Mémoire présenté devant l'ENSAE Paris
pour l'obtention du diplôme de la filière Actuariat
et l'admission à l'Institut des Actuaires
le 15/03/2022

Par : **Doudou Cissé**

Titre : **Construction de micro-zoniers en assurance MRH à l'aide d'outils de Data Science**

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

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

Membres présents du jury de la filière

Nom :

Membres présents du jury de l'Institut des Actuaires

Entreprise : AXA Direct Assurance

Signature :

Directeur du mémoire en entreprise :

Nom : Alice TAN
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

Ecole Nationale de la Statistique et de l'Administration Economique (ENSAE)
5, avenue Henry Le Chatelier - 91120 PALAISEAU, FRANCE

Résumé

Ce mémoire a été réalisé dans le cadre d'une étude sur l'amélioration des zoniers en assurance multirisques habitation. Un zonier est un découpage d'un territoire en zones, en fonction de leurs expositions aux risques. Un zonier permet ainsi de prendre en compte l'environnement géographique où évolue le contrat. Notre étude présente la construction de nouveaux zoniers sur le coût moyen de la garantie dégât des eaux et sur la fréquence vol, ceci pour les appartements de la France métropolitaine. Notre zonier est appelé micro-zonier car étant réalisé à un maillage très fin autour de l'adresse.
La méthodologie ne s'appuie pas sur une approche classique mais plutôt sur la création, par nous-même, d'un découpage de la France et sur l'utilisation de modèle de Machine Learning pour la prédiction des résidus géographiques.
La segmentation du portefeuille des assurés est un facteur important dans la tarification des contrats d'assurance. Cette segmentation est réalisée grâce à des variables dites tarifaires. L'ajout de la dimension géographique à ces variables tarifaires est une technique de plus en plus explorée en assurance multirisques habitation. La prise en compte de l'effet géographique se fait souvent au niveau commune ou département. Dans ce mémoire, il s'agit de voir si les performances des modèles de tarification peuvent être améliorées lorsque le zonier est construit à une maille plus petite que les découpages administratifs habituels.

Mots-clés : zonier, machine learning, modèles linéaires généralisés, lissage spatial.

Abstract

This study is produced with the goal of improving zoners in household insurance. A zoner is a division of a territory into zones, according to their risk exposure. Thus, a zoner makes it possible to take into account the environment in which the contract operates. Our study presents the construction of a new zone on the average cost of the water damage guarantee for apartments. This new zone is called micro-zone because its construction is based on a very fine mesh.

The methodology is not based on a classic approach but rather on the creation, by ourselves, of a division of France and on the use of machine learning models for the prediction of geographic residuals.

The segmentation of the portfolio of policyholders into homogeneous groups is an important factor in the pricing of insurance contracts. This segmentation is carried out using so-called tariff variables. Adding the geographic dimension to these rate variables is a technique that is increasingly explored in property and casualty insurance. In the literature, the geographical effect is often taken into account at the municipality or department level.

In this thesis, it is a question of seeing if the segmentation performance can be improved when the zoner is built on a mesh smaller than the usual administrative divisions.

Keywords : zoning, machine learning, generalized linear models, spatial smoothing.

Remerciements

Je tiens tout d'abord à remercier AXA Direct Assurance pour m'avoir accueilli au sein de la Direction Technique. Je remercie également l'ensemble de l'équipe MRH pour son accueil chaleureux.
Je remercie tout particulièrement Alice TAN mon manager, Benjamin Wallyn et Jing Chen, pour m'avoir accompagné tout le long du stage.
Je tiens également à remercier Monsieur Nicolas Baradel, mon référent académique.
Enfin, je remercie toute personne ayant contribué, de près ou de loin, à la rédaction de ce mémoire.

Point de confidentialité

Les chiffres renseignés dans les parties statistiques descriptives et modélisation de ce mémoire sont obtenus en modifiant les vrais chiffres de Direct Assurance, ceci pour protéger la confidentialité des données de l'entreprise. Cependant, les conclusions et interprétations ne changent pas.

Table des matières

Note de synthèse

L'assurance est une opération par laquelle l'assureur s'engage à verser une prestation à l'assuré en cas de réalisation d'un risque, en contre partie du paiement d'une prime ou cotisation. En assurance multirisques habitation (MRH) qui constitue l'objet de ce présent mémoire, le risque porte principalement sur les dégâts des eaux, les vols, les incendies et explosions, les tempêtes, la grêle... Il apparaît donc que la prime est un élément essentiel dans le travail de l'assureur et sa bonne estimation est cruciale. La prime pure est définie comme le produit entre la fréquence moyenne de sinistre et le coût moyen de sinistre. Une bonne estimation de ces deux dernières grandeurs est nécessaire pour bien appréhender la prime. Avec le développement de l'open data, les entreprises ont maintenant accès à une importante quantité de données. Le besoin d'exploitation de ces données dites externes (car différentes des variables tarifaires habituelles) a boosté l'essor des techniques d'amélioration des modèles de tarification : c'est le cas des zoniers. Un zonier peut être défini comme un découpage d'un territoire en zones géographiques, en fonction de leurs expositions à un risque donné. En assurance habitation, la prise en compte de la dimension géographique du risque représente un grand enjeu. Dans la majeure partie des cas, les zoniers sont créés avec un maillage à un niveau commune ou code postal. L'inconvénient d'un tel maillage est que ces mailles sont très grandes et ne permettent donc pas de capter avec une grande précision le signal géographique. L'objectif de ce mémoire est de construire un micro-zonier, c'est-à-dire un zonier à une maille très fine en exploitant au mieux un grand nombre de variables externes. Cette présente section en constitue la synthèse. Elle sera structurée comme suit :
- d'abord nous allons présenter le cadre de l'étude,
- ensuite nous présenterons les principaux résultats obtenus.

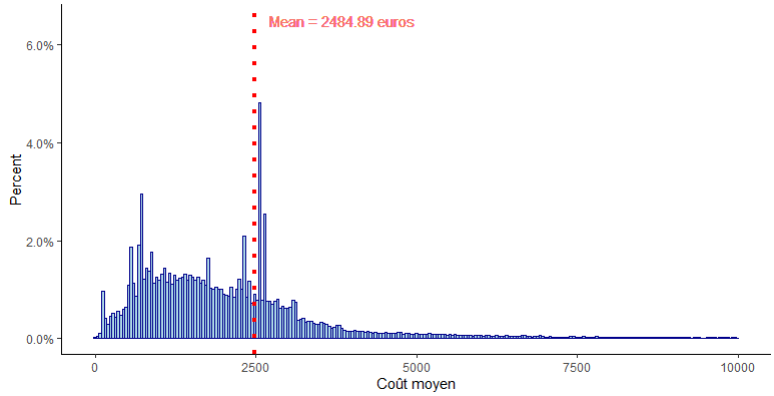
■ Cadre général de l'étude :

Nous avons effectué notre stage de fin d'études à Direct Assurance. Ce dernier est le nom commercial de la compagnie d'assurance directe Avanssur, une filiale du groupe Axa. Elle fut créée en 1992. A ses débuts, Direct Assurance était spécialisée en assurance auto. Actuellement la compagnie propose aussi des assurances multirisques habitation (MRH), des contrats moto et très récemment des assurances Santé. Notre stage réalisé dans cette structure s'inscrit dans le cadre de l'utilisation de données externes (open data) pour la construction de micro-zoniers. Concrètement, il s'agit de construire un micro-zonier sur le coût moyen de la garantie dégâts des eaux et sur la fréquence vol, en exploitant au mieux un grand nombre de variables externes. La complexité de cette étude nous a poussés à utiliser les techniques d'analyse géospatiale. Nous pouvons citer quelques exemples de manipulations spatiales utilisées dans ce travail : jointure de bases de données au niveau adresse, représentation de cartes, découpages de zones, gestion des systèmes de coordonnées géographiques, création de shapefiles, intersections de zones et agrégations de zones. L'algorithme de Voronoi nous a permis de découper la France en de petites zones ou polygones. Cet algorithme fournit un découpage d'un plan en des cellules à partir d'un ensemble discret de points appelés « germes ». Cette méthode permet ainsi le partitionnement d'un plan contenant n points

en n polygones de telle sorte que chaque polygone contienne exactement un point générateur et que chaque point d'un polygone donné soit plus proche de son point générateur que de tout autre point du plan.

- Résultats obtenus :

La figure suivante représente l'histogramme du coût moyen pour la garantie dégât des eaux appartements. En moyenne, le coût d'un sinistre s'établit à 2484 euros. Une bonne partie des sinistres ont un coût qui se situe entre 100 euros et 2000 euros environ. Les sinistres supérieurs à 5000 euros apparaissent faibles. Les coûts de sinistres nuls ont été enlevés de la base, pour les besoins de la modélisation.



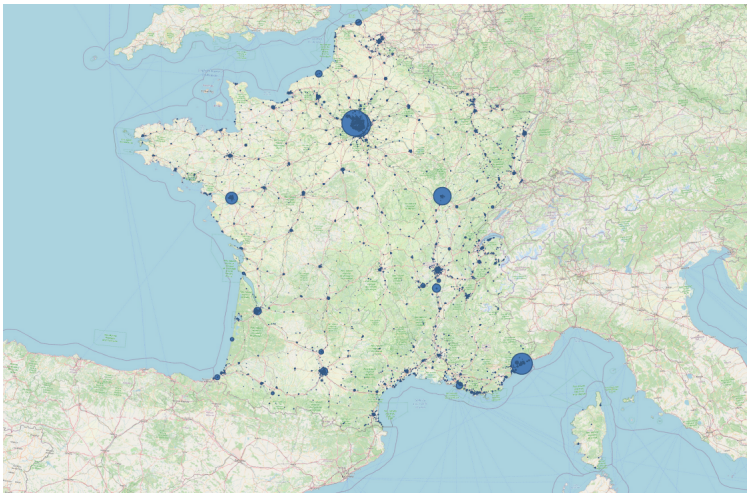
Histogramme du coût moyen de sinistre (en pourcent)

Les statistiques descriptives sur la fréquence vol révèlent qu'au niveau des contrats comptabilisant au moins un sinistre vol, 99,02% ont enregistré exactement un sinistre. Les contrats ayant deux sinistres représentent 0,96%. Les contrats faisant l'objet de trois sinistres ne pèsent que 0,02%. Aucun contrat de notre base de données n'a eu plus de 3 sinistres.

Nombre de sinistres VOL

Nous avons complété cette partie de statistiques descriptives classiques par des statistiques descriptives spatiales. L'idée est de faire ressortir les variations de montants de sinistres

suivant les différents endroits. La carte qui suit représente la répartition du coût moyen de sinistres sur le territoire français. L'ampleur du sinistre est proportionnelle à la taille des points en bleus, chaque point représentant un contrat. Plus un point est grand, plus le montant de sinistre qu'il représente est élevé. Pour améliorer la visualisation, nous avons ajouté le fonds de carte de France. Ce qui permet d'identifier plus clairement les endroits à forts coûts de sinistres.



Répartition spatiale du coût moyen dégât des eaux appartements (avec fond de carte)

Concernant la modélisation, la première étape consiste à modéliser le coût moyen de sinistre dégât des eaux et la fréquence de sinistre vol en excluant les variables géographiques, ceci avec les données de la base d'apprentissage. Ensuite, il s'agira d'isoler l'information géographique contenue dans les résidus d'un modèle linéaire généralisé de prime pure hors variables géographiques. Pour opérer un tel isolement, nous modéliserons ces résidus avec les variables externes à l'aide d'un modèle de machine learning, le catboost. Les prédictions de résidus obtenues avec le catboost, nommées résidus spatiaux dans la littérature, seront ensuite scindées en différentes classes. Ces dernières constitueront le zonier final.

- Modélisation hors variables externes :

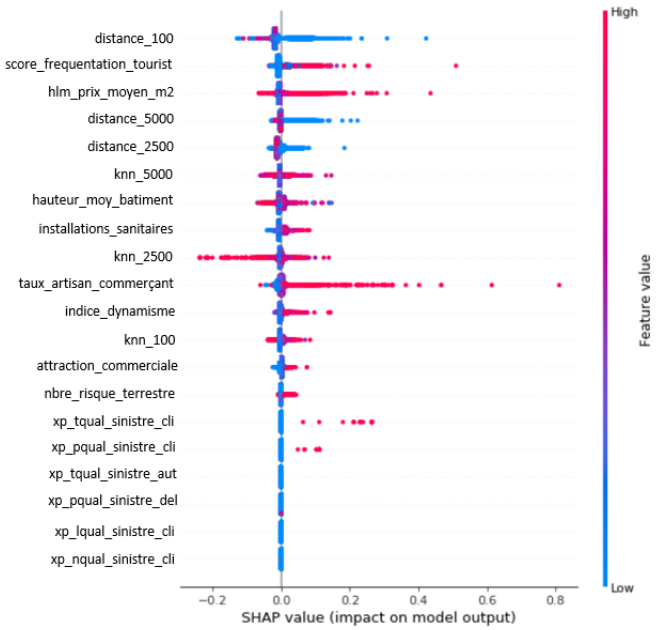
Globalement, les résultats du modèle confirment les statistiques descriptives. Comparées aux résidences principales, les résidences secondaires ont tendance à enregistrer des coûts moyens de sinistre dégât des eaux plus élevés. Le nombre de pièces de l'appartement est positivement corrélé avec le montant moyen de sinistres. Lorsque l'occupant est propriétaire, il a plus de chance d'avoir des coûts de sinistre élevés en moyenne. Les années 2017 et 2018 comptabilisent des sinistres plus couteux, relativement à l'année 2016. Les sociétés enregistrent en moyenne des coûts de sinistres plus élevés que les particuliers. La loi lognormale présente les meilleurs résultats.

En ce qui s'agit des résultats du glm de la fréquence vol expliquée par les variables internes, la loi binomiale négative fournit les meilleures performances. Pour la variable explicative "âge du client" la classe des clients dont l'âge est inférieur à 39 ans représente la modalité

de référence. Il ressort du glm que les coefficients de toutes les autres modalités de la variable "âge du client" sont négatifs et significatifs au seuil de 5%. On en déduit que les assurés jeunes ont tendance à subir plus de sinistres vol en moyenne. Les résidences principales ont des fréquences de sinistre vol plus élevées comparées aux résidences secondaires. Comme suggéré par nos statistiques descriptives, la fréquence de sinistre vol augmente avec le nombre de pièces de l'appartement. De même, lorsque l'occupant est le propriétaire les sinistres vol sont plus nombreux en moyenne. Comparés aux autres types de distributeurs, les courtiers voient leurs contrats distribués enregistrer les fréquences de sinistre les plus grandes.

■ Modélisation de l'effet géographique :

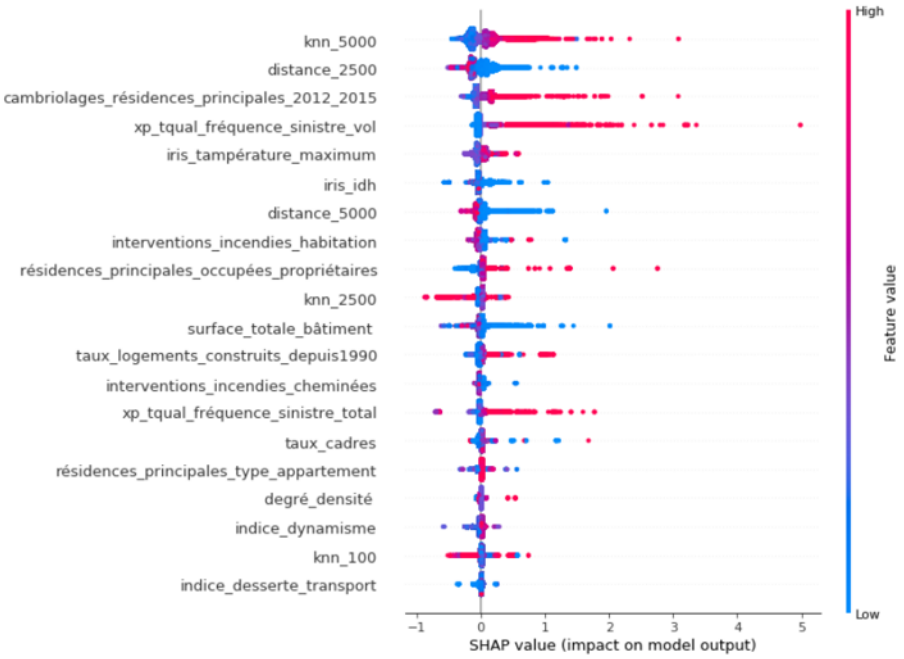
Pour modéliser l'effet géographique, nous avons utilisé un modèle de machine learning appelé Catboost. Le CatBoost est un algorithme de gradient boosting basé sur les arbres de décision. L'idée est d'effectuer un apprentissage séquentiel qui fonctionne sur le principe d'un ensemble, où chaque modèle tente de corriger les erreurs du modèle précédent. L'avantage du catboost réside dans son pouvoir de prédiction. C'est l'un des modèles de machine learning les plus puissants à ce jour. La figure suivante fournit les résultats de la modélisation pour le coût moyen dégât des eaux. Les variables externes qui contribuent le plus à l'explication du résidus sont la distance par rapport aux plus proches voisins, le score de fréquentation touristique autour de l'adresse (segment de rue à fort trafic de touristes et avec points d'intérêts touristiques à proximité), le prix moyen au m2 à l'adresse, la hauteur moyenne de tous les bâtiments référencés aux alentours, les installations sanitaires en dehors du logement, le taux d'artisans, de commerçants et de chefs entreprises logeant dans la zone, l'indice de dynamisme Particuliers, l'indice d'attraction commerciale de l'endroit, le nombre risque terrestre et les variables expérience client.



Résultats du catboost pour le coût moyen dégât des eaux

Concernant les performances du catboost, le gini s'élève à 16,13% sur la base d'apprentissage

et 10,29% sur la base test. Les prédictions sur les résidus obtenues sont découpées en 20 classes suivant les quantiles pour constituer le zonier final .
Les résultats du catboost qui explique les résidus de la fréquence vol par les variables externes sont très intuitifs. En effet, les variables explicatives géographiques relatives à la criminalité et au vol sont sorties significatives. La variable "cambriolages des résidences principales entre 2012 et 2015" est positivement corrélée aux résidus du glm vol. Ce résultat comporte une certaine logique. En outre, les milieux à températures élevées comptabilisent les résidus de sinistre vol les plus grands en moyenne. Ceci pourrait être expliqué par le fait qu'en milieu froid, les gens ont tendance à rester davantage dans leurs domiciles, ce qui peut dissuader les voleurs. D'autres variables criminogènes telles que les interventions pour incendies dans les habitations, les interventions pour incendies causés par les cheminées apparaissent aussi significatives. Les résidences principales occupées par les propriétaires comptent aussi des résidus élevés.

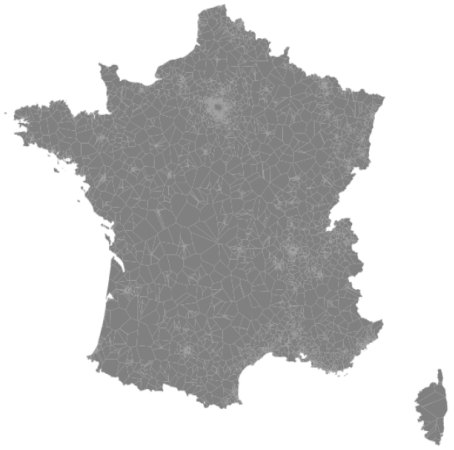


Résultats du catboost fréquence vol

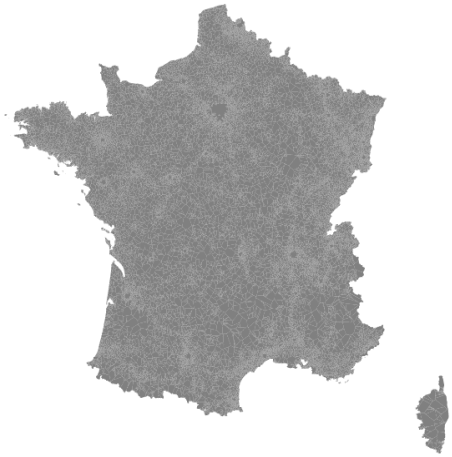
■ Découpage de la France en polygones par la méthode de Voronoi :

Cette partie aborde la construction de la maille du zonier. Le choix de la maille la plus adaptée est important puisqu'il détermine les tailles des zones sur lesquelles l'effet géographique sera estimé. Nous avons choisi, dans le cadre de cette étude, d'explorer les mailles fines. Il convient de noter que la taille des découpages (ou polygones) dépend de la densité des contrats d'assurance. Les zones à fortes densités de contrats auront des superficies très petites car comme dit plus haut chaque contrat va engendrer un polygone constitué des points de l'espace qui lui sont plus proches. Les zones à faibles densités auront des polygones assez grands. Les cartes qui suivent présentent nos découpages spatiaux de la France grâce à l'algorithme de Voronoi. Ces cartes sont obtenues après intersection spatiale entre nes découpages bruts de Voronoi et le contour de la France. Le découpage pour la fréquence vol

est beaucoup plus fin que celui du coût moyen. Ceci est dû est au fait que le découpage pour la fréquence vol est réalisé sur toutes les observations de la base d'apprentissage alors que pour le coût moyen le découpage est fait sur les contrats à coûts de sinistre non nuls.

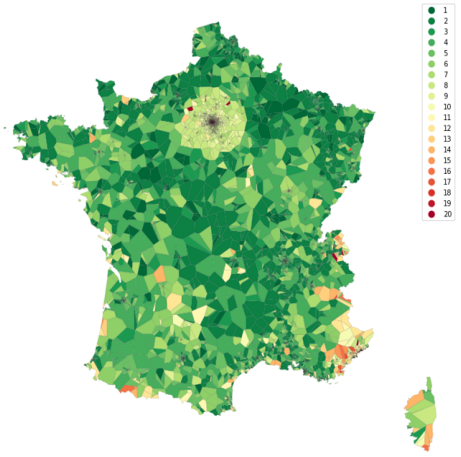


Notre découpage de la France en des polygones de Voronoi (coût moyen dégât des eaux)

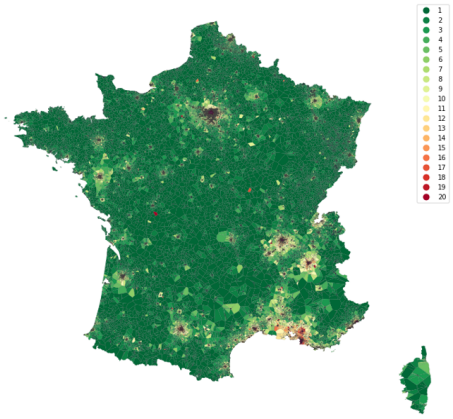


Notre découpage de la France en des polygones de Voronoi (fréquence vol)

Les résidus spatiaux découpés en classes puis représentés sur une carte donnent le zonier final. Les cartes qui suivent représentent nos micro-zoniers obtenus. A gauche, nous avons le micro-zonier pour le coût moyen dégâts des eaux et à droite le micro-zonier pour la fréquence vol. Concernant la légende, plus le numéro de la classe est élevé, plus le signal géographique associé est grand. Ainsi la couleur verte est associée aux zones à faible intensité du phénomène étudié (coût pour le dégât des eaux et fréquence pour le vol). Les zones à plus grandes intensités sont celles coloriées en rouge. Pour le micro-zonier vol, on recense les fréquences de sinistre les plus élevés autour de Paris, Lyon, Grenoble, Marseille, Nice, Toulouse, Bordeaux, Nantes et Rennes. Ces tendances sur les vols dans les domiciles sont similaires aux chiffres du ministère de l'intérieur français. Pour le micro-zonier vol, l'intensité du phénomène reste importante autour de Paris mais la tendance n'est plus la même sur toute l'étendue de la France. Le micro-zonier dégâts des eaux est moins lisse que celui de la fréquence vol. En effet, on note beaucoup plus de disparités au niveau du micro-zonier coût moyen. Ceci peut s'expliquer par fait que ce micro-zonier compte beaucoup moins de polygones car étant construit uniquement sur les contrats à coûts de sinistre non nuls.



micro-zonier coût moyen dégât des eaux



micro-zonier fréquence vol

- Intégration des micro-zoniers dans les modèle de coût moyen et de fréquence vol sur la base de test :

Nous avons mesuré l'apport du zonier sur la base de validation. Pour ce faire, une première modélisation du coût moyen est effectuée, sans le micro-zonier. Ensuite, une deuxième modélisation du coût moyen et de la fréquence col est réalisée, cette fois-ci en ajoutant le micro-zonier dans les variables explicatives tarifaires. La comparaison des performances de ces deux modèles a permis de mesurer l'apport du micro-zonier. Ces performances sont aussi comparées à celles du zonier actuel d'AXA qui est construit à la maille commune (ou code Insee).

Comparaison des indices de Gini normalisés :

a) Pour les modèles de coût moyen dégât des eaux :
Nous avons évalué l'indice de Gini d'une part sur toutes les observations de notre base de données de test et d'autre part sur uniquement les observations bien géocodées de la base de test. L'avantage d'une telle segmentation permet d'appréhender l'importance de bien géocoder les domiciles des assurés dans la construction des micro-zoniers. On remarque que pour le coût moyen dégât des eaux, le zonier d'AXA est meilleur que le micro-zonier sur toutes les observations. Mais lorsqu'on considère uniquement les appartements bien géocodés, notre micro-zonier donne de meilleures performances que le zonier actuel d'AXA. Ce qui justifie l'importance de bien géocoder les contrats. On note aussi l'importance du lissage qui a permis de diminuer le sur-apprentissage de nos modèles.

	Zonier actuel d'AXA	Notre Micro-zonier avant lissage	Notre Micro-zonier lissé
Gini normalisé (base d'apprentissage)	0,2987	0,2998	0,3028
Gini normalisé (base de test)	0,2874	0,2785	0,2787
Gain p/r au zonier d'AXA (sur test)	-	-0,88%	-0,87%

Performances des différents zoniers sur toutes les observations de la base de test

	Zonier actuel d'AXA	Notre Micro-zonier avant lissage	Notre Micro-zonier lissé
Gini normalisé (base d'apprentissage)	0,3332	0,3289	0,3104
Gini normalisé (base de test)	0,2965	0,2976	0,3056
Gain p/r au zonier d'AXA (sur test)	-	0,11,88%	0,91%

Performances des différents zoniers sur uniquement les observations bien géocodées de la base de test

b) Pour les modèles de fréquence vol :
Grâce à son niveau de détail granulaire, le micro-zonier fréquence vol se révèle plus performant sur tous les cas de figures (observations globales comme observations bien géocodées). Le modèle avec le micro-zonier constuits sur des polygones très petits cible les risques de manière plus précise et offre une segmentation du risque géographique réel plus homogène à l'intérieur des zones. Ce qui témoigne de l'importance d'étudier les risques au niveau local, surtout pour le cas de l'assurance habitation.

	Zonier actuel d'AXA	Notre Micro-zonier
Gini normalisé (base d'apprentissage)	0,4208	0,3953
Gini normalisé (base de test)	0,3611	0,3669
Gain p/r au zonier d'AXA (sur test)	-	0,58%

Performances des différents zoniers sur toutes les observations de la base de test

	Zonier actuel d'AXA	Notre Micro-zonier
Gini normalisé (base d'apprentissage)	0,3707	0,3961
Gini normalisé (base de test)	0,3293	0,3399
Gain p/r au zonier d'AXA (sur test)	-	1,07%

Performances des différents zoniers sur uniquement les observations bien géocodées de la base de test

Executive summary

Insurance is an operation whereby the insurer undertakes to pay a benefit to the insured in the event of the occurrence of a risk in return for the payment of a premium or contribution. In multi-risk home insurance (MRH), which is the subject of this brief, the risk mainly concerns water damage, theft, fire and explosion, storms, hail, etc. It is therefore clear that the premium is an essential element in the daily life of the insurer and its correct estimation is crucial. The pure premium is defined as the product of the average loss frequency and the average loss cost. A good estimate of the latter two quantities is necessary to understand the premium. With the development of open data, companies now have access to a very large amount of data. The need to exploit this so-called external data (because it is different from the usual tariff variables) has boosted the development of techniques for improving tariff models : this is the case of zoners. A zoning can be defined as a division of a territory into geographical zones, according to their exposure to a given risk. In home insurance, taking into account the geographical dimension of the risk is a major issue. In most cases, the zonings are created with a mesh at the commune or postcode level. The disadvantage is that these meshes are very large and therefore do not allow the geographical signal to be captured with great precision.

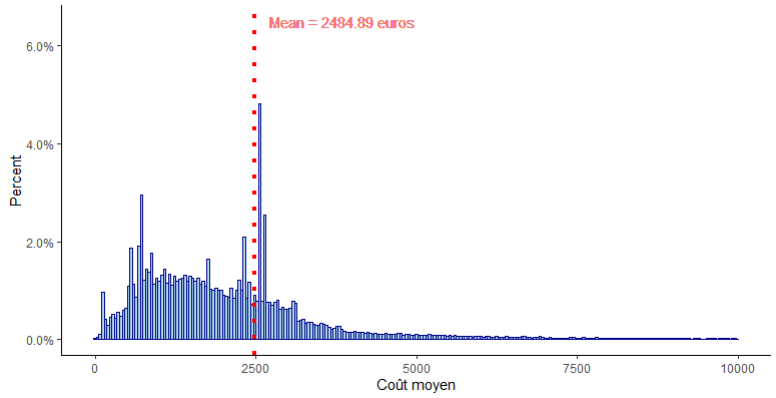
The objective of this thesis is to build a micro-zoning, i.e. a zoning at a very fine mesh size, by making the best use of a large number of external variables. This part constitutes the synopsis. It will be structured as follows
- first we will present the framework of the study
- then we will present the main results obtained.

■ General framework of the study :

We carried out our end-of-study internship at Direct Assurance. This company is the commercial name of the direct insurance company Avanssur, a subsidiary of the Axa group. It was created in 1992. At the beginning, Direct Assurance was specialised in car insurance. Today, the company also offers comprehensive home insurance, motorbike insurance and very recently health insurance. Our internship in this structure is part of the use of external data (open data) for the construction of zoning. In concrete terms, it is a question of building a micro-zoning on the average cost of water damage cover and on the frequency of theft, by making the best use of a large number of external variables. The complexity of this study led us to use geospatial analysis techniques. We can cite some examples of spatial manipulations used in this work : joining databases at the address level, map representation, area partitioning, geographic coordinate system management, shapefile creation, area intersections and area aggregations. The Voronoi algorithm allowed us to divide France into small areas or polygons. This algorithm provides a division of the map into cells from a discrete set of points called "seeds". This method thus allows the partitioning of a plane containing n points into n polygons in such a way that each polygon contains exactly one generating point and that each point of a given polygon is closer to its generating point than to any other point of the plane. General framework of the study.

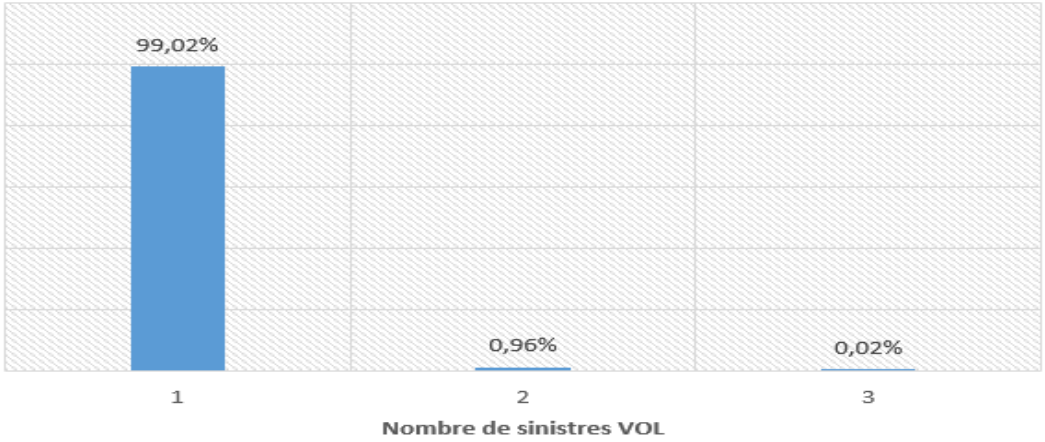
■ Results obtained :

The following gigure shows the histogram of the average cost for water damage cover for flats. On average, the cost of a claim is 2484 euros. A large proportion of claims have a cost of between 100 euros and about 2000 euros. Claims above 5000 euros appear low. The costs of zero claims have been removed from the base for modelling purposes.



Histogram of average claim costs (in percent)

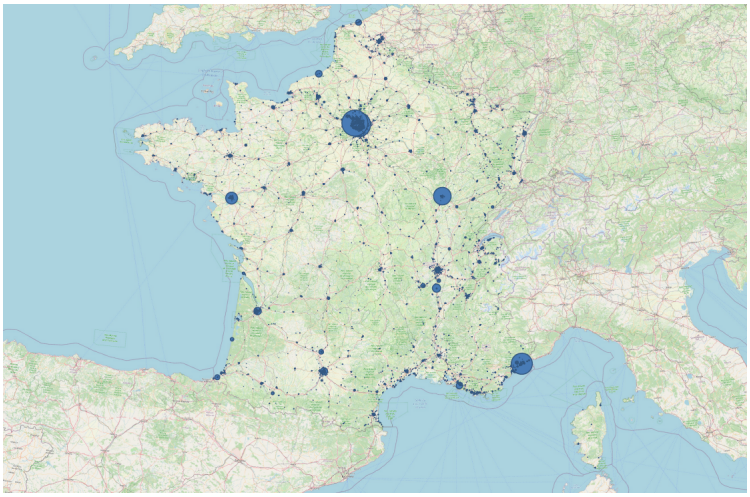
descriptive statistics on the frequency of theft show that 99.02% of contracts with at least one theft claim recorded exactly one claim. Contracts with two claims represent 0.96%. Contracts with three claims account for only 0.02%. No contract in our database had more than 3 claims.



Number of claims THEFT

We have supplemented this classical descriptive statistics section with spatial descriptive statistics. The idea is to highlight the variations in the amount of claims according to the different locations. The following map represents the distribution of the average cost of claims over the French territory. The size of the claim is proportional to the size of the blue points. The larger a point is, the higher the amount of loss it represents. To improve

the visualisation, we have added the background map of France. This makes it possible to identify more clearly the places with high claims costs.



Spatial distribution of the average cost of water damage in flats (with map background)

The first step in modelling is to model the average cost of claims excluding geographical variables, using data from the learning base. Then, we will isolate the geographical information contained in the residuals of a generalized linear model of pure premium excluding geographical variables. To perform such an isolation, we will model these residuals with the external variables using a machine learning model, the catboost. The predictions of residuals obtained with the catboost, called spatial residuals in the literature, will then be split into different classes. These classes will constitute the final zoning.

- Modelling without external variables :

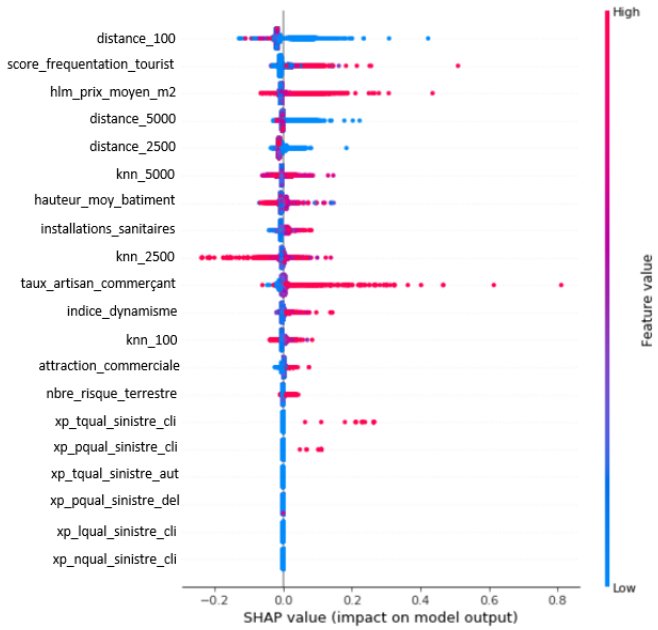
Overall, the results of the model confirm the descriptive statistics above. Compared to primary residences, secondary residences tend to have higher average water damage claim costs. The number of rooms in the flat is positively correlated with the average claim amount. When the occupant is a homeowner, he or she is more likely to have high claim costs. The years 2017 and 2018 show more expensive claims, compared to the year 2016. On average, companies have higher claims costs than individuals. It should be noted that extreme claims are discarded. The lognormal law shows the best results.

As regards the results of the glm of the theft frequency explained by the internal variables, the negative binomial distribution provides the best performance. For the explanatory variable "age of the customer" the class of customers with an age below 39 years is the reference modality. The glm shows that the coefficients of all the other modalities of the "age of the client" variable are negative and significant at the 5% level. This suggests that younger policyholders tend to have more theft claims on average. On average, primary residences have higher theft claim frequencies compared to secondary residences. As suggested by our descriptive statistics, the frequency of theft claims increases with the number of rooms in the flat. Similarly, when the occupant is the owner, the average number of theft claims is

higher. Compared to other types of distributors, brokers have the highest claim frequencies in their distributed policies.

■ Modelling the geographical effect :

To model the geographical effect, we used a machine learning model called Catboost. CatBoost is a gradient boosting algorithm based on decision trees. The idea is to perform sequential learning that works on the principle of an ensemble, where each subsequent model tries to correct the errors of the previous model. The advantage of catboost lies in its predictive power. It is one of the most powerful machine learning models to date. The following figure shows the results of the modelling. The external variables that contribute the most to the explanation of the residual are the distance to the nearest neighbours, the tourist traffic score around the address (street segment with high tourist traffic and points of tourist interest nearby), the average price per m2 at the address, the average height of all the buildings listed in the vicinity, the sanitary facilities outside the dwelling, the rate of craftsmen, shopkeepers and entrepreneurs living in the area, the dynamism index for individuals, the commercial attraction index of the place, the number of land risks and the customer experience variables.

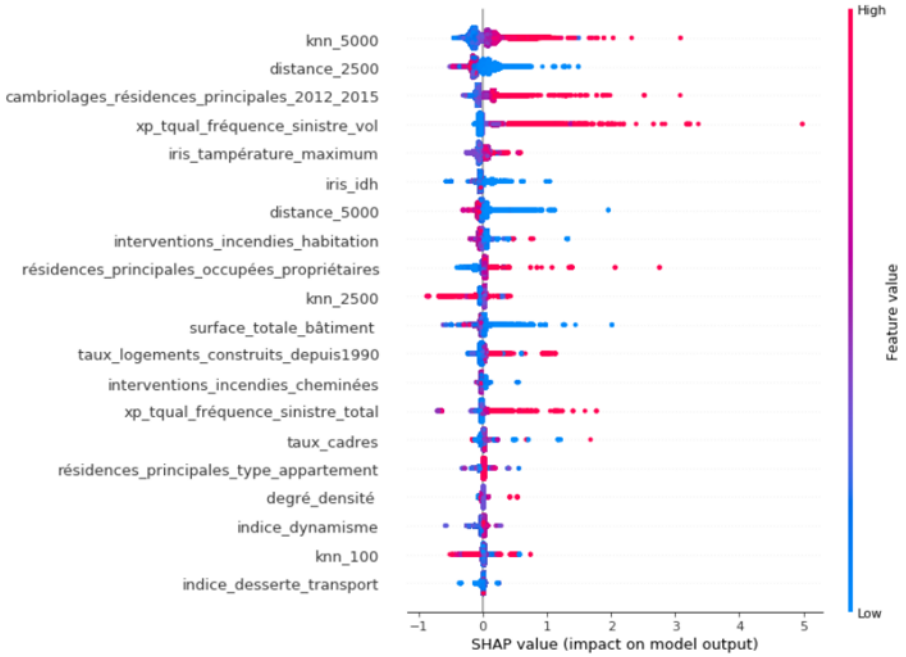


Catboost results for average cost

Concerning the performance of the catboost, the gini amounts to 16.13% on the learning base and 10.29% on the test base. The predictions on the residuals obtained are divided into 20 classes according to the quantiles.

The results of the catboost that explains the residuals of the theft frequency by the external variables are very intuitive. Indeed, the geographical explanatory variables for crime and theft are significantly outliers. The variable "burglaries of main residences between 2012 and 2015" is positively correlated with the residuals of the theft glm. There is some logic to this result. High-temperature environments have the highest theft residuals in the largest

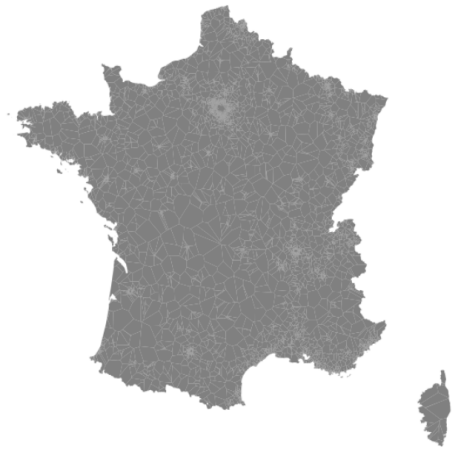
flats on average. This could be explained by the fact that in cold environments people tend to stay in their homes more, which may deter thieves. Other criminogenic variables such as fire interventions in dwellings, interventions for fires caused by chimneys also appear significant. Owner-occupied primary residences also have high residues.



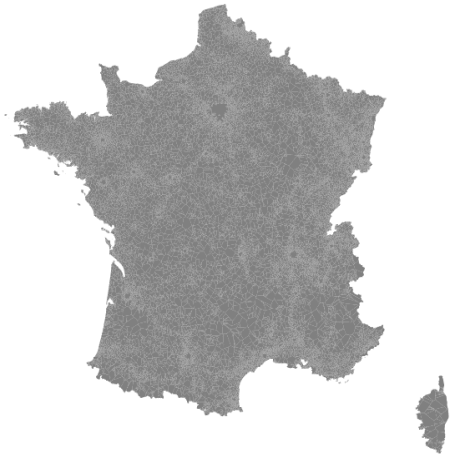
theft frequency catboost results

■ Division of France into polygons using the Voronoi method :

This part deals with the grid size of the area. The choice of the most suitable mesh size is important as it determines the sizes of the zones for which the geographical effect will be estimated. In this study, we have chosen to explore fine grids. It should be noted that the size of the grids (or polygons) depends on the density of insurance contracts with non-zero claims costs. Areas with a high density of contracts will have very small areas because, as mentioned above, each contract will generate a polygon made up of the nearest points in space. Low density areas will have fairly large polygons. This is a particularity of the average cost modelling. The following map presents our spatial division of France using the Voronoi algorithm. This map is obtained after the spatial intersection between the raw cut (see appendices) and the outline of France.

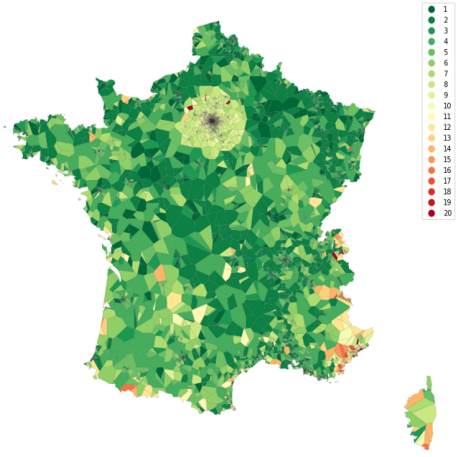


Our division of France into Voronoi polygons (average cost of water damage)

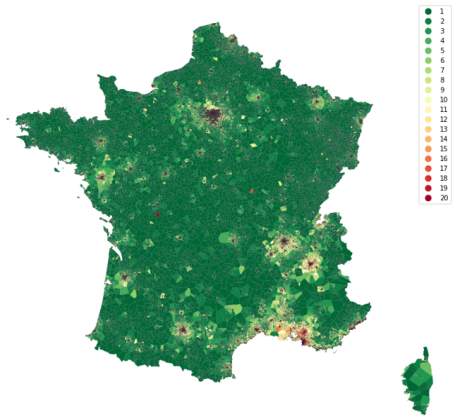


Our division of France into Voronoi polygons (theft frequency)

The spatial residuals cut into classes and then plotted on a map give the final zoning. The following maps represent our resulting micro-zoners. On the left, we have the average water damage cost micro-zoning and on the right the theft micro-zoning. Concerning the legend, the higher the class number, the greater the associated geographical signal. Thus the green colour is associated with areas of low intensity of the phenomenon studied (cost for water damage and frequency for theft). The areas with higher intensities are those coloured in red. For the micro-zone of theft, the highest frequency of damage is found around Paris, Lyon, Grenoble, Marseille, Nice, Toulouse, Bordeaux, Nantes and Rennes. These trends for theft from homes are similar to the French Ministry of the Interior figures (see Appendix 7). For the theft micro-zone, the intensity of the phenomenon remains high around Paris, but the trend is no longer the same throughout France. the damage micro-zone is less smooth than that of the theft frequency. In fact, there are many more disparities in the average cost micro-zones. This can be explained by the fact that this micro-zone has far fewer polygons because it is based solely on contracts with non-zero claims costs.



micro-zoning average cost of water damage



micro theft frequency

■ Integration of the zoning in the average cost model on the validation data set :

We have measured the contribution of the zoning on the validation basis. To do this, an initial modelling of the average cost is carried out, without the zoning system. Then, a second modelling of the average cost is carried out, this time adding the zoning in the tariff explanatory variables. The comparison of the performances of these two models has made it possible to measure the contribution of the zoning. These performances are also compared to those of AXA's current zoning system, which is constructed at the commune level (or Insee code).

Comparison of Normalized Gini indice :

a) For the average water damage cost models :
We evaluated the Gini index on all the observations in our test database and on only the well geocoded observations. The advantage of such a segmentation is that it allows us to understand the importance of a good geocoding in the construction of a micro-zoning. We notice that for the average cost of water damage, our AXA zoning is better than the micro-zoning on all observations. But when we consider only the well geocoded flats, our micro-zoning gives better performances than AXA's current zoning. We also note the importance of the smoothing which allowed to decrease the overlearning of our models.

	AXA's current zoning	Our Micro-zoning	Our Micro-zoning + Smoothing
Normalized Gini on train	0,2987	0,2998	0,3028
Normalized Gini on test	0,2874	0,2785	0,2787
Gain in relation to AXA's zoning (on test data)	-	-0,88%	-0,87%

Performance of the different zoning on all observations on the test base

	AXA's current zoning	Our Micro-zoning	Our Micro-zoning+ Smoothing
Normalized Gini on train	0,3332	0,3289	0,3104
Normalized Gini on test	0,2965	0,2976	0,3056
Gain in relation to AXA's zoning (on test data))	-	0,11,88%	0,91%

Performance of the different zonings on only the well geocoded observations of the test database

b) For theft frequency models :
to its granular level of detail, the theft frequency micro-zoning performs better in all cases (global observations as well as well geocoded observations). The model with the micro-zoning targets the risks more precisely and offers a more homogeneous segmentation of the real geographical risk within the zones. This underlines the importance of studying risks at the local level, especially in the case of home insurance.

	AXA current zoning	Our Micro-zoning
Normalized Gini on train	0,4208	0,3953
Normalized Gini on test	0,3611	0,3669
Gain in relation to AXA's zoning (on test data set)	-	0,58%

Performance of the different zonings on all observations in the test data base

	AXA current Zoning	Our Micro-zoning
Normalized Gini on train	0,3707	0,3961
Normalized Gini on test	0,3293	0,3399
Gain in relation to AXA's zoning (on test data set)	-	1,07%

Performance of the different zonings on only the well geocoded observations of the test database

Introduction

L'assurance est une opération par laquelle l'assuré reçoit, moyennant un paiement (prime ou cotisation), une prestation de la part de l'assureur en cas de réalisation d'un risque. En assurance multirisques habitation (MRH) par exemple, le risque porte principalement sur les dégâts des eaux, les vols, les incendies et explosions, les tempêtes, la grêle... A partir de cette définition, on voit que la prime est un élément essentiel dans le quotidien de l'assureur et sa bonne estimation est cruciale. Cette nécessité de bien estimer la prime est confortée par l'inversion du cycle de production, phénomène selon lequel contrairement à la situation classique où le producteur d'un bien connaît son coût de production et peut par conséquent proposer un prix de vente, l'assureur demande une prime d'assurance à l'assuré avant de connaître le montant réel des sinistres que l'assuré est susceptible de subir. La prime pure est définie comme étant le produit entre la fréquence moyenne de sinistre et le coût moyen de sinistre. Dès lors, une estimation correcte de ces deux dernières grandeurs est nécessaire pour bien appréhender la prime. Il existe plusieurs techniques d'amélioration de l'estimation de la fréquence et du coût de sinistre. Nous pouvons citer le recours à des modèles plus performants, l'ajout de variables externes... Du fait du développement de l'open data, les entreprises ont maintenant accès à une très importante quantité de données. L'exploitation de ces données dites externes (car différentes des variables tarifaires classiques) a permis l'essor de techniques sophistiquées d'amélioration des modèles de tarification : c'est le cas des zoniers. Un zonier est un découpage d'un territoire en zones géographiques, en fonction de leurs expositions aux risques. En assurance habitation, la prise en compte de la dimension géographique du risque est d'une grande importance. Elle permet de prendre en compte les spécificités géographiques de l'environnement dans lequel évolue le contrat, permettant ainsi un potentiel gain d'informations. L'objectif de ce mémoire est de construire des micro-zoniers, c'est-à-dire des zoniers réalisés à une maille très fine en exploitant au mieux un grand nombre de variables externes. Grâce à son niveau de détails, un micro-zonier offre une segmentation du risque géographique réel plus graduelle. Puis, il s'agira d'étudier la contribution de ces micro-zoniers dans l'amélioration des modèles de tarification. Ce qui permettra de décider de l'intégration ou non des micro-zoniers dans la tarification. Plus précisément, nous construirons un micro-zonier pour le coût moyen de sinistre de la garantie dégât des eaux et un autre micro-zonier pour la fréquence vol, ceci pour les appartements.
Le mémoire est structuré comme suit :
- Le premier chapitre aborde le cadre général de l'étude. Nous présenterons brièvement notre structure de stage Direct Assurance et le secteur de l'assurance Multirisques Habitation. Nous expliquerons ensuite notre méthodologie d'élaboration de micro-zonier.
- Le second chapitre présente les données utilisées dans cette étude et nous mettrons ensuite en évidence quelques statistiques descriptives.
- Le troisième chapitre constitue la partie modélisation.

1 Cadre général de l'étude

Dans ce premier chapitre, nous commençons par une brève présentation de Direct Assurance. Ensuite nous parlerons de l'assurance MRH de manière globale. Nous aborderons également la notion de Système d'Information Géographique et son importance dans le domaine de l'assurance. La dernière partie de ce chapitre présentera la méthodologie utilisée pour construire notre micro-zonier.

1.1 Brève présentation de Direct Assurance

Direct Assurance est le nom commercial de la compagnie d'assurance Avanssur qui est une filiale du groupe Axa. Historiquement elle a été créée en 1992. Spécialisée dans l'assurance auto, Direct Assurance propose également des assurances multirisques habitation (MRH), des assurances moto et très récemment des assurances Santé. Direct Assurance opère en assurance directe. Une compagnie d'assurance directe est une structure d'assurance qui conclut ses contrats uniquement en vente directe sur Internet.

1.2 Généralités sur l'assurance non vie

Dans cette partie, il s'agira de passer en revue la branche de l'assurance non vie. Il est minutieux de commencer par définir notre domaine d'étude et les concepts utilisés dans ce mémoire. Leur compréhension permet de faire ressortir l'aspect actuariel de notre travail. L'Article 1101 du Code Civil français définit le terme « contrat » comme étant « une convention par laquelle une ou plusieurs personnes s'obligent, envers une ou plusieurs autres, à donner, à faire ou à ne pas faire quelque chose. ». Cette définition englobe toutes les formes de contrat de manière générale. Le contrat d'assurance obéit à cette définition générale, tout en ajoutant quelques précisions.
Plus spécifiquement, plusieurs définitions existent pour un contrat d'assurance. Un contrat d'assurance peut être défini comme un contrat par lequel une partie (le souscripteur) se fait promettre, pour son compte ou celui d'un tiers par une autre partie (l'assureur), une prestation généralement pécuniaire en cas de réalisation d'un risque, moyennant le paiement d'une prime. De leur côté, Fromenteau et Petauton (2017) définissent un contrat d'assurance comme un accord financier passé entre un organisme (assureur) [qui prend l'engagement irrévocable de verser des prestations monétaires en cas de sinistres] et un souscripteur (assuré) [prenant l'engagement de verser à dates convenues des primes ou cotisations]. Il s'agit donc d'un engagement consensuel, car reposant sur la rencontre des volontés des parties. Il ressort de cette définition trois termes clés : prime, risque et prestation. Une prime d'assurance est le montant que paie le souscripteur à l'assureur en échange de la garantie fournie par le contrat. Autrement dit, il s'agit de la somme payée par le souscripteur pour bénéficier de la couverture des risques prédéfinis avec la compagnie d'assurance. Le risque est un événement aléatoire dont la réalisation ne dépend pas de la volonté des signataires. L'événement déclencheur du ou des sinistres (déterminés) est extérieur et indépendant de l'assuré. L'aléa réside ainsi à la fois sur le montant et la date de versement des flux. La pres-

tation représente ce que verse l'assureur au bénéficiaire en cas de réalisation du risque. La prestation dans l'assurance est donc liée au dommage (sinistre). En outre, le contrat d'assurance fait intervenir plusieurs acteurs : l'assureur, le souscripteur, l'assuré et le bénéficiaire. Suivant les contextes, le souscripteur peut coïncider avec l'assuré et le bénéficiaire. Mais ce n'est pas toujours le cas. L'assureur est la partie qui s'engage à indemniser le bénéficiaire du contrat d'assurance en cas de réalisation du sinistre. Du point de vue juridique, l'assureur peut être une société ou un intermédiaire (courtier, agent général). De son côté, le souscripteur est la partie qui signe le contrat et s'engage à effectuer les versements de prime. L'assuré, quant à lui, est l'entité qui subit le risque assuré. Dès lors, on voit que l'assuré n'est pas nécessairement le souscripteur, par exemple dans le cas d'une assurance contracté par un tiers. Le bénéficiaire est la partie qui reçoit les prestations en cas de réalisation du sinistre. Le contrat d'assurance a aussi un caractère aléatoire. L'aléa réside dans le fait que l'événement qui déclenche la prestation est incertain. En assurance non vie, le caractère aléatoire porte sur la réalisation d'un évènement. Une fois ratifié, le contrat d'assurance oblige les deux parties à respecter leur engagement. Par exemple, l'assureur est tenu de verser une indemnité si jamais le risque spécifié par le contrat se réalise. De son côté, l'assuré se doit de payer la prime et de renseigner des informations exactes, non erronées. La bonne réussite d'un contrat d'assurance dépend donc la bonne foi des engagés. Par bonne foi, on entend des facteurs tels que l'honnêteté et la loyauté.

1.2.1 Les branches de l'assurance non vie

Le Code des Assurances dénombre, dans l'article R 321-1, définit les branches d'assurance. D'après cette classification, l'assurance non vie regroupe 18 branches :
1- Accidents corporels : comprend les accidents du travail, les maladies professionnelles et les assurances de type « garantie des accidents de la vie » ou « garantie corporelle du conducteur ».
2- Maladie : concerne les assurances complémentaires santé et les contrats couvrant les garanties en cas d'incapacité temporaire de travail, d'invalidité partielle ou définitive, des suites de maladie ou d'accident.
3- Corps de véhicules terrestres : couvre les dommages auxquels font face les véhicules terrestres. Il est important de noter que cette branche ne concerne pas les véhicules ferroviaires.
4- Corps de véhicules ferroviaires : regroupe des dommages que subissent les véhicules ferroviaires.
5- Corps de véhicules aériens : regroupe les dommages auxquels les véhicules aériens sont exposés.
6- Corps de véhicules maritimes, lacustres et fluviaux : assure les sinistres liés aux véhicules fluviaux, véhicules lacustres et véhicules maritimes.
7- Marchandises transportées : regroupe les marchandises transportées ou bagages, indépendamment du moyen de transport utilisé.
8- Incendie et éléments naturels : couvre les dommages, subis par les biens (à l'exception de ceux cités dans les branches 3, 4, 5, 6 et 7), dus à un incendie, une explosion, une tempête, un élément naturel autre que la tempête, une énergie nucléaire et un affaissement de terrain.

9- Autres dommages aux biens : couvre les dommages, subis par les biens (à l'exception de ceux cités dans les branches 3, 4, 5, 6 et 7), provoqués par la grêle, le vol...
10- RC véhicules terrestres automoteurs : couvre les dommages subis par les véhicules terrestres automoteurs (RC automobile) et la responsabilité du conducteur.
11- RC véhicules aériens : assure les dommages liés à l'utilisation des véhicules aériens.
12- RC véhicules maritimes, lacustres et fluviaux : concerne la responsabilité des conducteurs de véhicules fluviaux, lacustres et maritimes.
13- RC générale : prend en compte toute responsabilité autre que celles mentionnées par les points 10, 11 et 12.
14- Crédit : cette branche couvre le crédit agricole, la vente à tempérament, le crédit à l'exportation, l'insolvabilité générale et le crédit hypothécaire.
15- Caution : la caution est la personne qui s'engage à assurer l'exécution d'une obligation prise par une personne (le débiteur) envers une troisième personne (le créancier), en cas de défaillance du débiteur.
16- Pertes pécuniaires diverses : cette branche comprend les pertes de bénéfices, les pertes de loyers ou de revenus, les pertes commerciales indirectes autres que celles mentionnées précédemment, les pertes pécuniaires non commerciales, les dépenses commerciales imprévues, les insuffisances de recettes, les risques d'emploi, la persistance de frais généraux, la perte de la valeur vénale, le mauvais temps et les autres pertes pécuniaires.
17- Protection juridique : permet à l'assuré, en cas de différend avec un tiers ou en cas de procédure judiciaire, de bénéficier d'une aide de la part de son assureur.
18- Assistance : assistance des individus en difficulté.

1.2.2 Les grandes phases de l'exercice d'activité d'assurance non vie

L'assurance non vie est rattachée à de nombreuses tâches diverses et variées. Les principales sont l'élaboration des contrats, la souscription, la tarification, le versement des primes, l'avenant éventuel au contrat, l'indemnisation, la gestion des recours et le calcul des provisions, la résiliation...
-élaboration des contrats : c'est une étape importante car elle permet de définir les droits et devoirs des parties qui s'engagent.
-souscription : à ce niveau, le souscripteur rempli un formulaire de déclaration et est tenu de fournir des réponses exactes à un certain nombre de questions. Ces réponses doivent permettre à l'assureur de se faire une idée sur l'ampleur du risque qu'il va endosser.
-tarification : c'est l'estimation de la valeur de prime que chaque souscripteur devra payer en moyenne. Pour y arriver, l'assureur essaie de segmenter au mieux son portefeuille. Le but étant d'identifier les critères pouvant expliquer la sinistralité future des assurés.
-versement des primes : la prime est réglée par le souscripteur qui s'engage depuis la signature du contrat à la verser. A cette étape, tout manquement de la part du souscripteur peut conduire l'assureur à résilier le contrat par l'envoi d'une lettre de mise en demeure.
-avenant au contrat : un avenant est une clause ajoutée au contrat d'assurance pour tenir compte de toute modification par rapport à ce qui était prévu dans le contrat initial. L'ave-

nant est donc la preuve que le contrat initial a été modifié. Ce qui permet d'éviter que le contrat de base soit complètement réécrit.
-indemnisation : en assurance non vie, l'indemnisation survient à la suite d'un sinistre. C'est le montant que l'assureur verse à l'assuré en cas de réalisation du risque couvert par le contrat. Un chargé d'indemnisation intervient pour comprendre les circonstances, situer les responsabilités et évaluer le montant de l'indemnisation.
-gestion des recours : le souscripteur a le droit de s'opposer aux modalités d'indemnisation s'il juge qu'il a été lésé (refus d'indemnisation de la part de l'assureur ou indemnisation insuffisante). Il peut même aller jusqu'à déclencher une action pénale contre l'assureur.
-calcul des provisions : les provisions représentent le montant total, évalué par l'assureur, suffisant pour permettre le règlement intégral de ses engagements vis à vis des assurés.
-résiliation : Dans la plupart des cas, les contrats d'assurance sont tacitement renouvelés d'une année sur l'autre. Cependant, l'assureur comme l'assuré peut choisir de résilier le contrat d'assurance sous certaines conditions et conformément aux règles prévues par le code des assurances. La possibilité de résilier un contrat est encadré par la loi. Par exemple, pour résilier le contrat, l'assureur doit envoyer une lettre recommandée à l'assuré au moins deux mois avant la date d'échéance. Pour les contrats conclus à des fins professionnelles, l'assureur doit résilier le contrat par lettre recommandée papier ou par électronique. Pour les autres cas de figure, l'assureur doit utiliser le support papier. La résiliation, de la part de l'assureur, doit être motivée, à l'exception des contrats relatifs aux activités professionnelles.

1.2.3 Mutualisation et segmentation

Deux concepts antagonistes s'opposent en assurance : mutualisation et segmentation. La mutualisation repose sur la loi des grands nombres. L'idée est de considérer un grand nombre de contrats afin de réduire le risque moyen. Ce qui permet de faire face à une grande variabilité de la réalisation des risques. Ainsi, la mutualisation suppose le partage des risques entre les individus de la population. De son côté, la segmentation est le fait de regrouper les risques en des ensembles homogènes. L'objectif étant d'avoir des profils de risques assez différents entre les groupes et homogènes dans les groupes. Le développement de l'open data ouvre la voie à de nouvelles techniques sophistiquées permettant des segmentations de plus en plus fines. Notre étude (élaboration de zonier avec l'open data) s'inscrit dans ce contexte.

1.3 cas de l'assurance MRH

1.3.1 Présentation de l'assurance MRH

Le contrat multirisques habitation (MRH) est une offre d'assurance qui propose aux assurés qui le souhaitent un large choix de garanties qui protégera leurs habitations. Pour les propriétaires de bâtiment, ce type d'assurance reste facultatif à l'exception de ceux qui sont en copropriété. La copropriété est définie comme étant « tout immeuble bâti ou groupe

d'immeubles bâtis dont la propriété est répartie, entre plusieurs personnes». Par contre, l'assurance MRH s'avère obligatoire en France pour les locataires qui entrent dans un logement. La loi française impose à ces derniers de souscrire, au minimum, un contrat d'assurance couvrant les risques relatifs au logement c'est-à-dire les dommages qui sont causés par un dégât des eaux ou un incendie. Il convient de noter que ce minimum légal ne couvre pas la responsabilité civile de l'assuré en cas de dommages corporels ou matériels causés à un tiers. Il ne protège pas non plus les biens personnels de l'assuré. Pour cette raison, l'assurance multirisques habitation propose plusieurs formules. D'une manière générale, les garanties des contrats d'assurance multirisques habitation se divisent en deux grands groupes. D'un côté, la garantie responsabilité civile vie privée qui couvre les dommages causés aux tiers par le souscripteur et de l'autre, le contrat qui protège le logement et les biens mobiliers qui s'y trouvent. Pour cette dernière catégorie, les garanties peuvent être de plusieurs types.
Notre description des garanties s'inspire de la documentation disponible sur les sites internet de la Fédération Française de l'assurance et du service public français. Nous avons les garanties dites de base et qui regroupent :
la garantie dégât des eaux :
Cette garantie couvre les dommages dûs à l'eau provenant d'une fuite d'eau, d'une rupture ou d'un débordement de canalisations non enterrées ; d'une fuite ou d'un débordement d'un radiateur, d'un appareil électroménager ou d'une baignoire ; d'un engorgement ou d'un débordement de gouttières ; des infiltrations d'eau sous les toits. Des exclusions existent. On peut citer les dégâts des eaux suite à un défaut d'entretien ou de construction (par exemple mauvais entretien d'une machine à laver à l'origine du sinistre), les dommages provoqués par la négligence de l'assuré, les dégâts des eaux résultant d'événements climatiques et naturels, les dégâts des eaux issus de l'humidité, de la condensation ou de la porosité.
La garantie incendie :
Selon les contrats et les compagnies d'assurances, les risques de base couverts peuvent varier légèrement. Généralement, cette garantie prend en charge les dommages consécutifs à un feu qui a une origine accidentelle, des explosions dues à un incendie, des feux déclenchés par la foudre, des dommages causés par la fumée, des implosions... Le périmètre couvert est l'habitation et son contenu dans les limites prévues par le contrat. Les dommages subis par les arbres et plantations ne sont généralement pas couverts. Quant aux exceptions, on peut citer les dommages dus à la négligence (par exemple lorsque l'incendie a pour origine un mégot de cigarette), les dommages subis par les biens tels que les télévisions et les matériels électroniques Toutefois, l'assuré peut supprimer ces exceptions en précédant à une extension des biens garanties. Ce qui se traduit par une majoration de la prime qu'il devra s'acquitter.
La garantie bris de glace :
Cette garantie prend en charge les dommages subis par les parties vitrées (baies vitrées, fenêtres, portes, cloisons, meubles constitués de verre, vitrines ...) du logement lorsqu'ils sont dus à un accident ménager, à une chute de grêle, à une tentative d'effraction... Pour que l'indemnisation ait lieu, il est nécessaire que l'ampleur des dégâts soit considérable. Les dégradations mineures telles que les rayures ne sont pas indemnisées. Il faut aussi noter que tous les objets vitrés ne sont pas systématiquement inclus dans cette garantie, certains

devront faire l'objet d'une extension. Concernant les exclusions, on peut noter les ampoules et autres éléments d'éclairage, la vaisselle, les objets en verre optique, les télescopes et les lustres.
La garantie vol et vandalisme :
Elle couvre les dégâts matériels dus à un vandalisme ou à un vol commis dans le logement. Plus précisément, cette garantie endosse les dommages causés par un vol par effraction ou escalade, un vol par menace ou violence, un cambriolage clandestin, un vol des clés du domicile, un vandalisme. Soulignons que certains assureurs proposent de garantir le vandalisme indépendamment du vol. En revanche, certains faits ne font pas l'objet d'une indemnisation, il s'agit notamment d'un vol commis avec la complicité du locataire, de vol ayant lieu dans une dépendance du domicile (caves et sous-sols, abris de jardins, garages...). Il convient de bien lire le contrat pour bien identifier tous les éléments non couverts par le contrat.
La garantie catastrophes naturelles :
Cette garantie est encadrée par la législation du pays en question. En France, le régime d'indemnisation des catastrophes naturelles a été créé par la loi du 13 juillet 1982. Avant cela, ces risques étaient très rarement couverts par les assureurs. Pour pouvoir faire l'objet d'un dédommagement, l'évènement doit être reconnu comme catastrophe naturelle par arrêté interministériel publié au journal officiel, ou par une demande de reconnaissance de l'état de catastrophe naturelle par le maire. La prime payée par les souscripteurs est uniforme sur l'ensemble du territoire géographique. Les dégâts habituellement couverts sont les inondations, les seimes, les sécheresses, les ouragans et cyclones, les tsunamis, les volcans. En plus de ces garanties de base, les compagnies d'assurance proposent aussi des garanties complémentaires pour couvrir des risques plus spécifiques : les dommages électriques, l'assurance de zones extérieures (jardin, piscine, véranda) . . .

1.3.2 Le marché du MRH en France

Dans la catégorie d'assurance MRH en France, la garantie incendie enregistre la part de coût moyen de sinistre la plus élevée en 2020, soit 53%. Elle est suivie de la garantie Tempête, grêle, neige qui représente 18% du coût moyen des sinistres MRH. Viennent ensuite la garantie vol et la garantie dégâts des eaux avec respectivement 12% et 7% du coût de sinistre total d'assurance MRH de 2020.

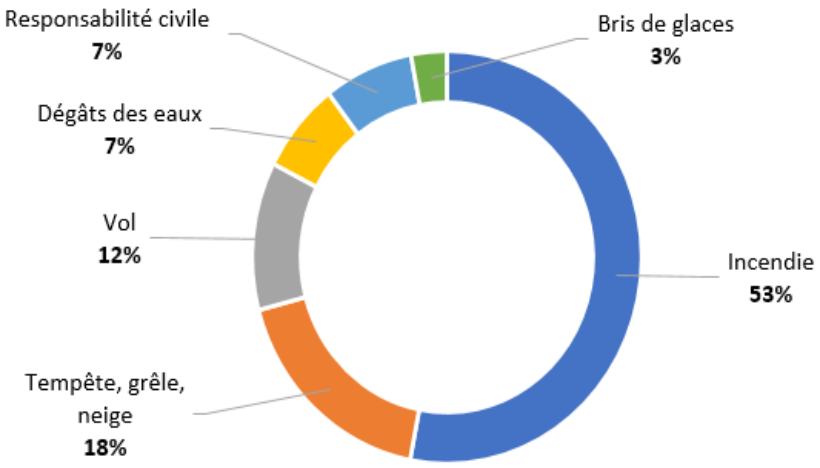


FIGURE 1 – MRH : coût moyen de sinistre par garantie en 2020 (Chiffres de la Fédération Française de l'assurance, figure de l'auteur)

Concernant les fréquences de sinistre, c'est la garantie dégât des eaux qui enregistre le niveau le plus élevé en 2020 en France avec un taux de 34,3 sinistres pour 1000 contrats dégâts des eaux. Ce taux correspond à une hausse des fréquences de sinistres dégâts des eaux de 3,1% par rapport à l'année 2019. La garantie tempête,grêle, neige présente la baisse de fréquence de sinistres la plus accentuée entre 2019 et 2020, soit -49,9%.

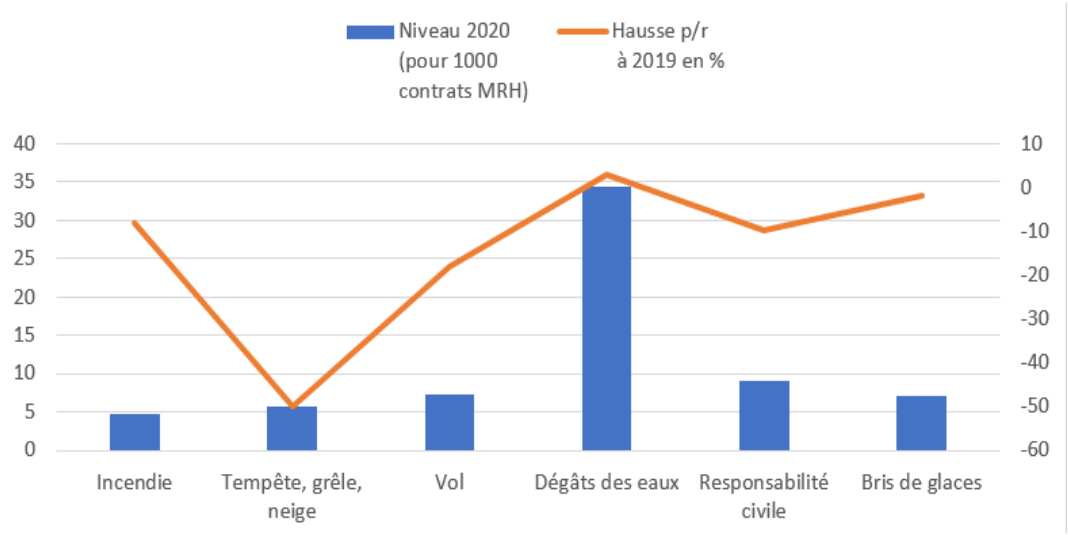


FIGURE 2 – MRH : fréquence de sinistre par garantie en 2020 (Chiffres de la Fédération Française de l'assurance, figure de l'auteur)

1.4 La tarification en assurance non vie

Notre modélisation de la sinistralité se base sur l'approche fréquence-sévérité, très utilisée en tarification. L'idée est de modéliser séparément le nombre de sinistre et le coût de sinistre. Notons X le risque associé à la garantie couverte. L'approche fréquence-sévérité définit X comme une somme aléatoire :

$$X = \begin{cases} \sum_{i=1}^{N} B_i, N > 0 \\ 0, N = 0 \end{cases}$$

avec N le nombre de sinistres et B_i les montants de sinistres individuels.
En supposant que B_i indépendant de N $\forall i$ et sous condition d'intégrabilité de B_i et N, la prime pure est donnée par $E[X] = E[N]E[B]$.
Dans cette sous-partie, on présente les principales distributions pour le nombre de sinistre et pour le montant de sinistre. Les notations suivantes seront utilisées. L'espérance mathématique représente la valeur moyenne d'une variable aléatoire. C'est une caractéristique de tendance centrale, c'est-à-dire permet de résumer un ensemble de données. De son côté, la variance est une mesure de dispersion. Elle s'interprète comme la variation moyenne autour de l'espérance. Mathématiquement, la variance est la moyenne des carrés des écarts à la moyenne. Une variance est toujours positive.

Notations	Significations
$F_X(x) = P(X \leqslant x)$	fonction de répartition
$E[X] = \int_R x dF_X(x)$	espérance
$Var[X] = E[X^2] - (E[X])^2$	variance
$f_X(x) = F'_X(x)$	densité
$p_X(x) = P(X = x)$	fonction de moment de probabilité
$M_X(t) = E[e^{tX}]$	fonction génératrice des moments
$P_X(t) = E[t^X]$	fonction génératrice des probabilités

Table 1 – Notations

■ Les lois de fréquence :

Dans ce qui suit, on présente les lois de fréquence les plus usuelles :

Distribution binomiale
Cette loi compte le nombre de succès parmi n expériences mesurées de façon indépendante. Elle correspond à la loi de la somme de n variables indépendantes identiquement distribuées de Bernouilli.
Lorsque N suit une loi binomiale de paramètres n,p, notée $N \sim \mathcal{B}(n, p)$ nous avons :
$P(N = k) = \binom{n}{k} p^k (1-p)^{n-k}$ avec $k \in 0, ..., n$.
$P_N(t) = (1 - p + pt)^m, E[N] = np$ et $Var[N] = np(1-p)$.
Dans ce cas, X suit une loi binomiale composée.

$$F_X(x) = (1-p)^n + \sum_{k=1}^{n} \binom{n}{k} p^k (1-p)^{n-k} F_B^{*k}(x), x \in \mathbb{N}$$

avec $F_B^{*k} = F_{B_1+B_2+...B_k}$ appelée convolution des B_i.
$E[X] = npE[B], Var[X] = npVar[B] + np(1-p)(E[B])^2$ et $M_X(t) = (1 - p + pM_B(t))^m$.

Approximations :
-Lorsque n est assez grand et p pas trop proche de 0 ou 1, la loi binomiale $\mathcal{B}(n,p)$ peut être approchée par la loi normale $\mathcal{N}(np, np(1-p))$.
-Pour les petites valeurs de p, une bonne approximation de la loi binomiale $\mathcal{B}(n,p)$ est la loi de poisson $\mathcal{P}(np)$.

Distribution de Poisson
La loi de Poisson de paramètre λ modélise le nombre de fois qu'un événement aléatoire, qui arrive en moyenne λ fois, se produit. Une variable aléatoire qui suit une loi de Poisson prend des valeurs entières 0,1,2.... La loi de probabilité est définie par :

$$p_X(k) = P(X = k) = e^{-\lambda}\frac{\lambda^k}{k!}$$

où λ un nombre réel strictement positif. $P_N(t) = exp(\lambda(t-1))$.
De plus, $E[N] = \lambda = Var[N]$.
La fonction de répartition de X est donnée par :

$$F_X(x) = e^{-\lambda} + \sum_{k\geq 1}\frac{\lambda^k}{k!}e^{-\lambda}F_B^{*k}(x), x \in \mathbb{N}$$

et les moments :
$E[X] = \lambda E[B], Var[X] = \lambda E[B^2], M_X(t) = exp(\lambda(M_B(t)-1))$.

Approximations :
Lorsque λ est assez grand, la loi de poisson de paramètre λ peut être approximée par la loi normale $\mathcal{N}(\lambda, \lambda)$.

Distribution binomiale négative
Considérons l'expérience consistant à faire des tirages indépendants donnant un succès avec une probabilité p et un échec avec une probabilité (1-p). On poursuit cette expérience jusqu'à obtenir n succès. La variable aléatoire représentant le nombre d'échecs obtenus avec la réalisation des n succès suit une loi binomiale négative de paramètres n et p. Nous avons les formules suivantes.

$$p_N(x) = \binom{n+x-1}{x}(1-p)^x p^n, x \in \mathbb{N}; P_N(t) = (\frac{p}{1-(1-p)t})^n.$$

La fonction de répartition de X s'écrit :

$$F_X(x) = p^n + \sum_{k\geq 0}\binom{n+k-1}{k}(1-p)^k p^n F_B^{*k}(k), x \in \mathbb{N}.$$

En outre, $E[X] = n\frac{1-p}{p}E[B], Var[X] = n\frac{1-p}{p}Var[B] + n\frac{1-p}{p^2}(E[B])^2$ et $M_X(t) = (\frac{p}{1-(1-p)M_B(t)})^n$.

Approximations :
Pour n assez grand, la loi binomiale négative BN(n,p) est approximée par la loi normale $\mathcal{N}(n(1-p)/p, n(1-p)/p^2)$.

■ Les lois de coût :

En assurance non vie, il est usuel d'utiliser des lois à supports dans $\mathbb{R}_+$ pour modéliser les montants de sinistres. Il en existe plusieurs, nous présentons ici les lois continues les plus classiques.

Distribution Gamma
Du fait de ses bonnes propriétés, la famille de lois gamma ou d'Euler est utilisée assurance pour modéliser les montants de sinistre, le temps écoulé entre deux sinistres... De manière générale, pour des distributions fortement asymétriques avec une décroissance rapide en queue de distribution, une loi gamma peut fournir une bonne modélisation.
B suit une loi gamma de paramètres p,θ si :

$$f_B(x) = \frac{\theta^p x^{p-1}}{\Gamma(p)} e^{-\theta x} \mathbb{1}_{x>0}$$

.
Dans ce cas, nous avons : $M_B(t) = (\frac{\theta}{\theta - t})^p$.
Du fait que la fonction génératrice des moments existe en zéro, tous les moments existent :

$$E[B^n] = \frac{\prod_{j=0}^{n-1}(p+j)}{\theta^n} \implies E[B] = \frac{p}{\theta}, Var[B] = \frac{p}{\theta^2}.$$

Dans ce cas X a pour fonction de répartition :

$$F_X(x) = p_N(0) + \sum_{k=1}^{\infty} p_N(k) F_{Gamma(kp,\theta)}(x), x > 0.$$

La densité de X s'obtient par dérivation de F :

$$f_X(x) = \sum_{k=1}^{\infty} p_N(k) f_{Gamma(kp,\theta)}(x), x > 0.$$

Cas particuliers de loi gamma :
-Pour p=1, on obtient la loi exponentielle de paramètre θ avec pour densité :

$$f_{Be}(x) = \beta e^{-\beta x} \mathbb{1}_{x>0}$$

-Lorsque p $\in \mathbb{N}_+$, on a la loi d'Erlang :

$$f_{BE}(x) = \frac{\beta^n x^{n-1}}{(n-1)!} e^{-\beta x} \mathbb{1}_{x>0}$$

Distribution log-normale

Souvent en assurance, les risques couverts conduisent à des sinistres assez élevés. Il convient donc d'utiliser des lois à queues épaisses pour modéliser ces risques. Une manière de procéder est de considérer la variable aléatoire e^Y au lieu de Y. En effet, la fonction exponentielle permet de rehausser les queues des distributions.

Si Y a pour densité f_Y, alors e^Y admet comme densité :

$$f_Y(y) = \frac{f_X(ln(y))}{y}$$

pour $y \in \mathbb{R}_+^*$ et $f_Y(y) = 0$ pour $y \in \mathbb{R}_-$.

Dans le cas où Y suit une loi normale, e^Y a pour distribution une loi log-normale $\mathcal{LN}$.

Pour $B \sim \mathcal{LN}(\mu,\, \sigma^2)$ on a :

$$f_B(x) = \frac{1}{x\sqrt{2\pi\sigma^2}} exp(-\frac{(ln(x)-\mu)^2}{2\sigma^2}), F_B(x) = \Phi(\frac{ln(x)-\mu}{\sigma}), x \in \mathbb{R}_+.$$

Concernant les moments,

$$E[B^n] = exp(n\mu + n^2\sigma^2/2) \implies E[B] = exp(\mu + \sigma^2/2), Var[B] = exp(2\mu + \sigma^2)(exp(\sigma^2) - 1).$$

Les moments sont donnés par :

$$E[Z] = -\frac{\alpha\nu}{2(\tau-1)}, \tau > 1.$$

$$Var[Z] = \frac{\alpha^2}{\tau^2(\tau-1)}, \tau > 3/2.$$

Distribution inverse gaussienne

Statistiquement, la loi inverse gaussienne (ou loi gaussienne inverse ou encore loi de Wald) s'interprète comme étant le temps en lequel le mouvement brownien avec une dérive positive atteint une valeur fixée.

Dans le cas où B est modélisé avec une inverse gaussienne $\mathcal{IG}(\mu, \lambda)$, la densité s'écrit :

$$f_B(x) = \sqrt{\frac{\lambda}{2\pi x^3}} exp(-\frac{\lambda}{2\mu^2 x}(x-\mu)^2), x > 0, \mu > 0, \lambda > 0.$$

La fonction de répartition s'écrit en fonction de celle de la loi normale centrée réduite.

$$F_B(x) = \Phi(\sqrt{\frac{\lambda}{x}}(x/\mu - 1)) + exp(\frac{2\lambda}{\mu})\Phi(\sqrt{-\frac{\lambda}{x}(x/\mu + 1)})$$

et les moments :

$E[B] = \mu, Var[B] = \frac{\mu^3}{\lambda}$.

La somme aléatoire X a pour fonction de répartition

$$F_X(x) = p_N(0) + \sum_{k=1}^{\infty} p_N(k) F_{\mathcal{IG}(k\mu, k^2\lambda)}(x)$$

et par dérivation de la fonction de répartition, on obtient la formule de la densité :

$$f_X(x) = \sum_{k=1}^{\infty} p_N(k) f_{\mathcal{IG}(k\mu, k^2\lambda)}(x).$$

1.5 Le traitement de l'information géographique

Comme dit dans l'introduction, le contexte de ce rapport s'inscrit dans la construction d'un micro-zonier sur le coût moyen de la garantie dégâts des eaux et un autre micro-zonier sur la fréquence des vols, en exploitant au mieux un grand nombre de variables externes. Face à l'accès, de plus en plus facile à l'open data, une question importante se pose : comment intégrer les bases de données externes aux données internes de l'entreprise ? La solution à ce problème n'est pas évidente. Souvent, il n'existe pas d'identifiant commun (clé de jointure commune) entre une base open data et la base de données de l'entreprise. Et même lorsque cette variable commune existe, elle correspond par exemple au code postal, au code INSEE, au code département... Ces mailles assez larges ne sont pas adaptées lorsqu'on désire ajouter à nos données internes des informations à un niveau très graduel (c'est-à-dire à une maille très fine). Une manière de résoudre ce problème d'ajout de données est de recourir aux coordonnées géographiques (latitudes, longitudes). En effet, l'essentiel des data externes contiennent ces deux variables. Lorsque la base de données de l'entreprise contient les coordonnées géographiques (des contrats par exemple), il est possible de faire une jointure spatiale pour récupérer des données au niveau adresse ou dans des polygones très petits de manière générale. Cela explique le fait que les entreprises ont de plus en plus recours au géocodage et également l'apparition de star-ups qui développent des techniques de géocodage. Une bonne jointure spatiale nécessite des compétences d'analyse géospatiale (conversion au même système de coordonnées géographiques, création de shapefiles, découpages de zones...). Tous ces traitements spatiaux sont regroupés dans une discipline nommée "Systèmes d'Informations Géographiques (SIG)".

1.5.1 Introduction aux SIG

Un système d'information géographique (SIG) est défini comme un ensemble de techniques permettant, à partir de sources diverses de données, de rassembler, de traiter, d'analyser et de représenter des informations localisées géographiquement. Un SIG utilise principalement l'Informatique, la Statistique et la Géographie pour analyser et représenter tous les éléments qui existent sur la terre ainsi que tous les événements qui y ont lieu. Les premiers travaux de SIG remontent en 1854 avec le docteur John Snow qui a utilisé de la cartographie pour identifier la source d'une épidémie de choléra à Londres. L'essor de l'informatique a permis le développement des SIG car ces techniques sont essentiellement basées sur l'utilisation des logiciels et matériels informatiques. La prolifération des technologies portables a permis de faciliter les géolocalisations et d'en accroître la précision. Les SIG font l'objet d'une grande variété d'applications. La plupart des enjeux auxquels le monde fait face aujourd'hui (environnement, politiques économiques, démographie...) sont étroitement liés à la géographie, et donc aux SIG. Par conséquent, les SIG sont utilisés par plusieurs structures (entreprises, administrations, services de renseignement, collectivités territoriales, secteur public...). De manière générale, les utilisations les plus courantes sont :
-Planification territoriale :

Dans le domaine de l'aménagement du territoire, les SIG permettent le calcul de population présente par zones, la ventilation d'équipements... En France par exemple, les études spatiales réalisées durant la pandémie du covid 19 par l'INSEE ont permis la ventilation efficiente des lits d'hôpitaux et la répartition des matériels médicaux. Dans les pays en développement, l'utilisation des SIG permet aux organismes internationaux de donner une alerte précoce sur les risques de pénurie alimentaire et à mettre en oeuvre des plans d'action pour l'environnement. D'autres utilisations des SIG en planification territoriale existent : l'optimisation du choix d'emplacement de nouveaux commerces, la définition des zones de vente, l'analyse foncière (étude d'impact d'une construction...), l'analyse archéologique, la prospection minière, la gestion des parcelles d'exploitation forestière, l'amélioration du recouvrement des taxes, l'allocation géographique des nouveaux investissements et l'identification des sites qui conviennent le mieux à l'établissement de nouvelles installations, la gestion de réseaux (surtout en télécommunication), le suivi en temps réel de véhicules, de flotte..., l'intégration tout type d'informations pour enrichir les bases de données des entreprises...

-Visualisation spatiale :

En matière de divulgation de données géographiques, les cartes sont souvent beaucoup plus parlantes que les tableaux. Les SIG sont utilisés pour visualiser, sur des cartes, des données géographiques : zones d'activités, réseaux routiers, statistiques sur des clients... La représentation peut aller d'une simple solution de cartes à de la Business Intelligence géospatiale. Il esxiste plusieurs logiciels de SIG prévus à cet effet : Arc GIS, QGIS, MapInfo, Postgis... Les logiciels de SIG offrent des options intéressantes lors de la visualisation : représentation de plusieurs couches de cartes, gestion de l'ordre d'empilement des couches et des paramètres de visibilité, association de symboles aux entités (symbole ligne pour les routes, symbole point pour les domiciles...), choix d'une échelle, ajout de légendes... On peut aussi mettre en évidence une couche ciblée et cacher les autres, rendre plus ou moins transparentes les cartes, modifier les couleurs... Pour une étude plus avancée donnant une plus grande marge de manoeuvre sur les données, on peut recourir aux logiciels tels que Python et R qui offrent la possibilité de programmation et de traitements spatiaux avant la représentation sur des cartes. Dans ce mémoire, nous avons beaucoup utilisé les packages de traitement géospatial sous Python qui offrent plusieurs possibilités (telles que les requêtes et les analyses statistiques) sur nos données géographiques.

-Géoréférencement :

Le géoréférencement consiste à attribuer des références géographiques à un emplacement de la terre (nouveau quartier, nouveau cours d'eau...). Autrement dit, il s'agit de caler un emplacement sur la carte de la terre en lui attribuant des coordonnées géographiques. Les techniques de géocodage interviennent à ce niveau. Le géocodage consiste à transformer la description d'un emplacement (exemple une adresse ou un nom de lieu) en des coordonnées géographiques. La qualité du géoréférencement dépendra donc de la précision de la méthode de géocodage.

Une fois un objet géoréférencé, une grande quantité d'informations peuvent lui être associées via ses coordonnées géographiques par jointure spatiale. Le géoréférencement est un aspect fondamental dans l'analyse des données spatiales. La fiabilité de toutes les opérations

géospatiales dépend de la précision du géoréférencement.

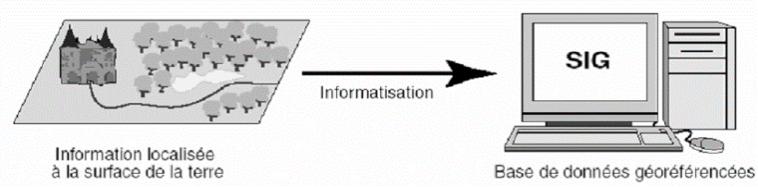


FIGURE 3 – Géoréférencement

-Geomarketing :
Dans un milieu de concurrence de plus en plus rude, les départements marketing cherchent de nouvelles pistes à explorer pour se démarquer. L'intégration de l'aspect géographique s'inscrit dans ce sens. Le géomarketing est une discipline qui s'appuie sur les SIG pour apporter une dimension géographique aux stratégies de marketing. Il aide à prendre de meilleures décisions en considérant l'aspect géographique. L'objectif étant de mettre en place des outils permettant la concrétisation spatiale des activités économiques. Le géomarketing apporte ainsi des solutions aux problématiques géographiques des comportements et mobilités des personnes et objets. Le géomarketing est utilisé par les entreprises pour visualiser les potentialités commerciales à exploiter sur les zones géographiques. Il est devenu une composante indispensable du suivi des clients et des opérations commerciales des biens et services. Par exemple, une segmentation de la clientèle basée sur l'emplacement géographique permet d'effectuer un marketing ciblé. Les services marketing des compagnies d'assurance par exemple, font de plus en plus recours à ces solutions géospatiales pour optimiser le chiffre d'affaires en localisant les zones à forte concentration de clients insatisfaits ou les zones à fort niveau de sinistre. En apportant une information supplémentaire géographique, le geomarketing permet de procéder à une segmentation plus fine des clients.

1.5.2 Enjeux de l'utilisation de l'information géographique en assurance

L'utilisation des techniques d'analyse géospatiale représente un enjeu stratégique. Malgré son importance, ce domaine reste encore un peu méconnu. En plus des utilisations classiques qu'en font la plupart des assureurs (visualisation sur des cartes avec des logiciels comme map info...), il est possible d'utiliser des techniques d'analyse géospatiale pointues pour mener des analyses plus fines de nos données. L'utilité justifiant le recours à ce type d'analyse en assurance est très vaste. A partir de l'étude de la position géographique des clients, les services marketing des compagnies d'assurance peuvent mettre en place des campagnes publicitaires ciblées. Autre exemple, un assureur peut veiller à répartir sa couverture de manière uniforme sur le territoire. En outre, l'analyse spatiale permet de localiser des phénomènes ou bien des groupes d'individus spécifiques (les clients insatisfaits par exemple). Les SIG peuvent aussi être utilisés pour identifier les zones à niveau de sinistres élevés (cambriolages, incendies, catastrophes naturelles...). Ils aident aussi à améliorer la diffusion des données via la visualisation sur des cartes qui sont souvent beaucoup plus parlantes que

les tableaux. Dans le cadre de ce mémoire, nous utiliserons les techniques de traitement d'informations géographiques à des fins de jointure de bases de données, de représentation de cartes, de découpages de zones et pour d'autres manipulations présentées dans la partie «Quelques exemples de manipulation de shapefiles». Ceci dans le but de fixer de manière optimale la prime.

1.5.3 Les systèmes de coordonnées géographiques

Un système de coordonnées géographiques est une représentation des emplacements de la Terre sur un datum. Un datum est l'expression mathématique de la surface de la terre (Ellipsoïde) qui vise à minimiser l'erreur de géolocalisation par rapport à la surface réelle de la terre (Géoïde). Au cours des siècles, diverses combinaisons géométriques ont été élaborées pour représenter la surface incurvée de la terre sur des coupures cartographiques. Ces combinaisons sont connues sous le nom de projections cartographiques : système de Mercator, système de projection transverse, la projection Universelle Transverse de Mercator (UTM)... Le système le plus fréquemment utilisé est le Système géodésique mondial de 1984 (WGS784). Il est utilisé pour mesurer des emplacements au niveau international. Les mesures GPS (Global Positionning System) par exemple sont établies à partir du WGS84. Il existe des systèmes de coordonnées projetés qui permettent la conversion mathématique des latitudes et longitudes pour transformer la surface ellipsoïde (tri- dimensionnelle) de la Terre en surface bidimensionnelle. Dans notre étude, la plupart des données mises à notre disposition sont codées dans le système projeté lambert 93. Bien que nous avons eu à manipuler des shapefiles codés dans d'autres systèmes et que nous avons eu à convertir au système lambert 93. Le lambert 93 est une projection cartographique développée par le mathématicien Johann Heinrich Lambert. C'est le système de projection officiel utilisé pour représenter la France métropolitaine.

1.5.4 La notion de shapefile

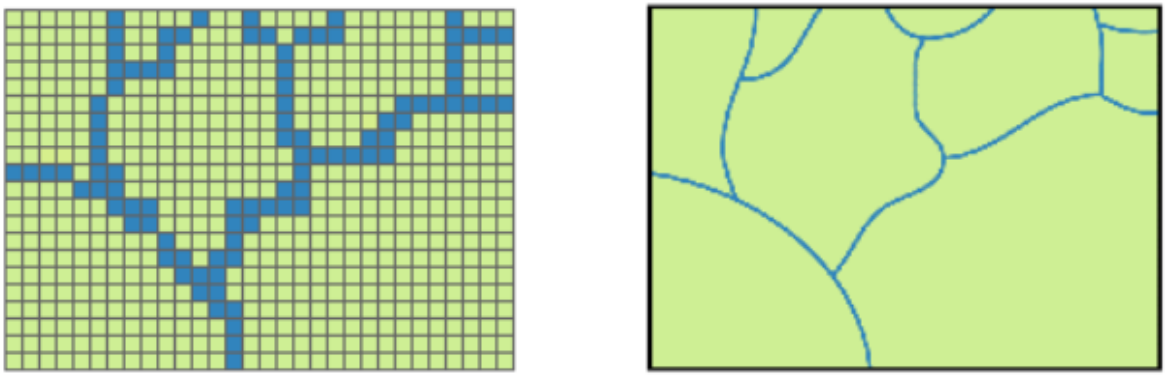
Un shapefile est un format de données vectorielles géospatiales. Les données ou variables que contient un shapefile sont reliées à des géométries spécifiées. Ces géométries qui définissent l'endroit de la terre sur lequel les données sont rattachées peuvent être des points, des lignes ou des polygones. Un point peut par exemple représenter un contrat (le domicile d'un assuré). Comme exemple de lignes, nous avons les routes (en assurance auto par exemple). Les polygones sont les délimitations des régions, des départements, des communes ou d'autres découpages beaucoup plus petits. Un exemple simple de shapefile peut être une base de données contenant le nombre de sinistres par communes avec les coordonnées géographiques et les géométries de ces communes. L'extension de ce type de fichier est « shp ». Il peut être représenté sur une carte.

1.5.5 Les types de données spatiales

Il existe deux types de données spatiales : le type raster et le type vecteur.

-raster :
Un raster est un ensemble de pixels qui constituent une image. Chaque pixel a ses propres attributs ou valeurs. Par définition, un raster est donc une image (photographie, images satellitaire, scan...).
-vecteur :
Le format vecteur utilise des points, des lignes et des polygones pour représenter les entités géographiques. On parle ainsi de dessins vectoriels ou d'images vectorielles. Une donnée de type vecteur a deux composantes : une table attributaire qui regroupe les informations associées et une composante graphique qui contient la géométrie.

Raster Vecteur



1.5.6 Quelques exemples de manipulation de shapefiles

Lorsqu'on a en face de nous deux jeux de données spatiales (deux shapefiles avec leurs géométries), on se trouve souvent dans le besoin de créer de nouvelles formes en fonction des endroits où ces jeux de données se chevauchent (ou ne se chevauchent pas). Ces manipulations peuvent être de plusieurs types : intersection, union, différence symétrique, différence.

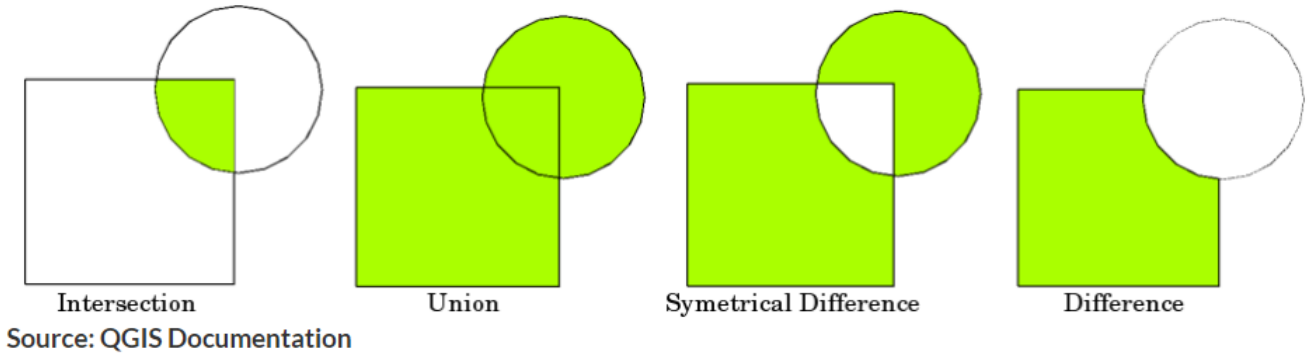


FIGURE 4 – Exemples de manipulation de données spatiales

Dans ce mémoire nous avons utilisé l'intersection. Plus explicitement, nous avons fait l'intersection entre le shapefile de la France avec d'autres jeux de données qui couvrent la France et

d'autres zones. Ce qui permet de conserver uniquement les données concernant le territoire français. Ces opérations peuvent être réalisées avec des logiciels de Système d'Information Géograpique tels que QGIS mais aussi avec des packages de python (shapely, rtree, geopandas...). D'autres opérations sur les données spatiales sont utilisées dans ce rapport : agrégation de polygones en une seule zone. Par exemple, nous avons obtenu la carte du contour de la France en agrégeant la carte des communes de la France en un seul polygone. Ce qui a motivé une telle pratique est que la carte des communes s'est révélée être celle qui couvre mieux la carte réelle de la France (surtout au niveau des frontières). Le fond de carte réel de la France qui nous a permis d'opérer de telles vérifications est disponible sous Qgis mais non utilisable dans l'étude car étant au format image Google maps. Une telle opération peut être réalisée avec la fonction dissolve du package geopandas de Python.
En plus de cela, nous avons beaucoup fait de jointures de base de données. Deux manières de combiner des jeux de données ont été utilisées dans cette étude : les jointures d'attributs et les jointures spatiales. La première catégorie correspond à la manière classique de fusionner des bases de données. Il s'agit de combiner des jeux de données sur la base d'une variable commune (exemple numéro de contrat). Par contre, pour une jointure spatiale, les observations des deux jeux de données sont combinées en fonction de leur relation spatiale les unes avec les autres. Supposons que l'on dispose d'une base de données contenant les assurés avec les coordonnées géographiques (latitudes et longitudes) de leurs domiciles, et d'autre part une base de données contenant le prix moyen au m2 à une maille très fine (par iris pour illustrer) des terres. Alors la jointure spatiale entre ces deux bases fournira un jeu de données unique contenant les assurés et le prix moyen au m2 de leurs habitations. Ce genre d'information additionnelle peut être utilisé pour affiner la tarification des contrats d'assurance habitation par exemple. Ce type de jointure peut être réalisé avec Postgis ou Python.

1.6 Méthodologie d'élaboration de nos micro-zoniers

L'intégration des données géographiques dans les modèles de tarification en assurance habitation se fait via la mise en place d'un zonier. Un zonier est un découpage qui permet de segmenter les zones géographiques en fonction de leur exposition aux risques. L'effet géographique est contenu dans les résidus d'un modèle de prime pure hors variables géographiques. Ensuite, cet effet est lissé sur l'ensemble du territoire.

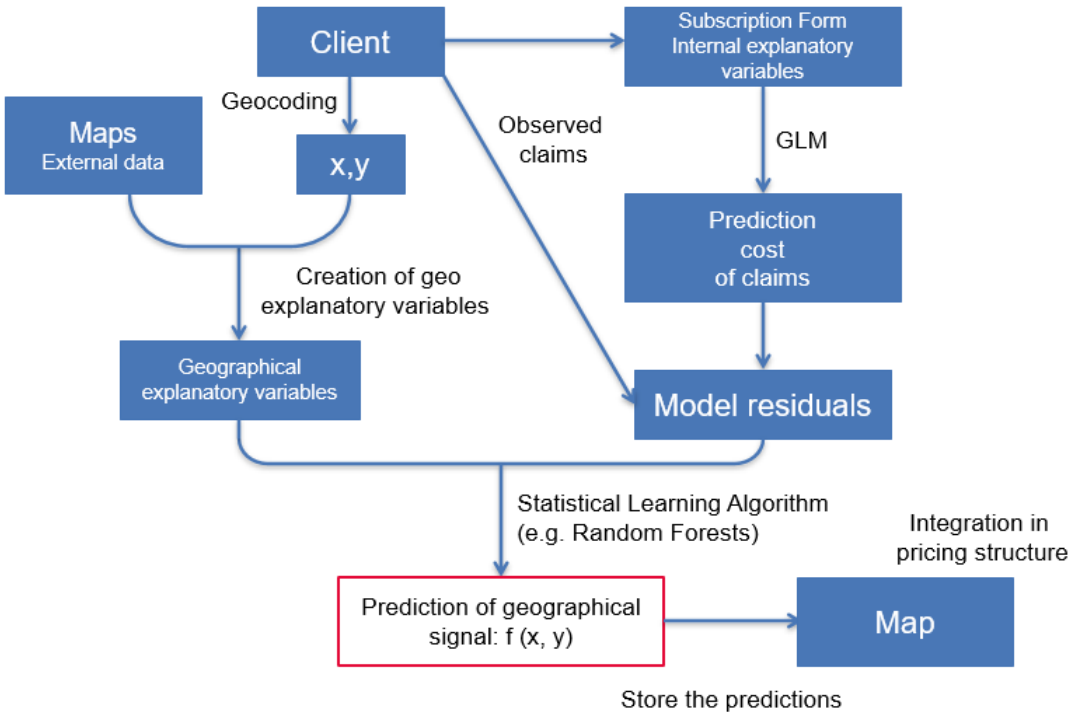


FIGURE 5 – Méthode classique de création d'un zonier

Le plus souvent, les zoniers sont créés en résumant le risque géographique par maille INSEE ou code postal. L'inconvénient est que ces mailles sont très grandes et ne permettent donc pas de capter avec une grande précision le signal géographique.

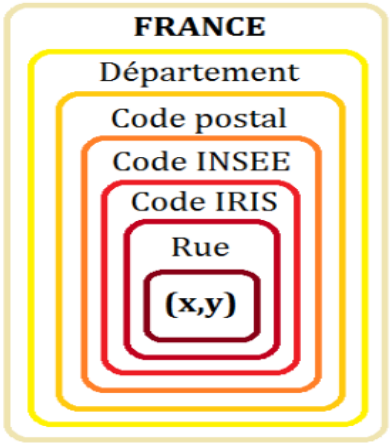


FIGURE 6 – Les niveaux de découpage en France

Dans cette étude, nous allons plutôt associer un risque géographique à chaque point (x,y) représentant un contrat. Pour étendre le signal géographique sur l'ensemble du territoire français, nous utiliserons l'algorithme de Voronoi. Ce dernier crée des découpages beaucoup plus petits que les communes en regroupant pour chaque point (x,y) de la base de données l'ensemble des points de l'espace qui sont plus proches de ce point que des autres points. L'avantage d'une telle méthode est d'avoir une maille très fine. On aura autant de polygones que de contrats. Pour prédire le signal géographique, nous utiliserons une méthode

algorithmique d'apprentissage statistique, plus précisément le modèle catboost. Les étapes de la réalisation d'un micro-zonier se définissent comme suit :
-découpage de la base de données en apprentissage, validation et test : 40% train, 40% test et 20% validation ;
-modélisation du coût moyen dégât des eaux et de la fréquence vol avec un modèle linéaire généralisé (glm) et les variables classiques de tarification, ceci sur la base d'apprentissage ;
-récupération des résidus de ces deux premiers modèles ;
-modélisation de ces résidus avec les variables externes géographiques à l'aide d'un catboost ;
-les prédiction issues de ces modèles représentent le signal géographique ;
-création de la maille (les polygones de Voronoi) avec la base d'apprentissage. Chaque polygone contient un seul contrat de la base appelé germe. L'ensemble des points géographiques d'un polygone donné ont la même valeur de signal géographique, celle du contrat autour duquel le polygone est formé. Ce qui permet d'avoir la valeur du signal géographique pour tous les points de France ;
-fit du zonier sur la base de validation pour évaluer les performances du modèle catboost du signal géographique. L'intérêt d'un tel découpage de la base de données permet d'éviter que le modèle de machine learning sur-apprenne.
-création des clusters : le signal géographique sera découpé en plusieurs classes, la représentation de ces différentes classes du signal géographique sur une carte représente un micro-zonier,
-fit du zonier sur la base test : chaque contrat de la base de test sera projeté sur le micro-zonier pour récupérer sa valeur de cluster, ceci grâce à une jointure spatiale entre le micro-zonier et la base de test.
-intégration des micro-zoniers dans les modèles de tarification dégâts des eaux et vols : un modèle de coût moyen avec les variables classiques de tarification sans le zonier et un autre modèle de coût moyen avec les variables classiques de tarification et le zonier seront comparés. Ce qui permettra de juger l'apport du zonier dans la tarification. On fait de même avec la fréquence vol.

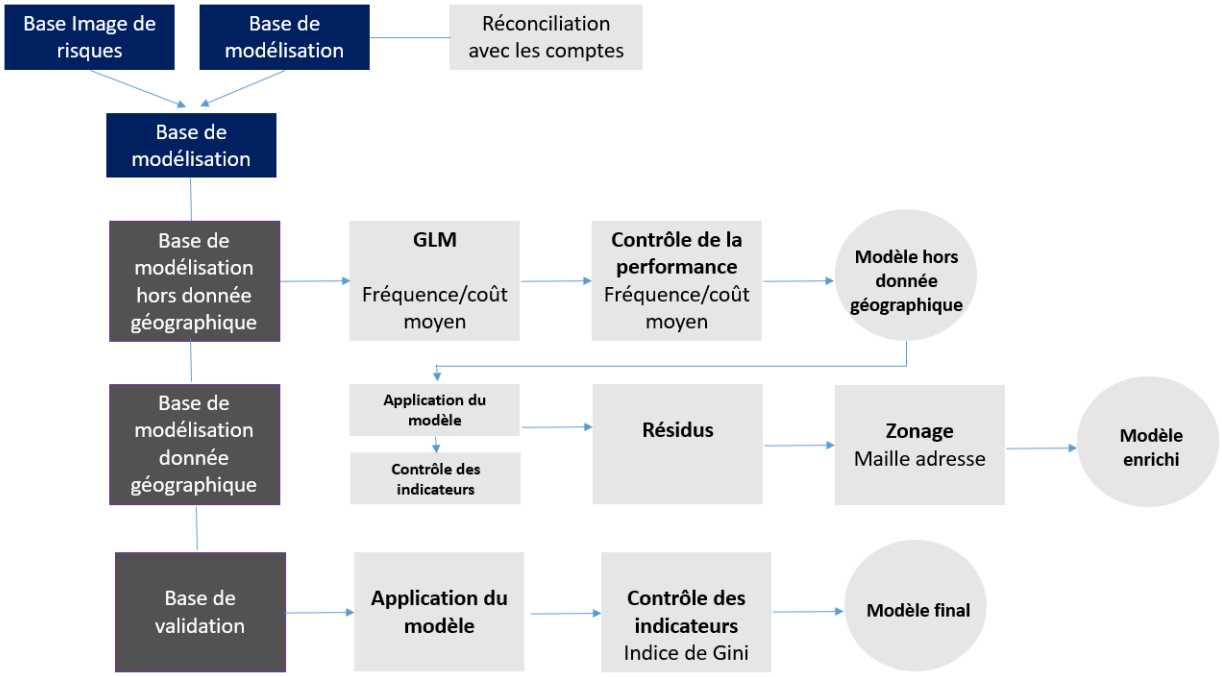


FIGURE 7 – Notre méthodologie de création de zonier

1.7 Présentation de la méthode de Voronoï

L'algorithme de partitionnement de Voronoï, formalisé par le mathématicien russe Georgi Voronoï (1868 – 1908), est un découpage du plan en cellules à partir d'un ensemble discret de points appelés « germes ». Concrètement, cette méthode permet le partitionnement d'un plan contenant n points en n polygones de telle sorte que chaque polygone contienne exactement un point générateur et que chaque point d'un polygone donné soit plus proche de son point générateur que de tout autre point du plan. Cette méthode est aussi appelée diagramme de Voronoï ou parfois pavage de Dirichlet. Les cellules/polygones obtenus sont appelés polygones de Voronoï. Ce procédé permet donc de découper le territoire français en des polygones beaucoup plus petits que les découpages administratifs habituels (régions, départements, communes, iris...).

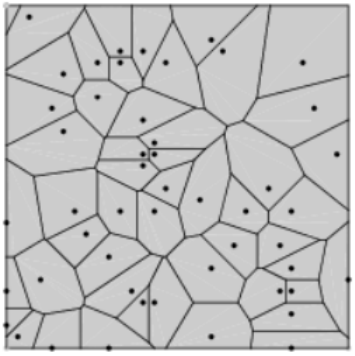


FIGURE 8 – Exemple de découpage d'un plan par Voronoï

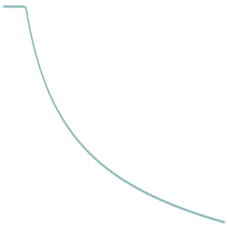
■ Formalisation mathématique :

Problème :
Soit S un ensemble de n points du plan E, n ≥ 3 .
Le but est de décomposer l'espace en régions autour de chaque point p de S, telles que tous les points dans la région contenant p soient plus près de p que de n'importe quel autre point de S. Ce qui conduit donc à s'intéresser aux médiatrices des segments formés par les points de S.
Création du diagramme :
La médiatrice des points p et q sépare le plan en deux demi-plans. Si on note D(p,q) le demi-plan contenant p alors on a :
$D(p,q) = \{M \in E : d(p,M) < d(q,M)\}$, avec d la distance euclidienne.
La cellule ou polygone de Voronoï engendrée par le point p est : $R(p) = \bigcap_{q \in S} D(p,q)$.
On a alors $R(p) = \{M \in E : \forall q \in S, d(M,p) < d(M,q)\}$.
Le diagramme de Voronoï est : $D = \bigcup_{p,q \in S} \overline{R(p)} \cap \overline{R(q)}$.
La frontière commune à deux polygones de Voronoï est appellée arête et les extrémités des arêtes sont les sommets de Voronoï. Le diagramme de Voronoï définit donc bien une partition du plan. Chaque zone est définie comme l'ensemble des points les plus proches d'un point donné.

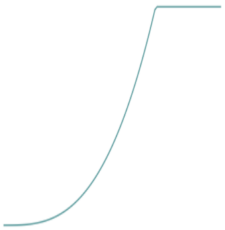
1.8 Présentation de la méthode de lissage

Mathématiquement, le lissage spatial est une fonction d'intensité d'un phénomène. De manière générale, le lissage spatial est utilisé pour estimer des valeurs en différents points de l'espace à partir de valeurs connues en un nombre limité de points. Le but étant de révéler des structures spatiales sous-jacentes et de régionaliser l'information. La valeur du phénomène géographique est calculée pour chaque point de l'espace en tenant en compte des valeurs enregistrées par les autres points de l'espace. Il s'agit donc de représenter non pas la valeur observée en un point, mais une moyenne pondérée des valeurs observées au voisinage de ce point. Le système de pondération de la méthode de lissage utilisée dans ce mémoire est tel que les points très proches se voient attribuer des coefficients élevés ; les points éloignés ayant des poids plus faibles. L'idée est que deux zones très proches n'aient pas des niveaux très différents. En tarification, le lissage spatial a une importance supplémentaire particulière. Il permet d'assurer une certaine équité. Par exemple, le lissage permettra que deux habitations ayant des caractéristiques similaires et se situant dans des zones géographiques très proches, aient des montants de primes pas très différents. Il existe plusieurs méthodes de lissage (kernel, krigeage...). Nous avons utilisé une méthode de lissage développée par Axa. L'avantage de cette méthode est de prendre en compte la similarité des zones en termes de niveau d'exposition, en plus de la proximité géographiques. Notre méthode de lissage utilise trois fonctions : une fonction de similarité en distance, fonction de similarité en exposition et une fonction de crédibilité.
- une fonction de similarité en distance : deux zones proches doivent avoir des valeurs de signal géographiques similaires. Un rayon de voisinage peut être défini. Un grand rayon

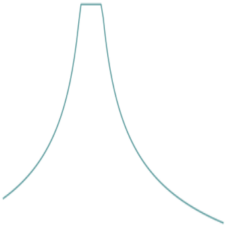
conduit à un niveau de lissage très élevé, avec un biais important. Un rayon petit fournit un niveau de lissage faible, synonyme d'une grande variance.
- fonction de similarité en exposition : deux zones similaires en termes d'exposition doivent avoir des valeurs de signal géographiques similaires.
- fonction de crédibilité : pour affecter des coefficients aux polygones, nous utilisons l'approche de crédibilité. Plus une zone a une exposition élevée, plus la crédibilité qui lui est associée est importante.



Exemple de fonction de similarité en distance



Exemple de fonction de crédibilité



Exemple de fonction de similarité en exposition

2 Présentation des données, traitements préliminaires et statistiques descriptives

2.1 Présentation des données

Cette section présente les données à notre disposition et les premiers traitements effectués. Il s'agira de vérifier l'intégrité des données utilisées. Comme dit plus haut notre périmètre d'étude concerne la garantie dégâts des eaux pour les appartements. Nous construirons notre zonier sur ce périmètre.

- Données internes :

La base de données interne utilisée dans cette étude contient des contrats de la garantie dégâts des eaux, ceci pour les appartements. Nous disposons de plusieurs variables tarifaires pour chaque contrat. Les variables tarifaires retenues pour le premier glm sur le coût moyen sans variable géographique sont :
- resid_type : le code (type) de résidence avec comme modalités P = principale et S= secondaire ;
- resid_nb_pieces : le nombre de pièces de l'appartement ;
- resi_qualite : la qualité occupant (P = propritaire et L=locataire) ;
- annee_vision : l' année de vision ;
- mnt_obj_valeur : le montant d'objets de valeur déclaré ;
- res_statut : le statut du contractant : société ou particulier.
Pour le glm fréquence vol, les variables explicatives sont les suivantes :
- CLI_age_cl : l'âge du client contractant ;

- resid_type : le code (type) de résidence avec comme modalités P = principale et S= secondaire ;
- resid_nb_pieces : le nombre de pièces de l'appartement ;
- resi_qualite : la qualité occupant (P = propritaire et L=locataire) ;
- POL_DISTRI : le distributeur de la police d'assurance (agents, couriers...) ;
- mnt_obj_valeur : le montant d'objets de valeur déclaré.

■ Données externes :

Les données externes mises à notre disposition sont de sources multiples et à différentes mailles. L'essentiel des données disponibles à l'échelle des découpages administratifs (communes, iris...) sont issues de l'INSEE et des sites de données publiques open data. Les données disponibles à la maille adresse proviennent de la Poste, d'Annuaire France Télécom, de l'Institut National de l'information Géographique, des directions régionales de l'environnement et d'autres entreprises et start-ups fournisseurs de données géographiques. Les bases de données externes peuvent être regroupées comme suit :
-les données socio-économiques : le nombre de ménages, le revenu moyen par ménage, le score d'attraction commerciale autour de l'adresse, le score de fréquentation touristique, le taux d'individus de 0 à 14 ans, la proportion de chômeurs dans la population majeure ... ;
-les données immobilières disponibles au niveau adresse : la surface au sol du bâtiment, l'altitude maximum du bâtiment, la hauteur du bâtiment, l'année de construction du bâtiment, le type de construction (collectif/logement étudiant/individuel), le prix moyen du m^2 dans l'iris ... ;
-les données environnementales : le nombre d'inondations enregistré, la différence de température moyenne de la zone, l'aléa tornade, l'aléa foudre, le nombre d'heures d'ensoleillement moyen, le niveau de précipitation moyen par jour en mm, les températures maximales et minimales de la zone... Des variables supplémentaires ont été crées, telles que les distances par rapport aux plus proches voisins.
Au total nous disposons de 238 variables externes.

2.2 Statistiques descriptives

Il est minutieux de d ebuter avec une analyse descriptive afin de mieux comprendre les caractéeristiques de nos donnéees. Il s'agira dans un premier temps de décrire la distribution de la variable coût moyen dégâts des eux et d'observer graphiquement sa liaison avec les variables explicatives internes. Ensuite nous ferons une analyse descriptive similaire avec la fréquence vol. Cette phase est fondamentale car permet de bien comprendre nos données avant d'appliquer les modèles.Le but étant de résumer au mieux nos données et de faire ressortir les tendances primaires. Cette analyse sera par la suite affinée par des modèles sophistiqués

■ Statistiques descriptives relatives au coût moyen dégâts des eaux :

La figure 10 représente l'histogramme du coût moyen pour la garantie dégât des eaux appartements. En moyenne, le coût d'un sinistre s'établit à 2484 euros. Une bonne partie

des sinistres ont un coût qui se situe entre 1000 euros et 3000 euros environ. Les sinistres supérieurs à 5000 euros sont rares. Les coûts de sinistres nuls ont été enlevés de la base, pour les besoins de la modélisation.

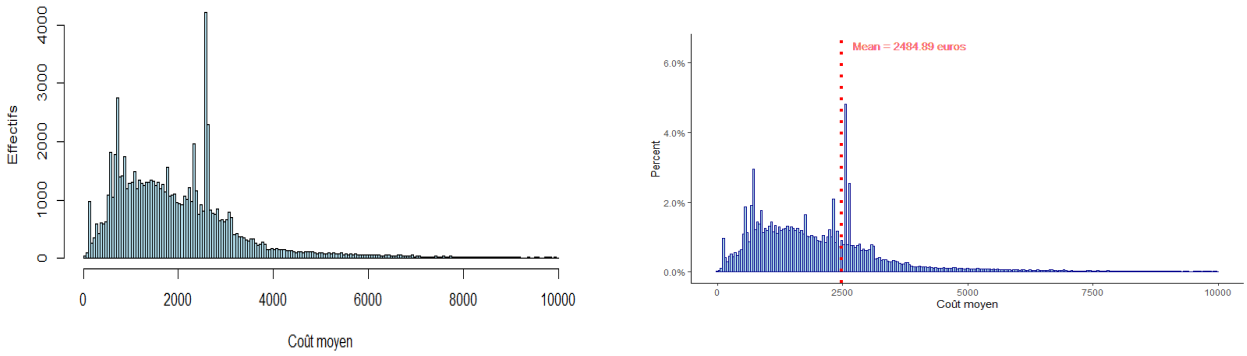


FIGURE 10 – Histogramme du coût moyen

Ensuite, nous avons étudié la distribution de notre variable d'intérêt (le coût moyen de sinistres) en fonction des variables tarifaires. Concernant la qualité de l'occupant de l'appartement (propriétaire ou locataire), ce sont les propriétaires qui enregistrent en moyenne les coûts de sinstres les plus élevés (voir figure 11). Le coût de sinistre moyen pour un propriétaire est de 3100 euros, contre 1965 euros pour un locataire. Les propriétaires ont donc en moyenne des coûts de sinistres plus élevés.

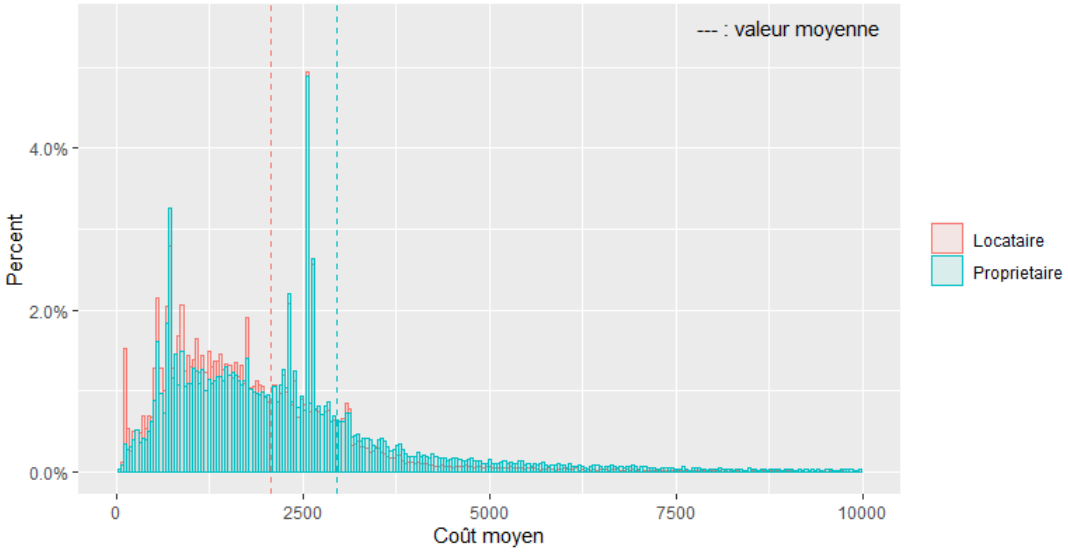


FIGURE 11 – Histogramme du coût moyen de sinistre par type de résident

Cette remarque est confirmée par la figure 12 qui montre une forte prépondérance de la sous population des propriétaires au niveau des tranches de coût de sinstres élevés. Ce graphique représente la proportion des types d'occupants suivants différents intervalles du

coût moyen. Ces intervalles sont construits avec les quantiles de la variable coût de sinistres. Il apparait clairement que la proportion de propiétaires croît avec le montant de sinistres. Plus l'intervalle de sinistre est élevé, plus la proportion de propriétaires est grande.

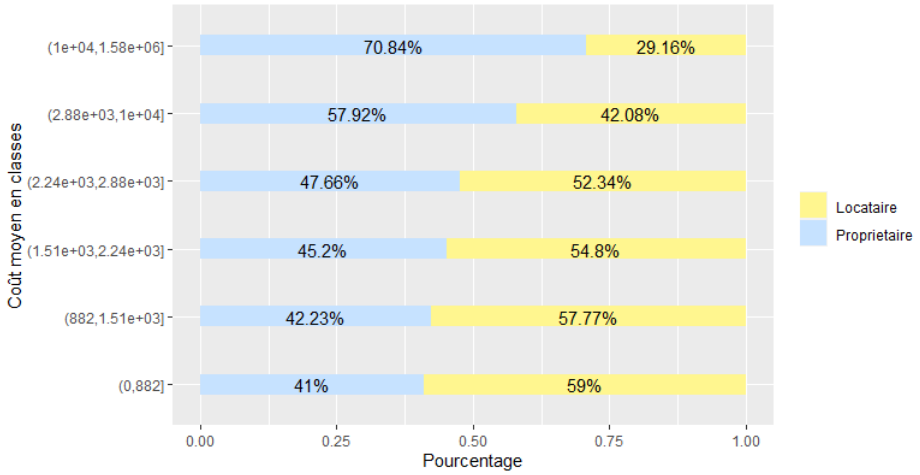


FIGURE 12 – Coût moyen de sinistres par type de résident

Au niveau de la figure 13, nous avons représenté l'histogramme de l'évolution du coût moyen de sinistres suivant le type de résidence. On remarque que les résidences secondaires ont tendance à enregistrer des coûts de sinistres élevés. Les résidences principales font plus l'objet de sinistres relativement plus faibles. En effet 20,4% des résidences principales ont un coût moyen de sinistre inférieur à 498 euros. 26,28% des résidences secondaires ont un coût moyen de sinistres qui se situe entre 1630 euros et 10000 euros.

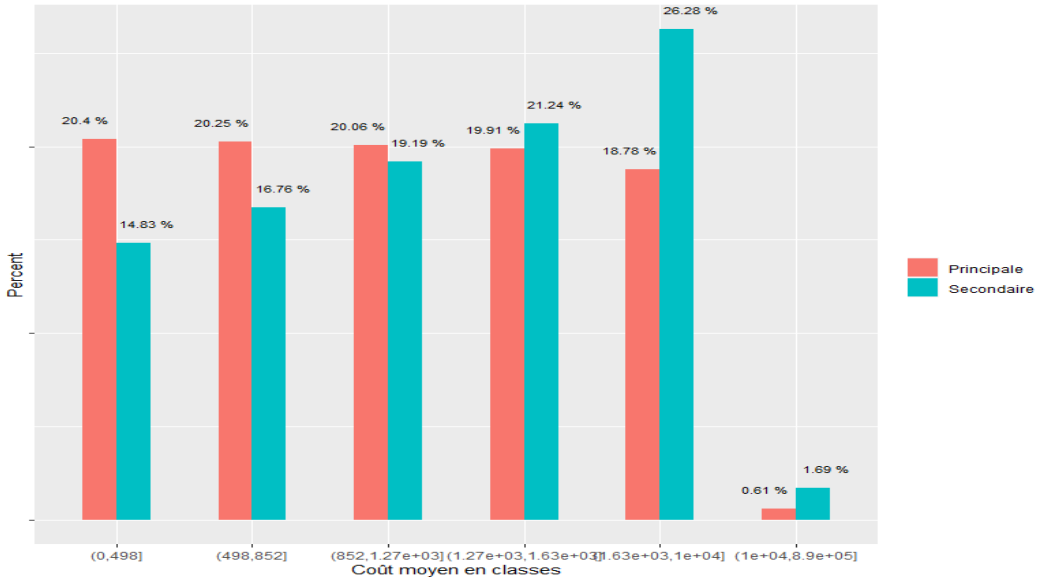


FIGURE 13 – Coût moyen de sinistre par type de résidence

La figure 14 représente la dispersion relative de la variable statut du contractant suivant le coût de sinistres. Il en ressort que le pourcentage de sinistres enregistrés par les particuliers

décroit avec le coût de sinistres. Plus précisément, 93,2% des sinistres de moins de 498 euros sont subis par les particuliers, au moment où ce pourcentage est plus petit pour les sinistres de plus de 10000 euros, soit précisément 85,24%. Pour les sociétés, on observe une tendance inverse. Le pourcentage de clients ayant le statut de société est plus élevé au niveau des sinistres graves et est plus petit pour les sinistres à coûts faibles.

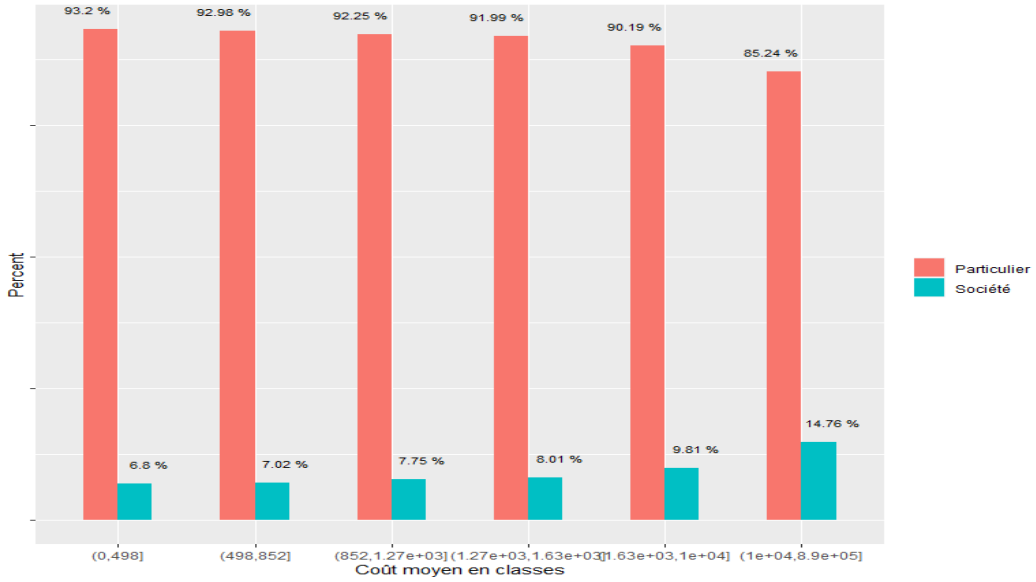


FIGURE 14 – Coût moyen de sinistre par statut du résident

Au niveau de la figure 15, nous avons représenté les boxplots par année de vision. Suivant les années, on remarque de légers contrastes. Les dispersions apparaissent assez stables pour les trois années. Le montant moyen d'un sinistre apparait plus élevé pour l'année 2018.

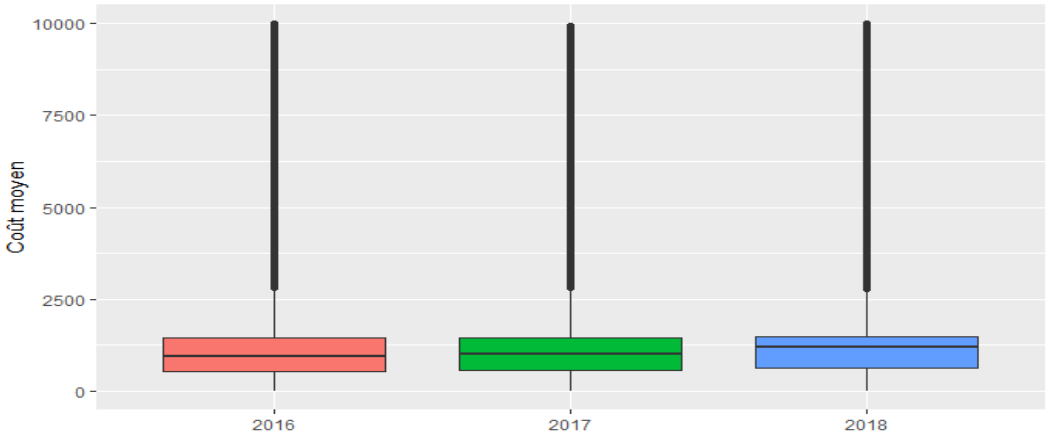


FIGURE 15 – Coût moyen de sinistre par année de survenance

De manière générale, le coût moyen de sinistres a tendance à augmenter avec le nombre de pièces de l'appartement. Les appartements comportant 7 pièces ont en moyenne un coût de sinistre plus élevé, tournant autour de 2720 euros. Ils sont suivis des appartements à 6 pièces et à 5 pièces avec respectivement 1867 euros et 1584 euros de coûts moyens de sinistre

dégât des eaux. Les appartements à petits nombres de pièces présentent, en moyenne, des coûts moyens de sinistres les plus faibles.

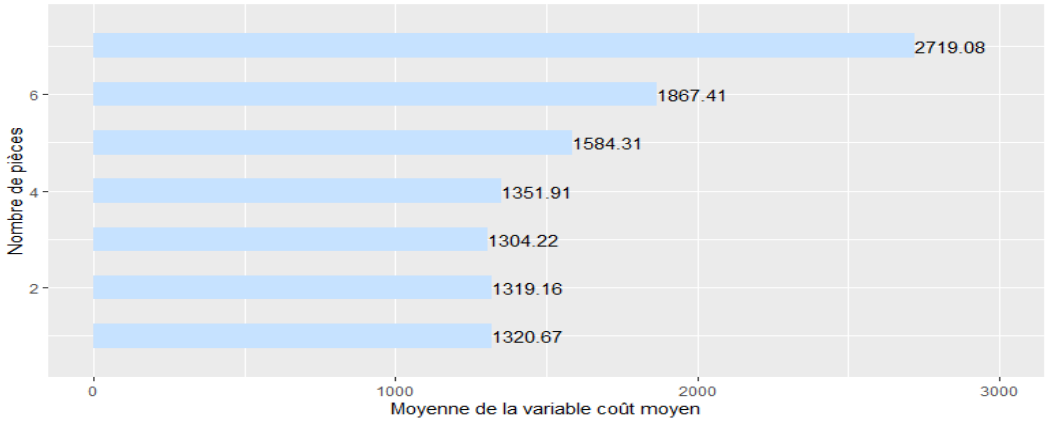


FIGURE 16 – Coût moyen de sinistre par nombre de pièces de l'appartement

■ Statistiques descriptives relatives à la fréquence vol :

Au niveau des contrats comptabilisant au moins un sinistre vol, 99,02% ont enregistré exactement un sinistre. Les contrats ayant deux sinistres représentent 0,96%. Les contrats faisant l'objet de trois sinistres ne pèsent que 0,02%. Aucun contrat de notre base de données n'a eu plus de 3 sinistres.

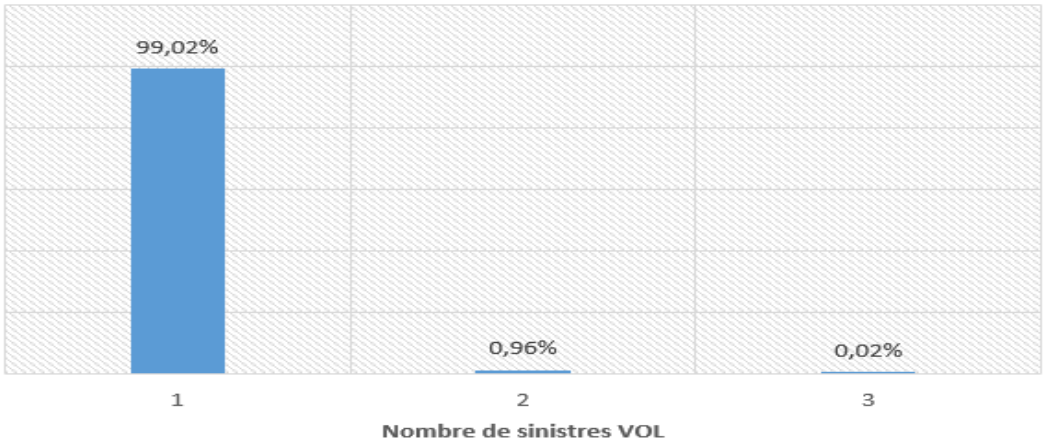


FIGURE 17 – Nombre de sinistres VOL

Une répartition selon le type de résidence montre que ce sont les résidences principales qui ont le nombre moyen de sinistres vol le plus élevé, à savoir 0,0050 sinistre. Au niveau des résidences secondaires, le nombre moyen de sinistres vol observé est 0,0030 sinistre. Suivant le statut des résidents, les propriétaires semblent plus susceptibles de subir des sinistres avec un nombre moyen de sinistres vol de 0,0069. Chez les locataires ce chiffre est nettement moins élevé, soit 0,0042 sinistre vol en moyenne.

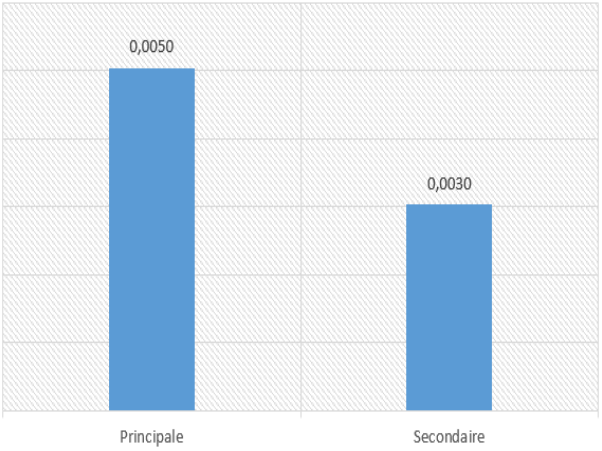


FIGURE 18 – Nombre moyen de sinistres vol par type de résidence

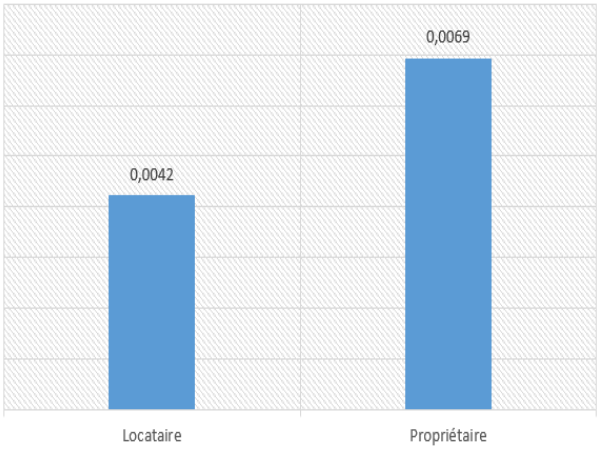


FIGURE 19 – Nombre moyen de sinistres vol par qualité du résident

Les graphiques qui suivent représentent le nombre moyen de sinistre vol par type de distributeur de police et le nombre moyen de sinistre vol par nombre de pièces de l'appartement. Il en ressort que les contrats distribués par les courtiers se voient attribués le nombre moyen de sinistres vol le plus élevé, soit 0,0066. Ils sont suivis par les contrats fournis par les agents avec une moyenne de 0,0047 sinistre vol et ceux conclus via les salariés qui enregistrent 0,0045 sinistre en moyenne. En outre, le nombre moyen de sinistre vol semble augmenter avec le nombre de pièces de l'appartement avec un seuil maximal autour de 6 pièces.

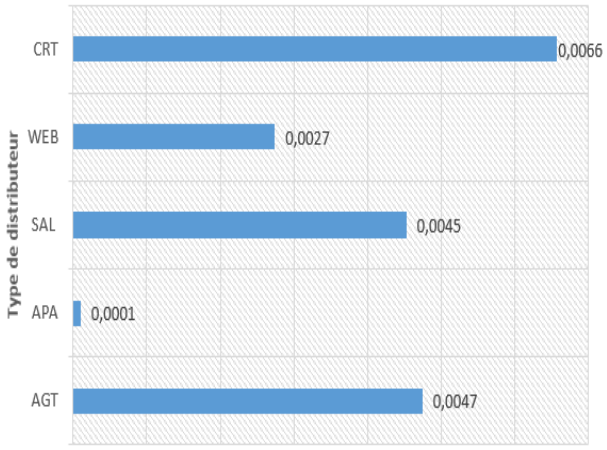


FIGURE 20 – Nombre moyen de sinistres vol par type de distributeur

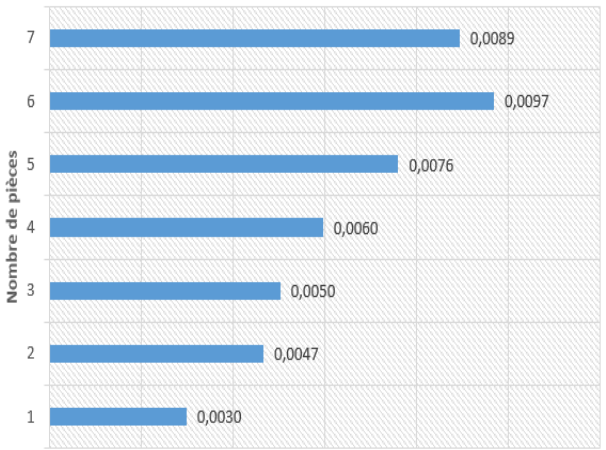


FIGURE 21 – Nombre moyen de sinistres vol par nombre de pièces

■ Statistiques descriptives spatiales :

Puisque notre objectif est de construire un zonier, il convient de compléter cette partie statistiques descriptives classiques par des statistiques descriptives spatiales. Nous allons visualiser la répartition du coût moyen de sinistres dégât des eaux appartements sur le territoire français. L'idée est de faire ressortir les variations de montants de sinistres suivant les différents endroits. La carte qui suit représente la répartition du coût moyen de sinistres

sur le territoire français. L'ampleur du sinistre est proportionnelle à la taille des points en bleus, chaque point représentant un contrat. Plus un point est grand, plus le montant de sinistre qu'il représente est élevé. Pour améliorer la visualisation, nous avons ajouté le fonds de carte de France. Ce qui permet d'identifier plus clairement les endroits à forts coûts de sinistres.

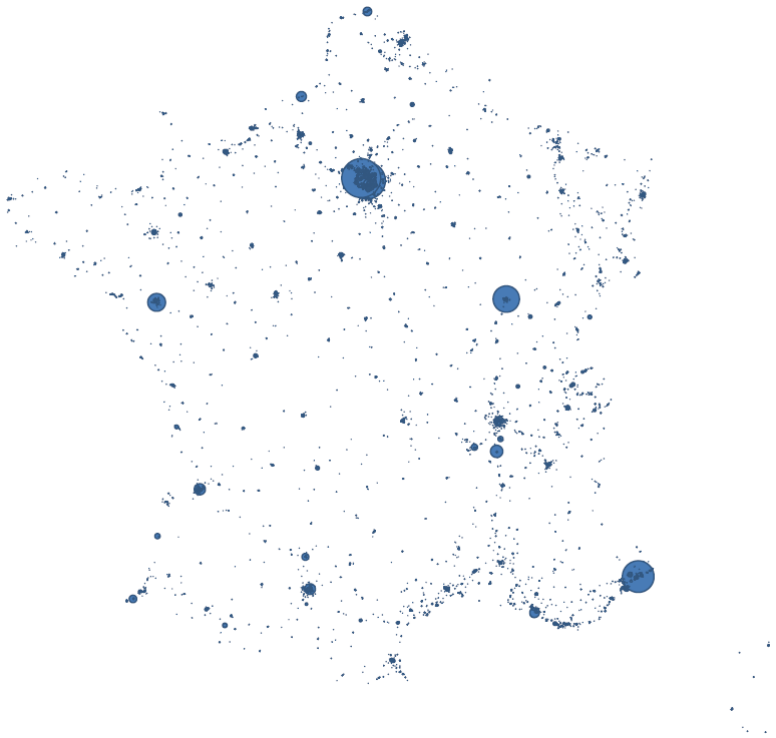


FIGURE 22 – Répartion du coût moyen de sinistres dégât des eaux sur le territoire français

Pour améliorer la visualisation, nous avons ajouté le fonds de carte de France (voir carte ci-dessous). Ce qui permet d'identifier plus clairement les endroits à forts coûts de sinistres. Un zoom sur la carte a permis de voir que les coûts de sinistres moyens par contrat les plus élevés sont observés au niveau de Paris, Nice, Dijon, Nantes, Lyon, Bordeaux et Toulouse. La ville de Paris connait un nombre important de sinistres dégâts des eaux à forts montants, comparé aux autres zones (voir carte zoom sur Paris).

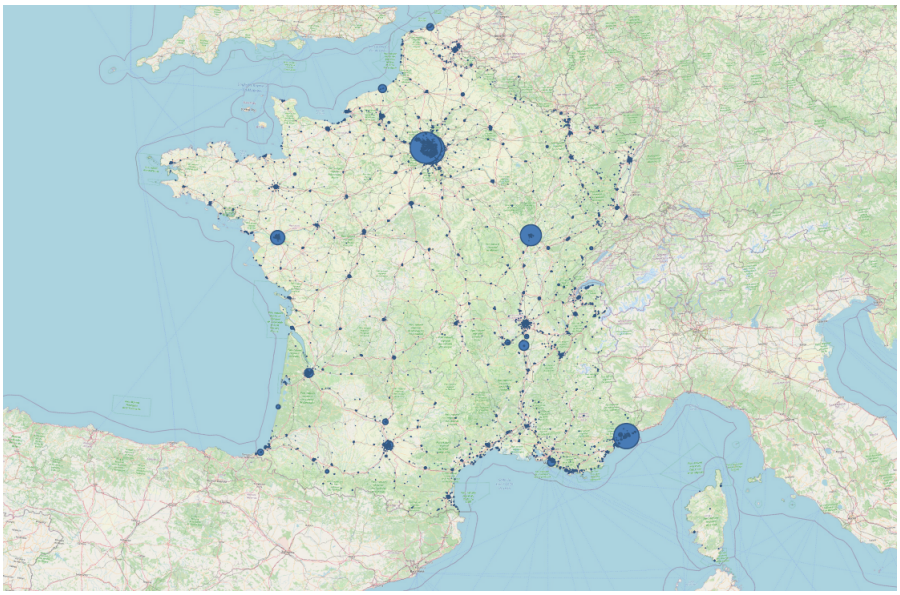


FIGURE 23 – Répartition spatiale du coût moyen dégât des eaux appartements (avec fond de carte)

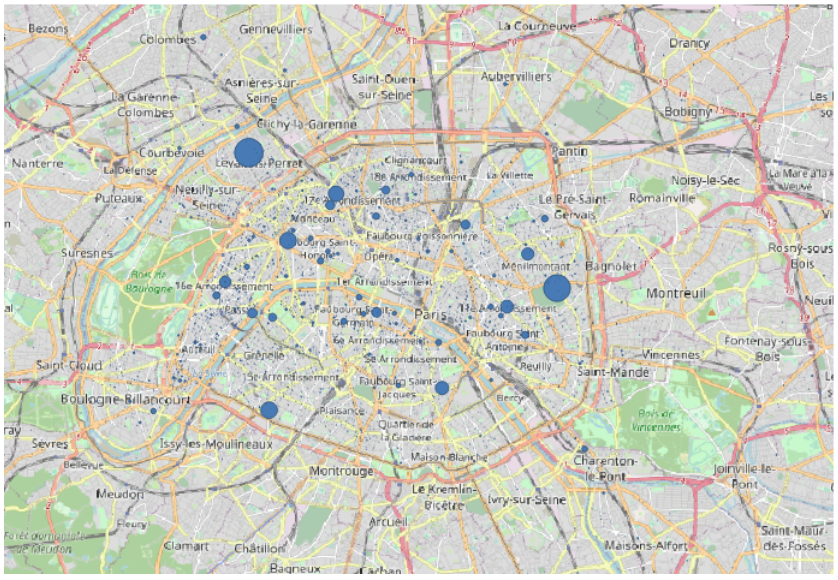


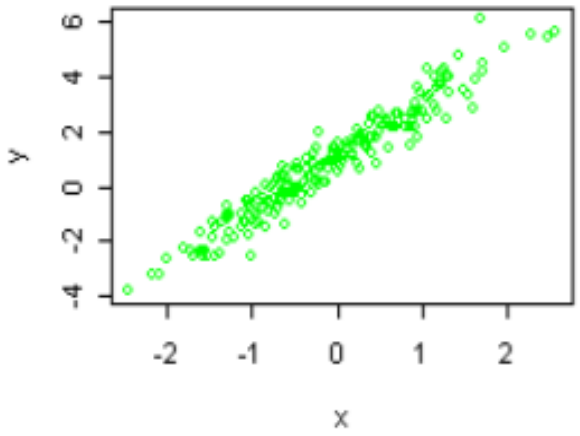
FIGURE 24 – Zoom sur Paris

2.3 Sélection des variables

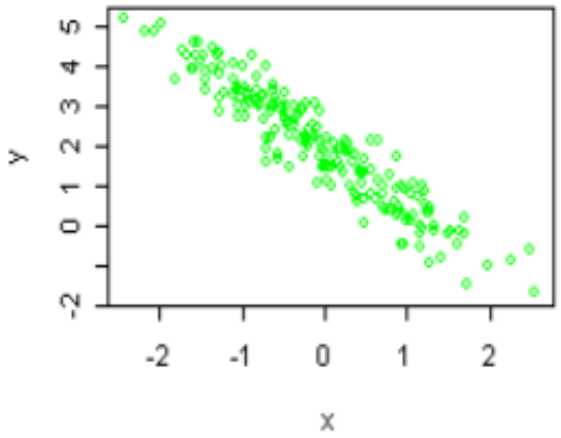
2.3.1 Analyse de la corrélation des variables

L'étude de la corrélation permet de mettre en évidence une éventuelle liaison (relation, dépendance) entre les variables. Nous disposons d'un nombre important de variables. L'analyse de la corrélation permettra d'éliminer les variables très corrélées, ce qui évitera d'avoir des vecteurs redondants dans nos modèles. Il convient tout d'abord d'éclaircir certaines terminologies. Deux variables sont dites liées si les variations de l'une dépendent de celles de l'autre variable. Lorsqu'une liaison existe, elle peut être positive ou négative. Une liaison positive traduit le fait que deux variables évoluent dans le même sens. Par contre, deux variables sont négativement liées si les valeurs de l'une des variables tendent à augmenter lorsque celles de l'autre variable diminuent. Deux variables sont dites indépendantes si elles varient indépendamment l'une de l'autre.
Dans un premier temps, une analyse graphique peut donner une idée sur le type de liaison entre deux variables.

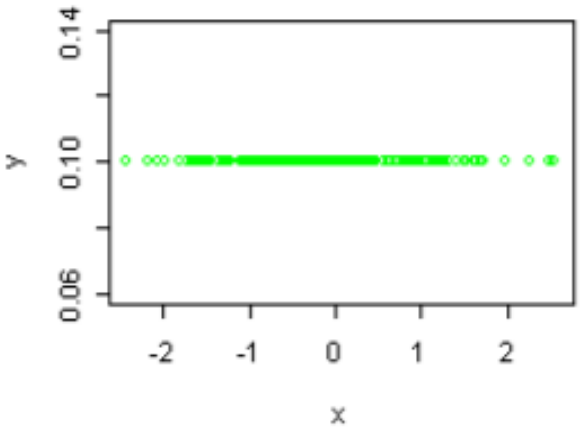
liaison positive



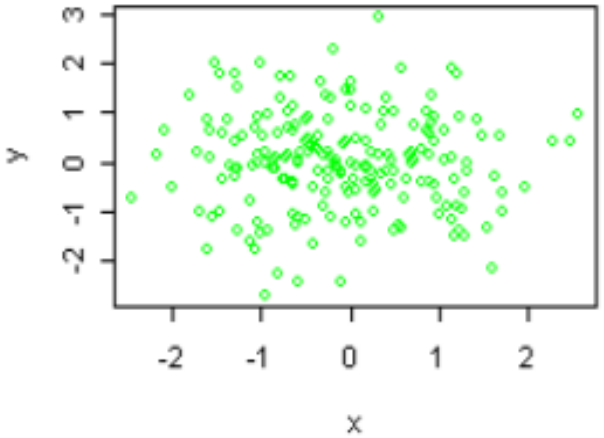
liaison négative



absence de liaison



absence de liaison



Pour mesurer l'ampleur réelle de la liaison, plusieurs indicateurs sont développés en

Statistique.
a) La covariance
Le calcul des coefficients de corrélation se base sur la notion de covariance. En Statistique, la covariance est une grandeur permettant de quantifier les déviations conjointes de deux variables par rapport à leurs espérances respectives. Elle permet ainsi de mesurer la liaison linéaire entre deux variables, de manière à déterminer le sens et l'intensité de la liaison.
La covariance peut s'écrire de deux manières :

$$\begin{aligned} COV(X,Y) &= E[X - E(X)][Y - E(Y)] \\ &= E[XY] - E[X]E[Y] \end{aligned}$$

Sur un échantillon de taille n, cette covariance est estimée par la covariance empirique :

$$\widehat{cov_{xy}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n}.$$

Interprétations :
-$COV(X,Y) > 0$ signifie que la relation entre X et Y est positive.
-$COV(X,Y) = 0$ indique une absence de corrélation linéaire entre X et Y.
-$COV(X,Y) < 0$ témoigne d'une relation négative entre X et Y.
Propriétés :
Soient X, Y et Z des variables aléatoires. La covariance vérifie les propriétés suivantes :
-$COV(X,Y) = COV(Y,X)$: propriété de symétrie.
-$COV(X,Y+Z) = COV(X,Y) + COV(X,Z)$: propriété de distributivité.
-$cov(X,a) = 0$ si a est une constante.
-$Var(X+Y) = Var(X) + Var(Y) + 2COV(X,Y)$: relation avec la variance
-X, Y indépendants $\implies COV(X,Y) = 0$.
La réciproque est fausse en général. Autrement dit, une covariance nulle ne permet pas d'affirmer l'indépendance entre des variables.
b) Le coefficient de corrélation de Pearson ;
Le coefficient de corrélation de Pearson est une normalisation de la covariance. Il permet de mesurer la liaison entre deux variables quantitatives. Il est de même signe que la covariance, avec les mêmes interprétations. Sa formule est donnée par :

$$\begin{aligned} r_{xy} &= \frac{COV(X,Y)}{\sqrt{Var(X)Var(Y)}} \\ &= \frac{COV(X,Y)}{\sigma_x \sigma_y} \end{aligned}$$

Ce coefficient de corrélation est approché sur un échantillon de taille n par le coefficient de corrélation empirique :

$$\widehat{r_{xy}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Puisqu'il est normalisé, le coefficient de corrélation est sans unité. Sa valeur est comprise entre -1 et +1.

Interprétations :
-Si $r_{xy} = +1$, la liaison linéaire entre X et Y est positive et parfaite.
-Si $r_{xy} = -1$, la liaison linéaire entre X et Y est négative et parfaite.
Test de significativité :
Souvent, on souhaite savoir si la laison entre deux variables est significative. Ceci revient à tester si r est significativement différent de 0. Le test d'hypothèse s'écrit :

$$H_0 : r = 0 \text{ hypothèse nulle}$$

$$H_1 : r \neq 0 \text{ hypothèse alternative}$$

,
avec comme statistique de test :

$$t = \frac{\hat{r}}{\sqrt{\frac{1-\hat{r}^2}{n-2}}}$$

Lorsque n est assez grand, t suit asymptotiquement une loi de Student à (n-2) degrés de liberté.
La région critique (ou de rejet) d'un tel est de la forme :

$$R.C. : |t| > t_{1-\alpha/2}(n-2)$$

où $t_{1-\alpha/2}(n-2)$ est le quantil d'ordre $1-\alpha/2$ de la loi de Student à (n-2) degrés de liberté. Dans notre étude, nous nous somme fixés comme seuil 0.9. Lorsque la valeur absolu de r est supérieure à 0.9, on considère que la corrélation entre les deux variables est forte et une seule des variables est conservée.

Variables	Coefficient de corrélation
Proportion de la population active ayant une formation Bac+5 — Proportion de cadres dans la population	0.99
Taux d'abonnés France Télécom professionnels — Taux d'abonnés France Télécom professionnels entreprise	0.99
Proportion de Cambriolages de Résidences Principales — Proportion de Cambriolages globale	0.98
Chauffage central du logement électrique — Combustible principal du logement électrique	0.98
Salaire Net Horaire Moyen — Salaire Net Horaire Moyen des cadres	0.98
Niveau de vie médian — Salaire Net Horaire Moyen	0.96
Niveau de vie médian — Salaire Net Horaire Moyen des cadres	0.93
Part d'interventions pour des incendies — Part d'interventions pour des incendies feux de cheminées	0.93

TABLE 2 – Les variables les plus corrélées

2.3.2 Sélection des variables

La partie « sélection des variables » est l'une des étapes préliminaires les plus importantes en machine learning. Le but est de supprimer les variables peu utiles qui risquent d'introduire un bruit supplémentaire dans notre modèle. Nous avions à notre disposition un grand nombre de variables externes. Pour sélectionner les plus pertinentes, nous avons utilisé le « boruta », un des plus puissants algorithmes de « features selection ».

■ Principe de l'algorithme Boruta :

La comparaison ne se fait pas avec les variables de base, mais plutôt avec des versions modifiées de ces variables. Ces nouvelles variables, appelées « shadow features » (variables/caractéristiques fantômes), sont obtenues par permutations aléatoires des valeurs de chaque variable. Pour illustrer considérons une base de trois features Var1, Var2 et Var3.

	Var1	Var2	Var3
1	$X_{1,1}$	$X_{2,1}$	$X_{3,1}$
2	$X_{1,2}$	$X_{2,2}$	$X_{3,2}$
3	$X_{1,3}$	$X_{2,3}$	$X_{3,3}$
4	$X_{1,4}$	$X_{2,4}$	$X_{3,4}$
5	$X_{1,5}$	$X_{2,5}$	$X_{3,5}$

TABLE 3 – Base des features

Notons Shadow_Var1, Shadow_Var2 et Shadow_Var3 les « shadow features » asssociés à features Var1, Var2, Var3. Les features de base et les shadow features sont combinés pour obtenir une nouvelle base de données appelée Boruta.

	Var1	Var2	Var3	Shadow_Var1	Shadow_Var2	Shadow_Var3
1	$X_{1,1}$	$X_{2,1}$	$X_{3,1}$	$X_{1,4}$	$X_{2,3}$	$X_{3,1}$
2	$X_{1,2}$	$X_{2,2}$	$X_{3,2}$	$X_{1,1}$	$X_{2,1}$	$X_{3,4}$
3	$X_{1,3}$	$X_{2,3}$	$X_{3,3}$	$X_{1,5}$	$X_{2,5}$	$X_{3,4}$
4	$X_{1,4}$	$X_{2,4}$	$X_{3,4}$	$X_{1,1}$	$X_{2,3}$	$X_{3,3}$
5	$X_{1,5}$	$X_{2,5}$	$X_{3,5}$	$X_{1,3}$	$X_{2,2}$	$X_{3,2}$

TABLE 4 – Base Boruta

Ensuite, un random forest est ajusté pour expliquer la variable d'intérée par les features de la base Boruta. L'importance de chaque features dans la modélisation est comparée à un seuil. Pour Boruta, ce seuil correspond à l'importance la plus élevée enregistrée parmi les shadow features. L'idée est qu'une variable n'est utile que si elle est capable de faire mieux que la meilleure variable parmi les shadow features. L'importance de cette approche est d'éviter de se fixer un seuil aléatoire de niveau d'importance permettant de décider de la pertinence ou non de retenir une variable. Illustrons la sélection de features avec un exemple.

	Var1	Var2	Var3	Shadow_Var1	Shadow_Var2	Shadow_Var3
Feature importance en %	51	30	18	21	24	29
Selectionné	oui	oui	non	-	-	-

TABLE 5 – Features selection

Dans notre exemple l'importance la plus élevée enregistrée par les shadow features est $max\{21\%, 24\%, 29\%\} = 29\%$. Les variables retenues pour expliquer la variable d'intérêt sont donc les features de base dont les niveaux d'importance dépassent 29, c'est-à-dire Var1 et Var2. Pour s'assurer d'avoir des résultats robustes, ce processus est répété plusieurs fois.

3 Modélisation

La première étape consiste à modéliser le coût moyen de sinistre dégât des eaux et la fréquence de sinistre vol en excluant les variables géographiques, ceci avec les données de la base d'apprentissage. Ensuite, il s'agira d'isoler l'information géographique contenue dans les résidus d'un modèle linéaire généralisé de prime pure hors variables géographiques. Pour opérer un tel isolement, nous modéliserons ces résidus avec les variables externes à l'aide d'un modèle de machine learning, le catboost. Les prédictions de résidus obtenues avec le catboost, nommées résidus spatiaux dans la littérature, seront ensuite scindées en différentes classes. Ces dernières constitueront le zonier final.

3.1 Modélisation hors variables externes

A ce niveau les modèles de coût moyen dégât des eaux et de fréquence vol calibrés par Axa France ont été utilisés.

3.1.1 Présentation du modèle linéaire généralisé

Les modèles linéaires généralisés sont utilisés pour étudier l'éventuelle liaison entre une variable dépendante Y et un ensemble de variables explicatives $X_1, X_2, ..., X_k$. Un modèle linéaire généralisé (glm) est caractérisé par une composante aléloire Y, une fonction déterministe et une fonction de lien. :
- composante aléatoire : les $(Y_i)_i$ sont indépendants et appartiennent à une loi de famille exponentielle $\mathcal{F}(\theta_i, \phi_i, a, b, c)$ [1].
- une fonction déterministe : le vecteur de variables explicatives X_i donne le prédicteur linéaire $\eta_i = X_i^T \beta$.
- une fonction de lien g : $\mathbb{R} \to \bar{\mathbb{X}}$ monotone, différentiable et inversible telle que $\mathbb{E}(Y_i) = g^{-1}(\eta_i) = \mu_i$, pour i $\in \{1, ..., n\}$, g est appelée fonction de lien.

Les fonctions de lien les plus utilisées sont résumées dans le tableau suivant.

Loi	Lien	Moyenne
Bernoulli $B(\mu)$	logit : $\eta = log(\frac{\mu}{1-\mu})$	$\mu = \frac{1}{1+e^{-x^T\beta}}$
Poisson $P(\mu)$	log : $\eta = log(\mu)$	$\mu = e^{x^T\beta}$
Gamma(α, β)	inverse : $\eta = \frac{1}{\mu}$	$\mu = (x^T\beta)^{-1}$
Normale $N(\mu, \sigma^2)$	identité : $\eta = \mu$	$\mu = x^T\beta$
Inverse Gaussienne (μ, λ)	inverse au carré : $\eta = -\frac{1}{\mu^2}$	$\mu = -(x^T\beta)^{-2}$

TABLE 6 – Quelques fonctions de lien

A partir de ces lois de base, d'autres lois peuvent être construites (lognormale, loggamma...).

1. La densité d'une loi de la famille exponentielle s'écrit : $ln(f_X(x)) = \frac{\theta x - b(\theta)}{a(\phi)}$ où θ_i est le paramètre d'échelle, ϕ_i le paramètre de dispersion, et a, b, c trois fonctions.

■ Estimation :

L'estimation du vecteur de paramètres β et du paramètre- ϕ se fait par la méthode de maximum de vraisemblance. La log-vraisemblance s'écrit :

$$ln\mathcal{L}(\beta,\phi) = \sum_{i=1}^{n} \frac{\theta_i(\beta)y_i - b(\theta_i(\beta))}{a(\phi_i)}.$$

Les conditions de premier ordre appelées équations du score sont :

$$\forall j = 1, ..., d, \sum_{i=1}^{n} \frac{(y_i - \mu_i)x_{ij}}{a(\phi_i)^2} = 0.$$

Pour évaluer les performances d'un glm, plusieurs indicateurs peuvent être utilisés. Dans ce rapport, nous avons utilisé l'AIC et la déviance.
$AIC(\hat{\beta}) = -2log\mathcal{L}(\hat{\beta}) + 2(p+1)$.
La déviance quant à elle est calculée suivant la formule :
$D(\hat{\mu}) = -2log(\mathcal{L}(\hat{\mu})/\mathcal{L}(N))$, avec $\mathcal{L}(N)$ la log-vraisemblance du modèle saturé.

3.1.2 Résultats des glm des modélisations sans variable géographique

Comme on pouvait s'y attendre toutes les variables tarifaires, retenues classiquement par Axa pour modéliser le coût moyen dégât des eaux et la fréquence vol, se sont avérées significatives. Globalement, les résultats du modèle confirment les statistiques descriptives effectuées plus haut.

■ Résultats de la modélisation du coût moyen dégât des eaux avec les variables internes :

Pour un contrat donné, le coût moyen est défini par :

$$\text{coût moyen} = \frac{\text{charge de sinistres du contrat}}{\text{nombre de sinistres du contrat} \times \text{exposition}}$$

Les résultats du glm retenu sont présentés dans le tableau suivant. Nous avons uniquement conservé les sinistres à coûts non nuls. La loi lognormale présente les meilleurs résultats. Comparées aux résidences principales, les résidences secondaires ont tendance à enregistrer des coûts moyens de sinistre dégât des eaux plus élevés. Le nombre de pièces de l'appartement est positivement corrélé avec le montant moyen de sinistres. Lorsque l'occupant est propriétaire, il a plus de chance d'avoir des coûts de sinistre élevés. Les années 2017 et 2018 comptabilisent des sinistres plus couteux, relativement à l'année 2016. Les sociétés enregistrent en moyenne des coûts de sinistres plus élevés que les particuliers. Il convient de noter que les sinistres extrêmes sont écartés.

	Estimate	Std. Error	t value	Pr(>|t|)	sig
(Intercept)	8.73411	0.0137	480.61	0.0000	***
resid_typeS	0.210007	0.0128	12.37	0.0000	***
resid_nb_piecesS	0.04788	0.0026	13.71	0.0000	***
resi_qualiteP	0.180348	0.0072	18.79	0.0000	***
annee_vision2017	0.054796	0.0078	5.28	0.0000	***
annee_vision2018	0.189658	0.0075	19.04	0.0000	***
mnt_obj_valeur[3000,8000[	0.052535	0.0126	3.14	0.0017	***
mnt_obj_valeur]0,2000[	-0.007714	0.0140	-0.42	0.6768	
mnt_obj_valeur>= 8000	0.195776	0.0131	11.23	0.0000	***
mnt_obj_valeur0	0.007714	0.0109	0.54	0.5915	
res_statut1	0.106799	0.0116	6.90	0.0000	***

TABLE 7 – Résultats du glm par une lognormale

Metrics	gamma	lognormal	loggamma
AIC	1099458.004	162136.420	165091.863
Deviance	48740.339	43902.591	47026.015

TABLE 8 – Performances

■ Résidus :

Les résidus du glm peuvent être calculés de plusieurs manières. Nous avons calculé les résidus sous forme de proportion entre le vrai coût moyen et le coût moyen prédit. Ce qui nous permet d'une part d'éviter les problèmes d'échelle lorsqu'il s'agira de représenter le zonier sur une carte. Les valeurs du zonier étant les prédictions de nos résidus fournies par la modélisation de l'effet géographique.
Plus précisément, les résidus sont définis par :

$$\text{résidus} = \frac{\text{coût moyen observé}}{\text{coût moyen prédit}}$$

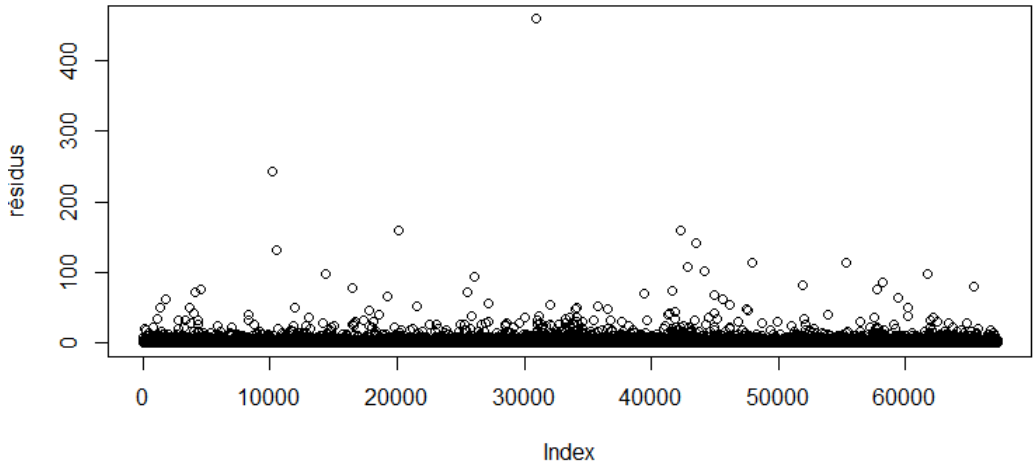


FIGURE 25 – Résidus du glm

■ Résultats de la modélisation de la fréquence vol avec les variables internes :

Le tableau qui suit présente les résultats du glm de la fréquence vol expliquée par les variables internes. La loi binomiale négative fournit les meilleures performances. Pour la variable explicative "âge du client" la classe des clients dont l'âge est inférieur à 39 ans représente la modalité de référence. Il ressort du glm que les coefficients de toutes les autres modalités de la variable "âge du client" sont négatifs et significatifs au seuil de 5%. On en déduit que les assurés jeunes ont tendance à subir plus de sinistres vol en moyenne. Les résidences principales ont des fréquences de sinistre vol plus élevées comparées aux résidences secondaires. Comme suggéré par nos statistiques descriptives, la fréquence de sinistre vol augmente avec le nombre de pièces de l'appartement. De même, lorsque l'occupant est le propriétaire les sinistres vol sont plus nombreux en moyenne. Comparés aux autres types de distributeurs, les courtiers voient leurs contrats distribués enregistrer les fréquences de sinistre les plus grandes.

	Estimate	Std. Error	t value	Pr(>|t|)	sig
(Intercept)	-4.8944	0.1466	-33.40	0.0000	***
CLI_age_client]39,45]	-0.1514	0.0634	-2.39	0.0170	**
CLI_age_client]45,50]	-0.1697	0.0625	-2.71	0.0067	***
CLI_age_client]50,55]	-0.2468	0.0638	-3.87	0.0001	***
CLI_age_client]55,61]	-0.3028	0.0624	-4.85	0.0000	***
CLI_age_client]61,66]	-0.3533	0.0709	-4.98	0.0000	***
CLI_age_client]66,72]	-0.4797	0.0708	-6.77	0.0000	***
CLI_age_client]72,81]	-0.5351	0.0736	-7.27	0.0000	***
CLI_age_client> 81	-0.7581	0.0764	-9.92	0.0000	***
HAB_CDRESIDS	-0.7020	0.0770	-9.12	0.0000	***
HAB_NBPIECS	0.1223	0.1399	0.87	0.0382	**
HAB_qualP	0.3831	0.0382	10.03	0.0000	***
POL_DISTRIBAGT	-13.7340	1428.5889	-0.01	0.9923	
POL_DISTRIBAPA	-14.3186	416.7414	-0.03	0.9726	
POL_DISTRIBSAL	-0.0509	0.0988	-0.52	0.6061	
POL_DISTRIBWEB	-0.2509	0.1678	-1.50	0.0134	**
POL_mtobv[3000,8000[	0.1943	0.0624	3.11	0.0018	***
POL_mtobv]0,2000[	-0.1559	0.0662	-2.35	0.0185	**
POL_mtobv>= 8000	0.5879	0.0674	8.73	0.0000	***
POL_mtobv0	-0.1480	0.0531	-2.79	0.0053	**

TABLE 9 – Résultats du glm par une binomiale négative

Metrics	poisson	binomiale négative
AIC	55678.5759	55639.5722
Deviance	46657.1947	40628.0570

TABLE 10 – Performances

3.2 Modélisation de l'effet géographique

3.2.1 Explication de la méthode

Le CatBoost est un algorithme de gradient boosting basé sur les arbres de décision. Le gradient boosting est une technique d'apprentissage automatique pour la régression, la classification et d'autres tâches, qui produit un modèle de prédiction sous la forme d'une combinaison de weak learners (apprenants faibles). L'idée est d'effectuer un apprentissage séquentiel qui fonctionne sur le principe d'un ensemble, où chaque modèle tente de corriger les erreurs du modèle précédent. Les weak learners sont ensuite agrégés pour former un prédicteur fort. Le CatBoost est une variante du XGBoost, qui est aussi basé sur le boosting, mais avec souvent des performances plus élevées. Le gradient boosting construit le modèle d'ensemble de manière itérative. A chaque itération, l'algorithme apprend un arbre pour réduire l'erreur d'apprentissage commise à l'itération précédente. En optimisant la fonction d'estimation à chaque étape (avec la possibilité pour l'utilisateur de choisir une fonction de perte de son choix), le gradient boosting a donc une complexité plus grande que les forêts aléatoires.
Le fonctionnement du CatBoost se résume en 3 étapes :
-Etape 1 : L'algorithme lit les données et attribue le même poids à chaque observation de l'échantillon.
-Etape 2 : Les pires prédictions du weak learner de base sont identifiées. À l'itération suivante, ces prédictions inexactes sont fournies au prochain learner avec un poids plus élevé, ce qui permet d'améliorer au fur et à mesure le pouvoir prédictif du modèle.
-Etape 3 : Répétez l'étape 2 jusqu'à ce que l'algorithme puisse prédire au mieux la variable cible.

Indice de gini :
Pour évaluer les performances du Catboost, nous avons utilisé le coefficient de gini normalisé. L'indice de Gini correspond au ratio d'aire A=(A + B) o'u A est l'aire entre la courbe de Lorenz et la première diagonale et A + B l'aire du triangle entre la première diagonale et l'axe des abscisses (valant 1/2). Mathématiquement, il est défini par :
$G = \frac{1}{2} \int_R \int_R |x - y| dF_X(x) dF_X(y)$.
En machine learning, la métrique souvent utilisée est le coefficient de Gini normalisé, qui est égal au rapport entre le coefficient de Gini du modèle calibré sur le coefficient de Gini du modèle idéal.

3.2.2 Résultats du catboost

Cette modélisation est effectuée sur la base d'apprentissage et ses performances sont évaluées sur la base de validation. Les performances du zonier final sont évaluées sur la base de test. L'avantage du catboost réside dans son pouvoir de prédiction. C'est l'un des modèles de machine learning les plus puissants à ce jour. La figure suivante fournit les résultats de la modélisation pour le coût moyen dégât des eaux. Les variables externes qui contribuent le plus à l'explication du résidus sont la distance par rapport aux plus proches voisins, le score de fréquentation touristique autour de l'adresse (segment de rue à fort trafic de touristes et

avec points d'intérêts touristiques à proximité), le prix moyen au m2 à l'adresse, la hauteur moyenne de tous les bâtiments référencés aux alentours, les installations sanitaires en dehors du logement, le taux d'artisans, de commerçants et de chefs entreprises logeant dans la zone, l'indice de dynamisme Particuliers, l'indice d'attraction commerciale de l'endroit, le nombre risque terrestre et les variables expérience client.
Concernant les performances du catboost, le gini s'élève à 16,13% sur la base d'apprentissage et 10,29% sur la base test. Les prédictions sur les résidus obtenues sont découpées en 20 classes suivant les quantiles pour constituer le zonier final.

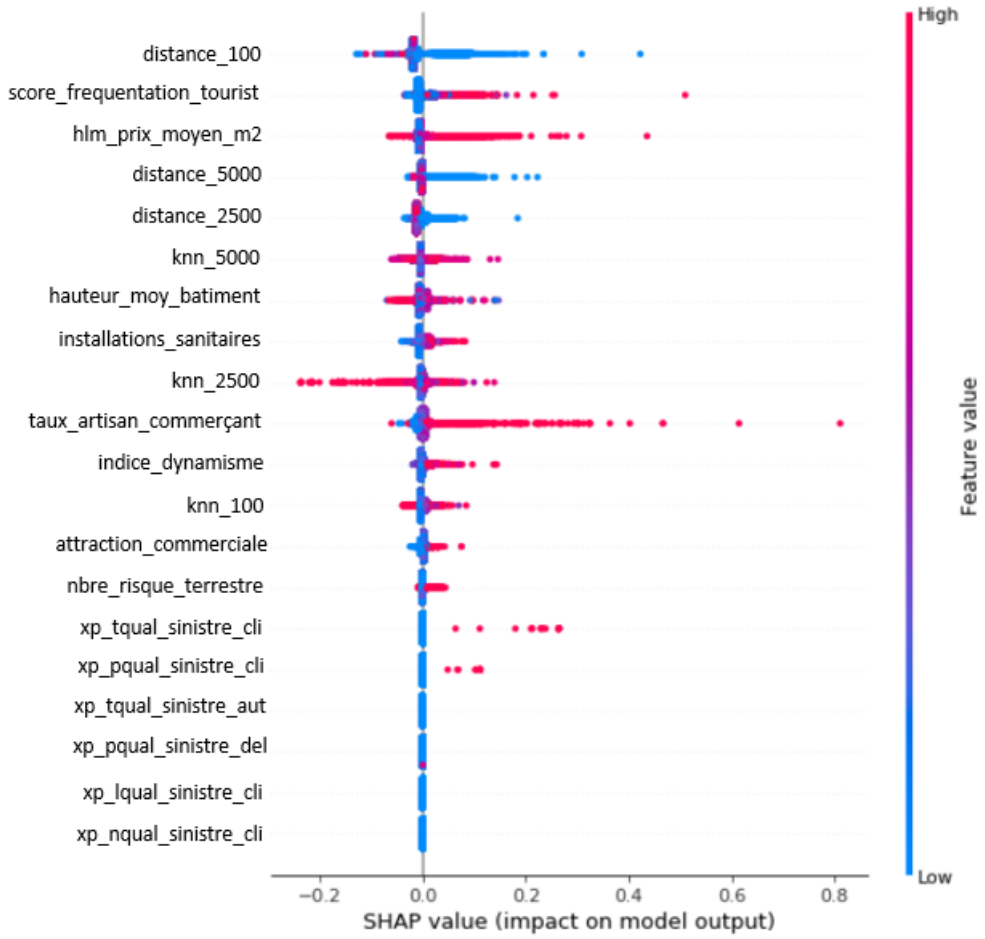


FIGURE 26 – Résultats du catboost coût moyen dégât des eaux

Les résultats du catboost qui explique les résidus de la fréquence vol par les variables externes sont très intuitifs. En effet, les variables explicatives géographiques relatives à la criminalité et au vol sont sorties significatives. La variable "cambriolages des résidences principales entre 2012 et 2015" est positivement corrélée aux résidus du glm vol. Ce résultat comporte une certaine logique. En outre, les milieux à températures élevées comptabilisent les résidus de sinistre vol les plus grands en moyenne. Ceci pourrait être expliqué par le fait qu'en milieu froid, les gens ont tendance à rester davantage dans leurs domiciles, ce qui peut dissuader les voleurs. D'autres variables criminogènes telles que les interventions pour incendies dans les habitations, les interventions pour incendies causés par les cheminées apparaissent aussi

significatives. Les résidences principales occupées par les propriétaires comptent aussi des résidus élevés.

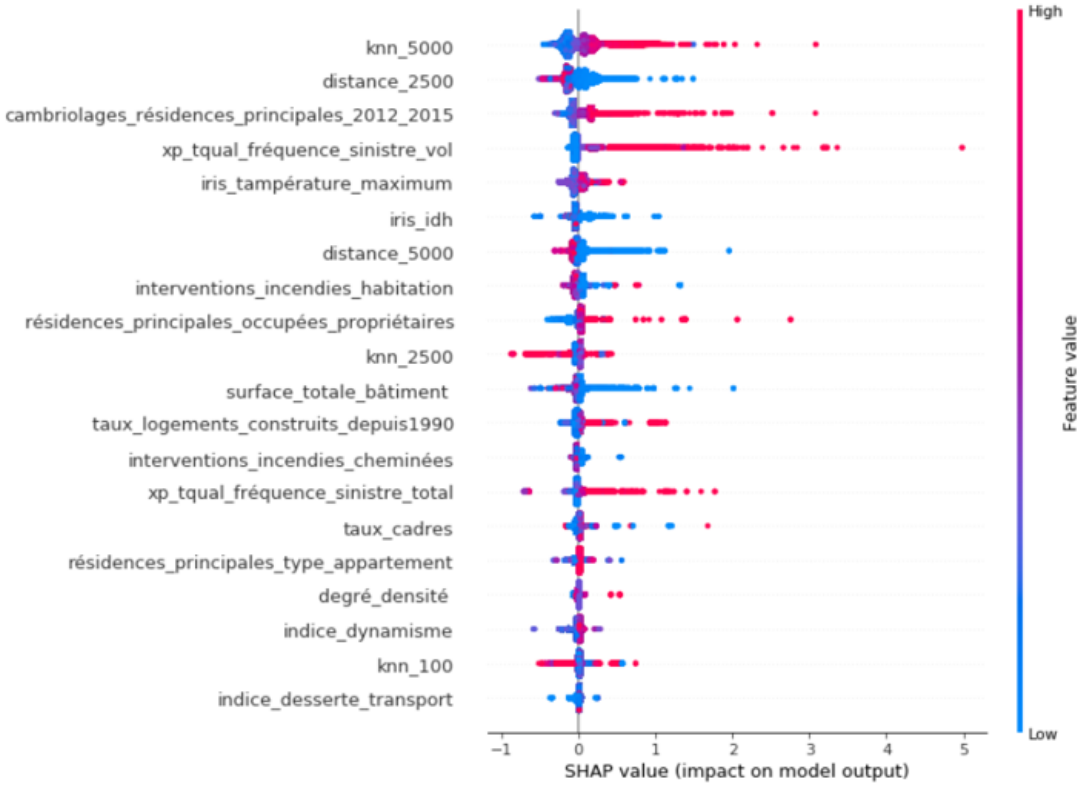


FIGURE 27 – Résultats du catboost fréquence vol

3.2.3 Découpage de la France en polygones par la méthode de Voronoi

Cette partie aborde la construction de la maille du zonier. Le choix de la maille la plus adaptée est important puisqu'il détermine les tailles des zones sur lesquelles l'effet géographique sera estimé. Nous avons choisi, dans le cadre de cette étude, d'explorer les mailles fines. Il convient de noter que la taille des découpages (ou polygones) dépend de la densité des contrats d'assurance. Les zones à fortes densités de contrats auront des superficies très petites car comme dit plus haut chaque contrat va engendrer un polygone constitué des points de l'espace qui lui sont plus proches. Les zones à faibles densités auront des polygones assez grands. Les cartes qui suivent présentent nos découpages spatiaux de la France grâce à l'algorithme de Voronoi. Ces cartes sont obtenues après intersection spatiale entre nes découpages bruts de Voronoi et le contour de la France. Le découpage pour la fréquence vol est beaucoup plus fin que celui du coût moyen. Ceci est dû est au fait que le découpage pour la fréquence vol est réalisé sur toutes les observations de la base d'apprentissage alors que pour le coût moyen le découpage est fait sur les contrats à coûts de sinistre non nuls.

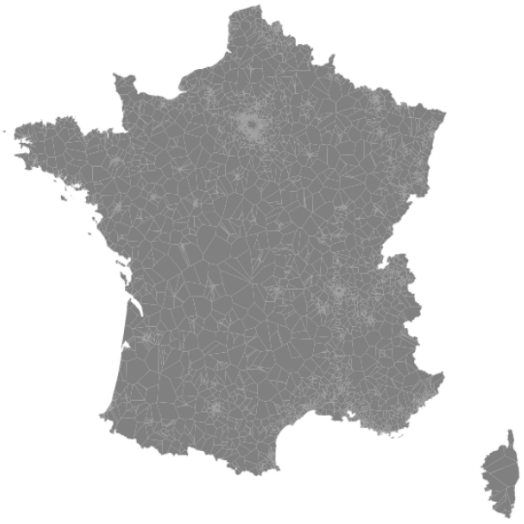


FIGURE 28 – Notre découpage de la France en des polygones de Voronoi (coût moyen dégât des eaux)

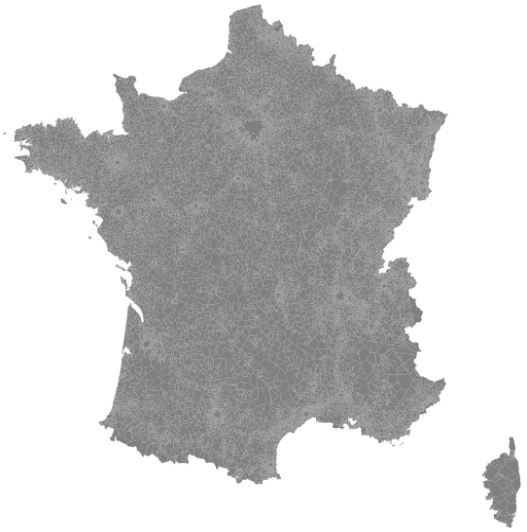


FIGURE 29 – Notre découpage de la France en des polygones de Voronoi (fréquence vol)

La comparaison qui suit montre que notre maille est beaucoup plus fine que les découpages administratifs. L'algorithme de Voronoi permet de découper une zone en des sous-zones aussi petites qu'on le souhaite, pourvu que l'on dispose d'un nombre de points important. Rappelons que le découpage iris est le plus petit découpage administratif de France (voir figures 6).

FIGURE 30 – Carte plan de zonage iris pour Paris

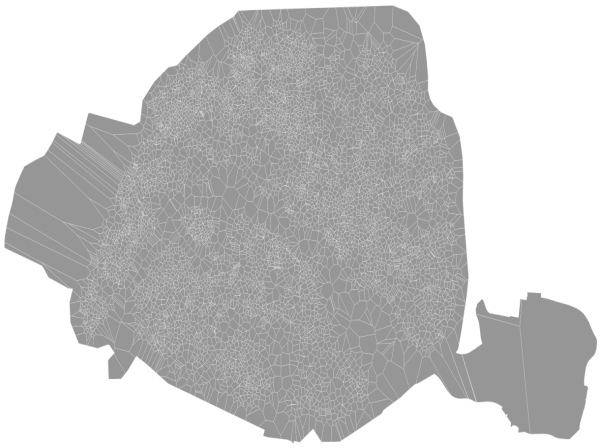


FIGURE 31 – Notre découpage de Paris par Voronoi

■ Interprétation de nos micro-zoniers :

Les résidus spatiaux découpés en classes puis représentés sur une carte donnent le zonier final. Les cartes qui suivent représentent nos micro-zoniers obtenus. A gauche, nous avons le micro-zonier coût moyen dégâts des eaux et à droite le micro-zonier vol. Concernant la légende, plus le numéro de la classe est élevé, plus le signal géographique associé est grand. Ainsi la couleur verte est associée aux zones à faible intensité du phénomène étudié (coût pour le dégât des eaux et fréquence pour le vol). Les zones à plus grandes intensités sont

celles coloriées en rouge. Pour le micro-zonier vol, on recense les fréquences de sinistre les plus élevés autour de Paris, Lyon, Grenoble, Marseille, Nice, Toulouse, Bordeaux, Nantes et Rennes. Ces tendances sur les vols dans les domiciles sont similaires aux chiffres du ministère de l'intérieur français (voir annexe 7). Pour le micro-zonier vol, l'intensité du phénomène reste importante autour de Paris mais la tendance n'est plus la même sur toute l'étendue de la France. Le micro-zonier dégâts des eaux est moins lisse que celui de la fréquence vol. En effet, on note beaucoup plus de disparités au niveau du micro-zonier coût moyen. Ceci peut s'expliquer par fait que ce micro-zonier compte beaucoup moins de polygones car étant construit uniquement sur les contrats à coûts de sinistre non nuls.

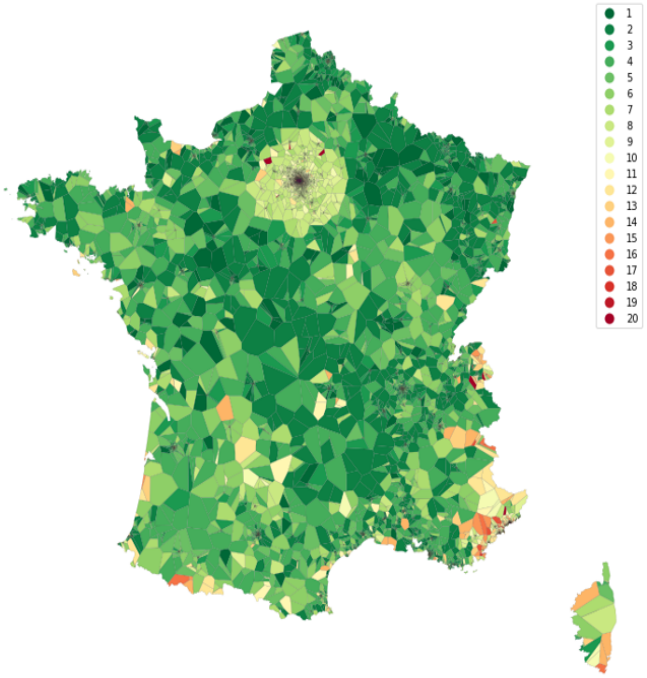


FIGURE 32 – micro-zonier coût moyen dégât des eaux

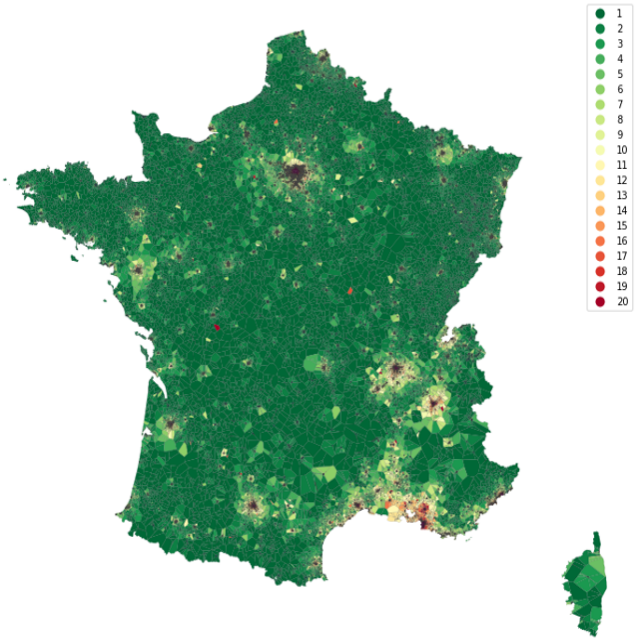


FIGURE 33 – micro-zonier fréquence vol

Pour mieux visualiser la structure granulaire de nos micro-zoniers, nous avons effectué un zoom autour de Paris. Ceci permet de mieux observer nos polygones. Les graphiques qui suivent représentent ces zooms. Comme nous pouvons le constater, les coûts moyens de sinistre dégâts des eaux sont plus importants dans Paris et dans ses environs proches. On note une concentration de sinistres à coûts élevés à ces endroits. De son côté, la fréquence de sinistre reste élevée dans Paris et environs mais la portée en termes de distance est plus faible autour de Paris. La fréquence vol apparait ainsi comme un phénomène plus local que le coût de sinistre dégât des eaux, la segmentation géographique des risques y étant plus nette.

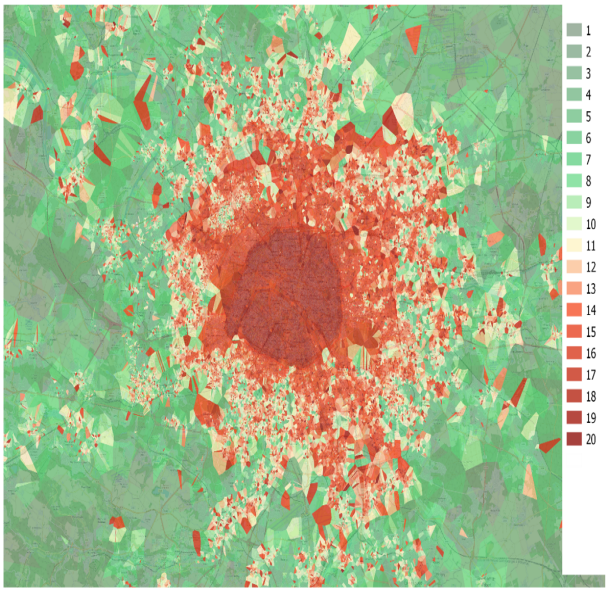


FIGURE 34 – micro-zonier du coût moyen dégât des eaux

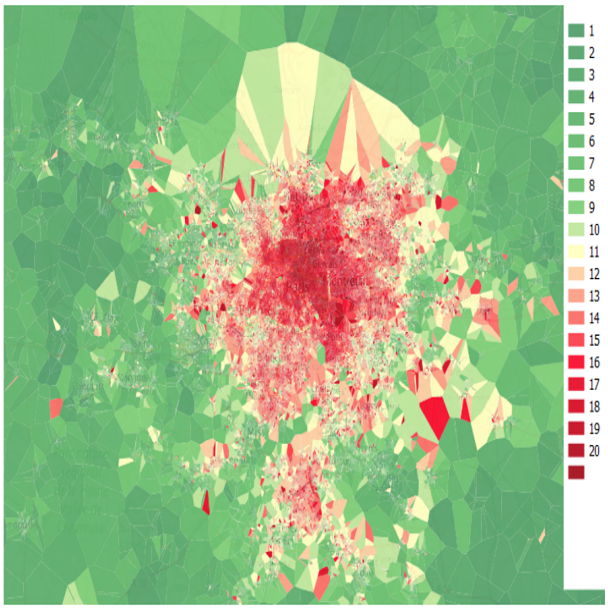


FIGURE 35 – micro-zonier de la fréquence vol

3.3 Intégration des micro-zoniers dans les modèle de coût moyen et de fréquence vol sur la base de test :

Nous avons mesuré l'apport du zonier sur la base de validation. Pour ce faire, une première modélisation du coût moyen est effectuée, sans le micro-zonier. Ensuite, une deuxième modélisation du coût moyen et de la fréquence col est réalisée, cette fois-ci en ajoutant le micro-zonier dans les variables explicatives tarifaires. Nous avons ilustré dans ce qui suit les résultats des glm sans le zonier puis avec le zonier, ceci pour le coût moyen dégâts des eaux. On remarque que le zonier représenté par la variable "cluster" est significative.

	Estimate	Std. Error	t value	Pr(>|t|)	sig
(Intercept)	8.708441	0.0234	280.27	0.0000	***
resid_typeS	0.147763	0.0205	5.42	0.0000	***
resid_nb_piecesS	0.054796	0.0045	9.07	0.0000	***
resi_qualiteP	0.207081	0.0124	12.57	0.0000	***
annee_vision2017	0.03325	0.0132	1.89	0.0582	**
annee_vision2018	0.191121	0.0127	11.28	0.0000	***
mnt_obj_valeur[3000,8000[	0.000532	0.0214	0.02	0.9854	
mnt_obj_valeur]0,2000[	-0.002793	0.0231	-0.09	0.9280	
mnt_obj_valeur>= 8000	0.136724	0.0228	4.50	0.0000	***
mnt_obj_valeur0	-0.002394	0.0183	-0.10	0.9230	
res_statut	0.119567	0.0198	4.53	0.0000	***

TABLE 11 – Résultats du glm sans zonier par une lognormale

	Estimate	Std. Error	t value	Pr(>|t|)	sig
(Intercept)	8.483006	0.0349	182.72	0.0000	***
resid_typeS	0.133399	0.0204	4.91	0.0000	***
resid_nb_piecesS	0.063707	0.0045	10.60	0.0000	***
resi_qualiteP	0.171703	0.0124	10.45	0.0000	***
annee_vision2017	0.029393	0.0131	1.69	0.0909	*
annee_vision2018	0.195111	0.0126	11.63	0.0000	***
mnt_obj_valeur[3000,8000[	-0.018088	0.0212	-0.64	0.5226	
mnt_obj_valeur]0,2000[	0.002793	0.0229	0.09	0.9272	*
mnt_obj_valeur>= 8000	0.0532	0.0229	1.74	0.0813	*
mnt_obj_valeur0	-0.011571	0.0181	-0.48	0.6314	
res_statut	0.111853	0.0197	4.28	0.0000	***
cluster2	0.118902	0.0391	2.29	0.0221	***
cluster3	0.100681	0.0365	2.07	0.0380	***
cluster4	0.106666	0.0373	2.15	0.0313	***
cluster5	0.076475	0.0374	1.54	0.1239	
cluster6	0.108661	0.0377	2.17	0.0304	***
cluster7	0.164122	0.0368	3.36	0.0008	***
cluster8	0.087115	0.0367	1.79	0.0741	**
cluster9	0.174496	0.0364	3.60	0.0003	***
cluster10	0.128079	0.0360	2.68	0.0074	***
cluster11	0.196574	0.0360	4.11	0.0000	***
cluster12	0.262808	0.0354	5.58	0.0000	***
cluster13	0.261345	0.0359	5.47	0.0000	***
cluster14	0.252833	0.0350	5.43	0.0000	***
cluster15	0.286349	0.0338	6.37	0.0000	***
cluster16	0.163856	0.1434	0.86	0.3903	
cluster17	0.186599	0.0354	3.96	0.0001	***
cluster18	0.34048	0.0329	7.78	0.0000	***
cluster19	0.37373	0.0315	8.92	0.0000	***
cluster20	0.478002	0.0312	11.51	0.0000	***

TABLE 12 – Résultats du glm avec zonier par une lognormale

Pour mesurer l'apport du micro-znoier, nous avons comparé les indices de Gini de ces deux modèles (sans le micro-zonier et avec le micro-zonier), tant pour le coût moyen que pour le vol. Nous les avons aussi comparés aux performances du zonier actuel d'AXA. Le zonier actuellement utilisé par AXA est construit à la maille commune.

■ Comparaison des indices de Gini normalisés :

a) Pour les modèles de coût moyen dégât des eaux :
Il convient de rappeler que tous les contrats de nos bases de données ne sont pas bien géocodés. Nous avons évalué l'indice de Gini d'une part sur toutes les observations de notre base de données de test et d'autre part sur uniquement les observations bien géocodées. L'avantage d'une telle segmentation permet d'appréhender l'importance du géocodage dans la construction d'un micro-zonier. On remarque que pour le coût moyen dégât des eaux, le zonier d'AXA est meilleur que le micro-zonier sur toutes les observations. Mais lorsqu'on considère uniquement les appartements bien géocodés, notre micro-zonier donne de meilleures performances que le zonier actuel d'AXA. Ce qui confirme l'importance de bien géocoder les contrats.On note aussi l'importance du lissage qui a permis de diminuer le sur-apprentissage de nos modèles.

	Zonier actuel d'AXA	Notre Micro-zonier avant lissage	Notre Micro-zonier lissé
Gini normalisé (base d'apprentissage)	0,2987	0,2998	0,3028
Gini normalisé (base de test)	0,2874	0,2785	0,2787
Gain p/r au zonier d'AXA (sur test)	-	-0,88%	-0,87%

TABLE 13 – Performances des différents zoniers sur toutes les observations de la base de test

	Zonier actuel d'AXA	Notre Micro-zonier avant lissage	Notre Micro-zonier lissé
Gini normalisé (base d'apprentissage)	0,3332	0,3289	0,3104
Gini normalisé (base de test)	0,2965	0,2976	0,3056
Gain p/r au zonier d'AXA (sur test)	-	0,11,88%	0,91%

TABLE 14 – Performances des différents zoniers sur uniquement les observations bien géocodées de la base de test

b) Pour les modèles de fréquence vol :
Grâce à son niveau de détail granulaire, le micro-zonier fréquence vol se révèle plus performant sur tous les cas de figures (obervations globales comme observations bien géocodées). Le modèle avec le micro-zonier cible les risques de manière plus précise et offre une segmentation du risque géographique réel plus homogène à l'intérieur des zones. Ce qui témoigne de l'importance d'étudier les risques au niveau local, surtout pour le cas de l'assurance habitation.

	Zonier actuel d'AXA	Notre Micro-zonier
Gini normalisé (base d'apprentissage)	0,4208	0,3953
Gini normalisé (base de test)	0,3611	0,3669
Gain p/r au zonier d'AXA (sur test)	-	0,58%

TABLE 15 – Performances des différents zoniers sur toutes les observations de la base de test

	Zonier actuel d'AXA	Notre Micro-zonier
Gini normalisé (base d'apprentissage)	0,3707	0,3961
Gini normalisé (base de test)	0,3293	0,3399
Gain p/r au zonier d'AXA (sur test)	-	1,07%

TABLE 16 – Performances des différents zoniers sur uniquement les observations bien géocodées de la base de test

Les performances du micro-zonier sont beaucoup meilleures pour la fréquence vol que pour le coût moyen dégât des eaux. **La construction de micro-zonier semble donc mieux adaptée pour les modèles de fréquence.**

3.4 Apports de ce travail

Une longue période de recherche a abouti au travail présenté dans ce mémoire. Les zoniers classiques, avec ajout de données externes au niveau commune, sont souvent réalisés avec le logiciel R car ne nécessitant pas beaucoup de manipulations géospatiales. Dans notre cas, le

besoin d'exploration à une maille beaucoup fine que les découpages administratifs habituels et l'ajout d'une quantité importante de données externes nous ont conduits à rentrer de fond en comble dans le monde de l'analyse spatiale. Cela nous a permis de réaliser notre propre découpage de la France. Les apports sont :
-la réalisation de zonier à une maille aussi fine qu'on le souhaite ;
-la possibilité de séparer nos données en apprentissage, validation et test puis d'appliquer des modèles de machine learning. Le micro-zonier est construit sur la base d'apprentissage puis évalué sur le test par jointure spatiale entre le micro-zonier et la base test. Cette jointure spatiale est réalisée avec les techniques d'analyse spatiale. On n'a plus besoin de construire le zonier sur l'ensemble de nos données. Ce qui permet de diminuer l'overfiting.
-la vérification de la précision des cartes. Beaucoup des cartes gratuites disponibles sur internet ne sont précises. Pour m'en rendre compte j'ai superposé plusieurs cartes sur la vraie carte de la France disponible sous Qgis pour ensuite zoomer au niveau des frontières. L'utilisation d'une carte peu précise entraine des biais sur la précision du zonier. La vérification de la carte est donc une étape importante pour garantir la fiabilité des résultats issus du zonier. Au niveau de la figure 36 nous avons superposé 3 cartes : en bleu foncé, une carte des frontières de la France utilisée par Direct Assurance, en rouge une carte des départements de la France utilisée par AXA France et en vert la vraie carte satellite de la France utilisée par le logiciel de SIG Qgis. Les points rouges représentent les domiciles des assurés. Au niveau de cette carte, nous avons effectué un zoom au niveau des frontières de la France. La surface en bleu clair représente ainsi l'océan. On remarque que seule la carte exacte de la France disponible sous Qgis contient tous les domiciles des assurés. L'utilisation des deux autres cartes provenant des sites internet entraine une perte d'information car ne couvrant pas les vraies frontières de la France de manière parfaite.

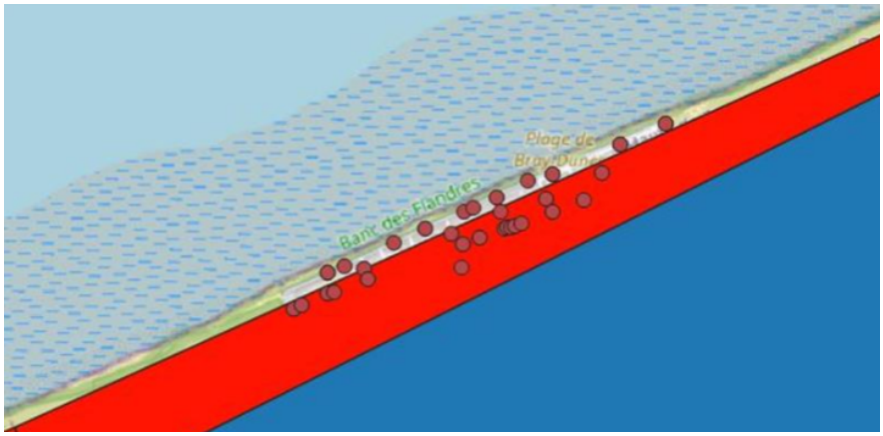


FIGURE 36 – Comparaison des cartes de la France disponibles sur internet

3.5 Limites et améliorations possibles

Bien que les apports soient conséquents, ce travail constitue une des premières versions de micro-zoniers réalisées en assurance. Les points d'amélioration à apporter peuvent se résumer comme suit :

-géocoder davantage les domiciles des assurés ;
-enrichir les données en explorant davantage les sources open data,
-proposer un zonier hybride qui utiliserait le micro zonier dans les zones avec beaucoup de sinistres et rester à la maille INSEE lorsque le nombre de sinistre est faible pour limiter tout problème d'overfitting que le micro zonier peut engendrer ;
-essayer éventuellement d'autres techniques de découpage géographique ;
-se procurer des cartes officielles satellites.

4 Conclusion

Notre stage de fin d'études s'est déroulé à la Direction Technique d'AXA Direct Assurance et s'inscrit dans le cadre de la création de micro-zoniers en assurance multirisques habitation. Plus précisément, nous nous sommes proposés dans ce travail de construire un micro-zonier sur le coût moyen de la garantie dégât des eaux et un autre sur la fréquence vol. La première étape du travail fut la préparation des bases de données, une manière pour nous d'explorer les données de natures et de sources différentes, et de vérifier leur cohérence. Ensuite, nous avons modélisé le signal géographique. Pour ce faire, nous avons effectué une première modélisation du coût moyen dégât des eaux et de la fréquence vol avec le modèle linéaire généralisé et les variables classiques de tarification, ceci sur la base d'apprentissage. Les résidus issus de cette première étape ont été modélisés avec les variables externes géographiques à l'aide d'un modèle de Machine Learning. La prédiction issue de ces modèles représente le signal géographique. Le signal géographique sera découpé en plusieurs classes, la représentation de ces différentes classes de signal géographique sur une carte représente le micro-zonier final. L'intégration des micro-zoniers dans les modèles de tarification a permis d'améliorer les performances. Grâce à sa maille fine, nos micro-zoniers ciblent les risques de manière plus précise et offrent une meilleure segmentation du risque géographique.

Bibliographie

[1991] Franz AURENHAMMER . Voronoi diagrams : a survey of a fundamental geometric data structure. ACM Computing Surveys (CSUR) 23.3, p. 345–405.

[2011] Yann MILOE. Tarification d'un produit MRH à l'aide de la méthode des marges. Mémoire Axéria iard.

[2012] Arnaud DONGUY. Contribution de l'information géographique aux métiers de l'assurance pour la gestion des événements d'ampleur. MINES ParisTech Centre de recherche sur les Risques et les Crises.

[2015] Clément FESQUET. Utilisation de facteurs exogènes pour les zoniers en tarfication MRH. Mémoire Actuaris.

[2017] Papa Djibril BA . FORMATION EN SYSTEME D'INFORMATION GEOGRAPHIQUE. Support de cours ENSAE Dakar, 2017

[2017] Jennifer PARIENTE. Modélisation du risque géographique en assurance habitation. Mémoire AXA France.

[2018] MARLIER Aurélia. Zonage d'un risque à évènement rares : l'inondation. Mémoire GMF.

[2020] Nassim DENNOUNI et Mhamed HADJ HENNI. Les Systèmes d'information géographique et l'aide à la décision. Faculty of exact sciences and Informatics, Hassiba Benbouali University of Chlef, Algeria

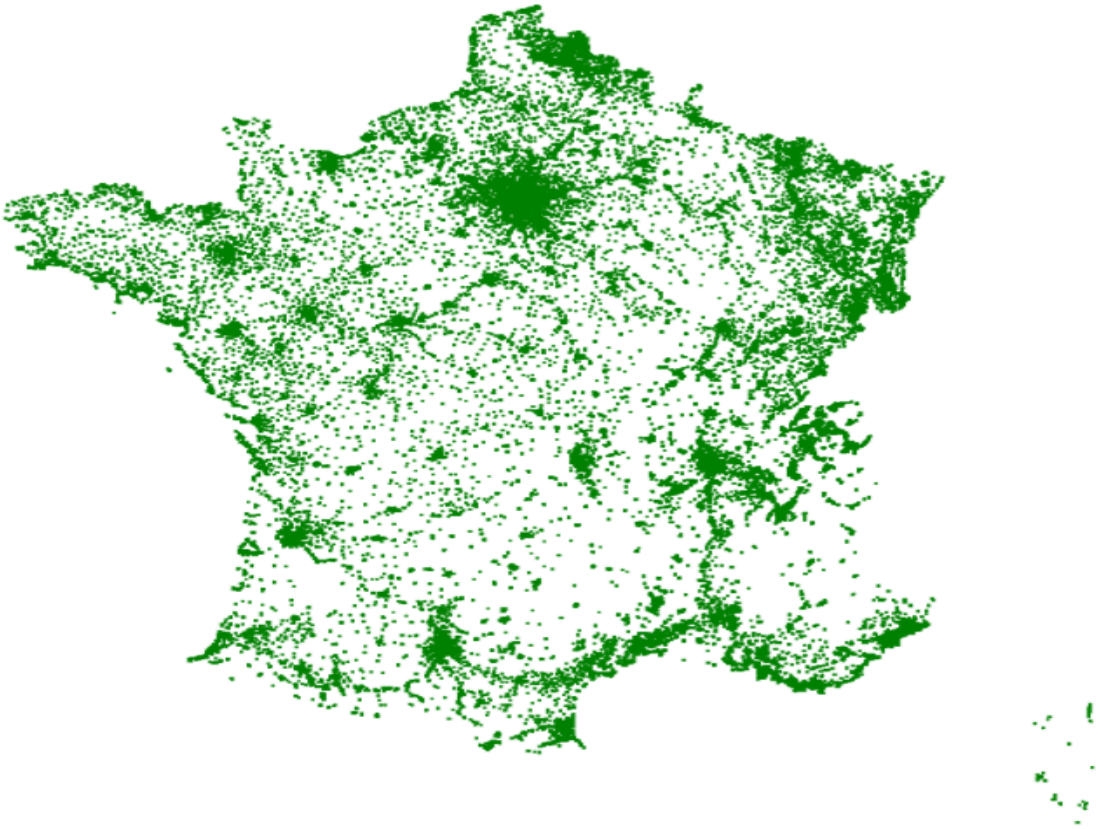
[2020-2021] Christophe Dutang. Actuariat de l'Assurance Non-Vie. Support de cours Ensae Paris.

Documentation ARCGIS. URL : https ://desktop.arcgis.com/fr/arcmap/10.3/guide-books/map-projections/about-geographic-coordinate-systems.htm

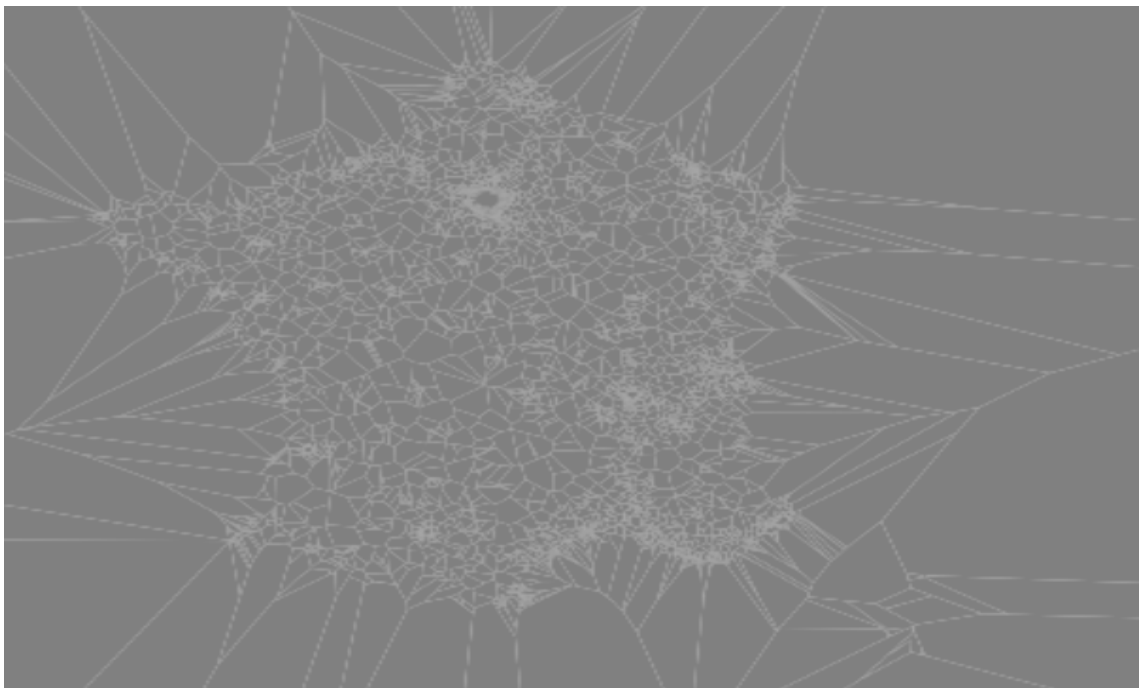
Catboost model. URL : https ://ichi.pro/fr/gradient-boosting-lightgbm-xgboost-et-catboost-kaggle-challenge-santander-268070598043800.

Annexes

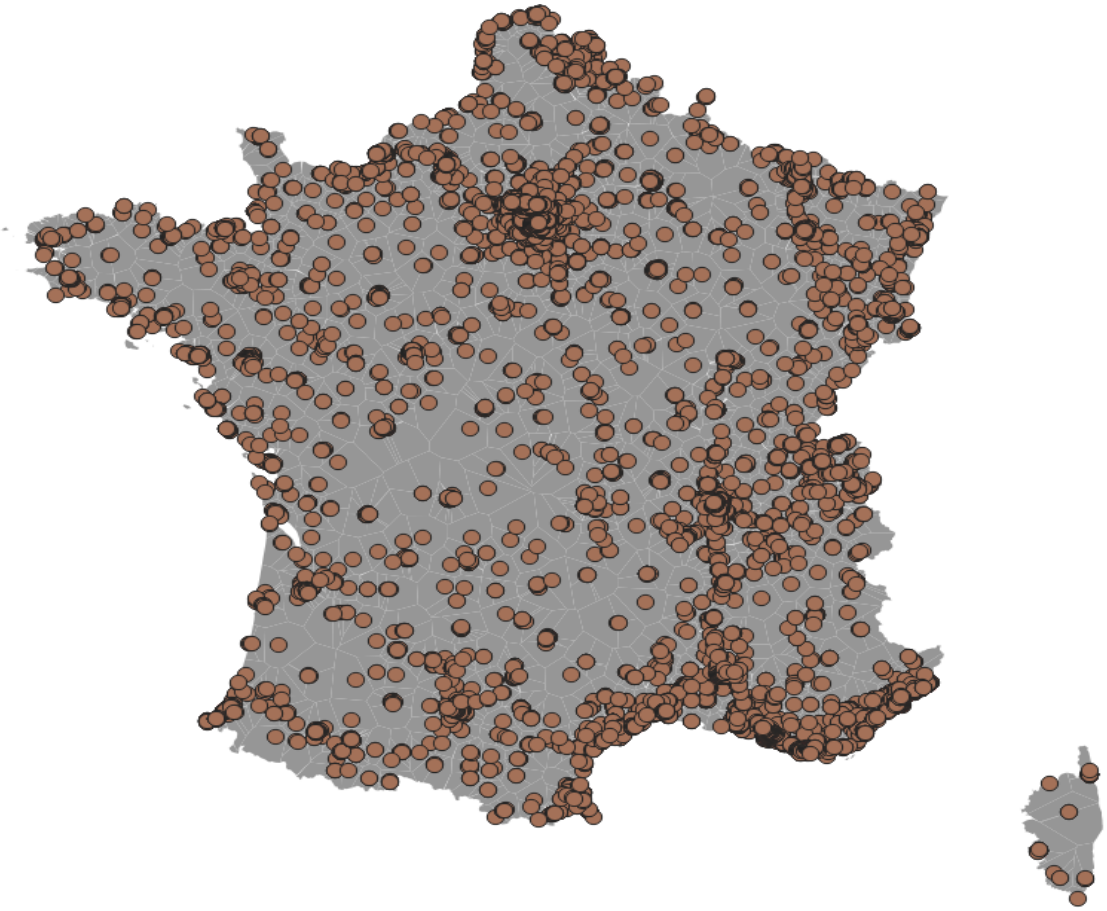
Annexe 1 : Répartition des contrats dégât des eaux appartements



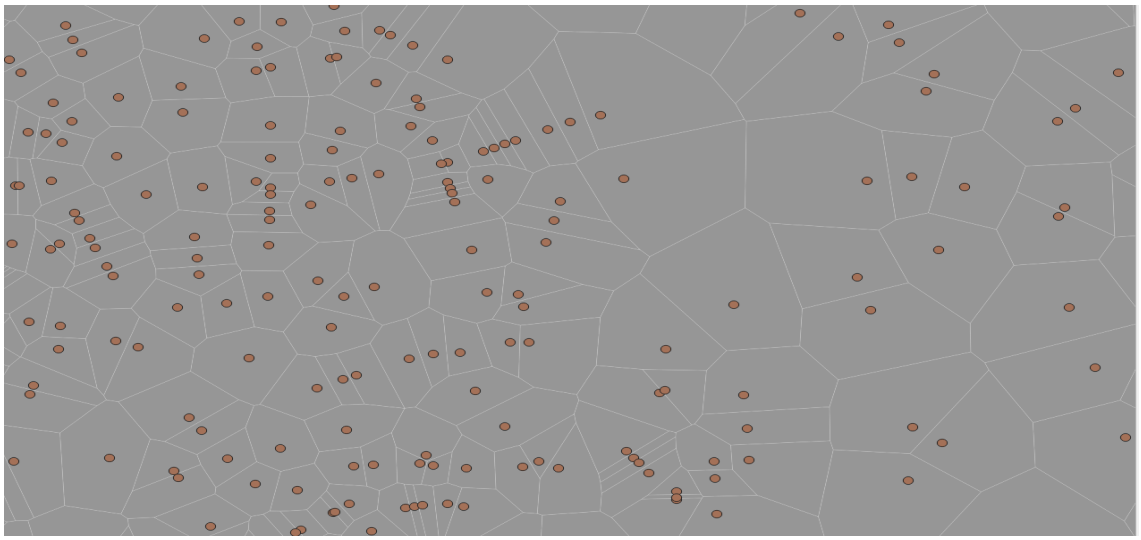
Annexe 2 : Notre découpage brut de la France par Voronoi pour le coût moyen



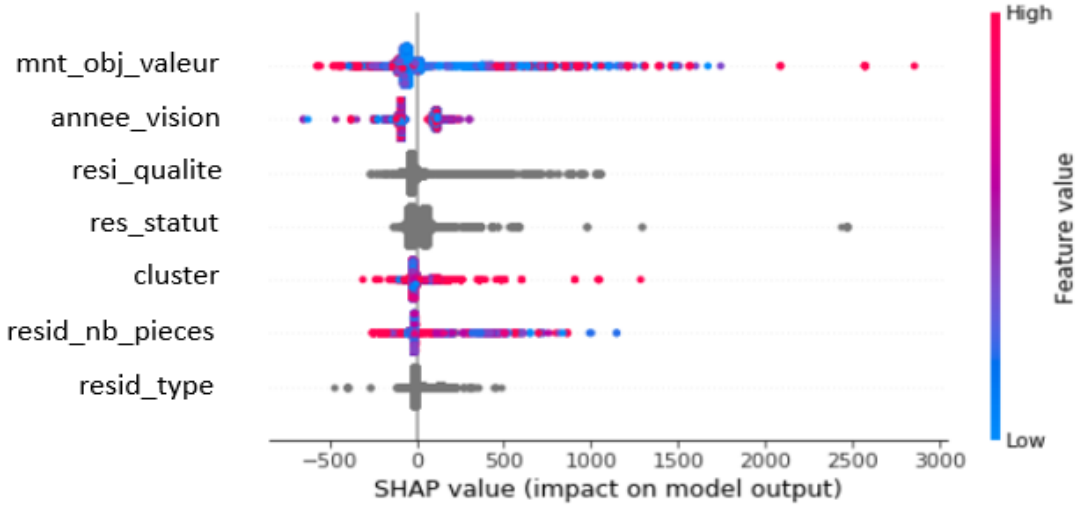
Annexe 3 : Voronoi + germes



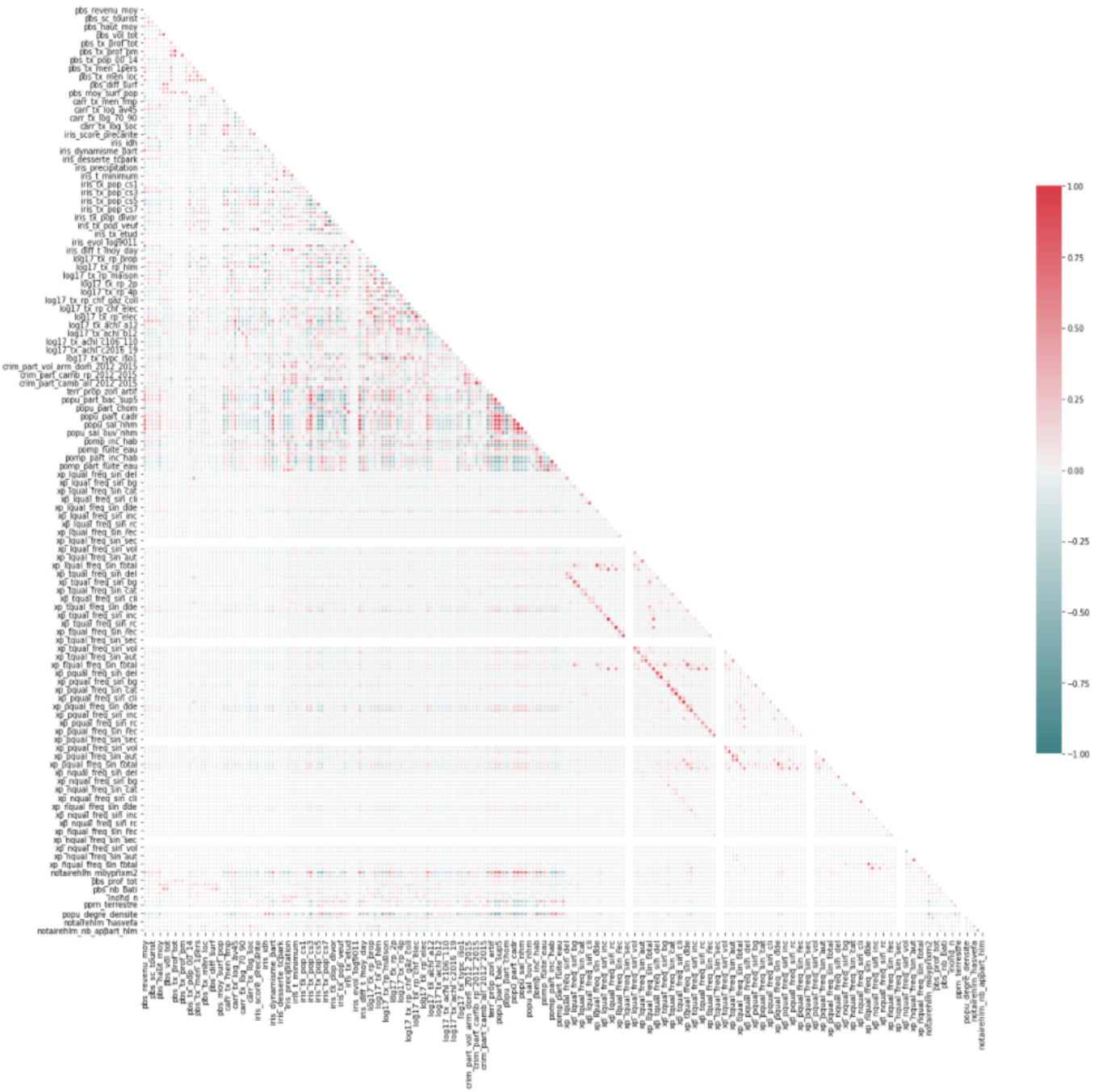
Annexe 4 : Zoom



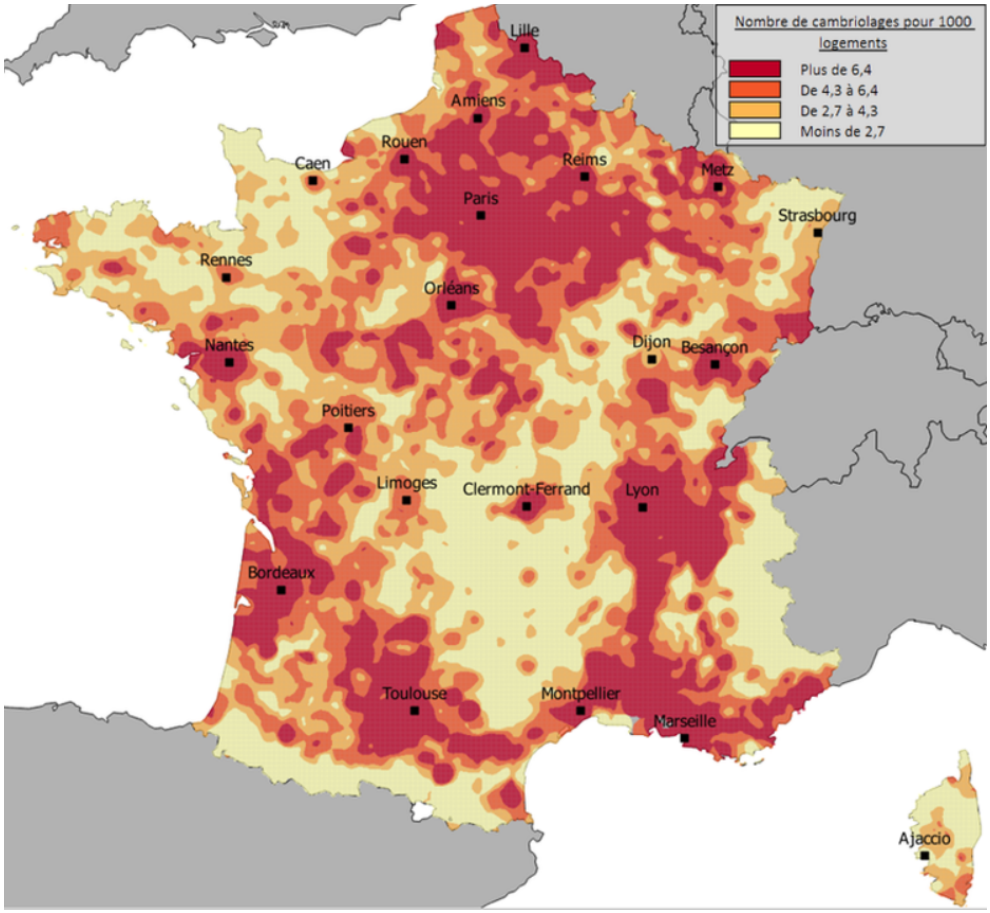
Annexe 5 : Résultats du catboost du modèle de coût moyen dégât des eaux avec le micro-zonier représenté par la variable "cluster"



Annexe 6 : Matrice de corrélation des variables explicatives externes



Annexe 7 : Répartition des cambriolages en France en 2020



Source : Ministère de l'intérieur

Annexe 8 : Résultats du catboost du modèle de fréquence vol avec le micro-zonier représenté par la variable "cluster"