PROJET DE STATISTIQUE APPLIQUÉE

Modèles espace-temps du risque de cambriolage : application à la Belgique

Étudiants :
Augustin Ador
Florian Héraud
Guillaume Salha

Encadrants :
M. Philéas Condémine
M. Guillaume Gorge

Mai 2015

Sommaire

Remerciements

Nous tenons à remercier l'ensemble des personnes qui ont participé, d'une certaine manière, au bon déroulement de notre projet de statistique appliquée. Plus particulièrement, nous tenons à remercier :

— Monsieur Philéas Condémine, actuaire et data scientist chez AXA Global P&C, qui nous a encadré avec enthousiasme durant sept mois, qui était toujours très disponible pour répondre à nos questions, et grâce à qui nous avons beaucoup appris ;

— Monsieur Guillaume Gorge, chef de groupe P&C Retail Lines chez AXA Global P&C, pour ses conseils très pertinents lors de notre présentation de mi-parcours et pour l'ensemble des documents fournis ;

— Messieurs Elie Dadoun et Ivan Herboch, étudiants à l'École Centrale de Paris, réalisant également un projet au sein d'AXA sur une problématique similaire à la nôtre (mais en France) et que nous avons eu le plaisir de rencontrer. Nous les remercions notamment pour leur coup de main durant notre travail de géocodage des points d'intéret via OpenStreetMap ;

— l'équipe pédagogique de l'ENSAE ParisTech qui assure la partie théorique de notre formation, et notamment Monsieur Vincent Cottet et Madame Malika Zakri qui ont géré toute l'année le module de statistique appliquée.

1 Introduction générale

1.1 Le risque de cambriolage en Belgique

Le cambriolage est défini comme l'effraction du domicile dans le but d'y commettre un vol. Il constitue une atteinte non autorisée au droit de propriété, le cambrioleur étant passible d'une peine d'emprisonnement pouvant aller jusqu'à dix ans selon le Code Pénal belge. Malgré cette peine théorique élevée, le risque de cambriolage est en augmentation nette depuis plusieurs années en Belgique. C'est une réelle problématique pour AXA qui, en tant qu'assureur, est directement impacté par cette évolution du risque.

1.1.1 Chiffres clés

La Police Fédérale belge a recensé plus de 75 000 cambriolages au sein d'habitations au cours de l'année 2012 [1]. Ce chiffre a augmenté de 7.5% par rapport à 2011, et de 28% par rapport à 2008. Cette hausse est l'une des plus élevées d'Europe de l'Ouest et les chiffres quasi-définitifs de l'année 2013, stables, tendent à confirmer cette tendance. La crise économique a, selon la Police Fédérale belge, un rôle important mais n'explique pas tout. L'envolée du prix de l'or rendant le vol de bijoux plus attractif ou encore le fait que les peines réelles (dépassant rarement deux ans de prison et n'étant parfois pas appliquées) soient inférieures aux peines théoriques peuvent également constituer des éléments explicatifs.

De même, une étude du bureau Research Solution datant de 2012 a mis en avant le relatif manque de prévoyance des Belges, et notamment des moins de 35 ans, vis-à-vis des cambriolages. Par exemple, 60% des Belges ne disposeraient que d'une simple serrure de sécurité pour leur entrée principale. Cette insouciance face à un risque réel, financier mais aussi psychologique, demeure aujourd'hui une réalité. Malgré tout, de plus en plus d'individus renforcent la sécurité de leur habitation, notamment en investissant dans des systèmes d'alarmes (faisant fuir 95% des voleurs) dont les ventes sont en hausse de 10% ces deux dernières années. Pour 2012 et 2013, notons enfin que la hausse se situe surtout dans les maisons, par opposition aux appartements, et surtout dans les zones rurales et en journée. Un tiers de ces cambriolages recensés étaient des tentatives qui n'ont pas abouti. La plupart des infractions ont eu lieu dans les grandes villes : les villes de Bruxelles (9252 cambriolages en 2013), Liège (9086), Charleroi (7231), Anvers (6904) et Asse (5428) regroupent en effet près de la moitié des cambriolages déclarés en 2013.

1.1.2 AXA et la Belgique

Avec un chiffre d'affaires de 154.6 milliards d'euros en 2013 d'après son rapport financier annuel, AXA est le premier groupe d'assurance dans le monde. Depuis 2010, la branche AXA Global P&C assure le pilotage de l'activité dommages du groupe. L'abréviation P&C renvoie aux termes "Property & Casualty" ; l'abréviation IARD pour "Incendie, Accidents et Risques Divers" est aussi employée. Les assurances IARD couvrent les dommages et la protection des biens, par opposition aux assurances de personnes. Nous retrouvons ici les assurances automobile, ainsi que les assurances habitation et donc notre problématique du risque de cambriolage. La réassurance

1. L'ensemble des sources d'où sont issus les chiffres de cette partie sont reportées en fin de rapport.

ainsi que les risques naturels sont également couverts par la branche AXA Global P&C.

En Belgique, AXA est leader du marché de l'assurance dommages avec 18.7% de part de marché en 2013. Cette année-là, le groupe a enregistré un chiffre d'affaires de 2,025 milliards d'euros dans ce domaine sur le territoire belge. En outre, en assurance vie, épargne et retraite, le groupe est troisième avec une part de marché de 14% et un chiffre d'affaires de 2,012 milliards d'euros en 2013. Ces résultats, notamment dans le domaine de l'assurance dommages qui nous intéresse plus particulièrement ici, montrent l'importance de la place qu'occupe AXA au sein de ce pays. Ils justifient la volonté des actuaires du groupe de s'améliorer encore davantage en comprenant mieux, grâce aux méthodes statistiques, les risques auxquels ils sont confrontés afin de les évaluer correctement. C'est en particulier le cas en ce qui concerne la problématique du risque de cambriolage, qui est étudiée au sein de ce projet. Analyser les caractéristiques spatiales et temporelles de la criminalité vol permettra ainsi de mettre en place des méthodes qui permettront de mieux appréhender ce risque.

Avant d'entrer dans le détail de la formalisation, une revue de la littérature théorique sur la criminalité est réalisée ci-dessous, afin de mieux cerner la problématique. Cette revue permet de faire ressortir un certain nombre de pistes pour expliquer le risque de cambriolage, à tester par la suite en utilisant les données réelles d'AXA.

1.2 Revue de la littérature sur la criminalité

Après une approche microéconomique des raisons pouvant pousser des cambrioleurs à l'acte, cette sous-partie montre que les aspects spatiaux et temporels sont primordiaux dans l'analyse du risque de cambriolage. Les facteurs socio-économiques occupent notamment une place importante.

1.2.1 Analyse microéconomique des comportements criminels

L'application de l'analyse microéconomique aux comportements criminels s'est développée au cours du XXème siècle. Le modèle le plus célèbre est celui de Gary S. Becker (1968). La criminalité y est présentée comme étant le choix d'agents rationnels qui maximisent leurs utilités. Les criminels essaient ainsi de maximiser leur bien-être physique ou la reconnaissance sociale que leur procurent ces activités criminelles. Pour Becker, un individu commet un crime dès que son espérance d'utilité est positive :

$$E(U) = qB - Cp > 0$$

$B \geq 0$ représente l'utilité apporté par le crime s'il réussit et $C \geq 0$ son coût (sa désutilité) en cas de condamnation. De plus, $p \in [0; 1]$ représente la probabilité que l'agent se fasse attraper, et $q \in [0; 1]$ la probabilité de succès du crime. Notons qu'un crime peut réussir mais que le criminel peut être condamné par la suite, l'un n'empêche pas l'autre.

La détermination de ces valeurs est bien évidemment complexe, même pour le criminel lui-même. Celui-ci est confronté à un raisonnement bayésien : il prendra la meilleure décision étant donné l'information qu'il possède et la valeur des variables qu'il pense être la bonne. Il est possible de remarquer, comme Cornish et Clark (1987) que p est une probabilité subjective qui varie selon les individus et leurs différentes classes sociales. Si p est très faible, la dureté de la peine n'aura

que peu d'effet dissuasif, ce qui explique qu'un assassinat peut être rationnel lorsque le criminel est certain de ne pas être condamné par la suite. Dans le même temps la valeur du bénéfice du crime est elle aussi subjective. Voler 100 euros peut être attractif pour les individus les plus pauvres mais pas pour les autres. Considérer ces aspects conduit à poser des restrictions sur les fonctions d'utilités des agents, telles que la concavité de celles-ci qui conduit à des utilités marginales décroissantes.

La classe sociale de l'individu est elle aussi très importante pour expliquer la décision ou non de réaliser un crime. L'obéissance aux lois est une norme plus ou moins acceptée chez les différents groupes sociaux ce qui influe sur la décision de l'individu de réaliser ou non un crime. Fishbein (1967) avance que le mimétisme avec les autres membres du groupe peut ainsi encourager au crime ; il prend l'exemple de la fraude à l'impôt. Suivre la norme du groupe est alors vu comme une stratégie rationnelle dans un jeu récurrent. Toutes ces influences sociales peuvent être appréhendées par l'équation ci-dessus.

Dans le modèle que nous allons développer nous ne pouvons que difficilement obtenir des données sur le criminel lui-même. Nous pouvons toutefois supposer, à partir de ce modèle, que les lieux sinistrés sont ceux où l'espérance d'utilité sera la plus importante i.e. ceux pour lesquels la facilité ou le bénéfice des cambriolages sera le plus grand. Nous pouvons alors penser à une liste de variables pertinentes à tester :

- **la richesse du quartier** qui peut augmenter le gain moyen pour un cambrioleur si le crime réussit ;
- **l'équipement de la victime potentielle en termes de sécurité** qui peut diminuer la probabilité de succès du vol : serrures, alarmes...
- **la proximité avec certains établissements publics** tels que les commissariats, mais aussi de manière plus générale **l'isolement** relatif de l'habitation par rapport au reste de la population, qui peuvent, de même, faire varier la probabilité de se faire attraper ;
- **la météo** ou encore **l'état de la chaussée** qui, s'ils sont "mauvais", peuvent décourager le cambrioleur. L'aspect météorologique peut être contrôlé à partir de la température au moment du crime ou d'informations sur les précipitations.

1.2.2 Importance de l'analyse spatio-temporelle

Essayer de prévoir la criminalité est un enjeu crucial afin de pouvoir tenter de la réduire. Néanmoins, modéliser ce type de risque sans étudier les aspects spatiaux et temporels conduirait à une analyse incomplète. En effet, l'influence spatiale de la criminalité est bien établie même si elle peut intervenir de façon complexe. L'analyse des données spatiales sur la criminalité trouve sa source au début du XIXème siècle en Europe avec Guerry (1833) et Quételet (1835). Il ressort alors de ces recherches deux grands points : la criminalité dépend bien des caractéristiques géographiques, l'occurence des sinistres n'est pas due au hasard.

La cartographie des crimes permet de connaître où vivent les criminels, où sont localisées les cibles privilégiées, comment les criminels se rendent sur les lieux du crime et où sont localisés les nouveaux points chauds. Toutes ces informations permettent d'avoir une approche de prévention typée par quartier.

À partir d'analyses spatiales il est possible d'aboutir à une réelle théorie des schémas de criminalité comme celle de Brantingham et Brantingham (1982) : **les criminels agissent souvent dans des lieux qu'ils ont l'habitude de traverser** que ce soit pour aller travailler, accom-

pagner les enfants à l'école ou retrouver des amis. Morenoff, Sampson et Raudenbush (2001) ont également démontré que **l'inégalité des ressources socio-économiques** entre les quartiers des villes est corrélée avec répartition spatiale de la criminalité. Parmi les variables qui peuvent être intéressantes à tester pour notre étude, citons :

- **la frontière zone riche/zone pauvre**, Rengert et al. (1999), Rossmo (1999) ;
- **la richesse du quartier**, Kennedy et Forde (1990), Bursik et Grasmick (1993), Pratt (2001) ;
- **le taux de chômage**, Roundtree et Land (2000), Hartnagel (2004),
- **le système de protection dans le quartier**, Paternoster et Bushway (2001) ;
- **l'évolution du nombre d'habitants dans la ville et la part de minorités**, Ceccato, Haining, et Signoretta (2002).

Attardons nous par exemple sur le premier point : la frontière entre une zone riche et une zone pauvre. Rengert et al. (1999) ainsi que Rossmo (2000) ont mis en avant que la sur-criminalité n'est pas directement liée à des caractéristiques socio-économiques mais davantage à la proximité de certaines zones riches avec des zones plus défavorisées. Ces études montrent que la sinistralité est maximale à quelque distance du lieu d'habitation du voleur, un phénomène désigné sous le terme de *Distance Decay Effect*. Par exemple, si nous considérons l'étude de Savoie (2008) prenant en compte des données sur la ville de Montréal en 2001, il apparaît que la distance médiane parcourue par les auteurs présumés est de 2.6 kilomètres.

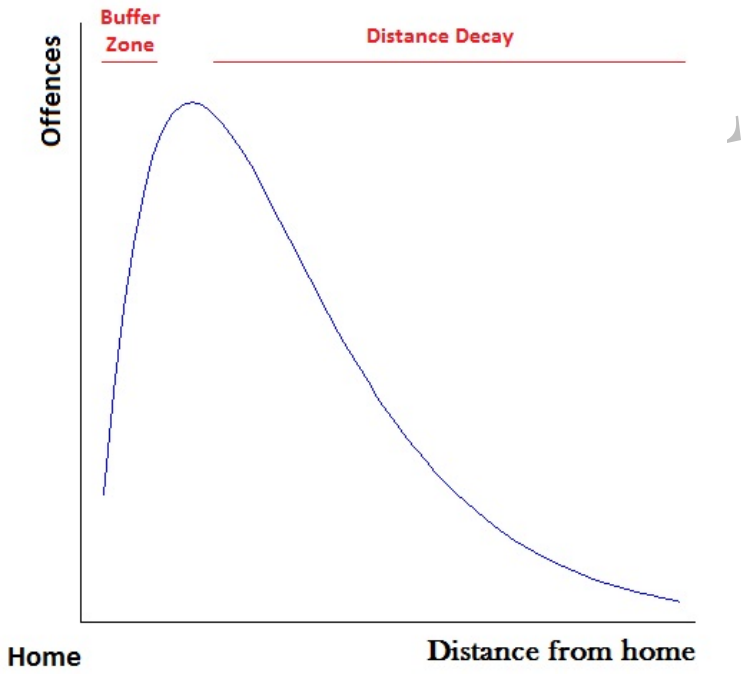

FIGURE 1 – Illustration du Distance Decay Effect

Coupler cette analyse spatiale à une analyse temporelle permet d'appréhender encore mieux les schémas de cambriolage. Les variables de **saisonnalité des événements** peuvent alors être étudiées : la météo comme cité précédemment, mais également **le jour de la semaine** voire **les horaires**. Par exemple dans la ville de Cincinnati, d'après les travaux de Feng (2010), il apparaît que les cambriolages sont plus nombreux le vendredi et le samedi soir.

Aussi, il peut être très intéressant d'analyser la présence de **répétition** et de **répétitions proches** qui illustre bien la pertinence du couplage spatio-temporelle de l'étude. Nous parlons de répétition lorsqu'une adresse précise est attaquée plusieurs fois et de répétition proche lorsqu'un sinistre a lieu à un emplacement très proche d'un précédent cambriolage. Ces phénomènes se retrouveront-ils au sein des données d'AXA ?

L'idée est que, lorsqu'un crime a été commis, il est plus facile de le répéter que d'identifier un autre endroit. Ericsson (1995) a mis en avant que 76% des voleurs questionnés sont retournés, après une certaine période de temps, cambrioler jusqu'à cinq fois le même lieu. Cependant dans les résidences aisées, il y a de grandes chances pour que le propriétaire investisse dans des systèmes de protection. Ainsi la probabilité de répétition diminue. Cela va favoriser la répétition proche qui peut être expliquée par les caractéristiques que les habitations proches ont en commun : même immeuble, même route d'accès... De plus, la rapidité d'exécution entre deux cambriolages peut se comprendre par ces caractéristiques qui peuvent changer rapidement (nouveau système d'alarme de l'immeuble, installation de caméras...), diminuant la probabilité de réussite du cambrioleur. Au contraire les propriétés dans les endroits les plus démunis auraient un risque élevé de répétition.

1.3 Orientation du projet

Au sein de ce projet, notre objectif principal est de savoir s'il y a un effet de dépendance spatio-temporelle de la sinistralité. Autrement dit, sachant qu'un cambriolage s'est produit à un certain lieu et à une certaine date, cela influe-t-il sur le risque de cambriolage pour le voisinage (spatial mais aussi temporel) ? Pour cela, nous avons à notre disposition des données AXA, présentées ci-dessous, ainsi que des données complémentaires obtenues par nous-mêmes. L'objectif de cette recherche complémentaire était de pouvoir tester le maximum d'intuitions possibles et de capter davantage les éléments ayant un impact sur ce risque de cambriolage, d'après la littérature théorique sur la criminalité du vol. Ainsi, nous pouvons être en mesure d'étudier l'effet de dépendance spatio-temporelle une fois pris en compte les effets structurels de sinistralité locale : est-ce une zone à risque ? Est-ce une période de temps à risque ? Voire un individu à risque ?

Après avoir présenté, dans un premier temps, les données utilisées au cours de ce projet, nous proposons un modèle spatial afin de tenter de répondre à la problématique. Celui-ci reprend les aspects théoriques évoqués en introduction. Le modèle est ensuite appliqué à nos données. Enfin, nous tentons d'élargir le cadre d'analyse en testant l'efficacité de modèles alternatifs non paramétriques issus du *machine learning*. Nous discutons les différents résultats obtenus, en terme de prédiction du risque de cambriolage.

2 Les données

Dans cette partie, nous présentons les données à notre disposition durant ce projet. Nous présentons non seulement les données fournies par AXA, mais également notre démarche de recherche de données externes afin de compléter celles-ci. Avant d'entrer dans le détail de la modélisation spatio-temporelle du risque de cambriolage, quelques statistiques générales sur ces données sont proposées.

2.1 Présentation des données d'AXA

La base de données fournie par AXA comporte des informations sur 374 397 assurés en habitation en Belgique, sur une période de quatre ans allant de 2010 à 2013. 232 variables sont répertoriées au sein de cette base.

Nous avons à notre disposition des informations générales sur chaque habitation assurée : son adresse, sa surface, une indicatrice précisant si elle est "adjacente" ou "isolée" des autres habitations (score réalisé en interne par AXA), une indicatrice précisant s'il s'agit d'une maison ou d'un appartement, mais également le nombre de cuisines ou encore la présence éventuelle d'une piscine extérieure. Notons que nous avons dû réaliser un travail d'anonymisation des adresses : pour chaque habitation, nous n'avons pas gardé le numéro exact de la rue. Nous avons toutefois créé une variable indiquant si celui-ci est pair ou impair, pour nous permettre d'étudier l'impact éventuel du côté de la rue sur le risque de cambriolage.

Nous avons également à notre disposition des informations sur l'assuré, comme son identifiant, son âge, l'âge du contrat ou encore une distinction propriétaire/locataire. Précisons que nous avions généralement, au sein de la base, plusieurs lignes pour un même individu. En effet, suivant le principe d'exposition du contrat en assurance, AXA a créé une nouvelle ligne pour l'individu dès lors que celui-ci subissait un sinistre ou bien modifiait son contrat.

De plus, chaque habitation a été regroupée au sein d'une zone dite *mosaic*, zones que nous étudierons plus en profondeur dans la suite du rapport. Nous avons des informations sur les caractéristiques socio-économiques de ces zones. Le revenu moyen et médian par ménage, le nombre d'habitants, le nombre de familles, le nombre de propriétaires, le taux de maison ou encore une décomposition par âge font parties de ces variables. Nous possédons également les mêmes informations à l'échelle de la rue, et nous possédons également certains quantiles associés à ces variables.

Enfin, nous avions également, pour chaque ligne, une variable indicatrice précisant si un cambriolage a eu lieu (0 ou 1). Comme nous avons regroupé les lignes par individus, cette variable devient une variable de comptage et il est donc tout à fait possible qu'un même individu se fasse cambrioler plusieurs fois dans la même année (en pratique, nous verrons plus loin que le nombre de cambriolages par an ne dépasse jamais 3 dans nos données). Nous avons bien entendu la date du sinistre, ainsi que des informations relatives aux montants estimés du sinistre et aux montants remboursés par l'assureur. Ceux-ci sont à manipuler prudemment, ils peuvent en effet grandement varier suivant le contrat de l'assuré.

2.2 Recherche de données externes

Ces données ont été complétées lors d'une étape de recherche de données. Nous présentons ici l'étape de géolocalisation de l'ensemble des assurés du portefeuille, la recherche de points d'intérêts (localisation des postes de police, des caméras de surveillance...) via OpenStreetMap et enfin le travail réalisé pour obtenir des données météorologiques.

2.2.1 Géolocalisation des assurés

Afin de pouvoir étudier l'aspect spatial du risque de cambriolage, nous avons besoin de pouvoir situer les habitations assurées sur une carte et repérer les points proches les uns des autres. Or, ceci n'était pas possible avec la base de données actuelle, au sein de laquelle nous disposions uniquement des adresses postales des habitations. Nous avions besoin des **coordonnées GPS** de celles-ci, c'est-à-dire les longitudes et latitudes.

Il est possible de les regrouper en passant par l'API Google Maps, via le package *RGoogleMaps* du logiciel R. Cependant, cette piste a vite été abandonnée car Google limite le nombre de requêtes quotidiennes à 2500 (rappelons que nous avons 374397 adresses différentes d'assurés dans la base).

Nous nous sommes ensuite tournés vers OpenStreetMap, qui est un projet collaboratif ayant pour but de constituer une base de données géographiques libre, à l'échelle de la planète entière. Un package utilisant l'API OpenStreetMap pour géocoder n'existant pas, nous avons conçu un code permettant, à partir de l'adresse d'une habitation, de lui associer ses coordonnées GPS telles qu'elles sont référencées sous OpenStreetMap. Ainsi, nous contournons le problème de la limite de l'API Google Map, pour un résultat restant très bon.

2.2.2 Recherche de données complémentaires via OpenStreetMap

Revenons à OpenStreetMap. Grâce à ce projet mettant à disposition l'ensemble des données, nous avons également pu obtenir les coordonnées GPS d'un grand nombre de points d'intérêts, pour l'ensemble du territoire belge [2]. Parmi ces points d'intérêts, nous obtenons par exemple les localisations des :

- postes de police ;
- caméras de surveillance ;
- lieux de culte ;
- supermarchés...

Il sera intéressant de vérifier plus tard si, par exemple, la proximité d'un poste de police réduit le risque de cambriolage.

Concernant le traitement des données, notons que celles-ci ont été obtenues directement à partir de : http ://download.geofabrik.de/. Elles étaient au format .shp (Shapefile) et ont été traitées sur R afin d'obtenir les coordonnées GPS en format .csv.

2. Avec, toutefois, plus ou moins d'informations suivant les lieux. Rappelons qu'il s'agit d'un projet collaboratif, ne garantissant absolument pas l'exhaustivité des données proposées.

2.2.3 Données météorologiques

Ayant à notre disposition les dates et localisations des sinistres nous pensions qu'il serait intéressant d'observer si la météo influe sur le risque de sinistre ou non. Pour pouvoir réaliser cette étude nous avions besoin des données quotidiennes de température moyenne et de précipitation au niveau de chaque commune. Ces données nous ont été fournies gracieusement par l'Institut Royal Météorologique de Belgique.

2.3 Limites des données

Plusieurs aspects ayant une influence probable sur le risque de cambriolage, parfois cités au sein de la littérature sur le sujet, ne pourront pas être étudiés au cours de ce projet. Notamment, nous n'avons pas de précision sur l'heure du cambriolage, ce qui aurait pu être intéressant. De plus, nous n'avons pas réussi à obtenir d'informations sur l'état de la chaussée au niveau de l'habitation assurée, un point que nous avions pourtant évoqué en première partie. Il est aussi possible de penser que le géocodage des habitations, obtenu via OpenStreetMap, est parfois légèrement incorrect. Enfin, il est très difficile d'obtenir davantage d'informations sur l'habitation voire sur l'assuré lui-même. Pourtant, il aurait été extrêmement précieux de savoir si la porte d'entrée de l'assuré est blindée ou si une alarme est installée !

3 Statistiques générales

Avant d'entrer plus en profondeur dans l'analyse spatiale du risque de cambriolage en Belgique, cette partie vise avant tout à faire un tour d'horizon des données à notre disposition. Nous commençons par des statistiques univariées et bivariées sur les principales variables, puis nous mettons en avant l'importance des caractéristiques de l'habitation, de la météorologie ou encore de l'aspect temporel dans l'étude du risque de cambriolage. Nous proposons finalement, en tant qu'outil de visualisation, une représentation cartographique de la criminalité en Belgique à partir de nos données.

3.1 Analyse descriptive

3.1.1 Tour d'horizon de la base de données

Nous allons commencer par analyser le portefeuille d'AXA, cela nous permettra d'obtenir une idée d'ensemble de nos données. L'âge moyen des assurés est de 53.5 ans, les assurés ont de 20 ans à 92 ans, avec un premier quartile à 41 ans et un troisième quartile à 66 ans. La plupart des habitations sont des maisons (c'est le cas de 75% d'entre elles) ; 34,7% des habitations sont jugées *isolées* par AXA. Par ailleurs, 74% des assurés sont propriétaires de leur habitation. Il est intéressant d'observer que les catégories minoritaires possèdent cependant un effectif assez important pour pouvoir par la suite analyser ces variables.

Le taux de sinistrés par an se situe à un peu moins de 1% : très exactement, il est de 0.84%. Quelques rares individus (104) se sont fait cambriolés deux fois dans la même année, un individu s'est même fait cambriolé trois fois. Regardons l'évolution temporelle du taux de sinistre et du nombre de contrats sur les quatre années de notre portefeuille. Nous constatons pour ces deux variables une légère augmentation de 2010 à 2012 puis une baisse en 2013. Il est donc difficile de pouvoir donner une conclusion sur une éventuelle tendance temporelle, les deux variables semblent au contraire assez stables.

Année	Taux de sinistrés	Nombre de contrats (% du total)
2010	0.81%	252 927 (23.3%)
2011	0.90%	272 054 (25.0%)
2012	0.90%	281 846 (25.9%)
2013	0.86%	279 960 (25.8%)

TABLE 1 – Évolution du nombre de contrats et des sinistres

Voici ci-dessous les dix villes de Belgique ayant le plus fort taux de sinistrés sur les quatre années, avec pour chacune d'être elles des informations, notamment sa localisation dans le pays si nécessaire. Nous n'avons pas souhaité afficher les villes avec moins de 10 000 assurés puisque les conclusions pourraient alors être hasardeuses. Sachant que le taux de sinistre moyen du portefeuille est de moins de 0.9%, ces villes semblent davantage sujettes au cambriolage. Charleroi possède le risque le plus élevé, suivi des banlieues de Bruxelles et de Liège. Nous notons donc qu'il y a plusieurs grandes villes du pays et que, en dehors de Bruxelles et sa banlieue, ce sont des villes situées en Wallonie.

Commune	Nb vols	Nb assurés	**Ratio**	Détails	Nb d'hab
Charleroi	716	33517	**2.1%**		204 670
Uccle	372	20745	**1.8%**	Banlieue Sud Brux	79 766
Anderlecht	220	13261	**1.7%**	Banlieue Ouest Brux	112 258
Ixelles	280	17335	**1.6%**	Banlieue Sud Est Brux	84 073
Molenbeek-Saint-Jean	162	10150	**1.6%**	Banlieue Ouest Brux	94 798
Liège	464	32886	**1.4%**		197 013
Seraing	142	10607	**1.3%**	Banlieue de Liège	63 968
Bruxelles	265	19801	**1.3%**		177 849
Woluwe-Saint-Lambert	161	12035	**1.3%**	Banlieue Est Brux	51 937
Mons	233	18133	**1.3%**	Ouest de Charleroi	93 366

TABLE 2 – Top 10 des villes aux taux de sinistrés les plus importants (2010-2013)

Notons aussi que, comme le montre le tableau ci-dessous, **la variable indiquant le nombre de cambriolages dans l'année est équidispersée** : la moyenne est en effet très proche de la variance.

Moyenne	Variance
0.0084	0.0085

TABLE 3 – Équidispersion du nombre de cambriolages par an

3.1.2 Influence des caractéristiques de l'habitation

Examinons désormais si l'isolement de l'habitation, le type (appartement ou maison) et le fait d'être propriétaire ont un impact sur le risque de cambriolage. Un diagramme en bâton est reporté ci-dessous afin de visualiser les résultats concernant l'isolement. Pour le type d'habitation et la distinction propriétaire locataire, ces graphiques semblables sont présentés en annexe 1.

La variable d'isolement semble très discriminante. Il est possible de penser que, **plus l'habitation est isolée, plus le cambrioleur est enclin à commettre un vol** puisqu'il aura sans doute moins de chance de se faire voir. Notons cependant que l'écart des taux de cambriolage semble avoir tendance à augmenter au cours du temps, et que l'année 2010 se démarque des autres par l'absence de différence entre habitations isolées ou non.

Par ailleurs, **les maisons ont un taux de cambriolage plus élevé que les appartements** (sauf en 2010 une nouvelle fois), cela peut peut-être s'expliquer par les caractéristiques des appartements :systèmes de sécurité plus nombreux (un ou plusieurs digicodes dans l'immeuble), présence d'une porte unique et proximité des voisins. Les maisons sont aussi habitées, en moyenne, par des personnes plus aisées ce qui pourraient également tenter les cambrioleurs. Enfin, **les propriétaires sont davantage sinistrés que les locataires**. La corrélation peut être due à une richesse plus élevée du cambriolé, lui permettant d'acheter un logement plus isolé par exemple. Nous pouvons noter que pour ces variables, les différences entre les modalités semblent aussi augmenter au cours du temps.

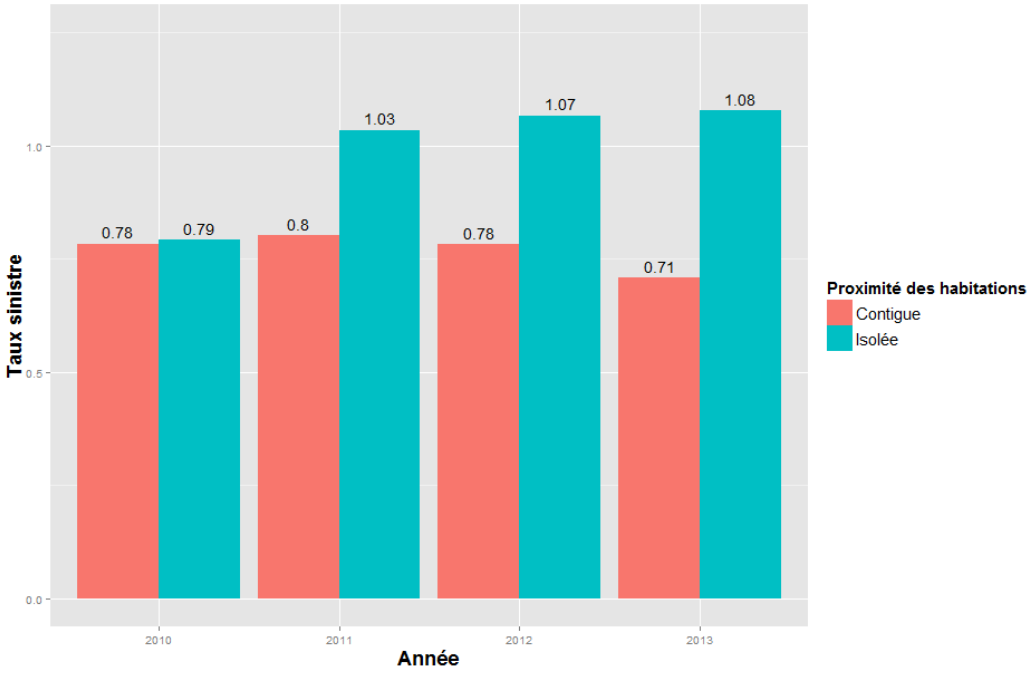

FIGURE 2 – Taux de sinistres et isolement de l'habitation

3.1.3 Influence de la météorologie

Analysons maintenant le lien entre météorologie et cambriolage. Nous pouvons tout d'abord constater que **37% de nos cambriolages ont été effectués lors d'un jour de pluie**, cela semble faible puisqu'il pleut plus d'un jour sur deux en Belgique (55% en moyenne). Nous pouvons penser qu'un cambrioleur préférera agir lorsqu'il fait beau (meilleure visibilité, sol non glissant, moins de risque de trace...).

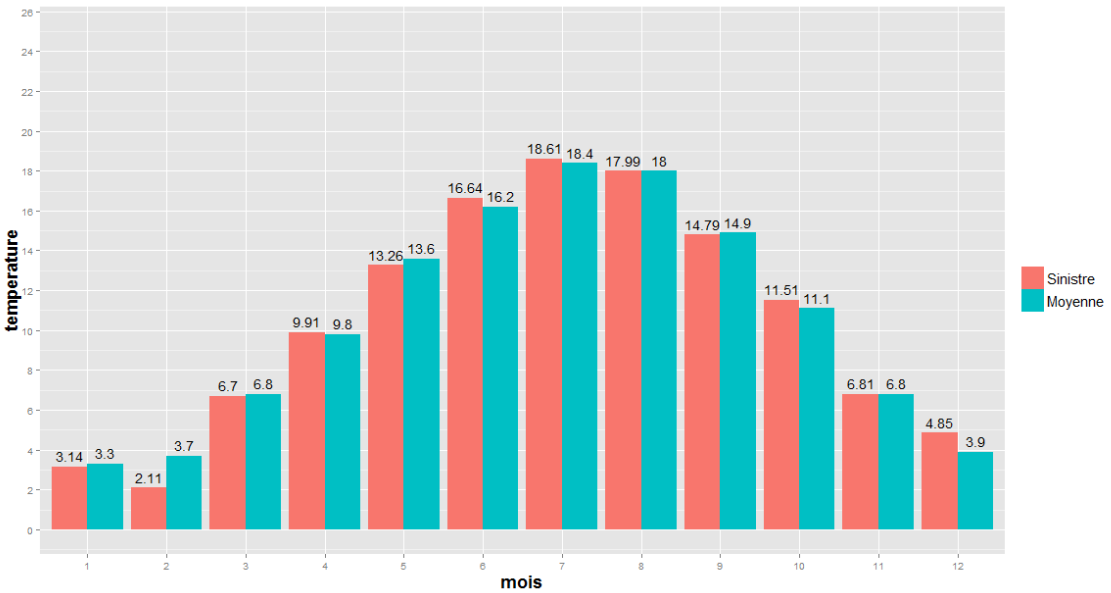

FIGURE 3 – Taux de sinistres et température

Afin d'affiner cette analyse, nous avons voulu comparer les **précipitations moyennes** des jours où ont eu lieu nos sinistres aux précipitations moyennes mensuelles. La corrélation entre cette variable et le risque de cambriolage semble peu claire, d'après le diagramme en bâton reporté en annexe 4. En outre, nous avons réalisé un autre graphique similaire présenté ci-dessus en prenant cette fois la **température** comme variable d'intérêt. Les valeurs sont très similaires pour chaque mois, à nouveau il est difficile de conclure.

3.1.4 Analyse en Composantes Principales des zones *mosaic*

Les habitations de notre base de données ont toutes été regroupées par AXA au sein de zones dites *mosaic*. Ce découpage est réalisé suivant un principe analogue à celui des zones IRIS en France. De nombreuses informations socio-économiques sur chacune de ces zones ont été retrouvées par AXA. Nous avons notamment le revenu moyen de la zone, l'âge moyen des habitants, la densité de population, le taux de maisons, le taux de propriétaires ou encore le taux d'habitations jugées isolées (ce dernier indicateur étant à prendre avec des pincettes, car il n'est construit que sur les assurés AXA et peut donc ne pas toujours refléter la réalité de la zone). Par ailleurs, grâce à notre géolocalisation de points d'intérêt via OpenStreetMap, nous sommes également en mesure de déterminer le niveau de proximité des zones avec ces points d'intérêt. Plus précisément, nous considérons ici les points d'intérêt suivants, qui sont relativement bien renseignés sur OpenStreetMap : les postes de police, les supermarchés, les écoles, les restaurants, les fastfoods et les bars.

Afin de mieux comprendre quels sont les principaux axes qui différencient ces zones entre elles, nous avons réalisé avec R une **analyse en composantes principales**, justifiée ici car les variables considérées sont toutes quantitatives. Les principales représentations graphiques de cette ACP, ainsi que l'explication de la manière dont les variables liées aux points d'intérêts ont été codées, sont reportées en annexe 2. Ici, nous nous concentrons seulement sur l'interprétation des principaux résultats.

Il apparaît que la première composante principale regroupe à elle seule plus de 40% de l'inertie totale. Elle discrimine clairement les zones suivant **la densité de population**. Nous retrouvons en effet, à droite de l'axe, les zones ayant des fortes valeurs pour la variable de densité de population, et ayant beaucoup de points d'intérêt (supermarchés, postes de police, restaurants...) à proximité. À l'inverse, les zones se retrouvant à gauche de cet axe sont celles ayant de faibles valeurs pour ces variables, et ayant un taux de maison et un taux d'habitations isolées élevés. Cela tend à confirmer la conclusion précédente, dans la mesure où il y a en moyenne davantage de maison et d'habitations isolées dans les zones moins habitées.

Nous avons également étudié le second axe, regroupant 13% de l'inertie totale. Il semble **séparer les zones riches des zones plus pauvres**, la densité intervenant aussi. Sur la projection des variables sur l'axe engendré par les deux premières composantes principales, le taux d'habitations isolées va dans le même sens que la variable indiquant le revenu moyen de la zone, ce qui n'est pas surprenant si nous considérons que les individus possédant des habitations à l'écart des autres sont plus aisés que les autres en moyenne. De même, la variable densité semble également contribuer relativement fortement à cet axe, en révélant que les zones plutôt en haut de l'axe ont en général de faibles valeurs pour la variable densité, c'est-à-dire qu'elles ont une densité de population faible.

Avec seulement deux axes, nous retrouvons donc plus de la moitié de la dispersion initiale

des données, et nous en savons davantage sur ce qui discrimine les zones *mosaic* entre elles. Davantage de détails sont présentés en annexe. Nous verrons dans les parties suivantes si la densité de population et la richesse du quartier ont un impact sur le risque de cambriolage. Les axes suivants n'ont pas été détaillés ici, à la fois à cause de leur plus faible part d'inertie et d'une interprétation de ces axes moins évidente qu'au dessus.

3.2 Importance de l'aspect temporel

3.2.1 Saisonnalité annuelle et hebdomadaire

Un découpage des taux de sinistres par jour est réalisé ci-dessous. Un découpage semblable sur les douze mois de l'année est également présenté en annexe 3. **Le taux de sinistre est plus faible le dimanche**, lorsque les gens sont en moyenne davantage chez eux (ainsi que dans une moindre mesure, le mercredi). On voit également que le vendredi et le samedi présentent un risque un peu plus élevé comme le prédisait la littérature. Mais en l'absence de données de l'heure du cambriolage, il n'est pas possible de constater un éventuel pic durant le soir ou la nuit. De plus, **les cambriolages semblent beaucoup plus fréquents en décembre**, certainement pendant les fêtes de Noël. Ils restent également nombreux en novembre. Dans tous les cas, cet aspect temporel semble avoir de l'importance.

Nous pouvons confirmer cela statistiquement en réalisant un test d'adéquation du χ^2 : si le jour n'avait aucun impact dans l'apparition d'un sinistre, les sinistres seraient distribués selon une loi multinomiale et auraient une probabilité $\frac{1}{7}$ de se situer chacun des jours. Or, lors de la mise en application d'un tel test, nous sommes amenés à rejeter l'hypothèse nulle selon laquelle nos données sont en adéquation avec cette distribution, aux seuils habituels de 5% et de 1%. Un test analogue sur les douze mois de l'année mène à la même conclusion. Cela tend à **confirmer l'importance de l'aspect temporel**.

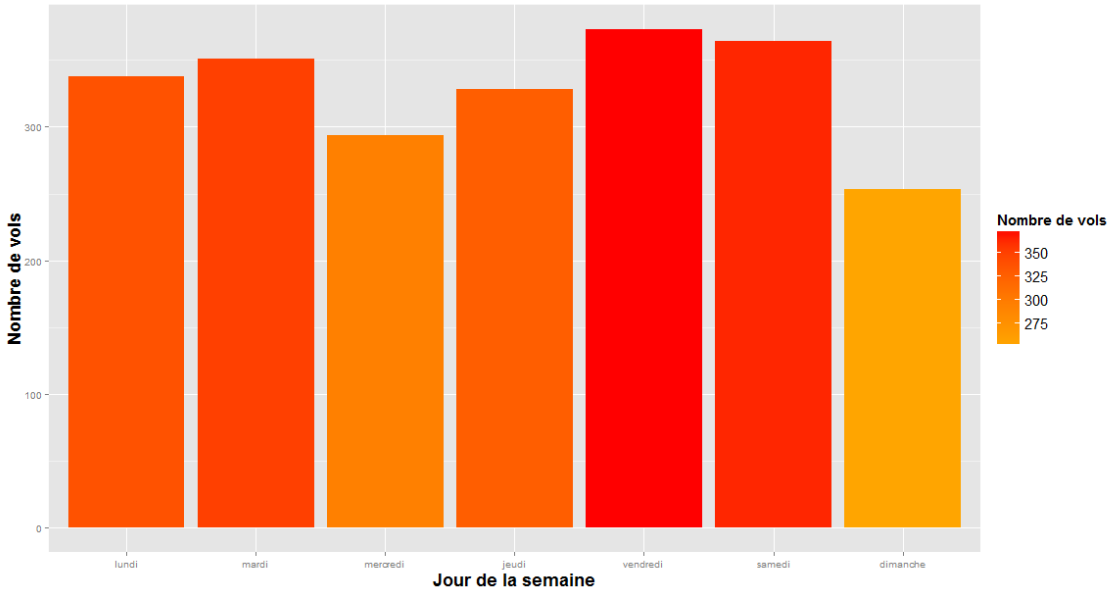

FIGURE 4 – Taux de sinistres journaliers

3.2.2 Phénomènes de répétition et de répétition proche

La littérature économétrique présentée ci-dessus montre qu'un cambriolage a un risque relativement élevé de se répéter peu de temps après dans la même zone (c'est ce que nous avions nommé les phénomènes de répétition proche), voire de se répéter au sein de la même habitation (phénomènes de répétition). Afin de confronter cette idée aux données d'AXA, nous avons décidé de créer une variable représentant le ratio du nombre de vols par assurés durant un certain mois $n-1$ dans un carré de 100 mètres autour d'une habitation. Pour observer l'effet de cette variable sur le risque de cambriolage dans l'habitation au mois n, nous l'avons intégré dans un modèle logistique avec plusieurs autres variables de contrôle.

Nous présentons en annexe 5 les résultats pour le mois de février 2010 (environ 250 000 observations), avec donc un ratio calculé sur janvier 2010. Nous observons que le coefficient de ce ratio n'est pas significatif. Il en est de même pour des modèles réalisés à d'autres périodes entre 2010 et 2013. La conclusion reste également identique si nous considérons, par exemple, un ratio calculé sur les trois derniers mois et non uniquement sur le dernier mois. Par conséquent, **il semble difficile de faire ressortir de tels phénomènes répétitifs au sein de nos données**, du moins à partir de notre approche.

Notons enfin que plusieurs variables de contrôle sont significatives : nous interpréterons les différents résultats associés à ces variables plus loin, lors de la construction d'un modèle spatial plus élaboré.

3.3 Représentation cartographique du risque de cambriolage

Les ratios de sinistralité par communes ne sont pas répartis uniformément sur tous les territoires de la Belgique. Au contraire on observe une certaine **autocorrélation spatiale positive de la sinistralité**, que nous tenterons de vérifier par la suite. Ainsi le nord de la Wallonie parait très sinistré, tout comme la région de Bruxelles et dans une part plus négligeable le centre nord de la Flandre (région autour de la ville d'Anvers). Plusieurs des intuitions précédentes ressortent sur cette carte : la densité de population a de l'importance (grandes villes) et la Wallonie, plus pauvre que la Flandre, semble davantage touchée. Nous devons noter que, pour les communes faiblement peuplées, où AXA est peu implanté, les ratios de sinistralités calculés peuvent ne pas tout à fait refléter la réalité et donc fausser légèrement la répartition en décile. Néanmoins malgré ces réserves l'autocorrélation spatiale positive apparaît nettement.

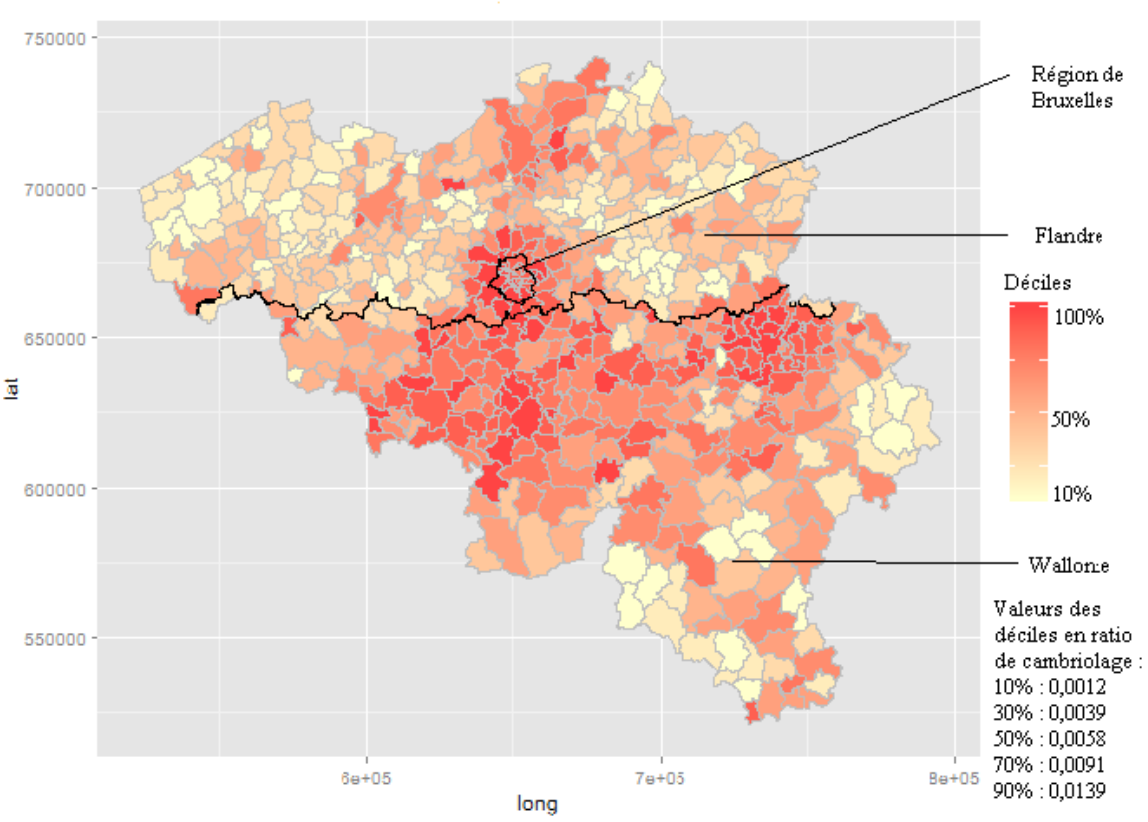

FIGURE 5 – Carte des ratios de sinistralité sur les 4 ans : regroupement par déciles

4 Modélisation spatiale

Nous débutons désormais la modélisation spatio-temporelle du risque de cambriolage. Avant de s'attaquer aux données, un aperçu des méthodes existantes et de la méthodologie que nous suivons ici est proposé.

4.1 Méthodologie de l'analyse

Notre point de départ est le modèle de régression linéaire classique où nous avons, pour chaque habitation indicée i, la relation suivante :

$$Y_i = \sum_{q=1}^{Q} X_{iq}\beta_q + \epsilon_i$$

Où Y_i correspond à la variable expliquée (le nombre de cambriolages en une année), $X_{i1}, ..., X_{iQ}$ les observations des variables explicatives pour l'individu i incluant une constante, et ϵ_i le terme d'erreur pour l'habitation i. Nous réalisons pour l'instant l'ensemble des hypothèses habituellement formulées dans le cadre d'un modèle linéaire, notamment l'indépendance entre les habitations et l'hypothèse cruciale d'espérance nulle des termes d'erreur conditionnellement aux variables explicatives. D'un point de vue matriciel, le modèle se réécrit ainsi :

$$Y = X\beta + \epsilon \tag{1}$$

Où les n observations de la variable expliquée sont regroupées dans le vecteur Y de taille $n \times 1$. X est une matrice de taille $n \times Q$, β un vecteur de taille $Q \times 1$ et ϵ un vecteur de taille $n \times 1$; nous avons $E(\epsilon|X) = 0$ et donc $E(\epsilon) = 0$. X regroupe alors l'ensemble des variables propres aux habitations assurées : on y retrouvera par exemple l'âge de l'assuré ou encore la richesse de la zone dans laquelle l'habitation est située [3].

Ce modèle est un point de départ, mais s'en limiter conduirait à totalement occulter les éléments spatiaux (a priori importants d'après la littérature) ayant un impact sur le risque de cambriolage. En cas de lien spatial entre les habitations, pourtant, celles-ci ne seraient plus indépendantes, le modèle ci-dessus souffrirait d'une mauvaise spécification et les résultats seraient biaisés.

Dès lors, nous sommes amenés à considérer d'autres modèles plus élaborés, prenant en compte **la dépendance spatiale** entre les assurés.

4.1.1 Modèle *spatial autoregressive* (SAR)

Les modèles dits *spatial lag models* sont une extension des modèles linéaires présentés ci-dessus, qui vont permettre à une habitation i de dépendre des observations de ses voisins. Le *spatial lag model* le plus classique est appelé **spatial autoregressive (SAR)** et prend la forme

3. Tout ceci est détaillé lors de la partie consacrée à l'application des modèles aux données AXA.

suivante :

$$Y_i = \rho \sum_{j=1}^{n} W_{ij} Y_j + \sum_{q=1}^{Q} X_{iq} \beta_q + \epsilon_i$$

Où les termes d'erreurs sont i.i.d. Ici, W_{ij} désigne l'élément (i, j) d'une matrice W dite **matrice de poids**. D'un point de vue mathématique, il s'agit d'une matrice contenant des valeurs positives ou nulles, la somme sur chaque ligne valant 1 : $\sum_j W_{ij} = 1$, $\forall i$. D'un point de vue pratique, elle nous permet de déterminer de quelle manière et dans quelle mesure les observations s'impacteront entre elles ou non. Nous attribuerons un poids $W_{ij} = 0$ à j s'il n'y a pas de « lien » entre les habitations i et j, et $W_{ij} > 0$ s'il y a un lien. Nous verrons lors de l'application aux données comment construire une matrice W adaptée à partir de l'algorithme de recherche des K proches voisins, l'objectif étant d'attribuer des poids plus importants aux voisins les plus proches. Enfin, nous avons $W_{ii} = 0$, $\forall i$.

Le scalaire ρ, quant à lui, est un paramètre à estimer qui nous permettra de **déterminer la force et la nature de la relation spatiale autorégressive** entre Y_i et les Y_j des voisins étudiés. Un coefficient positif correspondrait à une situation d'autocorrélation spatiale positive, un coefficient négatif correspondrait à une situation d'autocorrélation spatiale négative. Dans le cas $\rho = 0$, nous retrouvons le modèle (1). En notation matricielle, le modèle SAR devient :

$$Y = \rho W Y + X\beta + \epsilon \tag{2}$$

Il se réécrit ainsi :

$$Y = (I - \rho W)^{-1} (X\beta + \epsilon)$$

Et nous avons :

$$E(Y) = (I - \rho W)^{-1} X\beta$$

Car les termes d'erreurs sont d'espérance nulle. La matrice inverse est appelée **multiplicateur spatial** : elle montre que l'espérance des Y_i va dépendre d'une combinaison linéaire des X parmi les voisins, proportionnelle au paramètre de dépendance ρ. Remarquons que l'inversibilité de $I - \rho W$ est nécessaire et sera à vérifier, pour que tout ceci ait un sens (il existe quelques situations où ce n'est pas le cas).

4.1.2 Indice de Moran et test de dépendance spatiale

Devons-nous nous situer dans le cadre du modèle (1) ou bien dans celui du modèle (2) ? Nous évoquons des méthodes de visualisation de l'autocorrélation spatiale, puis nous proposons un test plus rigoureux pour répondre à cette question.

L'une des approches possibles est de considérer **l'indice de Moran** :

$$I = \frac{\sum_i \sum_j (Y_i - \bar{Y})(Y_j - \bar{Y})}{\sum_i (Y_i - \bar{Y})^2} \in [-1; 1] \tag{3}$$

Les valeurs positives et fortes de I indiquent une **autocorrélation spatiale** positive. On définit l'autocorrélation spatiale comme la corrélation d'une variable avec elle même provenant de la disposition géographique des données. Les valeurs négatives et fortes de I indiquent une

autocorrélation spatiale négative. Les valeurs proches de 0 indiquent une absence d'autocorrélation.

Nous pourrions aussi représenter un nuage de points de WY contre Y, dit **diagramme de Moran**, où Y est alors **centrée** et où les lignes de W sont toujours normalisées. Ceci implique que la moyenne empirique de WY est nulle. Superposons au nuage la droite de régression correspondante, qui passe donc par l'origine. La pente de celle-ci correspond à l'indice de Moran. Ceci permet d'apprécier l'intensité de l'autocorrélation, et également parfois de faire ressortir des points aberrants.

Définissons, enfin, I_{res} **l'indice de Moran appliqué aux résidus** : il s'agit de la formule (3) où les $Y_i - \hat{Y}$ sont remplacés par r_i les résidus obtenus à partir du modèle (2). Il est possible de réaliser des tests statistiques à partir de I_{res}, de la forme $H_0 : I = a$ contre $H_1 : I \neq a$ en retranchant I_{res} par a et en divisant par l'écart-type, et en comparant cela à une loi $N(0, 1)$. Ainsi, nous mettrons en avant la présence d'autocorrélation spatiale si nous sommes amenés à rejeter l'hypothèse nulle dans le test suivant : $H_0 : I = 0$ contre $H_1 : I \neq 0$. Ceci est réalisé automatiquement via la fonction R *moran.I* issue du package *ape*.

4.1.3 Modèle de Durbin spatial

Le **modèle de Durbin spatial** est le modèle SAR dans lequel nous prendrons en compte non seulement les Y des voisins les plus proches, mais également certaines des variables explicatives associées à ces voisins. Il se formalise ainsi :

$$Y = \rho WY + X_1\beta + WX_2\gamma + \epsilon \tag{4}$$

Où X se décompose désormais en X_1 et X_2, X_2 regroupant les variables explicatives pour lesquelles nous souhaitons récupérer les valeurs des plus proches voisins.

Cette amélioration du modèle SAR est très utile dans notre cas. Prenons l'exemple du revenu moyen de la zone dans laquelle est située une habitation assurée. Nous pouvons penser que celui-ci jouera sur le risque de cambriolage. Mais nous pouvons également penser que les revenus moyens des zones **voisines** auront également un impact ! Nous retrouvons ici l'idée des frontières zone riche/zone pauvre mise en avant dans Rengert et al. (1999) et Rossmo (2000). Le modèle de Durbin spatial permettra de capter ce phénomène. Ce n'était pas le cas du modèle SAR classique.

4.2 Application aux données AXA

À l'aide des packages *ape* et *spdep* de R, nous appliquons désormais les méthodes présentées ci-dessus aux données d'AXA. Nous avons montré l'importance, au sein de notre problématique, d'intégrer une matrice de voisinage pour certaines variables explicatives : nous avons donc choisi de travailler à partir d'un **modèle de Durbin spatial**. Pour des raisons computationnelles, mais également pour obtenir une réelle variable de comptage, nous avons décidé d'agréger les individus sur les quatre années : la variable expliquée devient donc nombre de sinistres entre 2010 et 2013. Nous avons pondéré les individus qui n'étaient pas assurés sur l'ensemble de cette durée.

4.2.1 Variables explicatives et matrice de voisinage

Les variables explicatives retenues pour l'analyse sont, pour chaque habitation, à la fois des variables correspondant aux spécificités de sa zone *mosaic* et des variables propres à l'habitation ou à l'assuré :

— le revenu moyen de la zone ;
— la densité de la zone ;
— l'âge de l'assuré ;
— le fait qu'il soit propriétaire ou locataire ;
— le fait que l'habitation soit isolée ou contiguë ;
— le fait que l'habitation soit une maison ou un appartement ;
— la présence ou non d'un poste de police dans les 2kms ;

De plus nous étudierons également l'effet du revenu moyen au sein des zones voisines. Enfin, nous aurions apprécié pouvoir ajouter davantage de points d'intérêt OpenStreetMap mais ceux-ci étaient pour la plupart soit mal représentés (par exemple les caméras de surveillance) soit peu pertinents (par exemple les pharmacies) donc nous les avons omis dans la régression.

la **matrice de voisinage** qui nous est utile à la fois pour l'indice de Moran et pour la régression spatiale est une matrice $n \times n$ avec n le nombre d'observations. La matrice à n lignes et n colonnes a comme coefficient $m_{ij} \neq 0$ si i et j sont voisins, $m_{ij} = 0$ dans le cas contraire. Nous avons décidé de choisir une distance euclidienne pour déterminer si les individus sont voisins ou non de 10 km ce qui est une distance assez importante pour ne pas obtenir d'observations sans voisin. Les coefficients sont pondérés comme étant égaux à l'**inverse de la distance**. Cette méthode permet de prendre en compte la valeur de la distance en appliquant des coefficients plus importants lorsque la distance entre deux observations est plus petite.

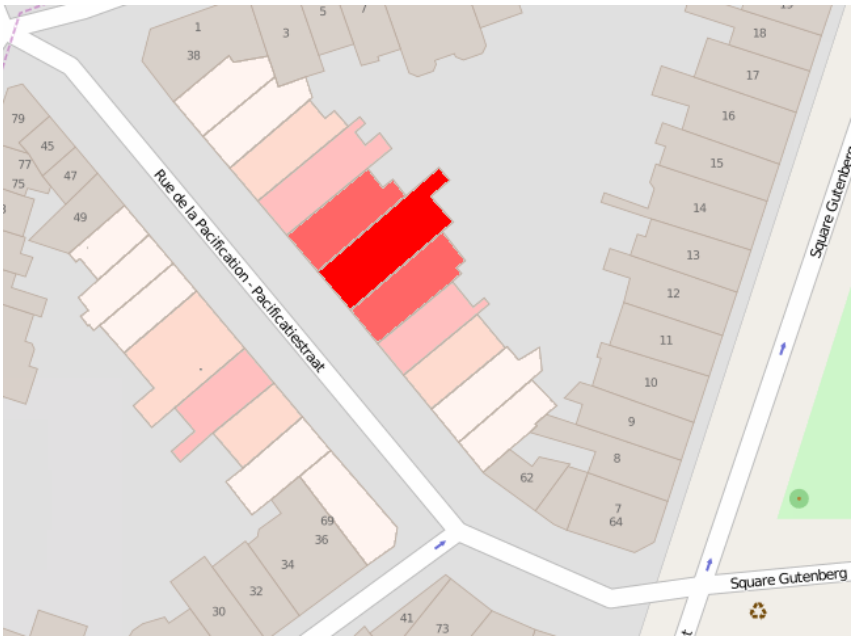

FIGURE 6 – Pondération du voisinage : visualisation

4.2.2 Présence d'autocorrélation spatiale positive

	Durbin Spatial
(Constante)	0.0011
	(1.2789)
Densité	0.0034***
	(0.0006)
Revenu	0.0005***
	(0.0001)
Age	−0.0061
	(0.0057)
Isolée	0.0043***
	(0.0013)
Propriétaire	0.0078
	(0.0053)
Maison	0.0091*
	(0.0037)
Police	−0.0062
	(0.0082)
ρ	0.0315
P-value (two-sided)	0.0317
γ_{Rev}	−0.0402
P-value (two-sided)	< 0.001
Test LM d'autocorrélation des résidus	0.222
P-value	0.700

$^{***}P(>|z|) < 0.001$, $^{**}P(>|z|) < 0.01$, $^{*}P(>|z|) < 0.05$

TABLE 4 – Modèle de Durbin spatial

Nous observons que la densité, le revenu, l'isolation et le fait que le logement soit une maison et non un appartement sont des variables de contrôle qui sont significatives à 5%. Avec des coefficients positifs, **une augmentation de la densité et du revenu moyen augmentent le risque de cambriolage** toutes choses égales par ailleurs. **Le fait d'être isolé et d'habiter dans une maison augmente aussi le risque de cambriolage**. ρ qui est **le coefficient de dépendance spatiale sur les cambriolages des voisins s'avère significatif à** 5% **et positif**, donc on a plus de chance d'être cambriolé lorsque son voisin est lui même cambriolé. Ensuite le coefficient γ_{Rev} est négatif et aussi significatif à 5%, ainsi **lorsque nos voisins ont un revenu faible nous avons plus de chance d'être victime d'un cambriolage**. Ce résultat est conforme avec la littérature économétrique sur le sujet.

La **dépendance spatiale du risque de cambriolage** du Durbin Spatial est confirmée par le test de Moran réalisé ci-dessous : ici l'hypothèse nulle d'indépendance spatiale est rejetée à 5%.

I de Moran	P-Value du test de dépendance spatiale à partir de I_{res} calculé sur les résidus $H_0 : I = 0$ vs $H_1 : I \neq 0$
0.164	0.045

TABLE 5 – Indice de Moran et test de dépendance spatiale

La modélisation spatiale est donc importante pour comprendre les risques de cambriolage, du fait de la dépendance spatiale des cambriolages. De plus une autre information importante de cette étude est l'augmentation du risque du cambriolage lorsque les voisins ont en moyenne des revenus moins élevés. Pour qu'un modèle d'apprentissage par *gradient boosting* tel que celui développé plus loin ait une chance d'être plus efficace dans la modélisation du risque de cambriolage, il ne faudra donc pas omettre cette information spatiale.

5 Apprentissage par gradient boosting

Au cours de cette dernière partie, nous tentons d'élargir le cadre d'analyse, en confrontant les résultats obtenus précédemment avec ceux de méthodes alternatives, issues du *machine learning*. Dans le modèle précédent, nous "forcions" l'aspect spatial et la correction des autocorrélations sur un pas fixe. Ici, nous allons considérer des méthodes traitant toutes les variables à égalité, et nous allons observer lesquelles sont réellement pertinentes. En particulier, nous verrons dans quelle mesure l'aspect spatial ressort.

Ces méthodes sont basées sur le concept d'arbre de régression. Nous rappelons les avantages et inconvénients d'un arbre par rapport à un modèle linéaire ou GLM plus classique, puis nous expliquons que le pouvoir prédictif d'un unique arbre est généralement moindre. Cela nous amène à considérer une méthode basée sur la considération simultanée de plusieurs arbres, censée améliorer nos prédictions : l'approche par **gradient boosting**. Nous comparerons les résultats obtenus sur les données AXA via les différents modèles.

5.1 Méthodologie de l'analyse

5.1.1 Arbres de régression

Les arbres de régression sont une alternative aux modèles linéaires et aux GLM. Un arbre divise l'espace des variables explicatives à plusieurs niveaux, obtenant plusieurs régions. Il s'agit alors de suivre les branches de cet arbre selon les valeurs d'un individu pour ces variables, pour déterminer l'estimation de la variable expliquée. Les individus "tombant" dans la même région reçoivent la même estimation qui correspond à la moyenne empirique de la variable expliquée dans cette région.

Nous ne rappellerons pas ici les détails de la construction d'un arbre de régression, le lecteur intéressé étant invité à se reporter par exemple à James et al. (2013). Précisons cependant qu'il est en général impossible de considérer l'ensemble des découpages possibles pour déterminer le meilleur selon une certaine fonction de perte, et que la solution s'obtient via un algorithme dit *recursive binary splitting*. Il est possible de considérer différentes fonctions de pertes, comme la classique fonction de perte quadratique mais également l'opposée de la vraisemblance d'une Poisson, suggérée par De Laet (2014), et qui est adaptée à un modèle de comptage tel que le nôtre (nombre de cambriolages). Notons également que des critères d'arrêts, sur la taille des régions et sur la profondeur de l'arbre (nous reviendrons sur ce point dans les parties suivantes), sont à définir pour éviter des phénomènes de **surapprentissage**. Ce terme désigne la situation où, par sa trop grande capacité à coller aux données à notre disposition, une structure aura du mal à obtenir des résultats satisfaisants sur de nouvelles données.

Les arbres sont généralement faciles à interpréter et à expliquer. Nous pouvons aussi facilement les visualiser, ce qui n'est en général pas le cas dans un GLM (Generalized Linear Model). Ils détectent aussi automatiquement les interactions entre les variables alors que, dans un modèle linéaire ou un GLM, nous devons les déterminer et les ajouter nous-mêmes au modèle. Cependant, le pouvoir prédictif d'un arbre est en général plus faible qu'un GLM. Plusieurs méthodes existent pour remédier à ce problème, dont le *boostrap aggregating* et le *random forest*, mais également l'approche par **gradient boosting** que nous allons utiliser ici.

5.1.2 Principe du boosting

Le principe du *boosting* a été développé il y a une trentaine d'années dans le monde du *machine learning*. Cette méthode est synthétisée ainsi dans Culp et al. (2006) : *boosting is an iterative algorithm that combines simple rules with mediocre performance in terms of misclassication error to produce a highly accurate rule*. De manière générale, au sein d'un modèle de régression nous voulions prédire une réponse y à partir de p variables explicatives $x = (x_1, x_2, ..., x_p)$. Il s'agit alors de déterminer $E(y|x) = f(x)$ de façon à minimiser une certaine fonction de perte notée $L(y, f(x))$. Rappelons que nous avions, pour un modèle généralisé : $f(x) = \sum_{j=1}^{p} \beta_j x_j$. Dans le cadre du *boosting*, nous aurons désormais : $f(x) = \sum_{t=1}^{T} f_t(x)$.

Les fonctions $f_t(x)$ ne dépendent pas d'une seule variable. Nous écrivons $f_t(x) = \beta_t h(x; a_t)$ avec h une fonction caractérisée par les paramètres $a_t = (a_1, a_2, ...)$. Par la suite, une telle fonction correspondra à un arbre de régression, mais nous aurions tout aussi bien pu considérer par exemple un réseau de neurones. En utilisant l'écriture précédente de $f(x)$, nous cherchons donc à minimiser par rapport à β_t et a_t de 1 à T :

$$\sum_{i=1}^{n} L(y_i, \sum_{t=1}^{T} \beta_t h(x_i; a_t))$$

La solution au problème précédent est obtenu par des méthodes dites de *forward stagewise additive modeling*. Elles consistent à :

- initialiser $f_0(x) = 0$;
- pour t de 1 à T
  - estimer β_t et a_t en minimisant $\sum_{i=1}^{n} L(y_i, f_{t-1}(x_i) + \beta h(x_i; a))$
  - poser $f_t(x) = f_{t-1}(x_i) + \beta_t h(x_i, a_t)$;
- renvoyer finalement $\hat{f}(x) = f_T(x)$.

La seconde étape n'est pas toujours facile à obtenir. Une procédure particulière est appliquée dans l'algorithme dit de *gradient boosting*, que nous détaillons ci-dessous dans le cadre d'arbres de régression.

5.1.3 Gradient tree boosting

L'algorithme du *gradient boosting*, inventé par Jerome H. Friedman en 1999, résout la seconde étape de la procédure précédente par descente de gradient. Concrètement, en appliquant celui-ci au cas particulier des arbres de régression, l'algorithme est le suivant :

- initialisation : $f_0(x) = argmin_\beta \sum_{i=1}^{n} L(y_i, \beta)$
- pour t de 1 à T
  - calculer l'opposé du gradient $r_i = -\left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f(x)=f_{t-1}(x)}$, pour $i = 1, ..., M$
  - ajuster un arbre de régression sur r_i, en obtenant ainsi les régions finales R_{jt}, $j = 1, 2, ..., J_t$
  - pour j de 1 à J_t
    - calculer $c_{jt} = argmin_c \sum_{x_i \in R_{jt}} L(y_i, f_{t-1}(x_i) + c)$
  - poser $f_t(x) = f_{t-1}(x) + \sum_{j=1}^{J_t} c_{jt} I_{(x \in R_{jt})}$
- renvoyer finalement $\hat{f}(x) = f_T(x)$

Sachant le modèle actuel, nous essayons de nous améliorer en réalisant un arbre, non pas directement sur la variable à expliquer y, mais sur les r_i. On dit que cette approche par *boosting*

permet un apprentissage en douceur : cela réduit le risque de surapprentissage. Afin de réduire encore davantage ce risque, nous aurons recours à un paramètre de rétrécissement ou **shrinkage** λ, dont le rôle sera de réduire le rythme d'apprentissage. Concrètement, nous remplacerons l'avant-dernière étape du *gradient boosting* par :

$$f_t(x) = f_{t-1}(x) + \lambda \sum_{j=1}^{J_t} c_{jt} I_{(x \in R_{jt})}$$

Il est possible de montrer qu'une petite valeur de λ, typiquement $\lambda \leq 0.1$, conduit à des améliorations significatives par rapport à la situation initiale où $\lambda = 1$.Trois paramètres seront donc à déterminer : λ, mais aussi le nombre d'arbres T et le nombre de découpages réalisés au niveau de chaque arbre J_t.

5.2 Application aux données AXA

5.2.1 Création d'un indice de richesse relative

Notre approche est désormais non-paramétrique et, contrairement au modèle spatial précédent, il n'est plus possible de définir de la même manière la matrice de poids afin de lier les plus proches voisins entre eux. Néanmoins, nous sommes toujours intéressés par les déterminants spatiaux du risque de cambriolage. En particulier, il serait regrettable de ne plus pouvoir capter les phénomènes de rapprochements zone riche-zone pauvre. Rappelons que la littérature économétrique montre en effet que ce n'est pas seulement le fait d'habiter dans un quartier riche qui augmente le risque, mais aussi voire surtout le fait d'habiter dans un quartier riche à proximité d'un quartier plus défavorisé.

Pour cette raison, nous avons été amenés à construire sur R un **gradient de richesse** indiquant dans quelle mesure l'habitation est dans un quartier mieux ou moins bien loti que ceux à proximité. Notons Rev_i le revenu moyen dans la zone *mosaic* dans laquelle l'habitation assurée i se situe. Le gradient de richesse est construit ainsi, pour chaque habitation :

$$Grad_i = 100 \times \frac{Rev_i}{\frac{1}{n} \sum_{j \in V} Rev_j}$$

Où V désigne l'ensemble des habitations situées dans un cercle de rayon R dont le centre est l'habitation i, et n le nombre d'habitations assurées dans ce cercle. Un indice proche de 100 correspond à une situation où il n'y a que très peu d'hétérogénéité de richesse autour de l'habitation. Un indice supérieur à 100 correspond à une situation où l'individu est dans un quartier plus riche que les voisins, et inversement pour un indice inférieur à 100. Plus l'indice s'éloigne de 100, plus la différence est nette (en pratique, l'individu est souvent en bord de zone).

Ce rayon R doit être assez grand pour regrouper un nombre suffisant de voisins. Cependant, il ne doit pas non plus être trop grand, au risque de perdre en pouvoir explicatif. En pratique, nous verrons par la suite que nous obtenons des résultats parlants pour $R = 5$kms, au sens des dépendances partielles du *gradient boosting*. Nous obtenons une variable centrée en 100, balayant avec une allure de courbe en cloche un intervalle de 70 à 130.

Cet indice peut sembler simple, mais il a l'avantage de maintenir la même distance maximum

indépendamment des densités des zones. De plus, deux individus habitant aux deux extrémités d'une même zone ne se verrons pas forcement attribuer le même indice : c'est une bonne chose car, même s'ils sont dans la même zone, l'un peut être situé à proximité d'une zone riche et l'autre à proximité d'une zone pauvre.

5.2.2 Variables explicatives et influences relatives

Les variables explicatives maintenues dans l'analyse sont, pour chaque habitation, à la fois des variables correspondant aux spécificités spatiales et des variables propres à l'habitation ou à l'assuré :

— le revenu moyen de la zone ;
— la valeur du gradient de richesse, avec $R = 5$kms ;
— la densité de la zone ;
— l'âge de l'assuré ;
— le fait qu'il soit propriétaire ou locataire ;
— le fait que l'habitation soit isolée ou contiguë ;
— le fait que l'habitation soit une maison ou un appartement ;
— la présence ou non d'un poste de police dans les 2kms ;
— la présence ou non d'une école dans les 2kms.

Le *gradient boosting* a été réalisé sur R via l'utilisation du package *gbm*, créé par Greg Ridgeway. Concernant le choix des paramètres, nous avons tout d'abord lancé l'algorithme sur 70% de la base de données, constituant notre **base d'entraînement**. Puis, à partir des résultats obtenus, nous avons tenté de prédire le risque de cambriolage sur la **base de test** constituée de 20% des données. Les 10% restants serviront plus tard à comparer finalement les différents modèles en terme de pouvoir prédictif. Les modèles ont été jugés à partir de la *validation deviance* et de la valeur de la fonction de perte sur la base de test, mais également en fonction de l'allure des représentations graphiques des effets marginaux (étaient-ils cohérents, et monotones ? Ou bien présentaient-ils beaucoup de bruit, signe de surapprentissage ?).

La profondeur de l'arbre est très importante : nous avons commencé par une petite profondeur de 4 en essayant d'augmenter progressivement. Imposer une taille de feuille minimale (définie par n.minobsinode sur R) empêche de faire des groupes trop petits, ce qui est intéressant car une feuille avec trop peu d'observations sera sans doute trop spécifique et nous surapprendrions. Le nombre d'arbres optimal va être choisi selon la meilleure itération. Quant au *skrinkage*, nous avons testé les principales valeurs habituellement retenues : 0.1, 0.05 et 0.01.

Sur ces critères, nous avons finalement retenu un taux de *skrinkage* de 0.05, et nous allons travailler avec une profondeur d'arbre de 4 et un nombre d'arbres égal à 100. Dès lors, nous avons relancé l'algorithme sur ces 90% des données. Le tableau suivant présente les influences relatives de chacune des variables dans la construction du gradient boosting.

Les variables qui ressortent en premières sont celles liées aux caractéristiques spatiales. La densité de population est la variable ayant le plus d'importance. Le gradient de richesse arrive en second... devant la richesse de la zone ! Son influence relative est donc plus grande, ce qui est conforme avec la littérature économétrique sur le sujet. L'isolement ou non de l'habitation a, sans grande surprise, également un rôle à jouer. Les autres variables se succèdent ensuite, avec des influences relatives plus faibles.

Variable	Influence Relative (%)
Densité de population	35.75
Gradient de richesse	19.02
Richesse moyenne	16.88
Isolement	14.94
Age de l'assuré	7.47
Appartement ou maison	4.22
Propriétaire ou locataire	1.46
Police à proximité	0.18
École à proximité	0.09

TABLE 6 – Influences relatives des variables du boosting

5.2.3 Analyse des dépendances partielles

Représentons désormais graphiquement les **dépendances partielles** de chacune de ces variables. Ces graphiques illustrent l'effet marginal d'une variable, une fois que les autres sont contrôlées.

Nous avions déjà montré que le risque de cambriolage était plus élevé dans les **zones à forte densité de population**, ce phénomène ressort à nouveau grâce au *boosting*. Ici, le risque atteint cependant assez rapidement son niveau maximum, ce qui pourrait davantage faire penser à un découpage entre zones rurales et urbaines. Quant au **gradient de richesse**, sans surprise, ce sont très nettement **les habitations dans des zones riches proches de zones pauvres** qui sont les plus exposées au risque.

Concernant l'indicateur de richesse lui-même, une forme en U assez intéressante apparaît, faisant ressortir **un risque élevé à la fois pour les zones les plus pauvres et les zones les plus riches**. Le minimum est atteint pour des zones à niveau moyen de richesse. Le *gradient boosting* nous confirme également que **les habitations jugées isolées** sont davantage exposées au risque, **de même que les maisons**.

Concernant l'âge, le risque semble décroître légèrement [4] à partir de la soixantaine, un résultat plus surprenant d'autant plus que nous avons bien fait attention à regrouper les valeurs trop extrêmes pour éviter d'obtenir des individus trop influents. Les individus âgés, se sachant plus faibles, pourraient-ils avoir tendance à mieux protéger leur habitation ? Nous pouvons aussi penser que les cambrioleurs préfèrent opérer lorsqu'il n'y a personne dans l'habitation, et que les personnes âgées sont plus souvent chez elles.

Même si l'influence relative est très faible, notons aussi que le résultat de la variable Police est assez cohérent : **les habitations proches d'un poste de police ont un risque de cambriolage légèrement plus faible**. Le fait que les variables issues de la recherche de points d'intérêts via OpenStreetMap soient peu explicatives peut se comprendre de plusieurs manières. Peut-être que ces variables ne sont effectivement que peu pertinentes. Sans doute aussi qu'elle sont très incomplètes : OpenStreetMap étant collaboratif, nous ne disposions pas des coordonnées de tous les postes de police et écoles de Belgique. C'est à cause de ce manque de données que d'autres points d'intérêts (par exemple les bars ou les caméras de surveillance publiques) n'ont finalement pas été inclus dans le *boosting*. Bien sûr, nous avons testé différents seuils que 2kms pour définir

4. Attention aux ordres de grandeur des axes qui ne sont pas gradués de manière identique !

la proximité d'un poste de police ou d'une école, en vérifiant que nous ayons assez d'habitations concernées à chaque fois. Cependant aucun seuil n'a donné de résultats bien meilleurs.

Enfin, les résultats ci-dessous sont les meilleurs que nous ayons pu obtenir mais nous pouvons noter qu'il reste malgré tout quelques légers signes de surapprentissage. Concernant les densités très élevées par exemple, il semble qu'il y ait un peu de surapprentissage sur Bruxelles. D'autres « pics » sont aussi présents dans certaines des dépendances partielles ci-dessous.

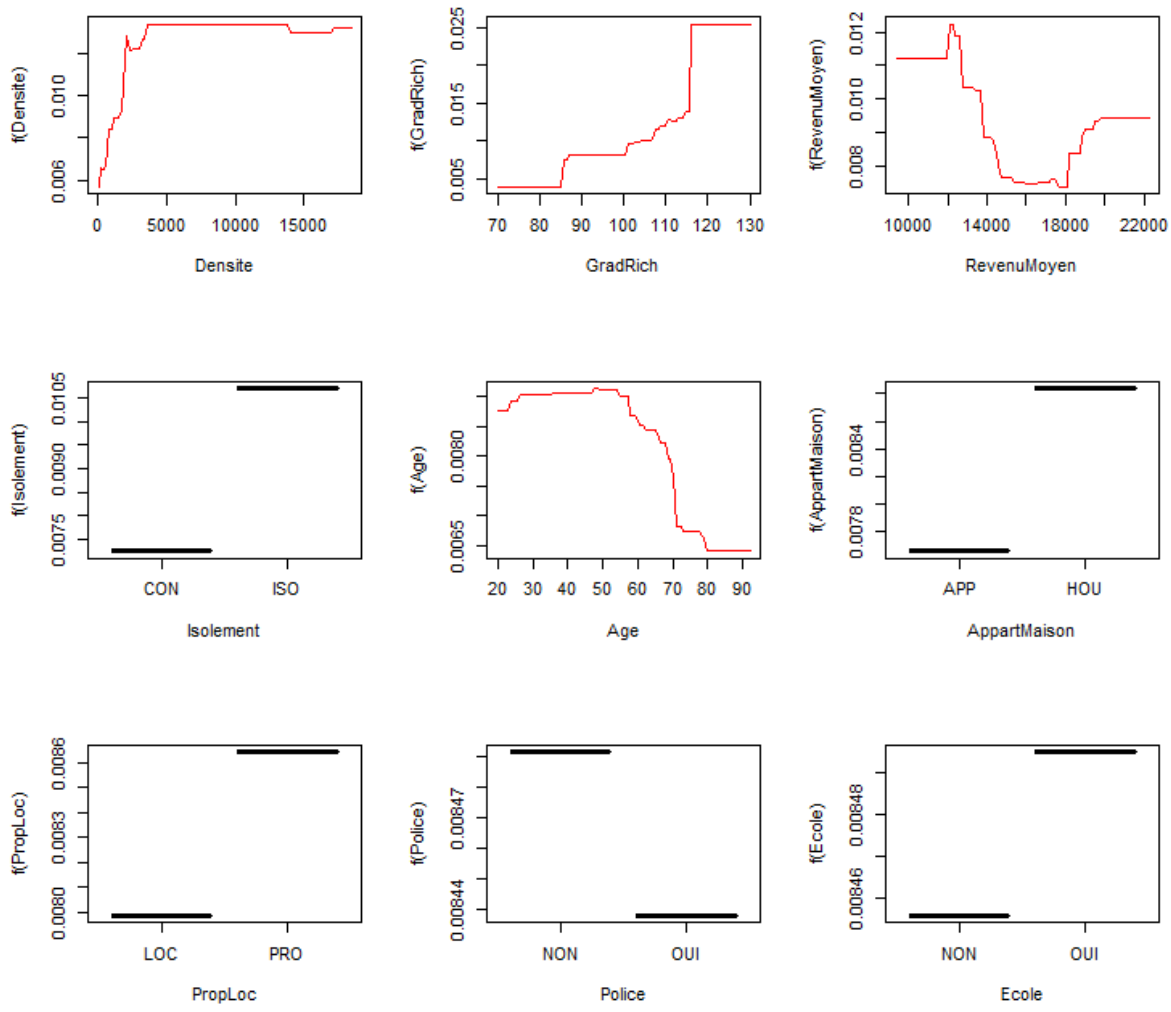

FIGURE 7 – Dépendances partielles des variables du boosting

5.3 Comparaison des méthodes en termes de prédiction

Dans cette partie, nous comparons les résultats du *gradient boosting* et ceux du modèle spatial précédent en termes de pouvoir de prédiction. Différentes approches sont étudiées, en reprenant la démarche de De Laet (2014). Pour cela, nous travaillons sur les 10% restants de la base de données AXA, à partir desquels les modèles n'ont **pas** été construits.

5.3.1 Nombre de sinistrés et non sinistrés modélisés et observés

De Laet (2014) propose une première méthode pour comparer nos modèles. Il s'agit de **regarder le nombre de zéros modélisés par chacun des modèles et le nombre de zéros effectivement observés**. Un zéro correspond à une situation où il n'y a eu aucun sinistre durant la période d'exposition. Dans le mémoire de De Laet, axé sur une problématique d'assurance automobile, il s'agit d'un assuré n'ayant eu aucun accident sur la période. Pour nous, il s'agit d'une habitation n'ayant pas connu de cambriolage. Le nombre de zéros réels est directement observé dans la base. Quant au nombre de zéros modélisés, il est défini ainsi :

$$E(nombre\ de\ 0) = \sum_{i=1}^{n} P(vol_i = 0) = \sum_{i=1}^{n} e^{-\hat{vol}_i}$$

La deuxième égalité venant du fait que nous nous plaçons dans une situation où la variable expliquée est une variable de comptage de type Poisson. Alternativement, il est aussi possible de comparer le nombre de sinistrés prédits par le modèle au nombre de sinistrés réellement observés, à partir de :

$$E(nb\ de\ sinistres) = n - \sum_{i=1}^{n} e^{-\hat{vol}_i}$$

Les tableaux suivants proposent la mise en application de cette approche d'abord sur la base à partir de laquelle ont été construits les modèles, puis surtout sur les 10% restants de la base de données qui constituent notre base de test et à partir desquels la qualité de prédiction sera jugée.

Modèle	Nombre de zéros modélisés / Nombre de zéros observés	Nombre de sinistrés modélisés / Nombre de sinistrés observés
Modèle spatial	0.9997	1.0299
Gradient boosting	0.9999	1.0037

TABLE 7 – Comparaison sur la base d'entrainement (90% des données)

Modèle	Nombre de zéros modélisés / Nombre de zéros observés	Nombre de sinistrés modélisés / Nombre de sinistrés observés
Modèle spatial	0.9974	1.3230
Gradient boosting	0.9996	1.0468

TABLE 8 – Comparaison sur la base de test (10% des données)

Les résultats sur la base d'entraînement sont très bons : les rapports sont en effet proches de 1, avec un léger avantage au *gradient boosting*. Mais ce résultat était attendu dans la mesure où nos modèles ont été construits à partir de cette base. Quant aux résultats sur la base de test, ils mettent à nouveau en avant de **meilleurs rapports pour le gradient boosting**, assez nettement. Ce modèle semble donc meilleur d'un point de vue prédictif, selon ce critère.

5.3.2 Courbes ROC

Une deuxième approche possible consiste à transformer notre variable de comptage du nombre de sinistres en une variable binaire séparant les sinistrés et les non-sinistrés, puis à tracer les courbes ROC correspondant aux deux modèles. Une courbe ROC, de l'anglais *Receiver Operating Characteristic*, mesure la qualité d'un classificateur binaire. Cette courbe met en avant le taux de vrais positifs (sinistrés qui sont effectivement détectés) en fonction du taux de faux positifs (non-sinistrés qui sont incorrectement détectés).

Plus l'aire sous la courbe ROC est grande, plus la courbe s'écarte de la bissectrice qui correspond à un « classificateur aléatoire », et meilleure est la prédiction. Ci-dessous, **la courbe ROC du gradient boosting est plus élevée que celle du modèle spatial**, ce qui confirme le meilleur pouvoir prédictif du *gradient boosting*.

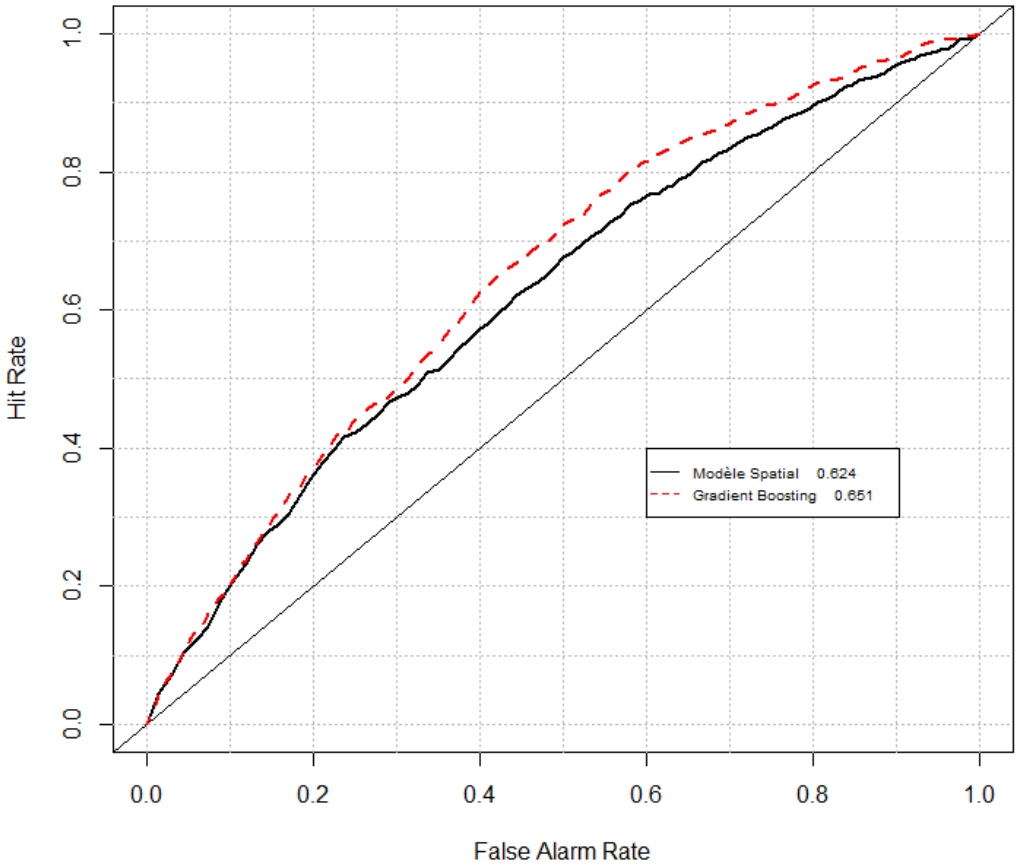

FIGURE 8 – Courbes ROC

5.3.3 Analyse des fonctions de perte

Un autre moyen de comparer les modèles consiste à observer directement la précision de leurs prévisions, sur les 10% des données constituant la base de test. À partir des modèles, nous récupérons les estimations $\hat{vol}_i = f(x_i)$ pour chaque individu i dans la base de test, x_i désignant le vecteur des variables explicatives. La méthode consiste à calculer pour chacun des modèles la valeur d'une certaine fonction de perte $L(vol, f(x))$. Dans notre cadre d'une variable de comptage, la fonction de perte que nous utilisons s'obtient à partir de l'opposée de la vraisemblance de la loi de Poisson. De Laet la présente sous cette forme :

$$L(vol, f(x)) \propto \sum_{i=1}^{n} (e^{f(x_i)} - vol_i f(x_i))$$

Ici, un terme constant de la vraisemblance est négligé. Par ailleurs, nous pourrions aussi considérer d'autres fonctions de perte, comme par exemple une fonction de perte quadratique :

$$L(vol, f(x)) = \sum_{i=1}^{n} (f(x_i) - vol_i)^2$$

Voici les résultats obtenus **sur la base de test**, pour chacun des modèles :

Modèle	Perte Poisson	Perte Quadratique
Modèle spatial	103482.6	840.26
Gradient boosting	103254.4	839.43

TABLE 9 – Comparaison des modèles à partir des fonctions de perte

Les fonctions de perte Poisson et quadratique sont toutes les deux plus faibles pour le gradient boosting. Ce troisième critère conduit donc à la même conclusion que les deux précédents : le gradient boosting a un meilleur pouvoir prédictif que le modèle spatial.

Notre démarche de construction d'un modèle alternatif non-paramétrique issu du *machine learning* n'a donc pas été vaine. Les résultats sont en effet meilleurs que ceux obtenus via une approche paramétrique spatiale plus classique. Dans le même temps, ce modèle nous permet également d'interpréter aisément le rôle et l'importance de chaque variable dans la détermination du risque de cambriolage, ce qui est aussi intéressant. Nous proposons donc à AXA d'utiliser plutôt un tel modèle qu'un modèle spatial, afin d'évaluer le risque des futurs assurés belges.

6 Conclusion

À partir d'une littérature économétrique importante et de statistiques descriptives pour nous guider, nous avons construit un modèle spatial de type Durbin ainsi qu'un *gradient boosting* pour évaluer le risque de cambriolage en Belgique et mettre en avant les principaux déterminants de ce risque. Nous pouvons désormais à l'aide de ces modèles identifier la présence :

- d'une zone à risque à l'aide de caractéristiques spatiales, en particulier grâce à la densité, le gradient de richesse (les habitations dans des zones riches proches de zones pauvres sont les plus exposées au risque), la richesse de la zone étudiée, la présence de postes de police à proximité...
- d'un individu ou d'une habitation à risque : en effet, des variables telles que l'âge de l'assuré, l'isolement de l'habitation, le fait d'être locataire ou propriétaire et de posséder un appartement ou une maison, affectent toutes le risque de cambriolage.

Par ailleurs, même s'il semble difficile de faire ressortir des phénomènes de répétitions temporelles proches au sein de nos données, nous avons toutefois mis en avant l'importance de l'aspect temporel. Par exemple, les cambriolages sont ainsi, conformément à la littérature, plus nombreux les vendredis et samedis, et moins nombreux les dimanches lorsque les individus sont en moyenne davantage présents chez eux. De même, une certaine saisonnalité mensuelle semble apparaître, et la météorologie a également son importance.

Le modèle spatial et le *gradient boosting* nous permettent de réaliser des interprétations très similaires. Cependant, rappelons que l'un des objectifs du projet était de mettre en place un système d'alerte au vol afin de mieux appréhender le risque de cambriolage. Ainsi, nous avons montré que le *gradient boosting* permettait d'obtenir de meilleures prédictions. Nous proposons donc à AXA d'utiliser plutôt un tel modèle qu'un modèle spatial, afin d'évaluer le risque des futurs assurés belges. Notons toutefois que le modèle Durbin montre qu'on a plus de risque d'être cambriolé lorsque son voisin est lui-même cambriolé, un résultat qui n'est pas présent dans le *gradient boosting* et qui est pourtant symbolique de l'importance de la modélisation spatiale du risque de cambriolage.

Ce projet, en plus de nous rapprocher du domaine des statistiques liées à l'assurance, nous a permis de découvrir les outils et méthodes de la statistique spatiale et de nous initier à l'apprentissage statistique. Ces deux méthodes ont un grand avenir dans un monde où les données numériques explosent et sont de plus en plus accessibles au point même que favoriser l'*open data* est devenu une priorité nationale. Nous espérons que ces résultats permettront à AXA d'avoir une meilleure connaissance du risque que portent ses assurés et ouvriront de nouveaux champs d'études. Des modèles plus poussés pourraient par exemple permettre de mieux lier les deux aspects spatiaux et temporels ensemble. De même, avec un budget plus important, une nouvelle étude pourrait récolter la distance réelle et non pas à vols d'oiseaux entre deux voisins en passant par exemple par l'API professionnelle Google Map moins limitée que sa version non payante. Ce type d'étude pourrait aussi être développé par les collectivités afin d'améliorer la sécurité des zones à risque. Ainsi déjà en 2002 au moins 13 % des forces de police des États-Unis utilisaient les statistiques spatiales pour élucider les crimes et les prévenir selon le National Institute of Justice (NIJ).

Références

M. A. Fishbein (1967), *Attitude and the prediction of behavior*. In M. Fishbein (Ed.), Readings in attitude theory and measurement. New York : Wiley, 1967, 477-492.

G. Becker (1968), *Crime and Punishment : An Economic Approach*. The Journal of Political Economy 76 : 169–217

P. J. Brantingham, P. L. Brantingham (1982), *Mobility, Notoriety and Crime : A Study of Crime Patterns in Urban Nodal Points*. Journal of Environmental System 11 :89-99

D. Cornish, R. Clarke (1987), *Understanding crime displacement : An application of rational choice theory*. Criminology, 25(4), 933–947

L. W. Kennedy, D. R. Forde (1990), *Routine activities and crime : an analysis of victimization in Canada*. Criminology, 28 : 137–152

R. J. Bursik, H. Grasmick (1993), *Economic deprivation and neighborhood crime rates, 1960-1980*. Law and Society Review, 27, pp. 263-283

U. Ericsson (1995), *Straight from the Horse's Mouth*. Forensic Update 43 :23-25

D. K. Rossmo (1999), *Geographic profiling system helps catch criminals*. GeoWorld, p. 41

P. Rountree, C. Kenneth (2000), *The Generalizability of Multilevel Models of Burglary Victimization : A Cross-City Comparision*. Social Science Research, 29, 284-305

J. Morenoff, R. Sampson, S. Raudenbush (2001), *Neighborhood Inequality, Collective Efficacy, and the Spatial Dynamics of Urban Violence*. Criminology 39 :517-560

R. Paternoster, S. Bushway (2001), *Theoretical and empirical work on the relationship between unemployment and crime*. Journal of Quantitative Criminology. 17, (4), 391-407

V. Ceccato, R. Haining, P. Signoretta (2002), *Exploring offence statistics in Stockholm City using spatial analysis tools*. Annals of the Association of American Geographers

T. Hartnagel (2004), *Correlates of Criminal Behaviour*. Criminology : A Canadian Perspective, 5th ed.

J. Savoie (2008), *L'analyse spatiale de la criminalité au Canada : résumé des principales tendances 1999, 2001, 2003 et 2006*. Série de documents de recherche sur

la criminalité et la justice, Centre canadien de la statistique juridique, Statistique Canada

T. Hastie, R. Tibshirani, J. Friedman (2008), *The Elements of Statistical Learning : Data Mining, Inference and Prediction.* Springer

Y. Feng (2010), *Robberies in Cincinnati* : http ://evasdatavisualization.weebly.com/

M. Fischer, J. Wang (2011), *Spatial Data Analysis : Models, Methods et Techniques.* Springer

Étude du bureau Research Solution sur les cambriolages en Belgique (2012) : un inquiétant manque de prévoyance

Criminalité en Belgique : les cambriolages en hausse de 7,5%, L'Avenir, 9 juillet 2013 : http ://www.lavenir.net/cnt/dmf20130709_00333956

Document de référence, rapport financier annuel, 2013, AXA

G. James, D. Witten, T. Hastie, and R. Tibshirani (2013). *An Introduction to Statistical Learning : with Applications in R.* Springer Texts in Statistics. Springer

B. De Laet (2014), *Regression trees and ensembles of trees in P&C pricing.* Master thesis, Master of Financial and Actuarial Engineering, KU Leuven

Site de la Police fédérale belge, pour les chiffres sur le cambriolage en Belgique : http ://www.polfed-fedpol.be/

Annexe 1 : Taux de sinistres et caractéristiques de l'habitation

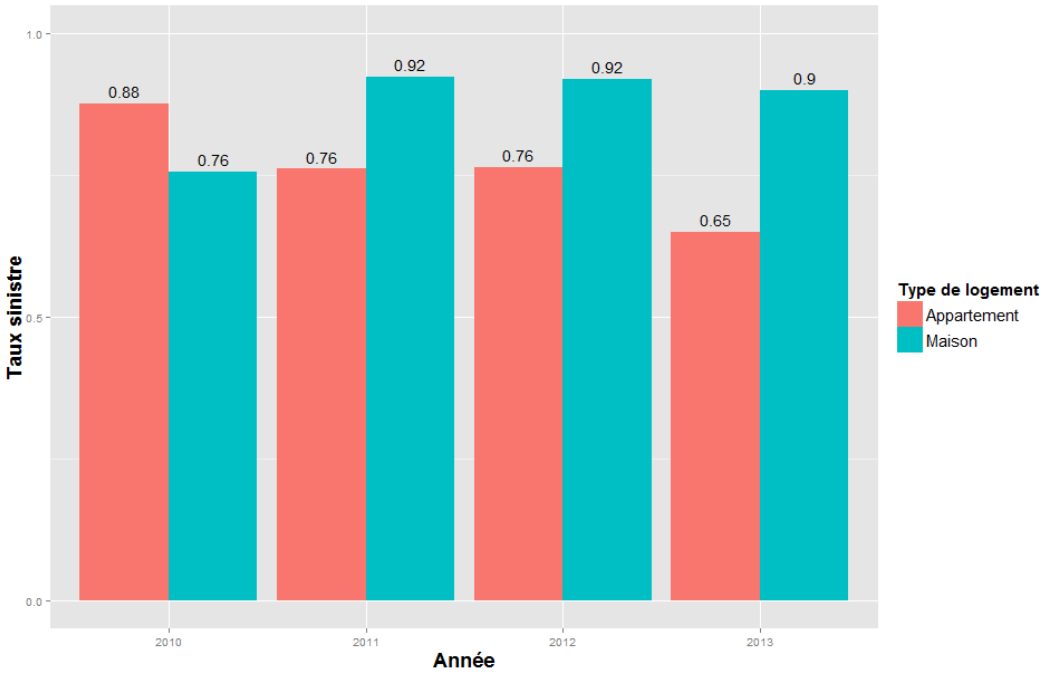

FIGURE 9 – Taux de sinistres : séparation appartement/maison

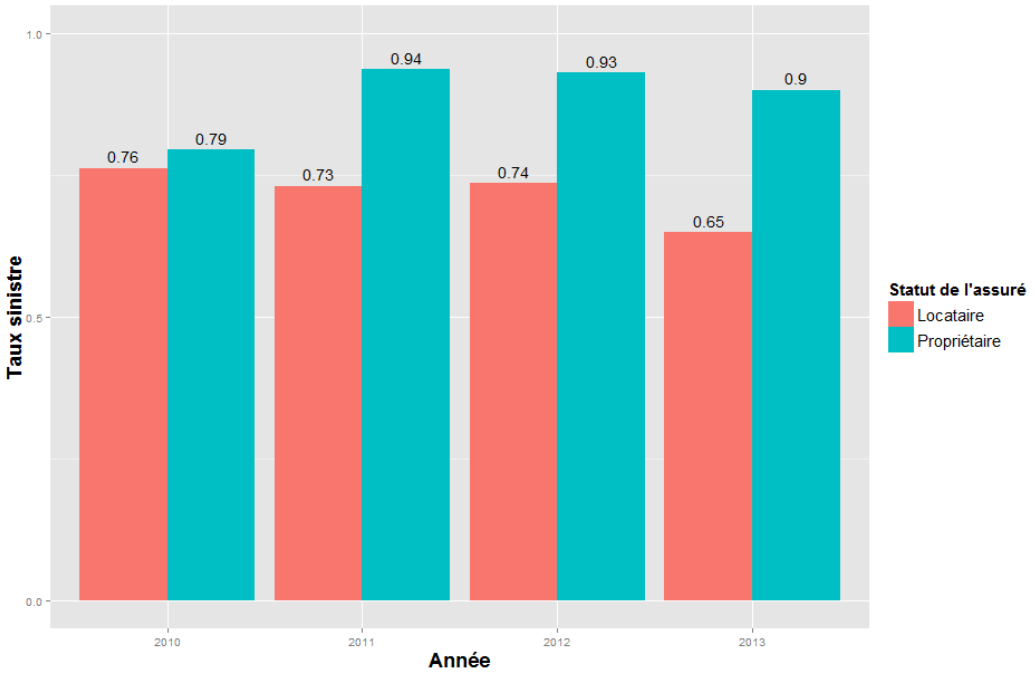

FIGURE 10 – Taux de sinistres : séparation propriétaire/locataire

Annexe 2 : ACP des zones *mosaic*

Cette annexe donne davantage de détails sur l'analyse en composantes principales des zones *mosaic* réalisée en partie 3.1.4. Les variables utilisées lors de cette ACP sont les suivantes :

- le revenu moyen de la zone ;
- l'âge moyen des habitants de la zone ;
- la densité de population ;
- le taux de maisons ;
- le taux de propriétaires ;
- le taux d'habitations jugées isolées.
- les six variables évaluant la proximité de la zone aux points d'intérêt OpenStreetMap suivants : les postes de police, les supermarchés, les écoles, les restaurants, les fastfoods et les bars.

Comment ces dernières variables sont-elles construites ? Pour évaluer la proximité d'une zone aux postes de police par exemple, nous avons toute d'abord calculé la moyenne de la distance euclidienne entre une habitation assurée dans la zone et le poste de police le plus proche. Puis nous avons découpé ces distances moyennes par décile, et attribué à chaque zone une valeur entre 1 et 10. Ainsi, une zone ayant un 1 pour la variable Police ci-dessous correspondra à une zone ayant très peu de postes de police à proximité, voire pas du tout. À l'inverse, une zone ayant un 10 en aura beaucoup. La même démarche a été adoptée pour les cinq autres points d'intérêts considérés ici.

Voici ci-dessous la matrice des corrélations entre les différentes variables de l'ACP. Les principales valeurs remarquables étaient attendues. Les cinq coefficients positifs et négatifs les plus élevés ont été reportés en couleur.

	Age	Rev	Den	TxIso	TxProp	TxMai	Police	Bar	Super.	FastF.	Restau	École
Age	1.00	0.06	-0.07	0.17	0.30	0.22	-0.10	-0.10	-0.09	-0.09	-0.09	-0.08
Revenu Moyen	0.06	1.00	-0.72	0.32	0.10	0.20	-0.40	-0.30	-0.14	-0.00	-0.17	-0.12
Densité Population	-0.07	-0.72	1.00	-0.34	-0.15	-0.38	0.51	0.35	0.32	0.21	0.38	0.29
Taux Hab. Isolées	0.17	0.32	-0.34	1.00	0.30	0.39	-0.39	-0.32	-0.30	-0.23	-0.30	-0.26
Taux Propriétaires	0.30	0.10	-0.15	0.30	1.00	0.57	-0.23	-0.25	-0.20	-0.22	-0.25	-0.21
Taux Maisons	0.22	0.20	-0.38	0.39	0.57	1.00	-0.42	-0.40	-0.36	-0.37	-0.40	-0.35
Postes de Police	-0.10	-0.40	0.51	-0.39	-0.23	-0.42	1.00	0.68	0.53	0.50	0.63	0.53
Bars	-0.10	-0.30	0.35	-0.32	-0.25	-0.40	0.68	1.00	0.54	0.60	0.65	0.51
Supermarchés	-0.09	-0.14	0.32	-0.30	-0.20	-0.36	0.53	0.54	1.00	0.62	0.60	0.59
Fast Foods	-0.09	-0.00	0.21	-0.23	-0.22	-0.37	0.50	0.60	0.62	1.00	0.68	0.55
Restaurants	-0.09	-0.17	0.38	-0.30	-0.25	-0.40	0.63	0.65	0.60	0.68	1.00	0.49
Écoles	-0.08	-0.12	0.29	-0.26	-0.21	-0.35	0.53	0.51	0.59	0.55	0.49	1.00

TABLE 10 – ACP : matrice des corrélations

Voici l'éboulis des valeurs propres. Une cassure très nette est présente entre la première et la seconde valeur. Nous aurions donc pu nous arrêter à l'interprétation du premier axe. Nous avons toutefois interprété également le second axe en partie 3.1.4.

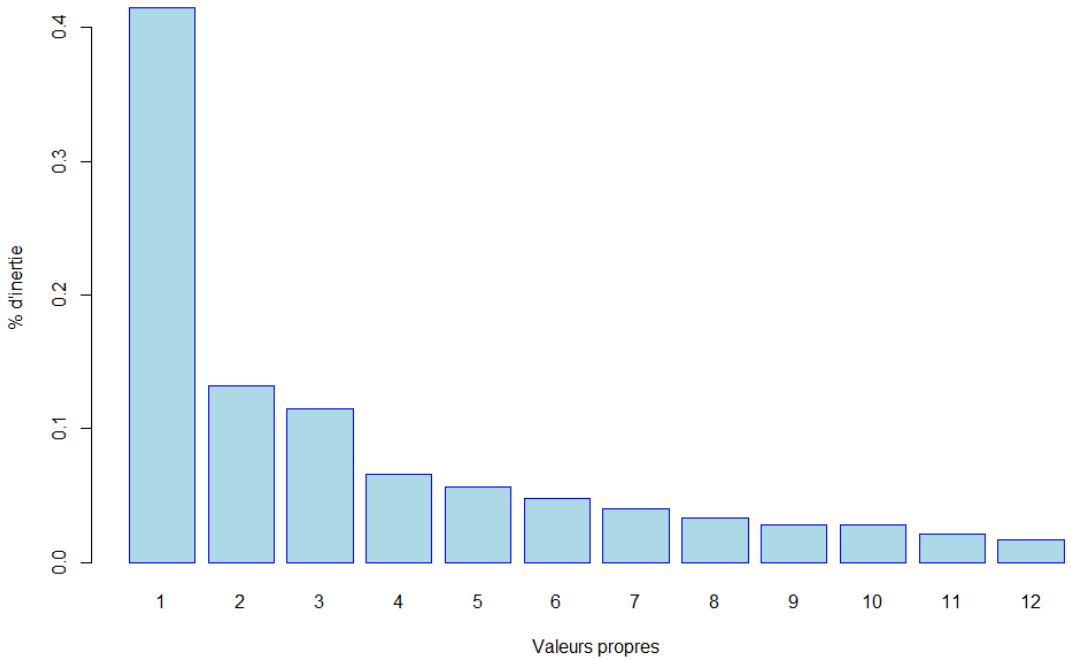

FIGURE 11 – ACP : ébouli des valeurs propres

Voici enfin la protection des variables initiales dans le plan engendré par les deux premières composantes principales. La position et l'ampleur des variables (c'est-à-dire la proximité avec le cercle unité) au sein de ce plan nous permettent d'aboutir à l'interprétation en partie 3.1.4.

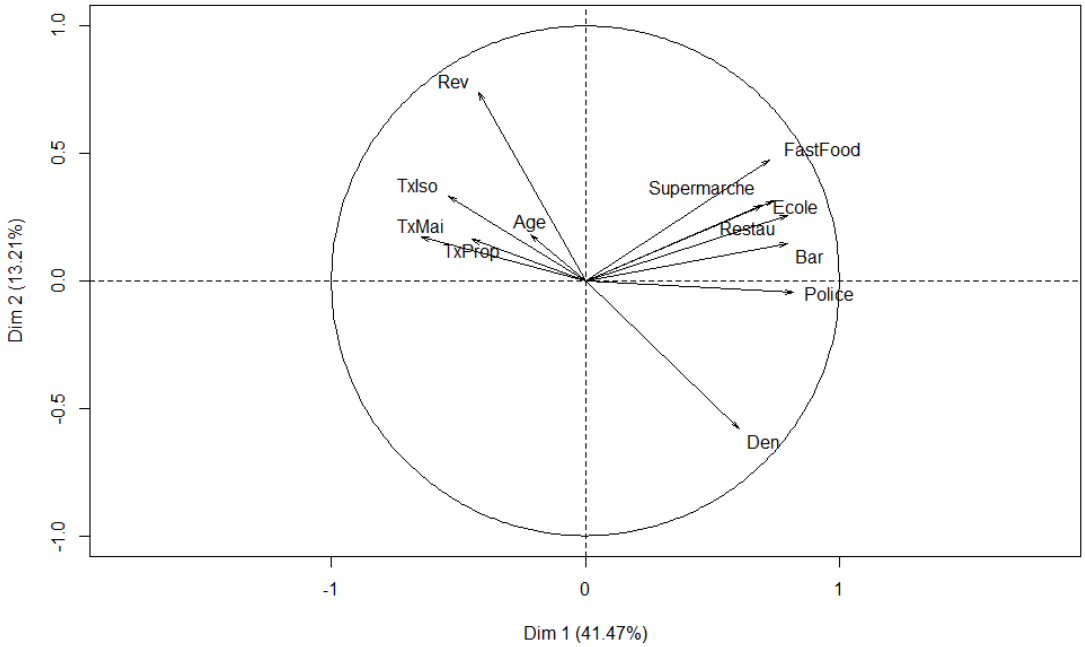

FIGURE 12 – ACP : projection des variables dans le plan engendré par les C.P. 1 et 2

Annexe 3 : Découpage des taux de sinistres par mois

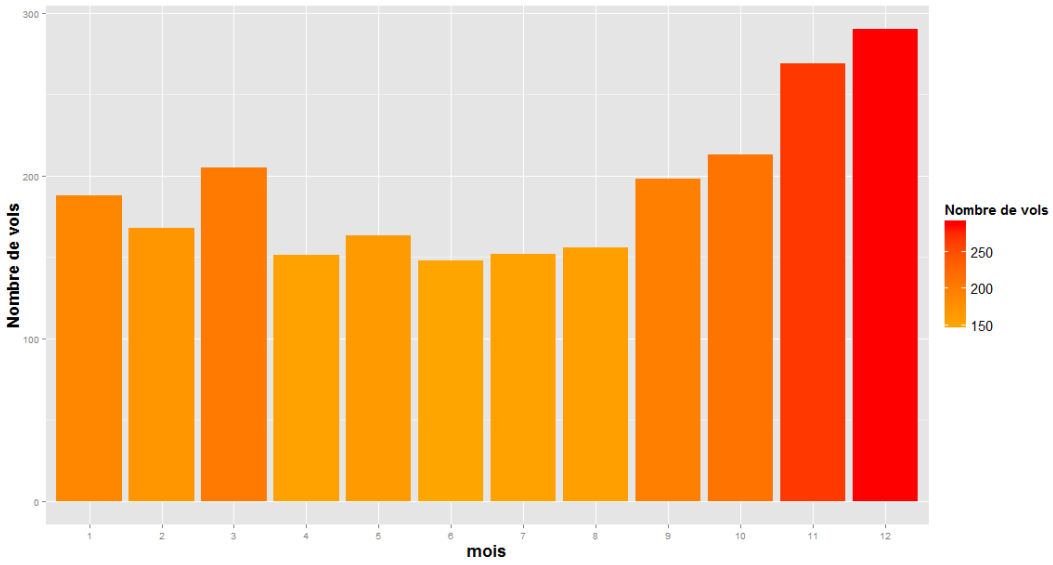

FIGURE 13 – Taux de sinistres mensuels

Annexe 4 : Sinistres et précipitations

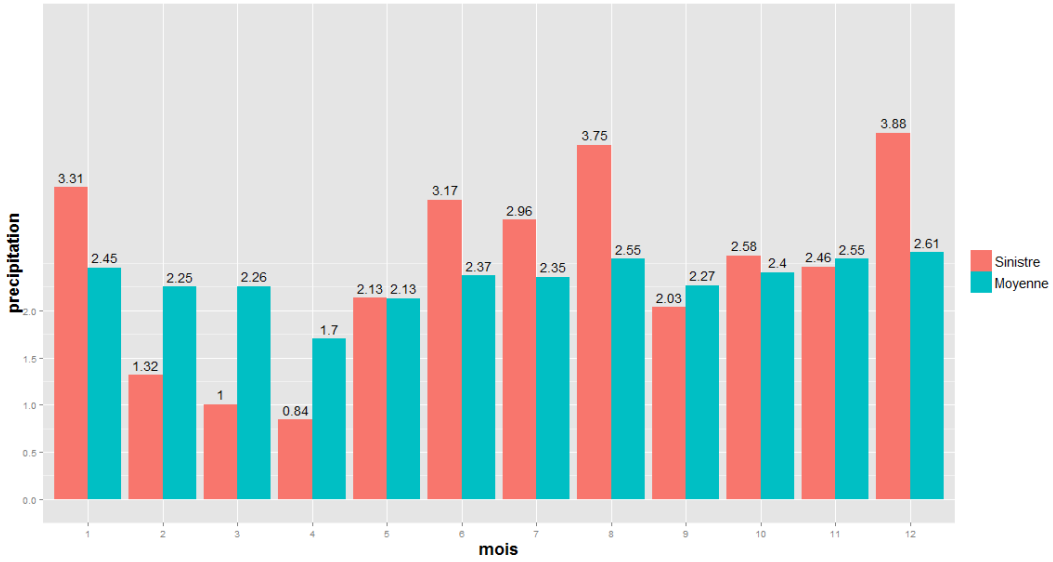

FIGURE 14 – Taux de sinistres et précipitations

Annexe 5 : Phénomènes de répétition proche : modèle logit

Le coefficient associé au ratio de nombre de vols sur le nombre d'assurés voisins au cours du mois précédent n'est pas significatif. Le tableau ci-dessous concerne février 2010, mais des résultats similaires s'obtiennent en prenant en compte d'autres périodes de temps, ou par exemple le ratio durant les trois derniers mois au lieu du dernier mois.

	Modèle logit
(Constante)	-7.09^{***}
	(0.44)
Ratio Nb Vols/Assurés	-0.37
	(2.32)
Age de l'assuré	-0.00
	(0.00)
Revenu Moyen	-0.00^{*}
	(0.00)
Densite population	-0.00^{***}
	(0.00)
Hab. Isolée	-0.01
	(0.09)
Propriétaire	1.09^{***}
	(0.18)
Maison	0.96^{***}
	(0.19)
AIC	7604.56
BIC	7761.18
Log Likelihood	-3787.28
Num. obs.	252921

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

TABLE 11 – Phénomènes de répétition proche : modèle logit (février 2010)