Master Actuariat de Dauphine

Mémoire présenté devant l'Université Paris Dauphine pour l'obtention du diplôme du Master Actuariat et l'admission à l'Institut des Actuaires

le 18 Janvier 2017

Par : Jennifer PARIENTE

Titre: Modélisation du risque géographique en assurance habitation

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

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

Membre présent du jury de l'Institut des Actuaires :

Florence PICARD

Fabien CHAILLOT

Signature :

Entreprise :

Nom : AXA France

Signature :

Directeur de mémoire en entreprise :

Nom : Doan NGUYEN TUAN

Signature :

Membres présents du jury du Master Actuariat de Dauphine :

Pierre CARDALIAGUET

Marc HOFFMANN

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

Signature du responsable entreprise :

Secrétariat :

Bibliothèque :

Signature du candidat :

Université Paris-Dauphine, Place du Maréchal de Lattre de Tassigny, 75775 PARIS Cedex 16

Résumé

Dans un contexte concurrentiel comme celui de l'assurance Multirisque Habitation, les assureurs doivent continuellement améliorer leur segmentation tarifaire afin d'adapter au mieux leur prime au risque assuré. L'aspect géographique du risque est très important en assurance habitation puisqu'il permet de prendre en compte les données relatives à l'environnement où évolue le contrat, ce qui constitue un gain d'information important notamment pour la garantie VOL. Un zonier peut alors être défini comme la prise en compte de ce signal géographique à l'intérieur du modèle, ce qui permet de segmenter au mieux les différents profils de risque pour créer des classes de risques homogènes en sinistralité.

Ce mémoire se décompose principalement en deux parties : la première partie permet d'analyser l'apport des données externes et en particulier l'impact lorsque ces données sont fournies à une maille très fine telle que les coordonnées GPS du contrat habitation. Quant à la deuxième partie, elle met en œuvre un processus de construction de micro-zonier à l'aide de ces nouvelles données externes et en ayant recours à des méthodes innovantes. L'objectif consiste ainsi à construire un zonier stable qui puisse garantir une information à la fois précise et robuste.

Mots-clés

Multirisque Habitation, Garantie VOL, Modèles Linéaires Généralisés, Gradient Tree Boosting, Risque géographique, Micro-zonier, Variables externes, Lissage spatial par interpolation spatiale, Géocodage

Abstract

In a competitive market such as that of household insurance, insurers constantly have to improve their pricing algorithm in order to optimally segment their risks and thus create a personalized price adapted to every risk profile. The geographic variables linked to a contract are very important in household insurance because a property's location is an extremely important factor in evaluating the underlying risk – this is particularly true for theft cover. Zoning can thus be defined as the combination of methods used to take a risk's location into account within a pricing model – i.e. using geographical variables to create classes of risks which can then be used as rating variables and which consequently allow insurers to better segment their risks.

This study is divided into two main parts : the first part involves analyzing an external dataset and especially the one that provides data about a risk which is accurate to a very small scale : those of GPS coordinates. The second part describes the process of building a micro-zoning structure using innovative methods applied to this new external data. The objective is thus to build a stable zoning process that provides information that is both accurate and robust.

Keywords

Household Insurance, Theft insurance, Generalized Linear Models, Gradient Tree Boosting, Geographic risk, Micro zoning, External data, Spatial smoothing with spatial interpolation, Geocoding

Synthèse

L'objectif de ce mémoire est de tester des **méthodes innovantes** dans la construction d'un micro-zonier. Le processus concerne la garantie VOL en Multirisque Habitation pour les appartements et traite un modèle de fréquence qui permet de prédire le nombre de sinistres annuel.

On définit le **zonier** comme le traitement du **signal géographique** lors de la création du modèle de tarification. Ce traitement est clé car la segmentation du risque habitation est portée pour beaucoup par la dimension géographique. L'adresse du contrat fournit alors une passerelle pour accéder à l'ensemble des données externes relatives à la localisation du contrat. On parle de **micro-zonier** lorsque l'étude de ce signal se fait à une maille très fine qui se détache des découpages administratifs usuels.

L'année passée, l'équipe Multirisque Habitation d'AXA France a reconstruit tous les zoniers de manière à capter le signal géographique de manière plus locale en utilisant un maillage très fin du territoire à partir de cellules de Voronoï. Cependant, du fait de la finesse que fournissait ce micro-zonier, des cas de sur-apprentissage sont apparus amenant de l'instabilité sur certains modèles et notamment sur le modèle de fréquence VOL appartement. C'est pourquoi ce mémoire établit un **processus alternatif** de création de micro-zonier qui résout les problèmes de stabilité survenus lors de la création du modèle précédent sur la fréquence de survenance d'un sinistre VOL en appartement, tout en garantissant un haut niveau de discrimination du risque.

L'étude réalisée dans ce mémoire a soulevé des questions concernant l'apport de la **donnée externe** et en particulier l'impact lorsque cette donnée est fournie à une maille très fine telle que l'adresse. Les données externes utilisées dans l'étude sont toutes des données géographiques relatives à l'environnement du contrat. L'analyse des données externes constitue une part non négligeable du mémoire et a été réalisée sur la garantie principale en assurance Multirisque Habitation : la garantie Dégât des Eaux.

Les données externes disponibles

Les données externes analysées sont issues de différentes sources et concernent plusieurs mailles géographiques, nous étudions en particulier :

- Des données départementales provenant de l'Observatoire National de la Délinquance et des Réponses Pénales,

- Des données à la maille INSEE concernant la période de construction des résidences principales,
- Des données à la maille IRIS concernant des variables socio-démographiques ou socio-économiques,
- Des données à la maille ADRESSE contenant notamment des typologies d'adresse et des classifications urbaines ainsi que des données de type points d'intérêt.

À partir de ces données externes, d'autres variables ont été créées telles que l'écart de revenu moyen avec les voisins ou encore la distance au point d'intérêt le plus proche comme une station de police, une clinique ou encore un lieu surveillé.

L'analyse des données externes géographiques permet de répondre à plusieurs questions importantes dans l'environnement actuariel actuel. En effet, l'étude montre l'importance des variables externes dans la tarification, car lorsqu'elles sont ajoutées aux variables tarifaires classiques, elles permettent d'obtenir un gain d'information important ce qui amène à une amélioration de la qualité des modèles. Par ailleurs, l'étude révèle que l'utilisation des seules variables externes en tant que variables tarifaires permettrait d'obtenir un modèle relativement aussi performant qu'un autre modèle qui serait constitué uniquement de variables récoltées à la souscription relatives au risque et à l'assuré lui-même. La **simplification du questionnaire client** est alors envisageable car en demandant uniquement l'adresse du contrat pour définir un premier devis rapide (*Quick quote*), l'ensemble des données externes relatives à la situation géographique du contrat est fourni, sous réserve que le géocodage de l'adresse soit réalisé avec une bonne précision.

L'utilisation des données externes permet également de substituer dans la tarification les variables récoltées à la souscription relatives aux montants assurés - souvent sous-estimés par le client - par une information possiblement plus robuste. En effet, l'étude a mis en œuvre un modèle de prédiction qui était tout aussi performant avec ou sans la présence des variables relatives aux montants assurés.

Quant à la problématique concernant l'apport des données à une maille aussi fine que celle de l'adresse, l'étude sur la base de données constituée et sur les données externes disponibles n'a pas été concluante puisque les modèles sont relativement aussi performants avec une information à l'adresse qu'avec la même information agrégée à une maille plus grossière mais qui reste tout de même fine.

Le signal géographique

Ces données externes permettent d'expliquer le signal géographique, il faut donc les exploiter de la meilleure manière possible. Nous disposons d'une quantité très importante de variables pour analyser notre risque et la **sélection de variables** est donc une étape importante. Afin de réaliser un premier tri parmi ces données, des modèles de **Gradient Tree Boosting** sont réalisés sur chacun des ensembles de variables afin de

détecter les variables possédant une quelconque influence sur la variable réponse. De la même manière, pour pré-sélectionner les variables, un tri par Gradient Tree Boosting est réalisé sur les variables « internes », c'est à dire les variables récoltées à la souscription relatives au risque et à l'assuré lui-même.

D'après les résultats d'une analyse comparative entre le modèle linéaire généralisé log-poissonien et le modèle de Gradient Tree Boosting, les performances des deux modèles restent relativement équivalentes, mais étant donné la quantité importante de variables dans l'étude, l'utilisation d'un Gradient Tree Boosting a semblé plus appropriée car il dispose d'une sélection automatique des variables tout en prenant en compte les possibles interactions existantes. Cette sélection automatique des variables réalisée par l'algorithme du Gradient Tree Boosting a été confrontée à d'autres méthodes de sélection dans l'étude ce qui a permis d'attester de sa qualité.

Les variables internes et externes conservées après les différentes étapes de sélection sont incorporées directement dans le modèle de Gradient Tree Boosting afin de capter une première partie du signal géographique. Ce premier modèle réalisé fournit des résultats plutôt encourageants puisque qu'il est d'ores et déjà plus stable et plus performant que ne l'était le précédent modèle linéaire généralisé utilisant les cellules de Voronoï.

De manière à réaliser l'étude du signal géographique plus en profondeur, les résidus issus du modèle - calculés comme étant la différence entre la fréquence observée et la fréquence prédite - sont analysés de manière à déterminer s'il demeure une part de ce signal présent. L'utilisation d'un **semivariogramme** a permis de confirmer la présence d'auto-corrélation spatiale entre les résidus, ce qui laisse sous-entendre que le signal géographique n'a pas été intégralement capté par le modèle de prédiction. De manière à capturer la part restante du signal géographique dans les résidus, un lissage spatial par interpolation spatiale est mis en œuvre afin de créer une nouvelle variable géographique. Plusieurs méthodes d'interpolation spatiale sont utilisées, notamment la méthode de **pondération par l'inverse à la distance** et la méthode géostatistique du **krigeage**. Ces deux méthodes considèrent que des points voisins se ressemblent, c'est pourquoi les observations les plus proches possèdent des poids plus élevés que les observations éloignées. La technique de pondération par l'inverse à la distance est relativement simple puisque les poids sont déterminés uniquement à partir de la distance entre le point considéré et les autres points, tandis que la méthode du krigeage, qui utilise un semivariogramme pour déterminer la valeur de ces poids, prend en compte la répartition spatiale de l'ensemble des observations (contrats). La méthode géostatistique du krigeage, traditionnellement utilisée dans les sciences de l'environnement, est ici employée à des fins de tarification.

Le schéma récapitulatif ci-dessous permet de visualiser les étapes réalisées dans cette étude afin de construire le processus d'intégration du signal géographique dans le modèle de prédiction.

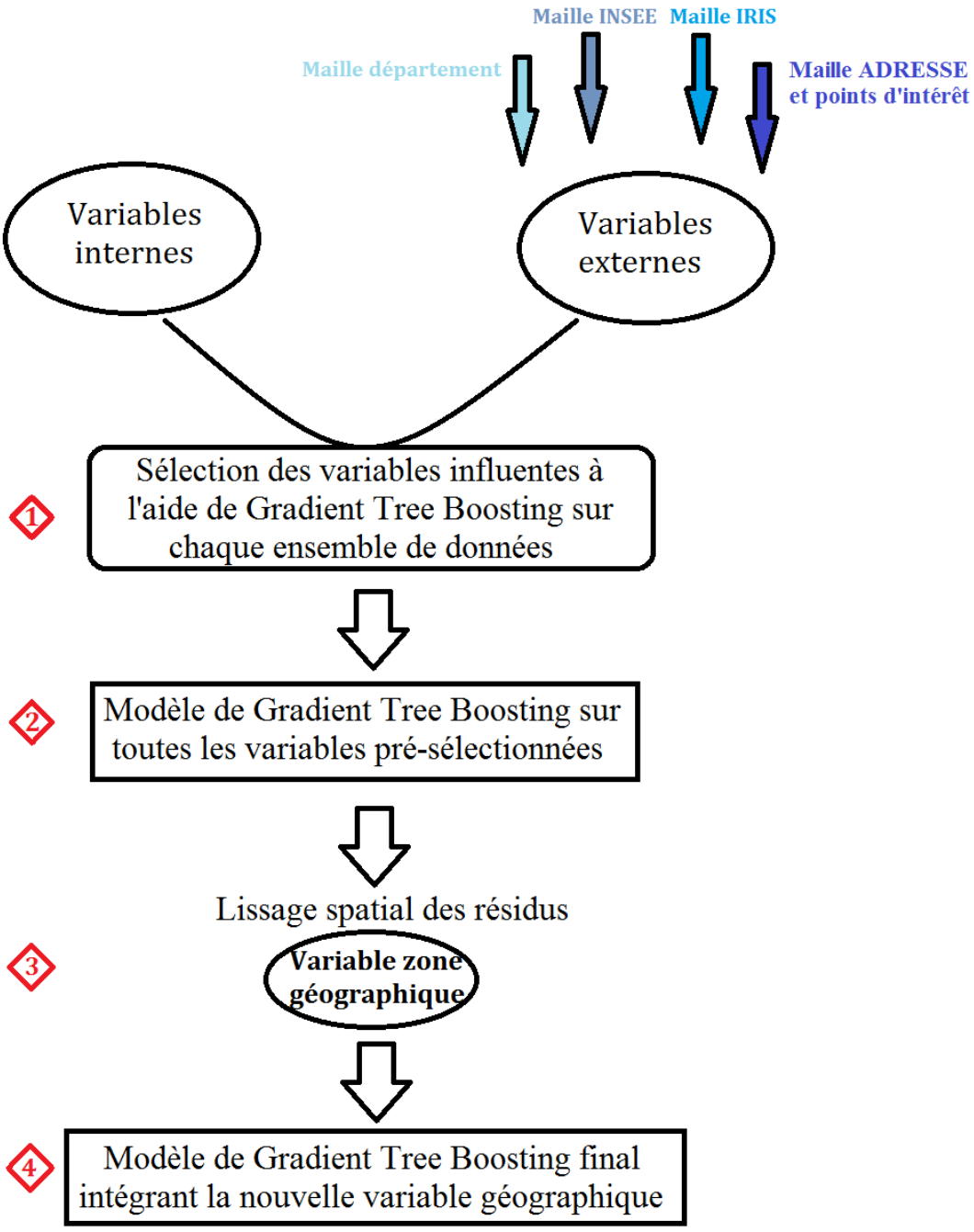

FIGURE 1 – *Étapes de construction de la structure de micro-zonier dans cette étude.*

La méthode éditée dans cette étude s'éloigne donc largement de celle utilisée l'année dernière, au sens où le modèle de prédiction ayant servi dans cette étude est un modèle de Gradient Tree Boosting et non un modèle linéaire généralisé. Les deux méthodes diffèrent également par la manière d'aborder les données géographiques. En effet, le traitement du signal géographique au sein du modèle créé se fait en plusieurs étapes (figure 1). La première partie du traitement consiste à pré-sélectionner les variables externes influentes pour le signal géographique (au même titre que les variables internes) et à les intégrer directement dans le modèle (cf. étapes 1 et 2 du schéma). En-

suite, une étape de lissage spatial des résidus permet de créer une nouvelle variable géographique qui est réintégrée au modèle (cf. étapes 3 et 4 du schéma). Concernant la méthode réalisée l'année passée, le modèle linéaire généralisé intégrait uniquement des variables « internes » relatives au risque et à l'assuré lui même et l'étude du signal géographique se faisait uniquement par l'intermédiaire de forêts aléatoires *(Random Forest)* appliquées aux résidus du modèle. Les variables externes étaient alors utilisées dans les forêts aléatoires pour prédire les résidus.

Les résultats d'une étude comparative entre les deux méthodes ont révélés que le nouveau modèle fournit une meilleure performance que l'ancien modèle tout en garantissant une certaine stabilité.

Synthesis

The purpose of this study is to test **innovative methods** in the construction of a micro-zoning structure. The process is applied to a frequency model for theft cover in household insurance.

Zoning is defined here as the set of processes used to include geographical signal within a rating structure. The geographical aspect refers to all external data related to the location of the contract through its address. We talk about **micro-zoning** when the variables used are accurate to a very small scale – that is one that is smaller than standard administrative categories such as postcodes.

Last year, AXA France household team's zoning structure was revised in order to capture the geographical signal more locally by using Voronoï polygons to create a very fine geographical grid. However, despite the greater degree of accuracy brought about by this micro-zoning structure, cases of over-fitting were observed and caused instability in some models - especially the frequency model for apartment theft. This is why the objective of this thesis is thus to establish an alternative process for building a micro-zoning structure which can remediate the stability problems which were encountered with AXA's previous frequency model for apartment theft while continuing to ensure a high level of risk discrimination.

The study in this paper also discusses issues concerning the utility of **external data** on a very small geographical scale : that of GPS coordinates. External data used in the study consists entirely of geographical data which relates to the contract's location and immediate surroundings. The analysis of external data constitutes a significant part of this thesis and was performed for Water Damage cover, which is the most widespread and important warranty in household insurance.

Available external data

The analyzed external data comes from different sources and is accurate to a differing range of geographical scales which depend on the source. It includes :

- Département (county) related crime data from the Observatoire National de la Délinquance et des Réponses Pénales,
- INSEE (Postcode) data concerning the construction period of buildings,

- IRIS related data containing socioeconomic and sociodemographic information,
- Data which depends on the risk's exact GPS coordinates (such as the social category of inhabitants in a particular street and information about neighboring points of interest).

Feature engineering is also carried out in order to create new variables such as the income gap with neighbors, or distances to various points of interest such as police stations, clinics or the nearest place with surveillance cameras.

The analysis of the external geographical data helps answer several important issues in the current actuarial environment. Indeed, this study illustrates the importance of external variables in pricing, because when added to conventional tariff variables, they provide a significant gain in information and consequently an improvement in the model's quality. Furthermore, we also show that using only external variables as tariff variables can produce a model which is almost as efficient as a model that would consist only of variables related to the customer or the risk itself. Practically speaking, this result could be used to **simplify the underwriting process**. It would, for instance, be possible to ask a customer for his or her address and, providing that address can be geocoded accurately, provide an approximate quote using external variables only.

Using external data can also be used to remove existing internal variables with low data quality from the model. For example, the customer's insured amount is often underestimated or declared incorrectly during the quotation process. We implemented a predictive model that was equally effective with or without variables relating to insured amounts.

On the other hand, this study does not allow us to reach a conclusion as to the practical utility of using data accurate to GPS coordinate level. When we aggregated the GPS coordinate data to a slightly larger (yet still very small) grid and used this in our models we obtained a model which seemed to be almost as predictive as one which used the actual GPS data itself.

The geographical signal

This external data helps explain the geographical signal and it is necessary to use it intelligently. As we have a very large choice of variables we can use to try and predict our risk, **variable selection** is therefore an important step. In order to achieve a first selection amongst all available variables, **Gradient Boosting Tree** models are generated on each of the sets of variables to detect variables which have a particularly strong influence on the response variable. Similarly, to sort variables, a Gradient Tree Boosting is also generated from « internal » variables, i.e. other, non-geographic, variables relating to the customer or the risk itself.

According to the results of a comparative analysis between the generalized linear model and Gradient Boosting Tree, the performance of both models remain relatively

equivalent but given the large amount of variables in the study, using a Gradient Boosting Tree seems more appropriate since it automatically selects variables and takes into account pre-existing interactions when doing so. The process of using Gradient Boosting to automatically select variables is also called into question in this study. Our analysis allows us to conclude positively as to the legitimacy of using this method for variable selection.

The internal and external variables which were selected through the processes described above are then incorporated directly into the Gradient Boosting Tree model to capture a first part of the geographical signal. The first Gradient Boosting model provides quite encouraging results as it is more stable and more predictive than the previous model which was based on Voronoï polygons.

In order to delve deeper into the study of the geographic signal, the residuals from this first model - defined as the difference between the observed frequency and the predicted frequency - are analyzed to determine if part of this geographic signal is still contained in the residuals. A **semivariogram** confirms the presence of spatial autocorrelation between the residuals, suggesting that the geographical signal has not been fully captured by the prediction model. In order to capture the remaining part of the geographical signal in residuals, spatial smoothing by spatial interpolation is implemented in order to create a new geographical variable. Several methods of spatial interpolation are used to do this, including the **inverse distance weighting** method and a geostatistical method known as **kriging**. Both methods consider that neighboring points are alike so that closer observations have higher weights than remoter observations. The technique of inverse distance weighting is relatively simple since the weights are determined solely from the distance between the considered point and the others, while the method of kriging, which uses a semivariogram to determine the value of those weights, takes into account the spatial distribution of all observations (here, one observation is equivalent to one contract). The geostatistical method of kriging, which is traditionally used in environmental sciences, is thus adapted here for pricing purposes.

The diagram below seeks to visually summarize the steps used in this study to integrate the geographical signal in the prediction model.

The new method differs from the one used in last year's model in that the prediction model used in this study is a Gradient Boosting Tree model and not a Generalized Linear Model. Both methods also differ in their approach to geographical data. Indeed, the treatment of geographical signal within the model created is done in several steps (Figure 2). The first part of this process involves pre-selecting the most influential internal and external variables and integrating them directly in the model (steps 1 and 2). Steps 3 and 4 then involve applying spatial smoothing to the residuals to create a new geographical variable that could then be reinjected into the model. Last year's method, on the other hand, simply consisted in fitting a generalized linear model which incorporated only « internal » variables relating to the customer or the risk itself and then

integrating the geographic signal by predicting the residuals of the first GLM with a random forest on external variables.

A comparative study of the two methods allows us to conclude that the new model is both more stable and more predictive than the previous one.

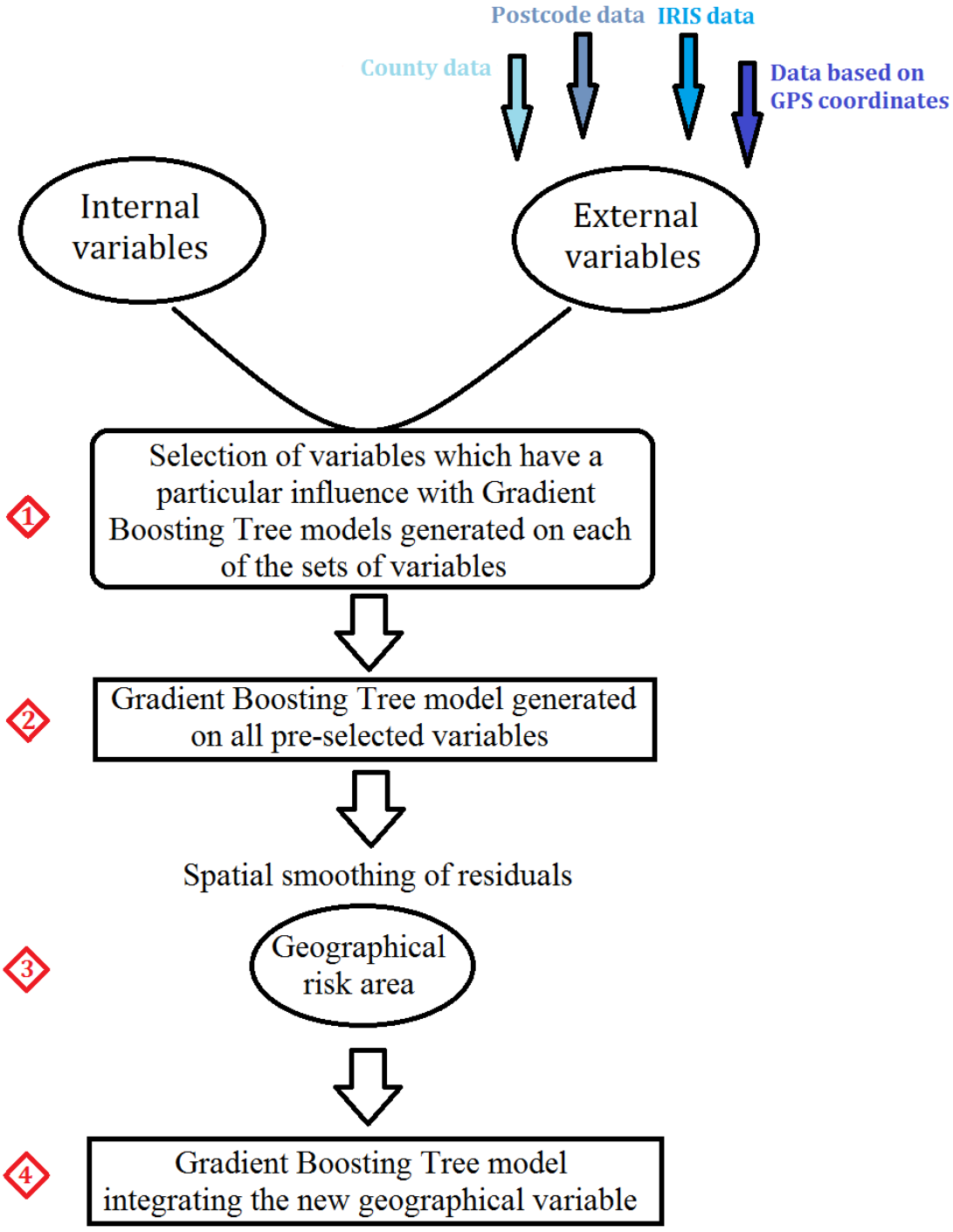

FIGURE *2 – Steps required for building the micro-zoning structure in this study.*

Remerciements

Je tiens tout d'abord à adresser mes remerciements à Guillaume GORGE, à l'ensemble de la Direction du Marché IARD de la branche Particuliers/Professionnels d'AXA France et notamment aux équipes « Innovation Pricing » et « Actuariat Non-Auto » pour leur accueil et leur sympathie tout au long de mon stage.

J'adresse notamment ma reconnaissance à Anne Laure LE GALLO et à Fleur LECARPENTIER, respectivement responsables des équipes « Actuariat Non-Auto » et « Innovation Pricing » pour m'avoir fait confiance sur ce sujet d'étude que j'ai beaucoup apprécié.

Je tiens à remercier tout particulièrement ma tutrice de stage, Doan Trang NGUYEN TUAN, pour son soutien, sa bonne humeur et son optimisme qui m'ont permis d'avancer durant la rédaction de ce mémoire. Elle a su répondre à mes interrogations et ses conseils m'ont permis d'alimenter ma réflexion pour construire le cheminement de mon mémoire.

Mes remerciements à Amaury RAULT, de l'équipe Auto d'AXA France et Selim RABOUDI, de l'équipe Data Science d'AXA Global P&C pour le temps qu'ils ont accordé à répondre à mes interrogations techniques et les conseils qu'ils m'ont fournis.

Je souhaite, de plus, exprimer mes meilleurs sentiments aux autres stagiaires et alternants.

Enfin, je remercie l'équipe pédagogique de Dauphine et en particulier Monsieur Marc HOFFMANN, qui m'a suivie jusqu'à la fin de ce mémoire. Et merci également à mes proches pour leur soutien et leurs encouragements à chaque étape de mon parcours académique et de mon mémoire.

Table des matières

Introduction

Un contrat d'assurance habitation couvre plusieurs risques tels que le dégât des eaux, le vol ou encore le bris des glaces. La tarification d'un tel contrat repose sur différentes variables, notamment des variables relatives au logement comme le nombre de pièces, des variables relatives à l'assuré lui même, mais il existe également des variables qui traduisent l'environnement géographique où évolue le contrat. En effet, l'aspect géographique du risque est très important en assurance habitation et il est primordial de le prendre en compte dans le calcul du tarif.

L'étude produite dans ce mémoire a été réalisée au sein de la Direction du Marché IARD de la branche Particuliers/Professionnels d'AXA France, dans l'équipe « Innovation Pricing » et le sujet s'applique à une problématique Multirisque Habitation. Cette équipe s'occupe d'identifier les besoins et de proposer des solutions techniques afin de permettre l'accélération de la sophistication tarifaire à l'aide de projets de recherche et de développement.

Suite aux travaux de refonte de la gamme Multirisque Habitation l'année dernière, l'équipe Multirisque Habitation d'AXA France a décidé d'améliorer la variable tarifaire qui constitue l'ensemble du signal géographique : le zonier. La précision du zonier réalisé était très fine et permettait de se détacher des découpages administratifs usuels (par département ou par code INSEE par exemple), cependant des problèmes de stabilité sont survenus. Le présent mémoire a donc pour ambition de créer et de tester une méthode alternative d'intégration du risque géographique dans le modèle tout en analysant l'impact de l'utilisation de nouvelles variables externes fournies à des mailles géographiques fines.

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

Le premier chapitre permet de donner au lecteur une vision claire du contexte dans lequel est réalisé le mémoire en détaillant notamment les étapes de réalisation du traitement du signal géographique et en comparant la méthode éditée avec celle utilisée précédemment par l'équipe Multirisque Habitation.

Le second chapitre traite de la création de la base de données utilisée pour réaliser cette étude en précisant les variables qui la composent, l'échelle à laquelle elles sont observées et leur provenance.

Dans le troisième chapitre, l'analyse des données externes est réalisée. L'accent est notamment marqué par l'étude de l'apport des données à la maille ADRESSE qui n'ont jamais été testées auparavant mais se penche aussi sur une problématique importante dans le contexte actuel : la simplification du questionnaire client. L'ensemble des analyses de la donnée externe est réalisé sur la garantie dégât des eaux. Ce chapitre permet également d'analyser et de pré-sélectionner les variables influentes sur le risque de survenance d'un sinistre VOL qui sont ensuite intégrées dans le modèle.

Quant au chapitre quatre, une comparaison entre le modèle linéaire généralisé log-poissonnien et le modèle de Gradient Tree Boosting y est réalisée. Le modèle linéaire généralisé est le modèle le plus utilisé en assurances dommages, mais le modèle de Gradient Tree Boosting représente une bonne alternative car il permet de faire face à la quantité massive de variables disponibles pour réaliser l'étude. C'est pourquoi le modèle de Gradient Tree Boosting est utilisé en tant que modèle de prédiction. Les variables internes et externes pré-sélectionnées sont alors directement intégrées dans le modèle qui vise à prédire la fréquence de survenance d'un sinistre VOL sur les appartements. Afin d'approfondir le traitement du signal géographique, une étude des résidus est réalisée pour déterminer s'il demeure du signal dans les résidus qui n'aurait pas été capté par les variables externes du modèle. Grâce à l'utilisation d'un semivariogramme, il est établit qu'il existe bien une auto-corrélation spatiale des résidus, ce qui justifie la dernière étape réalisée dans le dernier chapitre.

Le chapitre cinq représente la dernière étape du traitement du signal géographique, à savoir la création d'une nouvelle variable géographique à partir du lissage spatial des résidus à l'aide de techniques d'interpolation spatiale. La nouvelle variable ainsi créée est ensuite réintégrée au modèle.

Chapitre 1

Cadre et objectifs de l'étude

1.1 L'assurance Multirisque Habitation

1.1.1 Présentation du marché IARD

Dans le monde de l'assurance, l'abréviation **IARD** signifie « incendie, accidents et risques divers ». Il s'agit d'une famille d'assurances qui s'oppose à celle des assurances de personnes ou assurances VIE (santé, accident, décès...). Les assurances IARD couvrent les dommages et la protection des biens, par opposition aux assurances VIE qui protègent les personnes. Pour les particuliers, les contrats IARD reposent principalement sur l'assurance automobile et l'assurance habitation.

Un produit d'assurance Multirisque Habitation propose des garanties pour les évènements suivants :

- Le dégât des eaux
- Le vol et le vandalisme
- L'incendie
- Le bris des glaces
- Les catastrophes naturelles
- Les catastrophes technologiques
- Les évènements climatiques
- La responsabilité civile
- Les attentats et risques de terrorisme.

Ce sont ces risques que les équipes en charge de la tarification d'un contrat Multirisque Habitation doivent être en mesure d'anticiper afin de pouvoir offrir un tarif adéquat au profil de chaque client. Seules les garanties vol et dégât des eaux seront abordées dans ce mémoire. Pour plus de détails concernant ces garanties, le lecteur est invité à se référer à l'annexe A.

1.1.2 AXA France sur le marché Multirisque Habitation

La loi relative à la consommation, dite loi « Hamon », est entrée en vigueur le 1er Janvier 2015. Cette loi fixe de nouveaux droits pour les assurés, et de nouvelles obligations pour les assureurs. De nombreux domaines sont concernés par cette loi dont celui de l'assurance habitation, ce qui a renforcé l'environnement déjà concurrentiel dans lequel évoluaient les assureurs. Ces derniers doivent constamment se repositionner d'un point de vue tarifaire afin de stimuler leurs résultats et leur compétitivité car les risques et leurs différentes composantes explicatives évoluent continuellement.

Par ailleurs, il est également nécessaire de s'adapter au comportement des clients d'aujourd'hui qui sont de plus en plus présents sur Internet, les réseaux sociaux et qui comparent davantage les offres. Ainsi, les particuliers sont désormais très sensibles à la dimension prix et à la personnalisation des offres.

Avec un chiffre d'affaires de plus de 860 millions d'euros à fin 2015, AXA France possède environ 12% des parts de marché en Multirisque Habitation. La segmentation des risques demeure donc primordiale pour AXA ainsi que pour tous les assureurs du marché afin améliorer l'adéquation du tarif à chaque profil de risque pour ainsi attirer de nouveaux assurés mais également pour les conserver.

1.2 Présentation du mémoire

L'étude a été réalisée au sein de la Direction des Marchés IARD d'AXA France, au sein de la branche AXA Particuliers/Professionnels dans l'équipe Innovation Pricing sur une problématique orientée Multirisque Habitation.

En 2015, des travaux de refonte d'une variable segmentante du tarif, le zonier, ont été menés en assurance Multirisque Habitation des particuliers chez AXA France. Cette variable permet de diviser le territoire en zones de risque en fonction du risque géographique prédit. À cette occasion, beaucoup de zoniers ont été crées afin d'obtenir un zonier par garantie, en différenciant maison et appartement et le cas échéant, en séparant zonier Coût Moyen et zonier Fréquence. Cette étude a permis d'ouvrir des pistes de réflexion sur le traitement du signal géographique et un constat a été fait : l'instabilité d'un zonier réalisé à une maille géographique très fine de l'ordre des coordonnées GPS.

Le lecteur est invité à se familiariser avec la notion de **maille géographique**. En effet, tout au long du mémoire des références sont faites concernant des données ex-

ternes issues de plusieurs sources et correspondant à des mailles géographiques différentes. La figure 1.1 permet de visualiser la hiérarchie de ces mailles. Il est à noter que lorsqu'une référence à la maille ADRESSE est faite, il s'agit d'une maille aux coordonnées GPS (x,y). Ces deux termes seront employés sans distinction.

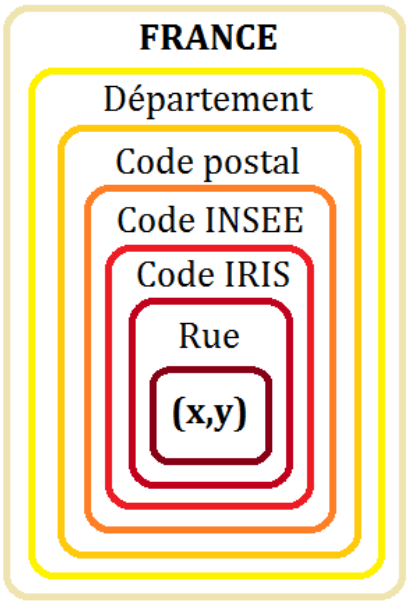

FIGURE *1.1 – Les différentes mailles géographiques.*

Suite au constat réalisé lors des précédents travaux, l'objectif de ce mémoire est d'établir une méthode alternative de construction d'une structure de micro-zonier, qui réponde à la double exigence de stabilité et d'apport d'information. L'étude est réalisée sur une base de données d'environ 800 000 observations relatives aux contrats appartements présents dans le portefeuille en 2013. Le périmètre des données utilisées pour cette étude est intégralement détaillé dans le chapitre 2.

Par ailleurs, ce mémoire de Data Science géographique possède une composante non masquée de recherche au sens où il a l'ambition, outre de répondre à la problématique métier de reconstruction d'un zonier robuste et prédictif, de définir une méthode générique d'exploitation de l'information géographique (résiduelle ou globale) qui est amenée à être développée à d'autres garanties habitation ou d'autres lignes métier telles que l'auto.

Une première partie de traitement des données a permis de recenser l'ensemble des variables externes à disposition, de les traiter et de les joindre avec la base de données des contrats Multirisque Habitation d'AXA. La base finalisée contient plus de 600 variables.

La démarche consiste ensuite à :

1. Établir un modèle prédictif du risque en fonction de données observées sur le contrat et de données géographiques externes. Dans cette étude, de la nouvelle donnée géographique à une maille adresse est testée. Elle est intégrée dans les modèles tarifaires de façon directe dans la régression (sélection des variables les plus influentes par Gradient Boosting). Une mesure est faite sur l'apport de ces données par rapport aux données agrégées à une maille moins fine.

2. Par la suite, mener une étude sur les résidus de cette première régression et diagnostiquer un reste d'information géographique à capter. Ainsi, un diagnostic de la corrélation géographique est réalisé à l'aide d'un semivariogramme, et si la corrélation géographique est établie, une méthode de traitement du risque géographique résiduel est mise en œuvre afin de tester différents types de maillage du territoire et tester différentes approches de lissage.

3. Une fois le risque géographique traité, il est nécessaire d'établir une méthode de diagnostic et de résolution d'instabilité de la prédiction du risque géographique puisque cette prédiction dans le modèle technique est amenée à varier beaucoup suivant la base d'apprentissage utilisée.

La première partie concernant l'étude de l'apport des données externes est réalisée sur la garantie principale en Multirisque Habitation : la garantie Dégât des Eaux (sur les appartements). Pour des raisons opérationnelles et après avoir mis en évidence le gain de l'apport des données externes, le processus de création de micro-zonier est réalisé sur la garantie VOL sur les appartements pour un modèle de fréquence afin de prédire le risque de survenance d'un sinistre VOL.

1.3 Définition d'un micro-zonier

L'explication d'un risque peut provenir de plusieurs sources : des caractéristiques de l'objet assuré, de celles de l'assuré lui-même, mais également de la localisation du risque. En particulier, sur un produit habitation où le bien assuré a une localisation fixe, les données géographiques sont fortement explicatives du risque. Il est donc essentiel de prendre en considération cet aspect du risque dans les modèles de prédiction. De cette manière, les assureurs peuvent alors adapter le montant des primes pour être rentable et compétitif sur chaque zone pour les classiques raisons d'ajustement de la prime au risque mais également pour rester aligné avec la stratégie de distribution par réseau d'agences, dont les zones de chalandises sont très marquées géographiquement.

On peut alors définir un **zonier** comme le traitement du signal géographique. Celui ci s'obtient à partir de l'adresse du contrat et se module par l'ensemble des données

externes relatives à sa situation géographique, comme par exemple la densité de population ou encore le revenu moyen à l'adresse. Le terme **micro-zonier** est employé lorsque le regroupement du risque se fait à des mailles plus fines que celles employées pour les découpages administratifs usuels.

Le zonier existant

Lors de la refonte de la gamme du produit Multirisque Habitation l'année dernière, tous les zoniers ont été refaits : un pour chaque garantie et en différenciant appartements et maisons. Le processus de construction de ce zonier consiste à réaliser un modèle linéaire généralisé log-poissonien uniquement avec les variables relatives au risque et à l'assuré lui-même. Aucune donnée externe n'est introduite directement dans le modèle. Ainsi, l'intégralité du signal géographique se trouve dans les résidus.

Ensuite, la méthode d'apprentissage statistique des forêts aléatoires *(Random Forest)* est réalisée pour obtenir un résidu prédit du risque à partir de l'ensemble des caractéristiques géographiques disponibles. Pour obtenir une prédiction sur l'ensemble du territoire français, il est nécessaire de construire une structure permettant d'accueillir l'information de ces prédictions pour former le zonier. De manière à adapter la précision du maillage à l'information détenue, des **cellules de Voronoï** ont été construites autour de chaque contrat.

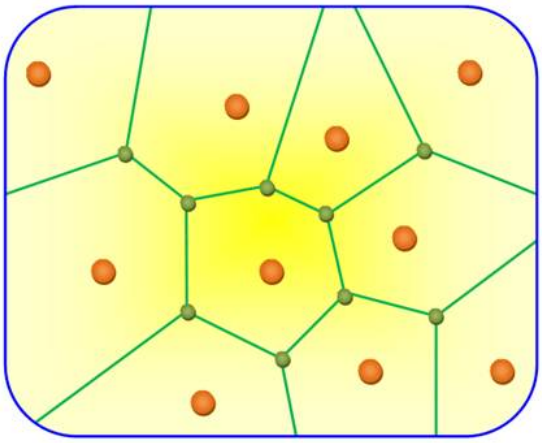

FIGURE *1.2 – Illustration de la structure des cellules de Voronoï. Source : http://villemin.gerard.free.fr/Geometri/Voronoi.htm*

Chaque cellule enferme une seule observation (contrat). Chaque observation est représentée par un point rouge dans la figure 1.2. Une cellule est ensuite construite comme l'ensemble des points plus proches de cette observation que de toutes les autres. Ainsi, pour les zones denses, les cellules seront plus petites et plus nombreuses, tandis que dans les zones où il y a moins de contrats observés, la granularité sera plus grossière. Afin d'obtenir une variable « zone » qui sera réintégrée dans le modèle linéaire généralisé, ces cellules sont ensuite agrégées de manière à obtenir 20 zones de risques.

FIGURE *1.3 – Comparaison du zonier VOL appartement à la maille INSEE (en haut) avec le zonier VOL appartement Voronoï (en bas) : région de Nanterre.*

En particulier, un zonier appartements pour la fréquence VOL a été crée par la même procédure [22] mais n'a pas été mis en production. Bien que le zonier ainsi construit ait l'avantage d'apporter un niveau de précision très élevé par rapport à un zonier à la maille INSEE (comme peut l'illustrer la figure 1.3), des problèmes de stabilité sont apparus dus à un sur-apprentissage : la prédiction étant trop fine, sur une garantie à la fréquence faible (donc peut exposée), le modèle s'ajustait trop a la base d'apprentissage mais n'était pas adaptable à un échantillon indépendant, pour lequel le niveau de prédiction chutait notablement.

Le zonier construit dans l'étude présente

Dans ce mémoire, l'étude du signal géographique est abordée de manière différente. Une première partie d'analyse du signal géographique est réalisée en incorporant directement dans le modèle de prédiction certaines données externes. Puis, dans une deuxième étape, les résidus du modèle sont analysés à l'aide d'un lissage spatial afin de traiter la part restante du signal géographique présente, non captée par le modèle lui-même.

1.4 Problématiques soulevées

Durant la réalisation de l'étude faite dans le présent mémoire, plusieurs problématiques ont été soulevées. Ainsi, nous tentons de répondre à ces interrogations parmi lesquelles on trouve notamment :

- L'utilisation de données à une maille ADRESSE, c'est à dire aux coordonnées GPS est-elle bénéfique ?
- Est-il possible de simplifier le questionnaire client en exploitant la multitude de données externes disponibles aujourd'hui dans un environnement où se développe le Big Data ?
- Y a t-il un gain à l'utilisation de méthodes non linéaires de Machine Learning pour la tarification par rapport aux techniques usuellement utilisées ?
- Comment modéliser le risque géographique au plus fin, mais sans être confronté à du sur-apprentissage ?

1.5 Résultats obtenus

Les constats tirés de la réalisation de ce mémoire permettent de mettre en évidence le pouvoir prédictif de données géographiques renseignées à des mailles fines dans le processus de tarification. En effet, outre le fait que les variables externes permettent d'améliorer nettement la qualité des modèles, l'étude révèle qu'elles permettent également de constituer un premier socle d'information permettant de réaliser un devis rapide *(Quick Quote)* et ainsi participer à la simplification du questionnaire client, car en demandant uniquement au client l'adresse du bien qu'il souhaite assurer, il est possible d'avoir accès à l'ensemble des données externes relatives à l'environnement géographique où évolue le contrat.

Par ailleurs, la méthode proposée dans ce mémoire suggère une approche différente dans le traitement du signal géographique par rapport à ce qui a pu être réalisé précédemment. En effet, le traitement du signal se fait ici en deux étapes dont la première étape consiste en l'intégration directe des variables externes dans le modèle de prédiction. Une fois le modèle réalisé, les résidus sont extraits afin de réaliser un lissage spatial. Différentes méthodes d'interpolation spatiale sont testées telles que la pondération par l'inverse à la distance ou bien encore la méthode géostatistique du krigeage. Cette deuxième étape amène à la création d'une nouvelle variable géographique qui est ensuite réintégrée au modèle. Cette méthode alternative du traitement du signal géographique fournit un zonier à la fois précis et robuste, ce qui répond à l'objectif premier de l'étude.

Chapitre 2

Construction de la base de données par contrat

2.1 Périmètre de modélisation

Lors de la tarification d'un contrat, plusieurs étapes sont nécessaires pour aboutir à la prime réellement payée par l'assuré. La première étape consiste à calculer l'espérance du risque que l'on appelle **prime pure**. Une fois la prime pure obtenue, la **prime technique** est obtenue en rajoutant les chargements de gestion et d'acquisition permettant de financer les coûts d'acquisition et d'administration supportés par l'assureur. Un chargement de sécurité peut également être intégré, ce qui permet de pouvoir résister à la volatilité naturelle des sinistres. C'est pourquoi ce chargement peut, par exemple, être proportionnel à la variance du sinistre. Finalement, la prime technique peut être modifiée en fonction de la politique commerciale de la compagnie d'assurance pour aboutir à la **prime commerciale**.

L'assurance non-vie se détache particulièrement de l'assurance vie de par la survenance même du sinistre mais également par le coût de ce sinistre. En effet, en assurance vie, la survenance du sinistre est certaine, seule sa date est inconnue ; tandis que pour l'assurance non-vie, la survenance du sinistre est juste probable (avec une probabilité comprise entre 0 et 1). De plus, en assurance non-vie, le coût du sinistre est rarement connu à l'avance. Par conséquent, en travaillant sur un modèle de sinistralité d'un portefeuille en assurance non-vie, il est nécessaire de pouvoir simuler à la fois la fréquence de sinistres sur le portefeuille, mais aussi leur coût. C'est l'objet même du modèle collectif (F. PLANCHET, 2003 [28]), très utilisé en actuariat.

Nous considérons donc le modèle collectif suivant :

$$S = \begin{cases} \sum_{i=1}^{N} X_i & \text{si } N > 0, \\ 0 & \text{sinon,} \end{cases}$$

avec :

- S la charge de sinistres,
- N la variable aléatoire à valeurs entières représentant le nombre de sinistres,
- La suite des variables aléatoires réelles X_i pour $i = 1, ..., N$ représentant les coûts des sinistres.

Ce modèle repose sur les hypothèses que les risques considérés sont des risques homogènes, que les variables N et les $(X_i)_{i\geq 1}$ sont indépendantes, et que les variables aléatoires $(X_i)_{i\geq 1}$ suivent la même loi. Si les hypothèses sont respectées, ce modèle à l'avantage de fournir une prime pure relativement simple à calculer puisque nous avons alors l'égalité suivante :

$$\mathbf{E}[S] = \mathbf{E}[N] \times \mathbf{E}[X].$$

L'étude proposée dans ce mémoire porte sur la modélisation du nombre de sinistres N sur des contrats Multirisque Habitation. Seuls les contrats concernant les appartements sont conservés afin d'avoir des contrats ayant un risque homogène. Les passages à la prime pure et aux primes technique et commerciale ne seront pas abordés dans ce mémoire qui traite uniquement de la création d'un micro-zonier afin d'intégrer le signal géographique dans la modélisation de la fréquence de survenance d'un sinistre et des problématiques soulevées durant l'étude.

La fréquence de survenance d'un sinistre correspond au nombre de fois où un sinistre est survenu durant la période d'exposition du client (par exemple, si le client est assuré depuis 3 mois, la période d'exposition est égale à 0.25, soit un quart d'année). Elle se calcule de la manière suivante :

$$\text{Fréquence} = \frac{\text{Nombre de sinistre}}{\text{Période d'exposition}}.$$

Les garanties qui seront abordées dans l'étude sont la garantie Dégât des Eaux et la garantie Vol. La garantie Dégât des Eaux permettra de mettre en évidence le gain de l'apport des données externes et de comparer les performances de différentes méthodes de régression, tandis que la garantie Vol sera utilisée afin de créer le processus de construction de micro-zonier en tentant de régler les problèmes de stabilité survenus dans l'étude réalisée par l'équipe Multirisque Habitation l'année dernière.

2.2 Les variables de la base initiale

La base de données utilisée pour l'étude du processus de construction du micro-zonier est une base de modélisation par contrat, contenant des variables caractéristiques du risque habitation et des variables clients observées dans les bases d'AXA France [22]. Cette base contient par exemple les variables suivantes :

Le nombre de pièces de l'appartement : il s'agit généralement d'une des variables principales des modèles. Elle représente un critère pour les assureurs afin d'appréhender la taille de l'habitation assurée.

L'inhabitation : est-ce une résidence principale ou une résidence secondaire ? Cette variable a une influence très significative sur la tarification de certaines garanties. Par exemple, en considérant le risque dégât des eaux : la fréquence de sinistre observée est plus importante dans les résidences principales que dans les résidences secondaires, mais la tendance s'inverse lorsqu'il s'agit du coût moyen. En effet, si un sinistre se déclare alors que le logement est vacant, si personne ne s'en aperçoit rapidement, le risque d'aggravation est non négligeable.

La qualité d'habitation : s'agit-il d'un propriétaire ou d'un locataire ? En ce qui concerne notre base de données qui se rapporte uniquement aux appartements, la répartition des contrats correspond à 73% de locataires et 27% de propriétaires.

La catégorie socio-professionnelle de l'assuré permet de donner une idée du mode de vie de l'assuré. On distingue une dizaine de catégories contenant notamment les étudiants, les cadres, les agriculteurs ou encore les retraités.

Les montants assurés : il s'agit de variables discriminantes pour le modèle car elles renseignent l'assureur non seulement sur le profil de risque de l'assuré mais également sur le montant auquel il s'engage à payer en cas de sinistre. On retrouve notamment dans ces variables : le capital assuré ainsi que le taux et le montant d'objets de valeur assurés. Cependant, souvent sous-estimés, ces montants sont relativement mal déclarés par les assurés. Il faut donc prendre garde à leur utilisation dans les modèles.

Les coordonnées géographiques du contrat : elles sont extraites à partir de l'adresse du contrat géocodée. Les variables de géolocalisation ne serviront pas à proprement parler dans les modèles mais seront utiles pour joindre la base initiale aux données externes.

Pour obtenir des résultats optimaux concernant l'apport des données externes, l'étude porte uniquement sur les contrats possédant un niveau de précision de géolocalisation « GCPREC » de 4 sur une échelle de 4. L'échelle de précision de la géolocalisation donnée en sortie par le géocodeur se décompose de la manière suivante :

- 4 : Adresse exacte,
- 3 : Numéro rue approché,
- 2 : Centroïde de la voie,
- 1 : Centroïde de la ville,
- 0 : Erreur.

Ainsi, pour réaliser l'étude, nous disposons d'environ 800 000 contrats appartements sur l'année 2013 avec une précision de géolocalisation maximale répartis sur toute la France comme peut l'illustrer la figure 2.1 ci-dessous.

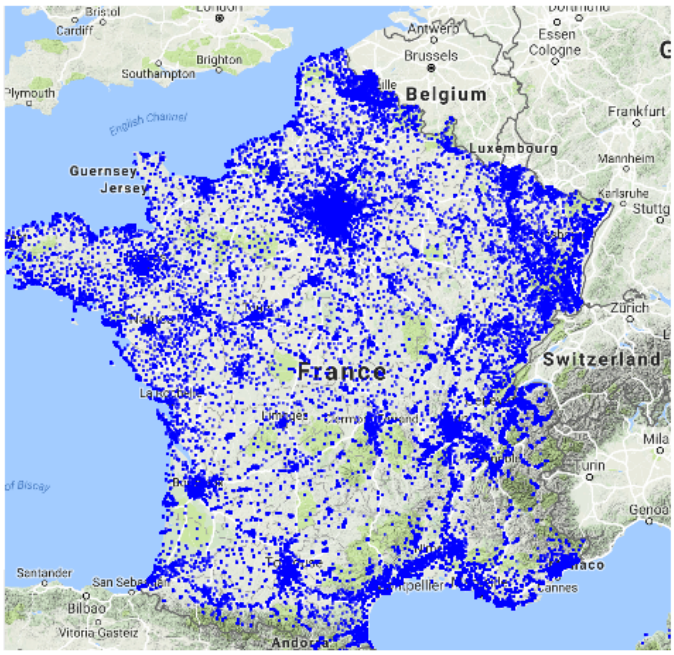

FIGURE *2.1 – Répartition des contrats appartements géocodés avec une précision « GCPREC » 4 au sein du portefeuille Multirisque Habitation d'AXA France pour l'année 2013.*

2.3 Apport de données externes

Beaucoup de données sont disponibles mais n'ont pas encore été explorées et incorporées dans les modèles de tarification. Cette partie traite de l'ensemble des données externes disponibles en fournissant un inventaire des variables dont l'influence est testée dans ce mémoire.

2.3.1 Données à la maille département

Les données à la maille département concernent la criminalité et la délinquance par département (atteinte aux biens, atteinte aux personnes, ...). Elles sont tirées des données publiques françaises de l'Observatoire National de la Délinquance et des Réponses Pénales (ONDRP) [26] qui est un département de l'Institut national des hautes études de la sécurité et de la Justice. Ces données concernent donc les crimes et délits enregistrés par les services de police et les unités de la gendarmerie nationale, on peut en particulier y trouver des variables correspondant aux vols de biens ou aux effractions par exemple.

2.3.2 Données à la maille INSEE

Les données INSEE [16] sont issues du site internet de l'INSEE. Elles sont triées par catégories : logement, socio-économique, socio-démographique et points d'intérêt. À la maille INSEE, seules les données concernant la part de résidences principales par période de construction sont conservées pour l'étude.

2.3.3 Données à la maille IRIS

AXA France dispose d'une base de données à la maille IRIS fournie par un prestataire externe. Elle regroupe en outre, des informations socio-démographiques et socio-économiques, des informations concernant les logements mais également une typologie socio-professionnelle (élites parisiennes, familles actives à la campagne, étudiants et jeunes actifs, ...).

2.3.4 Données à la maille ADRESSE

Un autre prestataire externe met à disposition d'AXA France des données à la maille ADRESSE pour les appartements et les maisons. Les données de cette base n'ont encore jamais été testées. L'étude réalisée dans ce mémoire teste l'apport de cette nouvelle donnée externe. Cette base de données contient des informations concernant la situation géographique de l'adresse (toutes les données de géolocalisation, l'attractivité de l'adresse et des données concernant le bâtiment). Elle fournit également des données à la maille IRIS.

Nous disposons aussi de données Open Street Map contenant la géolocalisation (c'est à dire les coordonnées GPS) de plusieurs points d'intérêt, dont notamment :

- Station de police
- Clinique
- Station de pompiers
- Université
- Théâtre
- Espaces surveillés
- Banque

Une fonction R a été créée pour calculer la distance au point d'intérêt le plus proche pour chaque adresse d'un contrat de notre base d'étude. La figure 2.2 représente un exemple de variable créée correspondant à la distance aux points d'intérêt les plus proches (en l'occurrence pour les points d'intérêt relatifs aux stations de police).

Le dégradé de couleur représente la distance à la station de police la plus proche. La figure fait également apparaître la localisation des stations de police dans Paris et ses alentours.

FIGURE *2.2 – Carte choroplèthe de la distance à la station de police la plus proche et géolocalisation des postes de police pour Paris et ses alentours.*

2.4 Création de variable

Chercher à créer de nouvelles variables à partir des données à notre disposition permet d'exploiter pleinement la richesse des données. Pour profiter entièrement de toutes les données à disposition, il est nécessaire de prendre possession de la base et travailler les données de manière optimale. Ainsi, nous avons décidé de créer une variable qui traduit l'écart de revenu entre l'adresse du contrat et l'ensemble des adresses voisines dans un rayon de 2km. En effet, des variables similaires ont été créées dans d'autres études réalisées au sein d'AXA ([1] et [9]) et leurs influences sur la fréquence VOL se sont révélées non négligeables.

Voici les étapes de construction de la variable qui a été créée dans notre étude :

- On crée un cercle centré sur le contrat avec un rayon de 2km comme présenté dans la figure 2.3 grâce à la fonction `gBuffer` du package `rgeos` à partir des coordonnées GPS du contrat [23] [24].
- On sélectionne les adresses appartenant à ce cercle.
- On agrège la donnée en prenant le minimum des revenus moyens sur les adresses sélectionnées.
- On calcule l'écart de richesse entre l'adresse du contrat et la donnée agrégée.

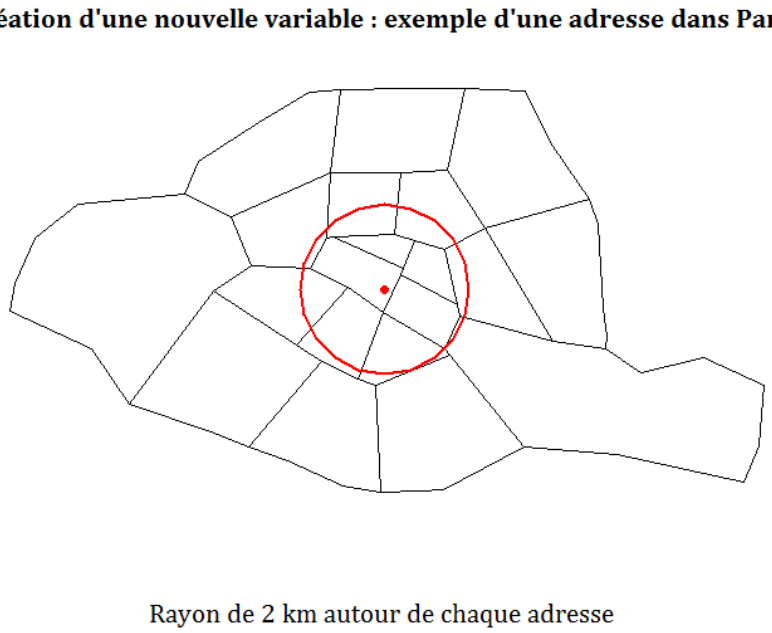

FIGURE *2.3 – Illustration de la technique utilisée pour la première étape de création de la nouvelle variable relative à l'écart de revenu moyen avec les voisins.*

2.5 Les données géographiques

Grâce au géocodage de la base de données, les adresses des contrats ont été normalisées de manière à pouvoir exploiter l'information concernant l'environnement géographique dans lequel évolue le contrat.

Le travail du signal géographique conduit à rencontrer et à manipuler plusieurs types d'information ou de structures spatiales telles que des coordonnées géographiques ou bien des polygones. Il est nécessaire de maîtriser le maniement de ces éléments, notamment lors de création de variables géographiques, de jointures spatiales ou encore de réalisation de cartes.

Système de coordonnées

Il est important lorsque l'on utilise des Systèmes d'Information Géographique (SIG), de bien comprendre et de gérer les différents systèmes de coordonnées auxquels on peut être confronté. En effet, il existe plusieurs référentiels qui permettent de représenter un point dans l'espace [33] , dans notre étude nous travaillons notamment avec :

- Les systèmes de coordonnées sous la forme de coordonnées géographiques en degrés : latitude, longitude et hauteur ellipsoïdale. Ce système ne nécessite pas l'utilisation d'une projection cartographique.

- Les systèmes de coordonnées sous forme de coordonnées projetées en mètre (représentation plane) munis d'une projection cartographique. Une projection permet de représenter sur une surface bidimensionnelle une partie d'une surface sphérique tridimensionnelle de la terre. L'utilisation d'un tel système de coordonnées et de sa projection permet de faciliter l'évaluation des distances puisqu'il s'agit de valeurs métriques plus facilement exploitables que les valeurs angulaires de latitude et longitude.

Il y a donc des systèmes de coordonnées avec et sans projection où on peut définir une **projection cartographique** comme étant le système de correspondance entre les coordonnées géographiques et les points du plan de projection. Il existe plusieurs types de projection permettant de représenter la surface de la Terre dans son ensemble ou en partie sur la surface plane d'une carte. Le choix d'une projection dépend de l'usage qui sera fait de la carte mais aussi de la position de la région à cartographier sur le globe. La figure 2.4 illustre les trois principaux types de projections.

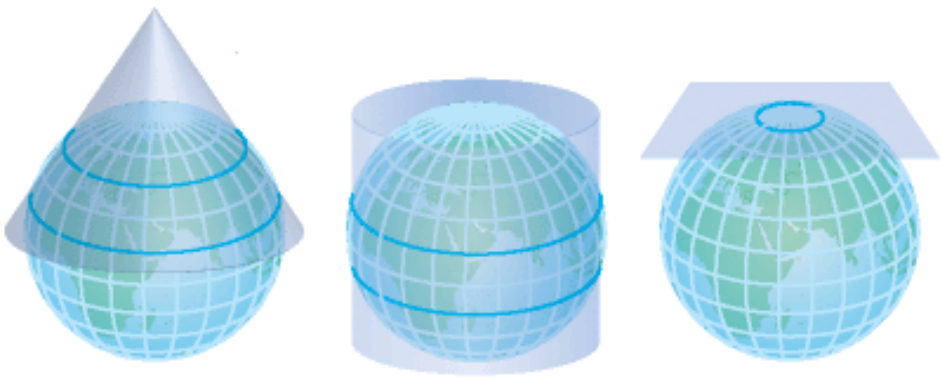

FIGURE *2.4 – Les différents systèmes de projection. Dans l'ordre : la projection conique, la projection cylindrique et la projection azimutale. Source : Blog technique de Nicolas Boonaert.*

Par exemple, les coordonnées de Paris peuvent être exprimées sous forme de coordonnées géographiques en degrés dans le système RGF93 (Réseau Géodésique Français), ou bien sous forme de coordonnées projetées en mètres dans le même système RGF93 avec en plus l'utilisation de la projection conique Lambert 93.

La projection Lambert 93 illustrée dans la figure 2.5 est utilisée pour représenter seulement une partie du globe en minimisant les déformations pour la France. Il s'agit donc de la projection officielle française. Elle a été conçue pour être utilisée uniquement avec le système RGF93.

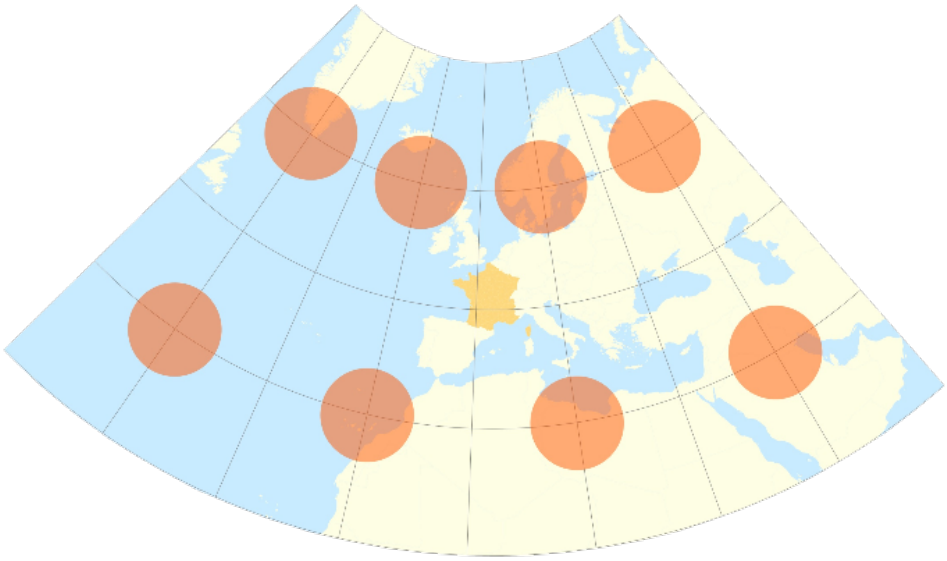

FIGURE 2.5 – *Projection cartographique Lambert 93. Source : pôle ARD, adess, domaine public.*

Chaque ensemble - système de coordonnées et projection cartographique s'il y en a une, ou simplement système de coordonnées - est identifié par un code, attribué par l'EPSG (European Petroleum Survey Group). En fonction des besoins et en particulier lors de la création de la distance au point d'intérêt le plus proche et lors de la création de la variable représentant l'écart de revenu moyen avec les adresses voisines, deux types de coordonnées ont été utilisés :

- Les coordonnées GPS (longitude et latitude) dans le système WGS84 (World Geodetic System 1984), code EPSG : 4326 ;
- Les coordonnées dans le système RGF93 avec la projection Lambert 93 (système à 2 dimensions), code EPSG : 2154.

Exemple des coordonnées de Paris pour ces systèmes

- Pour le système WGS84, en degrés décimaux : **(2.346614, 48.84535)** ;
- Pour le système RGF93 avec la projection Lambert 93 : **(E 652048.9, N 6860787.74)**.

Un même point aura donc des coordonnées différentes selon le système utilisé pour les mesurer. Les codes EPSG sont importants lors de la réalisation d'opérations spatiales car ils permettent d'identifier le système de coordonnées utilisé et de convertir les coordonnées dans le bon système si nécessaire.

Lors de l'utilisation du logiciel R en tant que Système d'Information Géographique, certaines fonctions nécessitent parfois d'employer un système de coordonnées particulier ou de spécifier le système utilisé. Ainsi, une erreur de gestion des systèmes de coordonnées peut conduire à des résultats aberrants ou altérer le processus de création des zoniers si par mégarde les jointures spatiales sont mal réalisées.

Les jointures spatiales

Il n'est pas toujours possible de réaliser une jointure attributaire s'il n'y a pas de colonne commune entre deux tables. Dans le cas de tables ayant une relation spatiale, il est alors envisageable de réaliser des jointures spatiales. Les jointures spatiales permettent de combiner les informations de plusieurs tables en utilisant une relation spatiale comme clé de jointure. Les deux tables doivent alors être exprimées dans le même système de coordonnées afin que la jointure réalisée soit juste.

Les jointures spatiales seront notamment utilisées lorsque nous voudrons joindre des contrats à des données renseignées par polygone ou par carreaux, telles que la zone géographique de risque. En effet, le découpage des zones peut s'avérer indépendant des découpages administratifs usuels et il n'existe donc pas de clé de jointure attributaire telle que le code INSEE ou le département, seulement une relation spatiale qui permet de déterminer dans quelle zone se situe le contrat à l'aide de ses coordonnées géographiques.

2.6 Base finalisée

Une fois le recensement de toutes les bases de données disponibles réalisé, il faut à présent établir une seule et unique base pour commencer l'étude. Les manipulations des bases de données ont été réalisées sous SAS puis R. Les données à la maille INSEE sont jointes à la base des contrats par code INSEE, tandis que celles à la maille IRIS sont jointes par code IRIS et celles à la maille département par numéro de département. La base de données à la maille adresse est, quant à elle, jointe à la base des contrats avec une triple clé de jointure dont les champs ont été préalablement normalisés (retrait des accents, tirets, apostrophes, etc...) :

[ADRESSE] X [CODE POSTAL] X [NOM DE LA COMMUNE].

La figure 2.6 permet de synthétiser l'information disponible que nous utiliserons dans l'étude.

La base finalisée contient plus de 600 variables. L'étude porte sur le portefeuille Multirisque Habitation d'AXA France pour les contrats appartements sur l'année 2013. La base est ensuite découpée aléatoirement en deux échantillons :

Un échantillon d'apprentissage contenant 70 % de la base et sur lequel les méthodes sont appliquées et les algorithmes apprennent. Il permet d'ajuster le modèle aux données.

Un échantillon test qui permet de valider le modèle sur une base de données totalement indépendante de la base sur laquelle le modèle a appris et ainsi simuler la réception de nouvelles données. Cet échantillon permet d'évaluer objectivement l'erreur réelle de prédiction du modèle en comparant valeurs prédites et valeurs observées.

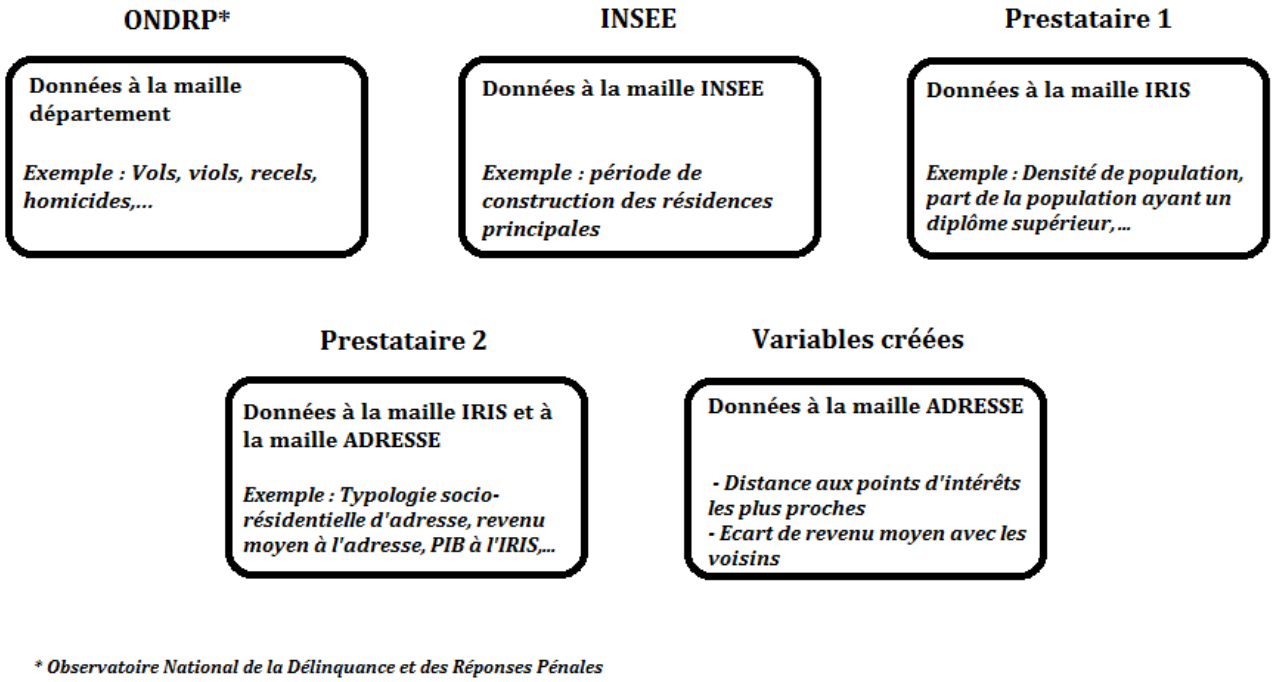

FIGURE *2.6 – Récapitulatif des données externes utilisées dans l'étude.*

Bilan et transition

La quantité de données disponibles pour réaliser l'étude est très importante. En plus des données issues du questionnaire client qui renseignent sur le risque assuré et sur le client lui même, nous disposons de beaucoup de données externes issues de sources diverses et concernant des mailles variées allant d'une précision grossière comme le département à une maille beaucoup plus fine : les coordonnées GPS (x,y) du contrat. C'est d'ailleurs par l'intermédiaire des coordonnées que nous pouvons également bénéficier de l'utilisation des Systèmes d'Information Géographique afin de manipuler les données référencées spatialement et ainsi profiter pleinement de la richesse des données, en créant des nouvelles variables qui seront testées par la suite.

La suite du mémoire amène à étudier l'ensemble de ces données disponibles en testant l'apport de la nouvelle donnée externe à la maille la plus fine : la maille ADRESSE. Cette étude est réalisée sur la garantie principale du produit Multirisque Habitation : la garantie dégât des eaux qui ne présente aucun inconvénient notable pour obtenir un modèle robuste puisque ce risque dispose de suffisamment d'observations pour avoir une bonne fréquence.

Une fois cette analyse réalisée, nous pourrons ensuite mettre en œuvre le processus de construction de micro-zonier en intégrant cette nouvelle donnée. Le processus portera sur la garantie VOL pour tenter de résoudre des problèmes de stabilité survenus sur le précédent zonier réalisé l'année passé par l'équipe Multirisque Habitation.

Chapitre 3

Analyse des données et sélection des variables

Dans cette section, une analyse des données disponibles est réalisée. La méthode d'apprentissage automatique du Gradient Boosting permet de réaliser un tri sur l'ensemble des variables et sert également en tant que modèle de régression pour prédire la variable réponse.

Ainsi, dans un premier temps la théorie du Gradient Boosting est développée, dont la technique nous permet de tester l'apport des données externes puis sert à analyser et sélectionner les variables qui apportent potentiellement le plus d'information dans le modèle.

3.1 Théorie du Gradient Boosting

Le Gradient Boosting est une technique d'apprentissage automatique utilisée pour des problèmes de régression ou de classification.

La méthode de boosting la plus connue est sans doute celle qui a été développée par Freund et Schapire en 1995 : l'**AdaBoost** (pour Adaptive Boosting), mais elle ne sera pas détaillée dans le présent mémoire. Cependant, le lecteur est invité à se reporter à l'article *A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting* [11] pour plus d'informations concernant cette méthode.

Comme les autres méthodes de boosting, le Gradient Boosting optimise les performances d'un ensemble de modèles de prédiction dits « faibles » en les assemblant en un modèle final. On appelle modèle de prédiction « faible », une méthode de classification ou de régression qui est à peine plus efficace qu'un tirage aléatoire. Le modèle de prédiction faible généralement utilisé avec un Gradient Boosting est un arbre de décision CART. De manière plus concrète, cette méthode du **Gradient Tree Boosting** consiste alors à réaliser une succession d'arbres de décision où chaque modèle est construit sur l'erreur résiduelle du précédent.

3.1.1 L'arbre de décision CART

L'arbre CART (*Classification and Regression Tree*) est une méthode d'apprentissage statistique supervisée visant à construire un modèle de prédiction. L'idée principale de l'arbre CART est de partitionner récursivement et de manière binaire l'espace des variables explicatives afin d'obtenir toutes les valeurs possibles de la variable à prédire.

Supposons que nous disposons d'un n-échantillon d'une variable Y avec p prédicteurs $x_1, ..., x_p$. L'objectif est de pouvoir prédire les valeurs de Y pour de nouvelles valeurs $x_1, ..., x_p$. Introduisons t_p, le nœud parent et t_g, t_d les nœuds fils respectivement gauche et droit. À partir de l'ensemble des observations sur la variable Y, l'arbre CART sélectionne à chaque itération et donc à chaque nœud t_p, la variable x_j et sa valeur x_j^* qui segmentent le mieux l'espace des variables en deux sous-espaces les plus homogènes possible dans les nœuds t_g et t_d. Cette notion d'homogénéité se traduit par une fonction d'impureté notée $i(\cdot)$. Il existe plusieurs fonctions qui permettent de définir l'impureté $i(t)$ d'un nœud t, mais dans le cas de l'étude d'une variable aléatoire Y continue, comme c'est le cas ici, la variance du nœud peut représenter un bon indicateur de l'hétérogénéité présente dans le nœud. L'impureté d'un nœud parent restant constante indépendamment du découpage réalisé pour créer les nœuds fils, maximiser l'homogénéité dans les nœuds fils est alors équivalent à maximiser la variation d'impureté Δi défini comme :

$$\Delta i = i(t_p) - \mathbf{E}\Big(i(t_{\text{fils}})\Big). \tag{3.1}$$

Ce concept de réduction d'impureté peut être appréhendé par la figure 3.1 où la réduction d'impureté est plus importante sur l'illustration de gauche que sur celle de droite.

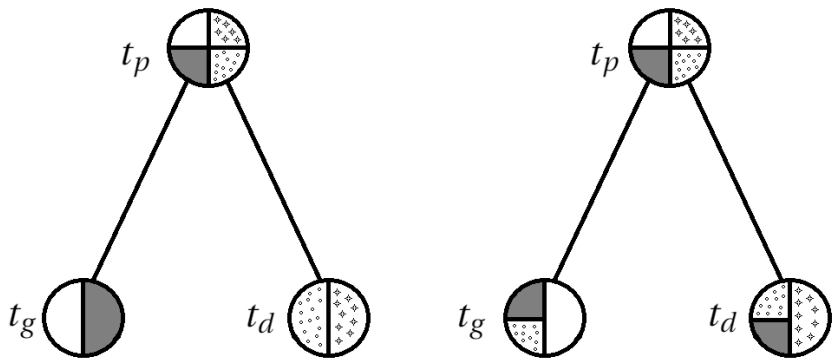

FIGURE *3.1 – Illustration du concept de réduction d'impureté de l'arbre CART. Idée d'illustration par M.Gahbiche*

Le problème de maximisation de l'impureté Δi définie dans l'équation 3.1 peut également s'écrire :

$$\max_{x_j<x_j^*, j=1,...,p} \Big[i(t_p) - P_g \times i(t_g) - P_d \times i(t_d) \Big], \tag{3.2}$$

où P_g et P_d sont respectivement les probabilités des nœuds gauche et droit. Ce problème traduit la manière de procéder de l'algorithme : il parcourt toutes les valeurs possibles pour toutes les variables x_j pour $j = 1, ..., p$ afin de trouver le meilleur découpage qui permettra de maximiser la variation d'impureté Δi.

Dans le cas où la fonction d'impureté est représentée par la variance du nœud, le problème de maximisation 3.2 peut alors se ré-écrire :

$$\min_{x_j<x_j^*, j=1,...,p} \Big[P_g \times \mathrm{Var}(t_g) + P_d \times \mathrm{Var}(t_d) \Big]. \tag{3.3}$$

Pour de plus amples détails concernant la méthode d'apprentissage automatique CART, le lecteur peut se référer à Roman Timofeev, *Classification And Regression Trees* [32].

3.1.2 Algorithme du Gradient Boosting

Le Gradient Boosting est un algorithme itératif qui distribue initialement des poids égaux à toutes les prédictions puis les adapte à chaque étape, de sorte que les mauvaises prédictions soient sur-pondérées à l'étape suivante pour que le modèle de prédiction « faible » y accorde plus d'attention. Revenons à l'article original de Friedman [12] en reprenant ses notations pour développer l'idée du Gradient Boosting plus en détails.

Nous disposons d'un échantillon de taille n constitué de p variables explicatives. Notons $\mathbf{x}_i = (x_{i,1}, ...x_{i,p}) \in \mathbb{R}^p$, le vecteur des p variables explicatives correspondant à l'observation i. L'objectif est de trouver une approximation $F_{\text{approx}}(\mathbf{x})$ d'une fonction $F^*(\mathbf{x})$ reliant les variables explicatives $\mathbf{x}$ à la variable réponse y et qui minimise l'espérance d'une certaine fonction de perte $L(y, F(\mathbf{x}))$ sur la distribution jointe de $\mathbf{x}$ et y.

Mathématiquement, la fonction $F^*(\mathbf{x})$ est alors définie par :

$$F^*(\mathbf{x}) \in \arg\min_F \; E_{y,\mathbf{x}} \Big[L(y, F(\mathbf{x})) \Big],$$

où les fonctions de pertes $L(y, F)$ généralement employées sont la perte quadratique $(y - F)^2$ et la perte absolue $|y - F|$.

Une procédure courante est de supposer qu'il existe une vision paramétrique $F(\mathbf{x}; \mathbf{P})$ de la fonction $F(\mathbf{x})$, où l'évaluation en $\mathbf{x}$ dépend à présent d'un ensemble fini de paramètres $\mathbf{P} = \{P_1, P_2, ..., P_M\} \in \mathbb{R}^M$. Dans son article, l'auteur se concentre sur les modèles additifs de la forme :

$$F\Big(\mathbf{x};\{\beta_m,a_m\}_1^M\Big)=\sum_{m=1}^{M}\beta_m h(\mathbf{x};a_m),$$

où la fonction h est le modèle de prédiction « faible » évoqué précédemment qui permet de relier les variables $\mathbf{x}$ aux paramètres $\mathbf{a}=(a_1,...,a_M)$.

Dans le cas d'un Gradient Tree Boosting, la fonction h est donc un arbre de régression (le plus souvent un arbre CART). Le cas échéant, les paramètres $\mathbf{a}_m$ correspondent alors aux variables utilisées pour la séparation des branches de l'arbre afin d'aboutir aux nœuds terminaux. En prenant l'exemple d'un arbre de régression à J feuilles, on obtient J régions distinctes $R_1,...R_J$ formant une partition de l'espace des variables. La fonction h peut alors s'écrire elle-même sous une forme additive :

$$h\Big(\mathbf{x};\{b_j,R_j\}_1^J\Big)=\sum_{j=1}^{J}b_j\mathbb{1}_{x\in R_j},$$

où b_j correspond à la valeur prédite par la région R_j.

Dans le cas où on considère une vision paramétrique de la fonction F, il faut prendre en compte un nouveau problème d'optimisation :

$$P^*\in\arg\min_{P\in\mathbb{R}^M}\ \phi(P),$$

où

$$\phi(P)=E_{y,\mathbf{x}}\Big[L(y,F(\mathbf{x}\,;\mathbf{P}))\Big],$$

et on considère ensuite que

$$F^*(\mathbf{x})=F(\mathbf{x}\,;\mathbf{P}^*).$$

La solution du problème d'optimisation est de la forme :

$$\mathbf{P}^*=\sum_{m=0}^{M}\mathbf{p}_m,$$

où $\mathbf{p}_0$ correspond à la première approximation grossière de $\mathbf{P}^*$ et les $\{\mathbf{p}_m\}_1^M$ aux incréments successifs ou *boosts* amenant à l'approximation de $\mathbf{P}^*$.

Le modèle non paramétrique initial devient alors :

$$F^*(\mathbf{x})\in\arg\min_F\ \underbrace{E_{y,\mathbf{x}}\Big[L(y,F(\mathbf{x};\mathbf{P}))\Big]}_{\phi(F(\mathbf{x}))}.$$

De la même manière que dans le problème paramétrique, la solution s'écrit sous la forme :

$$F^*(\mathbf{x})=\sum_{m=0}^{M}f_m(\mathbf{x}),$$

où $f_0(\mathbf{x})$ correspond à la première approximation grossière et les $\{f_m\}_1^M$ aux incréments successifs.

On définit également $\forall m \in \{0, ..., M\}$:

$$F_m(\mathbf{x}) = \sum_{i=0}^{m} f_i(\mathbf{x}),$$

ainsi, l'approximation cherchée vaut alors $F_{\text{approx}}(\mathbf{x}) = F_M(\mathbf{x})$.

Par l'algorithme de descente du gradient, on a :

$$f_m(\mathbf{x}) = -\rho_m g_m(\mathbf{x}),$$

avec

$$\begin{aligned} g_m(\mathbf{x}) &= \left[\frac{\partial \phi(F(\mathbf{x}))}{\partial F(\mathbf{x})}\right]_{F(\mathbf{x})=F_{m-1}(\mathbf{x})} \\ &= \left[\frac{\partial E_y[L(y, F(\mathbf{x}))|\mathbf{x}]}{\partial F(\mathbf{x})}\right]_{F(\mathbf{x})=F_{m-1}(\mathbf{x})}. \end{aligned}$$

En permutant dérivation et espérance, on obtient :

$$g_m(\mathbf{x}) = E_y\left[\frac{\partial L(y, F(\mathbf{x}))}{\partial F(\mathbf{x})}\middle|\mathbf{x}\right]_{F(\mathbf{x})=F_{m-1}(\mathbf{x})}.$$

Et le facteur multiplicatif ρ_m est donné par :

$$\rho_m \in \arg\min_{\rho \in \mathbb{R}} \; E_{y,\mathbf{x}}\Big[L(y, F_{m-1}(\mathbf{x}) - \rho g_m(\mathbf{x}))\Big].$$

Finalement, l'approximation $F_{\text{approx}}(\mathbf{x})$ peut se calculer à l'aide de quantités empiriques à partir des $(\mathbf{x}_i, y_i)$ et l'algorithme du Gradient Boosting peut alors s'écrire de la manière suivante [21] :

Algorithme 1 Gradient Boosting

Initialisation de $f_0(x)$ à une valeur constante

$$f_0(x) = \arg\min_{\rho} \sum_{i=1}^{n} L(y_i, \rho)$$

POUR m = 1 à M **FAIRE**

Calcul du gradient négatif

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

Réalisation d'un modèle de regression sur g_m par les moindres carrés à partir des covariables x_i pour obtenir une estimation de a_m dans $\beta h(x; a)$.

$$a_m = \arg\min_{a,\beta} \sum_{i=1}^{n} \left[g_m(x_i) - \beta h(x_i; a)\right]^2$$

Estimation de ρ_m

$$\rho_m = \arg\min_{\rho} \sum_{i=1}^{n} L\Big(y_i, F_{m-1}(x_i) + \rho h(x_i; a_m)\Big)$$

Mise à jour

$$F_m(x) = F_{m-1}(x) + \rho_m h(x; a_m)$$

FIN DE LA BOUCLE POUR

3.1.3 Les hyper-paramètres

Dans le but d'éviter un sur-apprentissage des données et de dégrader les prédictions futures, le Gradient Boosting possède plusieurs hyper-paramètres qui permettent d'optimiser son utilisation en limitant ce phénomène de sur-apprentissage.

Le nombre d'arbres et le nombre d'observations par feuille

Le nombre d'arbres M correspond au nombre d'itérations effectuées par l'algorithme. La fonction `gbm.perf` du package `gbm` de G.Ridgeway [30] permet de déterminer le nombre d'arbres optimal.

Le nombre minimal d'observations par feuille empêche de créer des groupes trop petits pour ainsi éviter d'obtenir une feuille avec trop peu d'observations et d'être trop spécifique dans l'apprentissage des données, ce qui conduirait sans doute à un sur-apprentissage. Dans notre étude, nous décidons d'utiliser comme taille de feuille environ 1.5% du nombre total d'observations de la base d'apprentissage utilisées pour créer le modèle.

Le shrinkage

La ligne de mise à jour dans l'algorithme du Gradient Boosting peut parfois être remplacée par :

$$F_m(x) = F_{m-1}(x) + \nu\rho_m h(x; a_m).$$

Le paramètre $\nu \in (0,1]$ appelé *shrinkage* permet de réguler la contribution de chaque arbre. À chaque itération, l'algorithme n'applique qu'une fraction ν du coefficient ρ à $F_m(x)$ ce qui permet de retarder la vitesse d'apprentissage de l'algorithme. Il a été montré qu'un facteur $\nu < 0.1$ augmentait considérablement la qualité du modèle par rapport à un modèle sans *shrinkage* ($\nu = 1$). Cela a pu être vérifié en pratique lors de notre étude. Le prix à payer pour l'utilisation de petites valeurs de ν est l'augmentation du nombre de d'arbres M requis. Il faut donc trouver un juste équilibre pour ces deux paramètres.

Le bagging

Une autre modification intéressante à apporter à l'algorithme du Gradient Boosting est une touche d'aléatoire : le *stochastic bagging* des données. On parle alors de *Stochastic Gradient Boosting* [13]. Le *bagging* (pour *bootstrap averaging*) consiste à tirer un échantillon aléatoire sans remise des données à chaque itération. L'échantillon aléatoire ainsi tiré est alors utilisé, à la place de l'échantillon total, par le modèle de prédiction faible et pour calculer la mise à jour de l'algorithme.

Un échantillon équivalent à 50% de l'échantillon total est généralement considéré à chaque itération. Cette technique aura pour conséquence d'augmenter la variance de chaque modèle de prédiction faible individuellement mais diminuera la corrélation entre les estimations des différentes itérations. Cependant, comme l'effet agrégé domine l'effet individuel, l'effet global correspond alors à une réduction de la variance du modèle combiné final :

$$F(\mathbf{x}) = \sum_{m=1}^{M} \beta_m h(\mathbf{x}; a_m).$$

Par ailleurs, utiliser la technique du *bagging* permet de réduire le temps de calcul de l'algorithme car à chaque étape l'arbre de régression ne doit s'ajuster que sur un échantillon plus petit de données. Il faut néanmoins garder à l'esprit que si l'échantillon utilisé pour ajuster le modèle aux données est trop petit, l'algorithme risque de voir son efficacité diminuer.

3.1.4 Interprétation des résultats

Les modèles de Gradient Boosting ne s'interprètent pas de manière aussi évidente qu'un modèle linéaire généralisé, cependant les résultats sont tout de même explicables par l'intermédiaire de l'influence relative des variables explicatives et de leurs graphiques de dépendances partielles.

Influence relative des variables dans le modèle Pour comprendre au mieux le rôle des différentes variables utilisées dans le modèle, il est judicieux de s'intéresser à l'influence relative de chaque variable dans le modèle.

En effet, l'importance des variables explicatives permet de les classer en fonction du nombre de fois où chaque variable est utilisée dans le découpage des arbres du boosting. Les importances des variables sont ensuite normalisées de manière à ce que la somme des nombres normalisés soit égale à 100 et ainsi obtenir un moyen de mesurer l'influence relative de chaque variable.

Graphiques des dépendances partielles Les graphiques de dépendances partielles permettent de visualiser l'effet marginal de chaque variable utilisée dans le modèle.

Ces deux méthodes d'interprétation du Gradient Tree Boosting sont utilisées dans le chapitre 4 lors de la mise en œuvre du modèle de régression.

3.2 Mesure de l'apport des données à la maille adresse par rapport aux mêmes données agrégées

Une des problématiques soulevées lors de l'analyse de la nouvelle donnée géographique externe a été de tester l'apport des données à la maille adresse par rapport aux mêmes données agrégées à l'IRIS. Pour plus de visibilité sur les résultats, l'étude est réalisée sur la garantie principale en Multirisque Habitation : la garantie Dégât des eaux.

3.2.1 Les mesures de performance

Plusieurs indicateurs ont été utilisés pour évaluer les modèles réalisés lors de la régression. L'indice de GINI est un indicateur performant mais il doit être combiné à d'autres mesures comme par exemple le Root Mean Squared Error.

L'indice de GINI, indicateur du pouvoir discriminant

L'indice de GINI est un indicateur mesurant la capacité de segmentation du modèle [6]. Il est calculé à partir de la fonction représentée par la courbe de Lorenz.

Développée en 1905 par Max O. Lorenz, la courbe de Lorenz, ou courbe de gain, permet de représenter graphiquement les inégalités de revenus au sein d'une population. La fonction qui lui est associée calcule la part des revenus par rapport à la part des détenteurs.

Dans l'étude proposée, la courbe de gain représente en abcisses, la part cumulée des contrats et en ordonnées, la part cumulée du nombre de sinistres.

L'indice de GINI permet de comparer deux modèles ou de tester l'apport de nouvelles variables. Il est calculé à partir de l'aire sous la courbe et correspond au ratio entre les aires A et B de la figure 3.2.

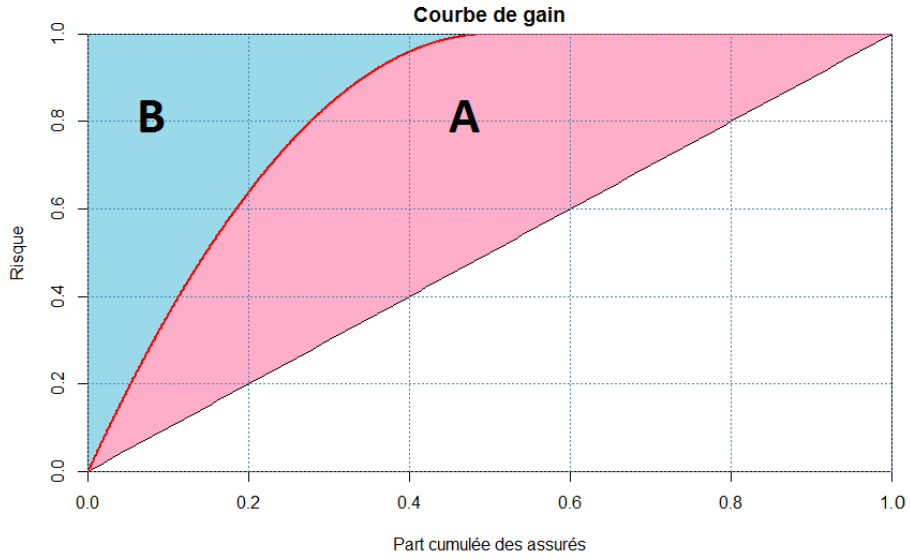

FIGURE *3.2 – Calcul de l'indice de GINI comme étant le ratio des aires A et B. La première bissectrice représente le modèle aléatoire, la courbe rouge représente le modèle testé et le modèle parfait est représenté par la courbe reliant les points (0,0), (0,1) et (1,1).*

La première bissectrice représentée dans la figure 3.2 décrit un modèle aléatoire ou encore une égalité parfaite lorsque l'on observe une mutualisation égale de la charge sur l'ensemble des assurés : « *x % des assurés détiennent x % du risque* ». Si, l'aire entre la courbe de gain et l'égalité parfaite vaut A, et que l'aire au dessus de la courbe de gain vaut B alors l'indice de GINI est défini comme

$$G = \frac{A}{A+B}.$$

Étant donné que $A+B=0.5$, l'indice de GINI vaut $G=2A$, ou encore $G=1-2B$.

En supposant que la courbe de Lorenz représente la fonction $y=L(x)$ alors la valeur de l'aire B peut s'exprimer à l'aide de l'intégrale :

$$B = 1 - \int_0^1 L(x)\,dx,$$

ainsi, l'indice de GINI peut alors s'exprimer comme :

$$G = 2\int_0^1 L(x)\,dx - 1.$$

Le modèle optimal ou saturé n'est pas nécessairement représenté par le carré supérieur. Il s'agit généralement d'une seconde courbe plus proche du carré supérieur. Il est alors intéressant de normaliser l'indice de GINI standard.

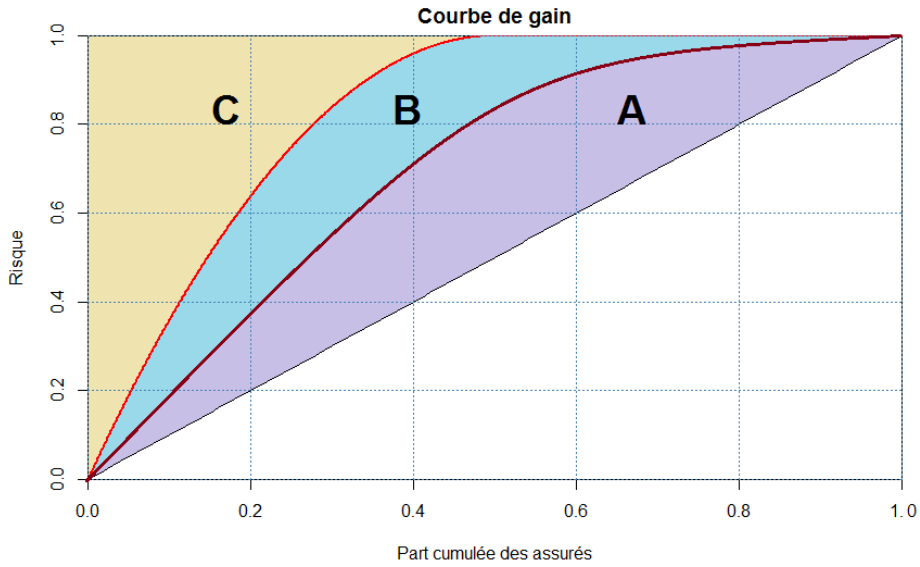

FIGURE 3.3 – *Calcul de l'indice de GINI normalisé en considérant une deuxième courbe représentant le modèle saturé.*

Considérons à présent la figure 3.3 et les GINI standards pour le modèle saturé

$$G_S = \frac{A+B}{A+B+C},$$

et pour le modèle testé

$$G_M = \frac{A}{A+B+C}.$$

Le GINI normalisé est alors défini par :

$$\tilde{G} = \frac{G_M}{G_S}.$$

En remplaçant par les définitions de G_M et G_S, on retrouve :

$$\tilde{G} = \frac{A}{A+B}.$$

Le RMSE, indicateur du pouvoir de prédiction

Le Root Mean Squared Error est utilisé pour évaluer la performance d'un modèle [15]. Calculé à partir des résidus issus de la différence entre les valeurs prédites par le modèle $X_{\text{modèle}}$ et les valeurs observées X_{obs}, le Root Mean Squared Error permet d'agréger toutes ces erreurs de prediction en une unique mesure au pouvoir prédictif :

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (X_{\text{modèle},i} - X_{\text{obs},i})^2}{n}},$$

avec n le nombre d'observations. Il s'agit d'une bonne mesure de prédiction pour comparer différents modèles. Néanmoins, il est important de noter que, dans notre étude, les Root Mean Squared Error s'appliqueront sur des fréquences (prédites et observées), c'est pourquoi les valeurs de cette mesure seront très faibles et les différences se feront sur les décimales ce qui rend l'interprétation des résultats moins évidente.

3.2.2 Mesure de l'apport à la maille adresse

Pour tester l'apport de la nouvelle donnée géographique à l'adresse, par rapport aux autres données, les variables sont ajoutées au Gradient Boosting au fur et à mesure, et selon leur provenance (variables internes, à la maille INSEE, à la maille IRIS, à la maille ADRESSE et points d'intérêt).

Plusieurs modèles de Gradient Boosting sont donc réalisés :

- **1er Gradient Boosting** : données internes récoltées à la souscription
- **2ème Gradient Boosting** : données internes récoltées à la souscription + données à la maille IRIS + données à la maille INSEE
- **3ème Gradient Boosting** : données internes récoltées à la souscription + données à la maille ADRESSE + Points d'intérêt
- **4ème Gradient Boosting** : données internes récoltées à la souscription + données à la maille IRIS + données à la maille INSEE + données à la maille ADRESSE + Points d'intérêt

Les performances de chaque modèle se trouvent dans la table 3.1.

Variables	GINI Train	GINI Test	Delta GINI	RMSE Train	RMSE Test
Internes	27,3 %	26,7%	0,6	0,2646045	0,2805371
Internes + maille INSEE et IRIS	33,2%	32,9%	0,3	0,2641126	0,2800929
Internes + maille adresse	33,6%	33,3%	0,3	0,2640436	0,280036
Internes + maille INSEE, IRIS et adresse	33,9%	33,5%	0,4	0,2640408	0,2800455

TABLE *3.1 – Tableau des performances des modèles permettant d'appréhender l'impact des données externes.*

La table 3.1 nous renseigne sur la qualité de segmentation des modèles et leur capacité de classement des risques par l'intermédiaire du coefficient de GINI. L'ajout des données externes permet d'accroître significativement le pouvoir segmentant du modèle tout en maintenant une bonne stabilité. Quant aux Root Mean Squared Error, les améliorations sont relativement faibles mais attestent également d'une amélioration du modèle lors de l'ajout des variables externes. Par ailleurs, l'utilisation de variables externes à la maille ADRESSE permet de créer un modèle plus performant que lorsque des variables externes à des mailles géographiques plus grossières sont utilisées.

3.2.3 Mesure de l'apport des données agrégées

Les données à la maille ADRESSE donnent des informations à la maille la plus fine possible : les coordonnées GPS. L'étude comparative réalisée dans cette section permet de tester si une maille aussi fine est nécessaire ou bien si le gain d'information est similaire lorsque que cette donnée est agrégée à une maille plus grossière. Le cas échéant, faut-il alors privilégier le principe de simplicité et utiliser la maille la plus grosse ?

Le prestataire externe qui fournit la base de données à la maille ADRESSE, fournit également certaines variables à la maille IRIS. Par ailleurs, pour les variables à la maille ADRESSE qui n'auraient pas d'équivalence à la maille IRIS existante dans la base, elles sont agrégées à ce même niveau géographique (à l'IRIS).

La comparaison de l'apport des données à la maille ADRESSE par rapport aux mêmes données agrégées à la maille IRIS est réalisée dans la table 3.2.

Variables	GINI Train	GINI Test	Delta GINI	RMSE Train	RMSE Test
Internes + maille adresse	33,6%	33,3%	0,3	0,2640436	0,280036
Internes + maille adresse agrégée à l'IRIS	33,3%	33,0%	0,3	0,2640802	0,280063

TABLE 3.2 – *Tableau comparatif des performances des modèles sur les données externes à la maille ADRESSE agrégée à la maille IRIS.*

Dans l'étude, nous tentons de capter un maximum d'information géographique par des données externes fines. La table 3.2 témoigne que l'utilisation des données à la maille ADRESSE est légèrement préférable à celle des mêmes données agrégées à la maille IRIS bien que l'apport semble limité. Ainsi, les données à la maille ADRESSE disponibles apportent un niveau d'information relativement équivalent à une maille agrégée plus grossière. Il est donc intéressant de faire remarquer que, certes, les données à la maille ADRESSE apportent naturellement davantage d'information, mais que les écarts de bon classement entre une donnée très fine - possiblement complexe à industrialiser - et une donnée agrégée à la maille IRIS témoignent qu'en cas d'opérationnalisation, les données agrégées à la maille IRIS représentent une bonne approximation de la maille ADRESSE.

Par ailleurs, ces résultats peuvent laisser présager qu'il demeure une information plus fine que la maille IRIS à capter, puisque la donnée à maille ADRESSE à disposition n'a pas suffi à sur-performer suffisamment la donnée à la maille IRIS. Ainsi, il reste peut-être de l'information à expliquer à la maille fine mais par un autre moyen que la donnée à disposition. On peut alors supposer que cela indique d'un résidu important de signal géographique restant.

3.2.4 Un modèle composé uniquement de variables externes ?

Dans le cadre de la simplification du questionnaire client lors de la souscription d'un contrat, il était intéressant de tester un modèle composé uniquement de variables externes et ainsi répondre à la problématique : « Est-il possible d'évaluer le risque rattaché au client uniquement à partir de son adresse ? ». Pour cela, une analyse comparative est effectuée entre un modèle composé uniquement des données fournies par le client : les variables internes, et un modèle composé uniquement des données externes obtenues grâce à la seule adresse de l'assuré.

Variables	GINI Train	GINI Test	Delta GINI	RMSE Train	RMSE Test
Internes uniquement	27,3 %	26,7%	0,6	0,2646045	0,2805371
Externes uniquement	26,6%	26,1%	0,5	0,2647610	0,2806844

TABLE 3.3 – *Comparaison des performances des modèles avec l'intégration les variables internes uniquement ou de variables externes uniquement.*

La table 3.3 laisse supposer que les données externes fournissent autant d'information que les seules données internes puisque les niveaux des Root Mean Squared Error sont sensiblement similaires bien que le modèle constitué exclusivement des variables externes soit légèrement moins segmentant que le modèle constitué des variables internes uniquement.

Ainsi, un modèle qui n'intègre que des données externes géographiques est performant car il est pratiquement équivalent à un modèle contenant uniquement des données internes. Ce constat montre le pouvoir prédictif des données géographiques externes. D'un point de vue opérationnel, cela permettrait par exemple d'établir un tarif de base au client reposant simplement sur peu de critères, l'adresse de la personne, plus quelques autres critères. Il est donc envisageable de simplifier le questionnaire client lors de la souscription.

3.3 Sélection et description des variables les plus influentes

Au regard de la quantité importante de variables provenant de sources diverses, plusieurs Gradient Tree Boosting sont utilisés afin de présélectionner les variables influentes sur la variable réponse. Ces variables sont ensuite analysées avant d'être intégrées dans le modèle de prédiction. En effet, bien qu'un maximum d'information aide toujours à mieux expliquer le risque à modéliser, il faut tout de même éviter la redondance d'information afin d'aider le modèle à mieux performer. Une analyse des corrélations est donc faite en amont avant de réaliser le modèle de prédiction.

Par ailleurs, bien que le Gradient Tree Boosting soit supposé intégrer la sélection des variables dans son algorithme, nous décidons tout de même de procéder à un tri des variables de manière à pouvoir confronter l'efficacité de l'algorithme lorsque toutes les variables sont intégrées sans traitement préalable avec un modèle où les variables intégrées sont sélectionnées par d'autres méthodes.

3.3.1 Analyse préliminaire des variables pré-sélectionnées

Afin de réduire l'incorporation de bruit dans le modèle, nous tentons d'éviter d'intégrer des variables inutiles. Pour cela une pré-sélection des variables est réalisée pour chaque ensemble de données provenant de sources différentes.

Pour rappel, les sources de données sont les suivantes :

- Les données internes relatives au contrat, au risque et à l'assuré.
- Les données de criminalité et de délinquance à la maille département provenant de l'Observatoire Nationale de la Délinquance et des Réponses Pénales.
- Les données à la maille INSEE concernant la période de construction du bâtiment.

- Les données à la maille IRIS provenant d'un prestataire externe
- Les données à la maille ADRESSE provenant d'un second prestataire externe fournissant également des données à la maille IRIS ; ainsi que les données concernant la distance aux points d'intérêt les plus proches et la variable d'écart de revenu moyen avec les adresses environnantes.

Ainsi, pour chaque ensemble de données, un Gradient Tree Boosting est réalisé pour sélectionner les variables influentes provenant de cette source. Une fois que l'ensemble des Gradient Tree Boosting aura été réalisé, toutes les variables seront ensuite intégrées ensemble dans un modèle de Gradient Tree Boosting final. (Les données à la maille IRIS et à la maille INSEE ont été incorporées ensemble dans un même modèle de Gradient Tree Boosting pour la pré-sélection des variables).

La sélection réalisée sur les variables à la maille département

À titre d'exemple, la figure ci-dessous correspond à la pré-sélection des variables provenant de l'Observatoire Nationale de la Délinquance et des Réponses Pénales réalisée par Gradient Tree Boosting.

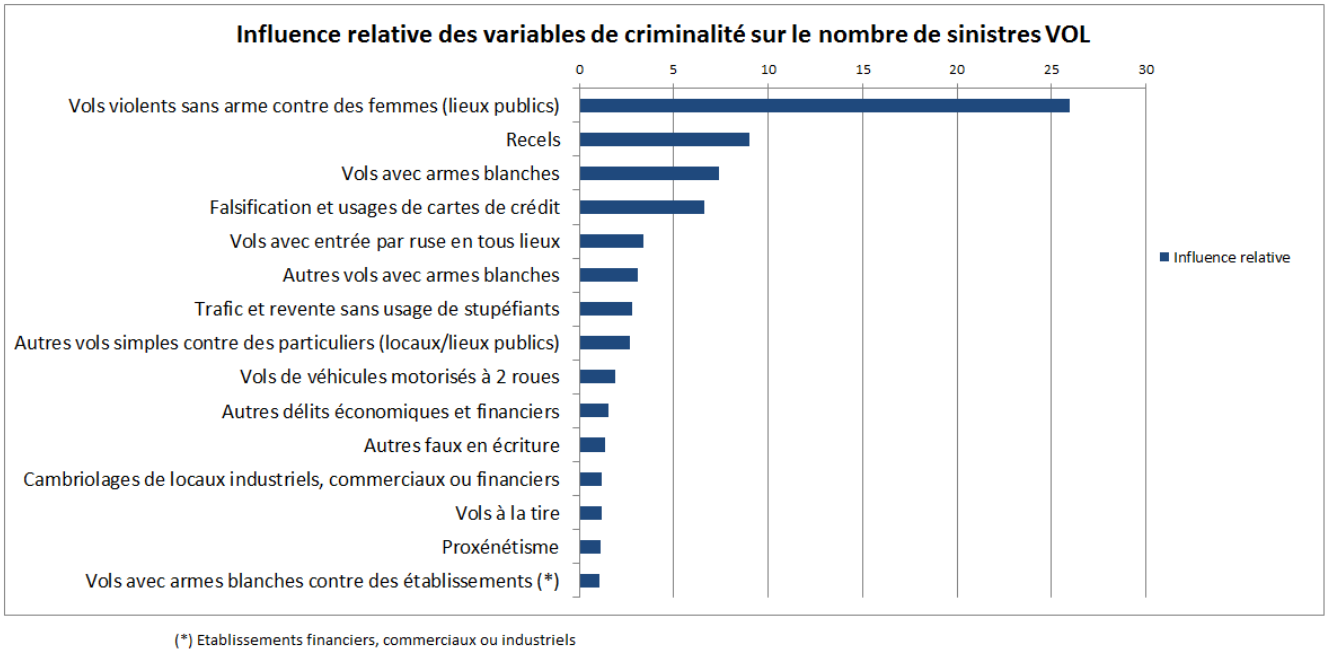

FIGURE 3.4 – *Influence relative des variables provenant de l'Observatoire National de la Délinquance et des Réponses Pénales.*

La figure 3.4 met en évidence l'influence relative des variables de criminalité et de délinquance ayant un impact sur le nombre de sinistre VOL pour les appartements. Le Gradient Boosting réalisé uniquement sur les variables issues de l'Observatoire National de la Délinquance et des Réponses Pénales indique que la variable la plus influente est celle concernant le nombre de vols violents sans armes contre des femmes dans les

lieux publics. Cependant, lorsque les variables sélectionnées par ce modèle sont incorporées avec les variables issues des autres sources, il apparaît que la variable la plus influente est la variable correspondant aux nombres de vols avec armes blanches.

Les variables concernant la criminalité et la délinquance sont très corrélées entre elles comme l'illustre le corrélogramme 3.5 ci-dessous.

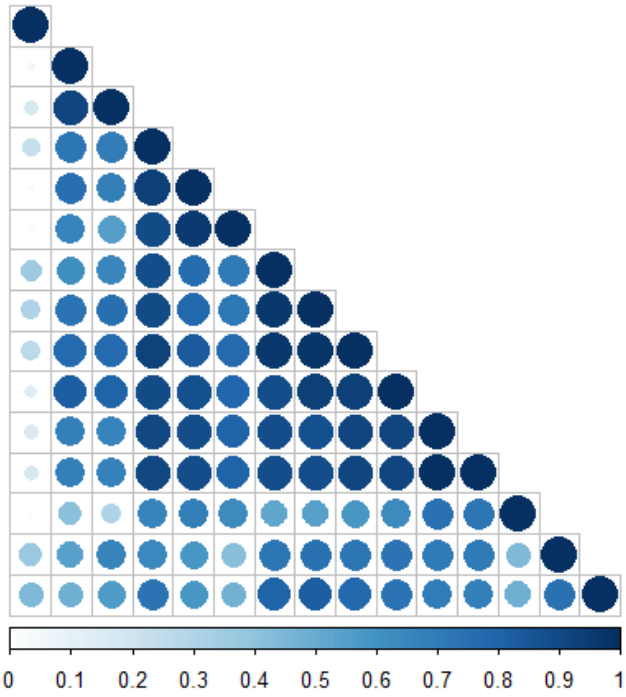

FIGURE 3.5 – *Corrélogramme des variables provenant de l'Observatoire National de la Délinquance et des Réponses Pénales.*

Par ailleurs, ces variables sont données à une maille département qui est donc très grossière. C'est pourquoi, nous décidons de garder uniquement une variable provenant de l'Observatoire National de la Délinquance et des Réponses Pénales : la variable du nombre de vols avec armes blanches. Ce choix repose sur le fait que cette variable apparaît dans les premières variables les plus influentes parmi toutes celles issues de la même source et c'est celle qui ressort en premier dans le modèle incorporant toutes les variables issues des différentes sources de données.

Les Gradient Tree Boosting réalisés sur les autres ensembles de données

De la même manière, un Gradient Tree Boosting est réalisé sur chaque ensemble de variables provenant des différentes sources. Nous rappelons que notre base de données contenait plus de 600 variables. Par conséquent, cette pré-sélection parallèle des variables pour chaque source permet de réduire considérablement le nombre de variables à intégrer dans le modèle tout en éliminant celles n'ayant pas d'influence sur notre variable réponse. A l'issue de cette étape, seules 145 variables restent candidates

à l'explication du risque de survenance d'un sinistre VOL pour les appartements. Dans la suite, une analyse descriptive de certaines de ces variables sera réalisée avant d'entreprendre une seconde sélection par différents procédés.

3.3.2 Description des variables influentes

Le but est de bien appréhender la donnée disponible et d'obtenir des intuitions sur les relations entre certaines variables explicatives et la variable réponse : la fréquence VOL sur les appartements.

Variables internes

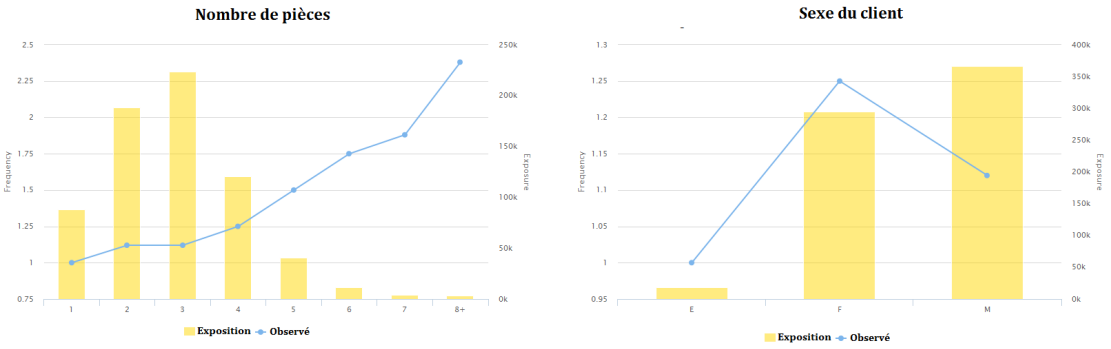

FIGURE 3.6 – *Fréquences relatives observées pour les variables internes.*

Variables externes

FIGURE 3.7 – *Fréquences relatives observées pour les variables externes.*

La première figure concerne le nombre de pièces de l'appartement. C'est une des principales variables utilisée lors de la tarification d'un contrat. Il apparaît que la survenance d'un sinistre VOL est croissante avec le nombre de pièces. Cela pourrait s'expliquer par le fait que le nombre de fenêtres augmente avec le nombre de pièces : il y aurait donc possiblement plus de « portes d'entrée » pour les malfaiteurs. Par ailleurs, un très léger écart d'exposition au risque en fonction du genre du client (homme ou femme) se dégage mais cette variable ne peut pas être prise en compte dans les modèles car il est interdit de discriminer le tarif avec le genre.

Les figures du dessous concernent des variables externes issues de différentes sources. La variable « Indice de Développement Humain » est donnée à la maille IRIS quant à la variable « Vol à l'arme blanche », elle provient des données issues de l'Observatoire National de la Délinquance et des Réponses Pénales à la maille Département. Enfin, la variable relative à la classification urbaine est une donnée qui agrège une information fournie à la maille ADRESSE.

Cette dernière variable est très intéressante puisqu'elle traduit la caractère urbain de l'environnement du contrat. La signification des modalités de cette variable est la suivante :

- **A** Agglomérations de Paris, Lyon et Marseille (plus de 1 100 000 habitants)
- **B** Très grandes agglomérations de 570 000 à 1 100 000 habitants
- **C** Grandes zones urbaines de 100 000 à 570 000 habitants
- **D** Zones urbaines de taille moyenne de 20 000 à 100 000 habitants
- **E** Petites zones urbaines de moins de 20 000 habitants
- **F** Zones rurales
- **0** Autres (immeubles vides, logements isolés sans qualification sociodémographique ...)

Il semblerait que les zones les plus peuplées soient les plus risquées et que le risque croit avec la densité de population. Ces constats sont probablement directement lié a l'exposition au risque car le risque est multiplié par plus de deux entre une zone rurale et une agglomération comme Paris, Lyon ou Marseille.

La tendance du risque est décroissante avec l'Indice de Développement Humain. Rappelons que l'Indice de développement humain (IDH) est une mesure sommaire du niveau moyen atteint dans des dimensions clés du développement humain : vivre une vie longue et en bonne santé, acquérir des connaissances et jouir d'un niveau de vie décent. L'IDH est la moyenne géométrique des indices normalisés pour chacune des trois dimensions. [1] A l'inverse, la tendance du risque est croissante pour la part de vol

1. Source : United Nations Development Programme, *Human Development Reports* `http://hdr.undp.org/fr/content/indice-de-développement-humain-idh`

à l'arme blanche. Ainsi les vols à l'arme blanche et les vols sur les habitations sont très liés. Enfin, la nouvelle variable créée qui décrit l'écart de revenu moyen avec les voisins possède également une nette tendance croissante.

La plupart des graphiques font apparaître des tendances croissantes ou décroissantes et l'observation d'une tendance traduit souvent une relation entre la variable étudiée et le risque, c'est pourquoi il est important de réaliser une première étude descriptive pour bien comprendre ces relations ainsi que le risque lui-même.

3.4 Analyse des corrélations des variables pré-sélectionnées

Lors de la décision dans le choix des variables à intégrer dans le modèle, l'étude des corrélations est importante car la présence de plusieurs variables corrélées dans un modèle est à éviter.

3.4.1 Étude graphique des corrélations

Les variables quantitatives

Cette section traite de l'étude des corrélations entre les variables quantitatives sélectionnées par les différents Gradient Tree Boosting.

Le coefficient de corrélation entre deux variables X et Y, défini par :

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y},$$

avec σ_X et σ_Y les écarts-types respectifs des variables X et Y, est compris entre -1 et 1 et traduit une forte corrélation positive dans le cas où il est compris entre $0,5$ et 1 et une forte corrélation négative dans le cas où est compris entre -1 et $-0,5$.

Nous ne cherchons pas à déterminer si les corrélations sont positives ou négatives mais seulement si elles sont importantes. Ainsi, la figure 3.8 ci-dessous illustre les corrélations des variables quantitatives pré-sélectionnées par les Gradient Tree Boosting dont la valeur absolue est supérieure au seuil de 0.8. Le seuil choisi permet de mettre en évidence le nombre important de variables très fortement corrélées entre elles. Il faut cependant remarquer que les corrélations des variables forment des « groupes » et les quatre groupes qui se distinguent correspondent aux différentes sources dont les variables sont issues :

- Les variables à la maille département fournies par l'Observatoire National de la Délinquance et des Réponses Pénales,
- Les variables à la maille IRIS fournies par un prestataire externe,
- Les variables à la maille IRIS fournies par un second prestataire externe,

- Les variables à la maille ADRESSE fournies par le second prestataire externe.

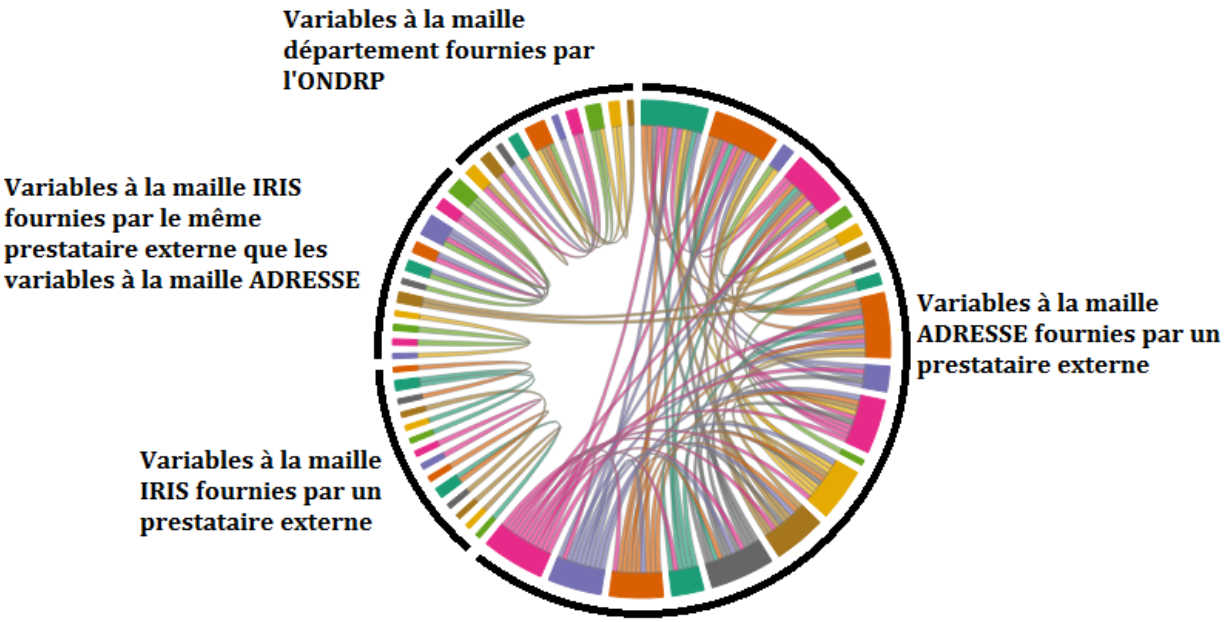

FIGURE 3.8 – *Illustration des corrélations des variables quantitatives, visiblement quatre groupes de variables se distinguent.*

Chaque variable est représentée par un bandeau de couleur et les liens qui relient ces bandeaux illustrent les corrélations existantes entre les variables. Il semble que chaque variable ne possède de fortes corrélations qu'avec les variables provenant de la même source. Seules deux liaisons, que l'on voit clairement traverser le diagramme, font exception. Il s'agit de la variable de revenu moyen à la maille IRIS fortement corrélée avec le revenu moyen à la maille ADRESSE ainsi qu'avec l'écart de revenu moyen avec les voisins (à la maille ADRESSE également).

Suite à l'étude des corrélations pour les variables quantitatives, nous ferons en sorte de limiter l'apport de données fortement corrélées dans le modèle de Gradient Tree Boosting final en réalisant une seconde sélection pour chacun des autres groupes de variables apparaissant sur la figure 3.8 (la sélection sur l'ensemble des variables provenant de l'Observatoire National de la Délinquance et des Réponses Pénales ayant déjà été réalisée dans la section précédente).

Les variables qualitatives

Pour l'analyse des variables qualitatives, l'indicateur de corrélation du V de Cramer est utilisé. Il permet de mesurer la corrélation entre deux variables qualitatives :

compris entre -1 et 1 pour les cas extrêmes de dépendance totale, il vaut 0 en cas d'indépendance.

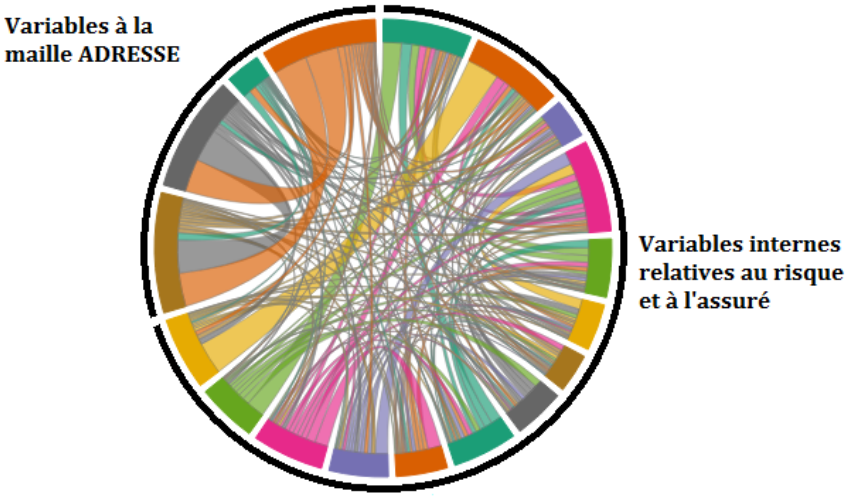

FIGURE 3.9 – *Corrélation des variables qualitatives (V de Cramer). Beaucoup de corrélations apparaissent mais leur importance reste suffisamment faible pour ne pas en tenir compte.*

Dans la figure 3.9 ci-dessus, aucun seuil de corrélation n'est fixé. Il est plus aisé d'apercevoir ici la variation d'épaisseur des traits, les plus épais symbolisant une corrélation plus importante entre les variables. Cependant, les corrélations les plus importantes représentées par les liens les plus larges correspondent seulement à des valeurs avoisinant 0.5. Ainsi, les variables qualitatives ne possèdent pas de très fortes corrélations comme c'était le cas pour les variables quantitatives.

Enfin, pour analyser les corrélations entre les variables qualitatives et quantitatives, ces dernières sont discrétisées et on utilise une fois encore le V de Cramer. Les corrélations qui ressortent sont celles déjà observées dans les deux précédentes figures.

3.4.2 Étude à l'aide d'une analyse des composantes principales (ACP)

Dans les précédentes sections, il a été établi que les variables quantitatives issues des différentes sources possédaient de fortes corrélations avec les variables de même provenance. En particulier, les variables à la maille IRIS provenant d'un des prestataires externes sont concernées. Dans le but d'éviter d'incorporer des variables trop corrélées entre elles, une Analyse en Composantes Principales (ACP) est réalisée sur ces données

afin d'obtenir des variables orthogonales résumant au mieux l'information contenue dans les variables initiales.

Les fondements de l'analyse en composantes principales

Comme il est clairement détaillé dans l'ouvrage de I.T Jolliffe, [18] *Principal Component Analysis*, l'Analyse en Composantes Principales permet de réduire la dimension d'un ensemble de données correspondant à un grand nombre des variables quantitatives corrélées entre elles, tout en conservant un maximum de l'information présente dans les données.

Notons $\mathbf{X}$ notre ensemble de données représenté par une matrice de taille $n \times p$ de n individus et p variables. Cet outil d'analyse de données va permettre de représenter les p variables dans un sous-espace F_m en explicitant au « mieux » les liaisons initiales entre ces variables, tout en réduisant la dimension de notre matrice de données en l'approximant par une matrice de même dimensions mais de rang m inférieur à p.

L'Analyse en Composantes Principales permet de transformer les données initiales en un nouveau jeu de variables, les composantes principales, décorrélées entre elles et ordonnées de manière à ce que les premières contiennent la plus grande partie de l'information contenue dans l'ensemble des variables initiales. Les composantes principales contiennent les coordonnées des individus sur les axes factoriels qui correspondent aux directions de l'espace qui expliquent au mieux la variance de l'échantillon. Ces axes factoriels sont en fait les vecteurs propres de la matrice de covariance (ou de corrélation) des variables. La valeur propre λ_k associée à l'axe factoriel k, pour $k = 1, ..., p$, représente le pourcentage de variance expliquée par l'axe.

L'inertie totale I_T peut alors se définir comme étant la somme des valeurs propres :

$$I_T = \sum_{k=1}^{p} \lambda_k,$$

elle mesure la dispersion totale du nuage de points.

Même si l'Analyse en Composantes Principales permet d'obtenir jusqu'à p composantes principales, on espère en général qu'une grande partie de l'information contenue dans $\mathbf{X}$ sera restituée par m composantes principales avec $m < p$.

Analyse des résultats

Dans notre étude, l'Analyse en Composantes Principales est utilisée sur les données à la maille IRIS et à la maille INSEE ressorties dans le modèle de Gradient Boosting.

Le choix de réaliser une Analyse en Composantes Principales sur ces données semble justifié car il s'agit uniquement de variables quantitatives dont certaines sont

très corrélées entre elles comme il a été montré dans la section précédente et comme l'illustre plus en détail le corrélogramme 3.10.

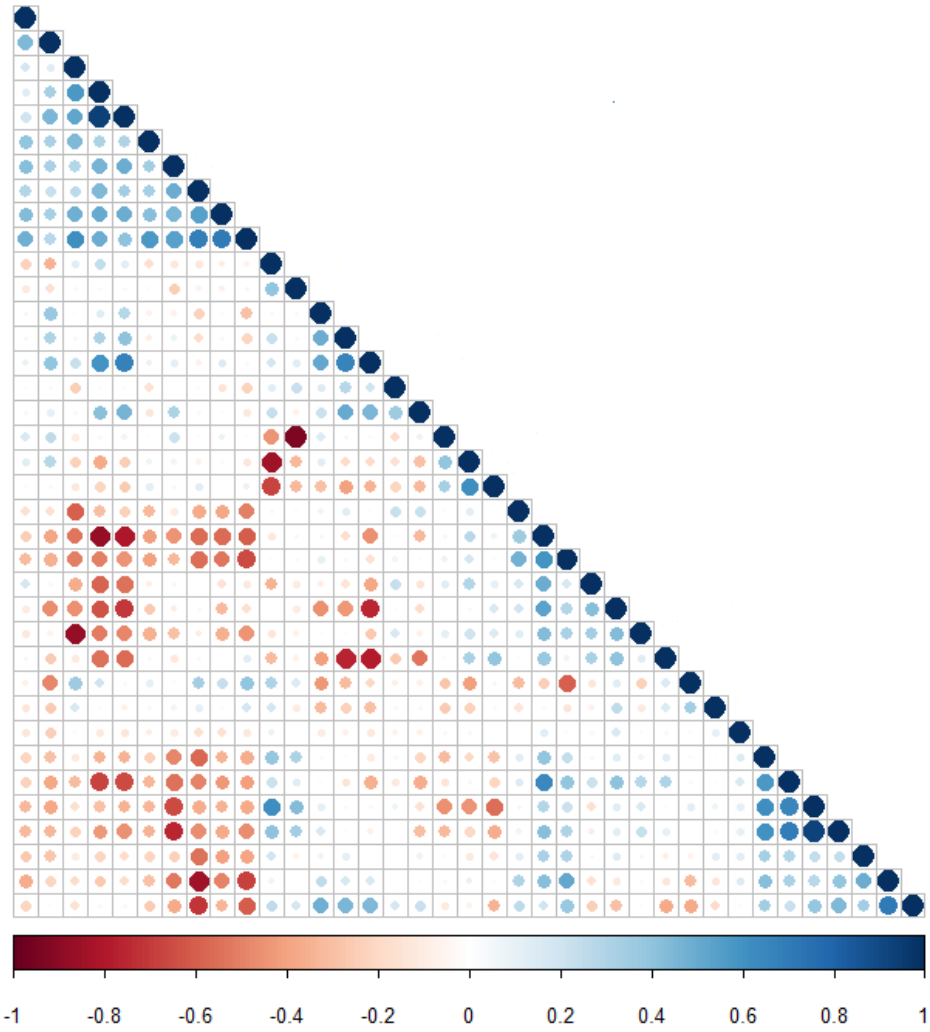

FIGURE *3.10 – Corrélogramme des variables à la maille IRIS et à la maille INSEE.*

Dans le cas présent, il faut conserver l'information présente dans ces 37 variables tout en évitant la redondance d'information due à la corrélation des variables entre elles et c'est bien le but d'une Analyse en Composantes Principales de conserver un maximum d'inertie avec un minimum de facteurs.

La sélection du nombre d'axes à conserver se fait à partir de la figure 3.11 et à l'aide de plusieurs critères :

Part d'inertie expliquée La valeur de m est choisie de sorte que la part d'inertie expliquée par les m premières composantes soit supérieure à une valeur seuil s fixée a priori par l'utilisateur :

$$\frac{\sum_{k=1}^{m} \lambda_k}{I_T} \geqslant s.$$

Critère du coude de Cattell (1966) Sur le diagramme d'effondrement des valeurs propres 3.11, un décrochement (coude) suivi d'une décroissance régulière apparaît. Les axes à conserver sont ceux qui se situent avant le décrochement.

Critère de Kaiser (1960) Les axes retenus sont ceux dont l'inertie est supérieure à l'inertie moyenne $\frac{I_T}{p}$. Dans le cas présent, l'inertie totale I_T vaut 37 et la valeur de p vaut également 37. Ainsi le critère de Kaiser retient donc ici les axes dont l'inertie est supérieure à 1.

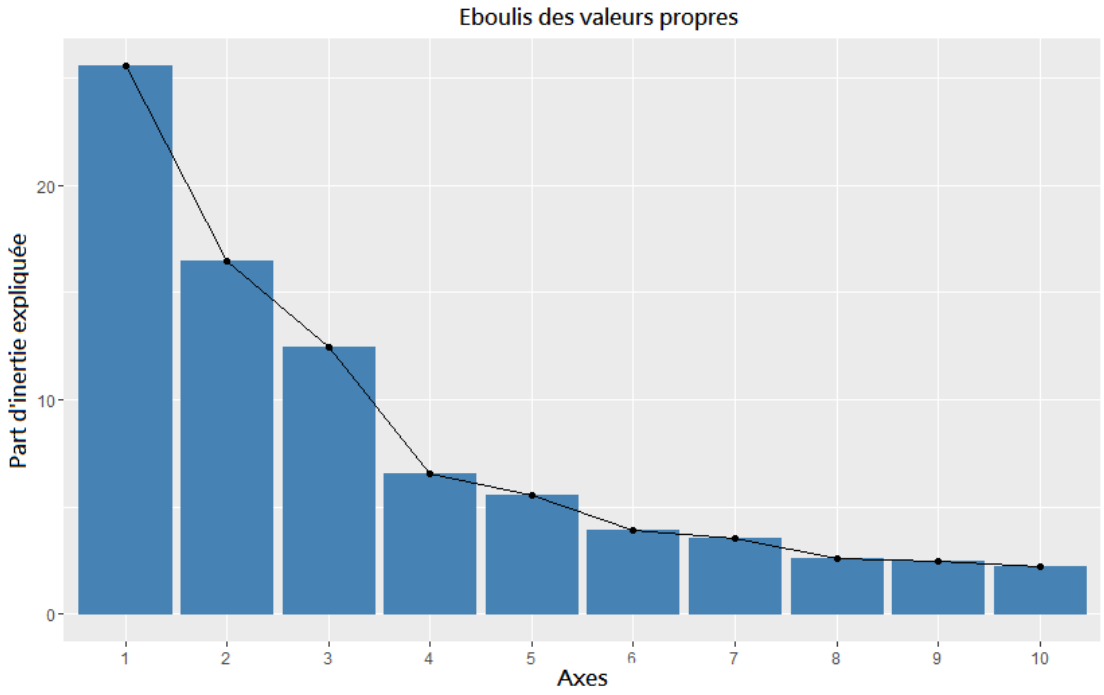

FIGURE *3.11 – Diagramme d'effondrement des valeurs propres. Le diagramme laisse apparaître deux « décrochements » après le 3^e et le 7^e axe. L'utilisation des autres critères permettra de choisir le nombre d'axes à conserver.*

Deux « décrochements » apparaissent sur le diagramme 3.11 d'effondrement des valeurs propres. Le premier décrochement se situe entre le 3^e et le 4^e axe, tandis que le deuxième se situe entre le 7^e et le 8^e axe. Ce premier critère ne permet pas de conclure sur le nombre d'axes à conserver. L'analyse des deux autres critères permettra d'approfondir cette étude.

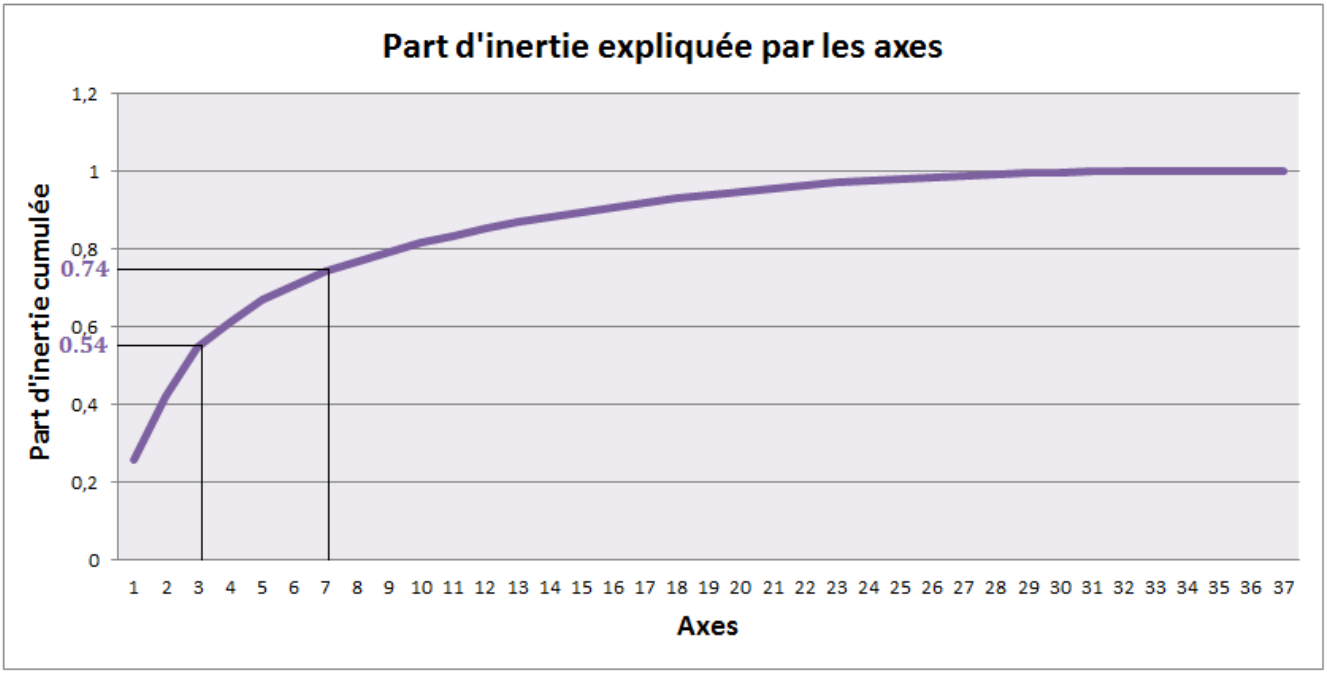

FIGURE *3.12 – Part d'inertie expliquée par les axes factoriels.*

D'après la figure 3.12, la part d'inertie expliquée par les sept premiers axes représente 74% de l'inertie totale tandis que celle pour les trois premiers axes n'en représente que 54%.

Enfin, le critère de Kaiser suggère de choisir les axes dont la valeur propre (c'est à dire l'inertie) est supérieure à 1. D'après la figure 3.13, les sept premiers axes possèdent des valeurs propres dont la valeur est supérieure à 1.

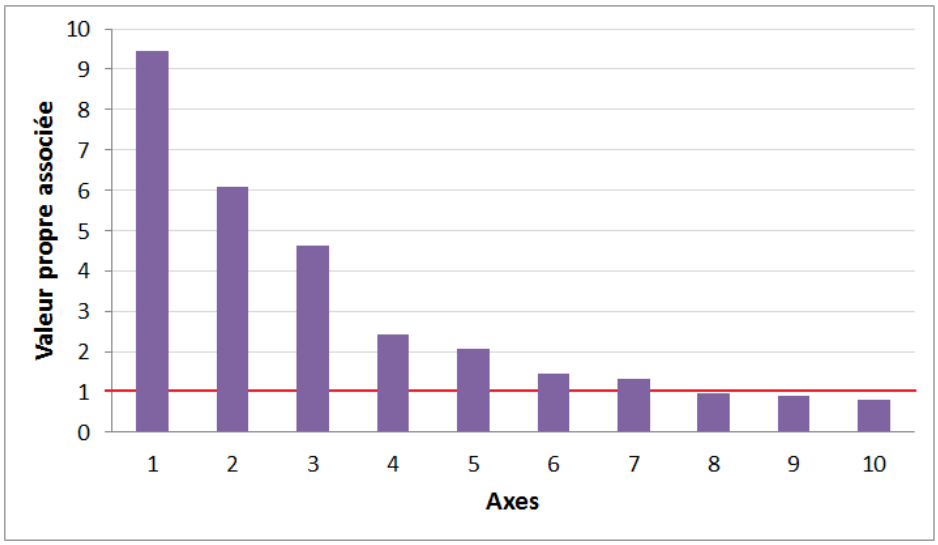

FIGURE *3.13 – Sélection du nombre d'axes à conserver par le critère de Kaiser.*

Ainsi à la vue de l'analyse de ces différents critères, nous décidons de garder les sept premières composantes principales qui représentent 74% de l'information contenue dans la base de données formée par les 37 variables initiales.

Interprétation des axes

Il peut être difficile de comprendre la signification d'un axe issu de l'Analyse en Composantes Principales, cependant il existe plusieurs indicateurs qui permettent d'établir une interprétation. Il est par exemple possible de visualiser les diagrammes des corrélations (voir annexe C) qui représentent les variables projetées sur les axes, mais à cause du nombre important de variables, l'analyse graphique est difficile. Ainsi, il est plus intéressant de se pencher sur l'indicateur de la contribution relative des variables à l'inertie d'un axe. Le package `factoextra` de R permet de réaliser l'Analyse en Composantes Principales mais également d'analyser les résultats. En particulier, il est possible de visualiser les contributions relatives des variables explicatives à l'inertie des axes.

Résultats

Suite à l'étude des trois critères de sélection du nombre d'axes à conserver, 7 axes ont été retenus dans cette Analyse en Composantes Principales, mais la compréhension de chacun de ces axes n'est pas aisée. Ainsi, l'interprétation des axes se limitera aux trois premiers. Les graphiques de contribution relative des variables pour les trois premiers axes sont disponibles dans l'annexe D. Un tableau synthétisant la contribution relative des variables pour chacun des 7 axes sélectionnés est également fourni dans cette annexe. Il apparaît que les variables qui contribuent le plus à l'inertie du premier axe sont les variables relatives au niveau d'études (CAP/BEP, diplôme supérieur). La densité de la population contribue également à cet axe. Concernant le deuxième axe, les variables qui participent le plus à l'inertie de l'axe sont celles relatives au revenu. À partir du troisième axe, l'interprétation devient plus compliquée car les variables qui possèdent le plus d'influence sur l'axe sont relativement différentes. On trouve notamment le type de résidence (résidence principale, secondaire ou de vacances) et le lieu de travail (dans la même commune, dans un autre département).

Bilan de l'analyse en composantes principales

L'Analyse en Composantes Principales a été utilisée dans le but de réduire le nombre de dimensions des 37 variables à la maille INSEE et à la maille IRIS ressorties dans le Gradient Tree Boosting de pré-sélection. À l'aide de seulement 7 axes, il est possible de conserver 74% de l'information contenue dans ces 37 variables. Cependant, l'interprétation des axes est relativement compliquée et d'un point de vue opérationnel, il est difficile d'expliquer la présence des composantes principales en tant

que variables explicatives dans le modèle si ces composantes ne sont pas totalement interprétables.

Néanmoins, ces composantes principales seront testées en les intégrant directement dans le modèle. Les calculs sont réalisés dans le chapitre suivant qui traite de l'intégration des variables dans le modèle de régression.

3.4.3 Mise en œuvre d'une sélection des variables par une procédure STEPWISE

Pour les variables provenant du prestataire externe fournissant des données à la maille ADRESSE, il s'agit de variables à la fois quantitatives et qualitatives, c'est pourquoi une procédure STEPWISE est utilisée afin de sélectionner les variables les plus explicatives parmi celles pré-sélectionnées par le modèle de Gradient Tree Boosting. L'utilisation de cette méthode permettra de comparer l'efficacité de sélection automatique des variables de l'algorithme de Gradient Tree Boosting avec la sélection réalisée par cette procédure.

Théorie

La procédure STEPWISE est une procédure de sélection de variables utilisant une régression pas à pas et basée sur un critère statistique. Les critères statistiques utilisés avec cette méthode sont généralement les critères AIC et BIC.

Le critère **AIC** ou *Akaïke Information Criterion* est un critère qui repose fondamentalement sur la minimisation de la perte d'information de Kullback-Leibler (K-L). D'après Burnham et Anderson [5], l'information de Kullback-Leibler peut être appréhendée comme une distance entre la réalité des données et le modèle utilisé pour leur approximation. Akaike (1973) a trouvé une relation qui liait cette information et la log-vraisemblance du modèle testé, en corrigeant cette dernière par le nombre de paramètres. Le critère AIC est alors défini de la manière suivante :

$$AIC = -2LL + 2q,$$

où LL représente la log-vraisemblance maximisée du modèle et q le nombre de paramètres du modèle. Il s'agit donc d'un arbitrage entre biais (qui diminue avec le nombre de paramètres) et variance (décrire les données avec le plus petit nombre de paramètres possible). Ainsi, plus le modèle contient de variables, plus le modèle est pénalisé.

En considérant qu'il existe N modèles $m_1, ..., m_N$ à comparer et que les valeurs individuelles de leur AIC sont respectivement notées $AIC_1, ...AIC_N$, il est plus cohérent de considérer les valeurs :

$$\Delta_i = AIC_i - AIC_{\min},$$

qui sont plus faciles à interpréter car elles ne sont pas dépendantes des constantes arbitraires présentes dans les valeurs individuelles. Le meilleur modèle sera alors celui dont $\Delta = 0$ tandis que les autres modèles posséderont des valeurs strictement positives.

Le critère **BIC** ou *Bayesian Information Criterion* est une variante du critère AIC au sens où il prend également en compte le nombre n d'observations du modèle :

$$BIC = -2LL + \log(n)q.$$

La taille de l'échantillon constitue donc un élément supplémentaire qui rend la pénalisation plus importante.

Le critère AIC sera utilisé pour la mise en application de la procédure STEPWISE de régression pas à pas.

Il existe deux types de méthodes pour les régressions pas à pas :

La méthode ascendante ou FORWARD Il s'agit d'une régression où le modèle initial ne possède souvent qu'une seule variable puis, à chaque itération, la variable qui améliore le plus le critère AIC est intégrée au modèle.

La méthode descendante ou BACKWARD Il s'agit d'une régression où le modèle initial comprend toutes les variables candidates puis, à chaque itération, la variable qui dégrade le plus le critère AIC est retirée du modèle.

La méthode STEPWISE est une méthode ascendante améliorée puisqu'à chaque ajout de variable dans le modèle, la méthode teste s'il y a des variables qui devraient être retirées.

Mise en application

Dans l'étude réalisée, la méthode est implémentée à l'aide de la fonction `stepAIC` du package `MASS` de R. La procédure STEPWISE est utilisée sur un modèle initial composé uniquement de la variable offset « exposition » et les variables testées sont les variables à la maille ADRESSE. A chaque itération, un modèle linéaire généralisé est crée et le critère AIC est calculé. Ainsi la sélection de variables est réalisée pour un modèle linéaire généralisé mais nous utiliserons les variables sélectionnées dans le modèle de Gradient Tree Boosting. Une comparaison des modèles linéaires généralisés et du modèle de Gradient Tree Boosting est réalisée dans le chapitre suivant.

Par ailleurs, la procédure a également été appliquée sur les données à la maille IRIS provenant du même prestataire externe que les données à la maille ADRESSE mais les résultats obtenus n'étaient pas concluants puisqu'ils dégradaient la qualité du modèle. Par conséquent, ces variables à la maille IRIS seront conservées comme telles.

Bilan et transition

Ce chapitre traite de la sélection des variables et de leur analyse. Étant donnée la quantité importante de variables dont nous disposions, une pré-sélection des variables était fortement requise. De manière à réduire le nombre de variables candidates à intégrer le modèle de régression, chaque ensemble de variables provenant des différentes sources est passé par une pré-sélection à l'aide d'un Gradient Tree Boosting. Les variables pré-sélectionnées présentaient néanmoins de fortes corrélations avec les variables provenant de la même source, c'est pourquoi différentes techniques ont été utilisées pour procéder à une seconde étape de sélection sur chaque groupe de variables.

Pour les variables issues de l'Observatoire National de la Délinquance et des Réponses Pénales, une sélection de variables en fonction de l'influence relative exercée sur la variable réponse a notamment été mise en œuvre. Quant aux variables quantitatives à la maille IRIS fournie par un des prestataires, une Analyse en Composantes Principales a été réalisée afin de réduire la dimension du nombre de variables tout en conservant 74% de l'information contenue dans les variables initiales. Grâce à cela, il a été possible de synthétiser l'information et d'éliminer les corrélations existantes pour cet ensemble de variables. Enfin, pour les variables à la maille ADRESSE et à la maille IRIS fournies par le second prestataire, plusieurs procédures STEPWISE ont été réalisées afin de sélectionner un nombre plus restreint de variables.

La pré-sélection a permis d'éliminer les variables inutiles, quant à la deuxième étape de sélection et de traitement des variables, elle a été mise en œuvre dans le but de tester la qualité de la sélection automatique des variables réalisée par l'algorithme du Gradient Tree Boosting. Les résultats de ce test sont énoncés et analysés dans le chapitre suivant. Dans la suite de l'étude, l'intérêt est porté par l'intégration des variables sélectionnées dans la régression et par l'analyse géographique des résidus issus du modèle réalisé.

Chapitre 4

Intégration dans la régression des variables influentes et analyse géographique des résidus

Ce chapitre met en œuvre l'intégration dans la régression des variables sélectionnées dans les étapes précédentes. Cette régression est réalisé par l'intermédiaire d'un modèle de Gradient Tree Boosting, qui en plus de posséder des qualités de sélection de variables, est également un modèle de prédiction.

Les modèles de prédiction généralement utilisés en assurances dommages sont les modèles linéaires généralisés, c'est pourquoi nous commencerons par énoncer la théorie de ces modèles avant de réaliser une comparaison qui permettra de mettre en parallèle le modèle de Gradient Tree Boosting et les modèles linéaires généralisés. L'efficacité de la sélection automatique des variables réalisée par l'algorithme du Gradient Tree Boosting sera ensuite remise en question en testant la performance de l'algorithme avec et sans les différentes traitements des variables réalisés dans le chapitre précédent. Enfin, le modèle final conservé sera examiné afin de réaliser la suite de l'étude puis nous nous pencherons sur l'analyse géographique des résidus.

4.1 Étude comparative d'un modèle linéaire généralisé et d'un Gradient Boosting

4.1.1 Les modèles linéaires généralisés

Le modèle linéaire généralisé (*Generalized Linear Model* ou *GLM* en anglais) est très répandu en assurances dommages [10][25]. Il permet d'étudier la relation entre une variable réponse Y et un ensemble de variables explicatives $X_1, X_2, ..., X_p$.

Le modèle linéaire généralisé est une extension de la régression linéaire, au sens où la relation qui lie l'espérance de la variable réponse et les variables explicatives n'est plus seulement linéaire mais est expliquée à l'aide d'une fonction de lien.

Rappelons tout d'abord que la régression linéaire consiste à estimer le paramètre β tel que

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

sous l'hypothèse que ϵ suit une loi normale centrée et de variance σ^2.

Ainsi, dans ce modèle de régression linéaire, Y suit une loi normale de moyenne $X\beta$ et de variance σ^2. Cette hypothèse de normalité de la variable réponse Y est peu recevable dans certains cas et notamment en assurance lorsqu'il s'agit de modéliser une fréquence, un coût moyen ou encore une prime pure. En effet, cela supposerait d'avoir des valeurs négatives ce qui ne correspond pas à la réalité des données. Les modèles linéaires généralisés permettent alors de s'affranchir de cette hypothèse en autorisant d'autres lois pour la variable réponse sous réserve qu'elles appartiennent à la famille exponentielle.

Un modèle linéaire généralisé se définit par trois composantes :

La composante aléatoire : la variable réponse

La variable réponse Y est la variable aléatoire que l'on cherche à expliquer. Le principe est d'émettre une hypothèse sur la loi suivie par Y à partir des n observations $y_1, y_2, ..., y_n$ dont on dispose. Cette loi doit appartenir à la famille exponentielle. La famille exponentielle englobe pratiquement toutes les lois connues (Bernouilli, Poisson, Normale, Gamma,...). La densité des lois appartenant à la famille exponentielle doit s'écrire sous la forme suivante :

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

avec

- $b(\cdot)$ une fonction de classe C^3 et de dérivée première inversible,
- $a(\cdot)$ [1] et $c(\cdot)$ des fonctions dérivables,
- θ le paramètre naturel lié aux deux premiers moments de la loi,
- ϕ le paramètre de dispersion.

Deux propriétés concernant le paramètre naturel θ découlent de l'écriture de la densité pour les lois appartenant à la famille exponentielle :

$$\mathbf{E}(Y) = b'(\theta)$$

$$\mathbf{V}(Y) = b''(\theta)\phi.$$

1. En général, on a $a(\phi) = \phi$.

Dans l'étude réalisée, l'intérêt porte tout particulièrement sur la loi de Poisson de paramètre λ qui modélise le nombre de sinistres, de densité :

$$\begin{aligned} f_\lambda(y) &= \exp(-\lambda)\frac{\lambda^y}{y!} \\ &= \exp(y\ln(\lambda) - \lambda - \ln y\,!), \end{aligned}$$

par rapport à la mesure de comptage, avec $y \in \mathbb{N}$, $\theta = \ln\lambda$, $a(\phi) = \phi = 1$, $b(\theta) = \exp(\theta) = \lambda$, et $c(y,\phi) = -\ln y\,!$.

La composante déterministe : les variables explicatives

Les variables explicatives à intégrer dans le modèle sont à choisir avec soin. En effet, le nombre de variables utilisé ne doit pas être trop important pour que le modèle soit utilisable en pratique mais il doit être suffisant pour que le modèle soit cohérent et performant. C'est pourquoi, il est nécessaire de sélectionner parmi toutes les variables à disposition, les variables dont le pouvoir explicatif est le plus important afin de réussir à obtenir un juste équilibre.

La fonction de lien

Il s'agit de la fonction qui va contenir le lien entre l'espérance de la variable réponse et les variables explicatives. Dans le cas des modèles linéaires simples, cette fonction de lien n'est autre que la fonction identité et le modèle est additif. Dans les modèles linéaires généralisés, au lieu de modéliser l'espérance de la variable réponse directement, une fonction de lien monotone et dérivable g est introduite de sorte que :

$$g(\mathbf{E}[Y]) = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p.$$

En pratique, la fonction de lien souvent utilisée est la fonction logarithme car elle permet d'obtenir un modèle multiplicatif : $g(x) = \ln(x)$. On obtient alors :

$$\mathbf{E}[Y] = \exp(\beta_0) \times \exp(\beta_1 X_1) \times ... \times \exp(\beta_p X_p).$$

L'estimation des paramètres $\beta_0, \beta_1, ..., \beta_p$ du modèle linéaire généralisé peut se faire par la méthode du maximum de vraisemblance à partir des données $y_1, .., y_n$. L'estimation de l'espérance de la variable aléatoire Y en découle alors à l'aide du calcul :

$$\mathbf{E}[Y] = g^{-1}(\beta_0 + \beta_1 X_1 + ... + \beta_p X_p).$$

Avec l'hypothèse d'indépendance des Y_i pour $i = 1, ..., n$, la log-vraisemblance du modèle est :

$$\mathscr{L}(y,\beta) = \sum_{i=1}^{n} ln(f_{\theta,\phi}(y_i)) = \sum_{i=1}^{n} \underbrace{\frac{y_i\theta_i - b(\theta_i)}{\phi} + c(y_i,\phi)}_{\mathscr{L}_i(\beta)}.$$

Bien que les paramètres β n'apparaissent pas directement dans la formule de la log-vraisemblance, leur présence est implicite et s'explique par les égalités suivantes :

$$\mathbf{E}[Y_i] = \mu_i = b'(\theta_i) = g^{-1}(X_i^T\beta).$$

La condition nécessaire pour être à l'optimum impose d'annuler la dérivée première de la log-vraisemblance $\frac{\partial\mathscr{L}}{\partial\beta}$. Considérons d'abord la log-vraisemblance $\mathscr{L}_i$ de l'observation i où, pour tout $j = 1,...,p$, on a :

$$\frac{\partial\mathscr{L}_i}{\partial\beta_j} = \frac{\partial\mathscr{L}_i}{\partial\theta_i}\frac{\partial\theta_i}{\partial\mu_i}\frac{\partial\mu_i}{\partial\beta_j}.$$

Décomposons chacune des dérivées partielles précédentes :

$$\begin{aligned}
&\frac{\partial\mathscr{L}_i}{\partial\theta_i} = \frac{y_i - b'(\theta_i)}{\phi},\\
&\frac{\partial\theta_i}{\partial\mu_i} = \left[\frac{\partial\mu_i}{\partial\theta_i}\right]^{-1} = \left[b"(\theta_i)\right]^{-1} = \left[\frac{\mathbf{V}(Y_i|X=X_i)}{\phi}\right]^{-1},\\
&\frac{\partial\mu_i}{\partial\beta_j} = X_{ij}(g^{-1})'(X_i^T\beta).
\end{aligned}$$

Finalement,

$$\frac{\partial\mathscr{L}_i}{\partial\beta_j} = \frac{y_i - b'(\theta_i)}{\mathbf{V}(Y_i|X=X_i)}(g^{-1})'(X_i^T\beta)X_{ij}, \tag{4.1}$$

et ainsi, pour tout $j = 1,...,p$:

$$\frac{\partial\mathscr{L}}{\partial\beta_j} = \sum_{i=1}^{n}\frac{y_i - b'(\theta_i)}{\mathbf{V}(Y_i|X=X_i)}(g^{-1})'(X_i^T\beta)X_{ij}. \tag{4.2}$$

Les équations de vraisemblance étant transcendantes, une solution pour approcher l'estimateur du maximum de vraisemblance est d'utiliser des procédures itératives d'optimisation telles que l'algorithme de Newton-Raphson ou l'algorithme IRLS (Iterative Reweighted Least Squares).

La démarche de telles procédures est la suivante :

1. Choisir un β^0 comme point de départ,
2. Puis $\beta^{k+1} = \beta^k + A_k\nabla\mathscr{L}(\beta^k)$, avec $k \in \mathbb{N}$ et où :

   Pour l'algorithme de Newton-Raphson $A_k = -\left[\nabla^2,\mathscr{L}(\beta^k)\right]^{-1}$

Pour l'algorithme IRLS $A_k = -\left[\mathbb{E}_\beta \nabla^2 \mathscr{L}(\beta^k)\right]^{-1}$.

3. L'algorithme s'arrête lorsque $\beta^{k+1} \approx \beta^k$ ou bien lorsque $\mathscr{L}(\beta^{k+1}) \approx \mathscr{L}(\beta^k)$.

Prise en compte de l'exposition [7] [29]

Lors de la tarification d'un contrat, il faut pouvoir prédire le nombre de sinistres qui surviendront, en moyenne, l'année suivante. Cependant, toutes les polices du portefeuille ne sont pas nécessairement observées sur un an mais certaines sur six mois ou neuf mois. Il s'agit généralement des contrats correspondant à des affaires nouvelles ou à des résiliations.

Il est important de tenir compte de ce paramètre pour ne pas modéliser de la même façon deux risques similaires qui auraient été exposés sur des durées distinctes. Une approximation souvent utilisée est de faire l'assomption forte que le risque est linéaire suivant l'exposition, ce qui permet de rationaliser sur une même assiette de temps le risque constaté. Afin de prendre en considération cet effet pour expliquer la variable aléatoire N d'espérance λ représentant le nombre de sinistres, il est possible d'intégrer l'exposition e du contrat (la durée d'observation mesurée en années) dans la régression qui utilise un modèle poissonnien et une fonction de lien logarithme. En effet, comme l'espérance de la variable considérée devient donc λe, la régression s'écrit alors :

$$\begin{aligned}\mathbf{E}[N|X,e] &= e \times \Big(\exp(\beta_0) \times \exp(\beta_1 X_1) \times \ldots \times \exp(\beta_p X_p)\Big) \\ &= \exp(X\beta + \ln(e)).\end{aligned}$$

Ainsi, la prise en compte de l'exposition e correspond simplement de l'ajout d'une variable explicative au sein du modèle dont le coefficient associé β est connu, fixé à 1 et n'a pas besoin d'être estimé. La variable $x_{p+1} = \ln(e)$ ainsi ajoutée s'appelle une variable **offset**.

4.1.2 Performances d'un Gradient Boosting par rapport à un modèle linéaire généralisé

Les modèles linéaires généralisés sont les modèles les plus couramment utilisés en assurance dommages, cependant les modèles de Gradient Boosting offrent une bonne alternative car leur pouvoir prédictif est intéressant tout comme la manière de les interpréter.

Dans le monde de l'actuariat, les modèles doivent être communiqués, compris et approuvés par un public non expert en statistiques et travaillant quelques fois dans d'autres domaines. Certes l'utilisation d'un modèle de Gradient Boosting pour traiter ce genre de sujet de tarification est innovant, cependant il faut avant tout prendre le

temps d'expliquer les principes sous-jacents de cette méthode afin de convaincre de son efficacité.

Comparaison théorique des deux modèles

Modèle linéaire généralisé :

Les modèles linéaires généralisés sont, par essence des modèles linéaires relativement simples et sont donc contraints sur la classe des fonctions qu'ils peuvent estimer. Un modèle linéaire généralisé sera donc très performant lorsque qu'il existe des relations linéaires mais insuffisant pour des relations non-linéaires. De plus, pour obtenir des résultats optimaux, les variables explicatives doivent être le plus indépendantes possible et il faut donc procéder à une sélection des variables avant de les intégrer dans le modèle en s'assurant également de traiter les valeurs manquantes.

Le véritable intérêt d'un modèle linéaire généralisé provient de sa facilité d'interprétation : notamment dans le cas d'un modèle multiplicatif lorsque la fonction de lien est la fonction logarithme. En effet, les coefficients $\beta_0, \beta_1, ..., \beta_p$ alors obtenus peuvent être analysés rapidement :

Le coefficient β_0 observé dans les précédentes formules correspond à l'intercept, c'est à dire au profil de référence. Ce profil regroupe l'ensemble des modalités les plus représentées des variables explicatives.

Puis, les β_j, $j \in \{1, ...p\}$ peuvent s'interpréter de la façon suivante :

- Si $\beta_j > 0$, alors l'individu possédant la modalité x_j a tendance à avoir une sinistralité plus importante que l'individu issu du profil de référence.
- Si $\beta_j < 0$, alors l'individu possédant la modalité x_j a tendance à avoir une sinistralité moins importante que l'individu issu du profil de référence.

Modèle de Gradient Tree Boosting :

Le modèle de Gradient Tree Boosting est une alternative intéressante au modèle linéaire généralisé car c'est un modèle non paramétrique qui ne nécessite généralement que de très peu de traitement de données. En effet, la sélection des variables est incorporée directement dans l'algorithme qui analyse également les interactions entre les variables ce qui permet de consacrer moins d'efforts au traitement des données pour obtenir un modèle satisfaisant.

Par ailleurs, la gestion des valeurs manquantes ou inconnues est très bien gérée de manière à ne pratiquement pas perdre d'information [14]. En effet, un utilisant un boosting sur l'arbre de décision CART, les valeurs manquantes ne sont ni supprimées, ni remplacées par la valeur la plus probable mais sont classées de manière cohérente. En effet, après avoir choisi la variable et sa valeur de séparation pour créer un nœud, l'algorithme sélectionne parmi les variables non utilisées pour créer le nœud celles qui expliquent le mieux la séparation ainsi créée.

Par exemple, si la séparation du nœud a lieu à un nombre de pièces égal à 3 et que la seconde variable qui explique le mieux cette séparation est la qualité d'habitation ; si le nombre de pièces n'est pas renseigné pour un contrat, l'algorithme regardera la valeur de la qualité d'habitation et classera l'observation dans la branche qui convient en fonction de la qualité d'habitation. Ainsi, aucune observation n'est laissée de côté, elles sont toutes utilisées de manière à construire le meilleur modèle possible.

Un autre avantage du modèle de Gradient Tree Boosting est qu'il peut être appliqué pour des problèmes de classification ou de régression sur une grande variété de distributions pour la variable réponse et il s'ajuste bien pour des relations non linéaires. Enfin, grâce aux graphiques d'importance relative et de dépendance partielle des variables, et grâce à la manière de présenter les interactions complexes, les résultats issus d'un tel modèle sont compréhensibles et exploitables.

Comparaison pratique des deux modèles

Nous décidons de comparer le modèle linéaire généralisé avec un Gradient Tree Boosting avec -pour l'exemple- un shrinkage à 0,03 et un *bagging* à 70%. De manière à réaliser une comparaison cohérente entre les deux modèles, la durée d'observation des contrats en portefeuille est prise en compte dans les deux modèles en mettant la variable d'exposition en *offset*. Par ailleurs, de façon à faire fonctionner correctement le modèle linéaire généralisé log-poissonien, les variables ont été préalablement sélectionnées de manière à supprimer les corrélations trop importantes.

Les résultats que nous évoquerons ci-après sont disponibles dans un tableau récapitulatif dans l'annexe B. Il apparaît que les deux modèles soient sensiblement performants bien que le Gradient Tree Boosting semble être légèrement moins bon en terme d'indice de GINI dans certains cas. Néanmoins les Root Mean Squared Error sur l'échantillon test sont toujours meilleurs dans le modèle de Gradient Tree Boosting.

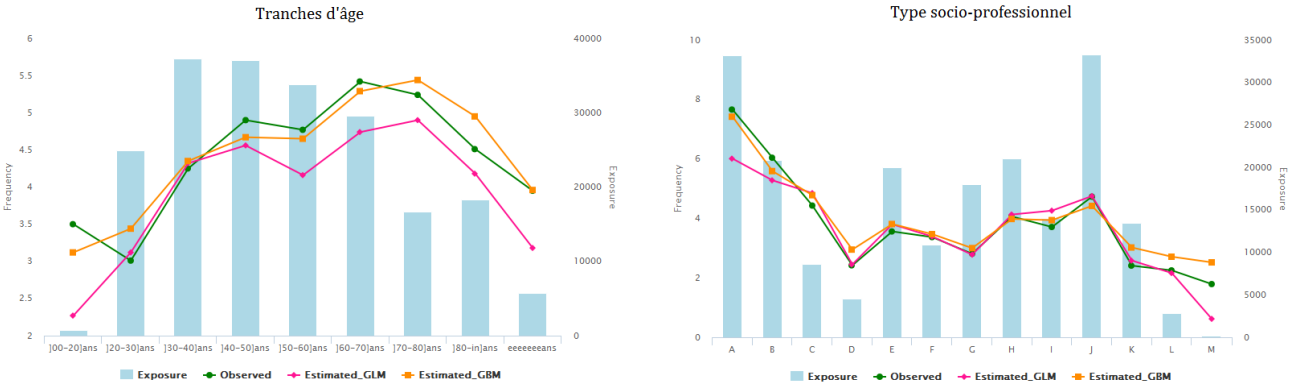

FIGURE *4.1 – Comparaison des prédictions réalisées sur l'échantillon test par les deux modèles pour une variable interne (à gauche) et pour une variable externe (à droite).*

Cette dernière constatation peut se vérifier en regardant la figure 4.1 qui exhibe les prédictions réalisées par les deux modèles sur l'exemple de deux variables et qui atteste du fait que les fréquences prédites par le Gradient Tree Boosting (courbes oranges) sont plus semblables aux fréquences observées (courbes vertes) que les fréquences prédites par le modèle linéaire généralisé log-poissonien (courbes roses). Il s'agit ici d'une variable client, la tranche d'âge, et d'une variable externe à la maille IRIS qui représente la typologie de population (élites parisiennes, étudiants, etc...) la plus représentée au sein de l'IRIS. On remarque notamment que pour la tranche d'âge, en plus de s'éloigner de la fréquence observée, le modèle linéaire généralisé sous-estime fortement le risque. Ainsi, cette étude comparative met en évidence la qualité de prédiction du modèle de Gradient Tree Boosting.

Par ailleurs, pour exploiter la quantité importante de données disponibles dans cette étude et afin de préserver la richesse d'information qu'elles contiennent, nous retenons pour la suite, l'utilisation de techniques d'apprentissage statistique plus automatisées telles que le Gradient Tree Boosting plutôt qu'un modèle linéaire généralisé qui nécessite une sélection plus « manuelle » des données.

4.2 Mise en œuvre de la régression en intégrant les variables sélectionnées

La première partie de l'étude du signal géographique est réalisée dans cette section en intégrant les variables « internes » avec certaines variables externes dans un modèle de Gradient Tree Boosting.

Les modèles ci-après sont réalisés sans intégrer les variables internes relatives aux montants assurés, qui, nous le rappelons sont fortement biaisées car sous-estimées par les assurés. Cette initiative a été prise après avoir vérifié que le modèle réalisé sans intégrer ces variables n'était pas dégradé. En effet, grâce aux variables externes incorporées directement dans le modèle, les performances avec et sans l'intégration des variables relatives aux montants assurés restent pratiquement identiques. Cela témoigne bien, une fois encore, du pouvoir prédictif des variables externes.

4.2.1 Analyse de la qualité de sélection automatique des variables par le Gradient Boosting

Comme il a pu être énoncé dans les parties précédentes, le Gradient Boosting permet de réaliser une sélection automatique des variables en prenant en compte les potentielles interactions entre les variables. Nous mettons cependant en œuvre une comparaison d'un modèle de Gradient Tree Boosting en intégrant toutes les variables pré-sélectionnées avec d'autres Gradient Tree Boosting où les variables intégrées sont

passées à travers une deuxième étape de sélection. Les différents procédés de sélection ont été mis en œuvre dans le chapitre 3 et concernent notamment une sélection de variables par une procédure STEPWISE ou encore une orthogonalisation des variables à l'aide d'une Analyse en Composantes Principales. La table ci-après permet de comparer la performance des différents modèles de Gradient Boosting avec et sans ce traitement des variables.

Modèles	**GINI Train**	**GINI Test**	**Delta GINI**	**RMSE Train**	**RMSE Test**
Avec sélection d'une seule variable criminalité	34,4%	33,1%	1,3	0,1291062	0,1306758
Avec ACP sur les variables à la maille IRIS	34,6%	33,1%	1,5	0,1291055	0,1306754
Avec STEPWISE données internes et à la maille adresse	34,0%	34,0%	0,0	0,1291074	0,1306703
Avec tous les traitements	34,2 %	33,7%	0,5	0,1291070	0,1306719
Avec la sélection automatique du GBM	34,4 %	33,4%	1	0,1291058	0,1306743

TABLE *4.1 – Mesure de l'impact de la sélection de variables.*

La table 4.1 montre que le travail de sélection automatique réalisé par modèle de Gradient Tree Boosting reste très satisfaisant même s'il peut être légèrement amélioré.

Il apparaît que le fait de conserver uniquement une variable issue de l'Observatoire National de la Délinquance et des Réponses Pénales dégrade légèrement le modèle puisque l'indice de GINI sur l'échantillon test diminue mais cela reste convenable.

Quant à la diminution de performance du modèle avec l'Analyse en Composantes Principales réalisée sur les variables à la maille IRIS, elle était relativement prévisible puisque les composantes principales conservées ne représentent que 74% de l'information totale contenue dans les variables initiales. Par ailleurs, il était également possible de sélectionner, parmi les variables initiales, celles qui contribuaient le plus aux axes, de manière à conserver l'interprétabilité de ces variables. Un modèle conservant uniquement les 7 variables initiales qui contribuaient le plus à l'inertie des axes sélectionnés a été réalisé et, les performances observées avec un tel modèle se rapprochent de celles réalisées avec le modèle contenant les composantes principales mais sont légèrement inférieures.

En revanche, la sélection provenant de la procédure STEPWISE réalisée sur les variables internes et les variables à la maille ADRESSE stabilise l'indicateur de GINI.

En conclusion, pris individuellement, ces traitements sur les variables détériorent ou améliorent le modèle. Cependant, en intégrant l'Analyse en Composantes Princi-

pales, la procédure STEPWISE et la sélection sur la variable criminalité dans un même modèle de Gradient Tree Boosting, la performance est tout de même améliorée puisque les Root Mean Squared Error sont légèrement inférieurs et que l'indicateur de GINI est plus important sur l'échantillon test réduisant ainsi l'écart en terme de GINI. Ainsi, à la vue des résultats précédents, et bien que l'efficacité de la sélection automatique des variables du modèle de Gradient Boosting soit vérifiée, nous décidons tout de même de conserver pour la suite de l'étude le modèle réalisé après tous les traitements de sélection des variables (Analyse en Composantes Principales, procédure STEPWISE,...).

Il est bon de constater que d'autres méthodes de validation de modèle auraient pu être mises en œuvre afin de chercher à optimiser davantage les paramètres de l'algorithme de Gradient Tree Boosting, cependant ce n'est pas l'objectif de ce mémoire qui se concentre principalement sur le traitement de l'aspect géographique du risque. Par exemple, la méthode des *K-folds* fait partie de ces techniques de validation de modèle qui permettent de s'assurer de la robustesse de la prédiction [14]. Le principe de la méthode de validation croisée des K-folds est de partitionner la base de départ en K sous bases de tailles égales puis, le modèle est calibré K fois sur (K-1) partitions (échantillon d'apprentissage) et validé sur la $K^{ème}$. De cette façon, chacune des K partitions servira de base de validation (échantillon test).

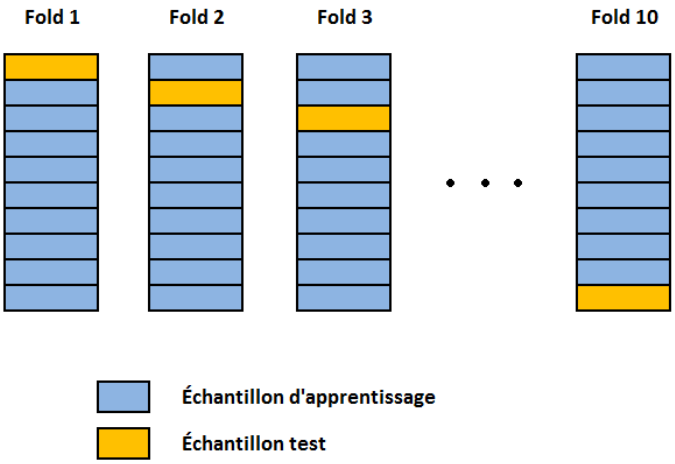

FIGURE 4.2 – *Illustration de la méthode de validation des K-folds.*

Cette technique a été appliquée en utilisant 10 *folds* sur le modèle de Gradient Tree Boosting sélectionné pour la suite de l'étude. Elle révèle que les indicateurs de performances du modèle restent relativement stables même lorsqu'il est calibré puis testé sur différentes partitions de la base d'étude.

4.2.2 Modèle final utilisé pour le processus de construction du micro-zonier

Le modèle de Gradient Tree Boosting utilisé pour réaliser la construction du micro-zonier est donc un modèle contenant 56 variables dont :

- 7 variables internes
- 1 variable à la maille département issue de l'Observatoire National de la Délinquance et des Réponses Pénales ayant le plus d'influence sur la fréquence VOL appartement dans le modèle de Gradient Tree Boosting,
- 7 composantes principales issues de l'Analyse en Composantes Principales réalisée sur les variables quantitatives à la maille IRIS fournies par le premier prestataire externe ainsi que les variables à la maille INSEE,
- 25 autres variables à la maille IRIS fournies par le deuxième prestataire,
- 16 variables à la maille ADRESSE sélectionnées par la procédure STEPWISE et fournies par le deuxième prestataire également.

Interprétation des résultats

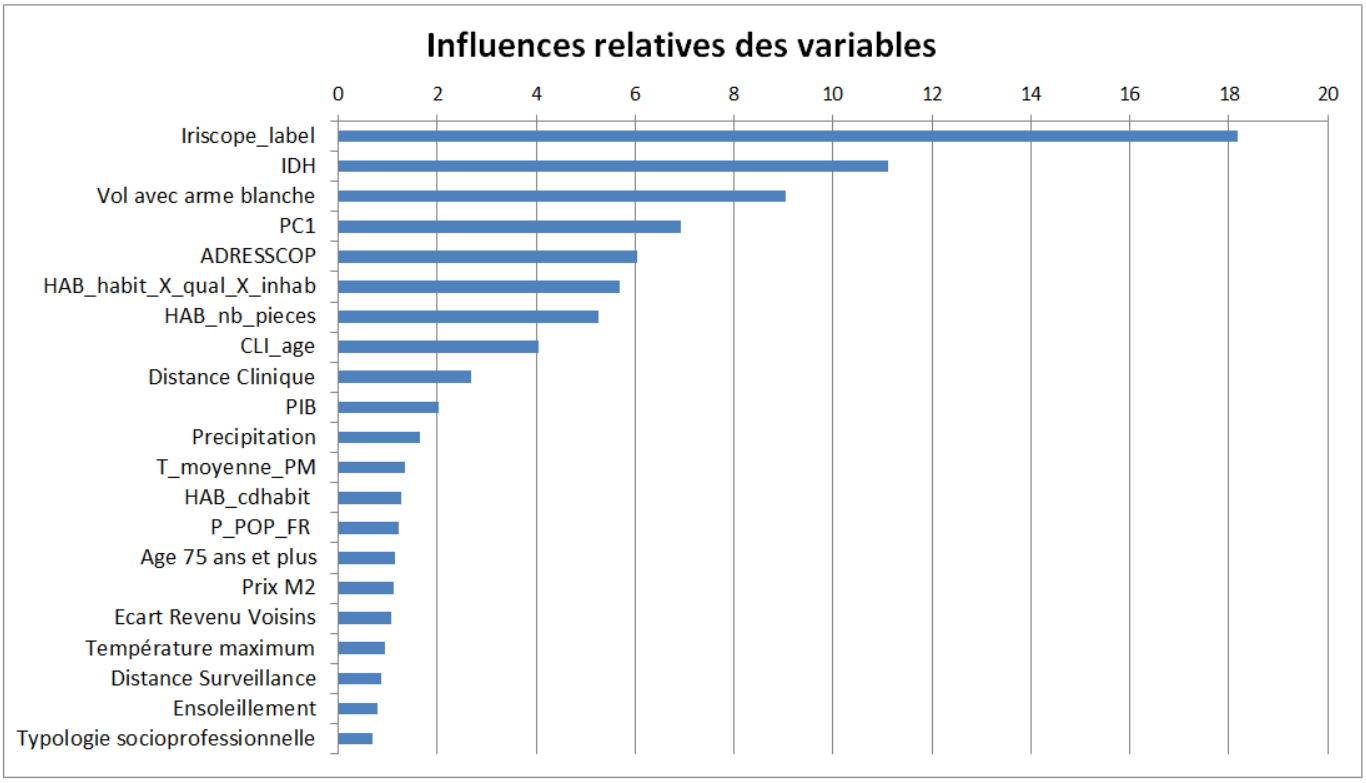

FIGURE 4.3 – *Influences relatives des variables explicatives dans le modèle de prédiction de la fréquence de survenance d'un sinistre VOL sur les appartements.*

L'analyse de la figure 4.3 permet de mettre en évidence l'influence relatives des variables dans la création du modèle.

La variable « iriscope_label » ayant la plus grande importance relative concerne une typographie d'adresse agrégée à la maille IRIS, tandis que la variable « ADRESSCOP » fait également référence à une typographie d'adresse mais à la maille adresse.

Il est intéressant de constater que la première composante principale issue de l'Analyse en Composantes Principales possède une influence relative importante puisqu'elle est dans les premières variables de la figure. Rappelons que 7 axes ont été sélectionnés à l'issue de ce procédé, cependant si seule la première composante possède une influence relative importante, il est alors possible de retirer du modèle les axes dont la signification est moins aisée.

Par ailleurs, les variables internes au contrat ne ressortent pas en premier comme on pourrait le penser. Ceci montre une fois encore l'importance des données externes. Mais il faut noter qu'il s'agit ici d'un modèle où les montants assurés ont été retirés étant donné leur manque d'objectivité.

Analysons à présent les graphiques de dépendances partielles des variables les plus influentes qui permettent d'observer la fréquence VOL en fonction de chaque variable explicative. Seules certaines variables parmi les plus influentes seront étudiées.

Variables internes

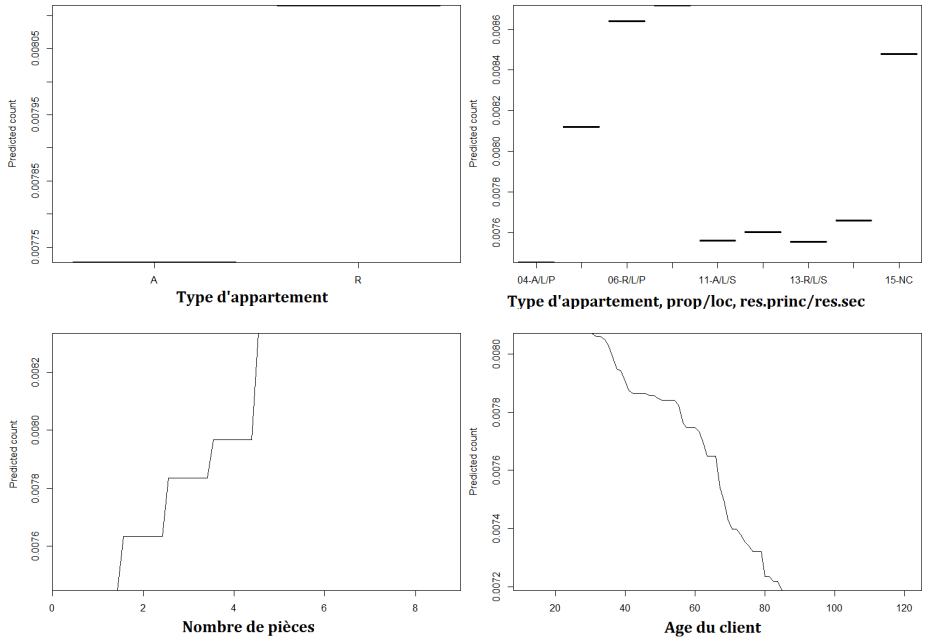

FIGURE *4.4 – Graphiques des dépendances partielles pour les variables internes.*

L'influence de certaines variables sur la fréquence VOL avait déjà été étudiée lors de l'analyse descriptive préliminaire. Les mêmes tendances que celles observées lors

de cette analyse apparaissent ici, à savoir que plus le nombre de pièces de l'appartement est élevé plus l'appartement est exposé au risque.

Concernant l'âge du client, il semblerait que le risque diminue avec l'âge. Il est possible que les individus aient tendance à mieux protéger leur habitation avec l'âge et qu'ils soient plus souvent chez eux, ce qui peut avoir un effet dissuasif pour les malfaiteurs. Concernant les autres variables internes, il est relativement normal de constater que les appartements situés au rez-de-chaussée sont plus exposés au risque que les autres. De plus, les propriétaires semblent plus exposés au risque que les locataires et les résidences secondaires paraissent moins risquées que les résidences principales.

Variables externes

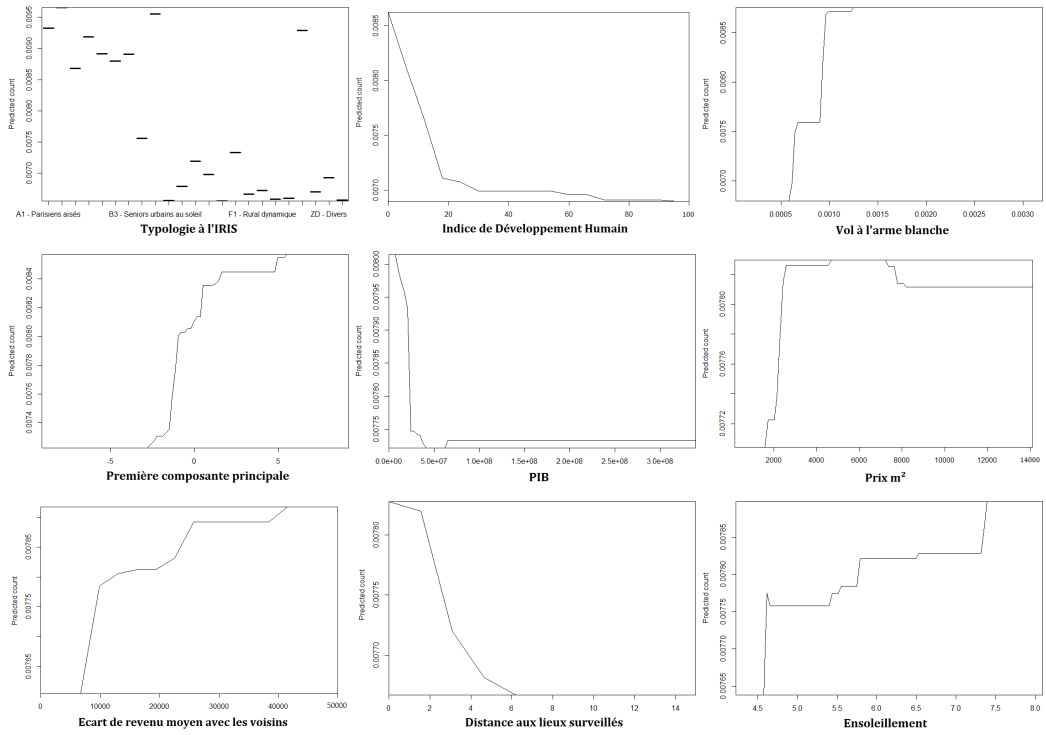

FIGURE *4.5 – Graphiques des dépendances partielles pour les variables externes.*

Les tendances observées lors de l'analyse préliminaires pour l'indice de développement humain ainsi que pour les variables d'écart de revenu moyen et de vols à l'arme blanche se retrouvent également ici.

Il est intéressant de constater également que la fréquence VOL en appartement possède une tendance croissante avec la première composante principale issue de l'Analyse en Composantes Principales, qui, nous le rappelons, est principalement constituée par les variables de densité de population et des proportions de diplômés dans la population.

Par ailleurs, il semblerait que la survenance d'un sinistre VOL en appartement soit plus importante dans des lieux de type A (milieu urbain dans la classification Iri-

scope_label) que dans les autres lieux. La fréquence semble également diminuer avec la distance à l'endroit surveillé le plus proche. Cette analyse de la décroissance du risque avec la proximité aux lieux surveillés semble un peu contre-intuitive mais il faut peut-être considérer la relation dans l'autre sens, c'est à dire que les endroits surveillés auront tendance à être aménagés aux endroits où il y a le plus de vols. Enfin, la fréquence possède une tendance décroissante avec le PIB mais croissante avec le prix du mètre carré (en 2014) et malgré son influence relativement faible dans la création du modèle, l'ensoleillement possède un effet marginal croissant sur la survenance du sinistre VOL.

Comparaison avec l'ancien modèle

La table 4.2 permet d'effectuer une comparaison entre les performances du traitement de l'information géographique réalisé l'année passée et celui construit dans ce mémoire.

Rappelons que l'ancien modèle a été construit à partir d'un modèle linéaire généralisé réalisé uniquement sur les variables relatives au risque et à l'assuré. Le zonier, qui contient l'ensemble de l'information géographique, est ensuite créé à l'aide de forêts aléatoires sur les variables externes pour expliquer les résidus du modèle. L'ancien modèle était très instable puisqu'il présentait un sur-apprentissage important amenant à un écart de 12.5 en terme d'indice de GINI. Il faut noter que le modèle ainsi que le zonier ont été réalisés sur une base de données différente de celle utilisée dans l'étude ici présente.

La table ci-dessous met en perspective les performances d'un modèle de Gradient Tree Boosting intégrant les variables internes et le zonier Voronoï de l'année dernière avec le modèle de Gradient Tree Boosting retenu dans l'étude de ce mémoire, réalisé en intégrant directement les variables internes et externes au modèle.

Modèle	**GINI Train**	**GINI Test**	**Delta GINI**
Modèle de Gradient Boosting avec zonier Voronoï	32,2%	31,3%	0,9
Nouveau modèle avec variables internes et externes	34,2%	33,7%	0,5

TABLE *4.2 – Comparaison des deux méthodes qui intègrent les données externes géographiques de façons différentes.*

Il semblerait que l'utilisation d'un modèle de Gradient Tree Boosting sur les variables internes et le zonier Voronoï soit beaucoup plus stable que lors de l'utilisation d'un modèle linéaire généralisé log-poissonnien puisque l'écart en terme de GINI est réduit considérablement. Par ailleurs, le nouveau modèle crée dans cette étude, basé sur l'intégration d'une partie du signal géographique directement dans le modèle de

Gradient Tree Boosting, est le modèle le plus stable tout en améliorant la qualité de segmentation.

Ainsi, le modèle réalisé dans l'étude est plus performant que le modèle créé l'année passée, d'autant plus qu'il faut rappeler que le modèle ici présent n'est pas totalement abouti puisque l'étape d'analyse de la part restante du signal géographique présente dans les résidus réalisée dans le chapitre 5 est susceptible d'améliorer encore les performances du modèle.

4.3 Diagnostic de la corrélation géographique des résidus

A l'issue du modèle de Gradient Tree Boosting réalisé, une prédiction de la fréquence VOL pour chaque contrat du portefeuille est disponible. Afin de déterminer s'il demeure un signal géographique à expliquer, les résidus du modèle sont analysés.

4.3.1 Visualisation de la carte des résidus

Un résidu est une mesure de la distance entre y_i la fréquence observée i et $\hat{y}_i$ sa prédiction par le modèle (fréquence prédite). Soit r_i, les résidus dits **« classiques »** :

$$r_i = y_i - \hat{y}_i,$$

en normalisant ces résidus, on peut obtenir les résidus dits de **Pearson** :

$$r_i^{Pearson} = \frac{y_i - \hat{y}_i}{\sqrt{\text{Var}(\hat{y}_i)}}.$$

Cependant, l'écart-type de la fréquence prédite est infime, de l'ordre de 10^{-3} ce qui amplifie les résidus classiques d'un facteur 300.

En pratique, et lors de la création du précédent zonier, l'équipe Multirisque Habitation considère simplement les résidus suivants :

$$r_i = \frac{y_i}{\hat{y}_i}.$$

Cependant dans notre étude de fréquence VOL, la fréquence moyenne observée est très faible, de l'ordre de 0,9%, et les fréquences observées sont donc majoritairement égales à 0. C'est pourquoi, l'utilisation de ce **rapport de l'observé sur le prédit** nous ferait perdre l'information de la prédiction. En effet, en considérant ce genre de résidus et lorsque la fréquence observée vaut 0, que la fréquence prédite ait une valeur proche de 0 comme 0,02 ou qu'elle ait une valeur proche de 1 comme 0,8, ce résidu vaudra 0.

Ainsi il est préférable de conserver les résidus classiques de manière à éviter de perdre de l'information. La carte illustrée dans la figure 4.6 montre les résidus agrégés à la maille IRIS pour la région de Paris et de ses alentours [8].

FIGURE 4.6 – *Carte choroplèthe des résidus agrégés à la maille administrative de l'IRIS pour la région de Paris.*

Avec une carte choroplèthe de ce genre, aucune interprétation ne peut être réalisée, c'est pourquoi la visualisation des résidus sera effectuée d'une autre façon et à la maille ADRESSE de manière à travailler à un niveau plus local.

4.3.2 Le semivariogramme

En regardant la carte précédente des résidus agrégés à l'IRIS, il n'est pas possible de constater s'il demeure un signal géographique dans les résidus. Il est alors possible de penser que ces résidus ne représentent que du bruit. Cependant, l'analyse est réalisée plus en profondeur en étudiant un semivariogramme réalisé sur les résidus à l'adresse pour tenter de démontrer que les résidus contiennent bien une part restante de signal géographique.

La théorie du semivariogramme

Pour analyser s'il demeure un effet géographique dans les résidus non expliqué par le modèle, un semivariogramme est utilisé afin de déterminer s'il existe une autocorrélation spatiale entre les résidus.

Le (semi)variogramme est un **outil géostatistique** permettant de caractériser la l'auto-corrélation spatiale de la variable étudiée [2]. Il s'agit d'un modèle de covariance ne dépendant que de la distance entre les observations. Le terme de semivariogramme

est utilisé lorsque le facteur ½ apparaît dans la définition. Le semivariogramme dépend donc uniquement du vecteur de translation h entre les points s et $s+h$ qui contient de l'information sur la distance entre ces deux points. Dans l'étude, il est alors défini de la manière suivante :

$$\gamma(h) = \frac{1}{2}\mathrm{Var}\left[r(s) - r(s+h)\right],$$

où $r(s)$ est la valeur du résidu pour le point s de localisation donnée par ses coordonnées (x, y) ; et $r(s+h)$ la valeur du résidu dans le voisinage à une distance h.

Par la suite, il est possible d'ajuster un modèle sur ce semivariogramme empirique. Cela permet de créer une passerelle entre description spatiale et prédiction spatiale. Plusieurs paramètres permettent de définir cette fonction afin qu'elle s'adapte au mieux à la forme du semivariogramme empirique.

Le premier élément qui permet l'ajustement concerne l'allure de la courbe elle même, il s'agit alors de choisir parmi plusieurs modèles : le modèle sphérique, circulaire, exponentielle ou encore gaussien.

Quelques exemples d'illustrations de l'allure des courbes de ces modèles sont donnés dans la figure 4.7 ci-dessous.

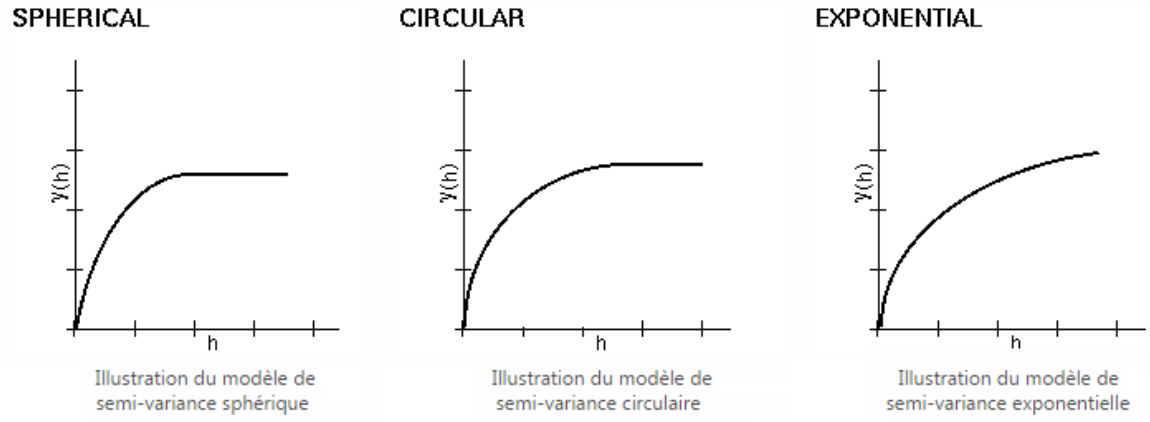

FIGURE *4.7 – Modèles de semivariogramme. Source : http ://desktop.arcgis.com/*

Les autres composantes qui définissent le modèle sont illustrées dans la figure 4.8 et sont expliquées ci-dessous :

L'effet de pépite Théoriquement, à une distance nulle entre deux observations, la variance doit elle aussi être nulle. L'effet de pépite met alors en évidence les potentielles variations spatiales à des distances inférieures à l'intervalle d'échantillonnage choisi pour réalisé le semivariogramme empirique.

La portée et le palier Lorsque la distance entre les observations augmente, le semivariogramme peut atteindre une valeur seuil : il s'agit du palier. La distance à laquelle ce palier est atteint correspond à la portée. Il est souvent considéré qu'au-delà de la portée, il n'y a plus de dépendance spatiale entre les observations.

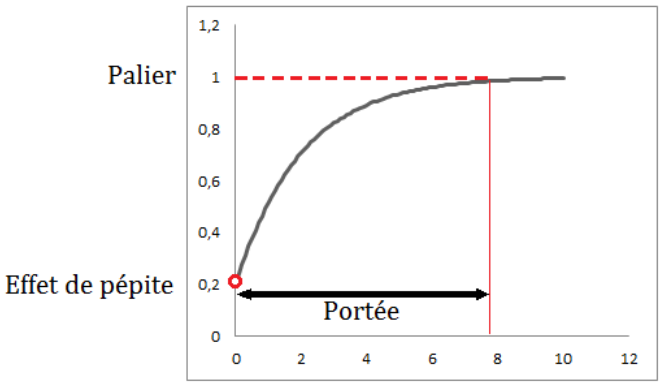

FIGURE 4.8 – *Illustration de l'impact des composantes sur le semivariogramme.*

En général, et lorsque le signal géographique est existant, il est attendu que la fonction ajustée soit croissante au sens où les points les plus proches ont une variance faible et plus la distance entre les points s'intensifie, plus la variance est importante. Dans le cas contraire, c'est à dire lorsque qu'il n'y a pas de signal géographique, la courbe est relativement plate.

Il est important de mentionner que les méthodes géostatistiques sont optimales lorsque les données sont stationnaires : c'est à dire que l'espérance et la variance de la variable étudiée ne varient pas significativement dans l'espace. Ces hypothèses sont d'abord un outil mathématique à l'élaboration d'estimateurs. Elles permettent de développer des formules pour évaluer les quantités à estimer.

Dans notre analyse, des méthodes géostatistiques sont appliquées sur les résidus issus du modèle. Ces derniers ne respectent pas nécessairement les conditions optimales d'application de ces méthodes. Malgré cela, ces techniques géostatistiques seront tout de même mises en œuvre dans notre étude en gardant en tête cette remarque et en procédant à un ajustement visuel des paramètres utilisés pour la modélisation du semivariogramme théorique.

Mise en application du semivariogramme et analyse des résultats [4]

Nous disposons des résidus pour chaque contrat, ainsi les résidus sont associés à une adresse, c'est à dire à des coordonnées (x, y). Dans notre étude, un semivariogramme est appliqué directement sur les résidus issus du modèle pour tenter de capter une part restante de signal géographique. En effet, si les résidus ne contenaient que du bruit, alors la tendance du semivariogramme serait relativement constante. Dans le cas contraire, une croissance de la semivariance en fonction de la distance jusqu'à un certain palier serait observable.

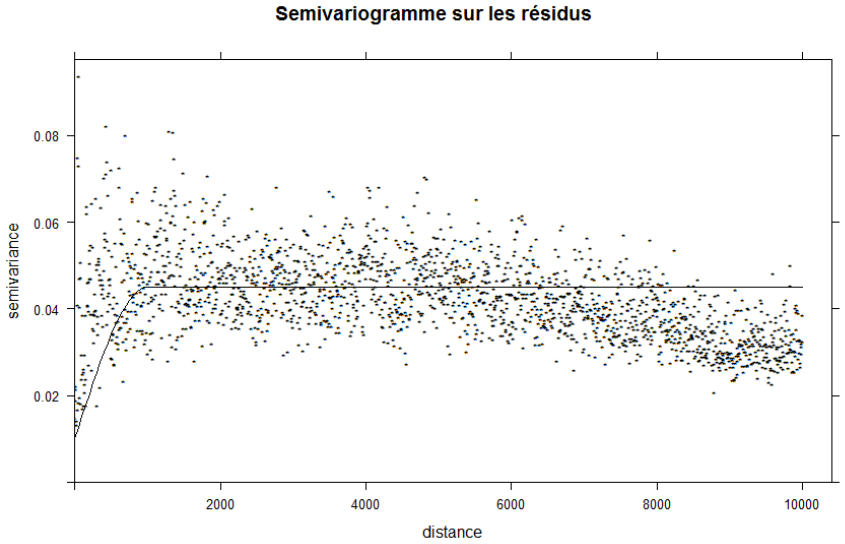

FIGURE 4.9 – *Semivariogramme empirique et modélisé sur les résidus à l'adresse. L'existence d'une auto-corrélation spatiale entre les résidus semble être une hypothèse visuellement acceptable.*

Le semivariogramme 4.9 permet de visualiser à la fois le semivariogramme empirique et le semivariogramme théorique qui correspond au modèle ajusté. La figure laisse apparaître une croissance de la semivariance jusqu'à environ 1 km (l'échelle des distances est en mètre) puis une stabilisation à un certain palier. L'analyse de l'allure du semivariogramme laisse donc présager que, jusqu'à une distance minimale d'environ 1km, il existe vraisemblablement une auto-corrélation géographique des résidus. La portée fournie par le semivariogramme - qui correspond à la distance limite où le palier est atteint - peut permettre de déterminer la fenêtre nécessaire lors du lissage spatial des résidus.

4.4 Prochaines étapes

Ce chapitre a permis de mettre en avant les avantages d'utiliser un modèle de Gradient Tree Boosting car sa sélection automatique de variables s'avère tout de même performante. De plus, grâce aux graphiques d'influence relative et de dépendances partielles des variables, son interprétation est facilitée.

Les performances du modèle réalisé jusqu'à présent fournissent déja des résultats encourageants, cependant nous souhaitons analyser plus en détails les résidus pour tenter de déceler une part de signal géographique restante non captée par le modèle lui-même. Le fait de disposer des coordonnées géographiques de nos contrats permet d'attribuer un résidu à une adresse. Dès lors, cette donnée est alors exploitable afin d'analyser et de visualiser à l'aide de cartes les variations spatiales des résidus.

Les cartes choroplèthes par région administrative sont les plus faciles à mettre en œuvre, cependant les frontières de ces régions ne sont pas nécessairement représen-

tatives des limites du risque. Il est alors nécessaire de se détacher du découpage administratif pour rendre compte du caractère spatial du risque [20]. De plus, pour les zones ayant une faible densité de population, l'échantillonnage est moins important voire inexistant tandis que dans les régions les plus denses, on dispose souvent de beaucoup d'observations (contrats) et il est donc important de ne pas perdre de l'information à cause des limites administratives en allant chercher les variations spatiales à un niveau plus local. L'objectif est donc de capter le caractère spatial résiduel du risque à travers les résidus du modèle tout en se détachant du découpage administratif du territoire pour conserver l'information à une maille très fine dont nous disposons.

Le chapitre suivant s'attardera alors sur les techniques utilisées pour réaliser un lissage spatial des résidus afin d'obtenir une donnée répartie sur l'ensemble du territoire français. L'étude du semivariogramme sera alors utile pour orienter le choix de la maille utilisée pour réaliser le lissage ou encore pour agréger la donnée.

Chapitre 5

Traitement du signal géographique résiduel et diagnostic de stabilité

À l'issue de la réalisation du modèle précédent, nous sommes à présent amenés à étudier la part de signal géographique restante au sein des résidus. Pour cela, plusieurs méthodes d'interpolation spatiale des résidus seront employées en utilisant différents maillages du territoire afin de capter le signal géographique résiduel. Plusieurs étapes de traitement des résidus seront réalisées, notamment une étape de lissage, une étape de choix du maillage et une étape d'agrégation. Ces étapes sont également amenées à être interchangées.

5.1 Lissage des résidus par interpolation spatiale

Les techniques d'interpolation spatiale permettent d'obtenir pour chaque adresse où il n'existe pas d'observation sur le résidu, une estimation de ce résidu à partir des adresses pour lesquelles l'information est disponible. Il est donc possible d'obtenir une estimation des résidus sur l'ensemble du territoire à partir du semis de points constitué par nos contrats. L'utilisation de ces techniques requiert de disposer pour chaque point observé de deux informations : ses coordonnées géographiques et la valeur du résidu associé à cette position. Grâce au géocodage de nos adresses, nous disposons bien de ces deux informations pour chaque contrat.

Ces méthodes estiment que les observations voisines sont similaires c'est pourquoi les observations proches sont pondérées avec une plus grande importance que les observations éloignées.

Dans notre étude, les résidus géolocalisés ainsi qu'une grille du territoire sont utilisés en entrée de ces méthodes et l'élément donné en sortie correspond à la grille du territoire où chaque carreau renvoie une valeur de résidu lissé.

5.1.1 Première approche : Pondération par l'inverse à la distance

La formule générale des techniques d'interpolation spatiale est détaillée dans l'ouvrage de Tomislav Hengl [15]. Dans l'étude présente, nous utilisons ces techniques pour déterminer la valeur estimée $\hat{r}(s)$ du résidu au point s de coordonnées (x, y) à l'aide des valeurs connues des résidus en n points s_i de coordonnées (x_i, y_i) pour $i = 1, ..., n$. La formule est alors donnée par :

$$\hat{r}(s) = \sum_{i=1}^{n} w_i(s) r(s_i), \qquad \text{avec} \qquad \sum_{i=1}^{n} w_i = 1,$$

où n correspond au nombre d'observations. Par défaut, les méthodes d'interpolation spatiale implémentées sous R utilisent l'ensemble des points disponibles mais il est possible de paramétrer le nombre `nmax` de points utilisés pour l'interpolation afin de conserver uniquement les `nmax` observations les plus proches dans l'espace du point s considéré. Quant à w_i, il représente le poids de l'observation i. Il s'agit généralement d'une fonction de pondération décroissante avec la distance de manière à accorder plus d'importance aux observations proches. La somme des poids est contrainte à être égale à 1 de manière à obtenir une interpolation non biaisée.

Dans notre étude, les poids w_i peuvent, par exemple, être obtenus simplement à partir de la fonction inverse $x \mapsto \frac{1}{x}$ et des distances entre le point s et les points s_i :

$$w_i(s) = \frac{\frac{1}{d^{\beta}(s,s_i)}}{\sum_{i=0}^{n} \frac{1}{d^{\beta}(s,s_i)}},$$

où $d(s, s_i)$ représente la distance entre le point s et le point s_i, et où $\beta > 1$ est un paramètre permettant d'ajuster les poids w_i afin d'accentuer la similarité spatiale des points les plus proches. Plus le coefficient β est élevé, moins les points éloignés auront de l'importance. Cette méthode est celle de la **Pondération par l'inverse à la distance** ou *Inverse Distance Weighting* (IDW) qui est une technique d'interpolation spatiale implémentée sous R par la fonction `idw` du package `gstat` [27]. Après avoir testé plusieurs valeurs pour le coefficient β, ce paramètre est finalement fixé à 3.

5.1.2 Seconde approche : le krigeage

L'utilisation d'outils géostatistiques intervient à nouveau dans l'étude. En effet, après l'utilisation du semivariogramme dans le chapitre précédent, c'est à présent la méthode du **krigeage** qui est employée. Cette méthode doit son nom à l'ingénieur minier sud-africain Danie.G Krige (1951), cependant c'est le français Georges Matheron (1962) qui a développé à l'École des mines de Paris l'approche qui sera utilisée ici et qui prend en compte la répartition spatiale des données observées.

Le krigeage est une autre méthode d'interpolation spatiale qui peut être vue comme une sophistication de la méthode basée sur la pondération par l'inverse à la dis-

tance étudiée précédemment. En effet, avec la technique du krigeage, les poids w_i ne sont plus déterminés à partir de l'inverse des distances aux voisins mais par l'intermédiaire du semivariogramme. Grâce à cela, les poids reflètent la vraie structure d'autocorrélation spatiale puisqu'ils tiennent compte de la distance entre les données et le point d'estimation, mais également des distances entre les données deux-à-deux.

De manière plus formelle, le vecteur des poids w_i est obtenu en résolvant le système matriciel suivant :

$$AW = B, \tag{5.1}$$

avec

$$A = \begin{pmatrix} \gamma(h_{11}) & \gamma(h_{12}) & \cdots & \gamma(h_{1n}) & 1 \\ \gamma(h_{21}) & \gamma(h_{22}) & \cdots & \gamma(h_{2n}) & 1 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \gamma(h_{n1}) & \gamma(h_{n2}) & \cdots & \gamma(h_{nn}) & 1 \\ 1 & 1 & \cdots & 1 & 0 \end{pmatrix}, \quad W = \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \\ \lambda \end{pmatrix} \quad \text{et} \quad B = \begin{pmatrix} \gamma(h_1) \\ \gamma(h_2) \\ \vdots \\ \gamma(h_n) \\ 1 \end{pmatrix}$$

où les $\gamma(h_{ij})$ sont calculés directement à partir du semivariogramme empirique et les $\gamma(h_i)$ proviennent, eux, du modèle ajusté sur les points du semivariogramme. Quant au coefficient λ, il découle de la contrainte portant sur la somme des poids. Il s'agit ensuite d'inverser la matrice A puis de multiplier de chaque coté l'équation 5.1 afin d'obtenir $W = A^{-1}B$.

Ce mémoire ne développera pas plus en détails la théorie sous-jacente à la mise en œuvre de cette méthode mais le lecteur est invité à consulter les références [2] et [15] de la bibliographie pour plus d'information.

Nous rappelons que les performances des outils géostatistiques tels que le semivariogramme et le krigeage sont optimales lorsque les données sont stationnaires. Ces hypothèses ont pour but de permettre l'estimation des paramètres statistiques. Nous sommes conscients que les conditions d'optimalité ne sont pas forcément réunies ici puisque nos résidus sont supposés contenir du bruit mais également une part de signal géographique. Par ailleurs, il est difficile de tester les hypothèses de stationnarité efficacement, c'est pourquoi nous décidons tout de même d'appliquer la technique du krigeage dont l'objectif répond totalement à la problématique présente qui est de prévoir la valeur de la variable à interpoler en un site non échantillonné. De plus, cette technique est la seule à tenir compte de la structure de dépendance spatiale des données et constitue donc un outil potentiel d'aide à la décision puisqu'elle présente des résultats intéressants concernant l'amélioration du modèle. C'est une fois encore à l'aide du package `gstat` de R [27] que la technique a pu être mise en application avec la fonction `krige`.

5.2 Agrégation des résidus

L'étape d'agrégation des résidus peut être vue sous plusieurs angles. En effet, l'agrégation des résidus peut être effectuée en étape primaire si l'on souhaite agréger les résidus avant de les lisser mais elle est également utilisée en étape finale après le lissage des résidus lors de la création d'un nombre fixe de zones géographiques de risque : 10 zones, 15 zones ou 20 zones par exemple.

5.2.1 Agrégation des résidus en étape primaire

Il peut être intéressant d'agréger les résidus avant de les lisser. L'agrégation peut alors se faire à une maille administrative plus grossière comme la maille IRIS ou se faire par *raster*, c'est à dire en utilisant des carreaux de même taille. Dans ce cas, les résidus ne sont plus observés à l'adresse mais à la maille considérée. Par exemple, en agrégeant les résidus observés par adresse à la maille IRIS, chaque centroïde d'IRIS est associé à une valeur de résidus correspondant à l'agrégation des résidus dont les adresses appartiennent à cet IRIS. Le centroïde de l'IRIS fait référence à un point fictif situé à l'intérieur du polygone associé à l'IRIS et dont les coordonnées correspondent au centre de ce polygone.

5.2.2 Agrégation des résidus en étape finale

L'agrégation des résidus en étape finale permet de créer les modalités de la variable « zone géographique de risque » qui sera réintégrée dans le modèle de Gradient Tree Boosting. Plusieurs techniques de classification ont été testées notamment une classification par **quantiles** et une classification par la méthode des **K-means**. Il est alors possible de choisir de découper la variable relative aux résidus en 10, 15 ou 20 classes. L'étude concernant le choix du nombre de classe le plus adapté sera réalisée un peu plus loin dans le mémoire.

Découpage par quantiles

La méthode d'agrégation des résidus par quantiles consiste simplement à créer plusieurs classes de risques qui correspondent aux différents quantiles de la variable relative aux résidus lissés. Cependant cette méthode n'est pas optimale puisque les classes formées peuvent contenir deux valeurs de résidus très différentes, en particulier les classes extrêmes. De plus, les classes peuvent être définies sur des intervalles très petits comme sur des intervalles très grands ce qui peut amener à regrouper des résidus peu semblables. Le défaut de la méthode est relativement visible lorsque l'on s'intéresse à l'effet marginal de la variable « zone » réintégrée dans le modèle de Gradient Tree Boosting, car en analysant le graphique 5.1 de dépendance partielle de la variable, aucune tendance ne se dégage réellement.

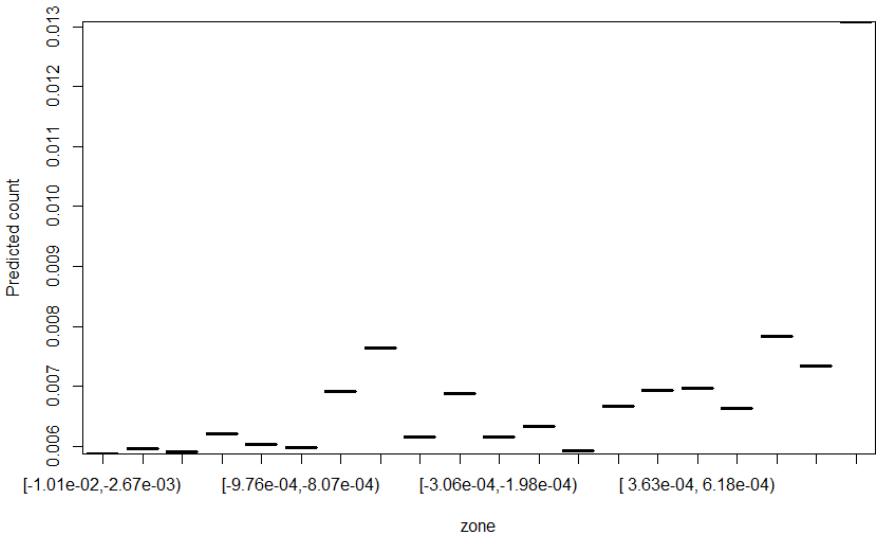

FIGURE 5.1 – *Effet marginal de la variable « zone géographique de risque » créée avec un découpage par quantiles.*

Par ailleurs, les résultats sur la performance du modèle après avoir intégré la variable zone étaient moins bons que lorsque le découpage est réalisé par la méthode des K-means. Ainsi, les résultats obtenus avec un découpage par quantiles ne seront pas détaillés mais nous nous concentrerons sur les résultats observés à l'issue d'un découpage par la méthode des K-means.

Découpage par la méthode des K-means

La méthode des K-means fait partie des techniques de *data clustering*. Il s'agit de méthodes statistiques permettant de diviser un ensemble de données en groupes homogènes. Cette homogénéité est évaluée à partir de critères de proximité définis en introduisant des mesures de distances entre les classes et les observations considérées. Le principe de l'algorithme des K-means consiste à répartir aléatoirement les observations dans K classes, puis de corriger itérativement les K classes afin qu'elles deviennent plus homogènes et qu'elles contiennent des résidus de plus en plus proches.

Modélisation mathématique [31]

Considérons l'ensemble fini des n observations représentées par d caractéristiques, appelons-les $X^{(1)}, ..., X^{(n)}$ et donnons quelques définitions afin de comprendre comment fonctionne l'algorithme :

Cluster Un cluster k est un ensemble non vide d'observations noté :

$$P_k = \{j \mid X^{(j)} \in \text{cluster } k\}.$$

Barycentre d'un cluster Le barycentre d'un cluster k est défini par :

$$m_k = \frac{1}{card(P_k)} \sum_{j \in P_k} X^{(j)}.$$

Cohérence d'un cluster La cohérence d'un cluster k se calcule à l'aide d'une distance d et est défini par :

$$C_k = \sum_{j \in P_k} d(X^{(j)}, m_k).$$

Clustering Un clustering $P = \{P_1, ..., P_K\}$ est une partition à K clusters des données.

L'algorithme des K-means revient alors à minimiser une fonction critère égale à la somme des cohérences des clusters :

$$\min_P \sum_{k=1}^{K} \sum_{j \in P_k} d(X^{(j)}, m_k).$$

Il est à présent possible d'énoncer les étapes réalisées dans l'algorithme des K-means. L'initialisation consiste à attribuer aléatoirement les $X^{(j)}_{j=1,...,n}$ à un des K clusters. Ensuite, et à chaque étape t de l'algorithme, les barycentres m^t_k des différents clusters sont calculés puis chaque $X^{(j)}_{j=1,...,n}$ est ré-affecté au cluster dont le barycentre lui est le plus proche comme l'illustre la figure 5.2.

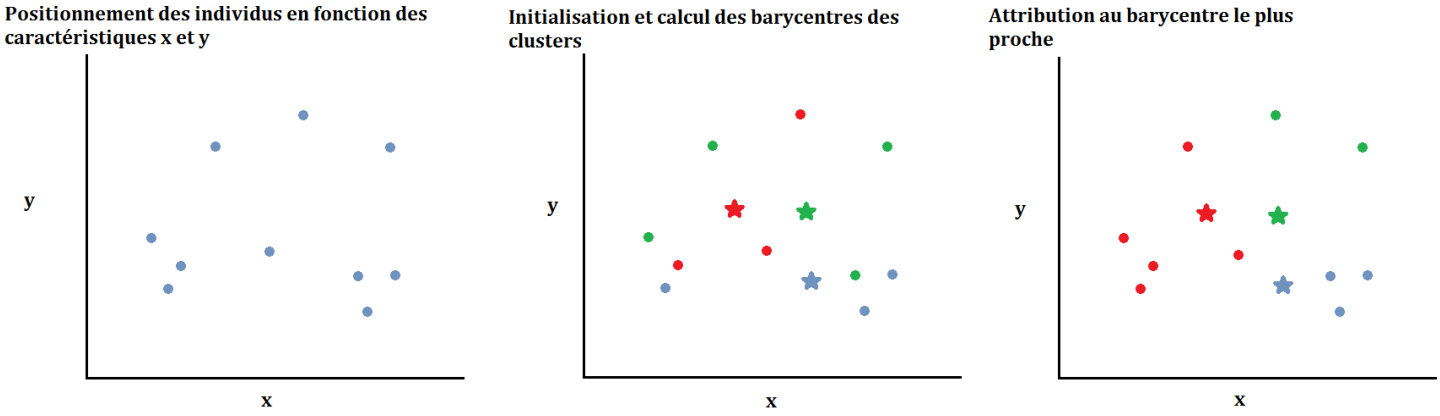

FIGURE 5.2 – *Illustration de l'algorithme des K-means.*

La fonction critère à l'étape t est alors évaluée :

$$A_t = \sum_{k=1}^{K} \sum_{j \in P^t_k} d(X^{(j)}, m^t_k).$$

Si le critère d'arrêt n'est pas respecté, l'algorithme passe à l'étape suivante. Le critère d'arrêt peut correspondre à un nombre d'étapes maximal prédéfini ou peut s'exprimer de la façon suivante : $|A_t - A_{t-1}| < \epsilon$, où ϵ est un seuil préalablement choisi par l'utilisateur.

La classification par la méthode des K-means semble être plus adaptée que le découpage par quantiles puisqu'en observant une fois encore l'effet marginal de la variable « zone » dans le graphique 5.3 de dépendance partielle, il apparaît à présent qu'une tendance croissante se dégage.

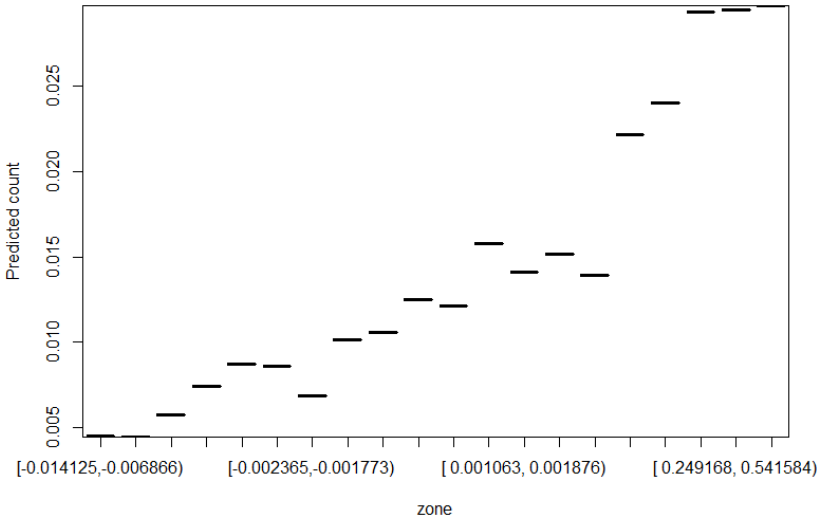

FIGURE 5.3 – *Effet marginal de la variable « zone géographique de risque » créée avec un découpage par K-means.*

Ainsi, le découpage réalisé par cette méthode est plus cohérent que celui réalisé par quantiles puisqu'il crée des classes de risque plus homogènes. Cependant, il faut garder à l'esprit que les classes obtenues par la méthode des K-means dépendent du choix aléatoire initial de l'affectation des observations aux clusters. En effet, deux partitions aléatoires peuvent amener à deux classifications finales différentes, il est donc préférable de réaliser plusieurs tirages aléatoires pour s'assurer de la robustesse de la classification. Par ailleurs, l'efficacité de la méthode repose également sur le choix du bon nombre de classes (*clusters*). Dans l'étude, plusieurs nombres de classes sont testés afin de trouver celui qui convient le mieux.

5.2.3 Maillage du territoire

Le choix de la taille du maillage du territoire est un paramètre important du lissage à considérer puisque les résultats du lissage spatial réalisé sur les résidus sont donnés sur une grille du territoire, c'est à dire par carreaux de taille prédéfinie. En choisissant un maillage fin, le risque sur l'échantillon d'apprentissage est susceptible d'être très bien capté mais l'efficacité du pouvoir prédictif sera réduit, ce qui conduira alors à observer la présence de sur-apprentissage. Par ailleurs, si le lissage n'est pas assez important, l'hétérogénéité dans les carreaux, qui est seulement due au bruit présent dans les résidus, sera conservée. Au contraire, en choisissant un maillage plus grossier, les résidus seront amenés à être trop lissés ce qui aura tendance à détruire la segmentation géographique et la précision dont on bénéficiait sera perdue. Ainsi, il est important de

choisir un maillage qui prend en compte ces deux effets opposés de manière à obtenir un maillage où le lissage des résidus constitue le facteur géographique d'explication le plus prédictif.

À en croire l'analyse du semivariogramme sur les résidus réalisée dans le chapitre 4, une bonne fenêtre de lissage avoisinerait 1km. Il s'agit du seuil de distance en dessous duquel les résidus sont auto-corrélés spatialement. Ainsi, le premier maillage du territoire qui sera considéré concernera des carreaux de taille 1km mais d'autres tailles de carreaux seront également envisagées.

5.2.4 Choix du nombre de zones

Un autre paramètre à prendre en compte dans l'étape de lissage des résidus est le choix du nombre de zones à créer lors du découpage en zones géographiques de risque. Il est donc nécessaire de déterminer combien de modalités doit posséder la variable « zone ». Il faut suffisamment de zones pour que la tarification soit efficace sans pour autant considérer un nombre trop important afin d'éviter de perdre l'effet de mutualisation.

De manière générale, les micro-zoniers construits lors de la refonte de la gamme l'année passée possédaient tous entre 10 et 20 zones de risque selon la garantie étudiée. En particulier, le précédent micro-zonier construit pour la fréquence VOL pour les appartements possédait 20 zones géographiques de risque. Ainsi, des découpages en 10,15 ou 20 zones seront testés afin de déterminer le nombre de zones le plus adapté à notre étude.

5.3 Études de sensibilités et performances

Dans cette section, la méthode de pondération par l'inverse à la distance abordée précédemment est mise en œuvre. Plusieurs sensibilités seront étudiés dont la sensibilité du modèle au choix du maillage du territoire ainsi qu'au nombre de zones choisi, mais également sa sensibilité aux potentiels résidus extrêmes. Chacun de ces leviers sera testé de manière univariée : un seul paramètre sera modifié à chaque fois de manière à bien appréhender son impact sur les performances du modèle. Seuls les indices de GINI seront analysés pour le moment car ils constituent un critère de choix suffisant pour témoigner de la stabilité du modèle dans un premier temps. L'étude des sensibilités de ces leviers aura pour but de fixer au mieux les paramètres et se fera par l'intermédiaire de la méthode de la pondération inverse à la distance. Une fois que tous les paramètres auront été convenablement choisis avec cette première approche, la méthode géostatistique plus élaborée du krigeage pourra alors être mise en œuvre avec ces mêmes paramètres dans l'espoir d'accroître la qualité du lissage des résidus et ainsi d'accentuer l'amélioration du modèle.

5.3.1 Étude de la sensibilité au choix du maillage du territoire

La première étude de sensibilité concerne le choix du maillage du territoire pour réaliser le lissage spatial. Les cartes ci-après permettent d'obtenir un aperçu géographique de ce que représente un carreau de 1 km (figure 5.4), un carreau de 5 km (figure 5.5) ou un carreau de 10km (figure 5.6). Le maillage du territoire choisi correspond à la taille de carreau de la grille utilisée pour le découpage du territoire. Cette grille est donnée en sortie de la méthode d'interpolation spatiale, c'est à dire qu'à chaque carreau de la grille est associée une estimation de résidu obtenue par interpolation spatiale.

Les cartes ci-dessous ne traduisent pas un niveau de risque mais plutôt l'amplitude du risque restant à expliquer puisqu'il s'agit des résidus obtenus par l'interpolation spatiale sur chaque carreau du territoire, pour différentes tailles de carreaux. Il apparaît qu'un lissage spatial sur un carreau de taille 1km est un lissage relativement fin puisqu'il n'est pas réellement possible de distinguer les carreaux à l'œil nu. Le lissage sur un carreau de 5km est un peu plus grossier mais reste tout de même difficilement discernable. Quant au lissage sur un carreau de 10km, les carreaux sont nettement plus visibles et le territoire français devient relativement trouble.

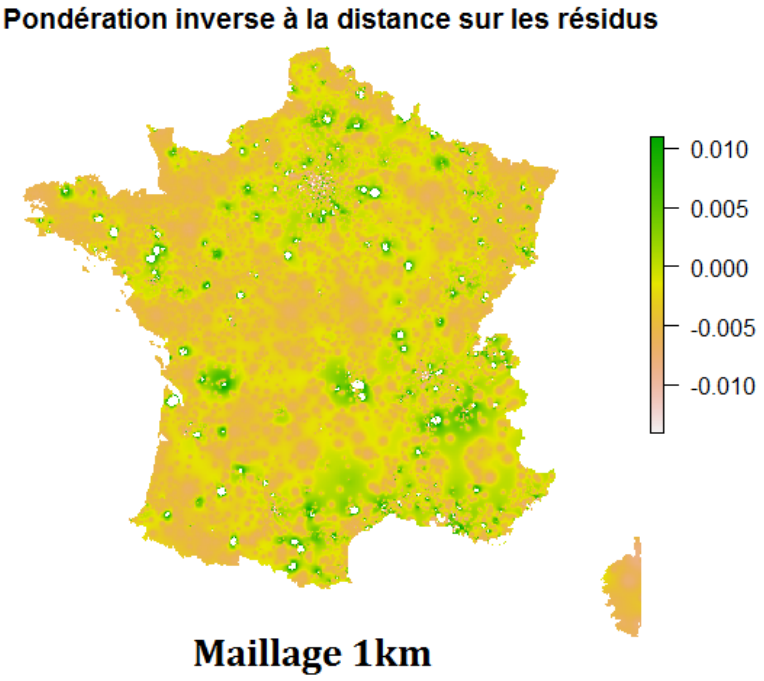

FIGURE 5.4 – *Lissage spatial des résidus avec un maillage de 1km.*

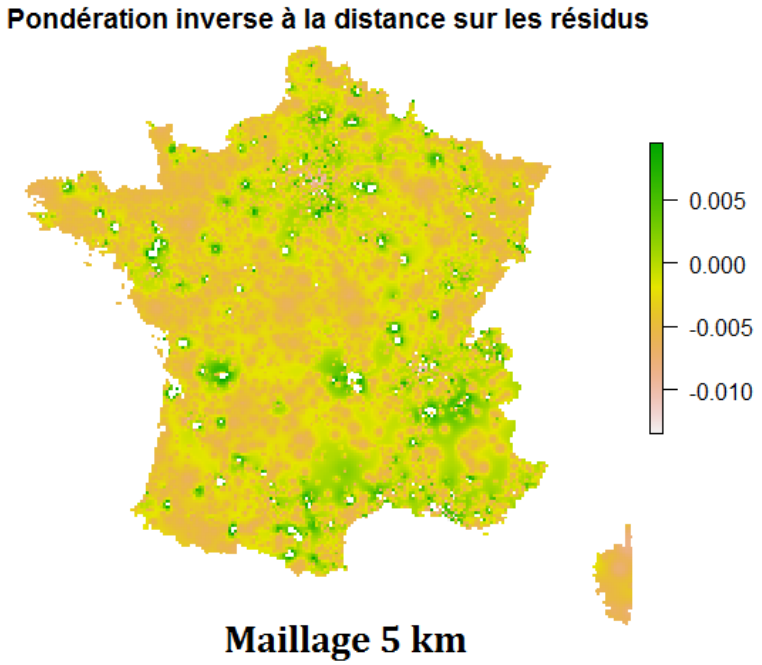

FIGURE 5.5 – *Lissage spatial des résidus avec un maillage de 5km.*

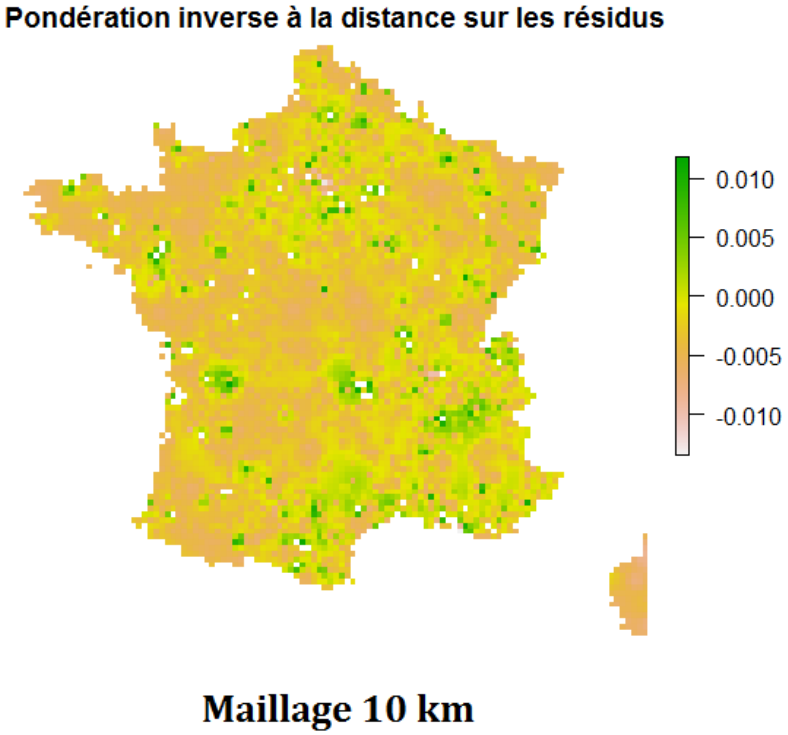

FIGURE 5.6 – *Lissage spatial des résidus avec un maillage de 10km.*

Comparaison des performances du modèle

Cette partie concerne l'étude de la sensibilité de la performance du modèle au maillage. Les deux autres paramètres pouvant potentiellement influer sur la performance sur modèle sont donc fixés. Le nombre de zones est alors fixé à 20 et nous travaillons sur l'ensemble des résidus à l'adresse (sans traitement des résidus extrêmes).

Rappelons qu'à l'issue de la réalisation du semivariogramme, appliquer un maillage du territoire avec un carreau de 1km semblait être la première approche à tester. Cependant, l'observation de la performance du modèle avec un tel maillage révèle que le maillage est trop fin puisque le modèle suggère la présence d'un important surapprentissage : le modèle performe mieux que le modèle initial (c'est à dire le modèle obtenu sans le lissage spatial des résidus) sur l'échantillon d'apprentissage mais est moins efficace sur l'échantillon test. Ainsi, des maillages plus grossiers de 5 km et 10 km sont alors utilisés de manière à réduire l'*overfitting*. Les résultats concernant les performances des modèles en fonction des maillages sont retranscrits dans la figure 5.7.

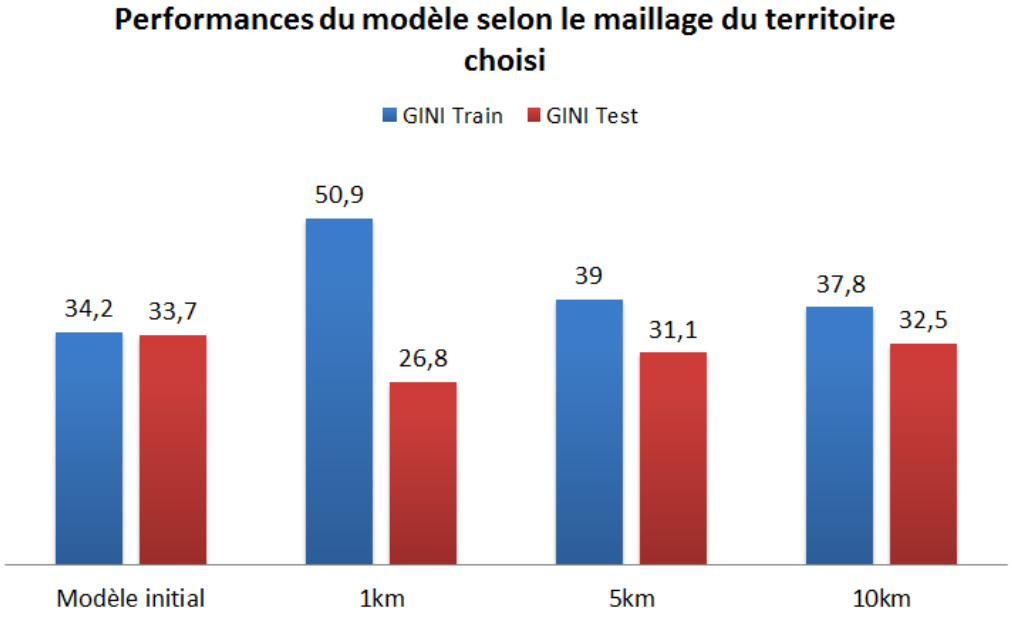

FIGURE 5.7 – *Performances des modèles en fonction du choix du maillage.*

Il semblerait que le modèle le plus stable soit celui créé avec un maillage de 10km, c'est à dire celui où les résidus sont le plus lissés. Il n'en demeure pas moins que le modèle qui intègre la variable « zone géographique de risque » est tout de même dégradé par rapport au modèle obtenu sans réaliser le lissage spatial des résidus.

5.3.2 Étude de la sensibilité aux résidus extrêmes

Les résidus issus du modèle initial sont calculés comme étant la différence entre la fréquence observée et la fréquence prédite par le modèle. Cependant lorsqu'un ou

plusieurs sinistres surviennent sur une période d'exposition relativement courte, la fréquence observée est alors très importante. La fréquence prédite n'est généralement pas du même ordre de grandeur, ce qui amène à observer quelques résidus très importants qui peuvent être interprétés comme des *outliers*. Ainsi, l'impact sur les performances du modèle de ces résidus extrêmes a été testé en les écrêtant très légèrement afin ne de pas les prendre en compte. Seuls les résidus de la base d'apprentissage inférieurs à un seuil de 2 sont conservés, ce qui représente plus de de 99% de la base d'apprentissage.

Pour cette étude, les paramètres concernant le maillage du territoire et le nombre de zones sont respectivement fixés à 5km et 20 zones.

La figure 5.8 montre que l'écrêtement des résidus impacte négativement les performances du modèle puisque l'indice de GINI diminue sur l'échantillon d'apprentissage mais également sur l'échantillon test.

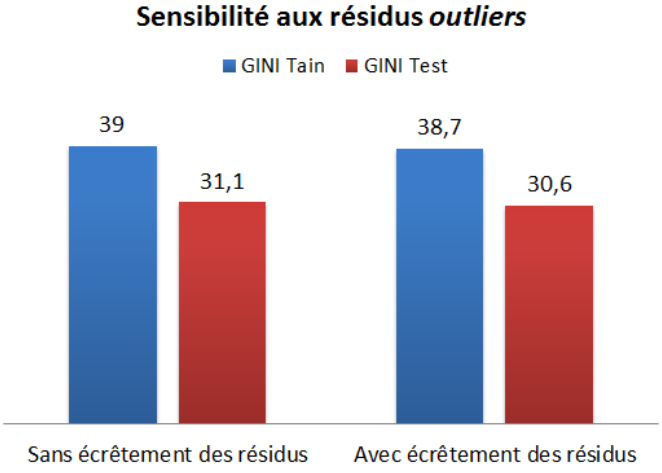

FIGURE 5.8 – *Performances du modèle avec et sans l'écrêtement des résidu.*

Bien que certains résidus peuvent être considérés comme extrêmes, les enlever pénalise le modèle. Par conséquent, le reste de l'étude conservera l'ensemble des résidus sans écrêtement.

5.3.3 Étude de la sensibilité au nombre de zones

Le choix du nombre de zones permet d'obtenir un nombre fini de modalités pour la variable « zone géographique de risque » qui sera réintégrée dans le modèle. Le nombre de modalité doit être choisi de manière à obtenir une bonne segmentation du risque pour que la tarification soit efficace tout en profitant de l'effet de mutualisation.

Pour cette étude, le paramètre de maillage du territoire est fixé 5km et nous travaillons sur l'ensemble des résidus observés. Plusieurs nombres de zones sont testés : 10 zones, 15 zones et 20 zones.

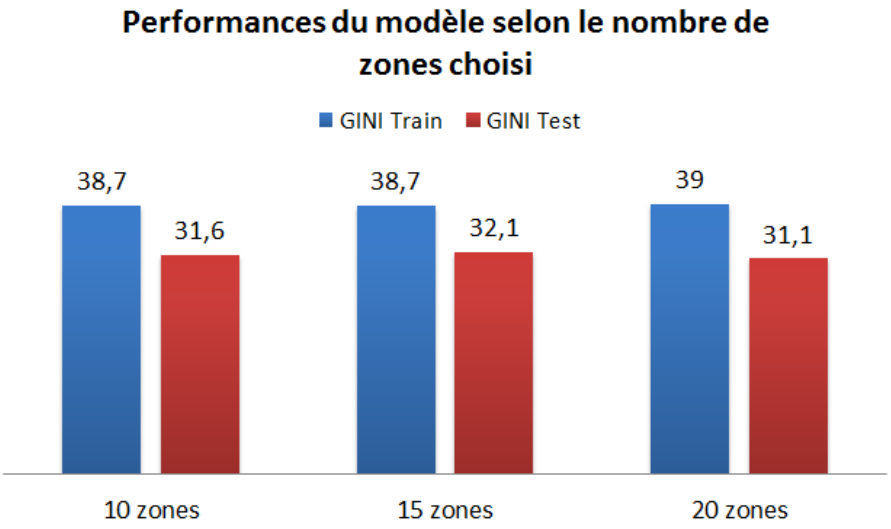

FIGURE 5.9 – *Performances du modèle en fonction du nombre de zones.*

Il semblerait que le choix de 20 zones soit celui le moins adapté puisque c'est dans ce cas que l'écart de GINI est le plus important : ce qui traduit une instabilité plus importante. Par ailleurs, l'indice de GINI sur l'échantillon test est également le plus faible, c'est pourquoi le choix de 20 zones n'est pas approprié. Concernant le choix de 10 zones ou de 15 zones, on remarque que les indices de GINI sur l'échantillon d'apprentissage sont les mêmes dans le deux cas cependant sur l'échantillon test, c'est le choix d'un nombre de zones égal à 15 qui permet d'obtenir un indice de GINI supérieur permettant également de réduire l'écart entre les indices des deux échantillons. Ainsi, l'utilisation de 15 modalités sera privilégié pour caractériser la variable concernant la zone géographique de risque.

5.3.4 Étude de l'impact de la maille d'observation des résidus

Jusqu'à présent les études de sensibilités ont été réalisées à partir des observations des résidus à l'adresse, cependant il est peut-être plus judicieux de ne pas considérer les résidus par adresse mais de les agréger à une certaine maille avant de les lisser. Plusieurs approches sont possibles. Il est par exemple possible de tenter d'agréger les résidus à la maille administrative de l'IRIS (plus fine qu'une maille à la commune). Ainsi, l'observation des résidus sera disponible à une maille plus grossière que la maille ADRESSE. Les résidus sont alors agrégés et associés aux coordonnées du centroïde de l'IRIS puis sont lissés par les mêmes méthodes que celles que nous avons vues jusqu'à présent.

Suite aux études de sensibilités précédentes, le paramètre concernant le nombre de zones (c'est à dire le nombre de modalités de la variable zone) est fixé à 15 et nous travaillons à nouveau sur l'ensemble des résidus observés.

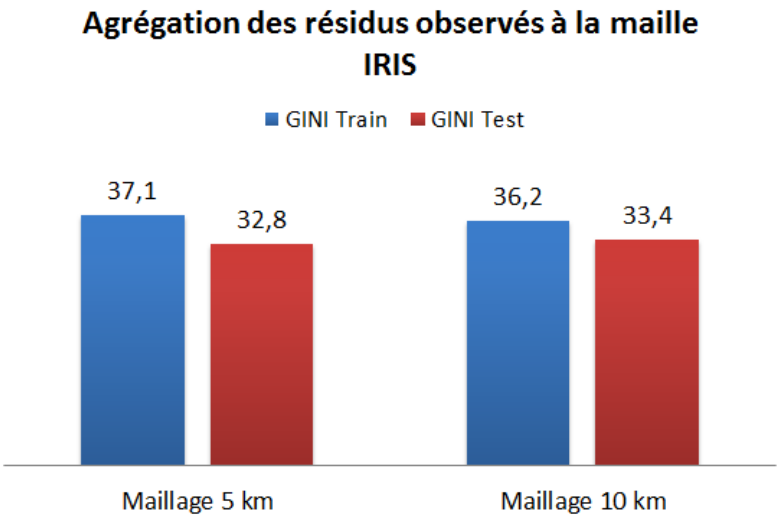

FIGURE *5.10 – Performances du modèle lorsque les résidus observés sont agrégés à la maille IRIS.*

La figure 5.10 montre que les performances des modèles sont meilleures dans le cas où les résidus observés sont à la maille IRIS et non à l'adresse. Les résultats obtenus dans ce cas par les modèles se rapprochent même sensiblement de ceux du modèle initial.

La conclusion de cette étude sur le choix de maille pour l'observation des résidus est intéressante puisqu'il semble cohérent de considérer que l'observation des résidus à une maille agrégée a plus de sens que l'observation des résidus à la maille adresse.

Bilan sur les études de sensibilités des paramètres utilisés lors du lissage

Plusieurs études de sensibilités ont été réalisées afin de pouvoir mesurer l'impact de chaque paramètre intervenant lors de l'étape de traitement du signal géographique présent dans les résidus. Elles ont permis de déterminer les valeurs les plus adaptées pour chaque paramètre.

La fenêtre du maillage du territoire choisie correspond à 10 kilomètres et 15 zones géographiques de risque ont été conservées. Concernant l'étude sur les valeurs aberrantes parmi les résidus, les conserver ne semble pas présenter d'inconvénient. L'étude concernant la maille d'observation des résidus a révélé que considérer des résidus agrégés à la maille IRIS plutôt qu'à la maille ADRESSE constituait un axe d'analyse plus intéressant. L'ensemble de ces constatations permet de fixer les paramètres au mieux pour ensuite tenter d'améliorer la prédiction avec une méthode d'interpolation spatiale alternative à la pondération inverse à la distance : le krigeage.

5.4 Mise en œuvre du krigeage et performances

Les études réalisées précédemment ont permis de déterminer les paramètres les mieux adaptés pour mettre en œuvre le lissage spatial par la méthode de pondération inverse à la distance. En fixant à présent ces mêmes paramètres, il est possible d'appliquer une méthode d'interpolation plus sophistiquée telle que le krigeage. La méthode du krigeage est donc réalisée sur les résidus agrégés par IRIS et permet d'obtenir une estimation sur chaque carreau de 10km du territoire.

Dans les chapitres précédents, il a été établi que le modèle initial était déjà plus performant que le modèle créé l'année passée. La figure 5.11 permet de rendre compte de l'impact du lissage sur la stabilité du modèle. Il semblerait que lorsque les paramètres sont bien choisis, la méthode de pondération par l'inverse à la distance permet de se rapprocher sensiblement des performances du modèles initial mais intègre tout de même trop de bruit au modèle ce qui conduit à plus de sur-apprentissage. En revanche, la méthode de krigeage semble plus appropriée puisqu'elle apporte une information supplémentaire au modèle car on peut observer une légère amélioration sur l'indice de GINI de la base test tout en limitant le sur-apprentissage.

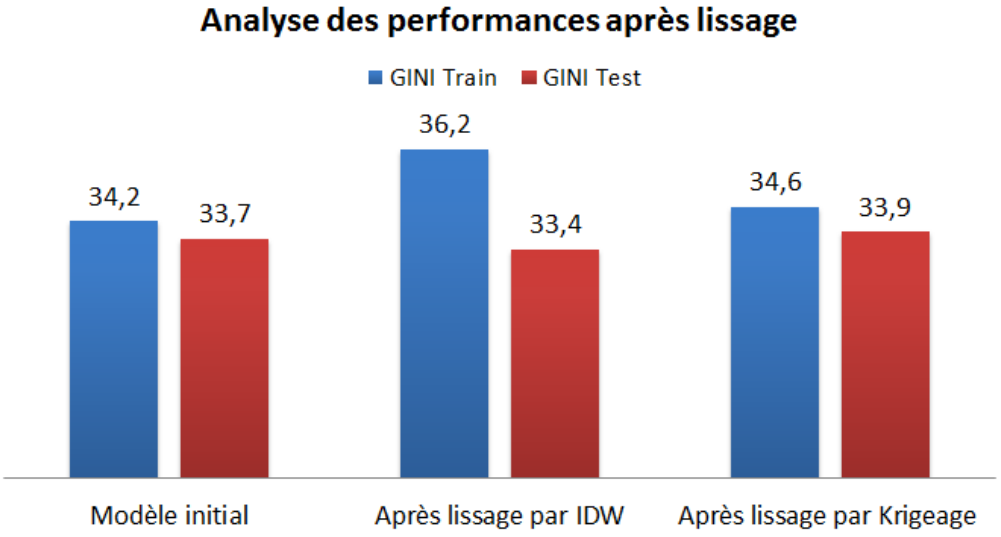

FIGURE 5.11 – *Étude de l'impact de l'intégration dans le modèle de la nouvelle variable géographique créée à partir du lissage spatial des résidus.*

Pour mesurer plus en profondeur le gain réalisé par le lissage spatial, intéressons-nous à présent aux valeurs des Root Mean Squared Error pour ces deux nouveaux modèles par rapport au modèle initial.

Modèle	Modèle initial	Après lissage par IDW	Après lissage par Krigeage
RMSE Train	0,1291070	0,1290980	0,1291043
RMSE Test	0,1306719	0,1306707	0,1306696

TABLE *5.1 – Analyse des performances de prédiction des modèles après lissage à l'aide des Root Mean Squared Error.*

La table 5.1 ci-dessus permet de consolider les résultats précédents puisque les Root Mean Squared Error induisent globalement les mêmes constatations que celles obtenues suivant les valeurs des indices de GINI. Cela confirme donc que l'étape de traitement du signal géographique présent dans les résidus permet d'améliorer la performance du modèle.

5.5 Résultats du diagnostic de performance

Ce chapitre traite de la part du signal géographique restante dans les résidus. Différentes approches ont été testées tout en étudiant plusieurs paramètres susceptibles d'influer sur la qualité de la variable zone créée dont notamment : le maillage du territoire ou encore le nombre de modalités de la variable.

Chaque test a permis de déterminer quel était le paramètre le plus adapté dans notre étude, cependant aucune des analyses réalisées sur les résidus observés à la maille ADRESSE n'a permis de créer une variable zone qui captait suffisamment de signal géographique pour améliorer le modèle. En effet, dans tous les tests réalisés sur l'observation des résidus à la maille ADRESSE, il a été constaté qu'en intégrant la nouvelle variable « zone » dans le modèle, de l'instabilité était introduite par la même occasion puisque l'écart en terme de GINI s'élargissait tout en diminuant la valeur de l'indicateur sur l'échantillon test qui demeure tout de même l'échantillon auquel il faut prêter attention afin de garantir une vision objective de la performance du modèle créé.

Les résidus sont constitués, d'après notre étude, d'une part de signal géographique mais également de bruit. Ainsi, il est probable que le bruit présent au sein des résidus prenne le dessus sur le signal géographique et qu'en incorporant la variable zone issue des résidus dans le modèle, trop de bruit soit également intégré, ce qui constituerait une explication probable de la dégradation de la performance du modèle. Une autre explication possible proviendrait du fait que, considérer des résidus à la maille ADRESSE n'est peut être pas le plus judicieux. C'est pourquoi la maille d'observation des résidus a été testée afin de considérer les résidus à une maille plus grossière telle que l'IRIS. L'intuition a été confirmée puisque l'étude des résidus observés à la maille IRIS se trouve être plus concluante car il semblerait que les modèles soient plus performants que lorsque les résidus sont observés à la maille ADRESSE.

Par conséquent, la section suivante tentera d'exploiter la piste du lissage des résidus observés à la maille ADRESSE et à la maille IRIS sur une base de données possédant un historique d'observations plus profond afin d'être en mesure de capter un meilleur signal géographique.

5.6 Résolution de l'instabilité de la prédiction géographique

Durant tout le début de l'étude, la base de données était uniquement composée d'observations sur l'année 2013. Cela s'explique notamment par le fait que la base de données utilisée est très volumineuse à cause de la quantité massive de variables externes disponibles ce qui constitue une limitation en terme de stockage dans la mémoire de l'ordinateur.

Cette dernière section permet d'appliquer les méthodes précédemment testées sur une base de données couvrant un historique sur les années 2009 à 2013. Nous espérons, dans cette optique, réussir à capter plus de signal géographique que de bruit dans les résidus. Cependant, l'étude sera restreinte à un seul département afin de réaliser une analyse à une échelle plus locale tout en limitant le volume de la base de données. L'échantillon d'apprentissage sera constituée des années 2009 à 2011 ainsi que d'une partie de l'année 2013, tandis que l'échantillon test sera constituée de l'année 2012 et de l'autre partie de l'année 2013.

5.6.1 Mise en application de la méthode éditée

Modèle initial

Pour disposer d'une base de comparaison cohérente avec les calculs qui suivront, le modèle réalisé dans l'étude est appliqué sur un nouvel échantillon d'apprentissage qui correspond à l'ensemble des observations des années 2009 à 2011 ainsi qu'une partie de l'année 2013.

Nous rappelons que le modèle de Gradient Tree Boosting utilisé correspond au modèle intégrant les variables internes et externes obtenues après les différentes étapes de sélection. Les résultats obtenus sur cette nouvelle base de données s'apparentent aux précédents, au sens où le modèle est relativement stable avec un écart de 1.6 en terme d'indice de GINI.

Krigeage

L'application de la méthode du Krigeage sur cette base est réalisée en testant plusieurs fenêtres de maillage du territoire : 5 km, 7 km et 10 km. Le précédent constat émis sur les résultats concernant l'observation des résidus à l'adresse est identique ici, c'est à dire que le modèle présente un sur-apprentissage quel que soit le maillage utilisé. Par ailleurs, lorsque les résidus sont observés à la maille IRIS, il est intéressant de

constater que le choix d'une fenêtre de 10 km pour le maillage du territoire complet ne demeure pas forcément le plus adapté lorsque l'étude concerne uniquement un département (figure 5.12).

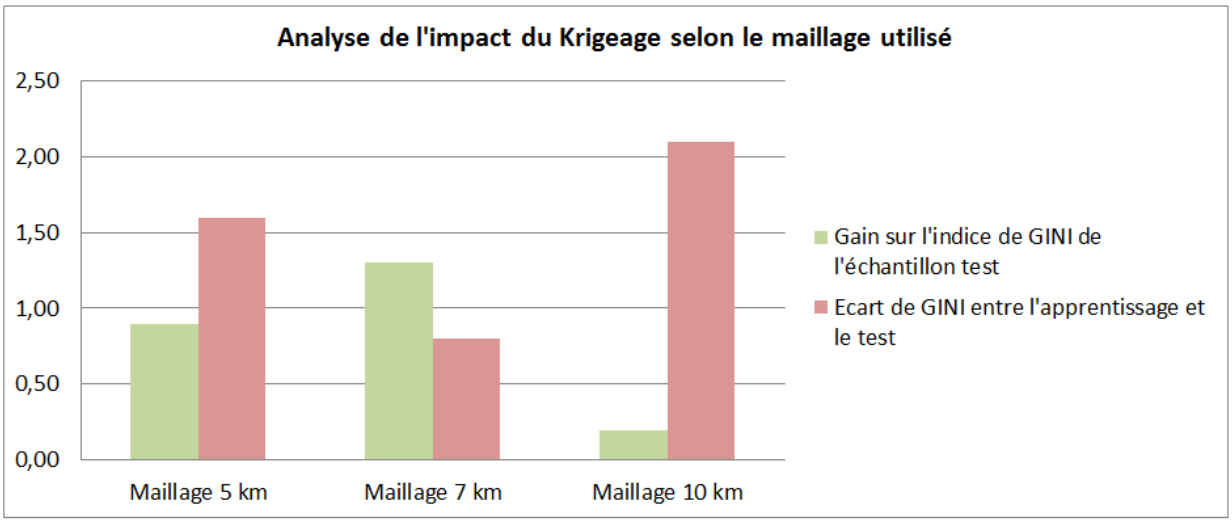

FIGURE 5.12 – *Étude de l'impact du krigeage selon le maillage du territoire choisi pour des résidus observés à la maille IRIS.*

En effet, sur l'étude du département des Hauts-de-Seine (92), le maillage du territoire le plus approprié semble être un maillage de 7 km. En utilisant ce maillage, cela permet non seulement de stabiliser davantage le modèle en réduisant l'écart en terme d'indice de GINI, mais également d'améliorer la segmentation du risque puisque l'indice de GINI sur l'échantillon test s'améliore de plus d'un point. Quant à l'utilisation d'une fenêtre de maillage de 10 km, il semblerait que ce soit la fenêtre la moins adaptée puisque, non seulement le gain sur l'indice de GINI est relativement faible mais la stabilité du modèle est dégradée. Ainsi, en fonction de l'échelle regardée - à l'échelle du territoire ou à l'échelle d'un département - la taille de la fenêtre ne sera pas la même. Il est également probable que l'observation d'un département autre que les Hauts-de-Seine nécessite d'une fenêtre de maillage encore différente.

5.6.2 Conclusion sur l'observation de plusieurs années sur un département

L'étude départementale réalisée permet de consolider les constats déjà établis sur la méthode éditée. Cependant, cette analyse a également permis de mettre évidence le fait que localement, lorsque la fenêtre de maillage est bien adaptée, un gain non négligeable de stabilité et de performance est effectivement obtenu en intégrant dans le modèle la nouvelle variable géographique créée à partir du lissage spatial des résidus.

Conclusion et extensions possibles

Dans l'intention de pallier des problèmes de stabilité survenus lors de la construction du précédent micro-zonier pour la fréquence VOL sur les appartements, le présent mémoire avait pour objectif d'établir un processus de construction de micro-zonier à l'aide de méthodes alternatives. La finalité de cette étude est donc de pouvoir répondre à la problématique de la modélisation du risque géographique tout en garantissant une information à la fois robuste et précise.

Une première partie non négligeable du mémoire se focalise sur l'exploitation de l'ensemble des données externes disponibles provenant des différentes sources et à plusieurs mailles géographiques. L'une des problématiques soulevées concernait l'apport de la donnée externe à une maille aussi fine que les coordonnées GPS. Il se trouve que, sur les données dont nous disposions, les résultats ne se sont pas avérés aussi révélateurs que prévus puisqu'en utilisant cette donnée à la maille adresse ou à une maille agrégée (mais qui reste néanmoins assez fine), les modèles obtenus fournissent des performances relativement équivalentes. Par ailleurs, le fait de disposer d'une bonne précision de géocodage des adresses de notre portefeuille peut constituer une étape délicate en pratique car il s'agit d'un processus complexe qui est toujours en train d'être amélioré aujourd'hui.

Cependant, l'étude réalisée a permis de répondre à plusieurs questions. Elle a notamment permis de mettre en évidence la nécessité d'utiliser les données externes disponibles pour capter plus d'information sur l'environnement où évolue le contrat et ainsi prédire au mieux le risque auquel est soumis l'assuré afin de lui proposer un tarif adapté.

De plus, l'étude a révélé qu'il était possible de simplifier le questionnaire client puisqu'en le restreignant uniquement à l'adresse du client, il est possible d'avoir accès à l'ensemble des données externes géographiques qui permettent d'obtenir un modèle pratiquement aussi performant qu'un modèle qui serait composé uniquement des variables relatives au risque et à l'assuré. La donnée externe permet alors de répondre à la nécessité d'obtenir rapidement un devis sur le tarif (*Quick quote*) et ainsi rester compétitif compte tenu des tendances d'aujourd'hui qui conduisent les assurés à comparer les offres sur internet.

Pour finir, l'analyse de cette donnée externe a également permis de souligner le

fait que l'information contenue dans les variables internes telles que les montants assurés, souvent sous-estimés par les clients lorsqu'ils les déclarent, peut être remplacée par l'information contenue dans certaines données externes qui peuvent constituer un élément plus fiable pour la tarification.

Au regard de la quantité importante de variables dont nous disposions pour réaliser le modèle, l'utilisation de méthodes d'apprentissage automatique telles que le Gradient Tree Boosting s'est imposée comme une évidence. En effet, les résultats d'une analyse comparative des performances d'un modèle linéaire généralisé log-poissonien avec un modèle de Gradient Tree Boosting a permis de montrer que les performances étaient similaires mais que le gain de temps sur le traitement des données était considérable lorsque l'on utilise un Gradient Tree Boosting. Par ailleurs, l'efficacité de l'algorithme de Gradient Tree Boosting dans sa sélection automatique de variables a également été remise en cause en réalisant une seconde sélection plus manuelle en parallèle à l'aide de techniques telles que l'étude des corrélations, l'Analyse en Composantes Principales ou encore une procédure STEPWISE. Cette étape a permis de confirmer la qualité de sélection automatique des variables réalisée par l'algorithme.

Une fois le modèle final de Gradient Tree Boosting réalisé en intégrant des données internes et des données externes, il a été établi que le modèle ainsi créé donnait des résultats plus intéressants que le précédent modèle élaboré par l'intermédiaire d'un modèle linéaire généralisé log-poissonien et à l'aide des cellules de Voronoï. En effet, en intégrant les variables externes - qui constituent une partie du signal géographique - directement dans le modèle de Gradient Tree Boosting, le modèle produit s'est avéré à la fois plus performant et plus stable.

Cependant, l'étude a été menée plus en profondeur en analysant les résidus issus du modèle pour tenter de capter un signal géographique résiduel. Les résidus ont alors été lissés par l'intermédiaire de méthodes d'interpolation spatiale telles que la pondération inverse à la distance ou encore le krigeage de manière à créer une variable de zone géographique de risque. Durant l'étape d'analyse des résidus, nous avons également été amenés à étudier l'impact des différents paramètres qui interviennent dans le lissage sur la qualité de la création de la variable « zone » tels que le maillage du territoire, le nombre de modalités de la variable mais aussi la maille d'observation des résidus. Les conclusions tirées de ces études sur les résidus ont révélées qu'il existait bien une part de signal géographique demeurant dans les résidus. C'est en appliquant la méthode géostatistique du krigeage sur les résidus agrégés à la maille IRIS que l'intégration dans le modèle de la variable « zone géographique de risque » créée s'est avérée la plus efficace.

Une fois ce constat réalisé, la méthode éditée a été appliquée sur une autre base de données contenant les observations sur les années 2009 à 2013 en restreignant l'étude à un département français afin d'étudier le phénomène plus localement et sur un historique plus profond. Les résultats de cette analyse ont confirmés ceux déjà évoqués précédemment. Par ailleurs, cette étude a également permis de mettre en évi-

dence le fait que le choix de la fenêtre de maillage du territoire dépend de la région observée. Ainsi, les prochaines études pourront s'intéresser à développer cette méthode en affinant le maillage utilisé en l'adaptant à la densité de population ou à la densité des observations.

Il est intéressant de noter que les conclusions obtenues dans ce mémoire ne présument en aucun cas d'une efficacité quelconque des techniques utilisées mais sont reliées à la base de données étudiée dans le cadre d'une étude de la fréquence de survenance d'un sinistre vol sur les appartements. Cependant, l'étude réalisée a permis de fournir une bibliothèque de fonctions qui a été mise à la disposition de l'ensemble du groupe AXA dans un environnement interne de partage de codes. Ainsi, les méthodes utilisées dans cette étude peuvent être exploitées sur d'autres bases de données ou sur d'autres garanties Habitation mais également sur d'autres lignes métier telles que l'assurance automobile par exemple.

Ce mémoire se concentre sur la modélisation du risque géographique dans le cadre de la réalisation d'un modèle de fréquence VOL sur les appartements, cependant il pourrait être judicieux d'allier cette méthode avec une étude temporelle du risque qui refléterait la saisonnalité du phénomène. Cela conduirait à une meilleure connaissance du risque et permettrait d'entreprendre des actions préventives comme, par exemple, l'envoi de messages pour avertir l'assuré qu'il se trouve dans une période et dans une zone où le risque de se faire cambrioler est non négligeable.

Lexique

Apprentissage automatique Le terme de Machine Learning peut également être employé. Il s'agit des méthodes qui permettent de concevoir des programmes capables de s'améliorer automatiquement avec l'expérience. De cette manière, la machine « apprend » en changeant sa structure de manière à être plus efficace la fois d'après. Le Gradient Boosting fait partie de ces méthodes.

Big Data Concept qui consiste à collecter et à gérer une quantité massive de données variées provenant de partout telles que les informations climatiques, la géolocalisation ou encore les objets connectés. La manipulation des ces données requière des techniques telles que le Machine Learning adaptées aux bases de données volumineuses.

Code INSEE Code correspondant au découpage administratif par communes. Il a été élaboré par l'Institut National de la Statistique et des Études Économiques (INSEE). On dénombre environ 36 000 codes INSEE en France.

Code IRIS Le mot IRIS est l'acronyme pour « Ilots Regroupés pour des Indicateurs Statistiques ». Il s'agit d'un découpage administratif infracommunal utilisé par l'INSEE pour les recensements de population. La plupart des communes de plus de 5 000 habitants sont divisées en IRIS. Ce découpage partitionne la France en environ 51 000 zones.

Data Science Discipline qui s'appuie sur des outils mathématiques, de statistiques, d'informatique et de visualisation des données afin de créer des modèles qui capturent au mieux la structure sous-jacente complexe des données.

Projection Représentation cartographique plane des coordonnées géographiques. Plusieurs systèmes de projection existent : cylindrique, coniques et azimutales. Une projection est toujours associée à un système de référence terrestre (SRT) et son ellipsoïde. En d'autres termes, les projections sont la transition d'une forme quasi sphérique (la terre en 3 dimensions) à une surface plane (la carte en 2 dimensions).

Signal géographique Part du risque assuré qui dépend uniquement de la localisation du contrat.

Système d'Information Géographique (SIG) Un système d'information géographique est un système d'information capable de stocker, d'organiser et de présenter des données alphanumériques spatialement référencées par des coordonnées dans un système de référence (CRS).

Systèmes de Référence de Coordonnées (CRS) Permettent d'identifier un jeu de coordonnées en précisant les éléments de définition nécessaires à leur positionnement. Des transformations permettent de passer de l'un à l'autre.

Système RGF93 (Réseau Géodésique Français 1993) Système global obtenu par densification des points du réseau mondial associé ETRS89. Il s'agit du système officiel français. Ce système est facilement compatible avec le WGS84 par exemple.

Système WGS84 (World Geodetic System 1984) Système global initialement créé par le département de la défense des États Unis en 1984, mis à jour en 2004. Son exactitude est métrique, et son ellipsoïde se nomme IAG-GRS80.

Referentiel EPSG Référentiel structuré permettant de décrire la position d'un objet sans ambiguïté par la définition du système de coordonnées de référence (CRS) et définissant les transformations et conversions qui permettent de modifier les coordonnées d'un CRS à un autre.

Zonier Traitement de l'ensemble du signal géographique qui constitue le risque assuré. Dans ce mémoire, le signal géographique est expliqué non seulement par les variables externes intégrées directement dans le modèle de prédiction mais aussi par les résidus qui sont ensuite lissés de manière à créer une nouvelle variable géographique qui contient les zones de risques.

Annexe A

Garanties Multirisque Habitation

Voici un descriptif, issu du Guide technique habitation d'AXA, des garanties dégât des eaux et vol du contrat Multirisque habitation abordées dans ce mémoire.

Le dégât des eaux

Principe de base : L'assurance Dégâts des eaux ne couvre pas la réfection des conduites ou des installations d'eau à l'origine du dommage **mais uniquement les conséquences des dommages causés par l'eau.**

L'assurance Dégâts des eaux couvre les conséquences des dommages causés par l'eau résultant :

• De la fuite, de la rupture ou du débordement des conduites d'eau non enterrées.

Les conduites enterrées correspondent aux conduites dont l'accès nécessite des travaux de terrassement. Les conduites encastrées sont garanties. Il s'agit des conduites situées à l'intérieur des murs et des planchers même si elles se trouvent au-dessous du niveau du sol ou si elles passent dans un vide sanitaire.

• De la fuite, de la rupture ou du débordement des appareils à effet d'eau.

Il s'agit des appareils auxquels il est ajouté un élément quelconque qui a pour but de permettre certaines opérations telles que l'arrivée de l'eau, son évacuation, son chauffage, son épuration, son aération, créant alors un certain mouvement d'eau, même s'il n'est pas continu. *Exemples : machines à laver le linge et la vaisselle, baignoires, lavabos, éviers…*

Il est nécessaire que l'équipement soit relié en permanence au dispositif d'alimentation, d'évacuation d'épuration ou de filtration.

Il faut distinguer l'appareil à effet d'eau du récipient qui est un simple réceptacle contenant de l'eau. C'est le cas notamment des vases, des piscines gonflables pour enfant…

• Des infiltrations d'eau ou de neige au travers des toitures, ciels vitrés, terrasses et balcons formant toiture

• Des infiltrations d'eau et de neige au travers des façades et murs extérieurs

Cette garantie ne s'applique pas aux murs de clôture.

• De la rupture accidentelle ou du débordement exceptionnel d'égout, non dus à un évènement climatique

Cette garantie concerne les égouts des voies publiques ou privées lorsqu'il y a une rupture accidentelle ou un débordement exceptionnel d'égout non dus à un évènement climatique.

• Des infiltrations par les joints d'étanchéité aux pourtours des installations sanitaires et au travers des carrelages

L'expression « par les joints d'étanchéité aux pourtours des installations sanitaires » est d'interprétation stricte : elle suppose l'existence d'un joint et ne vise que les joints horizontaux situés entre l'installation sanitaire et le mur.

L'expression « au travers des carrelages » signifie au travers des carreaux proprement dits, mais également au niveau des jointures entre les carreaux que celles-ci soient poreuses ou cassées. Cela s'applique pour les carrelages fixés aux murs et au sol, mais ne vise pas les infiltrations au travers des parquets, linoléum, dalles, carreaux plastiques ou moquette.

• Des dégâts des eaux subis dus à la faute d'un tiers

• Les frais engagés pour la recherche de fuites qui sont à l'origine d'un sinistre garanti à l'intérieur des biens assurés, ainsi que des frais de remise en état des biens dégradés par ces travaux de réfection

Le vol

Définitions

Le vol est la soustraction frauduleuse du bien d'autrui (*article 311-1 du Nouveau Code Pénal*). Le voleur s'arroge la détention matérielle d'une chose sans la volonté de son propriétaire. Le vol est consommé même si les objets soustraits sont ensuite abandonnés ou détruits.

La tentative de vol se définit comme tout acte accompli en vue de commettre un vol qui a reçu un commencement d'exécution mais qui a été suspendu ou manqué pour une cause quelconque.

Conditions d'application de la garantie

Sont garantis le vol et la tentative de vol commis **à l'intérieur des locaux privatifs clos et couverts** de l'assuré, dès lors que ce dernier peut en établir les circonstances détaillées.

Le vol peut avoir été commis notamment par :

• effraction,

• escalade : l'escalade est l'introduction par une ouverture située en étage. Il peut s'agir d'une fenêtre, d'un balcon. Il n'y aura pas effraction si l'ouverture n'était pas fermée,

• fausse clé : sont considérées, à titre d'exemple, comme des fausses clés, les crochets, passe partout, clés imitées, contrefaites ou altérées.

Le vol commis avec les vraies clés qui ont été volées chez le gardien ou perdues par l'assuré est assimilé à un vol commis avec fausses clés. En revanche, la garantie n'est pas acquise si l'assuré a laissé sa clé sous son paillasson, dans une boîte aux lettres, etc...

• maintien clandestin dans les lieux,

• menaces, violences ou intimidation de l'assuré, son entourage ou toute personne à son service : sont visés tous les coups et blessures, quel qu'en soit le résultat.

Annexe B

Performances Gradient Tree Boosting VS Modèle linéaire généralisé

	Modèle linéaire généralisé log-poissonnien			Modèle de gradient tree boosting		
Variables	**GINI Train**	**GINI Test**	**Delta GINI**	**GINI Train**	**GINI Test**	**Delta GINI**
Internes	26,9	26,7	0,2	27,3	26,7	0,6
Internes + maille IRIS et INSEE	33,5	32,6	0,9	33,2	32,9	0,3
Internes + maille ADRESSE	34,2	33,5	0,7	33,6	33,3	0,3
Toutes	34,4	33,8	0,6	33,9	33,5	0,4

TABLE *B.1 – Tableau comparatif des performances d'un Gradient Tree Boosting avec un modèle linéaire généralisé Poissonien avec une fonction de lien logarithme : GINI.*

	Modèle linéaire généralisé log-poissonnien		Modèle de gradient tree boosting	
Variables	**RMSE Train**	**RMSE Test**	**RMSE Train**	**RMSE Test**
Internes	0,2583866	0,2891271	0,2646045	0,2805371
Internes + maille IRIS et INSEE	0,2578160	0,2886096	0,2641126	0,2800929
Internes + maille ADRESSE	0,2577513	0,288509	0,2640436	0,280036
Toutes	0,2577319	0,2885088	0,2640408	0,2800455

TABLE *B.2 – Tableau comparatif des performances d'un Gradient Tree Boosting avec un modèle linéaire généralisé Poissonien avec une fonction de lien logarithme : RMSE.*

Annexe C

Exemple de diagramme des corrélations

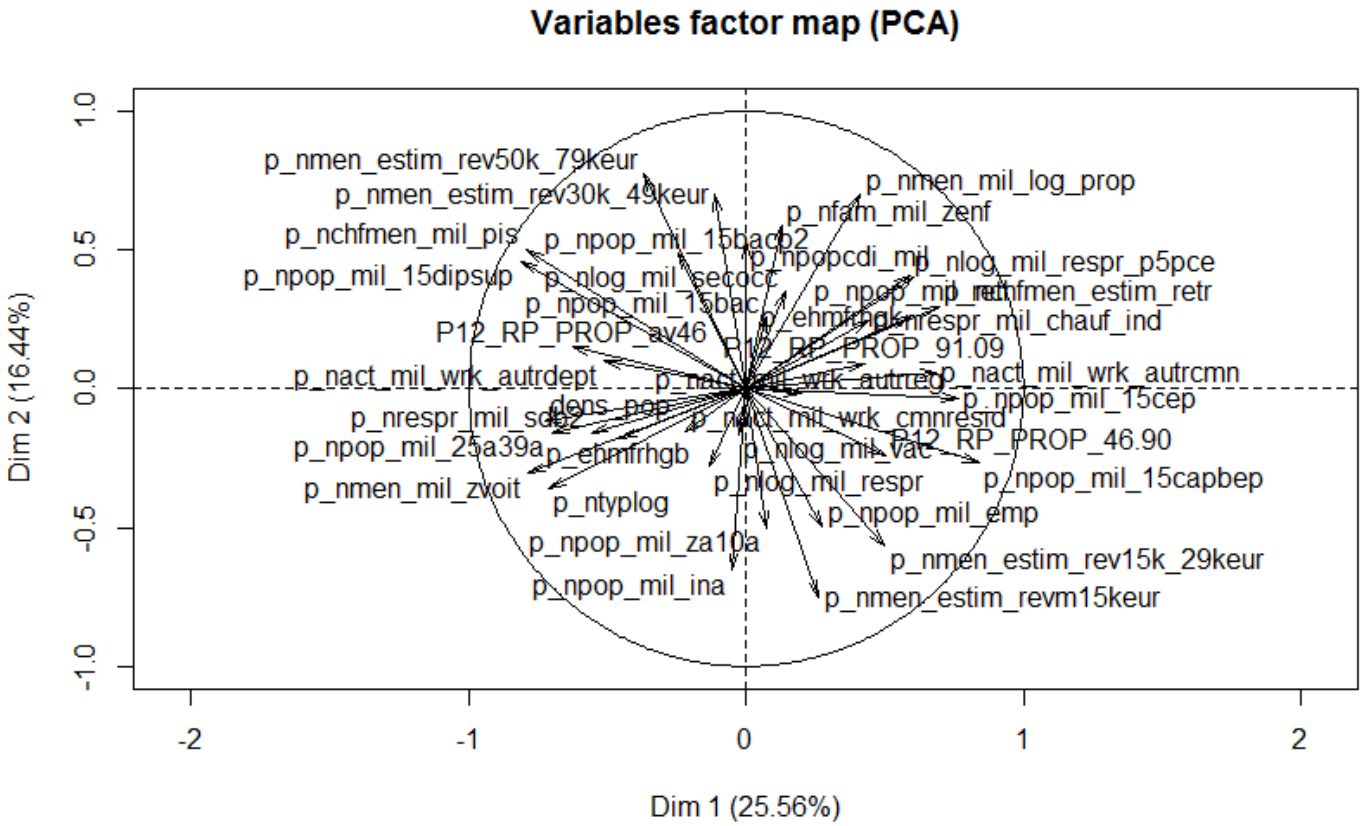

FIGURE *C.1 – Diagramme de corrélations des variables sur les axes 1 et 2.*

Annexe D

Contributions relatives des variables

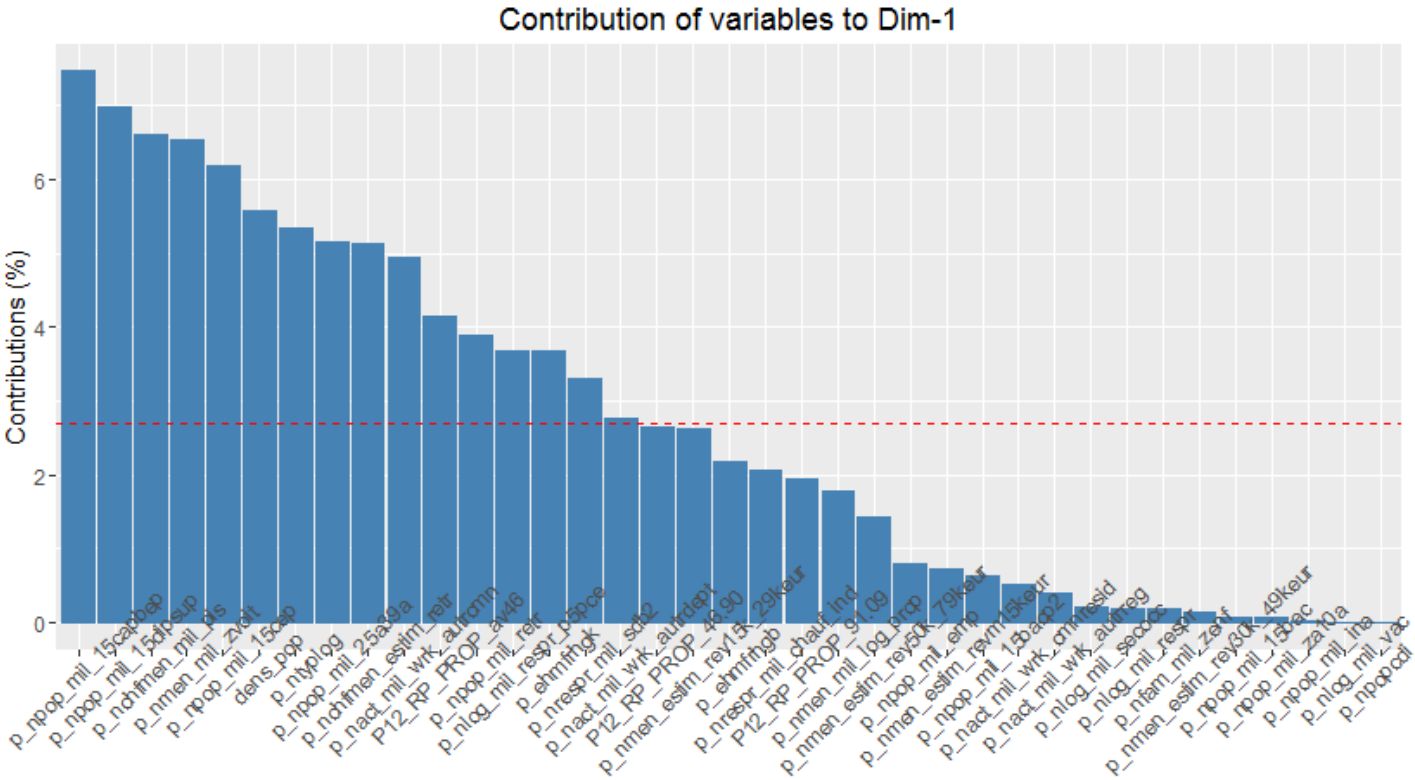

FIGURE *D.1 – Contribution relative des variables explicatives à l'inertie de l'axe 1.*

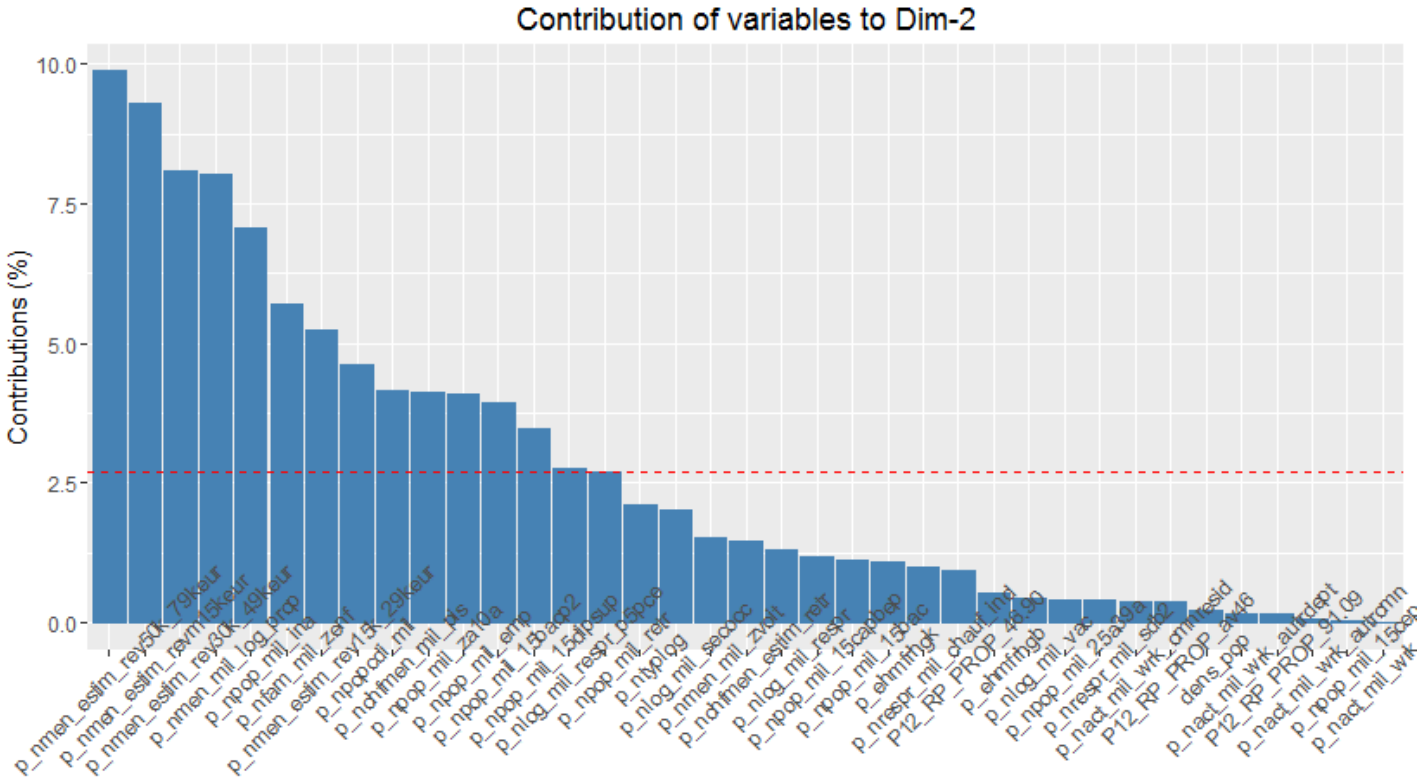

FIGURE *D.2 – Contribution relative des variables explicatives à l'inertie de l'axe 2.*

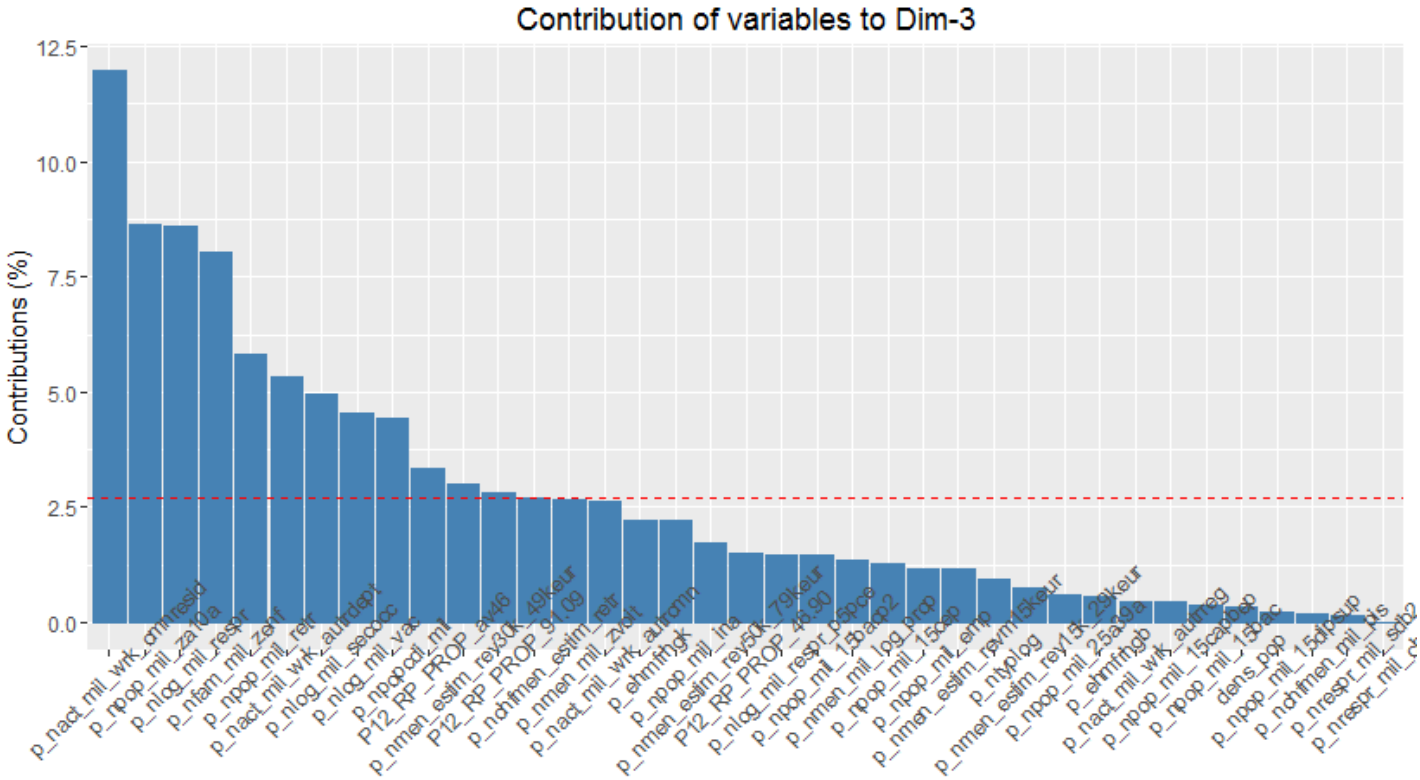

FIGURE *D.3 – Contribution relative des variables explicatives à l'inertie de l'axe 3.*

Variables	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5	Dim.6	Dim.7
Formation CAP BEP	**7,463**	1,184	0,433	0,020	0,036	1,295	1,574
Population travaillant dans une autre commune	4,956	0,058	2,629	0,113	0,958	0,176	1,183
Densité de population	**5,579**	0,211	0,351	0,617	1,589	1,304	0,081
K - Couples matures et seniors	3,676	1,093	2,229	2,521	0,596	1,372	1,922
Population travaillant dans la commune de résidence	0,510	0,381	**11,961**	0,408	1,889	0,125	0,958
Formation CEP	**6,186**	0,021	1,289	0,973	1,546	6,568	0,067
Population travaillant dans une autre région	0,409	0,005	0,450	0,177	3,476	0,041	0,016
Part des résidences prinpales construites en 1991 et 2009	1,940	0,141	2,818	4,370	0,464	2,806	6,094
B - Cosmopolites et branchés	2,180	0,536	0,545	0,001	**8,135**	**11,423**	6,603
Population travaillant dans un autre département	2,766	0,165	5,338	1,071	**8,094**	0,889	0,094
Population 0 - 10 ans	0,060	4,118	**8,644**	4,377	0,782	2,425	0,310
Revenu du ménage : entre 30k et 49k euros	0,134	**8,079**	2,991	1,763	0,235	2,129	0,041
Population inactive	0,026	7,055	2,200	2,432	1,735	**13,125**	0,916
Familke sans enfants	0,177	5,712	**8,036**	2,505	0,447	2,636	0,138
Formation Bac	0,061	1,108	0,378	**14,374**	2,320	1,265	3,002
G - Culture et héritage ouvriers	5,338	2,118	0,942	3,952	1,385	0,236	5,814
Type de contrat Salariés Fonction publique - CDI	0,000	4,613	4,448	1,375	3,696	2,134	0,895
Formation diplômes supérieurs	**6,987**	3,470	0,210	0,001	1,522	0,822	0,837
Chef du ménage : Professions intellectuelles supérieures	**6,613**	4,166	0,177	0,107	0,780	0,084	0,304
Part des résidences prinpales construites avant 1946	4,153	0,377	3,333	**8,505**	0,889	0,687	4,200
Résidences principales Chauffage Central Individuel	2,058	0,996	0,002	0,476	**9,554**	4,446	2,773
Ménages sans voiture	**6,529**	1,528	2,680	0,298	0,619	1,347	0,393
Population retraitée	3,893	2,689	5,807	0,078	3,077	1,389	3,086
Chef du ménage : retraité	5,124	1,465	2,693	2,019	4,396	0,538	4,376
Revenu du ménage : entre 15k et 29k euros	2,628	5,243	0,769	6,502	1,060	1,950	0,084
Revenu du ménage : moins de 15k euros	0,725	**9,294**	1,152	1,990	0,426	0,027	0,859
Résidence principale salle de bain baignoire douche autre	3,298	0,405	0,132	3,601	1,206	0,201	6,256
Résidences principales de plus de 5 pièces	3,678	2,744	1,457	6,165	3,941	0,882	1,426
Nombre de logements secondaires ou occasionnels	0,216	2,009	4,964	0,463	**10,826**	**16,242**	6,057
Part des résidences prinpales construites en 1946 et 1990	2,654	0,936	1,519	5,734	2,641	0,000	**16,721**
Formation Bac +2	0,638	3,942	1,456	**11,416**	3,932	0,256	0,356
Nombre de logements vacants	0,006	0,447	4,548	0,528	7,184	0,620	0,902
Population employée	0,793	4,088	1,163	4,828	4,220	3,683	6,571
Population 25 - 39 ans	5,147	0,406	0,600	3,477	0,946	3,211	6,707
Revenu du ménage : entre 50k et 79k euros	1,438	**9,873**	1,724	0,102	0,401	0,404	0,089
Statut du logement : proriétaire	1,785	**8,029**	1,351	1,811	0,001	0,065	0,834
Nombre de logements : résidences principales	0,180	1,293	**8,581**	0,849	4,996	**13,195**	**7,463**

TABLE *D.1 – Contributions relatives des variables à l'inertie des axes*

Bibliographie

[1] A. ADOR, F. HERAUD, and G. SALHA. Modèles espace-temps du risque de cambriolage : application à la belgique. *Document interne AXA*, Mai 2015.

[2] S. BAILLARGEON. Le krigeage : revue de la théorie et application à l'interpolation spatiale de données de précipitations. *Mémoire de la Faculté des études supérieures de l'Université Laval dans le cadre du programme de maîtrise en statistique*, Avril 2005.

[3] R. BELLINA. Méthodes d'apprentissage appliquées à la tarification non-vie. *Mémoire ISFA*, Janvier 2014.

[4] R. S. BIVAND, E. J. PEBESMA, and V. GÓMEZ-RUBIO. *Applied Spatial Data Analysis with R*. Springer, 2008.

[5] K. P. BURNHAM and D. R. ANDERSON. Multimodel Inference : Understanding AIC and BIC in Model Selection. *SOCIOLOGICAL METHODS & RESEARCH*, 2004.

[6] C. CESE. *Lorenz and GINI*. Document interne AXA, Janvier 2016.

[7] A. CHARPENTIER. Statistique de l'assurance. 2010. 3ème cycle. Université de Rennes 1 et Université de Montréal.

[8] B. COULMONT. Cartographie avec R. Septembre 2010.

[9] E. DADOUN and I. HERBOCH. Modélisation spatio-temporelle du risque de cambriolage. *Document interne AXA*.

[10] M. DENUIT and A. CHARPENTIER. *Mathématiques de l'assurance non-vie*. Economica, 2005. Tome II : Tarification et provisionnement.

[11] Y. FREUND and R. E. SCHAPIRE. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences 55*, 1997.

[12] J. FRIEDMAN. *Greedy function approximation : a gradient boosting machine*. The annals of Statistics, 2001.

[13] J. FRIEDMAN. *Stochastic gradient boosting.* Computational Statistics & Data Analysis, 2002.

[14] T. HASTIE, R. TIBSHIRANI, and J. FRIEDMAN. *The Elements of Statistical Learning,.* Springer, 2008.

[15] T. HENGL. *A Practical Guide to Geostatistical Mapping.* November 2009.

[16] INSEE. Données à la maille insee. `http://www.insee.fr/fr`. Site consulté en Avril 2016.

[17] Institut Géographique National. Systèmes de coordonnées. `http://www.ign.fr/sites/all/files/geodesie_coordonnees.pdf`. Site consulté en Mai 2016.

[18] I. T. JOLLIFFE. *Principal Component Analysis.* Springer-Verlag, 2002.

[19] A. KASSAMBARA and F. MUNDT. Factoextra : Extract and Visualize the Results of Multivariate Data Analyses. `http://www.sthda.com/english/rpkgs/factoextra/` Statistical tools for high-throughput data analysis. Site consulté en Juillet 2016.

[20] J. LARMARANGE, R. VALLO, Y. SEYDOU, P. MSELLATI, and N. MEDA. Méthodes pour carthographier les tendances régionales de la prévalence du VIH à partir des Enquêtes Démographiques et de Santé (EDS). *Cybergeo : European Journal of Geography*, 2011.

[21] S. LEE, S. LIN, and K. ANTONIO. Delta boosting : A boosting application in actuarial science. *Institute of Actuaries of Australia*, August 2015.

[22] C. LOIRET. Refonte du tarif multirisque habitation : construction de microzoniers et intégration de la sinistralité passée à l'adresse. *Mémoire ISFA,* Juillet 2016.

[23] R. LOVELACE. Creating maps in R. `https://github.com/Robinlovelace/Creating-maps-in-R`. Site consulté en Mai 2016.

[24] R. LOVELACE. Introduction to visualising spatial data in R. `https://cran.r-project.org/doc/contrib/intro-spatial-rl.pdf`.

[25] K. MEZIANI. *Modèles Linéaires et ses généralisations.* Cours de Master 1 Mathématiques Appliquées à l'Université Paris Dauphine, 2014.

[26] Observatoire National de la Délinquance et des Réponses Pénales. Données publiques départementales. `https://www.data.gouv.fr/fr/organizations/observatoire-national-de-la-delinquance-et-des-reponses-penales-ondrp/#datasets`. Site consulté en Juin 2016.

[27] E. PEBESMA. The meuse data set : a brief tutorial for the gstat R package. March 2016.

[28] F. PLANCHET. Le modèle collectif. *Cours d'Assurance Non-Vie ISFA*, Octobre 2003.

[29] F. PLANCHET and G. SERDECZNY. *Modèles fréquence – coût : Quelles perspectives d'évolution ?* Institut des Actuaires, Mars 2014.

[30] G. RIDGEWAY. Generalized boosted models : A guide to the gbm package. Août 2007.

[31] A. SURU. *Principes de base de l'assurance dommages*. Cours de Master 2 Actuariat à l'Université Paris Dauphine, 2015.

[32] R. TIMOFEEV. *Classification And Regression Trees*. Master thesis, Humboldt University, 2004.

[33] Tutoriel QGIS. Des coordonnées, oui mais dans quel système ? `http://www.ades.cnrs.fr/tutoqgis/02_02_coord.php`. Site consulté en Mai 2016.

Liste des figures

Liste des tables