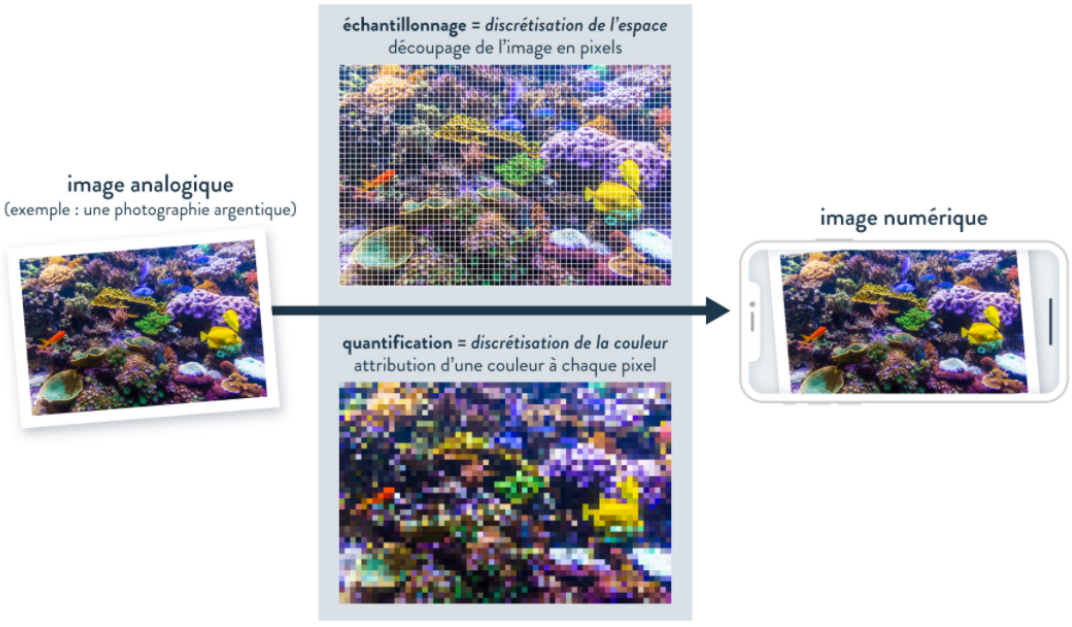


FIGURE 3.2 – Principales étapes d'une numérisation

Une étape d'encodage est ensuite effectuée pour traduire les valeurs du signal en binaire afin que l'ordinateur puisse lire les données. Par simplification, cette opération ne sera pas détaillée ici.

A elles seules, ces étapes de numérisation revêtent une complexité étonnante et peuvent devenir très théoriques. De manière simplifiée, il faut retenir que cette phase de numérisation peut être assimilée aux intégrales de Riemann qui discrétisent une fonction continue afin d'en mesurer l'aire sous la courbe. Par analogie, ici, l'image est numérisée afin de pouvoir la stocker.

En revenant aux images, avec un échantillonnage de bonne qualité (et sous certaines conditions), il est donc possible de représenter la quasi-totalité des informations présentes dans le signal analogique de base. C'est cet échantillonnage qui déterminera notamment la résolution de l'image. Ainsi, plus cet échantillonnage sera précis, plus la résolution sera grande mais la taille de l'image sera, en contrepartie, beaucoup plus volumineuse. Un compromis est donc à trouver en fonction des objectifs à atteindre.

Lors de l'échantillonnage, l'image analogique est donc convertie en image numérique.

Avec l'ensemble des informations présentées ci-dessus, il paraît évident qu'une image numérique soit représentée par une matrice. Plus particulièrement, il s'agit d'une matrice de pixel où chaque pixel contient une information. Les pixels sont représentés par leur numéro de ligne et de colonne (en commençant à 0 par convention).

FIGURE 3.3 – Image numérique

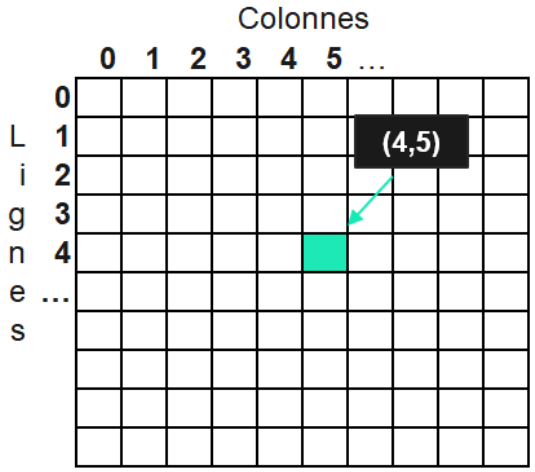


FIGURE 3.4 – Matrice de pixel

A chaque pixel est également associée une information sur la valeur du signal, usuellement la couleur. La couleur peut être codée selon différents référentiels :

- **Monochrome** : noir (0) ou blanc (1) ;
- **Niveaux de gris** : codage entre 0 (noir) et 255 (blanc) ;
- **Couleur** : pour chaque pixel, 3 informations sont présentes :
  - Si le système de couleur RGB est utilisé, la quantité de rouge, de vert, puis de bleu est renseignée (par un chiffre entre 0 et 255). La couleur finale étant obtenue par synthèse additive des trois couleurs primaires ;
  - Si le système BGR est utilisé, l'ordre des couleurs change : le bleu est d'abord renseigné puis vient le vert et ensuite le rouge ;
  - D'autres systèmes plus complexes comme le système HSV (Hue, Saturation, Lightness) appelé aussi TSL en français (Teinte, Saturation, Luminosité) peuvent être utilisés. Le système de représentation HSV est d'ailleurs privilégié lorsqu'il est souhaité isoler des nuances de couleurs. Dans ce cas, il s'agit alors de renseigner la teinte, codée suivant l'angle lui correspondant dans le cercle des couleurs, la saturation et la luminosité de la couleur (en pourcentage).

Une fois la numérisation effectuée, vient ensuite le traitement de l'image.

3.1.2 Quelques opérations utiles

Une fois ces opérations effectuées, vient le traitement de l'image. Le traitement d'image permet tant d'améliorer la qualité de l'image que d'en extraire des informations. De nombreuses opérations étant possibles, seuls les traitements ayant été utilisés dans le cadre du mémoire vont être présentés.

Détection de couleur

La détection d'image effectuée dans le cadre de ce mémoire repose essentiellement sur les nuances de couleurs. Il a donc fallu, en amont, apprendre à détecter une nuance spécifique.

Pour ce faire, il est nécessaire, avant toute chose, de définir un intervalle de couleurs que l'on souhaite conserver par la suite. Une fois cet intervalle défini, la couleur de chaque pixel va être comparée à l'intervalle. Si le pixel contient une nuance de couleur présente dans l'intervalle, ce dernier est conservé et est codé en blanc dans une image à part (appelée le masque), sinon ce dernier est codé en noir. A la fin de ce traitement, un masque de l'image est donc obtenu à partir de l'intervalle renseigné. Ce masque isole l'intervalle de couleurs renseigné.

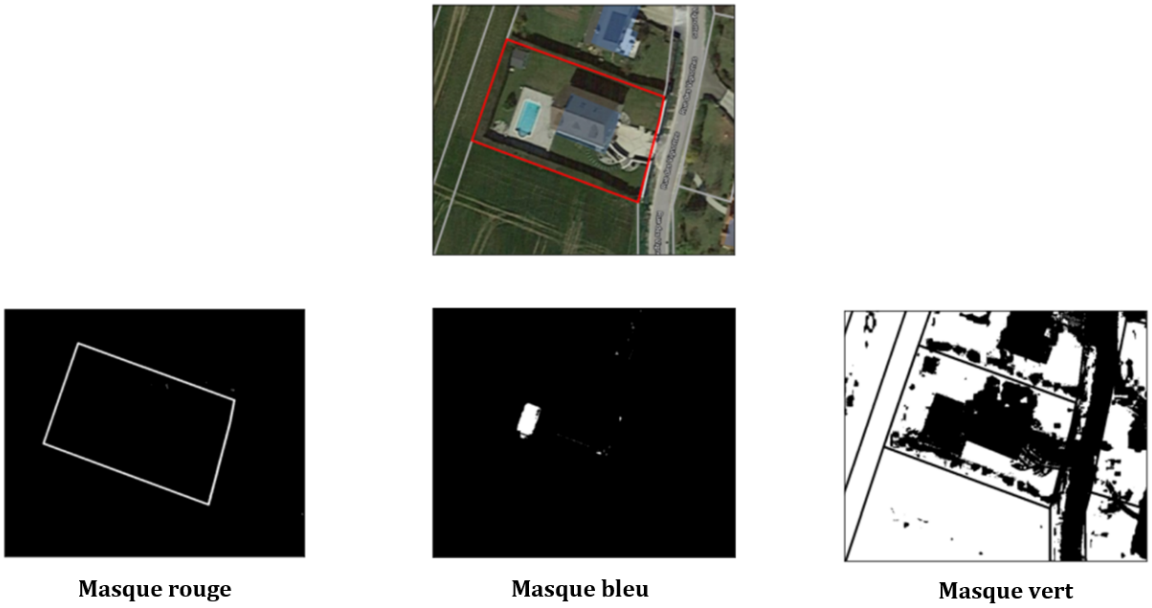


FIGURE 3.5 – Détection de couleur

La fonction **inRange** d'**OpenCV** de Python permet de réaliser ce traitement. Il est à noter cependant que cette fonction utilise le référentiel de couleurs HSV dans lequel il est plus simple de séparer des couleurs que dans l'espace RGB. L'intervalle de couleurs à conserver doit donc en conséquence être défini dans cet espace.

Une fois le masque créé, il est donc possible, soit de vouloir conserver uniquement les éléments à l'intérieur du masque (pour le masque rouge de la figure ci-contre par exemple) ou encore de nettoyer les formes détectées (pour le masque bleu de la figure ci-contre). Le détail de chacun de ces traitements est présenté ci-dessous.

Utilisation du masque de couleurs pour conserver l'intérieur

Afin de conserver uniquement l'intérieur d'une forme géométrique délimitée par un masque, deux opérations successives sont nécessaires.

Tout d'abord, le masque est transformé en remplissant l'intérieur du contour fermé en blanc.

Ensuite, les positions des pixels en blanc du masque sont récupérées pour modifier leurs couleurs avec ceux de l'image d'origine.

FIGURE 3.6 – Conservation de l'intérieur du masque

Nettoyage des formes détectées par le masque de couleurs

Le traitement explicité ici reposant sur des nuances de couleurs, il arrive parfois que les formes récupérées dans le masque ne soient pas bien délimitées ou présentent des résidus. Dans ce cas de figure, un nettoyage s'impose.

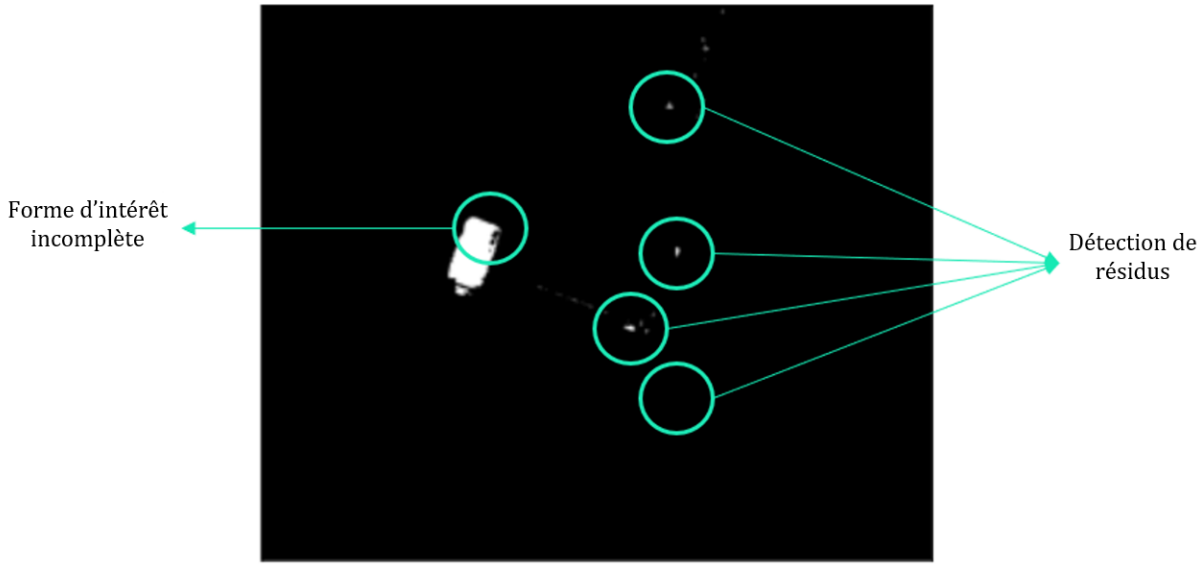


FIGURE 3.7 – Exemples de problèmes rencontrés lors de la détection de couleur

Pour ce faire, les contours de chaque forme détectée dans le masque vont être isolés avec la fonction **findContours**. Cet algorithme repose sur des notions complexes qui ne seront pas détaillées dans le cadre du mémoire mais l'idée générale derrière cette fonction est de détecter les contours par changement de nuances de couleurs.

Ensuite, pour chaque contour détecté, la surface du contour est calculée pour être comparée à un seuil. Si le contour est trop petit, ce dernier est effacé en le noircissant. Dans le cas contraire, il est conservé. Cette étape permet donc d'éliminer les résidus potentiellement présents.

Les contours conservés représentent donc maintenant des formes d'intérêt. Cependant, ces contours ne sont pas toujours bien remplis et une opération de fermeture des contours doit être

effectuée. La technique derrière cette fonction repose sur des méthodes de gradient d'image qui ne seront pas détaillées dans le cadre du mémoire.

FIGURE 3.8 – Nettoyage des formes détectées

Calcul de la surface

A chaque étape du traitement, il est également possible de calculer une surface pour la zone d'intérêt grâce à la fonction **contourArea**. Cette fonction repose sur le théorème de Green :

Soit C une courbe plane simple, positivement orientée et C^1 par morceaux, D le compact du plan délimité par C et $P\,\mathrm{d}x + Q\,\mathrm{d}y$ une 1-forme différentielle sur $\mathbb{R}^2$. Si P et Q ont des dérivées partielles continues sur une région ouverte incluant D, alors :

$$\int_C P\,\mathrm{d}x + Q\,\mathrm{d}y = \iint_D \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right) \mathrm{d}x\,\mathrm{d}y$$

Une simple conversion permet ensuite de passer à l'échelle réelle.

L'ensemble de ces méthodes a donc été utilisé pour le calcul des surfaces des maisons, des piscines, etc., surfaces qui seront nécessaires pour la suite.

Il est à noter que les données récupérées à l'aide de la reconnaissance d'images seront listées ultérieurement.

3.2 Les méthodes d'apprentissage statistique

Un autre sujet important à aborder au sein de cette partie est l'apprentissage statistique ou *machine learning*. En effet, il est très courant, pour les assureurs, d'utiliser des méthodes de *machine learning* lors du processus de tarification. Ce rapprochement entre méthodes d'apprentissage statistique et tarification peut être effectué en présentant les modèles collectifs.

3.2.1 Les modèles collectifs

Lors d'une tarification, l'assureur cherche à prédire le coût de ses potentiels sinistres sur une certaine période. Pour ce faire, la modélisation la plus courante en assurance non-vie consiste à

considérer un portefeuille dont les risques sont homogènes et indépendants. Dans ces conditions, la charge totale des sinistres du portefeuille, notée S est donnée par la formule suivante :

$$S = \sum_{i=1}^{N} Z_i$$

où

- N est une variable aléatoire représentant le nombre de sinistres du portefeuille considéré durant la période fixée ;
- $(Z_i)_{i\geq 1}$ est une suite de variables aléatoires représentant les indemnisations successives de chaque sinistre.

La charge totale des sinistres est donc perçue dans ce cas comme la somme du coût des sinistres et ce, toutes polices confondues. Il n'y a pas de distinction de police comme c'est le cas dans les modèles individuels.

En supposant que les coûts des sinistres, $(Z_i)_{i\geq 1}$, sont indépendants et identiquement distribués (i.i.d.) et qu'ils sont indépendants du nombre de sinistres, la charge totale moyenne des sinistres s'écrit alors :

$$\mathbb{E}[S] = \mathbb{E}[N] \times \mathbb{E}[Z]$$

lorsque les moments d'ordre 1 existent.

L'objectif de l'approche fréquence-coût, usuellement utilisée pour tarifier des produits non-vie, est alors d'approcher cette prime pure par des variables dites de tarification notées X.

Deux étapes sont donc nécessaires : la prédiction de la fréquence moyenne des sinistres sur la période considérée et celle du coût moyen d'un sinistre. Ces deux prédictions reposent très souvent sur des algorithmes de *machine learning* comme par exemple les modèles linéaires généralisés (ou GLM).

3.2.2 Quelques généralités sur l'apprentissage statistique

L'apprentissage statistique (ou *machine learning*) peut être défini comme l'apprentissage et la compréhension de comportements à partir d'exemples. Ce domaine permet donc de créer des modèles dans le but :

- d'**explorer**, de **représenter**, de **décrire** des variables et les liens entre ces dernières ;
- d'**expliquer** un phénomène par le biais d'un modèle ;
- ou encore de **prédire** un phénomène.

Il est également possible de diviser l'apprentissage statistique en deux grands sous-domaines : l'**apprentissage statistique supervisé** et l'**apprentissage statistique non supervisé**.

Si l'objectif, en apprentissage statistique supervisé, est plutôt de comprendre ou de prédire un phénomène, en apprentissage statistique non supervisé, le but principal est de trouver une certaine typologie dans les observations et de créer des classes homogènes.

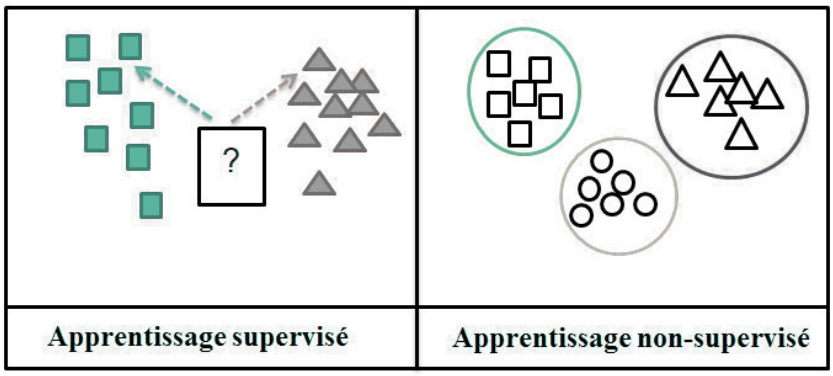


FIGURE 3.9 – Apprentissage supervisé versus apprentissage non supervisé

Ce mémoire nécessitant de se pencher sur l'explication d'un phénomène (la fréquence/le coût), seul l'apprentissage statistique supervisé sera développé.

3.2.3 Les principes majeurs de l'apprentissage statistique supervisé

Objectif

Au sein de l'apprentissage statistique supervisé, une variable aléatoire à valeurs dans $\mathcal{Y}$ et notée Y est à **expliquer**. Cette variable aléatoire peut être tant quantitative (problèmes de régression) que qualitative (problèmes de classification). Un ensemble de variables explicatives à valeurs dans $\mathcal{X}$ et noté $X = (X^1, X^2, .., X^p)$ avec p le nombre de variables explicatives, est également mis à disposition pour expliquer Y.

Il est ensuite supposé qu'il existe un lien réel entre X et Y. L'un des principaux but de l'apprentissage statistique supervisé est alors d'approcher ce lien. Plus formellement, il est supposé que :

$$Y = f(X)$$

où f est une fonction représentant le lien réel entre X et Y. L'apprentissage statistique supervisé cherche donc une fonction $\tilde{f}$ telle que :

$$Y = \tilde{f}(X) + \epsilon$$

où ϵ est l'erreur du modèle que l'on souhaite raisonnable.

Outre l'explication du phénomène considéré, d'autres éléments majeurs et tout aussi importants sont également recherchés lors de la création d'un modèle de *machine learning*.

Ainsi, il est possible de s'intéresser à l'**interprétation du rôle de chaque variable** dans le modèle. Cela est d'autant plus important en actuariat que l'assureur doit pouvoir être à même de justifier son tarif de manière simple. Pour les modèles de *machine learning* intrinsèquement interprétables comme les GLM ou les arbres de décision, le rôle de chaque variable est présenté de manière assez explicite en observant les paramètres du modèle. Cependant, pour les modèles plus complexes de type « boîte noire » (les modèles XGBoost par exemple), tout l'enjeu est alors de trouver des méthodes d'interprétabilité de ces modèles. C'est d'ailleurs cette explicabilité des modèles qui permettra à de nombreux assureurs d'utiliser plus couramment dans leur tarification des modèles de type « boîte noire ». Il est cependant à noter que ce choix d'utiliser des modèles complexes doit être fait de manière réfléchie. Les utiliser lorsque leur gain de performance est assez faible par rapport a un GLM classique ne paraît pas pertinent, une augmentation de la performance entraînant une augmentation de la complexité. Un compromis doit donc être trouvé.

Enfin, un modèle d'apprentissage statistique supervisé, en plus de devoir bien expliquer le phénomène, doit également pouvoir s'adapter à de nouvelles données et fournir des **prédictions** de bonne qualité. Pour ce faire, il est primordial de faire attention à un phénomène : le surapprentissage.

Le surapprentissage et le compromis biais/variance

Le surapprentissage apparaît lorsque le modèle construit explique parfaitement les données utilisées pour son calibrage mais est incapable de fournir des résultats pertinents sur de nouvelles données. Ce phénomène apparaît très souvent lorsque le modèle considéré est trop complexe.

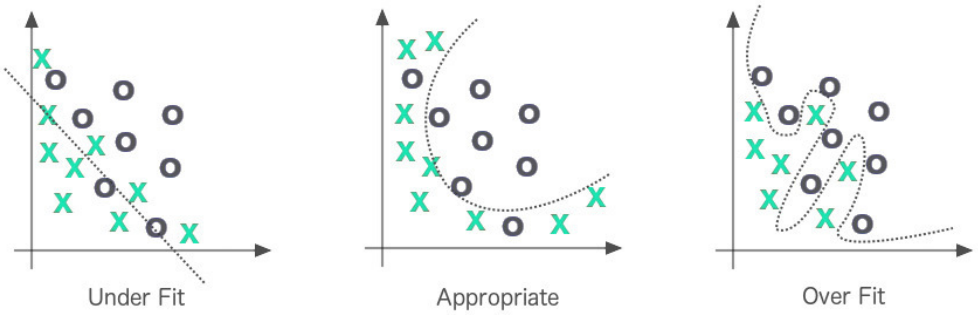


FIGURE 3.10 – Surapprentissage et sous-apprentissage

Une autre vision de ce phénomène peut s'illustrer avec le compromis biais-variance. En effet, lorsque le modèle s'ajuste parfaitement aux données d'apprentissage, une forte variance et un faible biais sont généralement observés : c'est le surapprentissage. L'erreur sur l'échantillon d'apprentissage sera donc très petite mais le modèle aura du mal à généraliser et l'erreur de validation sera élevée. A l'inverse, si le modèle est trop simple, la variance sera très faible mais le biais important : c'est le sous-apprentissage. Tant l'erreur d'apprentissage que de validation seront alors de piètre qualité. Idéalement, pour éviter le sous-apprentissage ou le surapprentissage, il faudrait donc un modèle ayant un biais et une variance raisonnable. La variance et le biais évoluant en sens inverse, il est courant de parler de compromis biais-variance.

Le graphe ci-contre résume l'ensemble de ces propos :

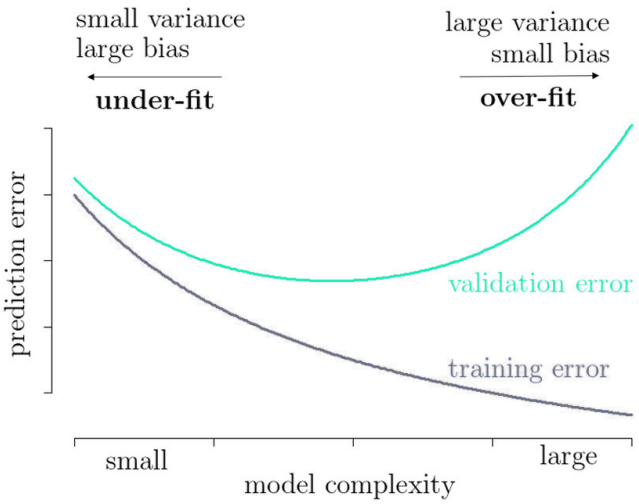


FIGURE 3.11 – Compromis biais-variance

Afin de surveiller et d'éviter ces phénomènes de surapprentissage et sous-apprentissage, il est important de s'intéresser à la complexité du modèle et à la qualité de la prédiction en fonction de cette complexité. Pour mesurer cette dernière, il est nécessaire de définir l'erreur de généralisation appelée parfois mesure de risque.

Mesure de risque ou erreur de généralisation

L'erreur de généralisation est un indicateur permettant de mesurer la qualité du modèle construit. Pour construire cet indicateur, une fonction de perte notée ℓ doit être définie en amont :

$$\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+$$

avec $\mathcal{Y}$ l'espace de définition de la variable aléatoire Y et ℓ tel que

$$\begin{cases} \ell(y, y) & = 0 \\ \ell(y, y') & > 0 \quad \text{pour} \quad y \neq y' \end{cases}$$

Cette fonction de perte peut être choisie selon le type de la variable à expliquer. Ainsi, si la variable est binaire ($\mathcal{Y} = \{-1, 1\}$), la fonction de perte utilisée usuellement est la suivante :

$$\begin{aligned} \ell : \{-1, 1\} \times \{-1, 1\} &\to \mathbb{R}^+ \\ (y, y') &\mapsto \mathbb{I}_{\{y \neq y'\}} \end{aligned}$$

Autrement dit, la fonction de perte mesure le taux de mauvais classement. Il est également possible d'utiliser une approche basée sur une fonction de score (Courbe ROC).

Si la variable à prédire est quantitative, ce seront alors des fonctions de perte comme l'erreur quadratique qui seront utilisées :

$$\begin{aligned} \ell : \mathbb{R} \times \mathbb{R} &\to \mathbb{R}^+ \\ (y, y') &\mapsto (y - y')^2 \end{aligned}$$

Une fois la fonction de perte ℓ choisie, l'erreur de généralisation associée à la fonction de prédiction $\tilde{f}$, notée $\mathcal{R}(\tilde{f})$, peut être définie par :

$$\mathcal{R}(\tilde{f}) = \mathbb{E}[\ell(Y, \tilde{f}(X))]$$

L'erreur de généralisation correspond donc au comportement moyen de la fonction de perte choisie. Dans le cas d'une régression, pour la fonction de perte d'erreur quadratique, l'erreur de généralisation correspond donc à l'erreur quadratique moyenne (MSE) et est donc liée à la racine de l'erreur quadratique moyenne (RMSE). Dans le cadre de ce mémoire, c'est la RMSE qui sera utilisée comme erreur de généralisation pour comparer les différents modèles entre eux.

Le but de l'apprentissage statistique supervisé est alors de trouver une fonction $\tilde{f}$ approchant la fonction f de manière à pouvoir expliquer au mieux le phénomène tout en minimisant cette erreur de généralisation pour avoir une bonne qualité de prédiction :

$$\text{argmin}_{\tilde{f} \in \mathcal{F}} \mathcal{R}(\tilde{f})$$

La fonction vérifiant cette équation est appelée fonction de prédiction optimale.

Le lien entre erreur de généralisation et qualité de prédiction peut ensuite être fait de manière aisée.

Qualité de prédiction

La qualité de prédiction peut ainsi être vue comme une estimation de l'erreur de généralisation. Il faut cependant faire attention à estimer de manière correcte cette grandeur.

En effet, l'approche simpliste consisterait à estimer l'erreur de généralisation sur les données ayant servi à entraîner le modèle. Ce faisant, une erreur serait commise, l'estimateur serait fortement biaisé et le risque serait sous-estimé.

Des méthodes plus adaptées ont donc dû être développées afin d'estimer correctement cette grandeur. Parmi elles :

- l'**estimation à l'aide d'échantillons indépendants** : décomposition des données en une base d'apprentissage pour entraîner le modèle et une base de test pour estimer l'erreur de généralisation ;
- l'**estimation par pénalisation** où un terme correctif est ajouté à l'estimation du risque (par exemple le Cp de Mallows, l'AIC ou encore le BIC qui seront détaillés dans la suite du mémoire) ;
- la **validation croisée** utile lorsque peu de données sont disponibles pour construire le modèle ;
- l'**estimation bootstrap**.

Dans le cadre de ce mémoire, l'estimation par échantillons indépendants sera privilégiée.

L'estimation de la qualité de prédiction permettra ensuite de comparer les différentes fonctions candidates entre elles (et donc les différents modèles entre eux) .

Et en pratique ...

En pratique, pour expliquer ou prédire un phénomène, plusieurs étapes sont nécessaires. De manière simplifiée :

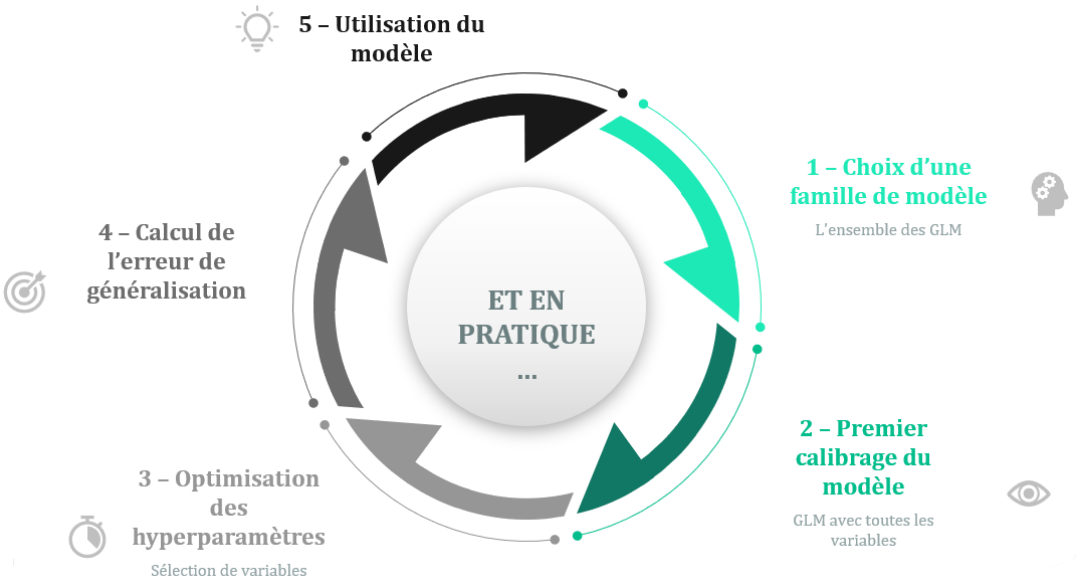


FIGURE 3.12 – Étapes principales d'un problème d'apprentissage statistique supervisé

1. Une hypothèse est faite quant à la **famille de modèles** à utiliser pour expliquer le phénomène d'intérêt (par exemple, il est possible de choisir les modèles GLM). Cela revient tout simplement à se limiter à une certaine catégorie de fonctions $\tilde{f}$ pour approcher f.
2. Au sein de cette famille de fonctions $\tilde{f}$, un **premier calibrage** du modèle peut être fait. Ce calibrage n'est généralement pas optimal et peut donner lieu à des phénomènes de surapprentissage.
3. Pour **résoudre le problème de surapprentissage**, il est nécessaire de trouver au sein de cette famille de fonctions $\tilde{f}$ (i.e. parmi tous les GLM possibles), la fonction (le GLM) permettant de minimiser l'erreur de généralisation. La plupart du temps, cela est effectué en calibrant les hyperparamètres du modèle de manière à minimiser l'erreur de prédiction par validation croisée (par exemple en sélectionnant les variables dans le cadre d'un GLM ou en optimisant le nombre de voisins dans le cas des k plus proches voisins). Le modèle devrait ensuite, en tout état de cause, ne plus comporter de surapprentissage et présenter un compromis biais/variance satisfaisant.
4. Une **erreur de généralisation** est ensuite calculée sur un échantillon de test (non utilisé pour l'apprentissage).
5. Le modèle, s'il est performant, peut ensuite être utilisé en **prédiction**.

D'autres familles de modèles peuvent ensuite être utilisées : il faut alors repartir de l'étape 1 et c'est l'erreur de généralisation qui permettra de comparer les différents modèles entre eux.

Dans le cadre de ce mémoire, deux familles de modèles seront plus particulièrement utilisées : les GLM et les Random Forest.

3.2.4 Le modèle linéaire généralisé

Aujourd'hui, les modèles linéaires généralisés font partie des méthodes de *machine learning* les plus utilisées en tarification. En effet, outre la performance du modèle, un tarif se doit avant tout d'être aisément explicable au client. Le modèle derrière la tarification doit donc être facilement compréhensible par tous les utilisateurs non avertis intervenant dans le processus de tarification. Les GLM faisant partie des modèles les plus facilement interprétables tout en étant performants, ces derniers sont donc couramment utilisés en assurance non-vie.

La théorie derrière les GLM va donc faire l'objet d'une présentation détaillée. Cependant, avant d'aborder le cas plus général des GLM, il semble plus opportun d'aborder la régression linéaire, un cas particulier de cette famille de modèles.

Pour la suite, les notations suivantes vont être adoptées :

- n représente le **nombre d'observations** à disposition ;
- $p-1$ représente le **nombre de variables explicatives** à disposition ;
- Y est un vecteur à valeurs dans $\mathcal{Y}$ composé des observations de la **variable à expliquer** : $Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$;
- X est une matrice à valeurs dans $\mathcal{X}$ constituée des observations des **variables explicatives** : $X = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1,p-1} \\ 1 & x_{21} & x_{22} & \dots & x_{2,p-1} \\ \vdots & & & & \\ 1 & x_{n1} & x_{n2} & \dots & x_{n,p-1} \end{pmatrix}$ où x_{ij} est l'observation de la variable explicative j pour l'individu i.

La régression linéaire multiple

Avant toute chose, il est à noter que le mot « linéaire » dans le cadre de la régression linéaire multiple fait référence à la linéarité des paramètres β_j, $j = 0 \dots p-1$ et non des variables explicatives.

Dans le cadre de la régression linéaire multiple, il est supposé que Y peut s'expliquer comme une combinaison linéaire des $p-1$ variables explicatives X_1,X_2,..., X_{p-1} selon la formule suivante :

$$Y = X\beta + \epsilon$$

avec :

- Y et X définis auparavant ;
- $\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{pmatrix} \in \mathbb{R}^p$, les **paramètres constants** du modèle à estimer ;
- $\epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix} \in \mathbb{R}^n$, les **erreurs** du modèle représentant l'écart entre la valeur observée et la valeur estimée par combinaison des variables explicatives. Les ϵ sont tels que ϵ_1,...,$\epsilon_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$, c'est-à-dire que les erreurs sont centrées, de même variance σ^2, non corrélées entre elles et suivent une loi normale.

Il est également supposé que la matrice X est de plein rang : $rg(X) = p$. Ce qui signifie qu'il ne doit pas y avoir de variable explicative qui soit une combinaison linéaire des autres.

L'objectif est alors d'estimer les paramètres β et σ du modèle de manière à modéliser correctement le phénomène observé. Une fois ces paramètres estimés et après validation du modèle (en vérifiant entre autres que les hypothèses sont bien vérifiées) et réduction du phénomène de surapprentissage (par sélection de variables par exemple), il est possible de prédire, pour de nouvelles observations, les valeurs de la variable à expliquer.

Bien que simple et intuitif, ce modèle fait cependant de fortes hypothèses qui impliquent que la variable à expliquer $Y \sim \mathcal{N}(X\beta, \sigma^2\mathcal{I}_n)$ et prend donc ses valeurs dans R (i.e. $\mathcal{Y} = \mathbb{R}$).

Cette contrainte n'est malheureusement pas envisageable dans de nombreuses applications actuarielles et notamment en tarification, où la variable à expliquer représente très souvent un nombre non négatif (par exemple le nombre de sinistres, le coût moyen d'un sinistre ou encore la prime pure d'un assuré). De plus, la relation entre les variables explicatives et la variable à expliquer est bien souvent non-linéaire. Ces deux critiques du modèle linéaire justifient la mise en place d'un modèle linéaire généralisé [18].

Les modèles linéaires généralisés

Trois éléments vont intervenir dans la définition d'un tel modèle :

- **la distribution supposée** de $Y|X$. Cette distribution doit appartenir à la famille exponentielle [1]. ;

1. Famille exponentielle : une variable aléatoire X dont la loi dépend des paramètres α et λ appartient à la famille exponentielle, si et seulement si, sa densité est de la forme $f(x;\alpha;\lambda) = exp(\frac{x\alpha - b\alpha}{a(\lambda)} + c(x,\lambda))$ avec a, b et c des fonctions, α le paramètre naturel et λ le paramètre de dispersion.

- **un prédicteur linéaire** $X\beta$;
- **une fonction de lien** g, inversible, permettant de relier le prédicteur linéaire à valeurs dans $\mathbb{R}$ et la moyenne de la distribution supposée de Y|X : $g(\mathbb{E}[Y|X]) = X\beta$.

En fonction des distributions supposées, différents modèles sont obtenus :

Modèle	Loi	Support	Fonction lien	Expression du modèle
Linéaire	Gaussien	$\mathbb{R}$	Identité : $g(x) = x$	$\mathbb{E}[Y\|X] = X\beta$
Logistique	Bernouilli	$\{0;1\}$	Logit : $g(x) = ln(\frac{x}{1-x})$	$\mathbb{E}[Y\|X] = (1 + exp(-(X\beta)))^{-1}$
Gamma	Gamma	$\mathbb{R}^+$	Inverse : $g(x) = \frac{1}{x}$	$\mathbb{E}[Y\|X] = (X\beta)^{-1}$
Log-Gamma	Gamma	$\mathbb{R}^+$	Ln : $g(x) = ln(x)$	$\mathbb{E}[Y\|X] = exp(X\beta)$
Log-Poisson	Poisson	$\mathbb{N}$	Ln : $g(x) = ln(x)$	$\mathbb{E}[Y\|X] = exp(X\beta)$

TABLE 3.1 – Modèles linéaires généralisés

A la différence de la régression linéaire multiple, les paramètres sont estimés par maximum de vraisemblance.

Dans le cadre de ce mémoire, ce seront respectivement les modèles Log-Poisson et Log-Gamma qui seront utilisés pour modéliser la fréquence et le coût. Il est à noter cependant que d'autres lois auraient aussi pu être utilisées pour la modélisation de ces grandeurs si elles s'étaient avérées plus adaptées.

Complexité du modèle

Que ce soit pour un modèle de régression linéaire multiple ou un modèle de régression linéaire généralisé, la complexité du modèle dépend du nombre de variables utilisées dans le modèle. Pour éviter les phénomènes de surapprentissage ou de sous-apprentissage, il est donc recommandé de procéder à une sélection de variables. Cette sélection de variables permettra une meilleure généralisation à de nouvelles données (et donc une meilleure performance) et contribuera à l'obtention d'un modèle parcimonieux, élément important en actuariat.

Pour procéder à cette sélection de variables, différentes méthodes existent. Parmi elles :

- **la méthode exhaustive** qui parcourt les 2^{p-1} modèles possibles ($p-1$ étant le nombre de variables à disposition) et sélectionne le meilleur modèle selon un critère défini en amont ;
- **la méthode backward** qui part du modèle complet (avec toutes les variables) et élimine de manière itérative la variable dont l'élimination permet d'améliorer au mieux le modèle selon un certain critère défini en amont. L'algorithme s'arrête lorsque le modèle ne peut plus être amélioré par élimination de variables ;
- **la méthode forward** qui part du modèle nul (avec juste une constante) et ajoute de manière itérative la variable dont l'ajout permet d'améliorer au mieux le modèle selon un critère défini en amont. L'algorithme s'arrête lorsqu'aucune variable ne permet d'améliorer le modèle par ajout ;
- **la méthode stepwise** qui mélange les méthodes backward et forward. Cette méthode part du modèle nul et ajoute/élimine de manière itérative la variable améliorant significativement le modèle. L'algorithme s'arrête lorsqu'aucune variable ne peut plus améliorer le modèle par ajout ou élimination.

Dans le cadre de ce mémoire, la méthode exhaustive n'est bien sûr pas envisageable au vu du nombre de variables à disposition. Les méthodes backward et forward présentent, quant à elles, toutes les deux leurs avantages. Ainsi, la méthode forward est plus rapide que la méthode

backward, les modèles ajustés étant moins complexes dans la méthode forward. Cependant, la méthode backward permet de capter les interactions en partant du modèle complet, ce qu'omet la méthode forward. Un juste compromis entre ces deux méthodes se dessine avec la méthode stepwise qui sera utilisée dans la suite du mémoire.

Concernant le critère permettant de départager les variables à ajouter/éliminer, il est courant d'utiliser l'AIC ou le BIC, deux grandeurs estimant le risque par pénalisation :

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

où k est le nombre de paramètres à estimer du modèle et L est le maximum de la fonction de vraisemblance du modèle.

$$BIC = -2ln(L) + kln(n)$$

où L est la vraisemblance du modèle estimée, n le nombre d'observations dans l'échantillon et k le nombre de paramètres libres du modèle.

Dans le cadre de ce mémoire, c'est l'AIC qui sera privilégié.

3.2.5 Les forêts aléatoires

Une autre famille de modèles utilisée dans le cadre de ce mémoire est la forêt aléatoire. Les forêts aléatoires font partie des méthodes ensemblistes dont le but est de mettre en commun plusieurs algorithmes d'apprentissage pour obtenir des prédictions de meilleure qualité. Deux grandes familles de méthodes ensemblistes existent : le boosting et le bagging (ou bootstrap aggregating). Les forêts aléatoires s'inscrivent dans le cadre de cette dernière famille, le bagging.

Le bagging

Le bagging, venant de la contraction de bootstrap aggregating, regroupe un ensemble de méthodes introduites en 1996 par Léo Breiman [2]. L'idée derrière cette méthode est de combiner plusieurs algorithmes d'apprentissage i.i.d. et de variance élevée (et donc de faible biais) afin d'obtenir un modèle plus performant de faible variance.

Mathématiquement, soit B le nombre d'algorithmes d'apprentissage indépendants (et de variance élevée) à disposition et $\hat{m}_i(X)$, $i = 1, \ldots, B$ la prédiction du modèle i.

Dans le cadre d'un problème de régression, le modèle agrégé s'écrira :

$$\hat{m}(X) = \frac{1}{B}\sum_{i=1}^{B} \hat{m}_i(X)$$

Sous l'hypothèse i.i.d. des modèles $\hat{m}_i$:

$$\mathbf{E}[\hat{m}(\mathbf{x})] = \mathbf{E}\left[\hat{m}_1(\mathbf{x})\right] \quad \text{et} \quad \mathbf{V}(\hat{m}(\mathbf{x})) = \frac{1}{B}\mathbf{V}\left(\hat{m}_1(\mathbf{x})\right)$$

Ainsi, le modèle agrégé aura le même biais que les modèles m_i (donc un faible biais) et une variance beaucoup plus faible que ces derniers. Le bagging permet donc de conserver la qualité des différents prédicteurs m_i (en ce qui concerne le biais) tout en en diminuant la variance.

Ce résultat repose cependant sur une hypothèse forte d'i.i.d. des modèles m_i. Le bagging va donc chercher à simuler cette indépendance supposée des m_i qui, en pratique, est difficile à obtenir.

Pour cela, le bootstrap [2] va être utilisé (d'où le nom bootstrap aggregating). Ainsi, B échantillons boostrap de la base d'apprentissage sont tout d'abord tirés pour ensuite ajuster sur chacun d'entre eux un modèle m_i. Ce faisant, de l'aléa est ajouté entre les différents modèles, augmentant ainsi l'indépendance artificielle des modèles.

2. Tirage aléatoire avec remise.

Cette méthode de bagging est particulièrement adaptée aux algorithmes à forte variance et peu stables comme les arbres de décision. Lorsque des arbres de décision sont utilisés, il est alors question de tree bagging. Cependant, une variante existe : les forêts aléatoires. Le principe derrière les arbres de décision va donc être explicité avant de revenir par la suite plus en détails sur les forêts aléatoires, certains ajouts étant présents par rapport au principe usuel du bagging.

Les arbres de décision

Les arbres de décision sont des méthodes de *machine learning* introduites en 1963 par James N. Morgan et John A. Sonquist [17]. Le but de cette méthode est d'expliquer/de prédire une variable aléatoire qu'elle soit quantitative ou qualitative. Pour construire des arbres de décision, différentes méthodes existent :

- la méthode **CART** (Classification and Regression Tree) introduite en 1984 par Breimann et al. [4] ;
- la méthode **CHAID** (CHi-squared Automatic Interaction Detector) publiée en 1980 par Gordon V. Kass. [14]

Dans le contexte de cette étude, c'est la méthode CART pour les problèmes de régression qui sera utilisée et donc présentée.

La méthode CART propose de créer des partitions par divisions successives dans le but final de séparer au mieux les données et de former des groupes de faible variance.

Au départ, l'ensemble des données est considéré et constitue le nœud racine. Une division de cet ensemble est ensuite effectuée. Pour effectuer cette division, représentée par un nœud dans l'arbre, il faut choisir une variable et un seuil de séparation (lorsque la variable choisie est quantitative, dans le cadre d'une variable qualitative, une division en deux groupes de modalités est choisie). Ces paramètres de division sont choisis/optimisés de manière à maximiser l'homogénéité des deux nœuds fils issus de la division et donc minimiser leur impureté. La figure ci-dessous permet de visualiser le fonctionnement d'un arbre de décision à 2 variables explicatives quantitatives : X_1 et X_2.

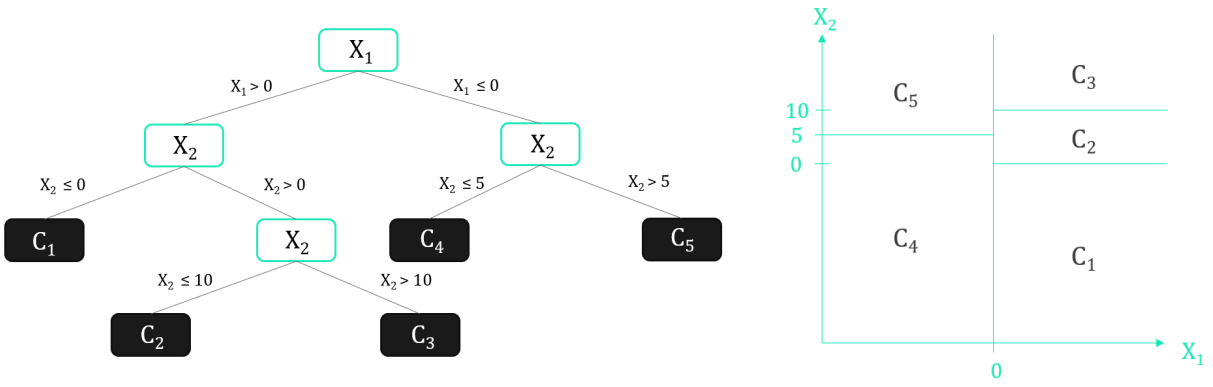


FIGURE 3.13 – Exemple d'arbre de décision

De manière plus mathématique, l'impureté d'un nœud, dans le cadre de la régression, est mesurée par la variance du nœud :

$$\mathcal{D}_{\mathcal{N}} = \frac{1}{|\mathcal{N}|} \sum_{i:X_i \in \mathcal{N}} \left(Y_i - \bar{Y}_{\mathcal{N}}\right)^2$$

où $|\mathcal{N}|$ est le cardinal du nœud $\mathcal{N}$ et $\bar{Y}_{\mathcal{N}} = \frac{1}{|\mathcal{N}|} \sum_{i:X_i \in \mathcal{N}} Y_i$.

Le but à chaque division est alors de chercher le couple (j, s) (variable de division, seuil de division) qui minimise l'impureté des deux nœuds fils $\mathcal{N}_1$ et $\mathcal{N}_2$ définit par :

$$|\mathcal{N}_1(j,s)|\,\mathcal{D}_{\mathcal{N}_1(j,s)} + |\mathcal{N}_2(j,s)|\,\mathcal{D}_{\mathcal{N}_2(j,s)}$$
$$\sum_{i:X_i\in\mathcal{N}_1(j,s)} \left(Y_i - \bar{Y}_{\mathcal{N}_1}\right)^2 + \sum_{i:X_i\in\mathcal{N}_2(j,s)} \left(Y_i - \bar{Y}_{\mathcal{N}_2}\right)^2$$

Le processus est ensuite réitéré et la construction de l'arbre prend fin lorsqu'un critère d'arrêt défini en amont est respecté. Ce critère d'arrêt est lié à la complexité de l'arbre, c'est-à-dire à sa profondeur. Ainsi, plus un arbre est profond, plus la complexité sera importante et il sera donc possible d'observer un faible biais mais une forte variance (phénomène de surapprentissage). A l'inverse, un arbre peu profond aura un fort biais mais une faible variance. Un juste compromis doit être trouvé. Pour trouver l'arbre optimal entre l'arbre trivial et l'arbre maximal, il est courant de procéder par pénalisation. Un paramètre de pénalisation de la complexité de l'arbre est donc optimisé par validation croisée de manière à observer les meilleurs résultats.

Une fois l'arbre construit, à chaque feuille de l'arbre (nœud terminal) est associée une valeur de la variable réponse correspondant à la moyenne des valeurs réponses du nœud (pour un problème de régression). Il est alors possible de prédire la valeur de la variable réponse pour de nouvelles données en suivant l'arbre de décision.

Cette méthode est grandement appréciée aujourd'hui de par sa facilité d'interprétation et est très utilisée dans le cadre du bagging.

Les forêts aléatoires

Lorsque le bagging est appliqué aux arbres de décision sans aucun changement, il est question de tree bagging. Cependant, une variante du tree bagging existe : les forêts aléatoires [3]. Ces dernières ajoutent davantage d'aléa entre les différents arbres du modèle en sélectionnant en plus, pour chaque arbre et à chaque coupure, un certain nombre de variables candidates pour réaliser la coupure. Ce faisant, l'indépendance entre les arbres augmente par rapport au tree bagging et la variance du modèle agrégé en est donc diminuée.

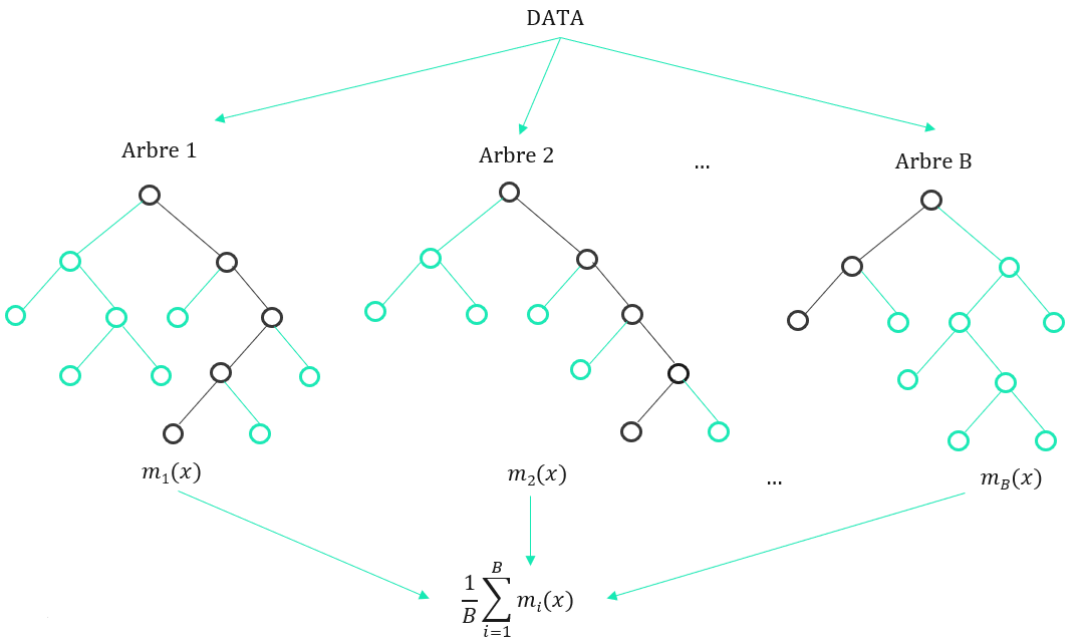


FIGURE 3.14 – Les forêts aléatoires/Le *tree bagging*

Cependant, une critique de ce modèle peut être formulée. En effet, lors de la construction d'un tel modèle, la fonction de perte/d'impureté utilisée est la fonction d'erreur quadratique. Or, cette fonction de perte s'adapte mal aux données actuarielles et ne prend pas en compte l'exposition des différents assurés ou l'excès de zéro de la loi du nombre des sinistres.

Les forêts aléatoires revisitées

Pour toutes ces raisons, un article de recherche intitulé *Boosting insights in insurance tariff plans with tree-based machine learning methods* de « Roel Henckaerts, Marie-Pier Côté, Katrien Antonio et Roel Verbelen » [13] propose une adaptation des forêts aléatoires aux problématiques actuarielles de tarification et notamment à la modélisation du coût et de la fréquence. L'idée générale de l'article est d'adapter la fonction de perte/d'impureté utilisée dans la construction des arbres aux données actuarielles en prenant en compte, par exemple, l'excès de zéro et l'exposition pour la modélisation du nombre de sinistres ou encore la possibilité d'une loi à queue longue pour la modélisation du coût.

Ainsi, au lieu d'utiliser la fonction d'impureté classique (usuellement la variance/l'erreur quadratique) lors de la construction des différents arbres, il a été décidé d'utiliser, pour les modèles de fréquence, la déviance de la loi de Poisson et, pour les modèles de coût, la déviance de la loi Gamma. Un lecteur désireux de disposer davantage d'informations sur ces choix pourra se référer à l'article de recherche initial.

Ce faisant, les arbres de décision de la forêt aléatoire s'adaptent mieux aux données et sont plus performants.

C'est cette dernière version des forêts aléatoires qui sera utilisée dans le cadre du mémoire.

Complexité des forêts alétoires

Concernant la complexité des forêts aléatoires, différents paramètres entrent en jeu :

- le nombre d'arbres dans la forêt ;
- la profondeur des arbres de la forêt (ou encore le nombre minimum requis d'observations dans un nœud donné afin de le diviser) ;
- le nombre de variables candidates à chaque coupure.

De par la loi forte des grands nombres, il n'est pas nécessaire en soi d'optimiser le nombre d'arbres dans la forêt aléatoire, mais il est utile de choisir ce paramètre assez grand pour que le modèle soit stable.

Les deux paramètres restants sont, quant à eux, à optimiser. Différentes approches existent pour ce faire, la plus répandue étant de procéder par validation croisée pour optimiser l'ensemble de ces paramètres et retenir ainsi les paramètres rendant la plus petite erreur de généralisation.

3.2.6 La comparaison des modèles

Une fois l'ensemble de ces modèles mis en place (GLM et forêt aléatoire) et la complexité adéquate trouvée pour éviter le phénomène de surapprentissage, les modèles peuvent être comparés entre eux à l'aide d'une erreur de généralisation commune aux deux modèles. Dans le cadre de ce mémoire, c'est la RMSE qui est retenue et qui permettra de départager tous les modèles optimisés.

Chapitre 4

Constitution de la base de données « à l'adresse »

La récupération des données en *Open Data* et le traitement de ces dernières sont au cœur de ce mémoire, ces deux étapes étant essentielles à la mise en place de modèles de tarification à l'adresse efficients. Dans ce chapitre, la constitution de la base de données « à l'adresse » va donc faire l'objet d'une présentation détaillée. Une réflexion sur la qualité des données sera également menée en parallèle, cette notion impactant directement la qualité des modèles proposés par la suite. Pour finir, une analyse de la base constituée, similaire à celle menée sur la base « assureur » dans le chapitre 2 , sera conduite.

4.1 La base de données

Afin de constituer la base de données « à l'adresse », les adresses de la base « assureur » ont fait l'objet d'une extraction. Cela représentait environ **26 000 adresses différentes** pour un total de 255 000 lignes. Puis, à partir de ces seules adresses, différentes informations ont été récupérées en *Open Data*. Ce sont ces informations qui constitueront par la suite les variables des modèles à l'adresse.

Ce mémoire se focalisant sur la tarification à l'adresse, une attention toute particulière a été consacrée à la récupération de données à la maille adresse. Cependant, afin de conserver une certaine exhaustivité dans cette étude, des données à la maille commune/département et des données météorologiques ont également été utilisées.

Par ne pas alourdir le texte, aucune notion technique quant à la récupération des données ne sera abordée dans ce chapitre bien que cette étape ait requis de s'approprier certaines méthodologies complexes de récupération de données et ait donc nécessité d'y consacrer un temps considérable.

Enfin, s'il est souhaité un jour utiliser ce processus de manière opérationnelle, il est également important de se rendre compte du temps de récupération des données. Ainsi, pour donner un ordre de grandeur, il a fallu environ 3 mois pour récupérer l'ensemble des données présentées par la suite. Des experts en informatique, de par leurs connaissances, auraient certainement pu réduire ce temps de récupération en optimisant le code ou en utilisant des serveurs, rendant ainsi le recueil des données plus accessible.

Les données récupérées dans le cadre de ce mémoire vont donc faire l'objet d'une présentation détaillée dans ce chapitre. Cependant, avant de pouvoir obtenir des données pouvant réellement impacter le tarif d'une assurance MRH, il est nécessaire de procéder au recueil de certaines données géographiques de base.

4.1.1 Quelques données géographiques préliminaires

Formalisation des adresses

Pour la plupart, les adresses extraites de la base « assureur » y figuraient sous la forme d'une seule variable prenant le format suivant :

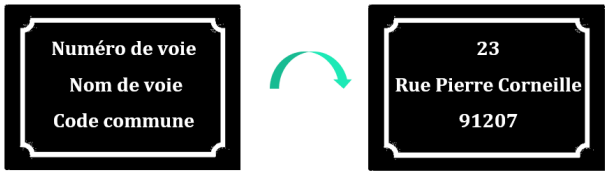


FIGURE 4.1 – Format initial de l'adresse

Après analyse des données, il est apparu qu'il n'était pas possible d'utiliser ces données en l'état et qu'un important travail de mise en forme et de normalisation des adresses devait être effectué. Ainsi, pour ne citer que quelques exemples de modifications apportées, il a été nécessaire de :

- rajouter un chiffre « 0 » devant tous les numéros de voie allant de 1 à 9 afin d'homogénéiser le format des adresses (certaines adresses présentaient en effet tantôt le numéro de voie « 1 » et tantôt le numéro « 01 » et ce, pour la même adresse) ;
- mettre en majuscules les adresses afin d'éviter les anomalies liées à l'accentuation ;
- supprimer les caractères superflus (tirets, points, ...) ;
- etc.

L'ensemble de ces changements a fait l'objet d'une automatisation.

Certaines adresses étaient cependant plus compliquées à traiter et présentaient des erreurs de saisie. Pour ces dernières, aucune règle de mise en forme n'était applicable. Il a donc fallu rentrer ces adresses de « mauvaise qualité » dans le moteur de recherche Google afin de les retrouver et de les restituer dans un format correct. Le graphique ci-dessous présente des exemples de traitements réalisés pour trois adresses différentes.

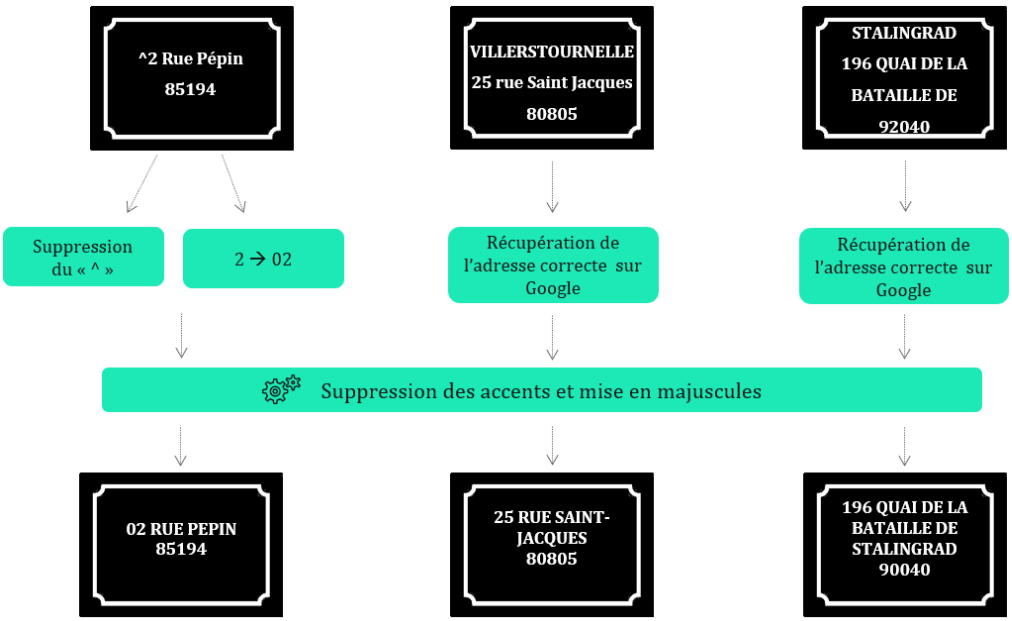


FIGURE 4.2 – Traitement des adresses

Une fois les adresses mises en forme, les codes officiels géographiques de l'INSEE ont été utilisés afin de retrouver les codes et noms des communes, régions et départements associés à chaque adresse. A ce stade, il s'est avéré nécessaire de prendre en compte un certain nombre d'éléments exogènes comme, par exemple, la fusion de certaines communes intervenue entre la date des données et les codes officiels géographiques actuels. Dans ce cas de figure, ce sont les codes géographiques actuels qui ont finalement été retenus.

A la fin de cette première étape, la base de données était constituée des variables suivantes :

- l'**adresse** selon le format suivant : numéro de voie + nom de voie + code postal + nom de la commune ;
- le **nom de la commune** ;
- le **code de la commune** ;
- le **type de la commune** ;
- la **commune parent** le cas échéant ;
- le **libellé d'acheminement** ;
- le **code postal** ;
- le **nom du département** ;
- le **code du département** ;
- le **nom de la région** ;
- le **code de la région**.

Le graphe ci-dessous résume ces informations.

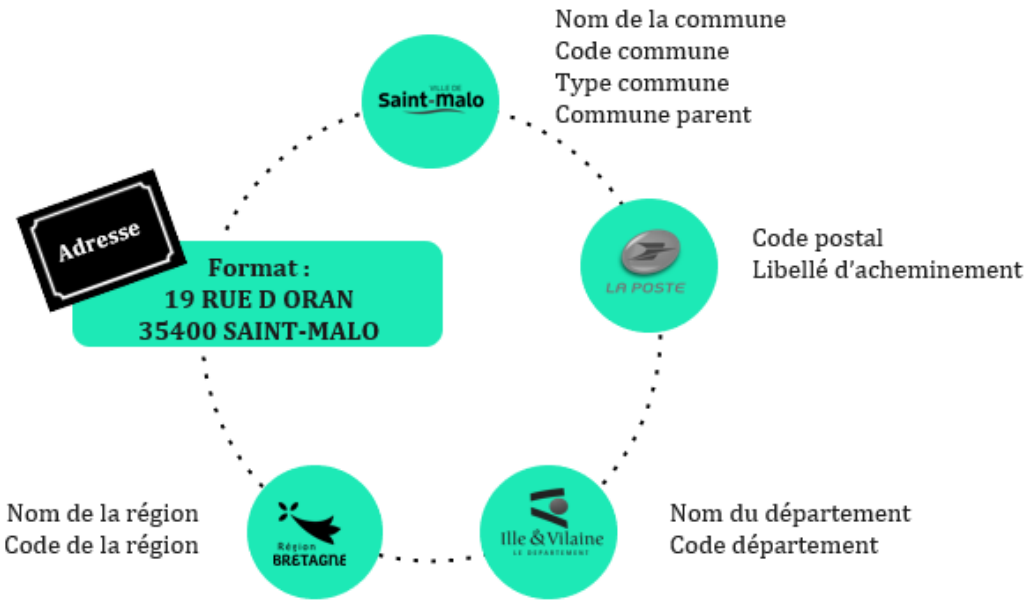


FIGURE 4.3 – Variables géographiques non tarifaires de la base « à l'adresse »

Géocodage

Un travail de géocodage a ensuite été entrepris, c'est-à-dire qu'à chaque adresse, un couple latitude/longitude a été affecté.

Pour effectuer ce géocodage, des données disponibles en *Open Data* ont été exploitées. En effet, dans le cadre de la mission Etalab citée précédemment, le gouvernement a souhaité jouer la carte de la transparence en facilitant l'accès à de nombreuses informations. C'est pourquoi il a, entre autres, mis à disposition du public la Base Adresse Nationale (BAN). Cette base, établie en partenariat avec des acteurs nationaux, des acteurs locaux et des citoyens, vise à réunir l'ensemble des adresses géolocalisées du territoire national. Elle est disponible sous licence ouverte depuis le 1er janvier 2020 et est fournie avec une *Application Programming Interface* (API)[1] gratuite permettant, entre autres, le géocodage.

C'est cette API qui a été utilisée dans le cadre du géocodage. Ainsi, itérativement, les différentes adresses de la base de données ont été renseignées sur l'API à l'aide du package **curl** de R. A chaque adresse, un fichier JSON contenant des réponses candidates (correspondant potentiellement à l'adresse d'entrée) était renvoyé. Pour chaque réponse candidate, différentes informations étaient fournies :

- l'adresse normalisée présente dans la BAN sous le format suivant : numéro de la voie + nom de la voie + code postal + nom commune ;
- une décomposition de l'adresse (respectivement le numéro de la voie, le nom de la voie, le code postal, le code commune, le nom de la commune, le code du département, le nom du département, le nom de la région) ;
- la latitude et la longitude associées à l'adresse ;
- un score de ressemblance comparant l'adresse en entrée avec celle retournée par l'API.

C'est ce score qui a permis de retenir, parmi l'ensemble des adresses candidates, celle qui semblait la plus pertinente par rapport à l'adresse entrée.

Certaines informations comme notamment la décomposition de l'adresse, qui étaient déjà présentes dans la base, ont été écartées. Les informations non redondantes, à savoir la **latitude** et la **longitude**, ont, quant à elles, été conservées.

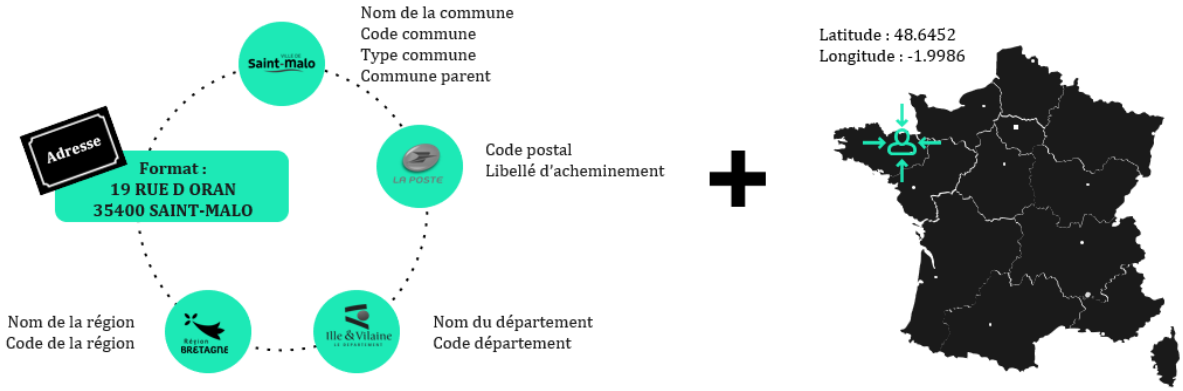


FIGURE 4.4 – Variables géographiques non tarifaires de la base « à l'adresse »

1. Interface de programmation d'applications qui permet de rendre disponibles les données ou les fonctionnalités d'une application existante afin que d'autres applications les utilisent.

☞ **Focus sur la qualité du géocodage**

Le géocodage va se révéler primordial pour la suite du mémoire, de nombreuses informations étant récupérées par le biais de la latitude et de la longitude. Au cas particulier, et comme évoqué précédemment, c'est l'API de la BAN qui a été utilisée. La qualité des données dépend donc fortement du travail effectué lors de la constitution de la BAN et lors de la mise en place de l'API. **Deux erreurs possibles** peuvent alors se distinguer.

La première concerne l'**utilisation de l'API**. En effet, comme précisé précédemment, cette API nécessite de renseigner une adresse en entrée pour ensuite retourner plusieurs adresses candidates correspondant potentiellement à cette adresse. Il semble donc possible que les adresses en sortie de l'API diffèrent de l'adresse en entrée. Afin d'identifier et de détecter facilement ce problème, les créateurs de l'API ont mis en place un score de ressemblance entre l'adresse en entrée et les adresses candidates en sortie. Par le biais de ce score, il est donc possible de maîtriser ce risque.

Pour le minimiser lors de la récupération des données, une procédure a donc été mise en place afin de ne conserver que les données ayant un score de ressemblance supérieur à 0.7. Ce seuil a été choisi après analyse manuelle de nombreuses adresses. Lors de ce processus, de nombreuses adresses ont donc été supprimées de la base « assureur » initiale (celle avec les 800 000 lignes) du fait d'une pauvre qualité de géocodage. Cependant, il est toujours possible que quelques erreurs se soient glissées.

La **précision** est également un élément clé du géocodage et constitue la deuxième source d'erreurs possibles. En effet, par manque de précision, il est possible d'obtenir les coordonnées de l'adresse voisine et non de l'adresse d'intérêt lors du géocodage. Ce risque est malheureusement difficilement quantifiable. Malgré tout, la BAN étant une base gouvernementale, il semble raisonnable de penser que la qualité des données a déjà fait l'objet de vérification en interne. Cependant, en l'absence d'informations complémentaires, il est important de garder ce point en tête.

Enfin, il faut également conserver à l'esprit que ce travail de géocodage est un métier à part entière : de nombreuses entreprises privées développent en effet leur propre logiciel/API pour le géocodage. Des API professionnels pourraient donc se révéler plus performantes (l'API Google par exemple). Dans le futur, lors de la commercialisation ou de la mise en place d'un réel processus de tarification à l'adresse, une étude comparative des différents acteurs du marché s'impose et devra être réalisée.

Les références cadastrales

La notion de cadastre est également primordiale pour la suite du mémoire et permettra de récupérer des données pour certains assurés.

Selon les services publics :

« Le plan cadastral est un document graphique. Il représente tout le territoire de la commune découpé en sections cadastrales (parties du territoire). L'emprise au sol des bâtiments est également représentée. Le tracé des principales voies de communication et des cours d'eau, la position des agglomérations, des hameaux, des fermes isolées, ainsi que le nom des communes limitrophes y sont indiqués. Les sections cadastrales peuvent être découpées en feuilles parcellaires et lieux-dits avec les numéros et les limites des parcelles. Le nom des propriétaires n'y figure pas. Ces documents sont accessibles à tous. »

Ainsi, à chaque propriété foncière est associée **une parcelle cadastrale**. L'idée a alors été de récupérer, pour chaque adresse, la référence cadastrale de la maison. Cela a été possible grâce au site geoportail.gouv.fr, un site du gouvernement proposant l'accès à des contenus géographiques émanant de différents contributeurs.

Il est à noter que, le parcellaire cadastral de l'IGN [2] étant en licence ouverte, la donnée émanant du site Géoportail n'est soumise à aucune restriction et peut ainsi être exploitée dans le cadre de cette étude.

Enfin, sur un plan plus technique, le processus de récupération présenté ci-dessus a été rendu possible grâce au package **RSelenium** de R.

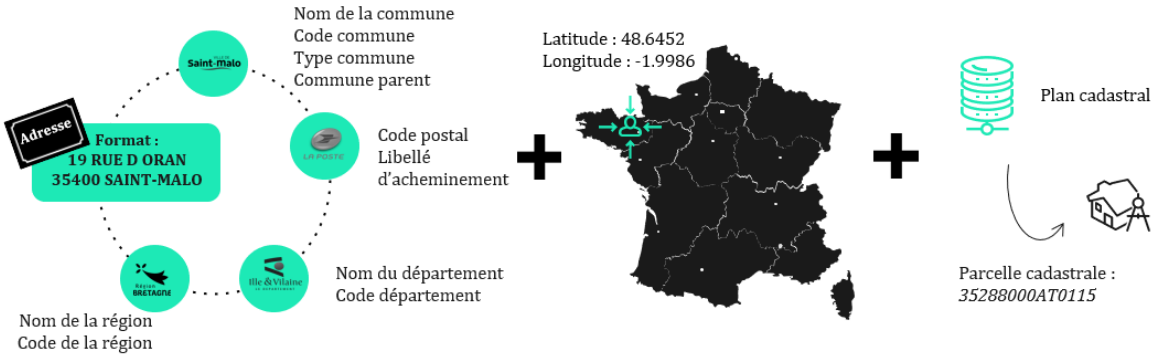


FIGURE 4.5 – Variables géographiques non tarifaires de la base « à l'adresse »

☞ Focus sur la qualité des références cadastrales

Là encore, il est important de réaliser que de nombreux phénomènes peuvent venir entacher la qualité de la donnée.

Ainsi, parfois, une maison peut se situer sur plusieurs parcelles cadastrales, rendant de ce fait les **données** récupérées **incomplètes**.

FIGURE 4.6 – Exemple d'une maison sur plusieurs parcelles

Par exemple, sur l'image ci-dessus, la même maison est présente sur les parcelles « 0116 » et « 0117 ». Or, Géoportail récupère le cadastre d'une adresse en géocodant lui-même l'adresse (grâce à la BAN). Un couple latitude/longitude ne pouvant être présent qu'en un seul endroit, seule une parcelle sera récupérée : la parcelle « 0117 » dans l'exemple. Il est donc possible par la suite de passer à côté de certaines informations associées à la parcelle « 0116 ».

Le nombre de maisons exposées à ce risque n'est malheureusement pas quantifiable de manière aisée. Dans le cadre de cette étude, il est donc supposé par la suite que cette situation reste exceptionnelle.

2. L'Institut national de l'information géographique et forestière ou IGN est un établissement public à caractère administratif ayant pour mission d'assurer la production, l'entretien et la diffusion de l'information géographique de référence en France.

Un autre phénomène notable et à garder à l'esprit pour la suite provient du fait que, sur une même parcelle cadastrale, plusieurs maisons peuvent être présentes (parfois même, plus d'une vingtaine de maisons sont présentes sur une même parcelle). La photo ci-dessous illustre ce cas de figure : deux maisons sont présentes sur la parcelle « 0077 ».

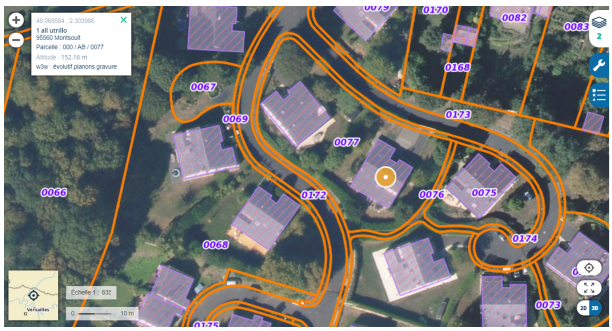


FIGURE 4.7 – Exemple d'une parcelle contenant plusieurs maisons

Dans ce cas, les données associées à la parcelle (par exemple les demandes de valeurs foncières présentées en 4.1.2) peuvent concerner n'importe quelle maison de la parcelle : des **données erronées** risquent donc d'être intégrées à l'étude en cas de mauvaise association.

Pour tenter de maîtriser ce risque, un critère de sélection des assurés sera défini lors de la mise en place de la reconnaissance d'images (en 4.1.2). Ainsi, l'ensemble des assurés dont la maison est présente sur une parcelle de plus de 4 bâtis sera supposé être concerné par le cas de figure ci-contre et sera donc retiré de l'étude (suppression de ces lignes dans la base initiale « assureur » avec les 800 000 lignes). Indirectement, par ce biais, le patrimoine de l'assuré sera limité à une maison et trois dépendances (i.e. trois autres petits bâtis).

Ce faisant, le risque est limité mais malheureusement toujours existant : le cas particulier de la photo ci-dessus ne sera, par exemple, pas détecté comme une anomalie étant donné que seuls deux bâtis sont présents sur la même parcelle. Ces cas seront toutefois considérés comme exceptionnels.

Les codes officiels géographiques, le géocodage et le cadastre sont des informations géographiques non tarifaires à proprement parler mais utiles à la récupération d'autres variables plus pertinentes. Ces variables plus pertinentes vont maintenant être exposées.

4.1.2 Les données à la maille adresse

Altitude

Au premier abord, une des premières données à laquelle il est naturel de penser est l'**altitude**. En effet, en lisant les conditions générales de l'assureur à l'étude, il est apparu que la garantie DDE couvrait tant les dégâts liés aux infiltrations, aux fuites que ceux causés par de fortes pluies ou des inondations non reconnues comme catastrophes naturelles. La variable altitude peut donc être potentiellement tarifaire, au moins pour la garantie DDE. Cette information a donc été récupérée en parallèle de la parcelle cadastrale sur le site geoportail.gouv.fr, en utilisant **RSelenium**. Concernant les mentions légales, les données altimétriques proviennent de la BD ALTI de l'IGN, une base encore une fois sous licence ouverte.

Open Street Map

➢ **Données existantes issues de travaux internes**

A cela s'ajoutent des données de type « point d'intérêt » issues d'Open Street Map.

Open Street Map est un projet collaboratif de cartographie en ligne qui vise à constituer une base de données géographiques libre du monde. Ce projet, lancé en juillet 2004 par Steve Coast, est aujourd'hui un incontournable de l'*Open Data* au sein duquel chacun est libre de contribuer en renseignant des informations.

Ainsi, grâce à ce projet, de nombreuses données sont désormais disponibles. Pour exemple :

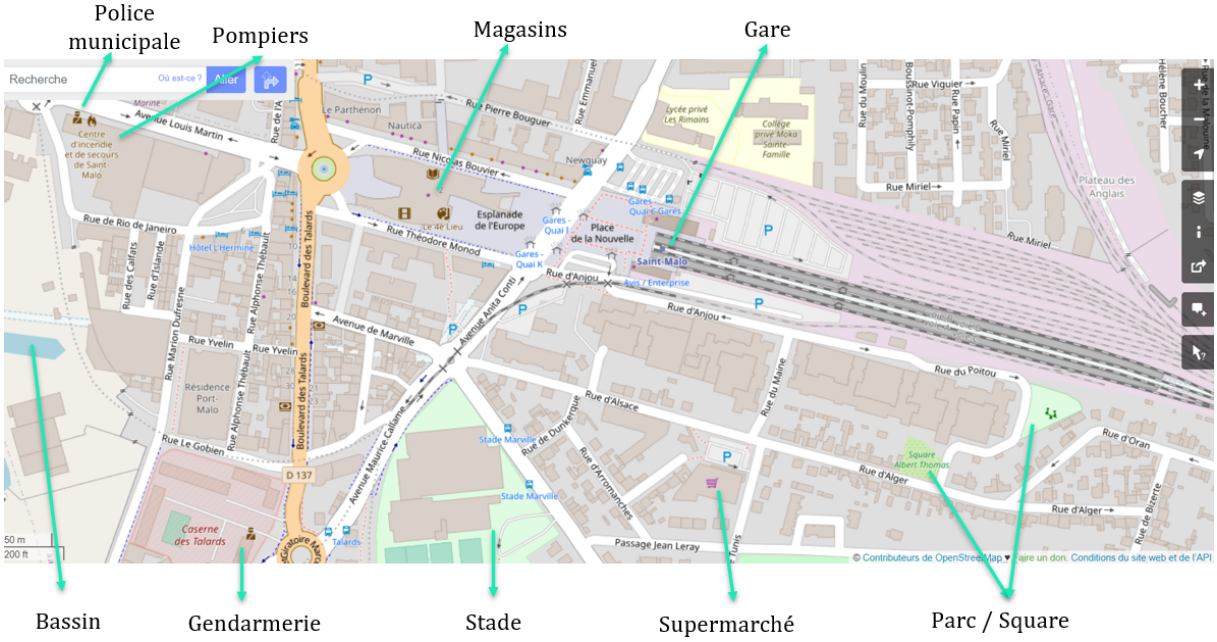


FIGURE 4.8 – Exemple d'informations issues d'Open Street Map

Sur l'image ci-dessus, figurent de nombreux éléments. Certains ont été mis en évidence comme, par exemple, la gendarmerie, la police municipale, les pompiers, la gare, les parcs, le bassin, etc. En zoomant, d'autres éléments apparaissent : les restaurants, les parkings, les arrêts de bus, les hôtels, les coiffeurs, les banques, les pressings, les garages, les boulangeries... et tout autre élément distinctif d'intérêt permettant de cartographier une ville. Cet exemple montre donc le potentiel d'Open Street Map et la richesse des informations contenues dans ce projet.

L'ensemble de ces informations peut être récupéré sous R à l'aide du package **OSMAR**. Sur un plan plus technique, ces informations sont représentées par des mots-clés et un élément peut disposer de plusieurs mots-clés : l'eau est par exemple représentée par un bassin, une rivière, un lac, un ruisseau ou encore un étang. Ainsi, pour détecter la présence d'eau, il est nécessaire d'utiliser l'ensemble des mots-clés qui lui sont associés.

Dans le cadre de ce mémoire, pour chaque point présent dans la base (représenté par un couple latitude/longitude), des périmètres de 50 mètres et 500 mètres autour de ce point ont été définis et l'ensemble des informations disponibles autour de ce point a été récupéré.

Puis, à partir de ces informations, différents éléments d'intérêt ont été calculés. Ainsi, il a semblé pertinent de récupérer dans les périmètres à 50 et 500 mètres :

- la **présence d'eau** ;
- la **surface d'eau** dans le périmètre ;
- la **présence d'un poste de police** ;
- la **présence d'un centre commercial** ;
- la **présence de transports en commun**.

➢ Ajout du mémoire

En plus de ces informations issues de travaux internes à Sia Partners, d'autres éléments sont venus les compléter.

Il a en effet semblé pertinent de calculer pour chaque adresse (et donc chaque couple latitude/longitude), la **distance au poste de police le plus proche** et la **distance à la caserne de pompiers la plus proche**.

Pour ce faire, la position de l'ensemble des casernes et des postes de police et de gendarmerie a été récupérée d'Open Street Map en passant par overpass-turbo [3]. Une fois ces données acquises, pour chaque adresse, la distance à chacun des postes de police/gendarmerie et des casernes a été calculée. Toutefois, afin de conserver un certain réalisme, il a été décidé de ne pas utiliser la distance à vol d'oiseau mais de calculer la distance « réelle » basée sur l'utilisation des voies existant réellement. Une API a donc été utilisée afin d'obtenir la distance donnée par le plus court chemin et le temps donné par le chemin le plus rapide. Cela a été possible grâce à l'utilisation d'Openrouteservice.

L'ensemble des données issues d'Open Street Map peut donc être récapitulé dans le graphe suivant :

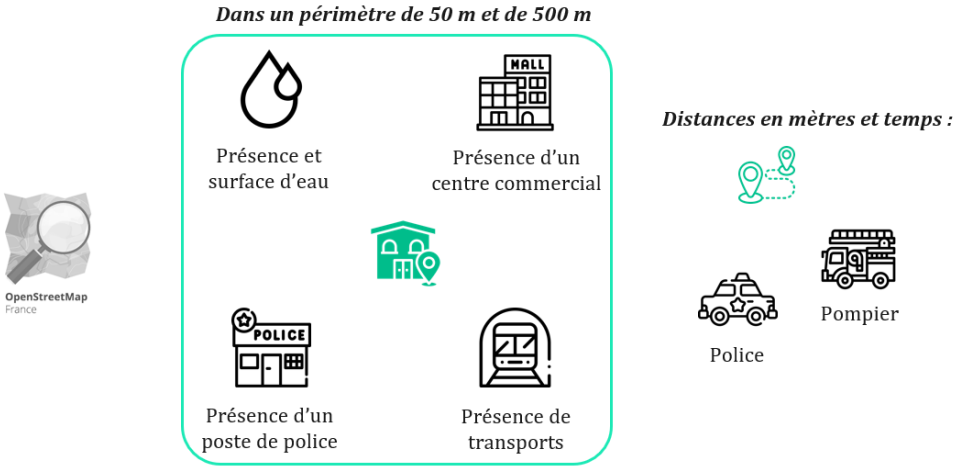


FIGURE 4.9 – Variables Open Street Map

Intuitivement, les variables associées à l'eau pourraient impacter la garantie DDE. Et enfin, les variables liées à la présence d'un centre commercial, de transports en commun ou de postes de police concerneraient, quant à elles, plutôt la garantie vol.

> ☞ **Focus sur la qualité des données d'Open Street Map**
>
> Les données d'Open Street Map étant issues d'un projet collaboratif en ligne, un risque potentiel se dessine quant à la fiabilité des données. En effet, toute personne tierce pouvant accéder aux données et les éditer librement, il ne peut être exclu qu'une personne distille, intentionnellement ou par erreur, des informations erronées. Les modifications/ajouts ne faisant l'objet d'aucun contrôle par Open Street Map, ces fausses informations sont dès lors accessibles au public, créant ainsi un problème de fiabilité des données concernées. Il est, par ailleurs, à noter que ce type d'erreur persistera dans le temps jusqu'à ce qu'un autre contributeur s'en aperçoive et résolve le problème. De la même manière que pour Wikipédia, la qualité des données dépend donc du nombre de contributeurs ainsi que de la qualité de leur travail.

3. Un outil de récupération des données d'Open Street Map

Images satellites

En outre, des informations issues de la reconnaissance d'images satellites ont, par la suite, été utilisées.

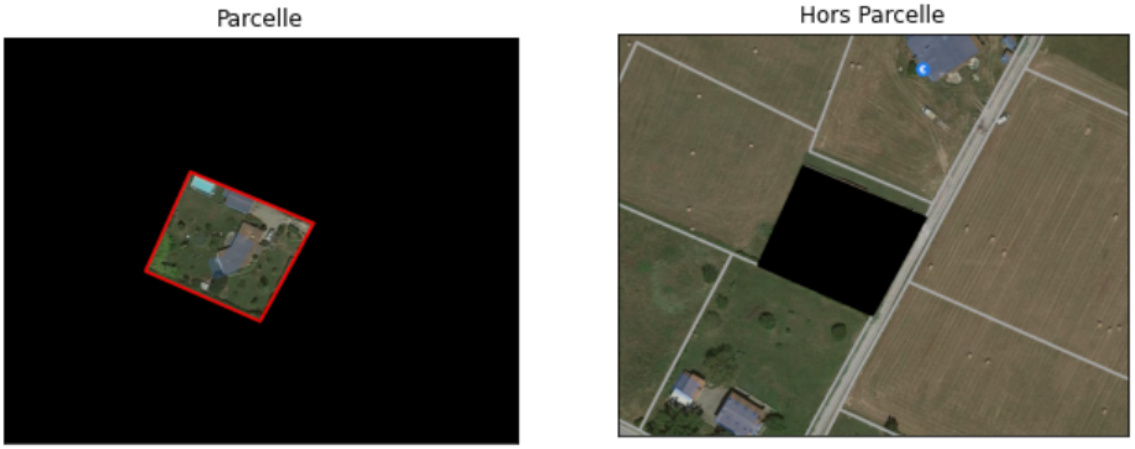
Pour cela, pour chaque adresse de la base de données, une image satellite de la parcelle associée a été récupérée. Sur ces images, il s'est avéré que l'ensemble des bâtis apparaissait en couleur bleutée et que la parcelle associée à l'adresse d'intérêt était encadrée en rouge. Ces particularités ont permis de mettre en place une reconnaissance d'images à l'aide de la librairie **OpenCV** de Python. Pour certaines données récupérées (présence de piscine par exemple), il aurait été possible de mettre en place des modèles plus complexes de type *deep learning* pour obtenir les mêmes informations. Cependant, pour des considérations de temps, d'apport et de complexité, il a été décidé d'utiliser la librairie **OpenCV** de Python pour cette première étude. Les méthodes de *deep learning* appliquées à la reconnaissance d'images restent cependant une piste très intéressante à creuser dans le cadre de futures études.

Les traitements automatisés sous **OpenCV** vont maintenant être explicités. Chaque image satellite est de la forme suivante :

FIGURE 4.10 – Image satellite

Sur cette image, plusieurs éléments se distinguent : la maison, les dépendances, la piscine, les parties arborées ou encore le jardin.

Tout d'abord, afin de ne conserver que les informations de la parcelle d'intérêt, un filtre rouge a été appliqué. Ce filtre a permis de récupérer le contour fermé de la parcelle pour ensuite séparer la parcelle du hors parcelle.



Parcelle

Sur l'image de la parcelle, différents traitements ont été appliqués. Tout d'abord, l'ensemble des bâtis a été détecté grâce à un filtre bleu foncé/gris. Le bâti le plus grand a été considéré comme étant la maison. Les autres bâtis, s'il y en avait, ont été considérés comme des dépendances. La surface plane de chaque bâti a ensuite été calculée selon différentes approximations. Ces approximations ont été réalisées grâce aux fonctions suivantes : **cv2.convexHull** pour une forme convexe et **cv2.minAreaRect** pour une forme rectangulaire. Le calcul de la surface, quant à lui, a été possible grâce à la fonction **cv2.contourArea**. Une simple échelle permet ensuite de convertir ces grandeurs en mètres carrés.

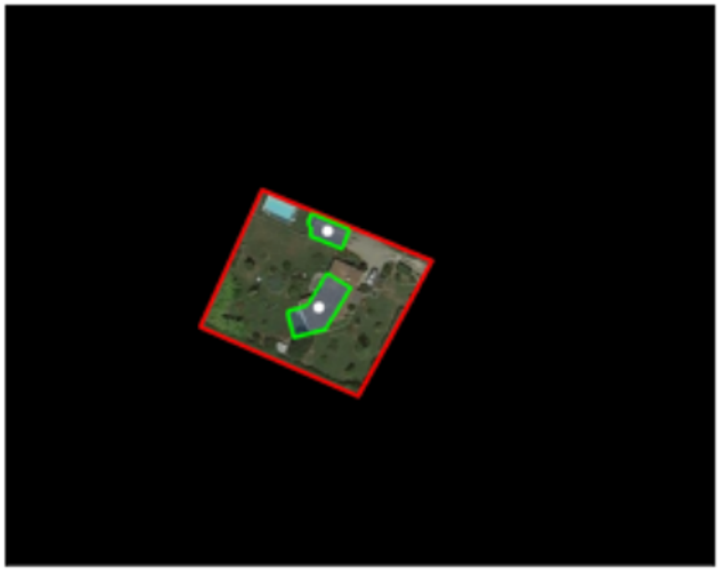


FIGURE 4.11 – Bâtis de la parcelle

Par la suite, la piscine a été détectée cette fois-ci en appliquant un filtre bleu clair. Ainsi, si des piscines sont présentes sur l'image, une indicatrice indiquera la présence de ces dernières.

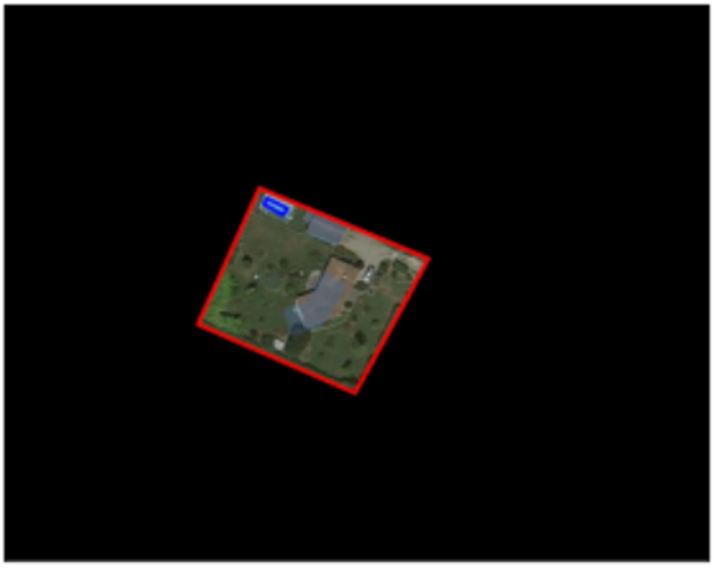


FIGURE 4.12 – Présence d'une piscine

Enfin, pour finir, la notion de parties vertes a été incorporée. En appliquant un filtre vert, il est

en effet possible de récupérer un pourcentage de parties vertes sur l'ensemble de la parcelle. Ce pourcentage peut être grossièrement associé au jardin de la maison.

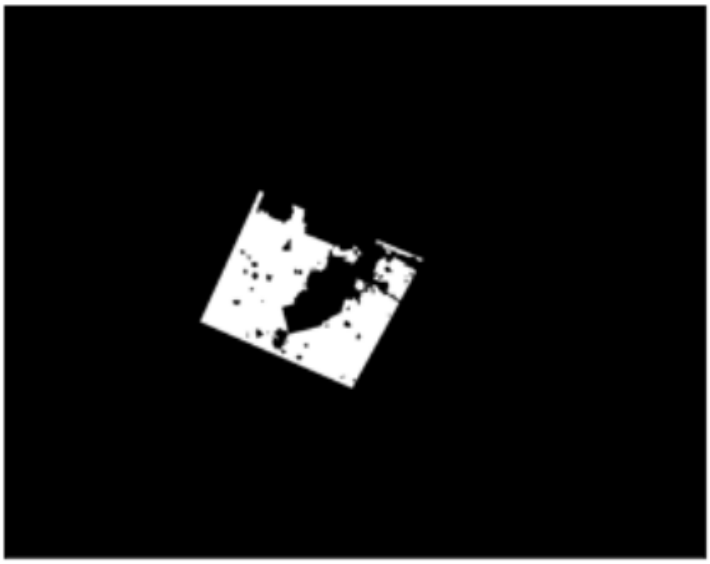


FIGURE 4.13 – Masque des parties vertes

Hors parcelle

Pour enrichir les données, il a également été décidé d'utiliser le hors parcelle pour déterminer la distance à la maison la plus proche. Cet indicateur permettra de disposer d'une variable dans la base représentant l'isolement de la maison (potentiellement tarifaire pour le vol).

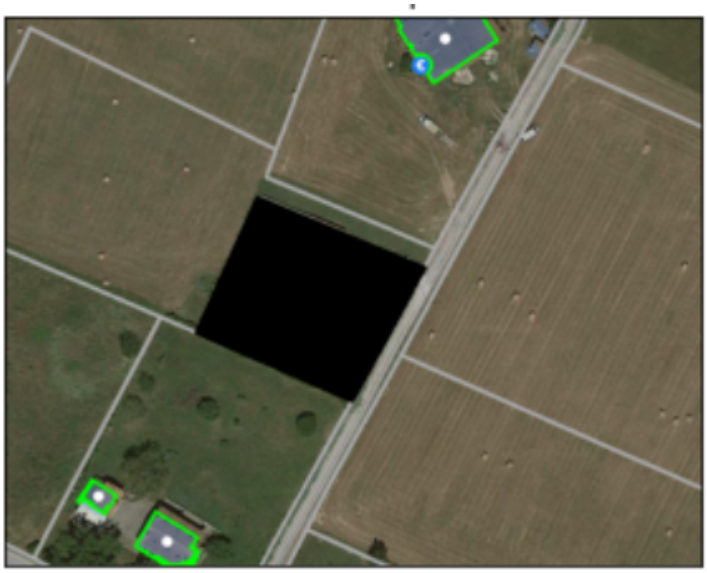


FIGURE 4.14 – Bâti hors parcelle

Pour résumer, la reconnaissance d'images permet d'obtenir les variables suivantes :

- la **surface de la parcelle** ;
- la **surface plane de la maison** (sans approximation, avec une approximation par une forme convexe et avec une approximation par un rectangle) ;
- le **nombre de bâtis** sur la parcelle ;

- la **surface plane de chaque bâti** (sans approximation, avec une approximation par une forme convexe et avec une approximation par un rectangle) ;
- la **présence d'une piscine** ;
- le **pourcentage de parties vertes** sur la parcelle.

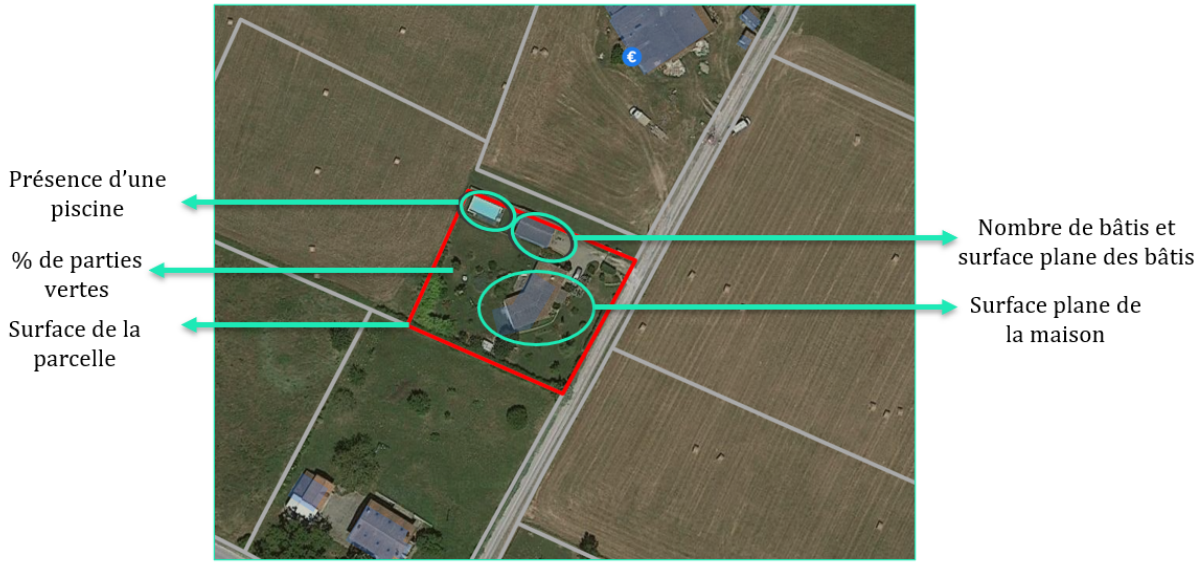


FIGURE 4.15 – Variables issues de la reconnaissance d'images

Par le biais de la reconnaissance d'images, il est espéré que les données de surface récupérées remplacent les informations déclaratives de surface et de nombre de pièces renseignées actuellement par l'assuré lors de la souscription et ce, pour toutes les garanties. Il est cependant important de noter que les surfaces récupérées sont des surfaces planes ne prenant pas en compte les étages.

☞ Focus sur la qualité de la reconnaissance d'images

La détection de couleur effectuée dans le cadre de cette reconnaissance d'images peut s'avérer ambiguë.

En effet, lors de la détection des bâtis par exemple, une palette de couleurs bleu foncé/gris est utilisée. Malheureusement, malgré cette palette assez large de couleurs, il arrive parfois que des bâtis ne soient pas détectés. Ce problème relève alors de nuances de couleurs difficiles à calibrer de manière parfaite pour toutes les images. C'est donc une erreur dont il faut être conscient mais qu'il faut accepter.

Il serait envisageable d'utiliser une palette plus large afin de détecter un plus grand nombre de nuances de couleurs mais cette solution conduirait à détecter des surfaces plus grandes de bâtis (prise en compte des ombres par exemple). De plus, il pourrait parfois y avoir un risque de confusion entre le bleu de la piscine et le bleu, plus mat, des bâtis. Un juste équilibre doit donc être trouvé.

Au cas particulier, plusieurs palettes de couleurs ont été testées afin de trouver un équilibre entre les bâtis et les piscines, et ce, tout en obtenant un taux de détection convenable pour ces deux éléments. C'est finalement la palette ayant donné les meilleurs résultats sur un petit échantillon qui a été retenue.

La reconnaissance d'images, bien qu'apportant un caractère innovant à l'étude, comporte cependant certaines limites liées aux nuances de couleurs.

Risques naturels et technologiques

Des données concernant les risques naturels et technologiques ont également été introduites. Pour cela, c'est encore un site gouvernemental qui a été utilisé : le site georisques.gouv.fr, un site proposé par le Ministère de la Transition Ecologique et réunissant de nombreuses bases de données sous une interface pratique et lisible. Les mentions légales du site indiquent que l'ensemble des informations fournies par le site est réglementé par une licence ouverte (sauf mention du contraire) : il n'existe donc, pour la plupart des données de ce site, aucune condition quant à la réutilisation de ces données. Après vérification concernant les données d'intérêt récupérées dans le cadre de ce mémoire, il s'avère que ces dernières sont bien sous licence ouverte.

Le package **RSelenium** de R a donc encore une fois été utilisé pour automatiser la récupération des données.

A la fin du processus, les informations suivantes sont accessibles :

Catégorie	Variables
Informations générales	Nombre de risques majeurs à la commune
	Nombre de reconnaissances de catastrophes naturelles à la commune ou au département
	Nombre de plans de prévention des risques naturels
	Nombre de plans de prévention des risques technologiques
	Détails sur les risques majeurs à la commune : une variable binaire pour chaque risque (O/N). Risques présents : Inondation, Rupture de barrage, Avalanche, Feu de forêts, Mouvements de terrain, Zone sismique 1, Zone sismique 2, Zone sismique 3, Zone sismique 4, Engins de guerre, Radon, Transport de marchandises dangereuses, Phénomène météo - tempêtes, grains et vent, Phénomène météo - neige, pluie verglaçante, Phénomène météo - grêle, Phénomène météo - foudre, Phénomène lié à l'atmosphère, Emissions de gaz de mine, Nucléaire, Risque minier, Risque technologique, Risque industriel
	Détails sur les arrêtés de catastrophes naturelles ayant eu lieu dans la commune : un nombre d'arrêtés par type de catastrophes naturelles. Catastrophes naturelles présentes : Inondations, Coulées de boue, Mouvements de terrain, Chocs mécaniques liés à l'action des vagues, Tempête, Glissements de terrain, Effets exceptionnels dus aux précipitations, Effondrements de terrain, Affaissements de terrain, Eboulements, Avalanche, Séisme, Poids de la neige - chute de neige, Tornade et grêle, Raz-de-marée, Tassements de terrain, Inondations par remontées de nappe phréatique, Inondations par remontées de nappe naturelle ...

Catégorie	Variables
Inondations	Territoire à risque important d'inondation à la commune (TRI)
	Nombre d'évènements historiques d'inondations au sein du département
	Plan de prévention des risques inondations au sein de la commune
	Programme de prévention PAPI (Programmes d'action de prévention des inondations)
Mouvements de terrain	Mouvements de terrain recensés dans les 500 m
	Plan de prévention des risques de mouvements de terrain à la commune
Cavités souterraines	Cavités souterraines recensées dans un rayon de 500 m
	Plan de prévention des risques liés aux cavités souterrai--nes à la commune
Séismes	Niveau du risque sismique dans la commune (1,2,3,...)
	Plan de prévention des risques sismiques à la commune
Radon	Niveau du potentiel radon à la commune
Retrait - gonflement des sols argileux	Exposition au retrait-gonflement des sols argileux
	Plan de prévention des risques liés au retrait-gonflements des sols argileux
Pollution des sols, sis et anciens sites industriels	Nombre de secteurs d'information sur les sols recensés dans un rayon de 1000 m
	Nombre de sites pollués ou potentiellement pollués recensés dans un rayon de 50 m
	Nombre d'anciens sites industriels recensés dans un rayon de 500 m
Installations industrielles	Nombre d'installations classées recensées dans un rayon de 1000 m
	Nombre d'installations rejetant des polluants dans un rayon de 5000 m
	Plan de prévention des risques technologiques d'installations industrielles à la commune
Canalisation de matières dangereuses	Canalisations de matières dangereuses recensées dans un rayon de 1000 m
Installations nucléaires	Installations nucléaires à moins de 10 km
	Installations nucléaires à moins de 20 km

Lorsque cela était possible, les données ont donc été récupérées à la maille adresse : installations nucléaires, sites industriels, etc. Malheureusement, beaucoup d'autres variables, de par la délimitation réglementaire des zones à risque, se trouvent à des mailles plus générales comme les plans de prévention des risques. Ces données, à des mailles plus larges, ont tout de même été conservées car elles pourraient se révéler tarifaires dans le cadre de l'étude.

Demandes de valeurs foncières

Enfin, pour compléter toutes ces informations, le fichier des demandes de valeurs foncières (DVF) a également été utilisé. Ce fichier, détenu par la Direction générale des Finances publiques (DGFiP), recense les transactions immobilières en France et fait l'objet d'une mise à jour trimestrielle. Il est accessible depuis avril 2019 en licence ouverte pour répondre à un objectif de transparence des marchés fonciers et immobiliers. Cependant, pour des raisons légales, il n'est

possible d'utiliser les DVF que sur les 5 dernières années écoulées. Il est également à noter que, pour des raisons de droit local, les transactions immobilières des départements du Haut-Rhin, Bas-Rhin, et de Mayotte ne sont pas présentes dans ce fichier.

Ce fichier a bien sûr fait l'objet d'une anonymisation avant publication de manière à ne pas pouvoir permettre la réidentification des personnes concernées. Toute réutilisation doit d'ailleurs respecter ce principe.

Concernant le fichier en lui-même, chaque transaction immobilière est identifiée par une parcelle cadastrale (parcelle déjà récupérée antérieurement pour chaque adresse). Une jointure permet donc de récupérer des informations pour l'ensemble des maisons de la base ayant fait l'objet d'une transaction foncière au cours des 5 dernières années. Parmi les informations récupérées figurent :

- la **valeur foncière** du cadastre lors de la dernière vente ;
- la **surface du terrain** déclarée lors de la dernière vente ;
- la **surface réelle de la maison** déclarée lors de la dernière vente ;
- le **nombre de pièces principales** déclaré lors de la dernière vente.

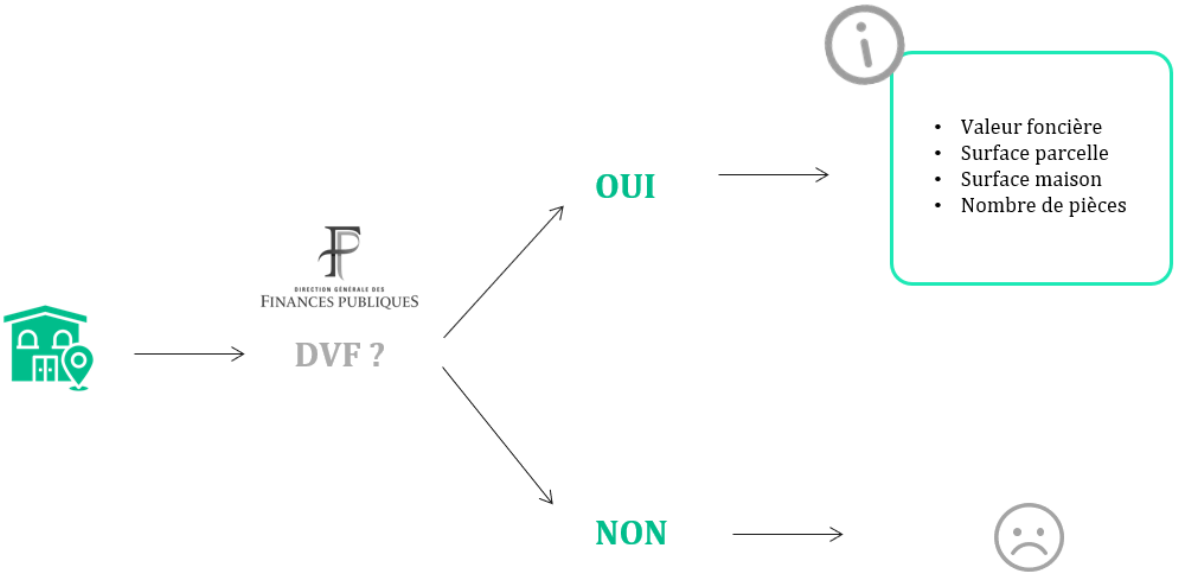


FIGURE 4.16 – Variables des demandes de valeurs foncières

Le fichier DVF, riche en informations, s'avère être de très bonne qualité. Malheureusement, pour des raisons légales, seules les informations des maisons ayant fait l'objet d'une transaction foncière ces 5 dernières années sont publiées. Bien que de très bonne qualité, ce fichier reste donc incomplet ce qui risque de s'avérer problématique pour la suite (présence de nombreux NA). Pour exemple, dans la base à l'étude, seules quelques maisons ont fait l'objet d'une transaction ces dernières années ce qui représente un total de 37 759 lignes sur les 255 075 lignes présentes dans la base. La base DVF bien que très intéressante ne pourra donc être utilisée dans le cadre d'un processus de tarification de par le nombre trop important de NA. A l'avenir, il serait cependant intéressant de voir si des partenariats ne pourraient pas être mis en place avec les notaires afin de récupérer des informations sur un périmètre plus large que les 5 dernières années.

4.1.3 Les données à la maille commune/département

Concernant les données à la maille commune, plusieurs sources ont été utilisées. Parmi elles :

- l'INSEE ;
- l'agence ORE ;
- les bases de données de data.gouv.fr ;
- Efficity.

Les données recueillies à partir de ces sites peuvent être regroupées par catégories :

FIGURE 4.17 – Variables à la maille commune/département

- **Données socio-démographiques** : elles permettent de cerner la population (nombre d'habitants, nombre de naissances/décès sur les 5 dernières années, nombre de ménages, etc) et sa répartition en fonction du sexe, de l'âge, du taux de scolarisation par classe d'âge ou encore du type de ménage (avec enfants, sans enfants, en couple, seul, etc). Des variables liées à la richesse/pauvreté de la population sont également disponibles dans cette catégorie (nombre de personnes redevables à l'impôt de solidarité sur la fortune (ISF), part des ménages imposables, part de pauvreté, niveau de vie des ménages, part d'allocataires au revenu de solidarité active (RSA)) ;
- **Données économiques** : elles sont liées à l'attractivité de la commune pour les entreprises (nombre d'entreprises, nombre d'entreprises créées sur la dernière année, etc), à la répartition de la population par catégories socioprofessionnelles (CSP) ainsi qu'au salaire net horaire moyen ;
- **Données liées aux logements** : elles représentent le marché immobilier actuel de la commune avec des informations comme le prix au m^2, les taxes (d'habitation ou foncière), les zones (Pinel ou ABC). Le nombre de logements et la répartition de ces derniers selon différents critères (maison/appartement, résidence principale/secondaire, locataire/propriétaire, nombre de pièces, période de construction, présence d'un parking ou non, année d'arrivée dans la commune, etc) sont également disponibles. Toutes ces variables peuvent s'avérer très importantes dans le cadre de l'assurance MRH ;
- **Données de quartier** : elles résument les différentes infrastructures présentes dans la commune (nombre de maternelles, d'écoles primaires, de collèges, de lycées, d'établissements supérieurs, d'hôtels, de campings, etc) ;
- **Données de criminalité** (les différents types de vols, les incendies, les attentats, etc) ;

- **Données de consommation énergétique** par secteur (agricole, industriel, tertiaire, résidentiel) tant pour le gaz que l'électricité ;
- **Données environnementales** telles la superficie, la densité, la part du territoire artificialisé, la part du territoire agricole, la part des surfaces d'eau, la part des forêts ou encore la part des zones humides.

Toutes ces données sont présentes à la maille commune à l'exception des données de criminalité qui sont, quant à elles, regroupées au niveau du département.

4.1.4 Les données météorologiques

Enfin, pour être le plus complet possible, il a également été décidé d'ajouter des données météorologiques à la base. Ces données ont été acquises et retraitées par la *Business Unit* Data Science de Sia Partners lors de travaux internes.

Le fichier fourni par la BU Data Science présente, pour chaque station météorologique à disposition, des informations sur la température, le vent, la pression, les précipitations et la neige et ce, sur plusieurs décennies. Concernant la France, aujourd'hui, une centaine de stations sont présentes dans ce fichier (sur les plus de 500 stations actives en France métropolitaine). Les stations du fichier sont réparties de la manière suivante :

FIGURE 4.18 – Stations utilisées

Afin d'utiliser ces données de manière simple, pour chaque station météo, un prétraitement des données a été effectué. Tout d'abord, la moyenne, le maximum, le minimum et la somme (lorsque cela était pertinent) de chaque variable (comme par exemple la température, les quantités de précipitations dans l'heure...) ont été calculés par année. Puis, chacune de ces grandeurs a été moyennée sur les dix dernières années. Ce travail avait déjà été réalisé lors de travaux internes.

Par la suite, il a été associé à chaque adresse la station météorologique la plus proche.

Cela a permis d'ajouter les variables suivantes dans la base :

- le numéro unique de la **station associée** ;
- la **somme annuelle des précipitations** dans la dernière heure (en mm) ;
- le **maximum annuel des précipitations** dans la dernière heure (en mm) ;
- la **moyenne annuelle des précipitations** dans la dernière heure (en mm) ;
- le **maximum annuel de température** à 2 mètres (en degrés) ;

- le **minimum annuel de température** à 2 mètres (en degrés) ;
- la **température moyenne annuelle** à 2 mètres (en degrés) ;
- la **vitesse maximale du vent** à 10 mètres (en mètres par seconde) ;
- la **vitesse moyenne du vent** à 10 mètres (en mètres par seconde) ;
- la **moyenne annuelle des chutes de neige** dans la dernière heure (en mm) ;
- le **maximum annuel des chutes de neige** dans la dernière heure (en mm) ;
- la **somme annuelle des chutes de neige** dans la dernière heure (en mm).

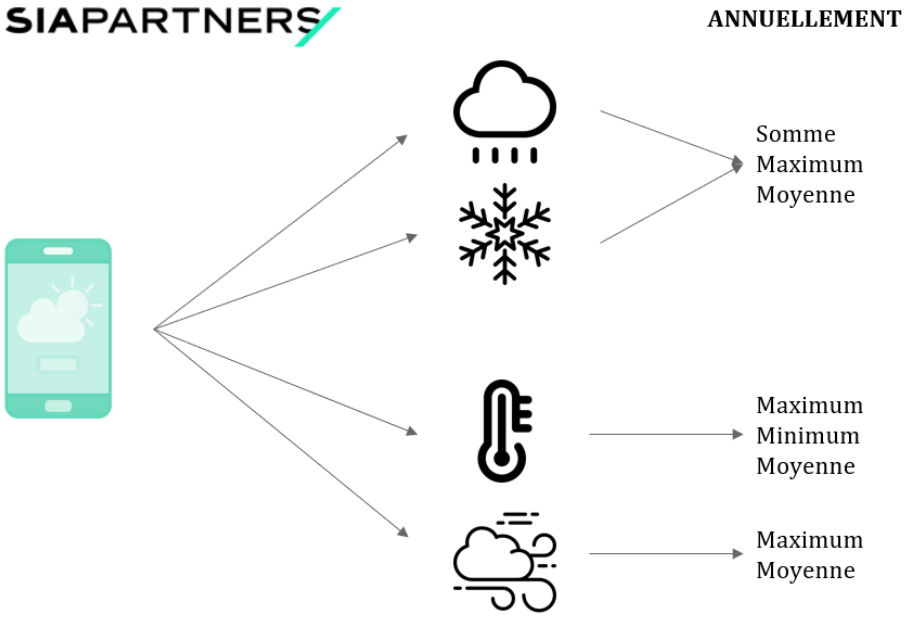


FIGURE 4.19 – Variables météorologiques

☞ **Focus sur la qualité des données météorologiques**

Concernant les données météorologiques, il est important de souligner le faible nombre de stations météorologiques disponibles dans le fichier interne à Sia Partners par rapport à l'ensemble des stations présentes sur le territoire français : l'étude présente seulement 1/5 des stations françaises. Pour aller plus loin, des données complémentaires seraient donc nécessaires, s'il est possible de les récupérer.

Il serait également intéressant d'associer les stations météorologiques aux adresses d'une manière plus fine. Pour ce faire, une compréhension des phénomènes météorologiques serait un plus. Cependant, à première vue, l'idée d'une triangularisation pourrait être testée.

4.1.5 Résumé de l'ensemble des données

Afin de mieux visualiser l'ensemble des données récupérées, un graphe a été créé.

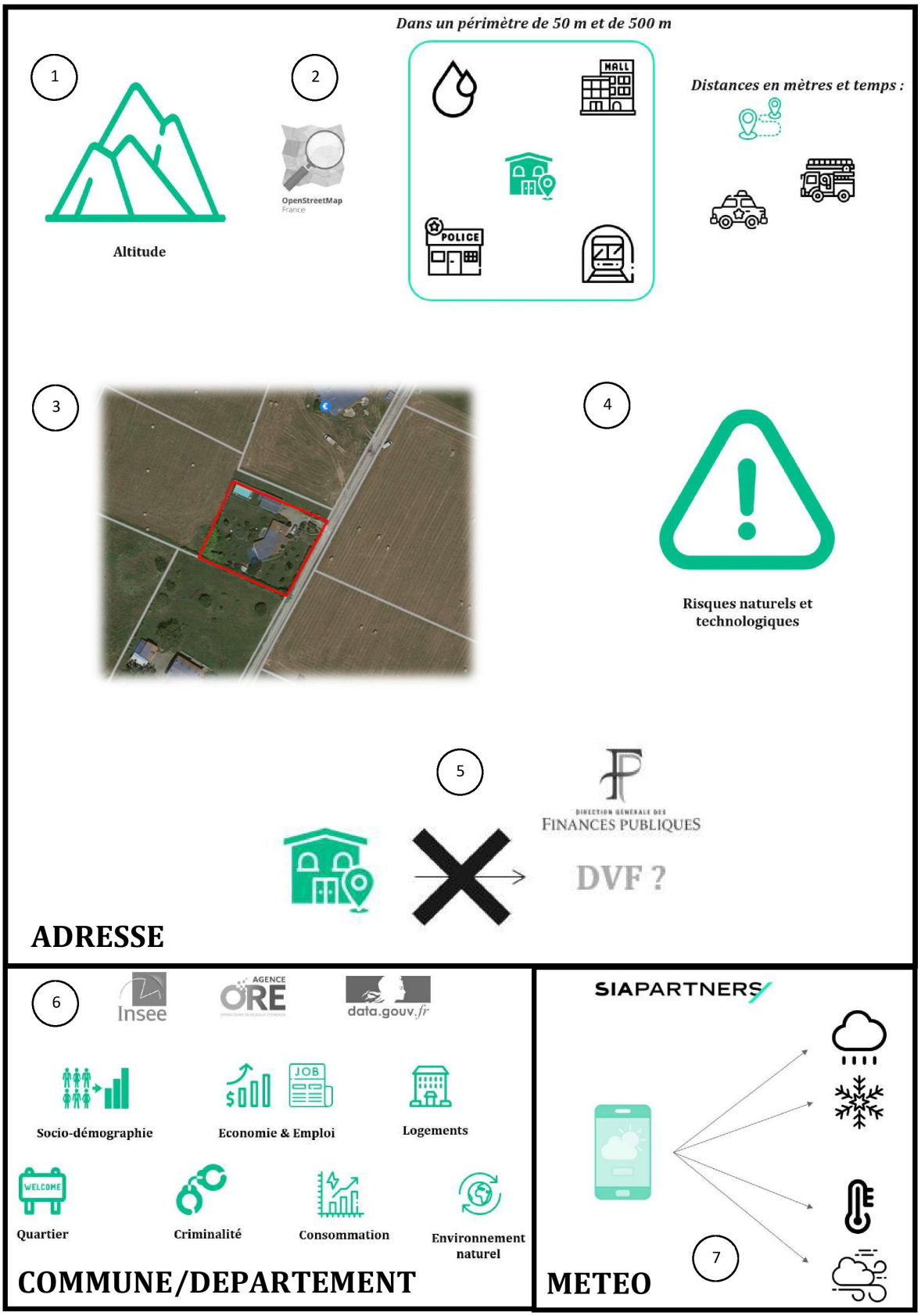


FIGURE 4.20 – Données récupérées

4.2 Les analyses univariées et bivariées

De la même manière que pour la base « assureur », une fois les données à l'adresse récupérées, des analyses univariées ont été effectuées pour la fréquence et le coût des garanties DDE et vol. En voici quelques-unes pertinentes démontrant l'intérêt des variables nouvellement recueillies.

4.2.1 L'analyse univariée de la garantie dégâts des eaux

Analyse univariée en fréquence

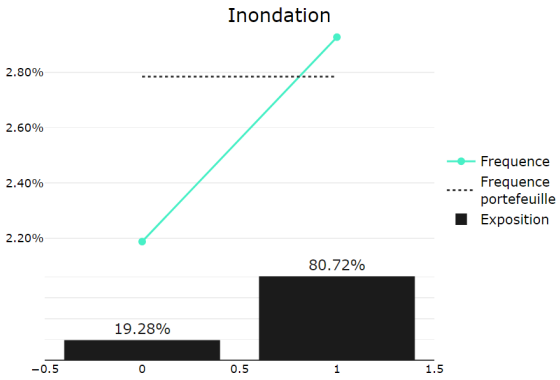


FIGURE 4.21 – Fréquence en fonction de la présence d'inondations à la commune

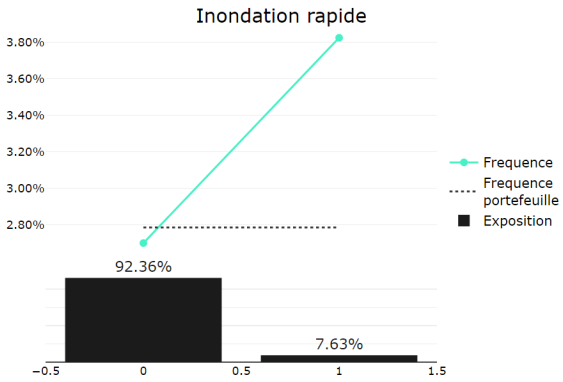


FIGURE 4.22 – Fréquence en fonction de la présence d'inondations rapides à la commune

Concernant la fréquence des dégâts des eaux, les variables inondation et inondation par crue torentielle ou montée rapide des cours d'eau (variables binaires concernant l'historique de la commune) se révèlent très tarifaires. Plus généralement, l'ensemble des variables liées aux inondations telles la présence d'un plan de prévention des risques ou encore l'appartenance à un territoire à risques importants d'inondation (TRI) a un impact non négligeable sur la fréquence. Cela reste cohérent et explique une des premières causes des dégâts des eaux : les inondations non reconnues comme catastrophes naturelles.

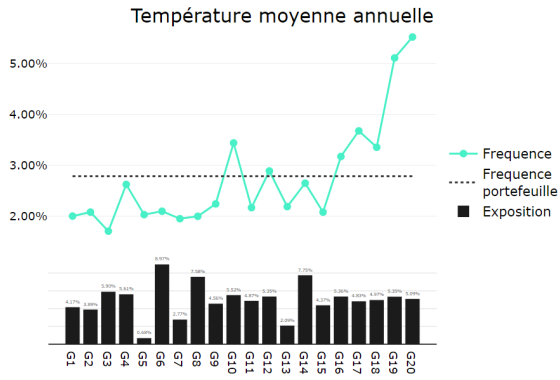


FIGURE 4.23 – Fréquence en fonction de la température moyenne annuelle à l'adresse

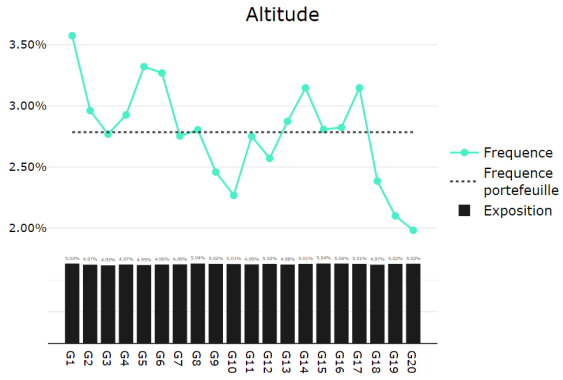


FIGURE 4.24 – Fréquence en fonction de l'altitude de l'adresse

Il est ensuite intéressant d'analyser la variable *meant*, représentant la température annuelle moyenne autour de l'adresse d'intérêt. L'effet sur le graphique est linéaire par morceaux. Ainsi, pour les températures moyennes annuelles allant de 6 degrés celsius à 13 degrés celsius (jusqu'au G15), la variable ne semble que peu influer sur la fréquence des dégâts des eaux. Cependant, pour des températures plus élevées (à partir de 13 degrés celsius/du G15), plus la température moyenne

annuelle est élevée, plus la probabilité d'un dégât des eaux augmente. Cela peut s'expliquer par le fait, qu'en période de forte chaleur (contribuant ainsi à l'augmentation de la température moyenne annuelle), de fortes pluies et d'importants orages surviennent fréquemment pouvant ainsi causer des inondations ou bien des infiltrations.

Pour l'altitude, le phénomène attendu est observé. Ainsi, plus l'altitude de la maison est élevée, moins il y a de dégâts des eaux (écoulement des eaux de pluies facilité donc moins d'inondations reconnues ou non comme catastrophes naturelles et moins d'infiltrations, etc).

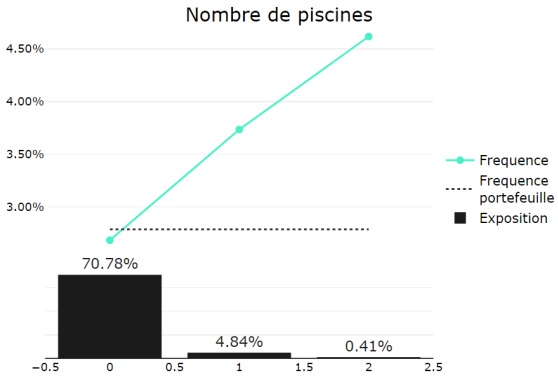


FIGURE 4.25 – Fréquence en fonction du nombre de piscines sur le terrain

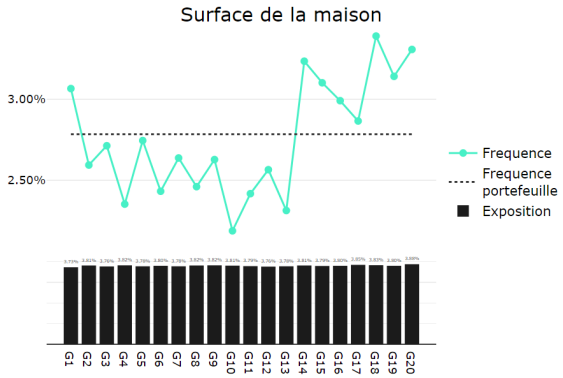


FIGURE 4.26 – Fréquence en fonction de la surface plane de la maison

Par ailleurs, le nombre de piscines présentes sur la parcelle semble également influer fortement sur la fréquence des dégâts des eaux. Il faut cependant interpréter ce graphe avec précaution vu la faible exposition des modalités 1 et 2. L'influence présumée du nombre de piscines sur la fréquence d'un dégât des eaux amène à considérer une troisième cause possible de dégâts des eaux (après les inondations et les infiltrations) : les dégâts provenant de fuites ou de ruptures de canalisations intérieures.

Enfin, pour finir, la surface de la maison présente également un effet sur la sinistralité. Cet effet est lui aussi linéaire par morceaux. Il apparaît ainsi que les maisons disposant d'une grande surface présentent davantage de risques de subir un dégât des eaux que les maisons à la surface plus modeste. Un parallèle avec la variable nombre de pièces de la base « assureur » peut être fait.

Analyse univariée en coût

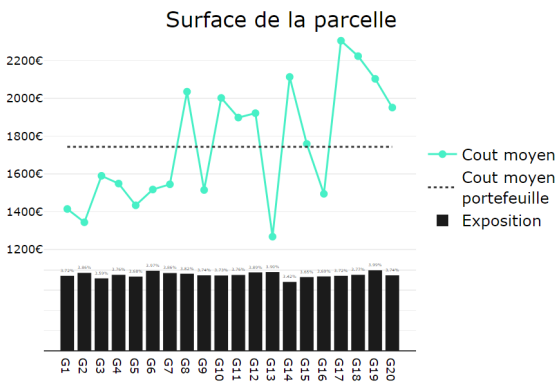


FIGURE 4.27 – Coût moyen en fonction de la surface de la parcelle

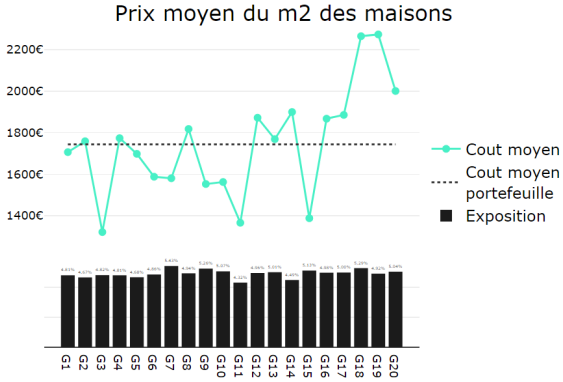


FIGURE 4.28 – Coût moyen en fonction du prix du m^2 des maisons à la commune

Concernant le coût moyen des sinistres de type dégâts des eaux, les premières variables d'intérêt sont la surface de la parcelle et le prix au m^2 des maisons de la commune.

Ainsi, pour ce qui est de la surface de la parcelle, l'effet est indéniable : plus cette dernière augmente, plus le coût d'un potentiel dégât des eaux est élevé.

En revanche, pour le prix au m^2 des maisons dans la commune, l'interprétation est plus complexe et un effet par morceaux se dessine. Ainsi, pour les maisons de moins de 2 000 euros par m^2 (G15), l'effet de la variable semble non significatif. Mais, à partir de 2 000 euros par m^2, un effet linéaire apparaît et plus le m^2 est cher, plus le coût d'un dégât des eaux est élevé.

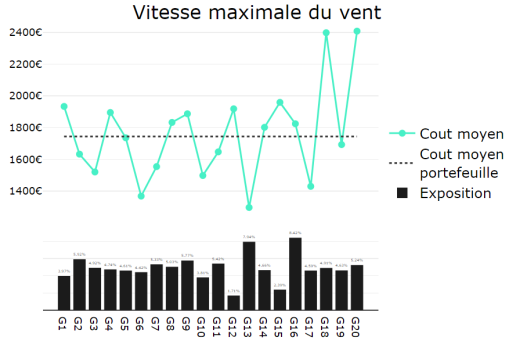


FIGURE 4.29 – Coût moyen en fonction de la vitesse maximale du vent à l'adresse

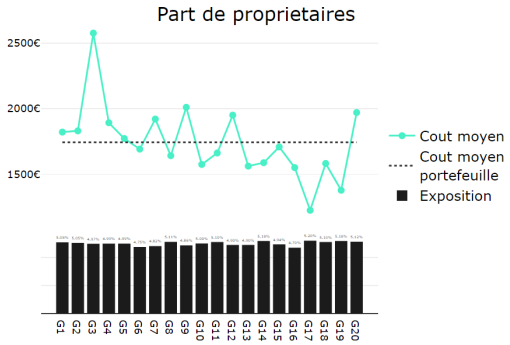


FIGURE 4.30 – Coût moyen en fonction de la part des propriétaires à la commune

Deux autres variables peuvent également être examinées : la vitesse maximale du vent autour de l'adresse considérée et la part des propriétaires de résidences principales dans la commune.

Pour le vent, un effet par morceaux est encore présent : c'est seulement à partir d'un certain seuil qu'il est possible d'attribuer une potentielle influence à cette variable. Ainsi, plus la vitesse du vent est élevée, plus le coût d'un dégât des eaux est important. Cela semble cohérent avec la réalité, un fort vent ayant tendance à amplifier tous les phénomènes naturels pouvant causer un dégât des eaux.

Enfin, concernant la part des propriétaires de résidences principales dans la commune, l'effet est clairement linéaire : moins il y a de propriétaires, moins le dégât des eaux est coûteux.

4.2.2 L'analyse univariée de la garantie vol

Analyse univariée en fréquence

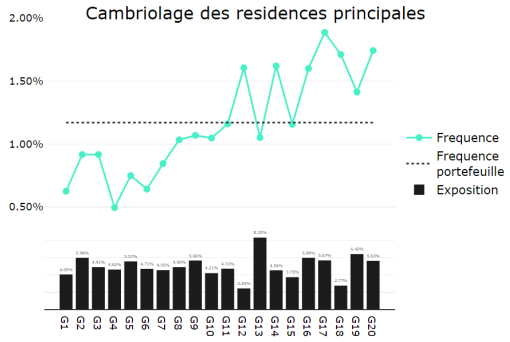


FIGURE 4.31 – Fréquence en fonction du taux de cambriolage des résidences principales au département

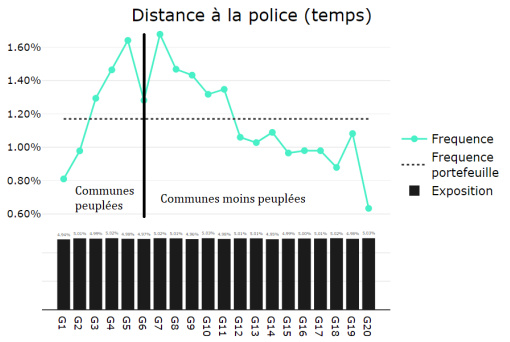


FIGURE 4.32 – Fréquence en fonction de la distance de la maison à la station de police la plus proche

Les deux premières variables impactant de manière non négligeable la fréquence des vols sont le taux de cambriolage des résidences principales à l'échelle du département et la distance de la maison considérée à la station de police la plus proche.

L'effet est très clairement linéaire pour le taux de cambriolage des résidences principales : ainsi, plus ce dernier est élevé, plus la fréquence des vols augmente.

Pour la distance de la maison considérée à la station de police la plus proche, une analyse plus poussée s'impose. Ainsi, il semble que la partie gauche du graphique puisse être associée aux communes denses et peuplées. Dans ce type de communes, de nombreux postes de police et de gendarmerie sont implantés : la distance à ces derniers est donc assez faible. A noter également qu'au sein même de ces communes, plus l'on s'éloigne des stations de police/gendarmerie, plus la fréquence des vols augmente (jusqu'au pic). En revanche, lorsque l'on s'éloigne du centre de ces communes denses, la tendance s'inverse (partie droite du graphique).

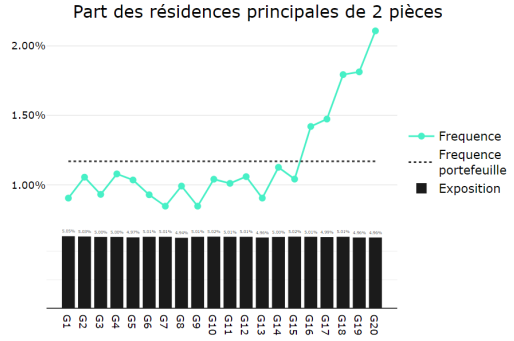


FIGURE 4.33 – Fréquence en fonction de la part des résidences principales de deux pièces à la commune

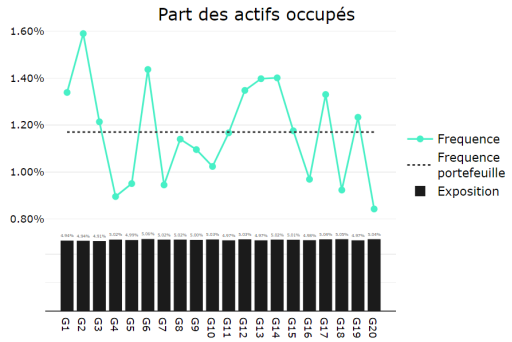


FIGURE 4.34 – Fréquence en fonction de la part des actifs occupés au niveau de la commune

Concernant la part des résidences principales de deux pièces à l'échelle de la commune, cette variable impacte également la fréquence des vols (par morceaux) et il semble que, sur une certaine portion du graphe, plus il y a de résidences principales de deux pièces, plus la fréquence des vols est élevée.

Il est également important de noter le lien entre la fréquence des vols et la part des actifs occupés. Une relation décroissante est, cette fois-ci, constatée : plus il y a d'actifs occupés, plus la fréquence des vols diminue. Autrement dit, plus il y a d'actifs non occupés, plus la fréquence des vols augmente. Cette interprétation est toutefois à prendre avec précaution. Il est en effet important de ne pas faire d'amalgames à partir de ces analyses.

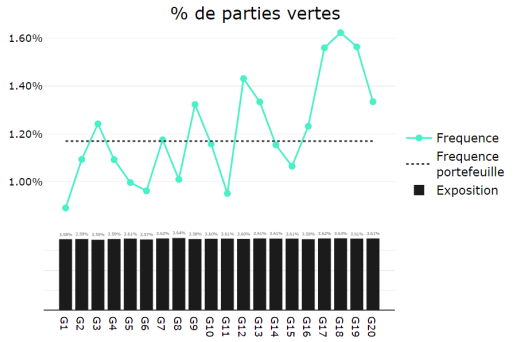


FIGURE 4.35 – Fréquence en fonction du pourcentage de parties vertes de la parcelle

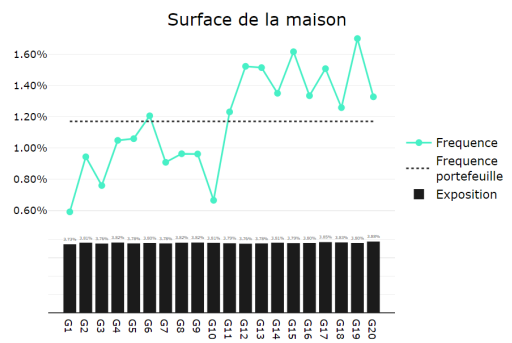


FIGURE 4.36 – Fréquence en fonction de la surface plane de la maison

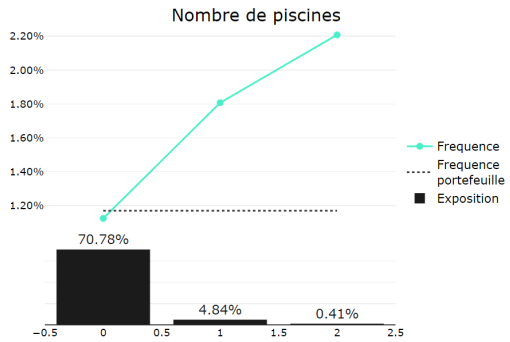


FIGURE 4.37 – Fréquence en fonction du nombre de piscines sur le terrain

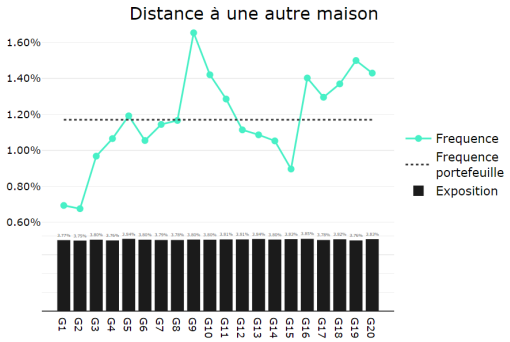


FIGURE 4.38 – Fréquence en fonction de la distance à une autre maison

De nombreuses variables externes sont également d'intérêt. Parmi elles, le pourcentage de parties vertes de la parcelle considérée ou la surface de la maison d'intérêt. Pour ces deux variables, plus la surface (de la maison ou des parties vertes) est grande, plus la fréquence des vols augmente. Le nombre de piscines influe, lui aussi, sur la fréquence des vols et augmente la probabilité de ces derniers. Enfin, l'isolement de la maison joue également un rôle important dans l'explication de la fréquence des vols. Il apparaît ainsi qu'une maison isolée aura plus de risques d'être cambriolée que les autres maisons.

Analyse univariée en coût

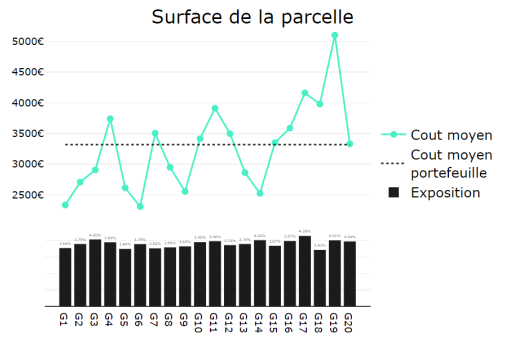


FIGURE 4.39 – Coût moyen en fonction de la surface de la parcelle

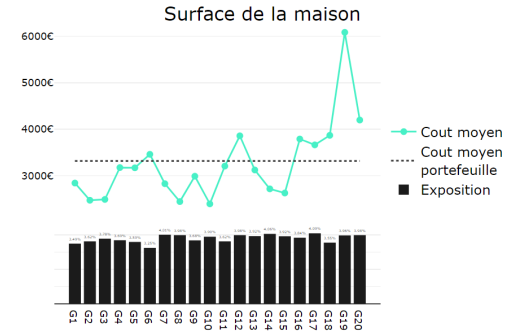


FIGURE 4.40 – Coût moyen en fonction de la surface plane de la maison

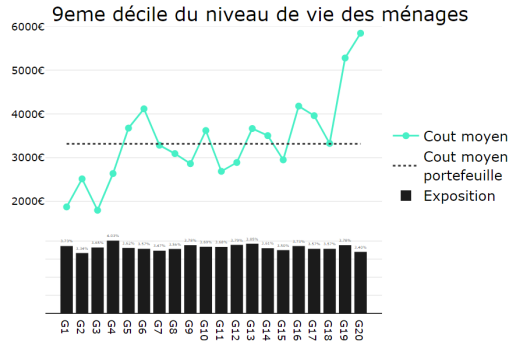


FIGURE 4.41 – Coût moyen en fonction du neuvième décile du niveau de vie des ménages à la commune

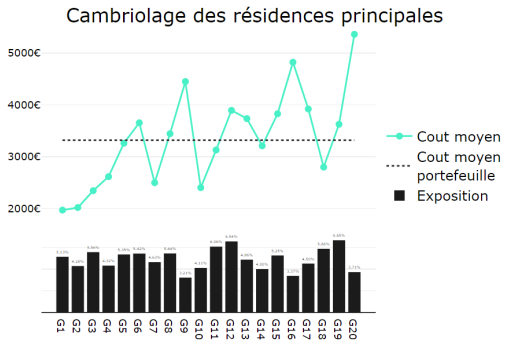


FIGURE 4.42 – Coût moyen en fonction du taux de cambriolage des résidences principales au département

Pour le coût moyen, tant les variables surfaciques (de la maison ou de la parcelle) que le niveau de vie des ménages impactent grandement le coût, et ce de manière croissante. Quant au taux de cambriolage des résidences principales à l'échelle du département, il influence, lui aussi, ce coût moyen qui croît au fur et à mesure que la part des cambriolages augmente.

4.2.3 Les analyses bivariées

En ce qui concerne les analyses bivariées, devant l'ampleur du nombre de variables, ces dernières n'ont pas été effectuées de manière exhaustive pour tous les couples. Elles ont cependant été réalisées pour certains couples de variables jugés plus intéressants que les autres. Afin d'alléger le contenu du rapport, ces analyses ne seront pas présentées ici.

4.3 Les corrélations entre les différentes variables

Le nombre de variables externes étant très important, il est également nécessaire de s'intéresser à la corrélation entre ces dernières afin d'éviter tout problème lors de la modélisation de la fréquence et du coût.

Différentes actions ont donc été entreprises afin de supprimer ces corrélations. Il a notamment été décidé de :

- supprimer une des deux variables lorsque la corrélation du couple était supérieure à 0.7. C'est la variable la moins intuitive qui a été supprimée à chaque fois. Pour citer quelques exemples :
  - la variable classe de densité a été supprimée, cette dernière étant étroitement liée avec la densité ;
  - la variable numéro de station météorologique a également été retirée, l'information étant redondante avec les données météorologiques ;
  - la variable zone Pinel a été éliminée, ces zones étant déterminées à partir du zonage immobilier ABC qui est, quant à lui, conservé ;
  - la variable police par population a aussi été supprimée de par son lien étroit avec la distance de la maison à la station de police la plus proche ;
  - les différentes variables surfaciques de la maison issues de la reconnaissance d'images ont été éliminées : seule l'approximation convexe de la surface de la maison a été conservée.
- supprimer une variable au sein d'un groupe de variables liées entre elles afin d'éviter des problèmes lors de l'implémentation des GLM. Par exemple, la part d'hommes et de femmes à la commune étant lié (la somme de ces deux variables fait 1), une des deux variables est donc supprimée. De la même manière, la part des 0-14 ans à la commune, la part des appartements à la commune, la part des logements d'une pièce à la commune, la part des locataires à la commune, la part des logements vacants à la commune, la part des agriculteurs à la commune ou encore la part des ménages seuls à la commune ont été retirées.

La taille de la matrice de corrélation ne permet malheureusement pas de l'afficher dans ce rapport.

Chapitre 5

Mise en place des méthodes de tarification

Une fois la base de données constituée, différents modèles vont être mis en place afin de tarifier les garanties DDE et vol d'un contrat d'assurance MRH. Pour chaque garantie, trois modèles vont ainsi être construits. Le premier modèle reprendra la méthode de tarification traditionnelle : un GLM fréquence-coût sera calibré sur la base « assureur » ne contenant aucune variable externe. C'est ce modèle qui constituera par la suite le modèle de référence auquel les deux autres modèles, à l'adresse cette fois-ci, seront comparés. Le deuxième modèle, à l'adresse, reposera, lui aussi, sur une tarification GLM fréquence-coût tandis que le troisième sera une forêt aléatoire à l'adresse modélisant séparément la fréquence et le coût moyen. Ces deux derniers modèles seront calibrés uniquement à l'aide des données à l'adresse (c'est-à-dire à l'aide des données externes évoquées au chapitre 4 mais aussi du zonier interne de l'assureur). Une comparaison entre ces modèles sera ensuite effectuée.

5.1 Le modèle de tarification traditionnelle

Ce premier modèle représente la tarification traditionnelle effectuée par les assureurs et est, en tant que tel, construit uniquement à partir des variables déclaratives dont dispose l'assureur, c'est-à-dire à partir des variables présentes dans la base « assureur » (aucune variable externe n'est donc ajoutée). La tarification se fera selon une approche GLM fréquence-coût, cette dernière étant la plus courante dans le secteur de la tarification non-vie. Une fois construit, c'est ce modèle qui servira de référence pour les comparaisons avec les modèles à l'adresse.

A l'origine, l'ensemble des variables déclaratives présentes dans la base « assureur » est déjà renseigné sous forme de classe. Aucune discrétisation de variables numériques ne doit donc être entreprise, ces dernières ayant déjà fait l'objet de traitements en interne. Certains regroupements de modalités pourront en revanche être effectués au fur et à mesure de la modélisation. Deux effets pourront être à l'origine de tels regroupements : la faible exposition de certaines modalités ou encore l'influence similaire de certaines modalités sur la variable d'intérêt (le nombre de sinistres ou le coût moyen). Il est à noter que ces regroupements ne seront pas forcément les mêmes en fonction du modèle (de fréquence ou de coût moyen).

Pour une garantie donnée, différentes étapes ont été nécessaires afin de construire les modèles de fréquence et de coût moyen. Une méthodologie commune a été adoptée, à savoir :

- 1) Analyse de l'adéquation des lois aux données et choix d'une loi : pour la modélisation du nombre de sinistres, les lois Poisson et négative binomiale ont été considérées et, pour la modélisation du coût moyen, ce sont les lois Gamma et log-normale qui ont été testées ;
- 2) Implémentation d'un premier modèle avec toutes les variables à disposition ;

- 3) Processus de sélection de variables : celle-ci a été réalisée avec la procédure *stepAIC* selon la méthode *stepwise* explicitée dans la partie 3.2.4 ;
- 4) Implémentation d'un modèle avec les variables retenues dans le *stepAIC* ;
- 5) Etude des possibilités de regroupements de modalités en fonction des résultats et de l'expérience métier. En cas de regroupements, implémentation du nouveau modèle ;
- 6) Validation et analyse du modèle final obtenu.

Dans le cadre de ce mémoire, et par souci de parcimonie, toutes ces étapes ne seront pas détaillées bien qu'elles aient toutes été réalisées : seuls les modèles optimisés seront présentés.

5.1.1 La modélisation de la garantie dégâts des eaux

En ce qui concerne la garantie DDE, c'est la loi Poisson qui a été choisie pour modéliser le nombre de sinistres et la loi Gamma pour le coût moyen.

Chacun des modèles (fréquence et coût moyen) a été calibré sur une base d'apprentissage et validé sur une base de test. Les effectifs de ces bases sont les suivants :

DDE	Fréquence	Coût
Base d'apprentissage	178 552	2 371
Base de test	76 523	942

FIGURE 5.1 – Répartition du nombre de lignes entre les bases d'apprentissage et de test pour les GLM traditionnels de la garantie dégâts des eaux

Un modèle complet a ensuite été calibré, d'une part pour la fréquence, d'autre part pour le coût moyen, avec les variables retenues dans l'analyse graphique présentée en partie 2.3.1. Seule la variable superficie a été écartée de l'étude pour être remplacée par le nombre de pièces, variable qui présente globalement une information plus fidèle que la superficie (qui a fait l'objet de nombreuses imputations).

Dans ces modèles complets, il est apparu, tant pour la fréquence que pour le coût moyen, que certaines variables n'étaient pas très significatives. Une sélection de variables a donc été entreprise. Les variables retenues pour l'explication de la fréquence et du coût moyen sont, après sélection, les suivantes :

Modèle	Fréquence	Coût
Qualité	X	X
Age	X	
Année de construction		
Nombre de personnes		
Nombre de pièces	X	X
Résidence secondaire	X	
Capital dépendance		
Capital mobilier	X	X
Capital bijou		
Sinistralité passée	X	
Franchise		
Zonier	X	
Formule		
Indicateur de recours		
Nombre de contrats		
Année		
Nombre de dépendances	X	

FIGURE 5.2 – Variables retenues dans les GLM traditionnels de la garantie dégâts des eaux

Outre cette sélection de variables, des regroupements de modalités sont effectués.

Toutes ces étapes ont permis de passer d'une RMSE de *0.11446195* à une RMSE de *0.11441531* pour la fréquence et d'une RMSE de *2367.92* à une RMSE de *2357.59* pour le coût moyen, démontrant ainsi l'utilité de la sélection de variables réalisée : le modèle devient plus parcimonieux sans pour autant altérer son pouvoir prédictif.

Une analyse plus fine des coefficients issus des modèles de fréquence et de coût moyen est maintenant proposée sous forme de graphiques. L'intervalle de confiance associé à chaque coefficient figure également sur ces derniers graphes.

Fréquence

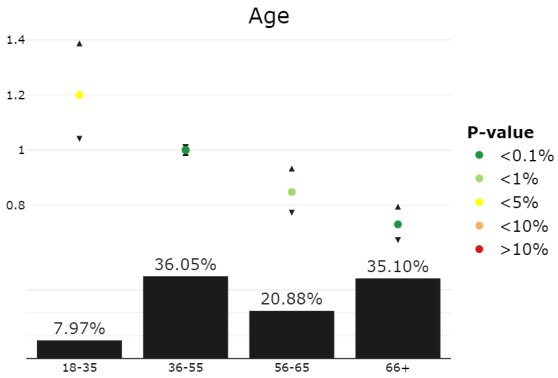


FIGURE 5.3 – Coefficients des classes d'âge

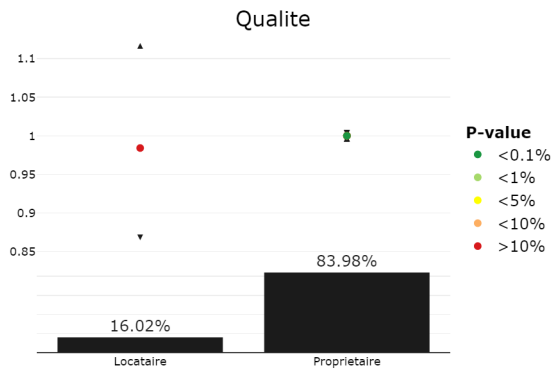


FIGURE 5.4 – Coefficients de la qualité

Les coefficients associés aux différentes classes d'âge sont ainsi tous significatifs. Aucune remarque particulière n'est donc nécessaire. En revanche, pour la variable qualité, le test de Student ne conclut pas à une grande significativité de la variable. Cela peut s'expliquer par le faible nombre de locataires vivant dans une maison dans le portefeuille considéré. Cette variable ayant cependant, la plupart du temps, un impact non négligeable en assurance MRH, elle sera toutefois conservée par la suite.

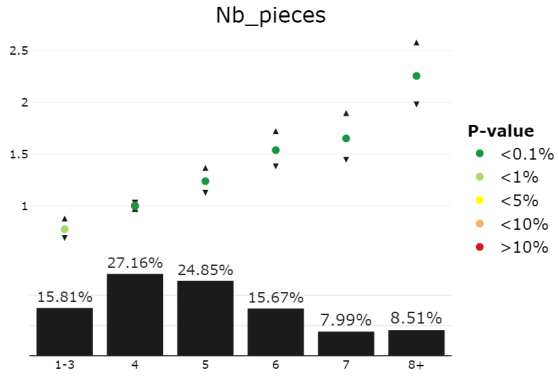


FIGURE 5.5 – Coefficients du nombre de pièces

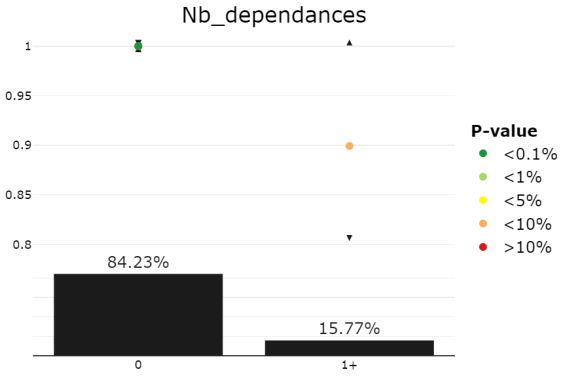


FIGURE 5.6 – Coefficients du nombre de dépendances

Concernant le nombre de pièces, là encore, aucune remarque particulière n'est utile, les coefficients évoluant de manière cohérente. Cependant, pour le nombre de dépendances, l'impact de la variable est moins évident.

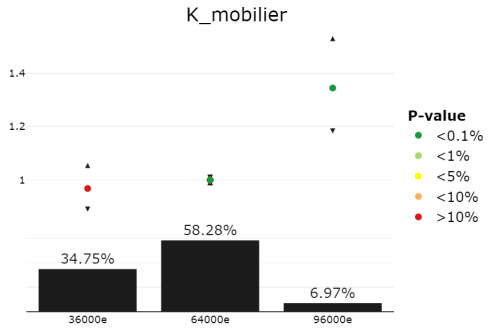


FIGURE 5.7 – Coefficients liés au capital mobilier

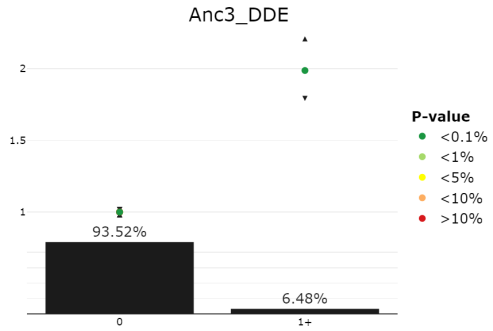


FIGURE 5.8 – Coefficients de la sinistralité passée

Les coefficients associés au capital mobilier sont, quant à eux, tous significatifs hormis celui de la modalité 36000e. Il serait donc possible d'envisager, pour cette variable, un autre regroupement de modalités. Cependant, en faire davantage nuirait possiblement au modèle et entraînerait de l'anti-sélection. Cette variable est donc laissée en l'état. Quant à la variable sinistralité passée, elle revêt, quant à elle, tout son intérêt.

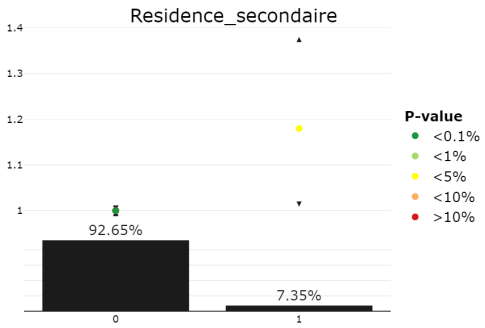


FIGURE 5.9 – Coefficients résidence secondaire

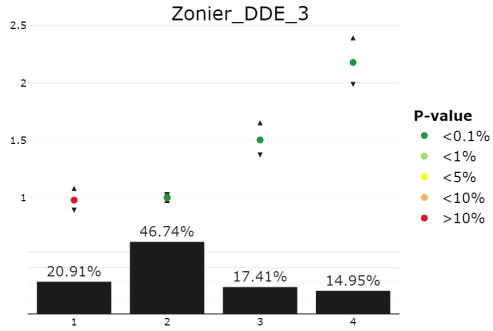


FIGURE 5.10 – Coefficients des modalités du zonier

Enfin, pour les coefficients associés à la variable résidence secondaire et au zonier, toutes les modalités semblent globalement significatives. Un doute peut subsister pour la modalité 1 du zonier. Cependant, le zonier provenant d'études internes antérieures, il a été décidé de ne pas le modifier : aucun regroupement n'a donc été réalisé.

Coût moyen

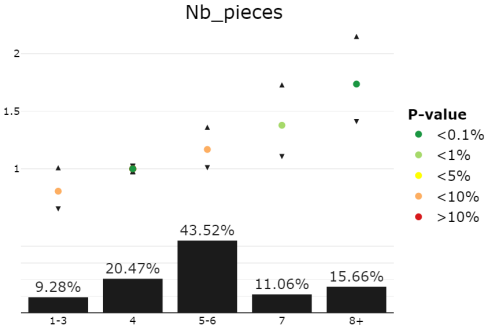


FIGURE 5.11 – Coefficients du nombre de pièces

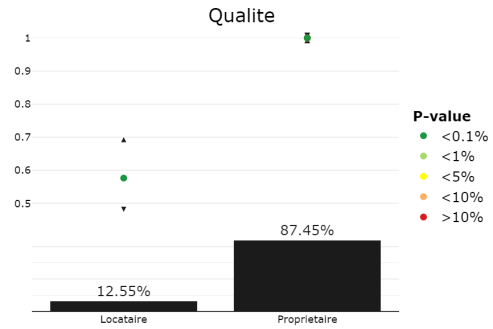


FIGURE 5.12 – Coefficients associés à la qualité

Pour la qualité et le nombre de pièces du modèle de coût moyen, les coefficients des différentes modalités sont globalement tous significatifs.

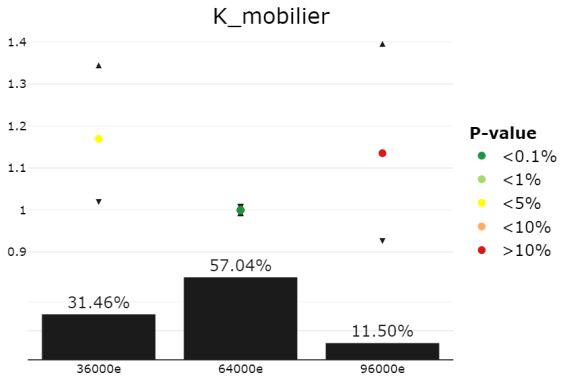


FIGURE 5.13 – Coefficients capital mobilier

Concernant le capital mobilier, une modalité (96000e) n'est pas significative. Cependant, la logique voudrait que le coût moyen d'un DDE varie avec le capital mobilier assuré : cette variable est donc conservée par avis d'experts. Cela permettra également d'éviter des problèmes d'anti-sélection sur ce segment.

Après analyse des coefficients, ces deux modèles (fréquence et coût moyen) ont également fait l'objet d'une étude de robustesse (sur la base de test). Les modèles sont apparus comme étant globalement robustes. Cependant, par souci de parcimonie, cette étude ne sera pas détaillée dans ce rapport. Des exemples de graphe réalisé sont tout de même présents ci-dessous :

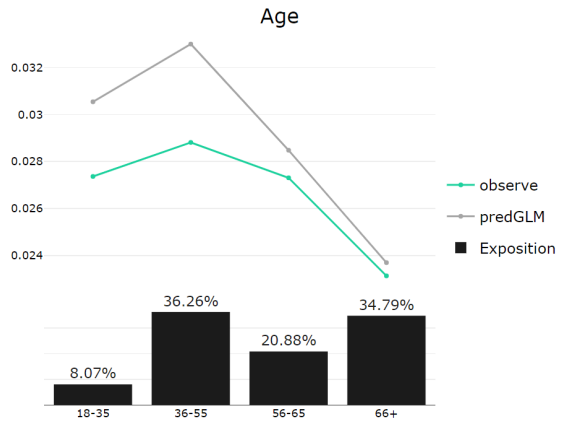


FIGURE 5.14 – Robustesse de l'âge pour la fréquence (sur la base de test)

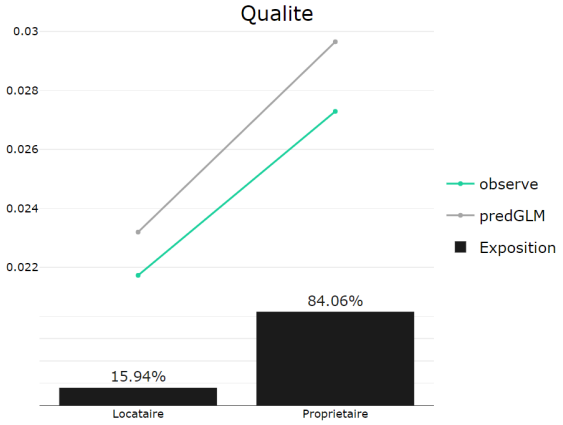


FIGURE 5.15 – Robustesse de la qualité pour la fréquence (sur la base de test)

Au global

Enfin, il est également possible d'analyser, de manière plus globale, les modèles précédemment construits. Ainsi, après application du modèle de fréquence et du modèle de coût moyen sur la base de test, la fréquence moyenne des sinistres et le coût moyen de ces derniers peuvent être calculés. Le coût total des sinistres peut alors être estimé. Toutes ces grandeurs sont ensuite comparables à la réalité connue de la base de test.

	Base d'apprentissage			Base de test		
	Réalité	*Prédiction*	*Erreur (%)*	*Réalité*	*Prédiction*	*Erreur (%)*
Fréquence	2,8463%	2,8463%	**0,00%**	2,6403%	2,8621%	**8,40%**
Coût moyen	1 739,9 €	1 738,3 €	**-0,09%**	1 752,9 €	1 720,5 €	**-1,85%**
Coût total	4 243 715 €	4 241 573 €	**-0,05%**	1 696 835 €	1 830 381 €	**7,87%**

FIGURE 5.16 – Prédictions avec les GLM traditionnels pour la garantie dégâts des eaux

Enfin, une mesure de la RMSE sur la base de test est également effectuée pour comparaison ultérieure avec les modèles à l'adresse.

	RMSE base d'apprentissage	RMSE base de test
Fréquence	0,1188	0,1144
Coût moyen	3 019,96	2 357,59
Coût total	**410,0700**	**364,5822**

FIGURE 5.17 – RMSE des GLM traditionnels pour la garantie dégâts des eaux

5.1.2 La modélisation de la garantie vol

Une analyse similaire à celle réalisée pour la garantie dégâts des eaux est effectuée pour la garantie vol.

Le nombre de sinistres est, là encore, modélisé par le biais d'une loi de Poisson et le coût moyen par le biais d'une loi Gamma. Les bases d'apprentissage et de test sont, cette fois-ci, constituées des effectifs suivants :

Vol	Fréquence	Coût
Base d'apprentissage	178 552	943
Base de test	76 523	407

FIGURE 5.18 – Répartition du nombre de lignes entre les bases d'apprentissage et de test pour les GLM traditionnels de la garantie vol

Ainsi, concernant la fréquence, aucun changement n'a lieu sur aucune des deux bases (apprentissage et test). Cependant, pour le coût moyen, il est possible de noter une moindre sinistralité des vols par rapport aux DDE.

Après réalisation des modèles complets, une sélection de variables a été entreprise conduisant à retenir les variables suivantes dans le modèle final :

Modèle	Fréquence	Coût
Qualité	X	X
Age		
Année de construction	X	
Nombre de personnes		
Nombre de pièces	X	X
Résidence secondaire	X	
Capital dépendance		X
Capital mobilier		
Capital bijou	X	X
Sinistralité passée	X	
Franchise		
Zonier	X	X
Formule		
Indicateur de recours		
Nombre de contrats		
Année		
Nombre de dépendances		

FIGURE 5.19 – Variables retenues dans les GLM traditionnels de la garantie vol

La RMSE du modèle de fréquence est ainsi passée de *0.07619188* à *0.07618394* et celle du modèle de coût moyen de *4707.35* à *4666.44*. Les modèles ont donc été simplifiés sans perte de pouvoir prédictif.

Il est maintenant possible d'analyser les différents coefficients issus des modèles de fréquence et de coût moyen ainsi que leur intervalle de confiance associé.

Fréquence

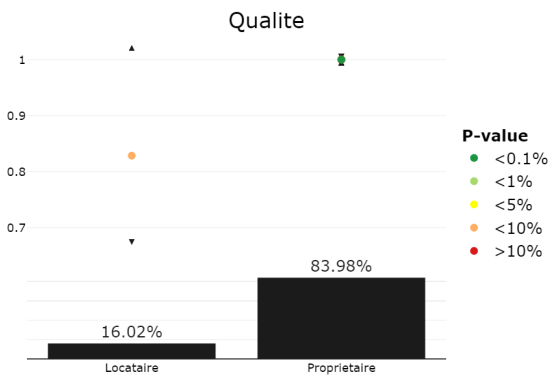


FIGURE 5.20 – Coefficients associés à la qualité

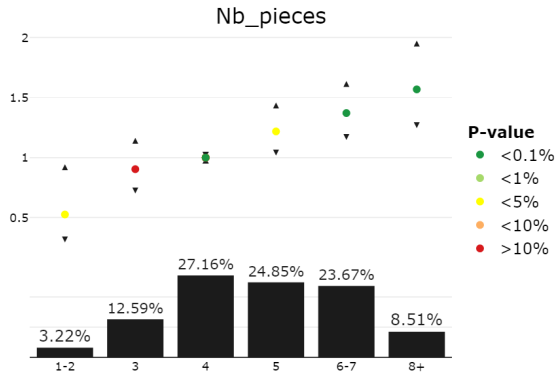


FIGURE 5.21 – Coefficients du nombre de pièces

La significativité de la variable qualité peut, une nouvelle fois, être remise en cause par le graphique ci-dessus. Cependant, pour les mêmes raisons qu'évoquées précédemment lors de l'analyse des coefficients de la garantie DDE, cette variable est conservée.

Concernant le nombre de pièces, l'ensemble des modalités est globalement significatif à l'exception de la modalité 3. L'évolution des coefficients est cependant très cohérente avec la réalité et, afin d'éviter l'anti-sélection sur ce segment, aucun regroupement n'est effectué.

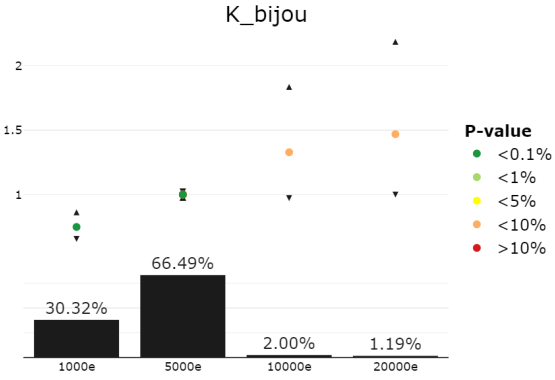


FIGURE 5.22 – Coefficients liés au capital bijou

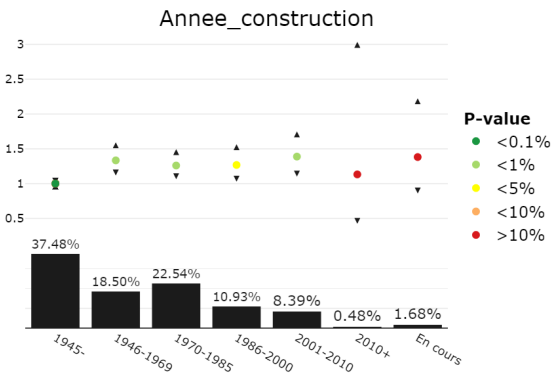


FIGURE 5.23 – Coefficients des années de construction

En ce qui concerne le capital bijou et l'année de construction, les modalités sont, là encore, toutes significatives à l'exception des modalités peu exposées comme les capitaux élevés (10000e ou 20000e) ou les années de construction récentes (2010+ ou en cours). Des regroupements auraient pu être envisagés. Cependant, il est supposé, ici, que l'absence de significativité des coefficients provient de la taille de la base qui est relativement faible et que, sur une base plus conséquente, ces modalités auraient pu prendre tout leur sens. C'est pourquoi, les regroupements ont été écartés.

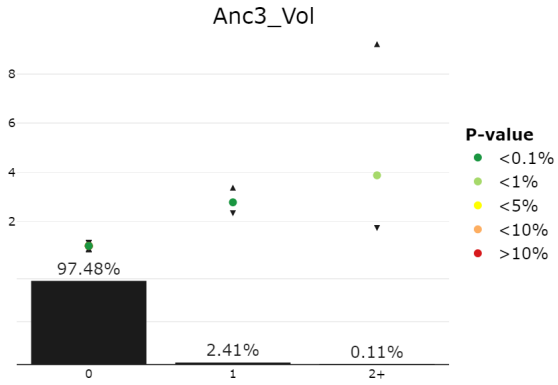


FIGURE 5.24 – Coefficients de la sinistralité passée

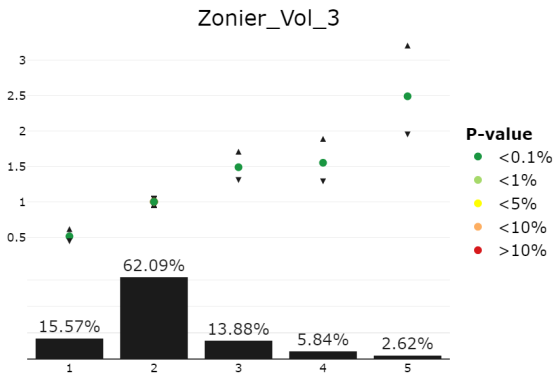


FIGURE 5.25 – Coefficients liés aux modalités du zonier

Les variables zonier et sinistralité passée ne soulèvent, quant à elles, pas de commentaire particulier, ces dernières apparaissant très significatives dans l'explication de la fréquence des vols.

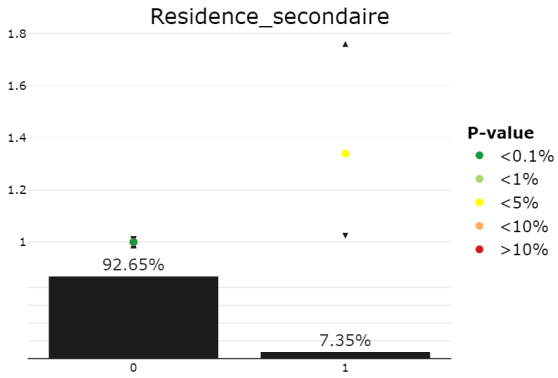


FIGURE 5.26 – Coefficients résidence secondaire

Enfin, la présence de résidences secondaires a un impact significatif sur la fréquence des vols ce qui reste cohérent avec la réalité.

Coût moyen

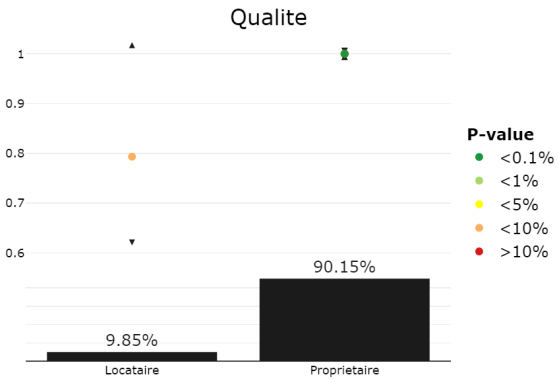


FIGURE 5.27 – Coefficients associés à la qualité

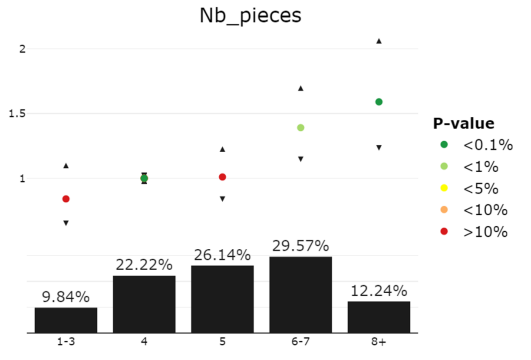


FIGURE 5.28 – Coefficients du nombre de pièces

Concernant le modèle de coût moyen des vols, la qualité apparaît, cette fois-ci, significative.

Pour le nombre de pièces, l'évolution des coefficients reste globalement cohérente bien que certains coefficients ne soient pas significatifs.

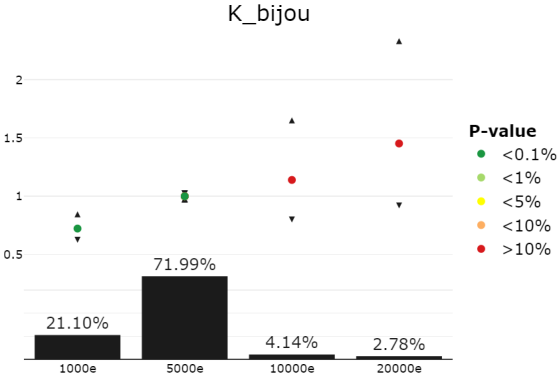


FIGURE 5.29 – Coefficients liés au capital bijou

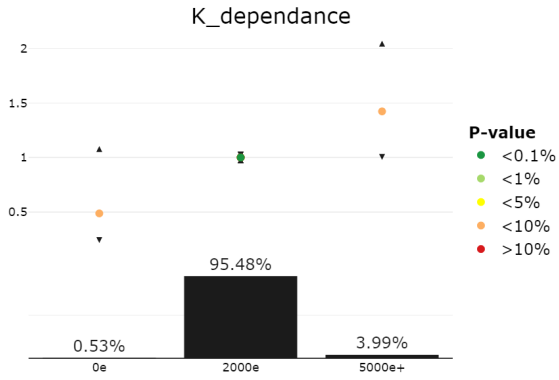


FIGURE 5.30 – Coefficients liés au capital dépendance

Pour les capitaux (bijou et dépendance), les mêmes remarques que précédemment peuvent être effectuées.

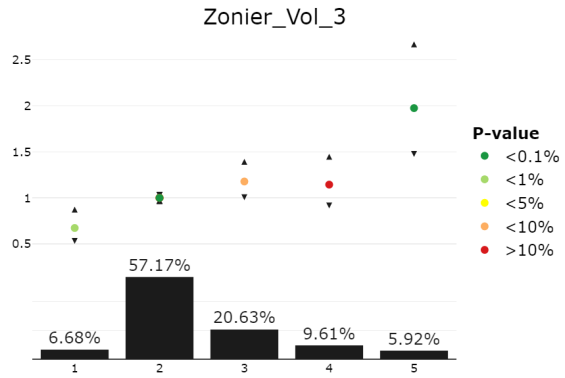


FIGURE 5.31 – Coefficients liés aux modalités du zonier

Enfin, le zonier, bien que constitué de certaines modalités non significatives, est cependant intégré au modèle. En effet, ce dernier ayant fait l'objet de travaux en interne, aucun regroupement n'est envisagé.

Au global

En analysant les modèles de manière plus globale, les résultats suivants sont obtenus :

	Base d'apprentissage			Base de test		
	Réalité	*Prédiction*	*Erreur (%)*	*Réalité*	*Prédiction*	*Erreur (%)*
Fréquence	1,1693%	1,1693%	**0,00%**	1,1729%	1,1748%	**0,16%**
Coût moyen	3 441,9 €	3 429,7 €	**-0,35%**	3 505,2 €	3 445,5 €	**-1,70%**
Coût total	3 293 880 €	3 464 172 €	**5,17%**	1 458 144 €	1 494 120 €	**2,47%**

FIGURE 5.32 – Prédictions avec les GLM traditionnels pour la garantie vol

Les RMSE du modèle final sont :

	RMSE base d'apprentissage	RMSE base de test
Fréquence	0,0757	0,0762
Coût moyen	4 468,86	4 666,44
Coût total	**436,5088**	**439,4151**

FIGURE 5.33 – RMSE des GLM traditionnels pour la garantie vol

5.2 Le GLM à l'adresse

Le premier modèle à l'adresse repose sur une approche GLM fréquence-coût. Ce faisant, une méthodologie similaire à celle présentée précédemment pour le modèle de tarification traditionnelle est utilisée.

Il existe cependant une différence notable en ce qui concerne les données utilisées. En effet, là où la méthode de tarification traditionnelle ne prend en compte que les variables déclaratives dont dispose l'assureur, la tarification à l'adresse va, quant à elle, utiliser uniquement les données récupérables à partir de l'adresse de l'assuré : c'est-à-dire l'ensemble des données récupérées en *Open Data* dans le chapitre 4 et le zonier interne (qui dépend de la localisation de l'assuré).

L'utilisation de ces données en *Open Data* amène donc à considérer de nouvelles problématiques.

5.2.1 Problématiques liées à l'utilisation de l'*Open Data*

➥ **Discrétisation des variables continues**

Lors de la récupération des données en *Open Data*, un grand nombre de données continues ont été recueillies. Or, il est courant, en tarification non-vie, de raisonner non pas sur la variable continue en elle-même mais sur des classes de cette dernière. L'objectif est alors de modéliser la non-linéarité de l'effet de la variable (par exemple, en tarification, l'âge n'est quasiment jamais utilisé directement dans les modèles mais des classes d'âge sont construites afin de prendre en compte la non-linéarité de cette variable). Pour revenir aux variables continues de l'*Open Data*, deux cas de figure sont donc à considérer :

- ☛ Soit la variable continue d'intérêt présente un effet plutôt linéaire dans l'analyse univariée présentée en section 4.2, auquel cas la variable n'a pas besoin d'être discrétisée ;
- ☛ Soit l'effet n'est clairement pas linéaire dans l'analyse univariée (section 4.2) et, dans ce cas, une discrétisation doit être effectuée.

En considérant le nombre de variables continues concernées par le deuxième cas de figure, une procédure automatique de discrétisation doit être mise en place. Pour ce faire, il est possible d'utiliser les quantiles de la variable en question. Ainsi, pour chaque variable continue présentant un effet non linéaire, la variable a été découpée en une vingtaine de classes délimitées par ses quantiles. De cette manière, les classes nouvellement construites sont assez équilibrées entre elles. Dans le cadre de ce mémoire, c'est cette solution qui a été retenue. Il est à noter cependant que d'autres solutions auraient, bien sûr, pu être mises en place.

Par ailleurs, il est important d'appeler l'attention sur un cas particulier rencontré plusieurs fois au cours de l'étude : le cas où l'effet de la variable continue sur la fréquence ou le coût moyen est composé de deux droites linéaires (linéarité par morceaux avec deux droites). Dans ce cas de figure, la variable continue peut être décomposée en deux autres variables continues en fonction d'un seuil.

Pour exemple, en considérant l'analyse univariée de la surface de la maison sur la fréquence d'un DDE (figure 4.26) il est possible d'observer un effet linéaire par morceaux qui peut être reproduit à l'aide de deux droites linéaires. Deux variables vont donc être créées à partir de la variable initiale : l'une prendra en compte l'effet de la surface lorsque cette dernière est inférieure à 125 m^2 (avant G10) et l'autre, l'effet de la surface lorsqu'elle est supérieure à 125 m^2 (après G10). Ainsi, si la surface de la maison est de 100 m^2, la première variable sera fixée à 100 m^2 et la deuxième sera, quant à elle, de 125 m^2. En revanche, si la surface de la maison est de 140 m^2, la première variable aura pour valeur 125 m^2 et la deuxième, 140 m^2.

➥ **Présence de nombreuses valeurs manquantes**

Une autre problématique peut être soulevée : le sujet des valeurs manquantes.

Ainsi, certaines variables récupérées présentent, pour diverses raisons, de nombreuses valeurs manquantes (du fait d'une indisponibilité de la donnée ou encore de problèmes de confidentialité). Ces données ne sont pas utilisables en l'état : il a donc fallu se séparer des variables présentant un nombre trop important de valeurs manquantes. C'est d'ailleurs la raison pour laquelle la base DVF, présentée en sous-section 4.1.2, n'a pas été conservée pour l'étude (plus de 217 000 valeurs manquantes étant présentes).

Les variables présentant un nombre raisonnable de valeurs manquantes ont, quant à elle, été conservées. Elles ont fait l'objet d'imputations simples lorsque cela était possible. Ainsi, par exemple :

- ☛ les variables premier et neuvième décile du niveau de vie des ménages à la commune présentaient quelques valeurs manquantes (82 631 pour chacune des variables soit environ 30% des valeurs). Une forte corrélation linéaire (environ 0.9) a toutefois été observée entre ces variables et la médiane du niveau de vie des ménages à la commune. Cette observation a été utilisée afin de construire un modèle linéaire simple prédisant le premier/neuvième décile du niveau de vie des ménages à partir du niveau de vie médian de ces derniers. Cela a permis d'imputer une valeur aux quelques NA présents pour ces variables ;
- ☛ la variable représentant le prix du m^2 des maisons à la commune présentait, elle aussi, un certain nombre de valeurs manquantes (précisément 5 058 soit 1% des valeurs). Afin d'imputer une valeur à ces NA, il a été décidé d'utiliser le prix médian du m^2 des communes avoisinantes.

Toutefois, les valeurs manquantes de certaines variables n'ont pu être imputées par manque de corrélations significatives avec d'autres variables. Une autre solution consiste à imputer ces valeurs manquantes par une simple moyenne ou médiane. Cependant, pour ces dernières variables, cette solution n'apparaissait pas satisfaisante (par exemple, pour la surface des maisons issue de la reconnaissance d'images). Aussi, pour calibrer de manière correcte les différents modèles, il a été décidé que seules les lignes ne contenant aucune valeur manquante seraient conservées lors de la modélisation. Cela a eu pour effet de réduire tant la base d'apprentissage que la base de test utilisées auparavant lors de la tarification traditionnelle : une perte de 57 700 lignes a ainsi été enregistrée pour la base d'apprentissage et une autre de 24 800 lignes pour la base de test. La comparaison des différents modèles sera par ailleurs effectuée sur ces dernières bases réduites qui sont communes à tous les modèles. Il est à noter qu'une RMSE sera donc recalculée ultérieurement sur ces bases réduites pour le modèle de tarification traditionnelle.

➥ **Sélection de variables avec plus de 200 variables**

Enfin, même après analyse des corrélations, la base de données « à l'adresse » présentait plus de 200 variables. Il est alors apparu difficile de fournir autant de variables à la procédure *stepAIC* afin de réaliser la sélection de variables. Pour contourner ce problème, une présélection de variables a donc du être entreprise.

Cette dernière a été effectuée visuellement, à l'aide des différentes analyses graphiques générées dans le chapitre 4. L'ensemble de ces analyses a ainsi permis de réduire le nombre de variables à une soixantaine pour chaque modèle (fréquence/coût moyen) et pour chaque garantie (DDE ou vol).

Un travail préliminaire conséquent a donc dû être entrepris, ces trois problématiques nécessitant l'examen de nombreuses analyses visuelles. Une fois ces traitements effectués, les modèles de fréquence et de coût moyen ont été entraînés pour chaque garantie selon la même méthode qu'adoptée précédemment pour la tarification traditionnelle.

5.2.2 La modélisation de la garantie dégâts des eaux

Les lois Poisson et Gamma ont, une nouvelle fois, été reprises pour la modélisation du nombre de sinistres et du coût moyen.

Les modèles ont été calibrés sur des bases réduites d'apprentissage et de test afin d'éviter la présence de valeurs manquantes ainsi qu'expliqué au point précédent. Les effectifs de ces bases sont présentés ci-contre :

DDE	Fréquence	Coût
Base d'apprentissage	120 862	1 559
Base de test	51 730	588

FIGURE 5.34 – Répartition du nombre de lignes entre les bases d'apprentissage et de test pour les GLM à l'adresse de la garantie dégâts des eaux

Une sélection de variables a ensuite été entreprise et, à la fin de cette procédure, les variables suivantes ont été retenues :

	Modèle	Fréquence	Coût
Adresse	Zonier	x	
	Nombre de piscines	x	
	Surface de la maison * (> 125 m2)	x	
	Surface de la parcelle *		x
Météo	Température annuelle moyenne (> 14°C) *	x	
	Vitesse annuelle maximale du vent (> 26 m/s) *		x
Commune	Part d'ouvriers (CS6) *	x	
	Part de retraités *	x	
	Part de résidences touristiques	x	
	Part d'auberges de jeunesse	x	
	Part industrielle de la consommation d'électricité *	x	
	Présence d'inondations dans le passé	x	
	Présence d'inondations rapides dans le passé	x	
	Part de propriétaires de résidences principales *		x
	Prix au m2 des maisons *		x
	Part des maisons (et résidences principales) construites entre 1991 et 2005 *		x

FIGURE 5.35 – Variables retenues dans les GLM à l'adresse de la garantie dégâts des eaux

* Il est à noter que les variables dont le nom est suivi d'une étoile ont été considérées de manière continue.

Le détail et l'interprétation des coefficients de chaque variable peuvent maintenant être exposés.

Fréquence

	Variables	Coefficient	Pvalue
Adresse	Surface de la maison * (> 125 m2)	0,036174	0,000585
Météo	Température annuelle moyenne (> 14°C) *	0,060527	0,002437
Commune	Part d'ouvriers (CS6) *	-0,02587	5,20E-06
	Part de retraités *	-0,013653	0,008894
	Part industrielle de la consommation d'électricité *	-0,010517	0,061631

Code couleur pvalue
<0,1%
<1%
<5%
<10%
>10%

FIGURE 5.36 – Coefficients et pvalue des variables continues du GLM fréquence à l'adresse pour la garantie dégâts des eaux

Ainsi, l'ensemble des variables continues est globalement significatif à l'exception de la variable représentant la part industrielle de la consommation d'électricité de la commune. Cependant, au vu de la faible *pvalue* de cette variable, cette dernière peut tout de même être considérée comme impactante dans la modélisation de la fréquence d'un DDE.

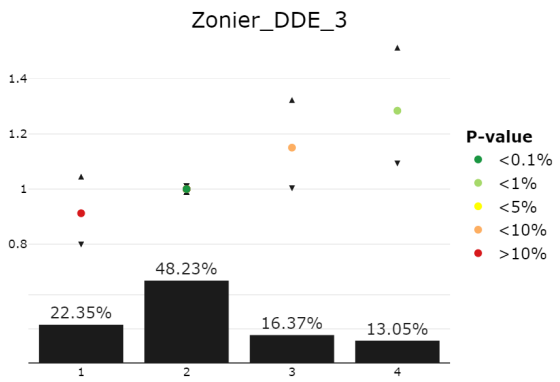


FIGURE 5.37 – Coefficients liés aux modalités du zonier

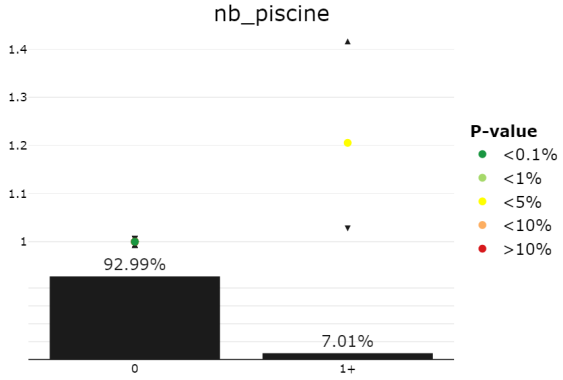


FIGURE 5.38 – Coefficients du nombre de piscines

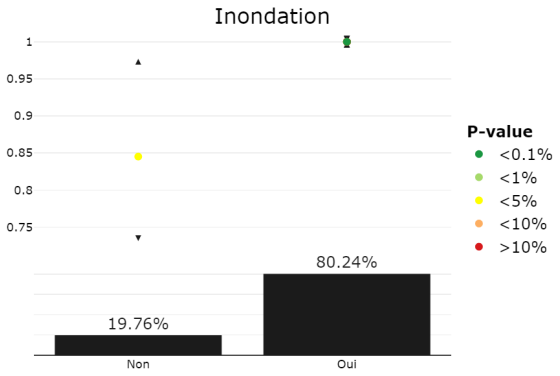


FIGURE 5.39 – Coefficients liés à la présence d'inondations à la commune

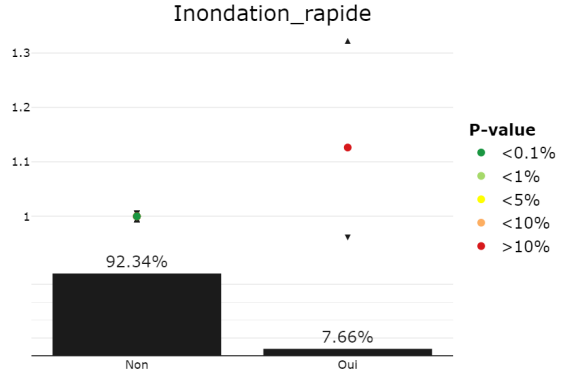


FIGURE 5.40 – Coefficients liés à la présence d'inondations rapides à la commune

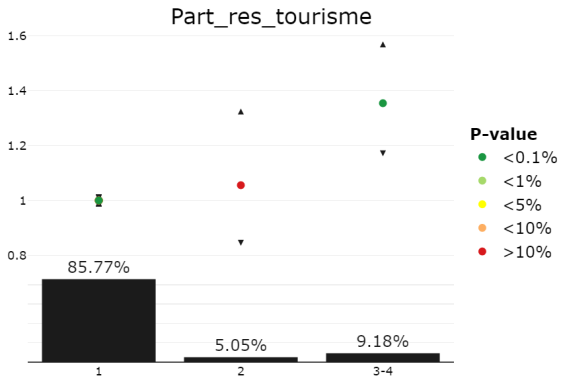


FIGURE 5.41 – Coefficients liés aux résidences touristiques

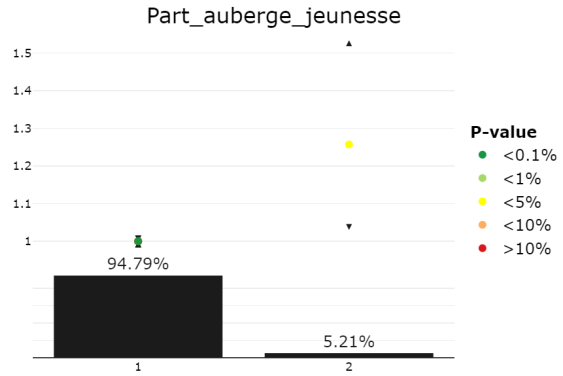


FIGURE 5.42 – Coefficients liés aux auberges de jeunesse

En revanche, pour les variables discrètes, l'interprétation est plus complexe. Ainsi, de nombreuses variables telles la présence d'inondations dans le passé, le nombre de piscines ou encore la part d'auberges de jeunesse sont clairement significatives. Cependant, pour d'autres, certaines modalités ne vérifient pas le test de Student. C'est par exemple le cas pour les modalités 1 et 3 du zonier. Une explication possible de ce phénomène est l'occultation de l'effet du zonier par d'autres variables externes du modèle. Concernant la part des résidences touristiques, il est également possible de remarquer que le groupe 2 n'est pas très significatif. De même, la variable présence d'inondations rapides dans le passé, bien qu'ayant été sélectionnée par la procédure *stepAIC*, ne semble pas, elle non plus, significative. Ainsi, pour évaluer la cohérence et l'apport de ces derniers coefficients paraissant, à première vue, non significatifs, il semble utile de se pencher sur l'interprétation des différentes variables.

Ainsi, une proposition d'interprétation est donnée ci-dessous :

- **Surface de la maison ($>125m^2$)** : la variable représentant la surface de la maison semble jouer le rôle usuel du nombre de pièces dans la tarification traditionnelle : ainsi, plus cette variable augmente, plus le risque de dégâts des eaux est important ;
- **Présence d'inondations dans le passé** : cette variable binaire indiquant la survenue antérieure d'inondations constitue une des causes principales de dégâts des eaux et, en tant que telle, sa présence est cohérente. Les territoires ayant déjà fait l'objet d'inondations sont en effet des zones à risque susceptibles de subir de nouvelles inondations ;
- **Présence d'inondations rapides dans le passé** : une réflexion similaire à celle menée pour la présence d'inondations dans le passé peut être menée ;
- **Température annuelle moyenne ($>14°C$)** : cette variable peut, en augmentant (en cas de fortes chaleurs par exemple), être liée à d'autres phénomènes météorologiques comme de fortes pluies, des orages, etc. Ces derniers phénomènes sont susceptibles, quant à eux, de causer des infiltrations dans la toiture des maisons ou des inondations et donc être responsables d'un dégât des eaux ;
- **Nombre de piscines** : par le biais de cette variable, une autre cause de dégâts des eaux peut être évoquée : les dégâts des eaux liés aux éléments internes de la maison. Ainsi, plus le nombre de piscines augmente, plus la probabilité d'avoir un DDE interne s'élève ;
- **Part d'ouvriers (CS6)** : cette variable, bien que significative, est plus difficile à interpréter. Il est donc opportun de faire davantage de recherches pour tenter de présenter une explication cohérente.

  Ainsi, en s'intéressant à la répartition des ouvriers en France (carte présentée ci-dessous), une fracture nord/sud est clairement observable : la part des ouvriers est plus importante dans le nord que dans le sud.

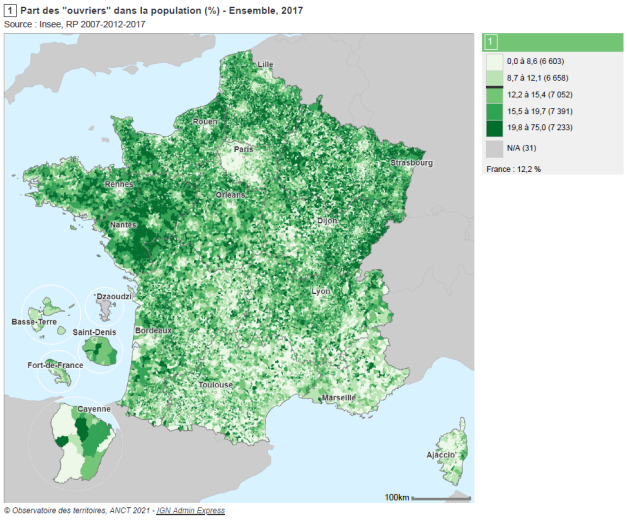


FIGURE 5.43 – Répartition des ouvriers par commune en France

En superposant cette carte aux cartes reprises ci-dessous et représentant deux des causes principales de dégâts des eux à savoir, respectivement, les inondations et les fortes pluies, une première interprétation apparaît.

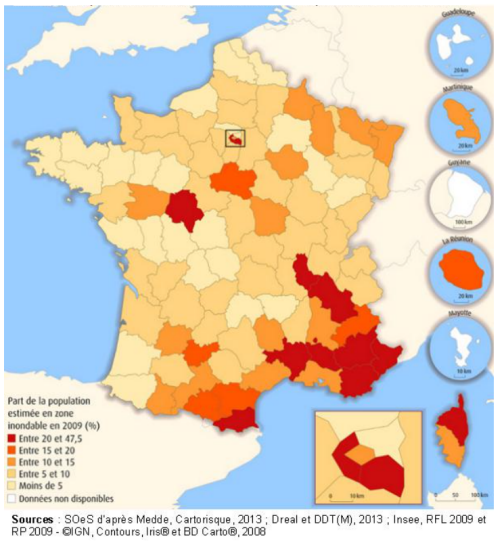


FIGURE 5.44 – Part estimée de la population en zone inondable

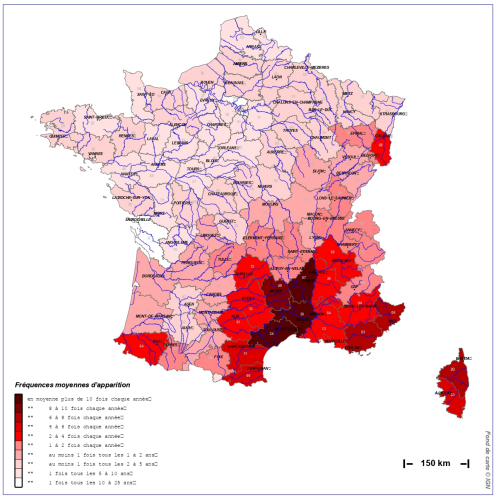


FIGURE 5.45 – Épisodes avec plus de 80 mm de pluies en un jour (pluies extrêmes) Période 1971/2020

Il semble en effet que la variable représentant la part des ouvriers à la commune n'explique pas en elle-même la fréquence d'un DDE mais permet de réaliser un découpage de la France en différentes zones plus ou moins à risque et ce, à la manière d'un zonier. Ainsi, le sud, plus exposé au DDE (inondations et fortes pluies), présente une part moins importante d'ouvriers à la commune que le nord qui est, quant à lui, moins sujet aux DDE. Le signe du coefficient associé à la part des ouvriers à la commune prend alors tout son sens.

➥ **Part de retraités** : un raisonnement similaire à celui mené pour les ouvriers est entrepris pour cette variable. Ainsi, une étude de la répartition des retraités en France est réalisée.

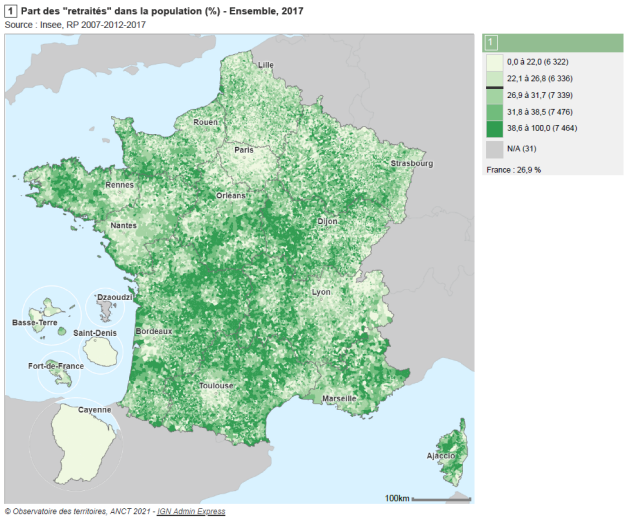


FIGURE 5.46 – Répartition des retraités par commune en France

Une autre division du territoire se dessine alors : il semble que les retraités vivent, pour la plupart, dans des zones peu risquées (au milieu de la France), d'où la valeur du coefficient dans le modèle de fréquence.

- **Part de résidences touristiques** et **Part d'auberges de jeunesse** : ces deux variables représentent globalement l'activité touristique de la commune et, de la même manière que pour les retraités et les ouvriers, un lien peut être établi entre la fréquence des dégâts des eaux et l'activité touristique d'une commune.

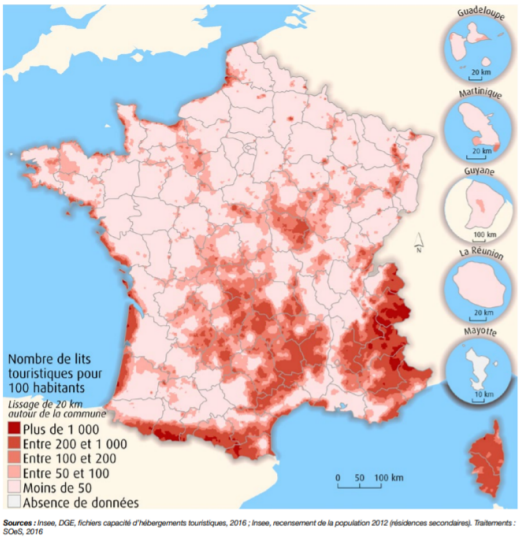


FIGURE 5.47 – Taux de fonction touristique en 2016, source : [8]

En effet, la carte ci-dessus, montre que la plupart des lieux touristiques se situent dans des zones sujettes à des inondations ou des fortes pluies pouvant causer des dégâts des eaux. Un nouveau paramètre vient donc, une nouvelle fois, délimiter le risque de fréquence.

- **Part industrielle de la consommation d'électricité** : enfin, plus la part des industries est importante dans la consommation d'électricité totale de la commune, plus il y a d'industries dans la commune. Or, il est fort probable que ces industries, avant de s'installer, ont réalisé une étude de risque concernant leurs implémentations. Ce faisant, elles se sont installées dans des zones peu risquées, d'où le signe du coefficient associé à cette variable.

Cette analyse permet donc de distinguer trois grands types de variables externes :

- ☛ celles se substituant aux variables déclaratives prises en compte usuellement dans la tarification traditionnelle (par exemple la surface de la maison qui remplace le nombre de pièces) ;
- ☛ celles apportant une valeur ajoutée à l'explication du phénomène (le nombre de piscines par exemple) ;
- ☛ celles jouant le même rôle que le zonier et qui contribuent à délimiter géographiquement les zones à risque (présence d'inondations dans le passé, présence d'inondations rapides dans le passé, température annuelle moyenne, part d'ouvriers à la commune, part de retraités à la commune, ...).

Coût moyen

Une analyse similaire est menée pour le modèle de coût qui ne comporte cependant que des variables continues.

	Modèle	Coefficient	Pvalue
Adresse	Surface de la parcelle *	0,026706	3,51E-05
Météo	Vitesse annuelle maximale du vent (> 26 m/s) *	0,069209	0,000475
Commune	Part de propriétaires de résidences principales *	-0,011647	0,070880
	Prix au m2 des maisons *	0,011712	0,073264
	Part des maisons (et résidences principales) construites entre 1991 et 2005 *	0,01366	0,041913

Code couleur pvalue
<0,1%
<1%
<5%
<10%
>10%

FIGURE 5.48 – Coefficients et pvalue des variables continues du GLM coût moyen à l'adresse pour la garantie dégâts des eaux

L'interprétation des variables est, au cas particulier, plus aisée :

- ➥ **Surface de la parcelle** : la variable représentant la surface de la parcelle semble jouer, encore une fois, le rôle usuel du nombre de pièces dans la tarification traditionnelle : ainsi, plus cette dernière augmente, plus le coût d'un dégât des eaux est important ;
- ➥ **Vitesse annuelle maximale du vent (>26m/s)** : cette variable représentant la vitesse maximale annuelle du vent influence le coût d'un dégât des eaux de manière assez logique. Ainsi, plus le vent est fort, plus les dégâts potentiellement causés peuvent s'avérer importants que ce soit dans le cadre d'une inondation, d'infiltrations ou encore de problèmes de canalisations dans la maison ;
- ➥ **Part de propriétaires de résidences principales** : les propriétaires, de par leur statut, sont généralement plus attentifs à la présence d'un possible DDE que les locataires. Ils découvrent donc les potentiels sinistres plus rapidement que les locataires : le DDE n'a donc pas le temps de s'aggraver et les dégâts sont donc moindres.
- ➥ **Prix au m^2 des maisons** : plus la maison est onéreuse (prix au m^2 coûteux), plus le coût d'un potentiel dégât des eaux sera important ;
- ➥ **Part des maisons (et résidences principales) construites entre 1991 et 2005** : plus il y a de maisons récentes au sein de la commune, plus la probabilité que la maison d'intérêt soit, elle même, récente est élevée. Or, les maisons récentes ont, la plupart du temps, une valeur élevée par rapport au marché et entraîne donc des coûts plus élevés en cas de dégâts des eaux.

Là encore, certaines variables tentent de remplacer les variables usuellement utilisées dans la tarification traditionnelle. Ainsi, outre la surface de la maison qui est fortement liée au nombre de pièces, le prix au m^2 des maisons peut être assimilé au capital assuré et la part des maisons construites entre 1991 et 2005 joue le rôle de l'année de construction. La vitesse annuelle maximale du vent, pour sa part, apporte une information supplémentaire sur la modélisation du risque.

Au global

Au global, la combinaison des modèles de fréquence et de coût donne les résultats suivants :

	Base d'apprentissage			Base de test		
	Réalité	*Prédiction*	*Erreur (%)*	*Réalité*	*Prédiction*	*Erreur (%)*
Fréquence	2,7713%	2,7713%	**0,00%**	2,4354%	2,7749%	**13,94%**
Coût moyen	1 724,4 €	1 722,9 €	**-0,09%**	1 755,1 €	1 707,0 €	**-2,74%**
Coût total	2 778 086 €	2 764 173 €	**-0,50%**	1 061 828 €	1 184 892 €	**11,59%**

FIGURE 5.49 – Prédictions avec les GLM à l'adresse pour la garantie dégâts des eaux

avec une RMSE de :

	RMSE base d'apprentissage	RMSE base de test
Fréquence	0,1180	0,1102
Coût moyen	2 872,08	2 552,58
Coût total	**390,3634**	**340,2812**

FIGURE 5.50 – RMSE des GLM à l'adresse pour la garantie dégâts des eaux

5.2.3 La modélisation de la garantie vol

Pour la garantie vol, les lois Poisson et Gamma ont également été utilisées. Les bases d'apprentissage et de test sont les suivantes :

Vol	Fréquence	Coût
Base d'apprentissage	120 862	674
Base de test	51 730	282

FIGURE 5.51 – Répartition du nombre de lignes entre les bases d'apprentissage et de test pour les GLM à l'adresse de la garantie vol

La procédure de sélection de variables permet de retenir les variables suivantes :

	Modèle	Fréquence	Coût
Adresse	Distance à la station de police la plus proche (< 155 s) *	x	
	Pourcentage de parties vertes *	x	
	Surface de la maison (> 100 m2) *	x	
	Surface de la maison (> 125 m2) *		x
	Nombre de piscines	x	
	Distance à une autre maison *	x	
	Zonier		x

FIGURE 5.52 – Variables retenues dans les GLM à l'adresse de la garantie vol (1/2)

* Il est à noter que les variables dont le nom est suivi d'une étoile ont été considérées de manière continue.

	Modèle	Fréquence	Coût
Commune	Zonage ABC	x	
	Part des résidences principales de deux pièces (> 0.1) *	x	
	Part des résidences principales de cinq pièces ou plus (> 0.35) *	x	
	Part des actifs occupés *	x	
	Part des ménages "couple sans enfant" *		x
	9ieme décile du niveau de vie des ménages *		x
	Part des propriétaires de résidences principales *		x
	Taxe foncière du bâti *		x
	Part des résidences secondaires *		x
Département	Taux de cambriolage des résidences principales *	x	
	Taux de vols simples envers les particuliers *	x	
	Taux de vols des véhicules *	x	
	Taux de vols simples envers les particuliers dans la sphère publique *		x
	Taux de cambriolage des résidences principales (> 0.002) *		x
	Taux de cambriolage des résidences secondaires *		x

FIGURE 5.53 – Variables retenues dans les GLM à l'adresse de la garantie vol (2/2)

Une analyse des coefficients peut maintenant être réalisée et une tentative d'interprétation de ces derniers proposée.

Fréquence

	Modèle	Coefficient	Pvalue
Adresse	Distance à la station de police la plus proche (< 155 s) *	0,100064	0,015222
	Pourcentage de parties vertes *	0,037824	1,68E-07
	Surface de la maison (> 100 m2) *	0,04068	0,000154
	Distance à une autre maison *	0,014427	0,053743
Commune	Part des résidences principales de deux pièces (> 0.1) *	0,068089	0,000938
	Part des résidences principales de cinq pièces ou plus (> 0.35) *	0,039397	0,002682
	Part des actifs occupés *	-0,026928	0,00106
Département	Taux de cambriolage des résidences principales *	0,024875	0,001383
	Taux de vols simples envers les particuliers *	0,015993	0,025668
	Taux de vols des véhicules *	0,020666	0,004716

Code couleur pvalue
<0,1%
<1%
<5%
<10%
>10%

FIGURE 5.54 – Coefficients et pvalue des variables continues du GLM fréquence à l'adresse pour la garantie vol

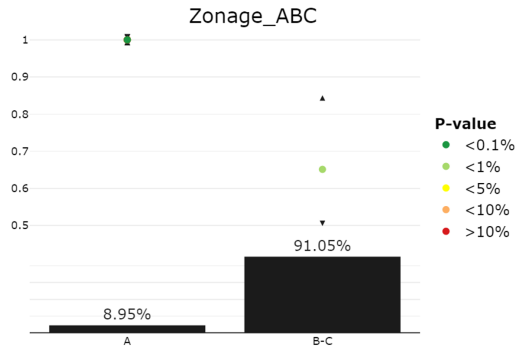


FIGURE 5.55 – Coefficients du zonage immobilier ABC

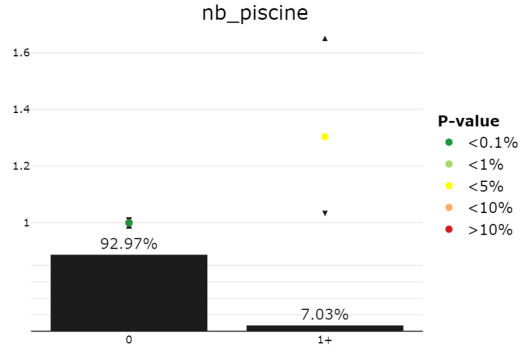


FIGURE 5.56 – Coefficients du nombre de piscines

L'ensemble des coefficients retenus dans le modèle, que ce soit pour les variables continues ou discrètes, apparaît significatif. Une interprétation de ces derniers est proposée :

- **Surface de la maison ($>100m^2$)** : cette variable se substitue au nombre de pièces dans la tarification traditionnelle : ainsi, plus la maison est grande, plus le risque de vol est important ;
- **Pourcentage de parties vertes** et **Nombre de piscines** : tant le nombre de piscines que la surface du jardin (c'est-à-dire le pourcentage de parties vertes) peuvent être des indicateurs d'une certaine richesse favorisant l'exposition aux vols ;
- **Distance à une autre maison** : une maison isolée aura plus de risques d'être cambriolée qu'une maison proche géographiquement de ses voisins ;
- **Distance à la station de police la plus proche ($<155s$)** : les maisons assez éloignées des stations de police ou de gendarmerie subiront, globalement, davantage de vols ;
- **Zonage ABC** :

Ce zonage, immobilier à l'origine, représente la tension du marché immobilier local. Par le biais de cette variable, des zones à risque (les zones A) sont mises en évidence à la manière d'un zonier. Une cartographie des zones est présentée ci-contre.

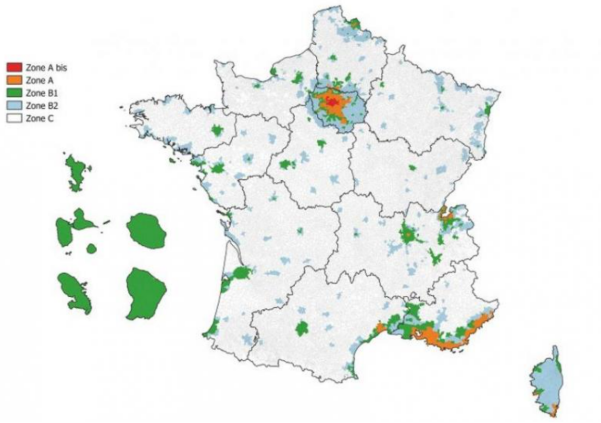


FIGURE 5.57 – Zonage ABC

La localisation des différentes zones reste, somme toute, cohérente avec les coefficients du modèle ;

- **Part des résidences principales de deux pièces (>0.1)** et **Part des résidences principales de cinq pièces ou plus (>0.35)** : intuitivement, plus le nombre de pièces est élevé, plus nombreux sont les biens susceptibles d'être volés : la fréquence des vols sera donc plus importante ;
- **Part des actifs occupés** : l'interprétation de ce coefficient est délicat et ne sera pas abordé dans ce mémoire. Il nécessiterait de faire l'objet d'une étude complémentaire.
- **Taux de cambriolages des résidences principales**, **taux de vols simples envers les particuliers** et **taux de vols des véhicules** : ces variables sont toutes des indicateurs de délinquance à la maille du département : il semble ainsi logique qu'une augmentation de ces variables traduise une insécurité globale et donc une augmentation du nombre de vols.

Dans le cadre de ce modèle, la surface de la maison et la part des résidences principales de deux pièces ou cinq pièces (voire plus) peuvent être assimilées à la variable nombre de pièces prise en compte dans la tarification traditionnelle. Le zonage ABC, les différentes variables liées à la délinquance et la part des actifs occupés tendent, quant à eux, à se substituer à un zonier. Toutes les autres variables interviennent en complément : il s'agit de variables qui viennent s'ajouter et qui n'ont pas d'équivalent dans la tarification traditionnelle (le pourcentage de parties vertes, l'isolement de la maison ou distance à une autre maison, la distance à la station de police ou de

gendarmerie la plus proche). Il est à noter que toutes ces variables sont issues de la reconnaissance d'images. La reconnaissance d'images semble donc générer de nouvelles informations très tarifaires dans le cadre de la garantie vol. Ce phénomène n'était pas observé pour la garantie DDE où, au mieux, les variables issues de la reconnaissance d'images approchaient les variables traditionnelles.

Coût moyen

	Modèle	Coefficient	Pvalue
Adresse	Surface de la maison (> 125 m2) *	0.026775	0.055857
Commune	Part des ménages "couple sans enfant" *	0.034116	0.001192
	9ieme décile du niveau de vie des ménages *	0.024376	0.005531
	Part des propriétaires de résidences principales *	0.024926	0.011332
	Taxe foncière du bâti *	-0.020036	0.025925
	Part des résidences secondaires *	-0.016678	0.068755
Département	Taux de vols simples envers les particuliers dans la sphère publique *	0.015303	0.093979
	Taux de cambriolage des résidences principales (> 0.002) *	0.117614	0.003733
	Taux de cambriolage des résidences secondaires *	0.026382	0.004255

Code couleur pvalue
<0,1%
<1%
<5%
<10%
>10%

FIGURE 5.58 – Coefficients et pvalue des variables continues du GLM coût moyen à l'adresse pour la garantie vol

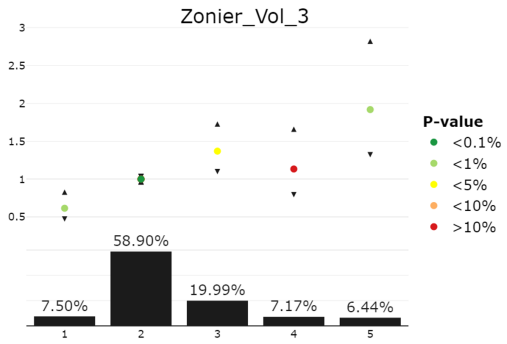


FIGURE 5.59 – Coefficients liés aux modalités du zonier

En ce qui concerne le coût moyen des vols, la plupart des coefficients sont significatifs. L'interprétation proposée est la suivante :

- **Surface de la maison ($>$125m^2)** : la variable peut se substituer, là encore, au nombre de pièces dans la tarification traditionnelle. Son impact est donc identique : plus la maison est grande, plus le coût du vol est important ;
- **Neuvième décile du niveau de vie des ménages** : plus le niveau de vie des ménages à la commune est élevé, plus le coût d'un vol dans cette commune est lourd ;
- **Part des propriétaires de résidences principales** : les propriétaires de résidences principales possèdent, généralement, davantage de biens et ces derniers sont souvent plus coûteux (que ceux des locataires). Par conséquent, le coût des vols sera plus important pour les propriétaires ;
- **Taxe foncière du bâti** : il est difficile d'établir une relation entre la taxe foncière du bâti et le coût moyen d'un vol, le taux de la taxe foncière étant étroitement lié aux politiques locales de fiscalité. Il faut également prendre en compte, lors de l'interprétation de cette variable, les exonérations possibles pouvant avoir lieu. Ces notions étant complexes, une étude plus approfondie serait nécessaire pour interpréter ce paramètre.

- **Part des résidences secondaires** : les résidences secondaires sont généralement moins équipées que les résidences principales. Ainsi, les vols réalisés dans des résidences secondaires sont souvent moins coûteux pour l'assureur ;
- **Part des ménages « couple sans enfant »** : les couples sans enfant, lorsqu'ils en ont les moyens, ont tendance à acquérir, et donc posséder, des biens globalement plus onéreux (produits high tech, oeuvres d'art) que les couples avec enfant, ces derniers dépensant davantage pour les achats de consommation courante. Ainsi, plus il y a de couples sans enfant dans la commune, plus le coût d'un vol peut être important ;
- **Taux de vols simples envers les particuliers dans la sphère publique**, **Taux de cambriolages des résidences principales (>0.002)** et **Taux de cambriolages des résidences secondaires** : ces indicateurs de délinquance sont susceptibles de fournir des informations concernant la fréquence des vols. Ils sont donc révélateurs d'une certaine aisance des lieux, de ce fait le coût des vols en est d'autant plus important.

Une fois encore, de nombreuses variables, similaires à celles prises en compte dans la tarification traditionnelle, se retrouvent dans le modèle de coût moyen. Ainsi, outre la surface de la maison qui s'assimile au nombre de pièces et le capital assuré qui peut se substituer au neuvième décile du niveau de vie des ménages, la qualité s'apparente, quant à elle, à la part des propriétaires de résidences principales. Un autre indicateur apparaît cependant : la part des résidences secondaires qui joue le rôle d'une variable dont il est parfois tenu compte dans la tarification traditionnelle. D'autres variables comme les indicateurs de délinquance viennent, quant à elles, compléter le zonier.

Au global

Les résultats finaux sont présentés ci-contre :

	Base d'apprentissage			*Base de test*		
	Réalité	*Prédiction*	*Erreur (%)*	*Réalité*	*Prédiction*	*Erreur (%)*
Fréquence	1,2104%	1,2104%	**0,00%**	1,1801%	1,2204%	**3,42%**
Coût moyen	3 431,7 €	3 437,8 €	**0,18%**	3 487,9 €	3 402,4 €	**-2,45%**
Coût total	2 347 312 €	2 444 374 €	**4,14%**	1 004 513 €	1 057 843 €	**5,31%**

FIGURE 5.60 – Prédictions avec les GLM à l'adresse pour la garantie vol

avec une RMSE de :

	RMSE base d'apprentissage	RMSE base de test
Fréquence	0,0770	0,0765
Coût moyen	4 183,79	5 194,52
Coût total	**435,1901**	**454,3478**

FIGURE 5.61 – RMSE des GLM à l'adresse pour la garantie vol

5.3 La forêt aléatoire à l'adresse

Pour terminer, un dernier modèle à l'adresse a été construit. Ce dernier modèle repose sur des méthodes de *machine learning* plus complexes : les forêts aléatoires. Il a été décidé de les utiliser afin d'étudier l'apport potentiel de telles méthodes dans le cadre d'une tarification à l'adresse. L'implémentation de ce modèle permet de répondre à une question : les méthodes plus complexes de *machine learning* peuvent-elles surpasser les GLM à l'adresse et concurrencer la tarification traditionnelle ?

La méthode différant des GLM, une nouvelle méthodologie est mise en œuvre.

Cependant, afin de faciliter l'implémentation du modèle, il a été décidé de partir sur la base du pré-traitement des données externes effectué lors de la mise en place du GLM à l'adresse. Ainsi, aucune problématique de discrétisation de variables numériques ou de valeurs manquantes ne sera exposée dans cette section et la présélection des variables antérieurement réalisée sera reprise en l'état.

Les étapes suivantes sont ensuite effectuées pour modéliser la fréquence et le coût moyen de chaque garantie :

- 1) Implémentation d'une première forêt aléatoire avec les paramètres par défaut recommandés dans l'article de recherche [13] introduisant les forêts aléatoires revisitées ;
- 2) Processus de sélection de variables : cette sélection de variables est réalisée en retenant les variables de plus forte importance dans le modèle mis en place à l'étape 1) ;
- 3) Implémentation d'un nouveau modèle avec les variables sélectionnées ;
- 4) Optimisation par validation croisée du paramètre *ncand* représentant le nombre de variables candidates choisies aléatoirement et à considérer, à chaque coupure, pour trouver la décomposition optimale. Le paramètre *minbucket*, représentant le nombre d'observations minimales pour effectuer une coupure est, quant à lui, fixé à 1% du nombre total d'observations ainsi que recommandé dans l'article de recherche. Enfin, le paramètre *ntree* est, lui, fixé à 500 ;
- 5) Analyse du modèle final obtenu.

Par ailleurs, il est important de rappeler que, dans le cadre de cette étude, c'est une version revisitée des forêts aléatoires (avec des fonctions de perte adaptées aux problématiques actuarielles) qui a été utilisée. L'implémentation de ces forêts est présente dans le package **distRforest** de R.

5.3.1 La modélisation de la garantie dégâts des eaux

Fréquence

Le paramètre retenu pour le modèle de fréquence est le suivant :

— $ncand = 8$.

Une fois le modèle implémenté, il est souhaité disposer de davantage d'informations sur la contribution de chaque variable dans l'explication du phénomène. Pour répondre à ce besoin, une mesure d'importance peut être calculée pour chacune des variables. L'importance des vingt premières variables les plus significatives du modèle de fréquence est présentée ci-après :

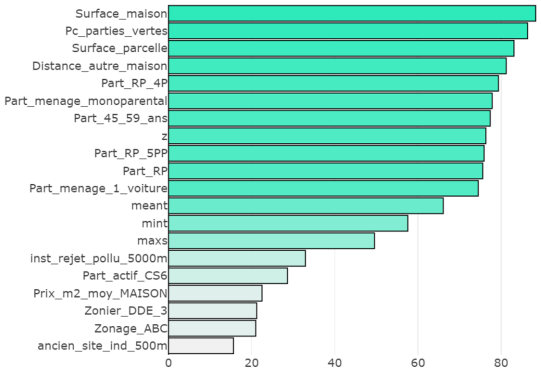


FIGURE 5.62 – Importance du modèle de fréquence pour la garantie dégâts des eaux

Ainsi, il s'avère que certaines variables telles la surface de la maison, la température moyenne annuelle ou encore la part des ouvriers à la commune (CS6), qui étaient déjà présentes dans les GLM à l'adresse, contribuent également, là aussi, et de manière importante, à l'explication de la fréquence des DDE. Les deux modèles à l'adresse s'accordent donc sur l'importance de ces variables. En revanche, d'autres variables, présentées comme significatives dans la forêt aléatoire, n'étaient, quant à elles, pas présentes dans les GLM. Plusieurs explications possibles à ce phénomène peuvent être avancées :

- soit la variable complète une information déjà présente dans les GLM : c'est le cas, par exemple, pour le pourcentage de parties vertes, la surface de la parcelle ou encore les parts de résidences principales avec quatre pièces ou plus, qui viennent compléter l'information donnée par la surface de la maison présente, quant à elle, dans le GLM à l'adresse ;
- soit la variable apporte une nouvelle information tarifaire non détectée par les GLM pour diverses raisons comme, par exemple, l'altitude, la température minimale annuelle, le maximum annuel des chutes de neige, etc.

Coût moyen

Concernant le coût moyen, le paramètre optimisé de la forêt aléatoire est le suivant :

— $ncand = 3$.

L'importance des vingt premières variables les plus significatives est présentée ci-contre :

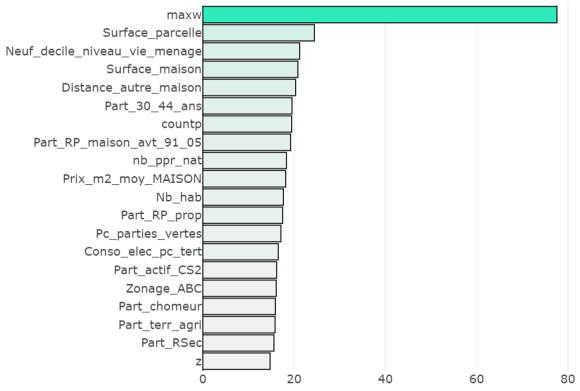


FIGURE 5.63 – Importance du modèle de coût moyen pour la garantie dégâts des eaux

Tant la vitesse maximale du vent, le prix du m^2 des maisons à la commune, la surface de la parcelle, la part des propriétaires de résidences principales ou encore la part des maisons construites entre 1991 et 2005 sont des variables, elles aussi, présentes dans le GLM de coût moyen à l'adresse. L'impact de ces variables est donc validé par les deux modèles à l'adresse. D'autres variables, de la même manière que pour la fréquence, viennent, quant à elles, compléter les informations déjà présentes dans le GLM à l'adresse. C'est le cas notamment pour la surface de la maison, intrinsèquement corrélée à la surface de la parcelle (corrélation linéaire de 0.7), ou encore pour le neuvième décile du niveau de vie des ménages, potentiellement lié au prix du m^2 des maisons. Enfin, d'autres informations, non présentes dans le GLM à l'adresse, sont captées par la forêt aléatoire : il s'agit de la somme annuelle des précipitations, la part des 30-44 ans à la commune, la part des résidences secondaires à la commune, etc.

Au global

La combinaison de ces deux modèles (fréquence et coût moyen) donne donc les résultats suivants :

	Base d'apprentissage			Base de test		
	Réalité	*Prédiction*	*Erreur (%)*	*Réalité*	*Prédiction*	*Erreur (%)*
Fréquence	2,7713%	2,7319%	**-1,42%**	2,4354%	2,7322%	**12,19%**
Coût moyen	1 724,4 €	1 726,1 €	**0,10%**	1 755,1 €	1 708,3 €	**-2,66%**
Coût total	2 778 086 €	2 706 919 €	**-2,56%**	1 061 828 €	1 160 812 €	**9,32%**

FIGURE 5.64 – Prédictions avec les forêts aléatoires à l'adresse pour la garantie dégâts des eaux

avec une RMSE de :

	RMSE base d'apprentissage	RMSE base de test
Fréquence	0,1176	0,1102
Coût moyen	2 641,55	2 572,42
Coût total	**389,4170**	**340,2605**

FIGURE 5.65 – RMSE des forêts aléatoires à l'adresse pour la garantie dégâts des eaux

5.3.2 La modélisation de la garantie vol

Fréquence

Pour le modèle de fréquence de la garantie vol, le paramètre suivant a été choisi :

— $ncand = 10$.

L'importance des vingt premières variables les plus significatives est présentée ci-contre :

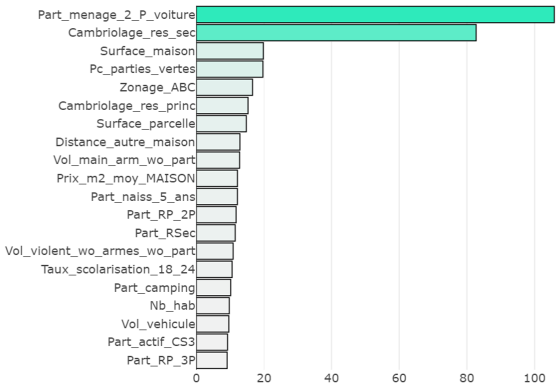


FIGURE 5.66 – Importance du modèle de fréquence pour la garantie vol

De nombreuses variables, déjà présentes dans le GLM de fréquence de la garantie vol, sont également considérées comme importantes dans ce nouveau modèle. C'est le cas notamment pour la surface de la maison, le pourcentage de parties vertes, le zonage ABC, l'isolement de la maison (distance à une autre maison), les indicateurs de criminalité, etc. D'autres variables viennent ensuite compléter l'explication de la fréquence des vols : la surface de la parcelle, le prix du m^2 des maisons ou encore la part des résidences secondaires. En revanche, il est important de souligner que les deux variables paraissant les plus importantes dans ce modèle (à savoir la part des ménages avec deux voitures ou plus et le taux de cambriolages des résidences secondaires à la maille département) n'ont pas été captées par le GLM à l'adresse. Cependant, leur présence dans ce nouveau modèle semble cohérente. En effet, les résidences secondaires, de par leur inoccupation fréquente, sont plus exposées aux cambriolages que les résidences principales. En ce qui concerne le nombre de voitures possédées par une personne, cette variable peut être considérée comme révélatrice d'une certaine aisance financière et exposer de ce fait davantage cette personne aux vols.

Coût moyen

Concernant le coût moyen, le paramètre est le suivant :

— $ncand = 2$.

L'importance des vingt premières variables les plus significatives est présentée ci-contre :

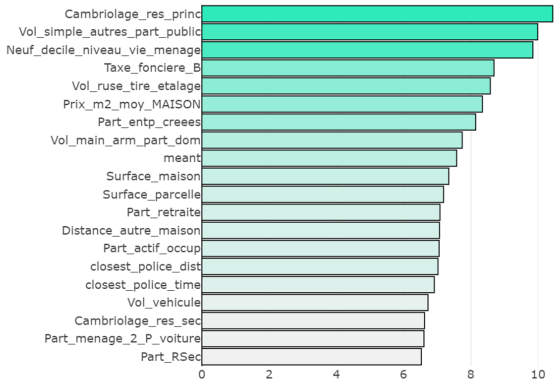


FIGURE 5.67 – Importance du modèle de coût moyen pour la garantie vol

Des remarques similaires à celles effectuées précédemment peuvent être formulées. Ainsi, de nombreuses variables, présentes initialement dans le GLM de coût moyen à l'adresse, sont retrouvées (la surface de la maison, le neuvième décile du niveau de vie des ménages, la taxe foncière du bâti, la part des actifs occupés, la part des résidences secondaires ou encore les indicateurs de criminalité) tandis que d'autres viennent compléter l'explication du phénomène (la part des entreprises créées sur la dernière année à la commune, le prix du m^2 des maisons à la commune, la température moyenne annuelle, la part des ménages possédant deux voitures ou plus, etc). L'ajout de ces dernières variables semble par ailleurs cohérente avec la modélisation du phénomène considéré à savoir le coût moyen d'un vol.

Au global

Ces deux modèles de fréquence et de coût moyen permettent d'obtenir les résultats suivants : avec une RMSE de :

	Base d'apprentissage			Base de test		
	Réalité	*Prédiction*	*Erreur (%)*	*Réalité*	*Prédiction*	*Erreur (%)*
Fréquence	1,2104%	1,1555%	**-4,54%**	1,1801%	1,1582%	**-1,86%**
Coût moyen	3 431,7 €	3 426,7 €	**-0,15%**	3 487,9 €	3 448,6 €	**-1,13%**
Coût total	2 347 312 €	2 239 052 €	**-4,61%**	1 004 513 €	961 013 €	**-4,33%**

FIGURE 5.68 – Prédictions avec les forêts aléatoires à l'adresse pour la garantie vol

	RMSE base d'apprentissage	RMSE base de test
Fréquence	0,0769	0,0765
Coût moyen	3 635,72	5 097,88
Coût total	**434,9070**	**454,1350**

FIGURE 5.69 – RMSE des forêts aléatoires à l'adresse pour la garantie vol

5.4 Comparaison des différents modèles

Une fois ces modèles calibrés et testés, ces derniers vont faire l'objet d'une comparaison par le biais de la RMSE calculée sur la base de test réduite (il est donc sous-entendu que la RMSE du modèle traditionnel a été recalculée sur cette base réduite : les résultats sont donc bien comparables).

Avant de présenter les résultats de ces comparaisons, quelques remarques d'ordre général peuvent être formulées sur les modèles précédemment construits.

Ainsi, concernant les modèles traditionnels calibrés précédemment, l'étude des résultats peut soulever des questions quant à la modélisation de la fréquence de la garantie DDE. En effet, sur la figure 5.16, une forte différence de prédiction de la fréquence est présente sur la base de test. Ce phénomène peut s'expliquer par la faible volumétrie des données ou encore par la potentielle absence d'une variable très tarifaire dans la modélisation de la garantie DDE. Il convient d'être conscient de ce phénomène même s'il ne peut lui être trouvé une solution dans le cadre de ce mémoire. Quant au modèle de coût moyen DDE, il paraît plutôt correct. Concernant la garantie vol, aucune remarque particulière n'est formulée, les modélisations de la fréquence et du coût moyen paraissant raisonnables sur la base de test.

Pour les modèles à l'adresse, basés essentiellement sur de l'*Open Data*, deux effets majeurs se distinguent. Ainsi, certaines variables tentent d'approcher voire de répliquer des variables déjà présentes dans la tarification traditionnelle ce qui apporte alors des informations cruciales à la modélisation des différents phénomènes et permet au tarif à l'adresse d'approcher la tarification usuelle. D'autres variables, quant à elles, apportent des informations complémentaires qu'une tarification traditionnelle n'aurait pu capter : ceci démontre l'apport indéniable de l'*Open Data* et de la reconnaissance d'images. Cela est particulièrement vrai pour la garantie vol.

Par ailleurs, il convient de noter que la pertinence de la tarification à l'adresse diffère en fonction de la garantie considérée. Ainsi, pour la garantie DDE, peu de variables en *Open Data* ont un réel impact sur les modélisations. En revanche, pour la garantie vol, l'*Open Data* et la reconnaissance d'images permettent d'ajouter de nombreuses informations qui viennent ainsi compléter l'explication de la fréquence des vols ou du coût moyen de ces derniers. Ce constat peut amener à penser que la tarification à l'adresse pourrait peut-être, dans le cas de la garantie vol, surpasser la tarification traditionnelle. Pour la garantie DDE, cela semble plus difficile.

Concernant la RMSE à proprement parler, les résultats suivants sont obtenus :

		Fréquence		Coût moyen		Total	
		RMSE app	*RMSE test*	*RMSE app*	*RMSE test*	*RMSE app*	*RMSE test*
DDE	*Tarification traditionnelle*	0,11792	0,11017	2 876	2 500	390,18	339,97
	GLM à l'adresse	0,11801	0,11023	2 872	2 553	390,36	340,28
	Forêt aléatoire à l'adresse	0,11759	0,11020	2 642	2 572	389,42	340,26
Vol	*Tarification traditionnelle*	0,07699	0,07651	4 185	4 987	435,14	454,39
	GLM à l'adresse	0,07698	0,07648	4 184	5 195	435,19	454,35
	Forêt aléatoire à l'adresse	0,07690	0,07649	3 636	5 098	434,91	454,13

FIGURE 5.70 – RMSE sur les bases réduites pour tous les modèles

Pour la garantie DDE, que ce soit pour le coût global des sinistres, leur fréquence ou leur coût moyen, c'est à chaque fois le modèle traditionnel qui surpasse les autres modèles.

En revanche, pour la garantie vol, c'est la forêt aléatoire qui obtient globalement la plus petite RMSE. Cependant, il convient de noter que, concernant la fréquence mais aussi le coût moyen, la meilleure RMSE n'est pas la forêt aléatoire : une effet de compensation entre les deux modèles (de fréquence et de coût moyen) s'est donc certainement produit. Ainsi, concernant la

modélisation de la fréquence en particulier, le critère de la RMSE amène à privilégier le GLM à l'adresse tandis que, pour le modèle de coût moyen, ce sera plutôt la tarification traditionnelle.

Aucun modèle de tarification ne semble donc clairement se démarquer à la lecture de ces résultats. Cependant, les valeurs très similaires des RMSE des différents modèles sont encourageantes et démontrent que les modèles à l'adresse, basés essentiellement sur des données en *Open Data*, peuvent très fortement approcher les modèles traditionnels. Pour répondre à la question centrale du mémoire, à savoir « La tarification à l'adresse peut-elle concurrencer la tarification traditionnelle ? » il n'est pas possible de se baser uniquement sur les RMSE des modèles, ces dernières n'intégrant pas la notion même de concurrence. Une étude plus approfondie doit donc être menée. Une comparaison plus réaliste des modèles va donc maintenant être réalisée à l'aide d'un marché concurrentiel.

Chapitre 6

Marché concurrentiel

6.1 Les hypothèses du marché concurrentiel

L'objectif du marché concurrentiel est de simuler un environnement se rapprochant au plus près de la réalité afin de mettre en concurrence plusieurs assureurs proposant des tarifs différents pour un même produit.

Pour constituer ce marché, différentes hypothèses doivent être posées. Certaines concernent les clients présents sur le marché ainsi que leur comportement vis-à-vis de ce dernier tandis que d'autres se rapportent aux assureurs.

6.1.1 Les clients

Dans le cadre de ce mémoire, il est supposé que l'ensemble des clients réalise leur devis en ligne.

Les clients concernés sont ceux repris dans la base de données réduite constituée lors de l'implémentation des modèles à l'adresse (base de données au sein de laquelle aucune variable à l'adresse ne présente de valeurs manquantes). Ils sont ensuite divisés en deux catégories :

- ceux dont les polices ont servi à apprendre les modèles et qui constituent le premier groupe de clients (**groupe d'apprentissage**) ;
- ceux dont les polices ont servi à tester les modèles et qui composent, quant à eux, le deuxième groupe de clients (**groupe de test**).

Ce faisant, la performance des assureurs sera principalement évaluée sur le groupe de test, là où les résultats ne seront pas biaisés. Quant aux résultats du groupe d'apprentissage, ils permettront de détecter la présence potentielle de surapprentissage.

Tous ces clients, quel que soit le groupe, choisiront leur assureur en fonction de la prime proposée. A ce stade, l'hypothèse la plus couramment émise, quant au comportement attendu du client, consiste à dire qu'il choisira l'assureur le moins cher. Cependant, dans le cadre de cette étude, cette hypothèse doit être nuancée et ne correspondra pas forcément à la réalité. En effet, ainsi qu'évoqué précédemment dans la sous-section 1.3.3, l'objectif des méthodes de tarification à l'adresse est de permettre au client de gagner du temps en n'ayant à renseigner qu'un seul élément : son adresse. Nul doute donc, qu'outre le niveau du tarif, le paramètre "gain de temps" sera également pris en compte au moment du choix de l'assureur. Pour tenter de modéliser ce phénomène, une certaine élasticité au prix va donc être calibrée.

Cette élasticité permettra de mesurer la différence de prix qu'un client est prêt à accepter en contrepartie du gain de temps offert par la tarification à l'adresse. Elle sera exprimée en pourcentage de la prime et dépendra de certains paramètres développés ci-dessous.

Afin de déterminer cette élasticité, des avis d'experts ont été pris en compte et des mémoires sur l'élasticité au prix en assurance MRH ont été étudiés (par exemple [6]). Il a finalement été décidé que cette élasticité au prix dépendrait de deux variables principales : l'âge et le capital mobilier assuré. L'âge est en effet l'une des variables ressortant le plus. Il apparaît ainsi que ce sont les personnes dont l'âge est compris entre 35 et 60 ans qui sont les plus susceptibles d'accepter un tarif un peu plus élevé en contrepartie d'un certain gain de temps. Grossièrement, ces personnes sont très souvent parents : ils doivent donc jongler entre leurs obligations professionnelles et familiales et disposent donc de moins de temps à consacrer au choix de leur assurance. Les personnes plus jeunes sont, quant à elles, plus soucieuses de leur budget qui bien souvent est moindre : ils sont donc prêts à consacrer un peu plus de temps pour pouvoir bénéficier d'un tarif plus avantageux leur permettant de réaliser quelques économies. En ce qui concerne les personnes plus âgées, ils se trouvent un peu dans le même cas de figure que les personnes de moins de 30 ans : ils ont souvent une petite retraite et disposent de davantage de temps. Ils tendent donc à privilégier le tarif au détriment du gain de temps. Pour ce qui est du capital mobilier assuré, des avis d'experts ont confirmé l'intuition selon laquelle, à partir de cette variable, une certaine forme de richesse peut être extrapolée. Ainsi, les clients les plus aisés, moins sensibles au prix de leur assurance, auront un capital mobilier assuré élevé tandis que les personnes dont le patrimoine (ou les revenus) sont plus faibles sont plus sensibles aux variations de prix et disposeront d'un capital mobilier assuré plus faible.

Cette élasticité prend donc la forme suivante :

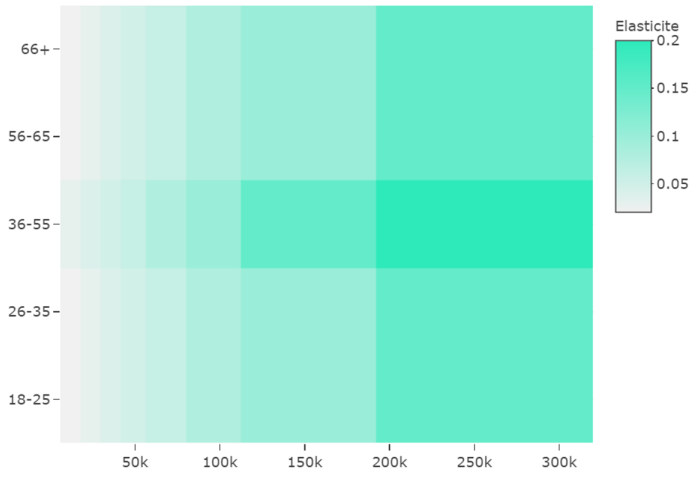


FIGURE 6.1 – Matrice d'élasticité

Les valeurs ont été calibrées en cohérence avec le montant de la prime totale (DDE + vol). Ainsi, une élasticité de 20% représente une différence de quelques euros (environ 20 euros).

Un terme correctif a également été ajouté aux locataires, la variable qualité ressortant comme impactante dans la modélisation de l'élasticité dans plusieurs mémoires. Par ce biais, il est supposé que les locataires, de par le caractère temporaire d'occupation de leur logement, sont moins attentifs au prix de leur assurance et privilégient ainsi le gain de temps.

6.1.2 Les assureurs

Concernant les assureurs, l'hypothèse est posée que tous commercialiseront un même produit constitué uniquement d'une garantie vol et d'une garantie DDE.

Dans le cadre de la vente de ce produit, ces derniers devront fixer un taux de frais permettant de couvrir tant les chargements de gestion que d'acquisition. Au cas particulier, il a été considéré que tous les assureurs avaient un même taux de frais de 20%.

L'ensemble de ces hypothèses peut, bien sûr, être discutée. Cependant, dans le cadre de ce mémoire, ce sont les valeurs présentées précédemment qui ont été retenues.

6.2 La tarification traditionnelle versus le GLM à l'adresse

Une fois les hypothèses du marché concurrentiel implémentées, il convient maintenant de procéder à l'analyse des résultats obtenus lors de la compétition entre le modèle de tarification traditionnelle (5.1) et le GLM à l'adresse (5.2).

		Groupe d'apprentissage			Groupe de test		
		Seul	En concurrence		Seul	En concurrence	
		Résultat	Résultat	Part de marché	Résultat	Résultat	Part de marché
DDE	*Tarification traditionnelle*	622K	-12K	52%	412K	23K	52%
	GLM à l'adresse	533K	-354K	48%	356K	-48K	48%
VOL	*Tarification traditionnelle*	527K	-50K	52%	239K	-53K	52%
	GLM à l'adresse	671K	-243K	48%	300K	-63K	48%

FIGURE 6.2 – Résultats du marché concurrentiel (GLM traditionnel versus GLM à l'adresse)

Les premiers résultats obtenus montrent une déficience du modèle à l'adresse par rapport à la tarification traditionnelle. Ainsi, pour le DDE, sur la base de test, un résultat de -48K est enregistré pour la tarification à l'adresse face à un résultat de 23K pour le modèle traditionnel. Un constat similaire est observé pour la garantie vol où les résultats des modèles traditionnel et à l'adresse sont respectivement de -53K et -63K. Les parts de marché sont, quant à elles, majoritairement prises par l'assureur effectuant une tarification traditionnelle, que ce soit pour la garantie vol ou DDE.

En outre, le résultat négatif de l'assureur procédant à la tarification traditionnelle pour la garantie vol montre une instabilité du tarif qui reste tout de même gênante. Cela provient probablement de la faible volumétrie de la base de données sur laquelle les modèles ont été appris.

Par ailleurs, l'observation de la répartition des différentes parts de marché par variable montre des phénomènes d'anti-sélection impliquant que des éléments-clés ont certainement été omis dans la modélisation. En effet, il existe, pour certaines variables un déséquilibre des parts de marché (que ce soit pour la garantie DDE ou vol). Cela concerne notamment les variables formule, sinistralité passée et capitaux assurés. Quelques exemples sur la base de test sont présentés ci-après :

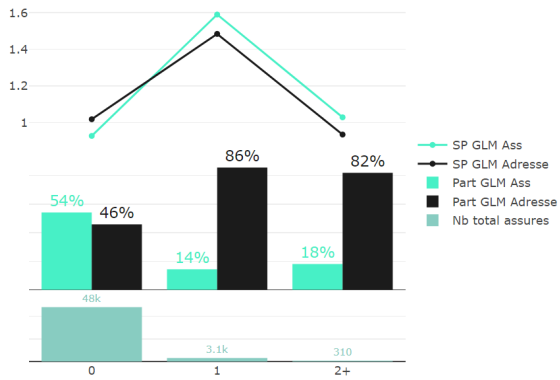


FIGURE 6.3 – Parts de marché de la sinistralité passée dégâts des eaux

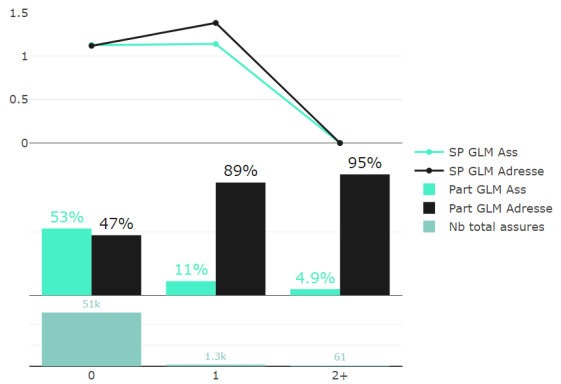


FIGURE 6.4 – Parts de marché de la sinistralité passée vol

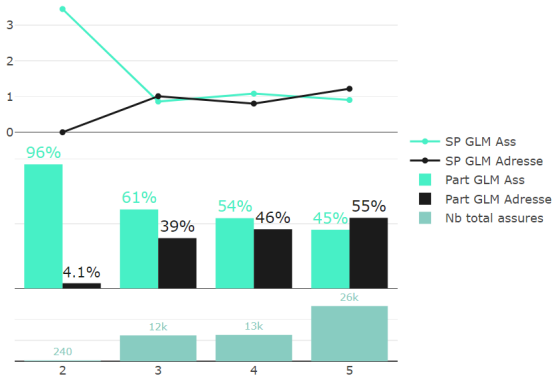


FIGURE 6.5 – Parts de marché de la formule pour la garantie dégâts des eaux

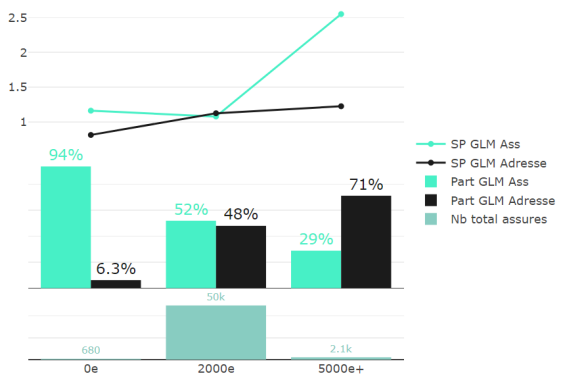


FIGURE 6.6 – Parts de marché du capital dépendance pour la garantie vol

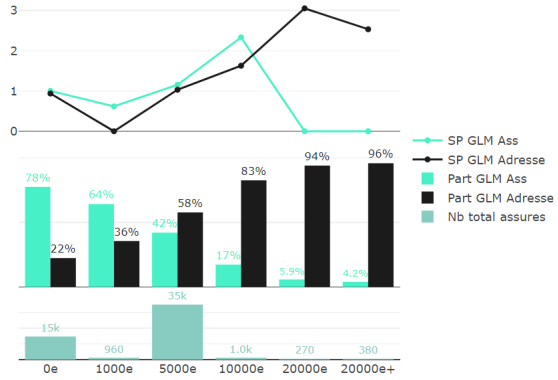


FIGURE 6.7 – Parts de marché du capital bijou pour la garantie vol

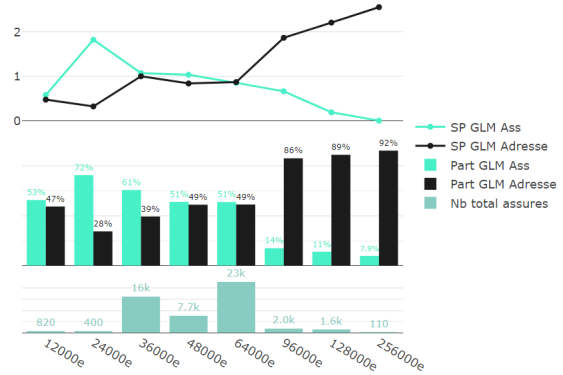


FIGURE 6.8 – Parts de marché du capital mobilier pour la garantie dégâts des eaux

Or, il semble logique qu'un assureur dispose de certaines de ces informations qui relèvent plus d'un choix du client que du risque en lui-même (la formule et les capitaux assurés). L'assureur a également le droit de refuser certains assurés ayant une sinistralité passée trop importante. Pour ce faire, il doit nécessairement avoir connaissance de la sinistralité passée du client. Un nouveau modèle à l'adresse est donc considéré : le GLM à l'adresse de base auquel sont ajoutées les variables formule, sinistralité passé et capitaux assurés (tant mobilier que bijou et dépendance). Les résultats sont ensuite mis à jour :

		Groupe d'apprentissage			Groupe de test		
		Seul	En concurrence		Seul	En concurrence	
		Résultat	Résultat	Part de marché	Résultat	Résultat	Part de marché
DDE	*Tarification traditionnelle*	622K	-109K	46%	412K	39K	46%
	GLM à l'adresse	547K	-161K	54%	376K	-18K	54%
VOL	*Tarification traditionnelle*	527K	-186K	47%	239K	-75K	47%
	GLM à l'adresse	640K	11K	53%	290K	10K	53%

FIGURE 6.9 – Résultats du marché concurrentiel avec ajout de la formule, de la sinistralité passée et des capitaux assurés dans le GLM à l'adresse (GLM traditionnel versus GLM à l'adresse)

Les résultats obtenus sur la base de test sont alors plutôt concluants pour la garantie vol avec un assureur à l'adresse qui présente un résultat de 10K comparé à l'assureur traditionnel enregistrant un résultat de -75K. La tarification à l'adresse semble donc être très efficace sur

cette garantie. Cela est certainement lié aux ajouts importants d'informations provenant de l'utilisation de la reconnaissance d'images.

En revanche, pour la garantie DDE, le GLM à l'adresse ne parvient pas à concurrencer de manière viable la tarification traditionnelle (résultat de -16K pour le GLM à l'adresse et de 39K pour la tarification traditionnelle) et ce, même si ce dernier enregistre plus de parts de marché que la tarification traditionnelle (54% contre 46% respectivement).

Une analyse des différentes parts de marché est donc à nouveau menée : il apparaît que la variable relative au nombre de pieces est assez déséquilibrée, entraînant un phénomène d'anti-sélection et ce, que ce soit pour la garantie DDE ou vol.

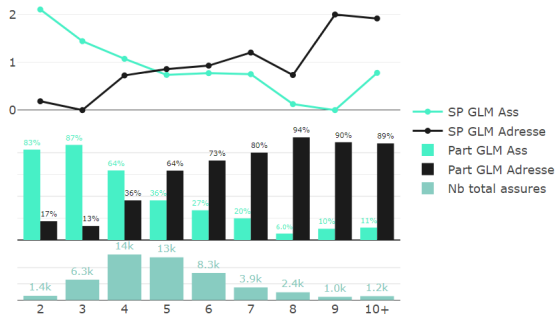


FIGURE 6.10 – Parts de marché du nombre de pièces pour la garantie dégâts des eaux

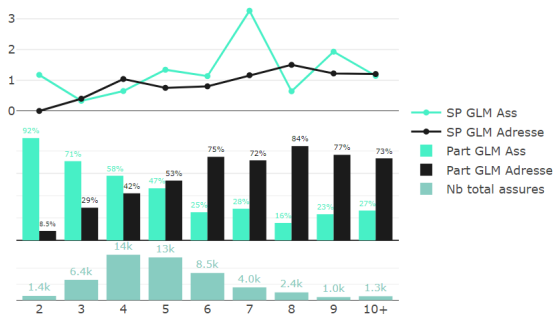


FIGURE 6.11 – Parts de marché du nombre de pièces pour la garantie vol

Quant aux autres variables (formule, sinistralité passée et capitaux assurés), elles ne présentent plus de déséquilibre.

Ainsi, après ajout de la variable nombre de pièces au modèle à l'adresse, le GLM comporte maintenant, en plus des données externes, les informations suivantes : formule, sinistralité passée, capitaux et nombre de pièces. Les résultats sont alors mis à jour :

		Groupe d'apprentissage			Groupe de test		
		Seul	En concurrence		Seul	En concurrence	
		Résultat	Résultat	Part de marché	Résultat	Résultat	Part de marché
DDE	*GLM Ass*	622K	-229K	34%	412K	103K	34%
	GLM Adresse	557K	184K	66%	386K	18K	66%
VOL	*GLM Ass*	527K	-289K	44%	239K	-85K	44%
	GLM Adresse	653K	215K	56%	294K	64K	56%

FIGURE 6.12 – Résultats du marché concurrentiel avec ajout de la formule, de la sinistralité passée, des capitaux assurés et du nombre de pièces dans le GLM à l'adresse (GLM traditionnel versus GLM à l'adresse)

Ils s'en trouvent améliorés que ce soit au niveau des parts de marché ou des résultats. Malheureusement, pour la garantie DDE, ce dernier ajout n'est toujours pas suffisant pour que la tarification à l'adresse surpasse la tarification traditionnelle.

6.3 La tarification traditionnelle versus la forêt aléatoire à l'adresse

Un raisonnement similaire à celui de la section 6.2 est adopté. Pour rappel, la forêt aléatoire à l'adresse figure en section 5.3. La comparaison du modèle traditionnel et de la forêt aléatoire à l'adresse donne les résultats suivants :

		Groupe d'apprentissage			Groupe de test		
		Seul	En concurrence		Seul	En concurrence	
		Résultat	Résultat	Part de marché	Résultat	Résultat	Part de marché
DDE	*GLM Ass*	622K	-391K	58%	412K	103K	58%
	RF Adresse	468K	42K	42%	328K	-125K	42%
VOL	*GLM Ass*	527K	-297K	58%	239K	-34K	58%
	RF Adresse	352K	-52K	42%	155K	-111K	42%

FIGURE 6.13 – Résultats du marché concurrentiel (GLM traditionnel versus forêt aléatoire à l'adresse)

		Groupe d'apprentissage			Groupe de test		
		Seul	En concurrence		Seul	En concurrence	
		Résultat	Résultat	Part de marché	Résultat	Résultat	Part de marché
DDE	*GLM Ass*	622K	-414K	57%	412K	96K	57%
	RF Adresse	465K	95K	43%	326K	-104K	43%
VOL	*GLM Ass*	527K	-296K	59%	239K	-32K	59%
	RF Adresse	367K	-20K	41%	162K	-98K	41%

FIGURE 6.14 – Résultats du marché concurrentiel avec ajout de la formule, de la sinistralité passée et des capitaux assurés dans la forêt aléatoire à l'adresse (GLM traditionnel versus forêt aléatoire à l'adresse)

		Groupe d'apprentissage			Groupe de test		
		Seul	En concurrence		Seul	En concurrence	
		Résultat	Résultat	Part de marché	Résultat	Résultat	Part de marché
DDE	*GLM Ass*	622K	-467K	56%	412K	91K	56%
	RF Adresse	393K	235K	44%	297K	-62K	44%
VOL	*GLM Ass*	527K	-333K	58%	239K	-34K	58%
	RF Adresse	361K	49K	42%	159K	-82K	42%

FIGURE 6.15 – Résultats du marché concurrentiel avec ajout de la formule, de la sinistralité passée, des capitaux assurés et du nombre de pièces dans la forêt aléatoire à l'adresse (GLM traditionnel versus forêt aléatoire à l'adresse)

La lecture de ces résultats fait ressortir un constat : un phénomène de surapprentissage est observé pour les forêts aléatoires et ce, malgré les optimisations effectuées. En effet, quel que soit le scénario (figures 6.13, 6.14 ou 6.15), il est possible de noter que les résultats en concurrence du modèle à l'adresse sur la base d'apprentissage sont bien meilleurs que ceux constatés sur la base de test. Une explication possible à ce phénomène pourrait être liée à la faible volumétrie de la base de données à disposition qui ne permet pas d'apprendre correctement et sans surapprentissage le modèle.

En ce sens, aucune conclusion certaine ne peut être apportée concernant l'apport potentiel du *machine learning* dans la tarification à l'adresse : des études supplémentaires devront être effectuées pour déterminer si une forêt aléatoire à l'adresse peut surpasser le GLM à l'adresse tout en concurrençant la tarification traditionnelle.

6.4 Conclusion et limites

Toutes ces analyses soulignent un certain nombre de points concernant la tarification à l'adresse :

- **1)** Certaines **variables (formule, sinistralité passée et capitaux assurés)** s'avèrent **indispensables** à l'établissement d'une tarification à l'adresse correcte : sans l'ajout de ces variables, il ne semble pas possible que la tarification à l'adresse puisse concurrencer de manière viable la tarification traditionnelle ;

- **2)** Une autre variable, le **nombre de pièces**, aurait dû être captée par la reconnaissance d'images mais fait malheureusement **l'objet d'un phénomène d'anti-sélection**. Cela vient probablement du fait que, dans le cadre de ce mémoire, seule la surface plane de la maison a été récupérée via la reconnaissance d'images, occultant ainsi une information tarifaire très importante de surface : le nombre d'étages, le nombre de pièces, la surface réelle de la maison, etc. Des travaux plus poussés pourraient cependant pallier ce problème (en allant plus loin dans le processus de reconnaissance d'images et en utilisant, par exemple, des images *Street View*) ;

- **3)** L'**impact de la tarification à l'adresse** est **différent en fonction des garanties**. Ainsi, si pour la garantie vol, les résultats sont plutôt concluants, ils le sont moins en ce qui concerne la garantie DDE. Cette différence s'explique notamment par le fait que, pour la garantie vol, les variables externes n'approchent pas seulement les variables tarifaires usuelles mais apportent aussi de nouvelles informations concernant le risque (et ce, notamment avec la reconnaissance d'images). Cela n'est pas le cas (ou peu) pour la garantie DDE. Pour tenter d'améliorer la tarification à l'adresse de la garantie DDE, il pourrait être envisagé d'utiliser et de tester des données autres que celles retenues dans le cadre de ce mémoire ;

- **4)** Il n'est pas possible de conclure de manière satisfaisante quant à l'apport des méthodes complexes dans le cadre de la tarification à l'adresse. Pour ce faire, il conviendrait de poursuivre et d'approfondir les travaux déjà réalisés.

Pour revenir à la problématique initiale du mémoire, à savoir « La tarification à l'adresse peut-elle concurrencer la tarification traditionnelle ? », les résultats présentés dans ce mémoire doivent être nuancés : bien qu'encourageants, ils montrent quelques limites et il semble, à ce stade, difficile d'avoir une tarification basée uniquement sur l'adresse de l'assuré. Enfin, compte tenu de la taille relativement faible de la base de données utilisée dans le cadre de ce mémoire, il ne peut être donné de conclusion générale et il paraît opportun de poursuivre et compléter ces travaux en utilisant une base de données plus conséquente pour confirmer ces premiers résultats.

Conclusion

Dans un marché toujours plus concurrentiel, les assureurs doivent faire face à un défi de taille : améliorer l'expérience de leurs clients dans le but de se démarquer. Une manière d'y parvenir est de simplifier le processus de souscription client, un processus long et fastidieux qui impose au client de répondre à de multiples questions ce qui le conduit très souvent à abandonner le processus avant sa fin.

Ce mémoire s'est donc intéressé à la simplification de ce processus pour un produit MRH maison constitué de deux garanties : vol et dégâts des eaux. L'objectif était de mettre en place une tarification avec pour seule variable l'adresse, et ce, à l'aide de données externes (majoritairement de l'*Open Data*). A cette fin, différents travaux ont été réalisés. Les résultats de ces travaux, bien qu'encourageants, ont toutefois montré quelques limites.

En effet, il est apparu, qu'à ce jour, il semblait encore difficile de mettre en place une tarification basée uniquement sur l'adresse de l'assuré, certaines informations relatives aux choix de l'assuré ou à sa sinistralité passée devant absolument être fournies pour que la tarification à l'adresse soit viable. Par ailleurs, le nombre de pièces, autre élément essentiel à la tarification, a, dans le cadre de ce mémoire, été difficilement capté par les données à l'adresse (reconnaissance d'images). Ce faisant, un phénomène d'anti-sélection est apparu au sein du marché concurrentiel. Ce phénomène lié à l'absence d'information sur le nombre de pièces pourrait cependant être supprimé par l'intégration du *Street View* dans la reconnaissance d'images. Enfin, d'autres ajouts de variables aux modèles à l'adresse pourraient également participer à l'amélioration de cette tarification à l'adresse notamment pour la garantie DDE qui reste tout de même peu satisfaisante : il pourrait, par exemple, être possible d'intégrer à l'étude la base des permis de construire, la base DPE ou encore la variable année de construction, trois éléments pouvant donner des informations quant à la vétusté des maisons.

Toutefois, bien que comportant des limites, ces travaux ont également démontré que la tarification à l'adresse pourrait être une solution d'avenir. En effet, il est, par exemple, apparu que les données à l'adresse permettaient tant d'approcher les variables tarifaires usuelles que de compléter la modélisation du risque par l'apport de nouvelles informations. Cela s'est révélé particulièrement vrai pour la garantie vol avec les données issues de la reconnaissance d'images (isolement de la maison, distance à la station de police la plus proche, etc).

Pour terminer, compte tenu de la taille relativement faible de la base de données utilisée dans le cadre de ce mémoire, il est important de souligner qu'il ne peut être tiré de conclusion générale quant à la performance de la tarification à l'adresse par rapport à la tarification traditionnelle. Il paraît en effet opportun de poursuivre et compléter ces travaux en utilisant une base de données plus conséquente pour confirmer ces premières tendances.

Bibliographie

[1] P. Aillot. *Modèle linéaire généralisé*. 2020.

[2] L. Breiman. Bagging predictors. *Machine Learning*, 24(2) :123–140, 1996.

[3] L. Breiman. Random forests. *Machine Learning*, 45(1) :5–32, Oct 2001.

[4] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and regression trees*. Wadsworth International Group, 1984.

[5] A. Charpentier. *Computational Actuarial Science with R*. 2016.

[6] O. CHIHI. Sensibilité du taux de résiliation au prix en assurance mrh occupant et simulation du portefeuille. 2011.

[7] I. Danton. Classement auto et mrh 2019. *L'Argus de l'assurance*, 2019.

[8] M. de la Transition écologique. *La fonction touristique des territoires : facteur de pression ou de préservation de l'environnement ?* 2017.

[9] F. F. de l'assurance. *L'assurance française - Données clés 2019*. 2019.

[10] F. F. de l'assurance. *Rapport 2019*. 2019.

[11] G. Eldin. *Construction d'un indicateur de Valeur Client et optimisation tarifaire en assurance non-vie*. 2018.

[12] C. Fort. *Tarification MRH à l'adresse*. 2020.

[13] R. Henckaerts, M.-P. Côté, K. Antonio, and R. Verbelen. Boosting insights in insurance tariff plans with tree-based machine learning methods. *North American Actuarial Journal*, 25(2) :255–285, 2020.

[14] G. V. Kass. An exploratory technique for investigating large quantities of categorical data. *Applied Statistics*, 29(2) :119, 1980.

[15] R. Laïly. *Tarification non-vie sur R*. 2021.

[16] G. Lucas. *Tarification Multirisques Immeuble, Le Big Data au service de la Simplicité Client et de la sophistication tarifaire de l'assureur*. 2019.

[17] J. N. Morgan and J. A. Sonquist. Problems in the analysis of survey data, and a proposal. *Journal of the American Statistical Association*, 58(302) :415–434, 1963.

[18] J. A. Nelder and R. W. M. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3) :370–384, 1972.

[19] OCDE. *Government at a Glance 2019*. 2019.

[20] E. D. Portal. *Open Data Maturity Report 2019*. 2019.

[21] F. Vermet. *Apprentissage statistique : une approche connexionniste*. 2020.