Modélisation spatio-temporelle du risque de cambriolage

Elie Dadoun et Ivan Herboch

Option Mathématiques Appliquées

Ecole Centrale Paris

Acknowledgements

And I would like to acknowledge ...

Abstract

This is where you write your abstract ...

Table des matières

Table des figures

Liste des tableaux

Chapitre 1

Généralités

1.1 Cambriolage en France : chiffres clés

Le cambriolage est l'effraction du domicile dans le but d'y commettre un vol. L'auteur de ce type de vol est passible de 3 ans d'emprisonnement et de 45 000 € d'amende.

En 2013, il s'est produit 382 000 cambriolages dont le coût moyen pour l'assureur est de 6500€[6]. Cela représente une hausse de 6,4% par rapport à 2012 en zone police (urbaine) et de 4,7% en zone gendarmerie (rurale). Les cambriolages dans les habitations principales ont respectivement augmenté, dans ces mêmes zones, de 7 % et de 1,3 % et ceux des résidences secondaires de 10 % et 17,7 %.

Un cambriolage dure en moyenne 5 minutes, sans jamais excéder 20 minutes. Une sonnerie d'alarme fait fuir 95% des voleurs. Il est intéressant de constater que 80% des cambriolages ont lieu en journée dont 55% uniquement entre 14h et 17h.

50% des cambriolages concernent la résidence principale, 6% les résidences secondaires et 44% les locaux professionnels. Cela peut éventuellement s'expliquer par la présence plus probable d'objets de valeur dans les résidences principales et locaux professionnels que dans les résidences secondaires qui sont la plupart du temps vides.

Les départements les plus touchés sont en premier lieu ceux d'Outre-Mer avec un taux de 6,5 cambriolages pour 1000 habitants en Guadeloupe. Viennent ensuite le Vaucluse, les Pyrénées-Orientales, l'Hérault, le Gard, les Bouches du Rhône et les Alpes-Maritimes qui ont des taux compris entre 5,4 et 6,5 pour 1000 habitants. L'Île-de-France est moins touchée (2,7 cambriolages pour 1000 habitants) mais son nombre d'habitants en fait l'une des régions dans lesquelles on compte le plus de cambriolages (environ 10% des cambriolages français).

La figure 1.1 permet de visualiser les différents taux de cambriolages selon les régions en France.

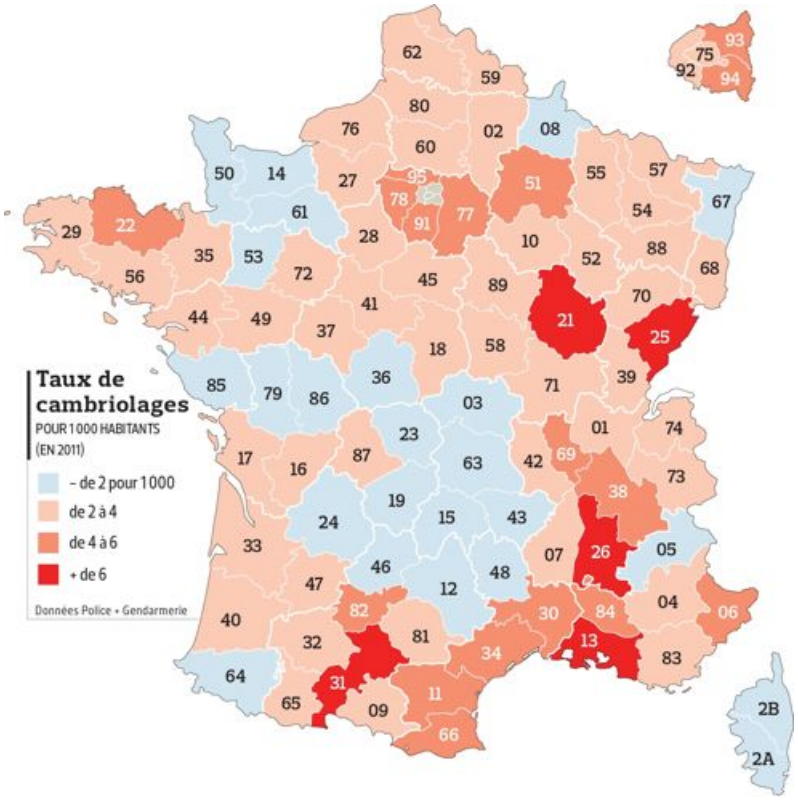


FIGURE 1.1 Taux de cambriolages pour 1000 habitants en 2011 (Source : Le Figaro)

1.2 Facteurs socio-économiques du cambriolage

Le but de cette partie est de donner les principales variables explicatives du risque de cambriolage, en commençant par les données socio-économiques.

1.2.1 Contexte

D'après une étude canadienne [14], les cambrioleurs ont tendance à commettre leur méfait à proximité de leur lieu de résidence. Cela suggère que des disparités en terme de classes socio-professionnelles entre deux zones tendent à augmenter la probabilité qu'un cambriolage soit commis dans la zone aisée par rapport à la probabilité qu'un cambriolage soit commis dans une zone aisée entourée de zones aisées.

De plus, les variables suivantes semblent bien expliquer le risque de cambriolage : la proportion de la population âgée de 15 ans ou plus titulaire au plus d'un CAP ou BEP [7][15], la valeur moyenne des logements dans le quartier [19][3], le revenu moyen par foyer [15], le taux de chômage [9], le nombre de foyers mono-parentaux, le nombre de personnes sans revenu, le pourcentage d'appartements loués, le pourcentage d'emménagements dans les 5 dernières années, le pourcentage de minorités dans la population [4].

1.2.2 Sources possibles de données

Il est possible d'utiliser des données socio-économiques, disponibles par exemple auprès de l'INSEE, pour alimenter le modèle d'évaluation du risque de cambriolage. Pour pouvoir prédire précisement le risque de cambriolage, il est nécessaire d'avoir une granularité suffisante dans nos données. Une fois ces données extraites, nous pourrons les utiliser dans un modèle prédictif.

Revenu moyen par foyer Le revenu fiscal des ménages par IRIS (Ilots Regroupés pour l'Information Statistique) est disponible sur le site de l'INSEE pour Paris et Marseille. Plusieurs informations liées peuvent être utilisées comme la part des ménages fiscaux imposés ou la valeur des déciles 1 à 9 du revenu fiscal moyen.

Classe socio-professionnelle par IRIS L'information est disponible par IRIS sur le site de l'INSEE.

Nombre de demandeurs d'emploi par arrondissement L'information sur le nombre de demandeurs d'emploi par arrondissement est disponible sur le site de l'INSEE. En recoupant avec le nombre d'habitants par arrondissement, on peut déterminer le taux de chômage.

1.3 Facteurs météorologiques

Le risque de cambriolage semble influencé par les conditions météorologiques. Nous exposons dans cette partie des pistes possibles à partir de références, et définissons des variables potentiellement explicatives du risque de cambriolage.

1.3.1 Variations saisonnières

Contexte La variation saisonnière de la fréquence d'activités criminelles a depuis longtemps été établie. Un rapport d'enquête nationale sur le crime aux Etats-Unis en 1980 [13] montre ainsi que le nombre de cambriolages est 30 à 40% plus élevé en été qu'en hiver, et que le motif de variations est assez reproductible d'une année sur l'autre, comme le montre la figure 1.2.

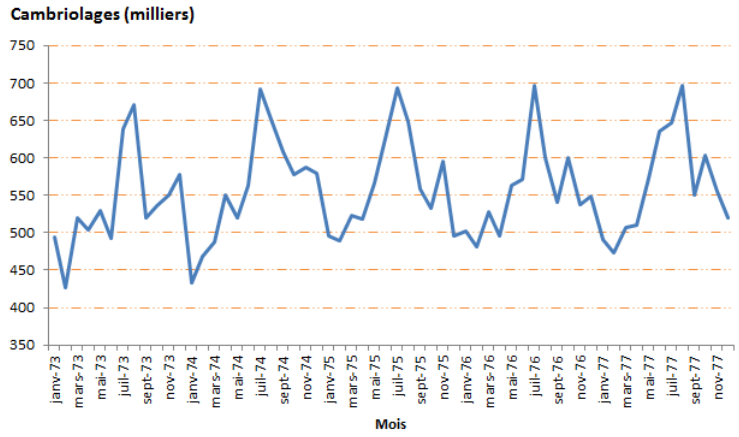


FIGURE 1.2 Nombre de cambriolages par mois entre 1973 et 1977 aux Etats-Unis[13]

Intuitivement, ces variations peuvent s'expliquer par le fait qu'en été, les habitations sont davantage inoccupées, notamment en raison des vacances, et les cambrioleurs ont plus d'opportunités pour commettre leurs crimes.
Plus généralement, on peut penser que par beau temps, les gens sortent plus souvent de chez eux et plus longtemps. Il y a donc vraisemblablement une explication dans le caractère opportuniste du cambriolage.

1.3.2 Influence de la température

Les variations présentées précédemment se retrouvent également pour d'autres types d'activités criminelles comme les agressions ou les vols. Cela suggère un effet psychologique du climat et notamment de la température sur le comportement criminel[2] . Des projections précises des hausses de températures dans le contexte du changement climatique permettent même de prédire la hausse de la criminalité pour les prochaines décennies[16].

1.3.3 Sources de données

Les données de certaines stations météorologiques sont en accès libre[1] et peuvent être utilisées pour décrire l'influence des conditions météorologiques sur le risque de cambriolage. Elles consistent souvent en des relevés journaliers des températures du jour et des précipitations.

1.4 Influence de la configuration spatiale

L'idée a priori est que des habitations "isolées" seront plus souvent cambriolées que des habitations "exposées". Il s'agit donc définir le niveau d'isolement d'une habitation et de trouver des variables qui permettent de le quantifier.

Une habitation est d'autant plus isolée qu'elle est peu visible et son environnement peu fréquenté. Nous envisageons quelques pistes permettant de quantifier le niveau d'isolement d'une voie.

1.4.1 Type de voie

Le type de voie nous paraît être une première variable explicative du niveau d'isolement d'une habitation. Par exemple, une habitation située dans une impasse est davantage isolée qu'une grande avenue, car elle est moins accessible.
Simon C.F.Shu et Jason N.H.Huang [5] ont montré que l'accessibilité d'une voie est un facteur influençant le risque de cambriolage, les voies les plus passantes étant moins vulnérables. Ils ont ainsi défini 4 types de voie, en fonction de leur largeur et de leur fréquentation possible par des piétons et véhicules. Leur étude montre qu'un cambriolage est 3 fois plus probable dans une voie privée que dans une rue autorisant le passage de piétons et de véhicules.

nom de la voie L'intitulé de la voie peut être un premier indice du niveau d'isolement de la voie. Les villes françaises utilisent une multitude de noms pour désigner différents types de voies : boulevard, rue, impasse, place, ...Des caractéristiques de la voie, comme sa largeur, ou son exposition, peuvent généralement être inférées à partir du nom de voie attribué. Ainsi, un boulevard est une voie assez large, comprenant en général au moins 4 voies de circulation, et est donc a priori une voie très fréquentée.

Une impasse, à l'inverse, et a priori très peu fréquentée, tandis qu'une place est a priori visible et exposée.

1.4.2 Distance à un point d'intérêt

L'éloignement d'une habitation vis-à-vis de certains points d'intérêt peut permettre de quantifier le niveau d'isolement d'une habitation. Par exemple, la distance au commissariat de police le plus proche est intuitivement une donnée intéressante.
La distance à d'autres points d'intérêt qui caractérisent un lieu en général animé et fréquenté, comme une boulangerie par exemple, nous semble aussi intéressante.

1.4.3 Sources de données

Le nom du type de voie peut être récupéré à partir des données fournies dans certains champs d'adresse de la base de données d'AXA Global P&C. La difficulté réside principalement dans la qualité des données et leur retraitement. Par exemple, l'abréviation pour 'boulevard' est parfois 'BVD' ou 'BD' ou encore 'BOUL' et doit être reconnue dans la chaîne de caractère décrivant l'adresse.

Une autre source de donnée est envisageable à partir de la plateforme 'OpenStreetMap'. Certaines caractéristiques des voies sont en général indiquées, comme par exemple le nombre de voies de circulation. Les localisations des points d'intérêt comme les commissariats ou les boulangeries peuvent aussi être récupérées.

Etude envisageable Définir différentes classes de voies à partir des noms de voies. Par exemple, on pourrait définir les classes suivantes pour regrouper :

- classe 1 : "boulevard", "avenue" ... : voies a priori très fréquentées tant par les piétons que par les véhicules
- classe 2 : "place", "carrefour", "square", ... : espaces très exposés
- classe 3 : "rue" ... : voies standards, autorisant le passage de véhicules
- classe 4 : "chemin", "traverse", ... : petites voies rurales
- classe 5 : "impasse", "passage", ... : voies privées/réservées urbaines
- classe 6 : "résidence", "cité", "cour", ... : quartiers résidentiels

Données A partir des données des assurés, on peut construire une classification représentative. Pour chaque type de voie, on peut calculer le pourcentage de sinistrés.

On peut moduler la classification pour éventuellement chercher à avoir des groupes bien distincts (différents niveaux de risques).

Chapitre 2

Etude du portefeuille

2.1 Etude du portefeuille

Avant d'extraire et d'étudier des données provenant d'autres sources, regardons les données à disposition et imprégnons nous des ordres de grandeurs.

2.1.1 Constitution du portefeuille

Le portefeuille dont nous disposons est le portefeuille Multi-Risque Habitation (MRH) d'AXA France pour les départements 13, 75, 92 et 93 et les années 2008 à 2013.
Les données relatives à ce portefeuille sont constituées de deux bases :

- la base *police*, qui regroupe les données sur les assurés et les risques (habitations) : adresse du risque, caractéristiques du logement, ...
- la base *sinistres*, qui regroupe les données sur les sinistres : date de survenance, type de sinistre, charge du sinistre, ...

Pour chaque année, nous avons ainsi une table *police* des contrats en vigueurs et une table *sinistre* des sinistres survenus au cours de l'année. Nous nous intéressons pour ce projet qu'à une partie des sinistres d'habitation : les cambriolages.

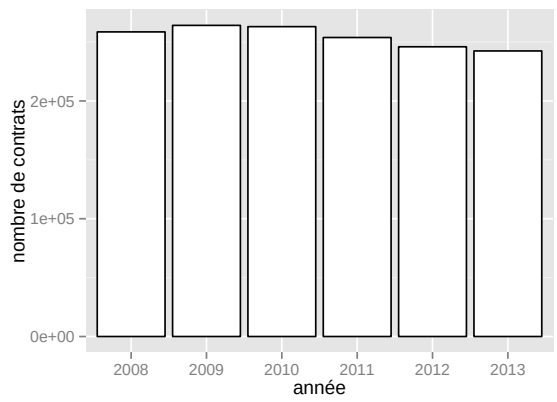
Nous allons préciser dans la suite les ordre de grandeurs pour les taux et nombres de sinistres, et nous montrons que certaines données de la base permettent d'expliquer dans une certaine mesure le risque de cambriolage.

2.1.2 Ordres de grandeur

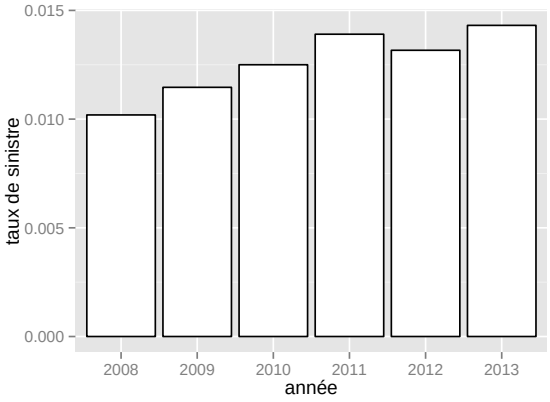
La base *police* contient environ 250 000 lignes par an. La base sinistre quant à elle contient environ 3000 lignes relatives à la catégorie 'vol' qui nous intéresse exclusivement ici.
L'ordre de grandeur du taux de cambriolage est donc environ $3000/250000 = 1,2\%$.
Cela veut dire qu'on observe en moyenne seulement 12 cambriolages sur 1000 adresses assurées.

2.1.3 Evolution du portefeuille et du taux de sinistre

Les graphes 2.1a et 2.1b montrent l'évolution du portefeuille et du taux de sinistres entre 2008 et 2013.



(a) Evolution du portefeuille



(b) Evolution du taux de sinistres

On observe que le portefeuille a légèrement diminué, et que le taux de sinistres est à la hausse. On va voir qu'on peut obtenir plus d'information si on regarde le taux de sinistres par département, ou selon les caractéristiques de l'habitation.

2.1.4 Taux de sinistres par département

Les départements à l'étude sont le 13, le 75, le 92 et le 93. Nous observons que la sinistralité en terme de cambriolage n'est pas la même dans ces départements comme on le voit sur la figure 2.2

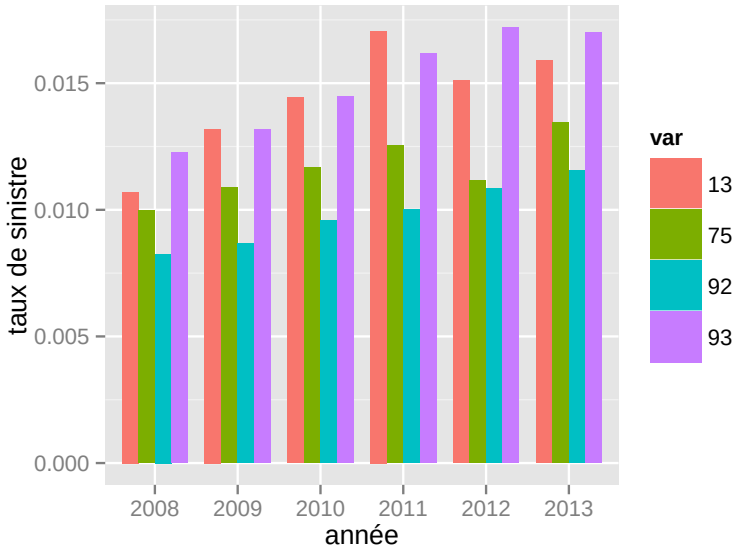


FIGURE 2.2 taux de cambriolage par département et par an

On voit par exemple que le 93 est en 2012 et 2013 50% plus sinistré que le 92. Le département de marseille présente aussi un taux de sinistres comparable et en tout cas plus important que la moyenne. On aimerait trouver des variables explicatives pour expliquer ces différences entre les départements.

On s'intéresse maintenant aux informations disponibles sur l'habitation et on regarde les variations du taux de sinistres en fonction des caractéristiques de l'habitation.

2.1.5 Informations sur l'habitation

Les habitations assurées sont diverses : des maisons ou des apparements, des résidences principales ou secondaires, ...

Il est intéressant de voir les taux de sinistres selon les caractéristiques des habitations, et de constater qu'ils diffèrent.

Propriétaires ou locataires On constate que les propriétaires sont nettement plus cambriolés que les locataires ou copropriétaires (figure 2.3)

Résidence principale ou secondaire Les résidences principales sont davantage touchées que les résidences secondaires (figure 2.4).

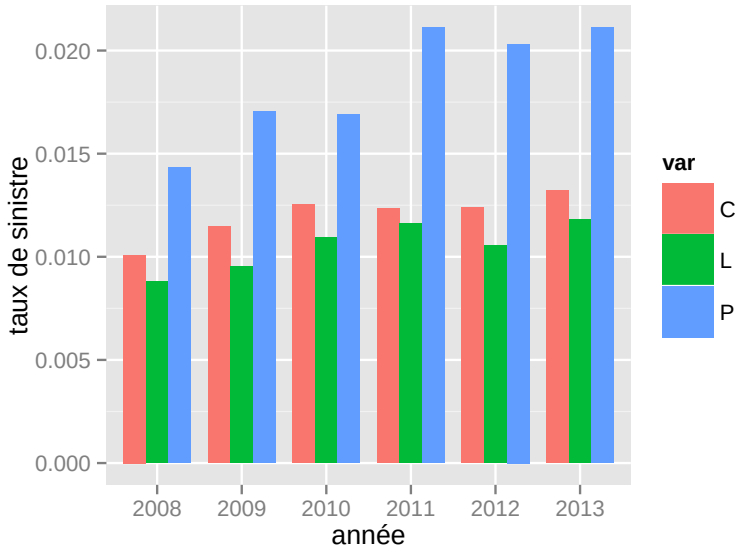


FIGURE 2.3 taux de cambriolage pour les propriétaires (P), locataires (L) ou copriétaires(C)

On peut penser ici à un biais. En effet, on peut concevoir que les assurés n'ont pas toujours connaissance des cambriolages survenus dans leur résidence secondaire et qu'une partie de ces cambriolages n'est pas signalée.

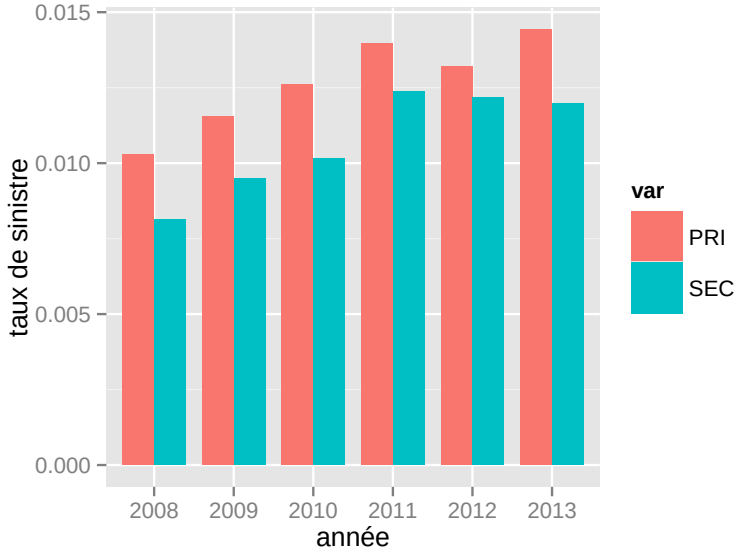


FIGURE 2.4 taux de cambriolage pour résidences principales (PRI) ou secondaires (SEC)

Maison, appartement ou Rez-de-chaussée Les différences des taux de sinistres en fonction du type d'habitation sont conformes à l'intuition : les maisons étant davantage isolées, elles sont plus souvent cambriolées. De même, les rez-de-chaussée sont plus accessibles et sont davantage cambriolés que les appartements (figure 2.5).

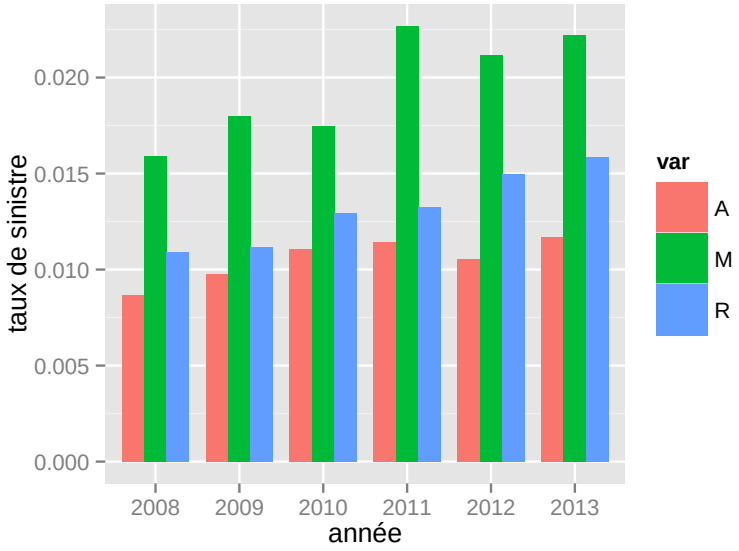


FIGURE 2.5 taux de cambriolage pour les maisons (M), les appartements (A) ou rez-de-chaussée (R)

Ancienneté de l'habitation Selon l'ancienneté de l'habitation, on observe également des différences, plus ou moins reproductibles d'une année sur l'autre.
Les habitations les plus anciennes semblent plus protégées. Cela peut s'expliquer par le fait que les immeubles anciens disposent plus souvent d'un gardien ou d'un concierge. Cependant, il y aurait un biais évident dans le renseignement de l'ancienneté du logement, surtout si cette indication n'est pas mis à jour d'une année sur l'autre.

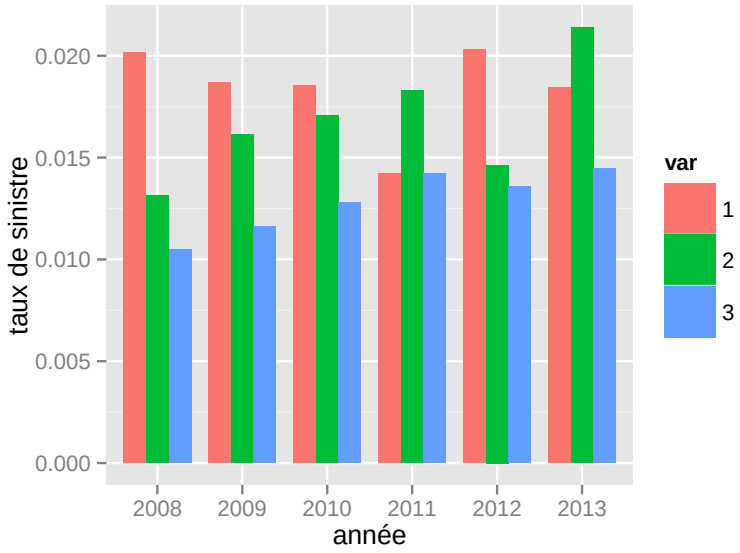


FIGURE 2.6 taux de cambriolage en fonction de l'ancienneté du logement : 1 (moins de 5 ans), 2 (5 à 10 ans), 3 (plus de 10 ans)

2.2 Variations temporelle des sinistres

On s'intèresse à la distribution temporelle du nombre de sinistres du portefeuille entre les années 2008 à 2013.

2.2.1 Variations mensuelle

On compte le nombre de sinistres survenus pour chaque mois de l'année calendaire. On obtient l'histogramme en figure 2.7

On voit que le mois de décembre est moins sinistré par rapport aux autres mois de l'année.

On peut aussi aggréger les sinistres par saison. Cela fait l'objet de la sous-section suivante.

2.2.2 Variations saisonnière

Les périodes des saisons sont définies par la donnée des solstices d'hiver et d'été et des équinoxes d'automne et de printemps.

On obtient la distribution suivante en figure 2.8

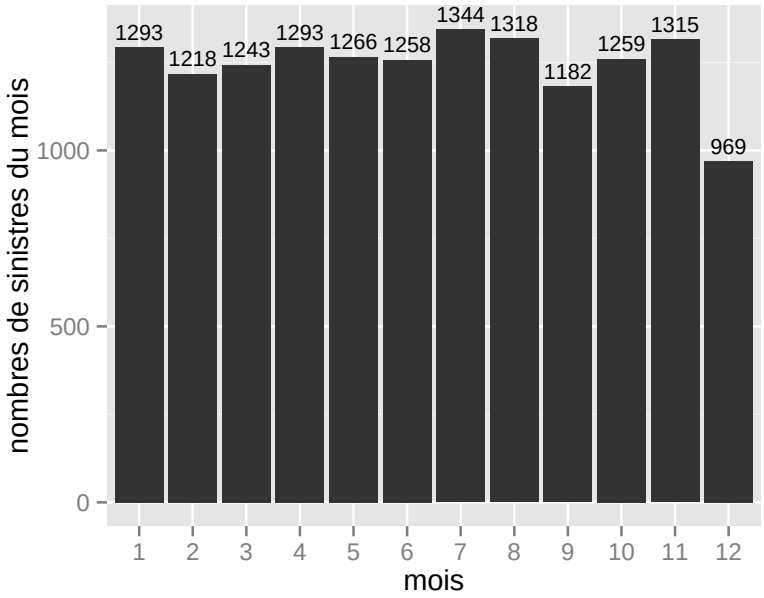


FIGURE 2.7 Variation mensuelle

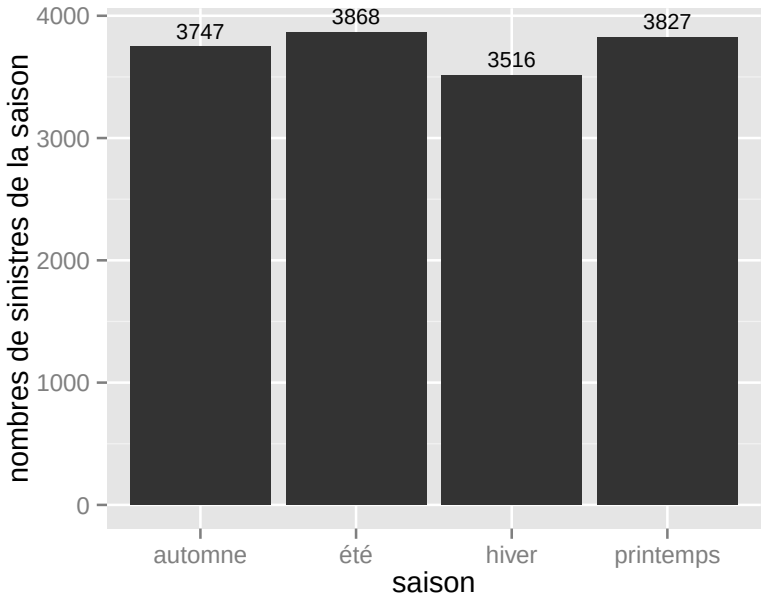


FIGURE 2.8 Variation saisonnière du nombre de cambriolages

Il y a près de 10% moins de sinistres pendant la saison hivernale par rapport aux autres saisons. Cela peut s'expliquer par le fait qu'il fait plus froid en hiver et que les gens restent plus souvent et plus longtemps chez eux, ce qui laisse moins d'opportunités de vol pour les cambrioleurs.

2.3 Représentation spatiale des risques

Tout au long de ce projet, il a été nécessaire de représenter spatialement les résultats afin de conforter une intuition ou de visualiser concrètement des résultats. Nous expliquons dans cette section les méthodes utilisées et les problématiques rencontrées.

2.3.1 Retraitement des adresses

Avant de représenter spatialement les risques, la première étape a été de leur attribuer une adresse postale sous une forme standard à partir des données.
Les données ont été retraitées afin de suivre le schéma suivant :

[Numero de rue] [nom de la rue] [code postal] [ville].

La chaîne de caractères ainsi obtenue peut facilement être interprétée par un moteur pour être convertie en une donnée de géolocalisation de type longitude/latitude.

Le retraitement des données n'a pas toujours été facile dans la mesure où ces 4 informations ne sont pas toujours indiquées dans les mêmes champs de la base, et que certaines lignes possèdent des compléments d'adresse renseignant par exemple l'étage où le palier, ce qui ne nous intéressent pas ici.
Notre méthode de retraitement a consisté à identifier des chaînes de caractères (par exemple 5 chiffres consécutifs entre 0 et 9 pour le code postal) et procéder par élimination.

Par exemple, on identifie un code postal à l'aide de l'expression régulière "$[0-9]\{5\}$" [12]. On a utilisé la librairie *stringr* [22] sous R.

Il faut noter que toutes les adresses n'ont pu être retraitées de façon automatique et qu'on a une perte évaluée à environ 5%.

2.3.2 Géocodage

Il est possible d'obtenir les coordonnées gps d'une adresse postale grâce à l'API Google maps via la librairie *RgoogleMaps* de R [10].

Exemple :

```
geocode('9 avenue de Messine 75008 Paris').
```

longitude	latitude
2.3132474	48.8761199

Limitation L'API Google autorise 2500 requêtes par jour. Le portefeuille d'AXA France est constitué de près de 90 000 adresses dans le 75, 40 000 dans le 93, 65000 dans le 13 et 45000 dans le 92.

Précision La précision renvoyée par Google est en 10^-7 de degré près pour la latitude et la longitude, ce qui correspond à une précision inférieure à 1 mètre. En pratique bien sûr, on a une précision de l'ordre de la taille de l'immeuble où se trouve le risque.

2.3.3 Convertion (Lambert)

Pour certaines considérations dont le calcul des plus proches voisins, il est nécessaire de projeter les coordonnées GPS longitude/latitude en coordonnées Lambert X/Y [23].

La librairie utilisée sous R est *sd* [18].

2.3.4 Notion de voisinage

Pour considérer un voisinage d'un risque, on peut procéder de deux façons :
— On se limite au K plus proches voisins du risque, K étant fixé.
— On se limite aux voisins situés à une distance inférieure à D fixée

Méthode des K plus proches voisins Cette méthode a l'avantage de considérer un même nombre de voisins pour chaque risque. Elle est illustrée en figure 2.9.

Selon la densité de population et la configuration spatiale, la distance entre le risque et son Kème plus proche voisin est plus ou moins élevée.

Méthode du disque de voisinage Cette méthode a l'avantage de fixer un voisinage de manière absolue. Elle est illustrée en figure 2.10

En revanche, le nombre de voisins est alors très variable selon la densité de contrats et la configuration spatiale.

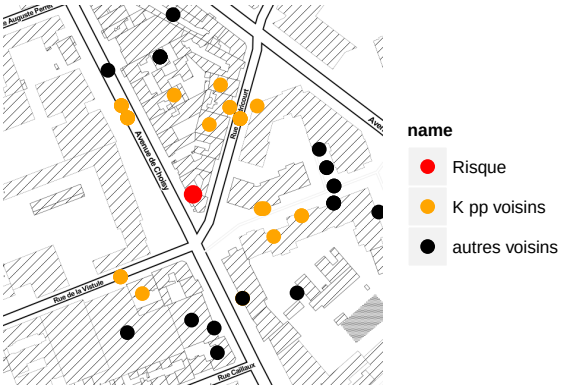


FIGURE 2.9 Illustration du voisinage d'un risque avec la méthode des K plus proches voisins

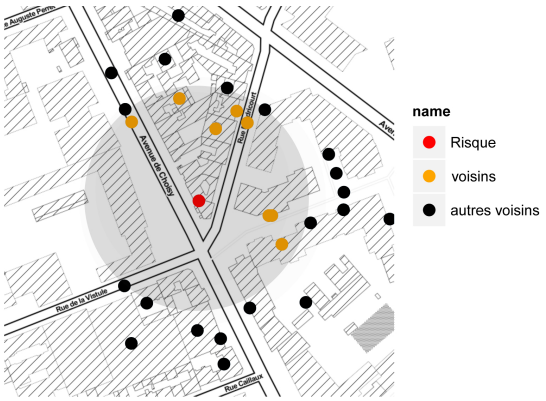


FIGURE 2.10 Illustration du voisinage d'un risque avec la méthode du disque de voisinage

Choix de méthode Pour le calcul de taux de sinistres, on préférera la méthode des K plus proches voisins, qui permet de s'affranchir de l'exposition du portefeuille. En effet, la sinistralité calculée sous la forme :

$$\frac{1}{K} \text{ (Nombre de sinistres survenu dans le voisinages de K voisins)}$$

est indépendante de la densité locale de contrats (exposition du portefeuille), et permet donc de comparer de façon juste des sinistralités locales.

2.3.5 Choix d'une distance

La distance utilisée est la distance euclidienne "à vol d'oiseau". Celle-ci est en effet directement accessible à partir des coordonnées Lambert. Une autre distance envisage est la distance à pied, calculée par l'API Google, mais nécessite des requêtes supplémentaires.

2.3.6 Représentation spatiale des sinistres et des contrats

On représente la répartition spatiale des sinistres et des contrats. Un calcul de densité spatiale nous permet de représenter une distribution dans l'espace : fonction *stat_density2d* de la librairie *ggplot2* [21].

Sur une première carte, on représente la distribution des sinistres, pour constater que certaines zones compte un plus grand nombre de sinistres que d'autres.

On constate qu'il y a un plus grand nombre de sinistres dans le nord-ouest de Paris et le centre, notamment dans les 16ème, 17ème, 18ème, 9ème et 3ème arrondissements.

Cependant, on doit garder en tête que la répartition des contrats n'est pas homogène et que certaines zones sont plus denses que d'autres. C'est pourquoi, il est aussi intéressant de représenter la distribution spatiale des contrats.

Le fait que certaines zone soient plus fortement sinistrées est partiellement expliqué par le fait que ces zones sont plus exposées. Cependant, on voit que certaines zones très exposées comme le sud-ouest de Paris comptent relativement peu de sinsitres, et que d'autres zones moins exposées comme une partie du 18ème et du 19ème arrondissement comptent relativement plus de sinistres.

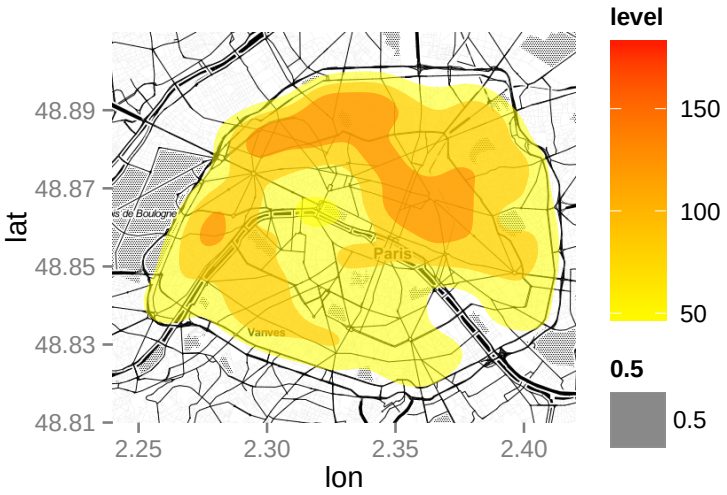


FIGURE 2.11 Répartition des sinistres à Paris

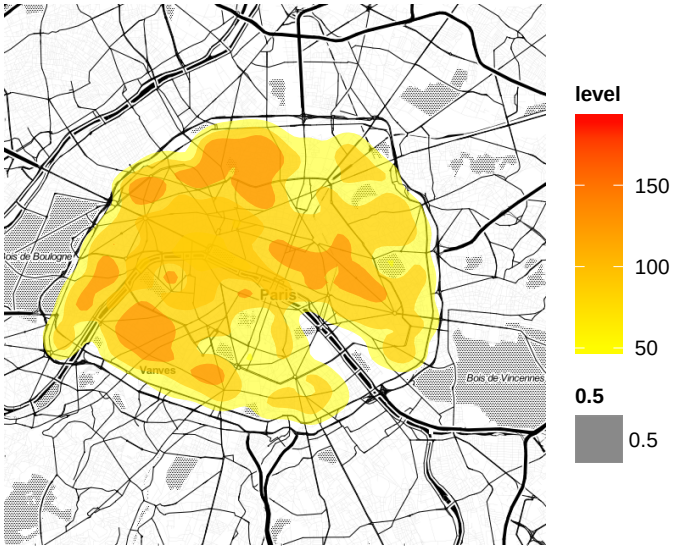


FIGURE 2.12 Répartition des contrats à Paris

Conclusion La répartition spatiale des sinistres ne peut donc être expliquée totalement par l'exposition du portfeuille. Il y a des zones plus vulnérables que d'autres.

Chapitre 3

Analyses et confrontation des intuitions

3.1 Odonymie

Nous avons vu au chapitre précédent que la configuration spatiale du lieu du risque joue un rôle sur sa vulnérabilité. Nous avions émis l'hypothèse que le type de voie où se trouve le risque est déjà un indice sur son *isolement*, qui caractérise sa vulnérabilité au cambriolage. Nous testons cette hypothèse dans cette section.

3.1.1 Récupération des données

A partir des adresses des risques, il s'agit de récupérer le mot qui correspond au type de voie. Pour cela, les étapes sont les suivantes :

— Récupérer la chaîne de caractères de l'adresse
— Enlever le numéro de rue qui se trouve en début de chaîne
— Reconnaître le mot en début de chaîne, par exemple 'BVD' pour 'Boulevard'

3.1.2 taux de sinistres selon le type de voie

Il ne reste alors qu'a compter le nombre de sinistres déclarés selon le type de voie et calculer les taux de sinistres associés.
Nous obtenons le tableau 3.1 et la figure 3.1. Nous ne donnons ici que les types de voie pour lesquels nous avons au moins 50 sinistres.

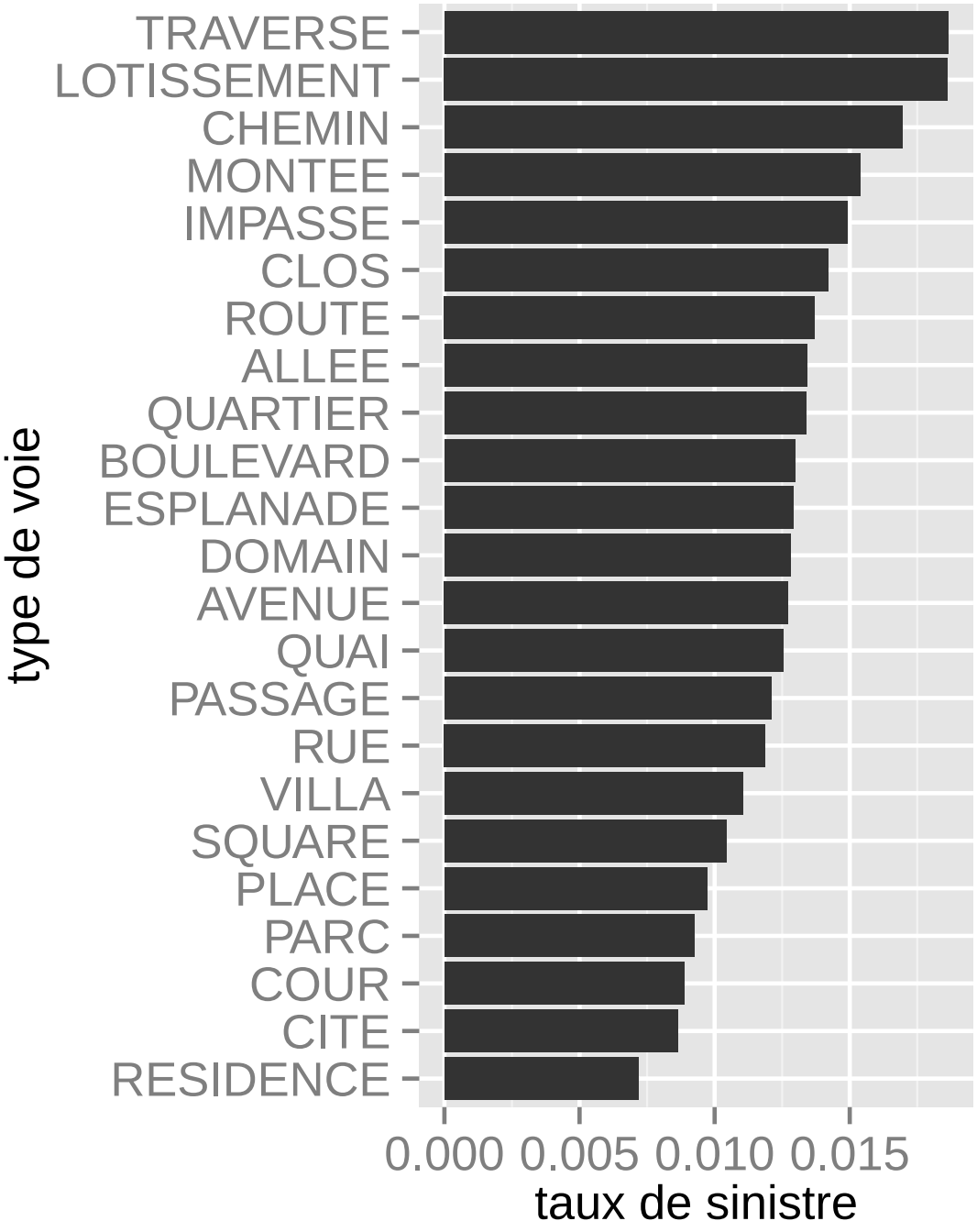


FIGURE 3.1 taux de cambriolage selon le type de voie de l'adresse

voie	ntot	nsin	taux
ALLEE	60233	809	0.01343
AVENUE	202568	2577	0.01272
BOULEVARD	114035	1481	0.01299
CHEMIN	41685	706	0.01694
CITE	6593	57	0.00865
IMPASSE	16777	250	0.01490
LOTISSEMENT	6765	126	0.01863
PASSAGE	6866	83	0.01209
PLACE	23073	224	0.00971
QUAI	7493	94	0.01255
QUARTIER	8676	116	0.01337
RESIDENCE	7252	52	0.00717
ROUTE	15467	212	0.01371
RUE	816312	9691	0.01187
SQUARE	15155	158	0.01043
TRAVERSE	10032	187	0.01864
VILLA	9512	105	0.01104

TABLE 3.1 nombre de risques, nombres de sinistres et taux correspondant selon le type de voie ; données 2008-2013

Il semble que le type de voie joue un rôle et que le nom récupéré dans l'adresse donne déjà une information importante puisqu'on a une distribution de sinistres non uniforme.

Validation statistique En supposant que l'échantillon ci-dessus provient des réalisations de lois binomiales, on détermine un intervalle de confiance.
Pour chaque type de voie, le nombre de sinistres est la réalisation d'une loi binomiale $\mathcal{B}(N,p)$ dont le paramètre p est estimé par : $\hat{p} = \frac{nsin}{ntot}$: le nombre de sinistres sur le nombre de risques de la voie.

Comme Np est grand, on peut utiliser l'approximation normale :

$$\mathcal{B}(N,p) \simeq \mathcal{N}(Np, Np(1-p))$$

Ainsi, les intervalles de confiance de niveau α sont obtenus par :

$$\left[\hat{p} - q_{\alpha/2}\sqrt{N\hat{p}(1-\hat{p})}, \hat{p} + q_{\alpha/2}\sqrt{N\hat{p}(1-\hat{p})}\right]$$

Pour les types de voies pour lesquelles on a suffisamment de données, on peut obtenir un intervalle de confiance raisonnable à 95% (figure 3.2).

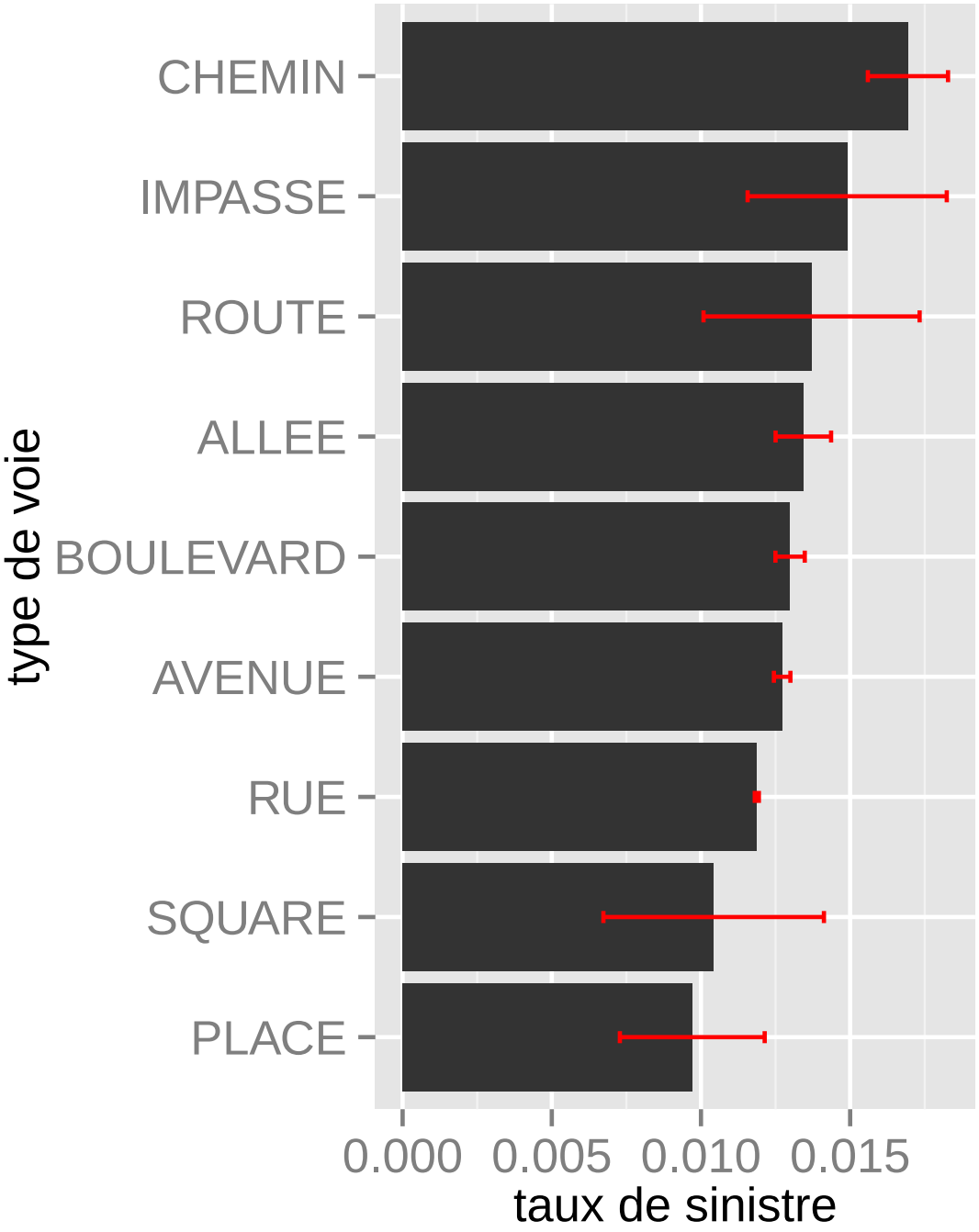


FIGURE 3.2 taux de cambriolage selon le type de voie de l'adresse

Pour améliorer l'intervalle de confiance, on peut décider d'aggréger certaines voies pour former des groupes. Par exemple, il semble que les squares et les places peuvent être regroupées, de même pour les avenues et les boulevards : même physionomie et sinistralités similaires. Nous obtenons l'histogramme en figure 3.3, avec des intervalles de confiances

Les risques situés dans les résidences ou des cités sont moins touchés. Cela peut s'expliquer par le fait qu'elles sont situées en général dans des zones gardées. Les risques situés au niveau d'une place ou d'un square sont beaucoup plus visibles donc protégées et ont aussi un taux de sinistres moindre.

A l'inverse, les risques situés dans des lotissements, des traverses ou des chemins, qui sont des noms qui font référence à un milieu plus rural, sont plus touchés.

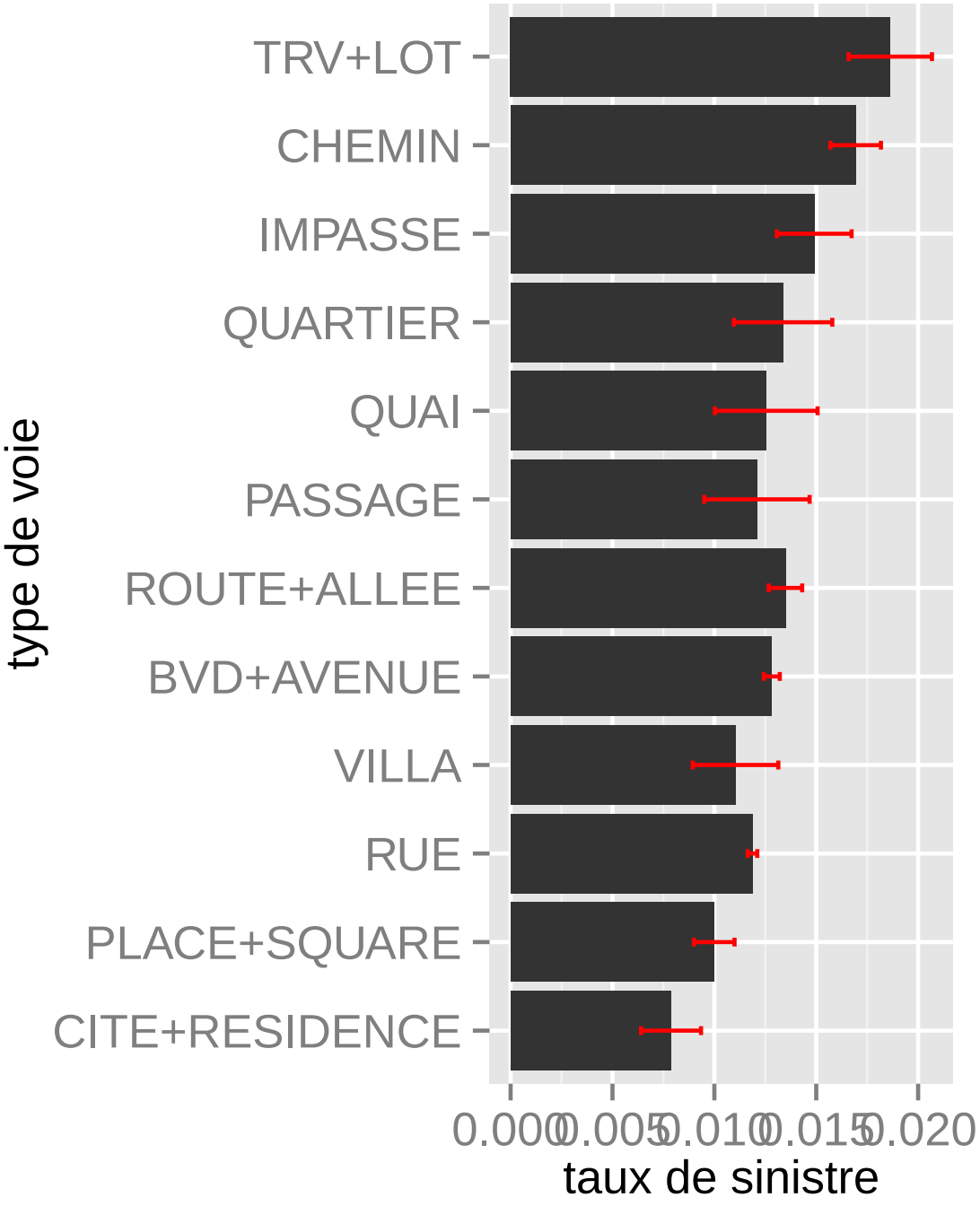


FIGURE 3.3 taux de cambriolage selon le type de voie de l'adresse

3.2 Analyse de l'influence des points d'intérêts

Par point d'intérêt, on entend l'emplacement d'un commerce, d'une caméra de surveillance ou de tout autre entité qu'il peut être intéressant d'étudier pour expliquer le taux de cambriolage

Méthode Haversine de calcul d'une distance entre deux points à la surface de la Terre L'étude nécessite de calculer la distance entre deux points connaissant leurs coordonnées géographiques (longitude et latitude).

On note $R \approx 6378137m$ le rayon de la Terre.

Soit M_1 et M_2 deux points de coordonnées géographiques respectives en (longitude,latitude) : (λ_1, ϕ_1) et (λ_2, ϕ_2).

Alors la distance entre M_1 et M_2 vaut :

$$d(M_1, M_2) = 2R \arcsin\left(\sqrt{\sin^2\left(\frac{\phi_2 - \phi_1}{2}\right) + \cos(\phi_1)\cos(\phi_2)\sin^2\left(\frac{\lambda_2 - \lambda_1}{2}\right)}\right)$$

On utilisera cette distance par la suite.

Etude envisageable/méthodologie La première étape de l'étude consiste à récupérer sur OpenStreetMap les coordonnées GPS des points d'intérêts. Ceux-ci sont par exemple les emplacements de :

— caméras de surveillance
— commissariats et gendarmeries
— boulangeries
— vétérinaires
— lieux de culte
— supermarchés
— entrées de métro

Il est possible de récupérer les informations d'OpenStreetMap à l'adresse http://download.geofabrik.de/. Les données sont au format *Shapefile* (.shp).

3.2.1 Taux de cambriolage en fonction de la distance à un point d'intérêt

Le but de cette méthode est de voir si la distance minimale à une classe de points d'intérêt influe sur le taux de cambriolage.

Au niveau des données, on dispose des coordonnées géographiques de tous les contrats et donc de tous les sinsitres de la base. De même, on dispose des coordonnées géographiques de chaque élémént d'une classe de points d'intérêt.

Pour un contrat donné, on calcule la distance (via la méthode Haversine) avec chaque élément de la classe. Le minimum de ces distances constitue la distance à la classe.

Parmi ces contrats, certains ont été sinistrés. On dispose donc à la fois d'une liste de distance de contrats à une classe de points d'intérêt et d'une liste de distance de sinistres à une classe de points d'intérêt.

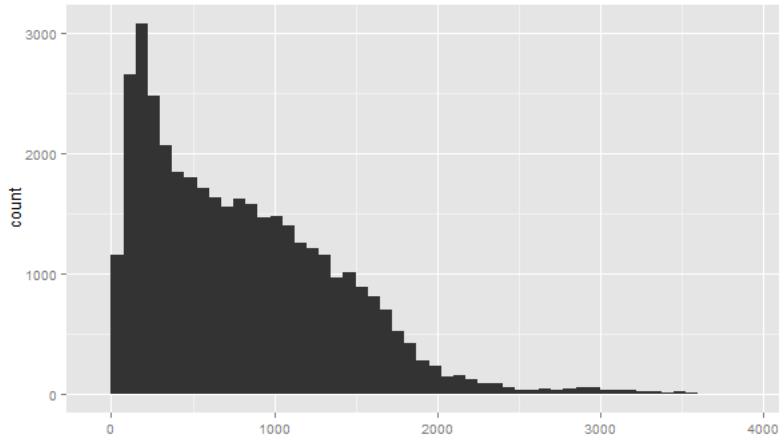


FIGURE 3.4 Distribution de la distance des contrats à la classe "boulangerie"

On ne peut tirer des informations seulement à partir du nombre de sinistres à une distance donnée d'une classe de points d'intérêt. En effet, prenons par exemple le classe "boulangerie". Chaque ville possède au moins une boulangerie. En zone urbaine, un assuré n'est donc jamais très éloigné d'un de ces commerces. Bien sûr, peu de contrats et donc peu de sinistres seront à une distance très petite d'une boulangerie car celles-ci ne sont pas non plus omniprésentes. Passé ce seuil, plus la distance augmente et plus le nombre de sinistres/contrats diminue car la probabilité d'être à une distance d d'une boulangerie diminue quand d augmente.

Par contre, ramener le nombre de sinistres à une distance donnée d'une classe au

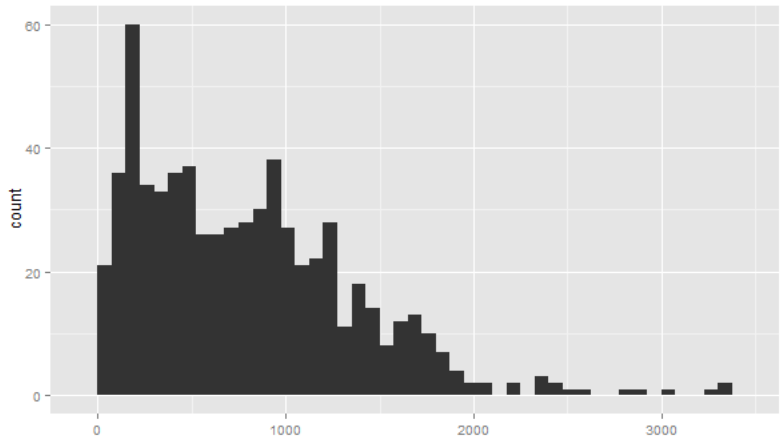


FIGURE 3.5 Distribution de la distance des sinistres à la classe "boulangerie"

nombre de contrats à une distance donnée d'une classe permet d'obtenir un taux de cambriolage empirique.

On va donc calculer le taux de cambriolage empirique à une distance x d'une classe donnée par la formule suivante :

$$\tau_{\text{cambriolage}} = \frac{\text{Nombre de sinistres entre } x \text{ et } x+D}{\text{Nombre de contrats entre } x \text{ et } x+D}$$

où D est un paramètre en mètres.

Cette formule permet de calculer un taux de cambriolage sur les contrats situé entre une distance x et une distance $x + D$ d'une classe donnée.

On choisira D ni trop petit pour éviter d'avoir un taux trop discontinu ni trop grand pour éviter d'avoir un taux trop lissé et donc une perte d'information.

On réalisera l'étude sur le portefeuille de contrats de la Seine-Saint-Denis (93).

Parmi les différentes classes testées, la classe "bureau de poste" donne des résultats intéressants. On trace les résultats pour les années 2010 à 2013 pour un coefficient D égal à $70m$.

On voit que le taux de cambriolage semble augmenter quand on s'éloigne des bureaux de poste. Le résultat semble se retrouver sur les 4 années testées, ce qui prouve la robustesse de l'étude, au moins au cours du temps. On peut interpréter ce résultat en disant que les bureaux de poste se trouvent souvent en centre-ville, dans des zones où il est sans doute compliqué pour un cambrioleur de commettre un méfait à cause de la surveillance mutuelle entre les habitants. Plus on s'éloigne du centre-ville et moins cette surveillance mutuelle existe d'où un taux de cambriolage qui augmente.

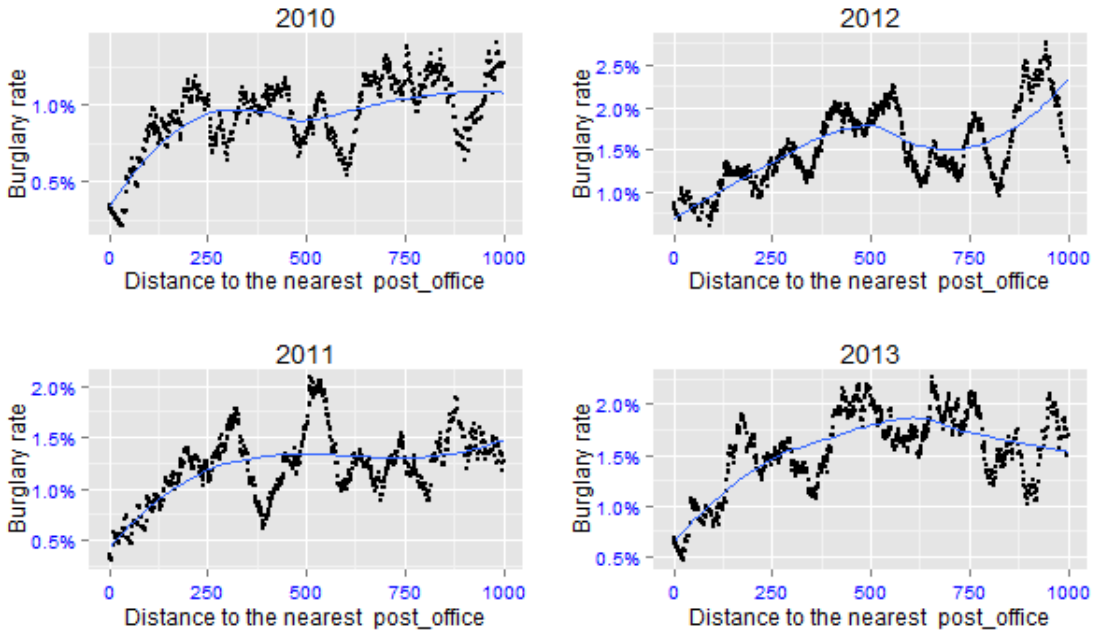


FIGURE 3.6 Taux de cambriolage empirique en fonction de la distance à la classe "bureau de poste"

3.2.2 Taux de cambriolage à l'emplacement d'un contrat

Le but est de créer un taux de cambriolage empirique en un point de coordonnées données. Cela permettra d'alimenter la régression que l'on fera par la suite.

A partir de ce point, on transforme les coordonnées géographiques données en longitude et latitude par leur projection en coordonnées Lambert 93.

Pour estimer un taux de cambriolage, on va se baser sur une approche des plus proches voisins.

Soit M un point où se trouve un contrat. On calcule alors quelles sont les k plus proches contrats $P_1, \cdots, P_k$ par un algorithme de recherche des plus proches voisins.

On définit alors l'estimateur du taux de cambriolage au point M par :

$$\hat{\tau}_{\text{cambriolage}}(M) = \frac{1}{k+1}\left[\mathbb{1}(M=1) + \sum_{i=1}^{k} \mathbb{1}(P_i = 1)\right]$$

avec $P_i = 1$ (resp. $M = 1$) si le contrat P_i (respectivement M) est sinistré.

Taux de cambriolage On obtient par la méthode des k-plus proches voisins le taux de cambriolage pour Paris ($k = 250$) :

FIGURE 3.7 Taux de cambriolage, Ville de Paris, $k = 250$

Où les couleurs du jaune au rouge représentent les taux de cambriolage des moins élevés aux plus élevés.

Sélection des variables de distance à un point d'intérêt par ACP On calcule pour chaque contrat la distance à chaque classe de points d'intérêt.

Utilisons la méthode d'Analyse en Composantes Principales pour déterminer quelles variables garder ou pas dans la régression à venir.

Les variables de distance à un points d'intérêt (82 en tout) sont du type :

— bureaux de poste
— commissariats de police
— boulangeries
— vétérinaires
— caméras de surveillance
— pharmacies
— ...

Rappel : La part de l'inertie expliquée par les k premières composantes principales est :

$$\frac{\lambda_1 + \cdots + \lambda_k}{\mathrm{Tr}(\Sigma)}$$

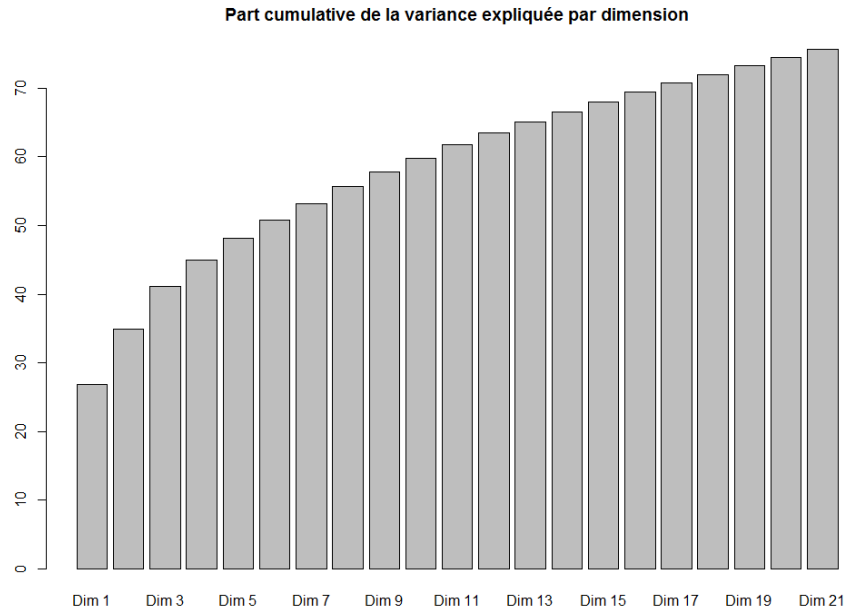


FIGURE 3.8 Analyse en Composantes Principales des densités de points d'intérêt

avec Σ la matrice de covariance de la matrice de données $X = (x_1, \cdots, x_n)$ de données $(n \times p)$ ($p = 82$ ici) et $(\lambda_1, \cdots, \lambda_k)$ les k valeurs propres associées aux k premières composantes principales.

On voit que garder 21 dimensions permet d'expliquer 75% du nuage de points, ce qui nous montre que l'on peut garder 21 variables sur les 82 disponibles, en gardant un pouvoir explicatif important.

On élimine d'office les variables dont on ne pourra expliquer l'impact sur le cambriolage. Les variables explicatives suivantes sont choisies parmi celles ayant la plus grande contribution à la variance :

— ralentisseurs (cassis, dos d'âne)
— stations-services
— vétérinaires
— points de vente d'alcool
— coiffeurs
— fleuristes

— cafés
— bouchers
— location de voiture
— bureaux de poste
— supermarchés
— kiosques à journaux
— bibliothèques
— bars
— banques
— bijouteries
— radars de vitesse
— entrées de parking
— commissariats de police

Remarques L'ACP est une méthode basée sur les statistiques d'ordre 2 du nuage de points. Le modèle sous-entendu est donc gaussien. Comme la régression linéaire, l'ACP est très sensible aux valeurs aberrantes ("outliers") donc on pourra affiner la sélection en utilisant d'autres méthodes comme l'Analyse Factorielle des Données Multiples. On n'utilise pas l'ACP pour la suite.

Gradient boosting

Théorie

L'algorithme du Gradient Boosting[8] répond au problème d'estimation d'une fonction donnant une variable cible y en fonction de $\mathbf{x} = (x_1, \cdots, x_n)$. A partir d'une base d'apprentissage de valeurs $(y, \mathbf{x}_i)^N$ connues, on cherche une fonction $\mathrm{F}^*(\mathbf{x})$ donnant y en fonction de $\mathbf{x}$ telle que, sur la base d'apprentissage, l'espérance d'une fonction de perte $\Psi(y, \mathrm{F}(\mathbf{x}))$ est minimale :

$$\mathrm{F}^*(\mathbf{x}) = \mathrm{argmin}_{F(\mathbf{x})} \mathbb{E}_{y,\mathbf{x}} \Psi(y, \mathrm{F}(\mathbf{x}))$$

Le boosting approche $\mathrm{F}^*(\mathbf{x})$ par une méthode additive de la forme :

$$\mathrm{F}(\mathbf{x}) = \sum_{m=0}^{M} \beta_m h(\mathbf{x}, \mathbf{a}_m)$$

avec $h(\mathbf{x}, \mathbf{a})$ des fonctions dites "base-learner" choisies comme simples fonction de $\mathbf{x}$ et $\mathbf{a} = (a_1, a_2, \cdots)$. On choisit alors les $(\mathbf{a}_m)_{0 \leqslant m \leqslant M}$ et $(\beta_m)_{0 \leqslant m \leqslant M}$ par itérations successives. On part d'un premier estimateur $F_0(\mathbf{x})$, et on poursuit ensuite pour $m = 1, 2, \cdots, M$

$$(\beta_m, \mathbf{a}_m) = \text{argmin}_{\beta, \mathbf{a}} \sum_{i=1}^{N} \Psi\left(y_i, \mathrm{F}_{m-1}(\mathbf{x}_i) + \beta h(\mathbf{x}_i, \mathbf{a})\right)$$

et

$$\mathrm{F}_m(\mathbf{x}) = \mathrm{F}_{m-1}(\mathbf{x}) + \beta_m h(\mathbf{x}, \mathbf{a}_m)$$

Cette résolution pouvant être très complexe, la méthode du Gradient Boosting introduit l'approximation suivante pour une fonction de perte donnée $\Psi(y, \mathrm{F}(\mathbf{x}))$. Cette approximation passe par deux étapes. On trouve d'abord la fonction $h(\mathbf{x}, \mathbf{a}_m)$ par la méthode des moindres carrés

$$\mathbf{a}_m = \text{argmin}_{\mathbf{a}, \rho} \sum_{i=1}^{N} \left[\overline{y}_{i,m} - \rho h(\mathbf{x}_i, \mathbf{a})\right]^2$$

appliquées aux pseudo-résidus :

$$\overline{y}_{i,m} = -\left[\frac{\partial \Psi(y_i, \mathrm{F}(\mathbf{x}_i))}{\partial \mathrm{F}(\mathbf{x}_i)}\right]_{\mathrm{F}(\mathbf{x}) = \mathrm{F}_{m-1}(\mathbf{x})}$$

Ensuite, connaissant $h(\mathbf{x}, \mathbf{a}_m)$, on détermine la valeur optimale de β_m par :

$$\beta_m = \text{argmin}_{\beta} \sum_{i=1}^{N} \Psi\left(y_i, \mathrm{F}_{m-1}(\mathbf{x}_i) + \beta h(\mathbf{x}_i, \mathbf{a}_m)\right)$$

Cette approximation remplace un problème d'optimisation difficile par deux problèmes, l'un basé sur un critère des moindres carrés, l'autre étant un simple problème d'optimisation dépendant de la fonction Ψ.

Le Gradient tree boosting est une Gradient boosting dans lequel $h(\mathbf{x}, \mathbf{a})$ est le noeud L-terminal d'un arbre de régression. A chaque itération m, un arbre de régression découpe le $\mathbf{x}$-espace en L régions $(R_{lm})_{1 \leqslant l \leqslant L}$ disjointes et prédit une valeur constante dans chacune de ces régions :

$$h\left(\mathbf{x}, (R_{lm})_{1\leqslant l\leqslant L}\right) = \sum_{l=1}^{L} \overline{y}_{lm} \mathbb{1}\left(\mathbf{x} \in R_{lm}\right)$$

Ici $\overline{y}_{lm} = \text{Moyenne}_{\mathbf{x}_i \in R_{lm}}\left(\overline{y}_{im}\right)$. Comme l'arbre prédit une valeur constante $\overline{y}_{lm}$ sur chaque région R_{lm}, la solution à l'équation précédente permettant de trouver β_m se réduit à :

$$\gamma_{lm} = \text{argmin}_{\gamma} \sum_{\mathbf{x}_i \in R_{lm}} \Psi\left(y_i, \mathrm{F}_{m-1}\left(\mathbf{x}_i + \gamma\right)\right)$$

On met alors à jour l'approximation $\mathrm{F}_{m-1}\left(\mathbf{x}\right)$ dans chaque région :

$$\mathrm{F}_m\left(\mathbf{x}\right) = \mathrm{F}_{m-1}\left(\mathbf{x}\right) + \nu \cdot \gamma_{lm} \mathbb{1}\left(\mathbf{x} \in R_{lm}\right)$$

Le paramètre de "shrinkage" $0 < \nu \leqslant 1$ contrôle le taux d'apprentissage de la procédure. Il a été montré empiriquement que des petites valeurs ($\nu \leqslant 0.1$) permettent un meilleur apprentissage.

Algorithm 1 Gradient TreeBoost

$\mathrm{F}_0\left(\mathbf{x}\right) = \text{argmin}_{\gamma} \sum_{i=1}^{N} \Psi\left(y_i, \gamma\right)$
for $m = 1$ to M do **do**
  $\overline{y}_{i,m} = -\left[\frac{\partial \Psi(y_i, \mathrm{F}(\mathbf{x}_i))}{\partial \mathrm{F}(\mathbf{x}_i)}\right]_{\mathrm{F}(\mathbf{x}) = \mathrm{F}_{m-1}(\mathbf{x})}, i = 1, \cdots, N$
  $(R_{lm})_{1\leqslant l\leqslant L} =$ L-terminal noeud d'un arbre $\left(\overline{y}_{im}, \mathbf{x}_i\right)_1^N$
  $\gamma_{lm} = \text{argmin}_{\gamma} \sum_{\mathbf{x}_i \in R_{lm}} \Psi\left(y_i, \mathrm{F}_{m-1}\left(\mathbf{x}_i + \gamma\right)\right)$
  $\mathrm{F}_m\left(\mathbf{x}\right) = \mathrm{F}_{m-1}\left(\mathbf{x}\right) + \nu \cdot \gamma_{lm} \mathbb{1}\left(\mathbf{x} \in R_{lm}\right)$
end for

Sous R, on utilise le package *gbm* [17] pour appliquer la méthode du gradient tree boosting.

Analyse On utilise la méthode de gradient boosting par arbres pour prédire le taux de sinistres calculé avec la méthode des k plus proches voisins pour $k = 250$. On garde les types de point d'intérêt suivants pour la méthode de gradient boosting. Ces types de points d'intérêt sont renseignés de manière assez exhaustive dans OpenStreetMap. De plus, on élimine les types de points d'intérêt qui ne sont pas commun à

Paris, comme les châteaux d'eau par exemple, et qui ne pourraient donc donner une quelconque information sur les sinistres.

Voici les 54 variables conservées :

— commissariats de police
— pressings
— opéras, théatres
— oeuvres d'art en extérieur (statues, etc $\cdots$)
— cinémas
— kiosques
— salons de beauté
— bijouteries
— vente de meubles
— vente de chaussures
— distributeurs (de nourriture, tickets, etc $\ldots$)
— pubs
— crèches
— écoles
— poubelles pour recyclage
— informations pour touristes
— points d'attente pour taxis
— opticiens
— cabines téléphoniques
— coiffeurs
— agences immobilières
— stations d'essence
— agences de voyage
— fontaines
— lieux de culte
— supermarchés
— location de voitures
— vétérinaires
— toilettes
— vente de cadeaux
— marchés (aux puces, hebdomadaires, etc $\cdots$)
— épiceries

— vente d'alcool
— entrée de métro
— boutiques de vêtements
— bibliothèques
— bureaux de poste
— librairies
— bouchers
— hôtels
— distributeurs de billets
— restaurants
— pharmacies
— boîtes aux lettres de la Poste
— fleuristes
— banques
— restaurants "fast-food"
— distributeurs d'eau (fontaines)
— boulangeries
— cafés

Traitement des valeurs aberrantes On utilise la base de données OpenStreetMap des points d'intérêt de Paris. Pour détecter les contrats qui ne sont pas situés à Paris (à cause d'un mauvais géocoding ou car l'adresse du contrat n'était pas en région parisienne), on regarde donc la distance des contrats à un point d'intérêt quelconque (les écoles par exemple). Si cette distance est supérieure à 10km, le contrat n'est pas géocodé en région parisienne.

Performances du boosting On lance le boosting. Voici les performances de cet apprentissage :

Ce graphe de performance est tracé via la fonction **gbm.perf**(., method="OOB"). On utilisera par la suite un gradient boosting sur 200 arbres afin d'éviter un sur-apprentissage.

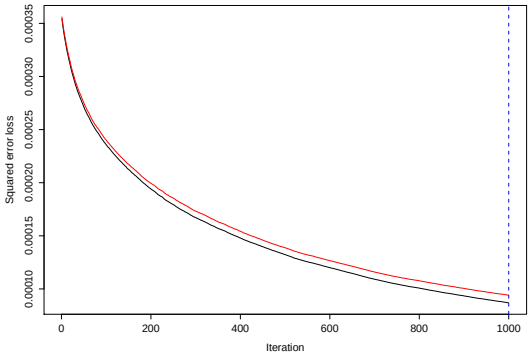


FIGURE 3.9 Performances du Gradient Boosting

Influence relative des variables explicatives

On trace ici l'influence relative des différentes fonctions, soit le nombre de fois (en proportion) où chaque variable est utilisée dans le découpage des arbres du boosting.

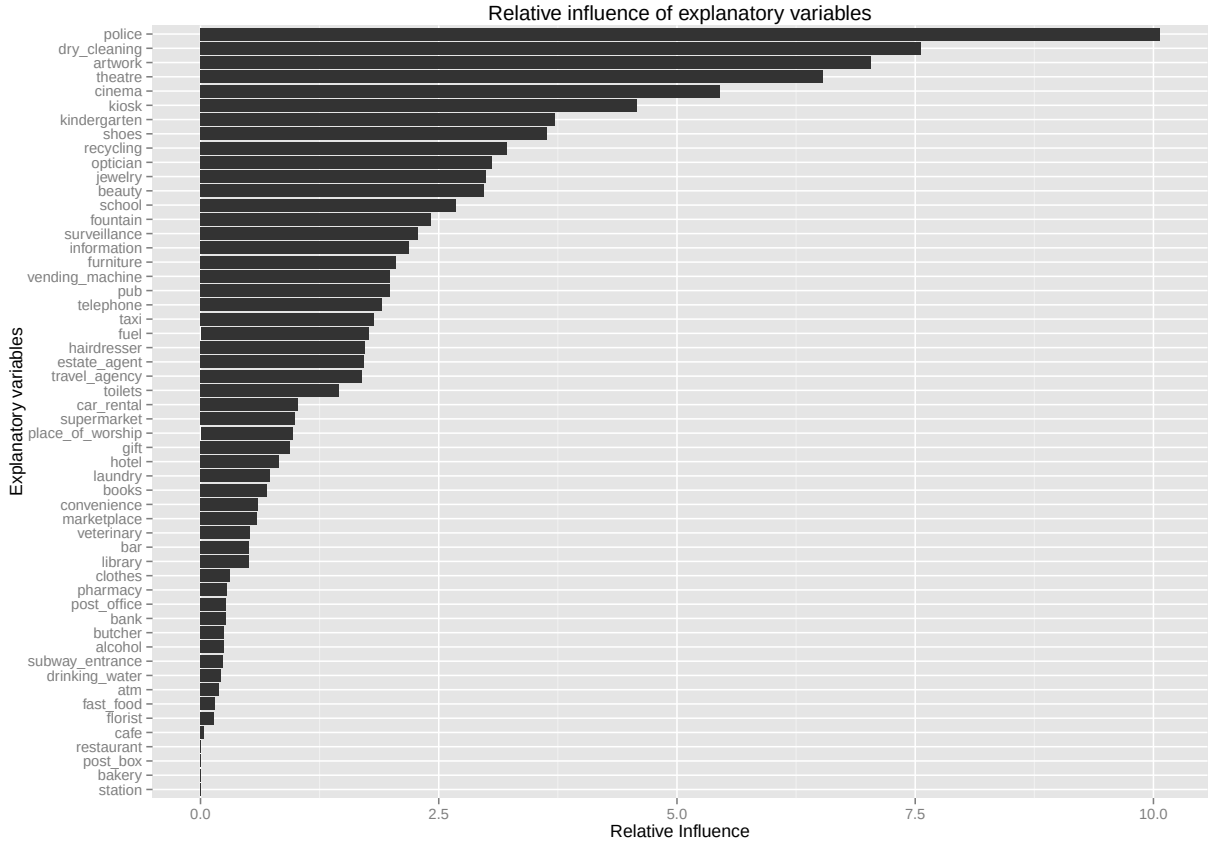


FIGURE 3.10 Gradient Boosting : Influence relative des types de points d'intérêts

On voit tout d'abord que la variable d'intérêt police est la plus utilisée. Cela montre que la présence d'un commissariat a un effet certain sur le taux de cambriolage, puisse-t-il être positif ou négatif. Pour les autres variables, il est assez difficile d'expliquer pourquoi certaines sont plus représentées que les autres. On note toutefois que la variable post_office analysée à la section précédente est ici peu représentée.

Dépendance partielle aux variables explicatives

Analysons maintenant la façon dont varie le taux de cambriolage en fonction de chaque variable prise séparément. Vu le nombre de variables, on se contentera d'afficher les 15 variables les plus représentées.

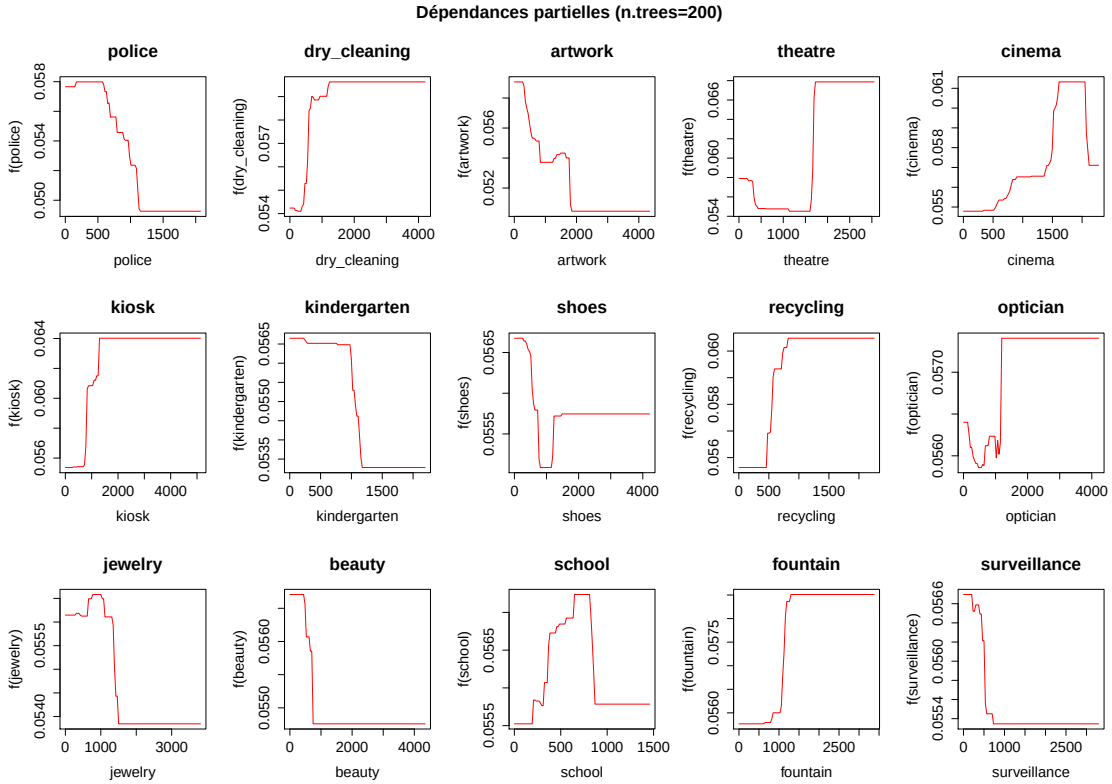


FIGURE 3.11 Gradient Boosting : dépendance partielle aux types de point d'intérêt

On remarque notamment que le taux de cambriolage semble décroître en fonction de la distance à un commissariat de police. C'est à priori contre-intuitif car on pourrait penser que les cambrioleurs seraient découragés de s'attaquer à un logement proche d'un commissariat. Cependant, si on regarde le graphe en question de plus près, on

se rend compte que le taux de cambriolage remonte légèrement pour une distance inférieure à 100 mètres. Analysons ce graphe de plus près. On choisit de pousser à 500 le nombre d'arbre pour avoir un découpage plus fin de la courbe de dépendance partielle.

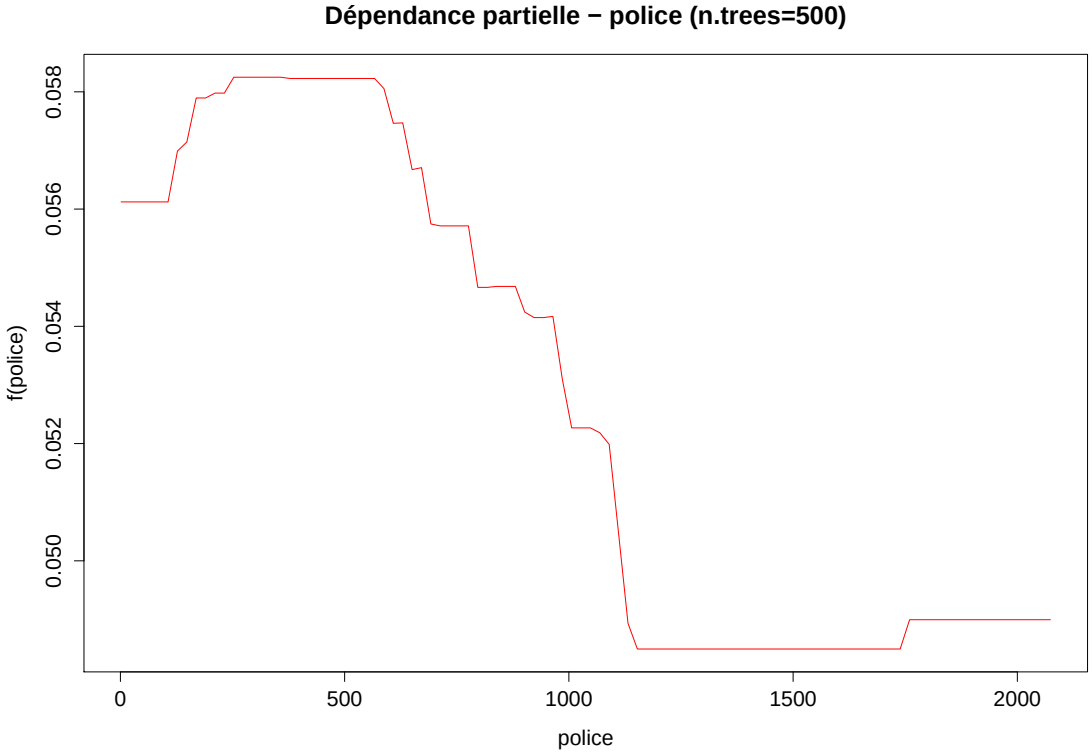


FIGURE 3.12 Dépendance partielle : distance à un commissariat de police

On voit mieux le phénomène de croissance du taux de cambriolage à une petite distance du commissariat. On peut tout d'abord supposer qu'il existe une distance d'influence en dessous de laquelle avoir un commissariat proche a une importance. Il nous semble voir cela sur le graphe même s'il faut toujours rester prudent.

3.3 Facteurs socio-économiques

Mettons maintenant en évidence l'influence de facteurs socio-économiques sur le taux de sinistres local.

Obtention des données socio-économiques On associe chaque contrat à l'IRIS (voir plus haut, Ilôt de regroupement pour l'Information Statistique) le plus proche. A partir de là, on met en relation les données extraites du site de l'INSEE avec les données de contrat. On a ainsi par contrat les informations suivantes :

- Revenu par ménage local - Quartile 1
- Revenu par ménage local - Quartile 9
- Revenu par ménage local - Médiane
- Revenu par ménage local - Moyenne
- Revenu par ménage local - Décile 1
- Revenu par ménage local - Décile 9
- Taux de population homme
- Taux d'actifs parmi les hommes
- Taux d'actifs en intérim
- Taux de cadres
- Taux d'ouvrier
- Différence $9^{\text{ème}}$ décile - 1^{er} décile - Revenu par ménage (variable calculée)

Traitement des données manquantes Les données de l'INSEE ayant tendance à présenter des valeurs non disponibles, on remplace systématiquement la valeur manquante par la moyenne de la variable dans l'échantillon afin d'avoir le biais le plus faible possible.

Etude On cherche à établir un lien entre le taux de cambriolage et des indicateurs sociaux économiques.

Supposons que l'on calcule le taux de cambriolage via la méthode des k plus proches voisins. En reprenant les notations précédentes, on a pour tout point M et $P_1, \cdots, P_k$ ses k plus proches voisins :

$$\hat{\tau}_{\text{cambriolage}}(M) = \frac{1}{k} \sum_{i=1}^{k} \mathbb{1}(P_i = 1)$$

Notions $I_1, \cdots, I_P$ chacun des centroïdes des zones IRIS de Paris. On a, pour

chaque zone IRIS, les données socio-économiques $\mathbf{x} = (x_1, x_2, \cdots)$ correspondantes grâce à la base Mosaïc que l'on a recupéré en interne chez AXA.

Soit $p \in [\![1, P]\!]$. On dispose via la formule précédente de $\hat{\tau}_{\text{cambriolage}}(I_p)$. On va donc étudier l'effet des variables explicatives socio-économiques $\mathbf{x}^p = (x_1^p, x_2^p, \cdots)$ sur $\hat{\tau}_{\text{cambriolage}}(I_p)$.

3.4 Analyse d'un facteur conjoncturel : les conditions météorologiques

Nous avons vu au premier chapitre que la température locale et sa variation temporelle nous paraît une variable explicative intéressante du risque de cambriolage. Il nous a été possible de récupérer des données météorologiques locales et temporelles. Nous analysons ce que ces données peuvent nous dire quant à l'impact des conditions météorologiques sur le cambriolage.

3.4.1 Données

Les données de 9 stations météorologiques pour l'île de France et celle d'une station pour Marseille ont été récupérées [1]. Nous donnons leurs noms et localisations dans le tableau 3.2.

nom	longitude	latitude	altitude	département
montsouris	2.34	48.82	75	75
villacoublay	2.21	48.77	174	78
toussus	2.11	48.74	154	78
orly	2.38	48.72	89	94
trappes	2.01	48.77	167	78
melun	2.68	48.61	91	77
roissy	2.53	49.02	108	95
lebourget	2.43	48.97	52	93
bretigny	2.32	48.60	80	91
marignane	5.23	43.44	5	13

TABLE 3.2 Stations météorologiques et leurs localisations

Comme nous disposons des données de contrats d'AXA pour les départements 13, 75, 92, et 93, nous avons uniquement utilisés les données des stations montsouris, lebourget et marignane.

Les données consistent en des relevés journaliers de températures et de précipitations. Un exemple est donné dans le tableau 3.3.

date	min	max	moy	etn	etx	etm	prec
2008-01-01	3.6	6.5	5.1	2	0.5	1.3	0
2008-01-02	0.1	3.6	1.9	-2.2	-2.8	-2.5	0
2008-01-03	-1.4	5	1.8	-3.4	-1.1	-2.3	0.2
2008-01-04	4.2	10	7.3	2	4	3.1	0.6
2008-01-05	6.2	9.5	7.9	3.9	2.7	3.3	7.4
2008-01-06	5.5	9.6	7.6	2.9	2.7	2.8	0.2

TABLE 3.3 Exemple de relevés pour la station montsouris

3.4.2 Première analyse sur la température

On fait une première analyse grossière pour avoir une idée du niveau d'influence de la température sur le nombre de cambriolages. Pour cela, on va utiliser la variation temporelle de la température.

Hypothèse : le risque de cambriolage est positivement corrélé à la température locale.

Principe de l'analyse Les étapes de la méthode sont les suivantes :

- On choisit une zone géographique G autour d'une station météorologique donnée, pour laquelle on a à disposition un grand nombre d'adresses géocodées.
- On choisit plusieurs plages temporelles similaires (par exemple P_1 : printemps 2010, P_2 : printemps 2011, P_3 : printemps 2012 ...)
- On calcule la température moyenne T_i de la période P_i
- On dispose du nombre n_i de sinistres survenus pendant la période P_i ainsi que le nombre d'adresses total m_i dans la zone sélectionnée G.
- On calcule le pourcentage de sinistres R_i dans la zone géographique sélectionnée aux différentes périodes. - On explique R_i avec la variable T_i.

Avantages a priori de la méthode :

- On évite la variation saisonnière de la criminalité pour étudier seulement l'influence de la variation de la température
- On évite la variation géographique de la criminalité, qui peut sans doute être semble davantage expliquée par des critères socio-économiques
- On minimise l'incertitude que l'on a sur la date de survenance et sur les températures

Limites :

— On a à disposition 7 années de sinistres depuis 2008, donc on aurait au plus 7 observations R_i

— On moyenne sur un grand nombre d'observations, ce qui a pour effet de lisser fortement les variations.

Résultats :

Les résultats obtenus pour le 93 (station météo : lebourget) avec cette méthode sont données dans le tableau 3.4.2
Période : 01/03 au 31/05 de chaque année (2008 à 2014)

annee	temp. moyenne	nb sinistres
2008	11.4	221
2009	11.8	234
2010	10.8	244
2011	13.1	286
2012	11.8	330
2013	9.2	307
2014	12.0	262

Nous avons calculé la température moyenne de la période en prenant la moyenne des températures moyennes des relevés du jour au cours de la période. On voit que cette analyse très grossière ne permet pas de mettre en évidence une corrélation positive entre la température et le nombre de cambriolages.

On est conscient ici que la température a une influence modérée sur le risque de cambriolage, que la seule donnée d'une valeur moyenne sur une période ne permet pas de capturer.

On propose maintenant de regarder le nombre de sinistres par niveau de température.

3.4.3 Deuxième analyse sur la température

Cette deuxième analyse va permettre d'utiliser plus d'information sur les données pour essayer d'établir un lien entre température du jour et sinistralité.

Principe de l'analyse Les étapes de l'analyse sont les suivantes :

— On regarde pour chaque sinistre, la température moyenne enregistré le jour de survenance du sinistre. On en prend son niveau (par exemple au °C près)

— On compte le nombre de sinistres survenus les jours de même niveau de température

— Ce nombre, ramené au nombre de jours de l'année présentant ce niveau de température, donne un taux de sinistre. On peut comparer les taux de sinistres obtenus pour différents niveaux de températures.

Classification et arbre de regression CART On construit un arbre de régression via la méthode CART [11] pour décrire le nombre de sinistres par jour en fonction de la température moyenne du jour.

Soit $x = (x_1, \ldots, x_p) \in \mathbb{R}^p$ l'ensemble des variables supposées décrire une variable cible Y. Soit ρ une fonction objectif, typiquement $\rho(x, y) = ||x - y||^2$. On cherche à minimiser $\mathbb{E}[\rho(Y, g_{tree}(x))]$ sur les fonction g_{tree} de la forme :

$$g_{tree}(x) = \sum_{i=1}^{M} \beta_r \mathbb{1}_{[x \in R_r]}$$

où $(P) = \{R_1, \ldots, R_M\}$ est une partition de $\mathbb{R}^p$. La fonction g_{tree} est ainsi modélisée constante par morceaux.

L'algorithme pour la recherche de l'arbre optimal est le suivant :

1. On commence avec $M = 1$ et $\mathcal{P} = \{R\} = \{\mathbb{R}^p\}$
2. On partitionne R en $R_{left} \cup R_{right}$ où :

$$R_{left} = \mathbb{R} \times \mathbb{R} \times \cdots \times]-\infty, d] \times \mathbb{R} \times \cdots \times \mathbb{R}$$

$$R_{right} = \mathbb{R} \times \mathbb{R} \times \cdots \times]d, \infty[\times \mathbb{R} \times \cdots \times \mathbb{R}$$

où l'un des axes est coupé au point d, avec d appartenant à l'ensemble fini des points observés. La recherche de l'axe à couper et de d est faite de telle sorte à maximiser la vraisemblance à la suite du découpage.

3. On redécoupe la partition actuelle $\mathcal{P}$ comme à l'étape 2 en découpant l'une des cellules de la partition courante. On cherche donc la meilleure cellule, puis le meilleur axe et la meilleure valeur médiane d comme à l'étape 2. Ensuite on

met à jour la partition :

$$\mathcal{P} = \mathcal{P}_{old} \backslash \{\text{cellule à découper}\} \cup \{\text{cellules découpées } R_{left}, R_{right}\}$$

4. On itère 3, jusqu'à $M = M_{max}$
5. Elagage : On enlève progressivement des branches afin de simplifier l'arbre, en introduisant une pénalisation selon la taille. Ce paramètre est déterminé par validation croisée.

Sous R, on utilise le package *rpart* [20] pour appliquer cet algorithme.

Résultats On obtient, pour l'ensemble des sinistres des départements 75, 91 et 93 sur les années 2008 à 2013 une représentation sous forme d'arbre (figure 3.13)

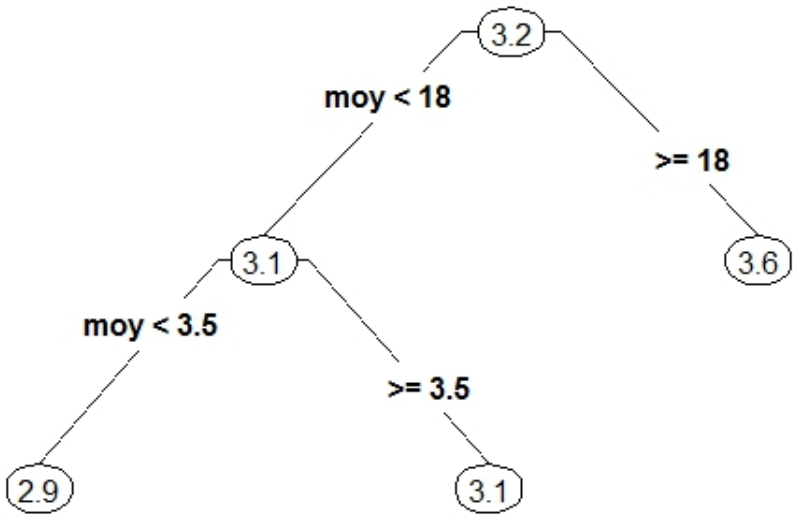


FIGURE 3.13 Variations du nombre de sinistres par jour en fonction de la température moyenne du jour

L'arbre de régression suggère que le taux de sinistres est plus important les jours chauds : 3.6 sinistres par jour en moyenne lorsque la température moyenne du jour est supérieure à 18°C, et plus faible les jours froids : 2.9 sinistres par jour lorsque la température moyenne du jour est inférieure à 3.5°C.

Biais On peut noter les biais suivants dans la méthode et dans les données :

— Tous les jours ne sont pas équivalents : en effet, pour une analyse plus fine, il faudrait considérer les jours ouvrés/non ouvrés (weekend, jour férié, période de vacances scolaires...)
— Les données météo sont disponibles à l'échelle du département seulement
— Les données sur la date de survenance du sinistre sont communiquées par les assurés : c'est plutôt la date constatée du sinistre et non la date de survenance du cambriolage.

3.4.4 Influence des précipitations

S'inspirant de l'analyse précédente pour la température, on regarde maintenant l'influence du mauvais temps (pluie) sur l'occurrence des sinistres. On dispose des données de sinistres pour le 75, 91 et 93 ainsi que les données journalières de précipitations. On construit encore un arbre de régression via la méthode CART (figure 3.14).

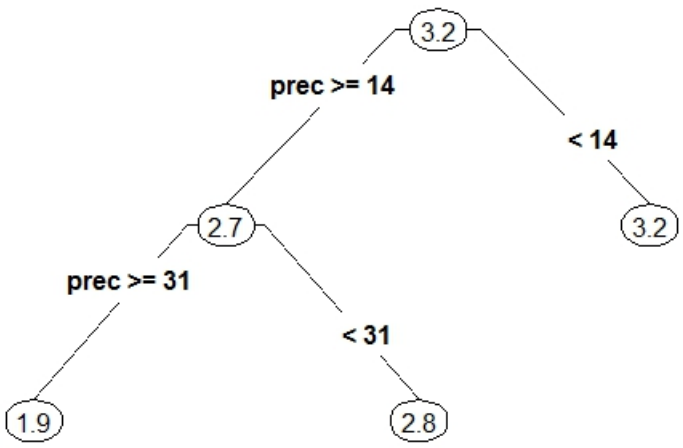


FIGURE 3.14 Arbre de régression : variation du nombre de sinistres par jour en fonction des précipitations du jour

L'arbre de régression suggère que le nombre de sinistres moyen en île de France est 40% plus faible les jours de fortes pluies - précipitations supérieures à 31mm^2 - par rapport aux jours de "beau temps" (précipitations inférieures à 14 mm^2).

Piste d'améliorations Prendre en compte les variations relatives de température sur une courte période. Que se passent-ils quand il fait subitement beau et chaud ?

3.5 Influence des jours fériés

Nous analysons dans cette section l'influence des jours fériés, vacances, weekend en tant que facteur conjoncturel du risque de cambriolage.

Pour cela, nous récupérons le calendrier des vacances scolaires et jours fériés de 2008 à 2013 dans la zone C, qui comprend les départements 75, 92 et 93. Pour chaque sinistre survenu dans la zone C, on peut ainsi décider, à partir de la date de survenance renseignée dans la base de donnée, si :

— le sinistre est survenu pendant les vacances scolaires
— le sinistre est survenu un jour férié
— le sinistre est survenu le weekend

Résultats Nous comptons les effectifs de sinistres selon les 3 critères précédents. Les résultats sont récapitulés dans les tableaux 3.4 et 3.5.

	oui	non
jour férié	69	2123
vacances	725	1467
weekend	626	1566

Table 3.4 Effectifs des jours entre 2008 et 2013

	oui	non
jour férié	257	9741
vacances	3278	6720
weekend	2942	7056

Table 3.5 Effectif des jours de sinistres entre 2008 et 2013

Pour comparer la sinistralité selon ce facteur conjoncturel, il suffit de calculer le rapport entre ces effectifs. Nous obtenons ainsi le nombre de sinistres moyen survenu les jours de weekend, les jours en semaine, les jours fériés, les jours non fériés, les jours de vacances scolaires, et les jours hors vacances scolaires. On obtient le tableau 3.6.

Conclusion Les sinistralités obtenues sont très comparables, hormis peut-être pour le nombre de sinistres les jours fériés, mais le faible nombre de jours fériés (environ 12

var	taux
txWeekend	4.70
txNonWeekend	4.51
txFériés	3.72
txNonFériés	4.59
txVac	4.52
txNonVac	4.58

TABLE 3.6 Moyenne du nombre de sinistres par jour selon le type de jour

par an) ne permet pas de valider statistiquement cet écart.

On peut donc conclure parmi les assertions suivantes :

— Le calendrier n'a pas d'influence conjoncturelle sur le risque de cambriolage

— Les dates de survenance sont renseignées avec un biais tel que l'erreur sur le jour réel du sinistre est bien plus importante qu'un éventuel effet du calendrier sur le risque de cambriolage

Chapitre 4

Analyse spatio-temporelle

Nous décrivons dans ce chapitre les méthodes envisagées pour modéliser de façon spatio-temporelle le risque de cambriolage.

4.1 Analyse d'une sur-sinistralité locale temporaire

Avant toute tentative de modélisation spatio-temporelle tenant compte des différents facteurs conjoncturels et structurels évoqués précédemment, nous analysons s'il est possible de constater et de valider statistiquement une dynamique de sinistralité qui impliquerait que des zones sinistrées ont temporairement un risque de cambriolage plus élevé suite à la survenance d'un sinistre.

Question Observe-t-on une sinistralité accrue dans le voisinage d'un contrat après que celui-ci a été sinistré ?

Pour répondre à cette question, nous allons envisager différentes méthodes.

4.1.1 Méthode à périodes fixées

Imaginons que l'on a repéré sur une période P_j les s_j sinistres au sein d'une population. On souhaite comprendre quelle pourrait être la répartition de sinistres à la période P_{j+1} suivante. Est-elle liée à celle de la période P_j ?

1ère étape : Construction de la table des K plus proches voisins Soit un entier K fixé, par exemple $K = 50$. A partir de la table des contrats localisés par

leur coordonnées lambert (X, Y), on construit la table **Voisinage** des K plus proches voisins avec la fonction *get.knn* du package *FNN*.

2ème étape : fonction de comptage La fonction de comptage prend en argument une période P sous forme de deux dates. Elle effectue la procédure suivante :

- récupérer la liste des contrats sinistrés pendant la période P
- récupérer dans la table **Voisinage** les références des K plus proches voisins pour ces contrats
- compter le nombre de sinistrés parmi ces références pendant la période P

3ème étape :itérations sur périodes On appelle la fonction de comptage sur les différentes périodes d'études. On obtient en sortie une table de la forme suivante :

contrat	P1 : 1/0	P1 : nb.sin(K)	P2 :1/0	P2 :nb.sin(K)	...	Pj :1/0	Pj :nb.sin(K)

Pour chaque contrat i et chaque période P_j, on a ainsi les informations suivantes :

- le contrat a été sinistré pendant la période P_j : oui/non (1/0)
- le nombre de voisins de ce contrat sinistré pendant la période P_j

4ème étape : calcul de taux de sinistres A partir de la table obtenue précédemment, on calcule le taux de sinistres.
Soit $x_{i,j}$ la variable aléatoire qui vaut 1 (respectivement 0) lorsque le contrat i a été sinistré (respectivement non sinistré) pendant la période P_j. Notons $n_{i,j}^{(K)}$ le nombre de sinistres survenus pendant la période P_j au sein du groupe constitué du contrat i et de ses K voisins.
Pour chaque couple de périodes P_j et P_{j+1}, on calcule les quantités suivantes :

$$P_{1,j}^{(K)} = \frac{1}{K+1} \frac{\sum_{i/n_{i,j}^{(K)}>0} n_{i,j+1}^{(K)}}{\#\{i : n_{i,j}^{(K)} > 0\}}$$

$$P_{0,j}^{(K)} = \frac{1}{K+1} \frac{\sum_{i/n_{i,j}^{(K)}=0} n_{i,j+1}^{(K)}}{\#\{i : n_{i,j}^{(K)} = 0\}}$$

On interprète $P_{1,j}^{(K)}$ comme une estimation de la probabilité d'occurence d'un sinistre au cours de la période P_{j+1} pour les contrats appartenant au voisinage d'un contrat sinistré au cours de la période précédente.

On interprète $P_{0,j}^{(K)}$ comme une estimation de la probabilité d'occurence d'un sinistre au cours de la période P_{j+1} pour les contrats n'appartenant pas au voisinage d'un contrat sinistré pendant la période précédente.
Lorsque le rapport $\frac{P_{1,j}^{(K)}}{P_{0,j}^{(K)}}$ est supérieur à 1, cela signifie que les sinistres surviennent plus souvent dans les voisinages de lieux déjà sinistrés.

On obtient donc une table contenant les valeurs $P_{1,j}^{(K)}$, $P_{0,j}^{(K)}$ et leur rapport pour chaque couple de périodes.

Remarque : on peut calculer ce rapport pour différentes valeurs de K et différentes longueurs de périodes.

Résultats On calcule le rapport $\frac{P_1}{P_0}$ pour différentes valeurs de K. On s'attend à ce que ce rapport tende vers 1 lorsque le nombre de plus proches voisins devient grand. On a fixé la durée d'une période à 60 jours, pour obtenir l'histogramme des valeurs donné figure 4.1

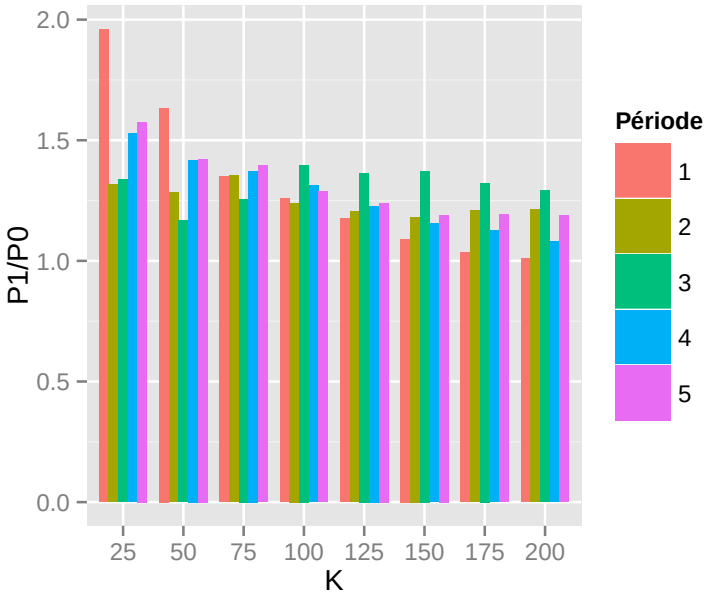


FIGURE 4.1 Rapport $\frac{P_1}{P_0}$ en fonction du nombres de plus proches voisins, pour les sinistres dans le 93 en 2013

On voit que le rapport $\frac{P_1}{P_0}$ diminue lorsque le nombre de plus proches voisins considérés augmente. Ce rapport est proche de 1 lorsque K devient grand. Pour le 93,

lorsqu'on considère un voisinage de 200 voisins, la sinistralité observée dans un voisinage suite à un sinistre est quasiment celle observée dans des zones non sinistrées.

Cette première méthode à périodes fixées semble montrer une sur-sinistralité locale. Néanmoins, elle a le désavantage de fixer les périodes de façon déterministe. On imagine ainsi une deuxième méthode pour s'affranchir de cette contrainte.

4.1.2 Méthode au jour par jour

On propose une autre méthode en regardant ce qu'il se passe dans les voisinages des contrats sinistrés un jour j, et on compare à ce qu'il se passe dans les voisinages des contrats pour lequel il n'y a eu aucun sinistre le jour j.

Modélisation Soit j un jour. On note $C^{j+1,j+d}$ la variable aléatoire qui prend la valeur 1 lorsque le contrat C a été sinistré entre les jours j et $j+d$ et 0 sinon. On note V_C le voisinage du contrat C.
Ainsi : $\exists C_1 : C_0^{j,j} = 1$ et $C \in V_{C_0}$, signifie que le contrat C appartient au voisinage d'un contrat sinistré le jour j.

On s'intéresse à probabilité :

$$\mathbb{P}_1^j(C^{j+1,j+d} = 1 | \exists C_0 : C_0^{j,j} = 1, C \in V_{C_0})$$

contre la probabilité :

$$\mathbb{P}_0^j(C^{j+1,j+d} = 1 | \forall C_0 : C_0^{j,j} = 1, C \notin V_{C_0})$$

Les paramètres du modèle sont encore la plage temporelle donnée par le nombre de jours d, ainsi que le paramètre de voisinage K (nombre de voisins considérés).

Ces deux probabilités sont respectivement estimées par :

$$\frac{\sum_{C, \exists C_0 : C_0^{j,j}=1, C \in V_{C_0}} C^{j+1,j+d}}{\#\{C, \exists C_0 : C_0^{j,j} = 1, C \in V_{C_0}\}}$$

et

$$\frac{\sum_{C, \forall C_0 : C_1^{j,j}=1, C \notin V_{C_0}} C^{j+1,j+d}}{\#\{C, \forall C_0 : C_0^{j,j} = 1, C \notin V_{C_0}\}}$$

On peut ensuite intégrer sur l'ensemble des jours pour lesquels il y a eu un sinistre.

Résultats On a estimé les quantités précédentes pour différentes valeurs des paramètres d et K. Le rapport des quantités donne une information sur le caractère local, spatialement et temporellement du cambriolage.

Un rapport de 1 indique que pour la durée d et le nombre de voisins K considérés, la probabilité d'occurrence d'un sinistre sous d jours pour un contrat appartenant au voisinage d'un contrat sinistré est la même que pour un contrat pour lequel il n'y a pas eu de sinistre dans le voisinage.
Lorsque le rapport est supérieur à 1, cela veut dire qu'un sinistre a plus de chance de se produire dans un voisinage d'un contrat déjà sinistré.

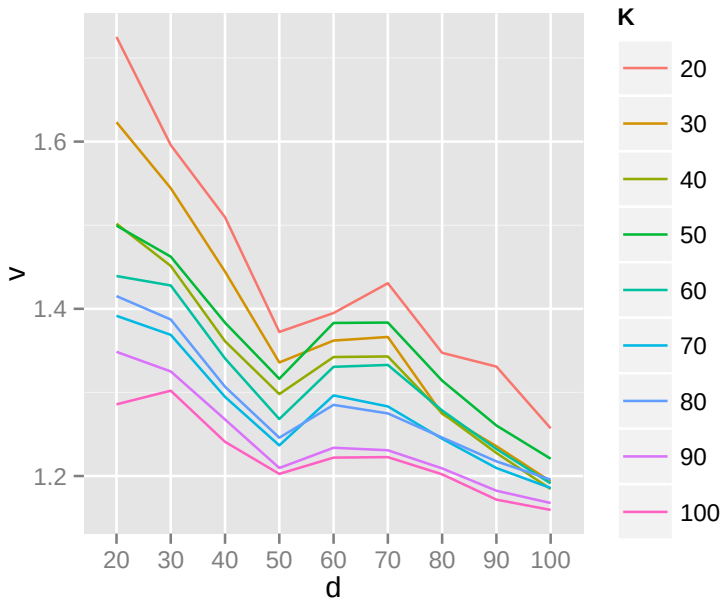


FIGURE 4.2 Variation du rapport en fonction du laps de temps d, pour différents K ; données du 93

On voit (figure 4.2) que le rapport est supérieur à 1, même pour un laps de temps long : il y a bien des clusters de sinistres, c'est à dire des zones où les cambriolages sont plus fréquents.

Il y a une première diminution du rapport pour $20 < d < 50$: de nouvelles zones sont sinistrées et les zones déja sinistrées sont moins sinistrées. Si on intègre un laps de temps plus long, entre 50 et 80 jours, le rapport augmente de nouveau : on peut penser que les cambrioleurs reviennent dans les voisinages des premiers contrats sinis-

trés.

Si on regarde l'évolution du rapport lorsqu'on augmente le paramètre spatial K, on se rend compte que le caractère local de la sinistralité s'estompe avec le laps de temps. En effet, on voit par exemple que le niveau d=100 est quasiment constant, ce qui signifie que sous 100 jours qui suivent un sinistre, il y a autant de sinistres dans le voisinage proche du contrat sinistré que dans un voisinage plus éloigné.

En revanche, pour des laps de temps plus réduits, la sinistralité est plus grande localement. Par exemple sous 20 jours après un sinistre, il se produit en moyenne 1.7 fois plus de sinistres dans le voisinage des 20 plus proches voisins, et seulement 1.3 fois plus dans le voisinage des 100 plus proches voisins.

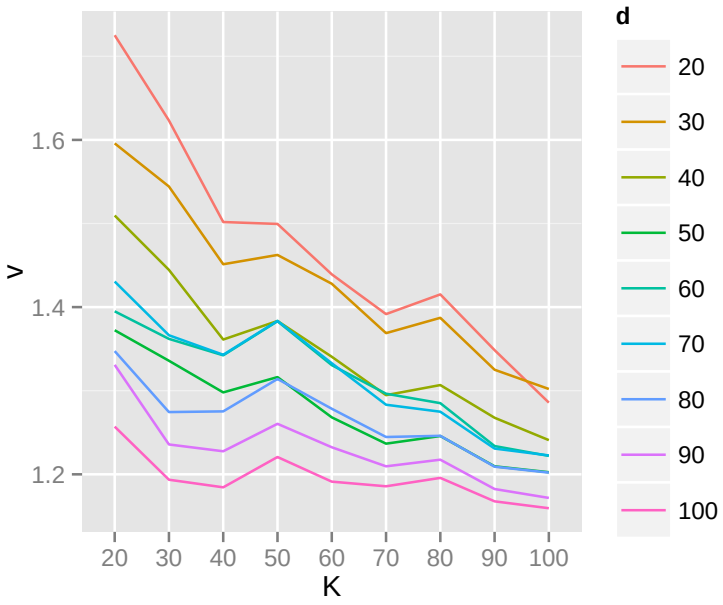


FIGURE 4.3 Variation du rapport en fonction du nombre de plus proches voisins K, pour différents laps de temps ; données du 93

Les résultats précédents ne peuvent être confortés par des intervalles de confiance qui valideraient la sur-sinistralité locale et temporelle après l'occurrence d'un sinistre. En effet, les écart-types sur le rapport $\frac{P_1}{P_0}$ sont particulièrement grands, notamment lorsque K et d sont faibles : nous n'avons pas assez de données à disposition pour valider avec un test statistique les résultats obtenus.

Pour réduire la variance, on a envie d'agréger un plus grand nombre de données. Plutôt que de calculer les fréquences des sinistres survenus après un jour J, on peut calculer les fréquences des sinistres survenus après une période. Cela revient à moyenner sur différentes valeurs de d proches. Cette idée fait l'objet d'une troisième méthode.

4.1.3 Méthode des fenêtres glissantes

Afin d'agréger plus de données, on considère les sinistres non pas survenus un jour donné mais dans une plage de jours.

On fixe les paramètres entiers suivants :

— d : longueur de la plage de jours

— D : durée sous laquelle on tient compte des sinistres survenus dans le voisinage

— K : nombre de plus proches voisins considérés pour constituer le voisinage

L'algorithme est le suivant :

1. $t = 0$: premier jour pour lequel on a constaté un sinistre. On regarde les contrats sinistrés dans la plage temporelle $[t, t + d[$
2. On récupère la liste L^1 des voisins de ces contrats sinistrés. On obtient ainsi un vecteur de n_1^1 contrats
3. On compte le nombre de voisins k_1^1 sinistrés sous D jours après la date de sinistre
4. On compte dans la liste complémentaire de L^1 (de taille n_1^1) le nombre k_1^1 de contrats sinistrés dans la plage temporelle $[t + d, t + d + D[$.
5. On répète les étapes 1 à 4 aux instants suivants $t + d$, $t + 2d$, ...,$t + (m - 1)d$
6. On obtient des réalisations appariées $(k_1^1, n_1^1, k_0^1, n_0^1) \dots (k_1^m, n_1^m, k_0^m, n_0^m)$

Remarque Outre la nécessité d'aggréger les données qui rend plus difficile l'interprétation des résultats, cette méthode nous contraint de choisir des fenêtres temporelles légèrement différentes pour les contrats des voisinages d'un sinistre et pour les autres.

Afin de vérifier la sur-sinistralité, on compare les quantités $k_0 n_1$ et $k_1 n_0$. Si la première est plus grande que la seconde, cela signifie qu'il y a sur-sinistralité.

Afin de valider statistiquement la sur-sinistralité, on peut utiliser un test de Wilcoxon, qui est approprié pour des réalisations appariées.

Test de Wilcoxon Le test de Wilcoxon a l'avantage d'être non paramétrique.

Soit des observations $(x^1, \dots, x^m)$ d'une loi $\mathcal{P}_x$ inconnue et des observations $(y^1, \dots, y^m)$ d'une loi $\mathcal{P}_y$ elle aussi inconnue. On teste l'hypothèse :

$$\mathcal{H}_0 : \mathcal{P}_x = \mathcal{P}_y$$

L'idée est que, si les probabilités sont égales, alors en regroupant les observations $(x^1, \dots, x^m)$ et $(y^1, \dots, y^m)$ et en les rangeant dans l'ordre, l'alternance des x^i avec les y^i est assez régulière. On écrit donc les statistiques d'ordre de l'échantillon global, et on calcule la somme des rangs W_x des x^i. Sous $\mathcal{H}_1$, la loi de W_x est :

$$\forall l \in [\![\binom{m}{2}, \binom{2m}{2} - \binom{m}{2}]\!], \mathcal{P}_{\mathcal{H}_0}(W_x = l) = \frac{k_l}{\binom{2m}{m}}$$

où k_l est le nombre de m-uplets d'entiers $r_1 \dots r_m$ dont la somme vaut l et qui sont tels que :

$$1 \leq r_1 < r_2 < \dots < r_m \leq 2m$$

Ainsi, si $\mathbb{P}(W_x = W_{x,réalisée})$ est petit (par exemple inférieur à 0.05), on rejette $\mathcal{H}_0$.

Résultats On a donc appliqué un test de Wilcoxon avec les échantillons $x = (x^1, \dots, x^m)$ et $y = (y^1, \dots, y^m)$, où $x^i{=}k_1^i n_0^i$ et $x^i{=}k_0^i n_1^i$.

Afin d'agréger suffisamment de données, on a choisi les paramètres suivants : $d = 40$, $D = 40$, $K = 140$.

On obtient :

```
wilcox.test(x,y)

Wilcoxon rank sum test with continuity correction

data:  x and y
W = 1313, p-value = 0.5653
alternative hypothesis: true location shift is not equal to 0
```

La p-value est telle que nous ne pouvons pas refuser l'hypothèse nulle selon laquelle les lois de probabilité $\mathcal{P}_x$ et $\mathcal{P}_y$ sont égales. Le test ne permet pas de confirmer une sur-sinistralité locale temporaire suite à un sinistre.

4.1.4 Conclusion

L'objectif de cette section était simplement de vérifier une sur-sinistralité locale temporaire suite à la survenance d'un cambriolage dans une zone donnée. Malgré différentes approches, nous ne sommes pas parvenus à valider ce point.

Nous nous sommes en effet confrontés aux difficultés et limites suivantes :

— Dès que nous nous restreignons à une fenêtre temporelle, nous n'avons accès qu'à très peu de sinistres. En effet, pour Paris par exemple, ayant près de 90 000 polices actives, il ne se produit qu'un peu plus de 1000 cambriolages dans l'année, soit moins d'une centaine par mois. Ainsi, même en considérant un nombre conséquent de voisins, nous ne comptons qu'une poignée de sinistres dans le voisinage pour une fenêtre temporelle de 30 jours. Cela implique une certaine variabilité statistique et il est alors impossible de confirmer les résultats avec un bonne confiance.

— Lorsque nous agrégeons les données, en augmentant le nombre de plus proches voisins K, ou en augmentant la durée des fenêtres temporelles, nous gagnons en confiance statistique mais nous nous éloignons de l'analyse locale souhaitée. Cela ne nous intéresse pas par exemple de prendre une fenêtre temporelle supérieure à 2 mois. D'une part parce que l'on peut considérer que deux cambriolages qui surviennent dans une même zone à plus de 2 mois d'intervalle restent indépendants, d'autre part parce que dans le cas de la prévention, il n'est pas réaliste de demander aux assurés de rester prudents pendant plus deux mois suite à un cambriolage survenu dans leur voisinage.

Bibliographie

[1] Données météo. http://meteo-climat-bzh.dyndns.org/.

[2] Anderson K.B. Dorr N. DeNeve K.M. & Flanagan M. Anderson, C.A. Temperature and aggression. chapter in advances in experimental social psychology, 32, 63-133. 2000.

[3] Raymond Paternoster ; Shawn D. Bushway. Theoretical and empirical work on the relationship between unemployment and crime. 2001.

[4] Vânia Ceccato. Exploring offence statistics in stockholm city using spatial analysis tools. 2002.

[5] Simon C.F.Shu and Jason N.H.Huang. Spatial configuration and vulnerability of residential burglary : A case study of city in taiwan. *4th International Space Syntax Symposium London*, 2003.

[6] Observatoire National de la Délinquance et des Réponses Pénales. La criminalité en France - Rapport Annuel 2013.

[7] Isaac Ehrlich. On the relation between education and crime. 1975.

[8] Jerome H. Friedman. Stochastic gradient boosting. *Department of Statistics Stanford University*, 1999.

[9] T F Hartnagel. Correlates of criminal behaviour. 1987.

[10] Markus Loecher. *RgoogleMaps : Overlays on Google map tiles in R*, 2014. R package version 1.2.0.6. URL : http://CRAN.R-project.org/package=RgoogleMaps.

[11] Wei-Yin Loh. Classification and regression trees. *Wiley Interdisciplinary Reviews : Data Mining and Knowledge Discovery*, 1, 2011.

[12] Mathieu Nebra. Les expressions régulières. 2015. URL : http://openclassrooms.com/courses/concevez-votre-site-web-avec-php-et-mysql/les-expressions-regulieres-partie-2-2-2.

[13] U.S. Department of Justice. Crime and seasonality : National crime survey report -sd-ncs-n15, ncj-64818. 1980.

[14] Malczewski ; Poetz. Residential burglaries and neighborhood socioeconomic context in london, ontario : Global and local regression analysis.

[15] Travis C. Pratt. Assessing the relative effects of macro-level predictors of crime : A meta-analysis. 2001.

[16] Matthew Ranson. Essays on the economics of climate change. *M - RCBG Associate Working Paper Series*, 8, 1980.

[17] Greg Ridgeway. *gbm : Generalized Boosted Regression Models*, 2013. R package version 2.1. URL : http://CRAN.R-project.org/package=gbm.

[18] Edzer Pebesma Roger S. Bivand. Applied spatial data analysis with r. 2013. URL : http://cran.r-project.org/web/packages/sp/index.html.

[19] Carol W. Kohfeld ; John Sprague. Urban unemployment drives urban crime. 1988.

[20] Terry Therneau, Beth Atkinson, and Brian Ripley. *rpart : Recursive Partitioning and Regression Trees*, 2014. R package version 4.1-8. URL : http://CRAN.R-project.org/package=rpart.

[21] Hadley Wickham. *ggplot2 : elegant graphics for data analysis*. Springer New York, 2009. URL : http://had.co.nz/ggplot2/book.

[22] Hadley Wickham. *stringr : Make it easier to work with strings.*, 2012. R package version 0.6.2. URL : http://CRAN.R-project.org/package=stringr.

[23] Wikipedia. Projections coniques conforme de lambert. URL : http://fr.wikipedia.org/wiki/Projection_conique_conforme_de_Lambert.