le cnam

INSTITUT DES ACTUAIRES

# Construction d'un zonier en MRH à l'aide d'outils de data-science

Mémoire d'actuariat présenté pour l'obtention du

**Master professionnel Sciences de gestion, mention finances de marché Spécialité Actuariat du CNAM**

Et l'admission à **l'Institut des Actuaires**

par Guillaume Beraud-Sudreau

Caractère confidentiel : NON

Jury :
Michel FROMENTEAU
Edith BOCQUAIRE
Gwenaël BILLIOTTE
David FAURE
Pierre PETAUTON
François WEISS

Mémoire encadré par Philippe Marie-Jeanne

**Résumé**

L'objectif de ce mémoire est de présenter rigoureusement des méthodes de modélisation innovante en actuariat, et donner un exemple d'application de celles ci en tarification Multi-Risque Habitation.
Les problèmes généraux de modélisation (définition d'un modèle prédictif, présentation de différentes métriques de performance de modèles et comparaison de celles ci) seront tout d'abord décrits, afin de fournir une base théorique à une seconde partie, présentant quelques algorithmes de régression. Ces algorithmes pourront finalement être appliqués dans le cadre pratique de la création d'un zonier de risque MRH. Dans ce cadre, nous aborderons aussi la problématique de l'acquisition de variables externes, leur traitement pratique ainsi que leurs caractéristiques statistiques.

**Remerciements**

Je tiens tout d'abord à remercier Philippe Marie-Jeanne, directeur du Data Innovation Lab d'AXA, d'avoir accepté de diriger ce mémoire. Je le remercie aussi, plus largement, pour l'ensemble de ses conseils et des expériences qu'il a pu partager avec moi lorsque j'ai travaillé au sein de son équipe.

J'ai également eu la chance de travailler au Data Innovation Lab avec Alexandre Gerbeaux et Arnaud Debergh sur l'utilisation de données géographiques en tarification MRH. Je les remercie chaleureusement pour les nombreuses discussions et leur contribution déterminantes pour le succès de ce projet.

Je remercie aussi Anne-Laure Le Gallo, qui m'a introduit dans le monde merveilleux de l'assurance Multi-Risque Habitation, et qui a eu la patience de partager avec moi sa connaissance de ce domaine.

Enfin, je remercie Julien Callard, Isabelle Antoine et les équipes MRH de AXA France et Direct Assurance pour les nombreux échanges que nous avons eus.

# Table des matières

# Première partie

# Introduction

# Chapitre I.1

# Objectifs du mémoire

## I.1.1 Présentation des objectifs

L'objectif de ce mémoire est de présenter différentes méthodes d'apprentissage statistique, ainsi qu'un exemple d'application dans le cadre de la création d'un zonier de risque en assurance Multi-Risque Habitation.

Il ne s'agit donc pas d'un mémoire dédié spécifiquement à l'assurance multi-risques habitation : les concepts techniques présentés (en particulier dans les parties II et III) peuvent être appliqués à d'autres types d'assurance - par exemple l'assurance auto - et les conclusions obtenues peuvent également se généraliser.
A contrario, certains éléments extrêmement importants de la détermination d'une prime pure - développement des sinistres, évolution temporelle des fréquences et des sévérités ("trending"), estimation des frais liés aux polices ou de la marge optimale - sont abordés très rapidement sans être développés.

Afin de construire un zonier en suivant les méthodes décrites dans ce document, il est nécessaire d'utiliser un certain nombre de méthodes d'intelligence artificielle.
Ces méthodes sont donc décrites, et leur efficacité comparée.
Dans ce but, diverses méthodes d'estimation des performances de modèles de régression seront présentées (ainsi qu'un cadre théorique nécessaire à leur présentation).
Enfin, toutes les méthodes décrites dans ce mémoire étant destinées, entre autres, à la tarification MRH, celle ci ainsi que les données utilisées pour nos applications numériques, seront brièvement présentées en introduction.

## I.1.2 Organisation du mémoire

En introduction, nous rappellerons tout d'abord rapidement la définition et l'importance de l'estimation de la prime pure en assurance habitation.
Ensuite, dans une première partie, nous présenterons certains principes fondamentaux de l'apprentissage statistique. En particulier, après avoir précisément défini ce qu'était un modèle, nous rappellerons le concept de compromis biais-variance, ainsi que de validation croisée, afin de justifier l'utilisation de ce type d'outils lors de la réalisation pratique de modèles prédictifs. Puis, dans

une seconde partie, nous présenterons trois types de modèles prédictifs : les Modèles Linéaires Généralisés (GLMs), outils couramment employé en actuariat, puis les arbres de régression, outil très populaire en machine learning, et enfin les forêts aléatoire, outil qui semble parmi les familles de modèles les plus performantes à l'heure actuelle.
Pour ces trois types de modèles, nous décrirons en profondeur les principes et hypothèses sur lesquels ils reposent, les algorithmes d'apprentissage utilisés pour les paramétrer, ainsi que les qualités et défauts qu'ils présentent. Enfin, dans une dernière partie, nous proposerons une application de ces modèles à la création d'un zonier de risque en assurance habitation.
Pour cela, nous décrirons les méthodes utilisées pour obtenir un grand nombre d'informations sur les polices utilisées pour la modélisation, puis les outils ainsi que l'approche utilisée pour construire un modèle prédictif du risque, en tenant compte des spécificités des variables manipulées (en particulier leur forte auto-corrélation).
Les techniques décrites dans ce mémoire seront enfin illustrées par des applications pratiques, en particulier sur la création d'un zonier MRH, dont les résultats seront présentés à la fin de ce mémoire.

### I.1.3 Apports du mémoire

Comme décrit ci dessus, l'objectif de ce mémoire est de présenter, dans le cadre de la tarification d'assurance IARD, différentes méthodologies d'apprentissage statistique. Les aspects théoriques de ces méthodes sont bien connus, et elles sont utilisées couramment dans diverses industries. Cependant, au moins lors de la rédaction du présent mémoire, leur connaissance et utilisation dans le milieu actuariel étaient marginales.

Les rappels d'apprentissage statistique (II.2), la présentation des différentes méthodes d'estimation des modèles II.3, ou la description des modèles linéaires généralisés (III.1), d'arbres (III.2) et de forêts aléatoires (III.3) ne sont donc globalement pas des innovations, mais il s'agit probablement d'une des premières tentatives de décrire ces techniques dans un cadre actuariel, et de mesurer les gains qu'elles peuvent permettre.
Au sein de ces parties, plusieurs éléments sont plus que de simples transpositions ou adaptations d'un savoir faire ou de théories au milieu actuariel ; en particulier, malgré mes recherches, je n'ai pu trouver d'autres descriptions explicites du modèle de Gini décrit chapitre II.3.4.c, ni d'autres illustrations de l'impact de la mesure de performance sur la complexité des modèles (II.3.1). De même, si le phénomène de bagging et la théorie sur laquelle reposent les forêts aléatoires est bien connue, je n'ai jamais pu trouver d'illustration directe de ces principes - comme celle que j'ai tenté de réaliser chapitre III.3.1. Enfin, la création d'arbres optimisant le critère de Gini (décrite chapitre III.2.2.b) et la mesure de sa performance sont probablement nouvelles.
Ces éléments sont modestes mais peuvent être généralisés à l'ensemble de la tarification IARD.

Enfin, la dernière partie, décrivant la création d'un micro-zonier, décrit un travail plus appliqué. A ce titre, elle représente donc dans son ensemble une contribution moins générale mais originale de ce rapport.
Au sein de cette partie, outre les méthodes de création et l'exploitation des variables explicatives, les remarques sur la non-indépendance des variables géographiques (partie **??**) ont une portée générale et semblent souvent ignorées des actuaires prenant part à la tarification des contrats.

Ce mémoire se cantonne donc à la segmentation du risque, et ne cherche pas à être un travail spécifiquement réalisé pour l'assurance multi-risques habitation.
Si celle ci fait partie des applications possibles des principes décrits ici - et sert d'exemple pour

les différentes applications numériques - l'objectif n'est pas de revoir en profondeur l'ensemble des éléments constituant les primes (ou même les primes pures).
A ce titre, nous rappelons donc que la plupart des éléments affectant la prime qui ne sont pas décrits dans ce rapport (développement des sinistres, estimations des divers frais et commissions, optimisation des marges etc...) n'ont pas d'impact direct sur la création du zonier (celui ci étant par nature centré).

# Chapitre I.2

# Principe d'assurance

## I.2.1 Mutualisation des risques

Le métier de l'assureur Dommages consiste à organiser, en mutualité, une multitude d'assurés exposés à la réalisation de certains risques, supposés indépendants les uns des autres, et d'indemniser ceux d'entre eux qui subissent un sinistre grâce à la masse commune des primes collectées. Le principe de mutualisation repose sur le fait que les sinistres survenus aux différents assurés étant indépendants, la variance de leur somme est relativement faible par rapport à celle ci (le ratio entre la variance et l'espérance des sinistres doit décroitre avec la racine du nombre de clients, si ceux ci présentent bien des risques indépendants et identiquement distribués - ce qui est approximativement le cas).
Ce principe de mutualisation des risques permet d'obtenir, pour qu'une société d'assurance soit rentable, l'inégalité suivante :

$$\sum_i P_i \geq \sum_i \mathbb{E}(R_i) + F_i$$

avec $P_i$ la prime versée par l'assuré $i$, $R_i$ le montant remboursé à l'assuré $i$ pour les sinistres suceptibles de l'impacter, et $F_i$ les frais de l'assureur liés à l'assuré $i$.

## I.2.2 Construction de la prime

Afin de garantir que cette inégalité soit bien vérifiée, l'assureur cherche à avoir, pour chaque assuré $i$ :

$$\forall i : P_i \geq \mathbb{E}(R_i) + F_i \tag{I.2.1}$$

En effet, la possibilité de vendre des polices à perte risque fortement d'entrainer une forte anti-sélection (si les autres assureurs proposent des prix plus élevés - ce qui est sans doute le cas) les personnes se voyant proposer des polices avec un prix plus faible que leur risque risquent fortement de souscrire, réalisant ainsi la perte.

### I.2.2.a Définition de la prime pure

La tarification de produits d'assurance est donc au cœur du métier d'assureur. Les compagnies d'assurance utilisent quotidiennement des modèles statistiques pour évaluer les risques auxquels elles doivent faire face. En particulier, les modèles de régression permettent de quantifier les relations entre l'espérance de la valeur des risques des contrats assurés $\mathbb{E}(R_i)$ et les variables

décrivant ces risques. Chaque police d'assurance se définit en fonction de variables de tarification qui vont permettre d'expliquer la sinistralité observée. Ces variables sont généralement des informations :

- sur l'assuré : par exemple, l'âge ou la catégorie socio-professionnelle (CSP) pour un particulier, le secteur d'activité ou le nombre de salariés pour une entreprise ;
- sur le bien assuré : par exemple, l'âge du véhicule, la puissance ou la marque en assurance auto, la surface du logement en MRH ;
- géographiques : Ces variables, liées au lieu où réside le client, seront décrites en détail dans la dernière partie de ce mémoire.

Dans certains pays, l'utilisation de certaines variables est restreinte pour des raisons éthiques (par exemple, en Europe, le prix de l'assureur ne peut dépendre du sexe ou de la religion, entre autres).

L'assureur a donc besoin de prédire l'espérance des sinistres attendus, $\mathbb{E}(R_i)$, appelée *Prime Pure*, à l'aide de variables connues. La création de ces modèles prédictifs est donc au cœur de la mission de l'actuaire.

### I.2.2.b Définition de la prime commerciale

Après avoir estimé correctement la prime pure et l'ensemble des coûts générés par un client (coûts de gestion de sinistres, coûts de gestion de police, coût d'acquisition, coût d'immobilisation du capital lié au risque...), l'actuaire peut estimer le prix optimal auquel il souhaite présenter le produit.
En effet, si on suppose que l'on dispose d'une estimation de la probabilité d'achat $d_i(P)$ de la police d'assurance par le client $i$ au prix $P$, on peut estimer le profit $B_i(P)$ réalisé sur un contrat :

$$B_i(P) = d_i(P) \times (P - \mathbb{E}(R_i) - F_i)$$

On cherchera donc la prime commerciale $P_i$ qui maximisera ce profit :

$$P^*{}_i = Argmax_P B_i(P) = Argmax_P d_i(P) \times (P - \mathbb{E}(R_i) + F_i) \tag{I.2.2}$$

(on remarquera que $B_i(P)$ est positif pour toutes les valeurs de $P$ qui respectent l'équation I.2.1, négatif sinon).
Cette optimisation du prix ne repose que sur une vision à 1 an du produit. Il est possible d'intégrer les bénéfices attendus les années suivantes (à partir de la probabilité de renouvellement du client $i$, qui dépend de la prime fixée la 1re année et des augmentations tarifaires proposées au client) dans l'optimisation du bénéfice, mais cette intégration soulève un certain nombre de difficultés qui dépassent le cadre de ce mémoire.

La construction de la prime commerciale $P_i$ permet de décomposer la prime d'assurance en 3 composantes distinctes :

- La prime pure : liée au risque du client.
- Les divers chargements liés aux frais de fonctionnement de l'assureur.
- La marge de l'assureur.

L'estimation de la prime pure est capitale dans la construction de la prime d'une police d'assurance. En effet, une erreur d'estimation de la prime pure risque fortement de générer des estimations des marges (et donc du prix total proposé au client) fantaisistes.

### I.2.2.c Modèles Fréquence-coût

Afin d'estimer l'espérance de la valeur totale $\mathbb{E}(R_i)$ des sinistres survenus au client $i$, on peut voir celle ci comme l'agrégation de deux nouvelles variables supposées indépendantes :

— la fréquence de sinistres : le nombre de fois où un sinistre est survenu durant la période d'exposition du client (le nombre d'années polices : par exemple, $0,25$ si le client est assuré depuis 3 mois)

$$fréquence = f_i = \frac{Nombre\ de\ sinistres}{Durée\ d'exposition}$$

— le coût moyen : le montant de sinistre moyen pour le client (s'il en a un)

$$coût = c_i = \frac{Charge\ totale\ des\ sinistres}{Nombre\ de\ sinistres}$$

La prime (annuelle) d'un client sera alors :

$$Prime\ pure = P_i = \frac{Charge\ totale\ des\ sinistres}{Dure\ d'exposition} = fréquence \times coût = f_i \times c_i \tag{I.2.3}$$

L'utilisation d'un tel découpage suppose une indépendance de la fréquence et des coûts de sinistre.
Soit $R_i = \sum_{j=1}^{N} C_j$ avec $N$ une loi de probabilité à valeur dans $\mathbb{N}$ et $C_j$ un ensemble de lois indépendantes et identiquement distribuées à support réel et strictement positif. On a : $R_i = 0$ si $N = 0$, ce qui correspond à un cas dégénéré. On suppose également que, sachant les informations liées au profil du client dont disposent les assureurs, les $C_j$ sont indépendants de $N_i$ et indépendants entre eux (cette hypothèse, très forte, est clairement discutable).
Bien entendu $N$ correspond au nombre de sinistres que subira le client $i$, et les $C_j$ le coût de ces $N$ sinistres (ce sont a priori des variables aléatoires).

$$\mathbb{E}(R_i) = \mathbb{E}(\mathbb{E}(R_i|N)) = \mathbb{E}(\sum_{j=1}^{N} \mathbb{E}(C_j|N)) = \mathbb{E}(N \times \mathbb{E}(C_j)) = \mathbb{E}(N) \times \mathbb{E}(C_j) \tag{I.2.4}$$

Nous obtenons donc que le résultat de l'espérance du montant de sinistre total, ou prime pure, est égal au produit des espérances de la fréquence et du coût moyen. Cela revient donc à modéliser la fréquence de sinistre et le coût moyen. Cette approche sera suivie dans ce mémoire.

## I.2.3 Assurance Multi-Risques Habitation (MRH)

Un portefeuille de polices d'assurance multi-risques habitation servira d'application numérique à ce mémoire. Nous rappellerons donc dans cette section les principales caractéristiques de ce type d'assurance.

### I.2.3.a Importance de l'assurance habitation

L'assurance Multi-Risque Habitation (abrégée MRH par la suite) protège l'habitation de l'assuré ainsi que son contenu.
L'assurance MRH est, par les volumes de primes collectés et de sinistres versés, la 2$^{\text{nde}}$ assurance la plus importante en France (derrière l'assurance auto).

Elle est cependant déficitaire en France. L'une des raisons de cette situation est une hausse des risques représentés par les habitations - dégradation du bâti - ainsi que la difficulté d'estimation du risque de catastrophe naturelle. Ces facteurs tendent à faire sous-estimer le risque assuré ; cette sous-estimation est importante dans le cadre d'un marché concurrentiel, où l'assurance habitation est souvent considérée comme un produit secondaire, permettant de fidéliser les clients possesseurs d'une assurance auto, rentable (technique de vente croisée).

### I.2.3.b Présentation des couvertures MRH

L'assurance habitation peut être décomposée en couverture sur plusieurs risques :
— Dégât des eaux
— Vol
— Incendie
— Responsabilité civile
— Bris de glace
— Catastrophe naturelle et évènements climatiques

Chez AXA France, les trois premières catégories représentent chacune environ un quart de la prime pure (ces proportions sont différentes selon le type de bien : les catastrophes naturelles et évènements climatiques représentent un risque important pour les maisons mais mineur pour les appartements, alors que les dégâts des eaux représentent quasiment la moitié du risque de ces derniers.
De plus, il est important de noter que les risques se réalisant les plus fréquemment (dégât des eaux ou vol, par exemple), ne représentent pas nécessairement les coûts les plus élevé (la sévérité d'un incendie étant en moyenne considérablement plus élevée que celle d'un dégât des eaux).

Ce mémoire illustre les méthodes utilisées dans le cas d'une assurance multi-risques habitation, en particulier sur les garanties dégâts des eaux et vol.
Ces deux garanties sont présentées ici ; les autres garanties, qui ne servent pas d'illustration dans ce mémoire, sont simplement rappelées.

#### Dégâts des eaux

L'assurance Dégâts des eaux couvre les conséquences des dommages causés par l'eau résultant :
— De la fuite, de la rupture ou du débordement des conduites d'eau non enterrées (les conduites enterrées sont des conduites dont l'accès nécessite des travaux de terrassement ; les conduites encastrées sont garanties).
— De la fuite, de la rupture ou du débordement des appareils à effet d'eau.
Les appareils à effet d'eau sont des appareils auxquels il est ajouté un élément quelconque qui a pour but de permettre certaines opérations telles que l'arrivée de l'eau, son évacuation, son chauffage, son épuration, son aération, créant alors un certain mouvement d'eau, même s'il n'est pas continu. Par exemple : machines à laver le linge et la vaisselle, baignoires, lavabos, éviers... Il est nécessaire que l'équipement soit relié en permanence au dispositif d'alimentation, d'évacuation, d'épuration ou de filtration.
— Des infiltrations d'eau ou de neige au travers des toitures, ciels vitrés, terrasses et balcons formant toiture.
— Des infiltrations d'eau et de neige au travers des façades et murs extérieurs.
— De la rupture accidentelle ou du débordement exceptionnel d'égout, non dus à un évènement climatique ; cette garantie concerne les égouts des voies publiques ou privées lorsqu'il

y a une rupture accidentelle ou un débordement exceptionnel d'égout non dus à un évènement climatique.
— Des infiltrations par les joints d'étanchéité au pourtour des installations sanitaires et au travers des carrelages.
— Des dégâts des eaux subis dus à la faute d'un tiers.
— Les frais engagés pour la recherche de fuites qui sont à l'origine d'un sinistre garanti à l'intérieur des biens assurés, ainsi que des frais de remise en état des biens dégradés par ces travaux de réfection.

**Vol**

**Définition** : Le vol est la soustraction frauduleuse du bien d'autrui (article 311-1 du Code Pénal). Le voleur s'arroge la détention matérielle d'une chose sans la volonté de son propriétaire. Le vol est consommé même si les objets soustraits sont ensuite abandonnés ou détruits.
La tentative de vol se définit comme tout acte accompli en vue de commettre un vol qui a reçu un commencement d'exécution mais qui a été suspendu ou manqué pour une cause quelconque.

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

Le vol peut avoir été commis notamment par :
— effraction,
— escalade : l'escalade est l'introduction par une ouverture située en étage. Il peut s'agir d'une fenêtre, d'un balcon. Il n'y aura pas effraction si l'ouverture n'était pas fermée,
— fausse clé : sont considérés, à titre d'exemple, comme des fausses clés, les crochets, passes-partout, clés imitées, contrefaites.
— maintien clandestin dans les lieux,
— menaces, violences ou intimidation de l'assuré, son entourage ou toute personne à son service : sont visés tous les coups et blessures, quel qu'en soit le résultat.

**Autres garanties**

Les autres garanties couvertes par l'assurances multi-risques habitation ne sont pas utilisées dans les applications numériques de ce mémoire, et ne sont donc pas présentées en détail.

### I.2.3.c Risques couverts dans ce mémoire

Dans ce mémoire, nous ne nous intéresserons qu'aux sinistres attritionnels (couvertures pour lesquels les survenances de sinistres sur différentes polices sont indépendantes). Ainsi, nous ne nous intéresserons pas aux sinistres de types évènements climatiques ou catastrophes naturelles. Ces couvertures peuvent être modélisées par ailleurs (par exemple en se basant sur des modèles physiques du territoire - hydrologique, dans le cas du risque d'inondation).
On remarque cependant que certains sinistres attritionnels peuvent être causés par un même évènement déclencheur (par exemple, on peut noter une relation entre température extrême et risque d'incendie). Ces relations seront négligées dans le cadre de ce mémoire, mais une étude approfondie de ce type de phénomènes est nécessaire à l'actuaire souhaitant mettre en œuvre les résultats présentés ici. Les exemples présentés dans ce mémoire concernent principalement les

garanties vol et dégât des eaux. Ce choix est motivé par plusieurs raisons :

- Les vols ou dégâts des eaux représentent la vaste majorité des sinistres ; il s'agit de risques dont la fréquence est relativement haute, ce qui les rend plus simples à étudier - disposant de plus d'observations, il est relativement aisé d'obtenir des résultats fiables.
- Les risques de vols ou de dégâts des eaux (et en particulier leurs sévérités) sont fortement marqués géographiquement : il s'agit donc d'un bon exemple pour construire un zonier.

A l'inverse, les autres garanties couvertes par l'assurance MRH ne sont pas étudiées dans ce mémoire, pour diverses raisons :

- Incendie : le risque d'incendie recouvre des évènements indépendants qui peuvent être modélisés aisément en utilisant les méthodes décrites dans ce mémoire ; cependant, du fait du faible nombre d'évènements, un soin particulier devra être apporté à la validation des résultats obtenus.
- Événements climatiques et catastrophes naturelles : ces évènements affectant simultanément un ensemble de polices, il n'est pas possible de construire un modèle décrivant ou prédisant le risque porté par une police : les observations disponibles n'étant pas indépendantes, les outils usuels d'apprentissages statistiques ne sont pas applicables.
- Bris de glace : Les modèles présentés ici peuvent s'appliquer aux bris de glace ; cependant, ils ne représentent qu'une part négligeable du risque des polices MRH.
- Responsabilité civile : Cette garantie peut être étudiée avec les méthodes proposées, après un traitement approprié des données.

### I.2.3.d Données utilisées

#### Base utilisée

La base utilisée pour les applications numériques de ce mémoire provient de l'historique des contrats d'AXA France.
Elle compte 900 000 observations, entre 2012 et 2014. Chaque ligne représente un et un seul contrat : l'année de souscription fait donc partie des variables utilisées, mais pas l'année d'exposition. La base correspond donc à environ 900 000 années-police.
Il s'agit d'un sous-ensemble de la base totale (qui compte plusieurs dizaine de millions de lignes). Ce seul sous-ensemble a été considéré dans la mesure où il permet de réaliser les tests plus rapidement, tout en gardant une taille - et donc une stabilité - très convenable.

Les données d'AXA France utilisées pour le test contenaient un grand nombre d'observation mais malheureusement relativement peu de variables (une vingtaine), ce qui n'a pas permis d'exploiter pleinement les algorithmes décrits dans ce rapport.
Parmi ces variables, les plus significatives, pour la plus-part des modèles créés, étaient le type d'occupation (propriétaire ou locataire), le type de résidence (principale ou secondaire), le type de logement (maison ou appartement), le nombre de pièces ou le montant d'objets de valeur déclarés par l'assuré.

#### Traitement réalisés

**Préparation des données** :
Avant d'être utilisées dans les expériences décrites dans ce rapport, les valeurs des sinistres ont été pré-traitées : en particulier les sinistres ont été développés, afin d'utiliser une estimation de la vision ultime des sinistres (calculée à l'aide de facteurs fournis par les équipes de reserving). De plus, les sinistres ont été écrêtés, afin d'éviter que les modélisations réalisées ne soient trop

fortement influencées par des sinistres atypiques et la charge due aux sinistres atypiques a été mutualisée sur l'ensemble des polices sinistrées.

**Offuscation des données** Pour des raisons évidentes de confidentialité, les données utilisées ont été légèrement offusquées (dégradées) : en particulier :

- la base a subi un échantillonnage des polices non-sinistrées, afin de (légèrement) gonfler artificiellement la fréquence de sinistralité dans la base observée.
- les sinistres ont été multipliés par un facteur aléatoire - proche de 1 - afin de faire perdre leur sens aux primes moyennes, tout en conservant intacts les problèmes de régression étudiés.

# Deuxième partie

# Modèles et Erreurs

# Chapitre II.1

# Objectifs du chapitre

Afin de pouvoir évaluer la pertinence des modèles proposés dans ce mémoire, nous devons définir des mesures de comparaison de la qualité des tarifs calculés. Dans ce chapitre, nous comparons donc plusieurs mesures de la qualité des différents régresseurs du risque proposés.

Il est important de signaler que ce problème, simple dans le cas de la comparaison de différents modèles issus d'une même famille dont on fera varier les paramètres et méta-paramètres, est plus complexe lorsqu'il s'agit de comparer deux modèles de structures différentes. Par exemple, dans le cadre d'un GLM doté d'une fonction de lien et d'une distribution données, on peut facilement juger de la pertinence de l'ajout d'une nouvelle variable explicative. En revanche, la comparaison d'un tel modèle avec une forêt aléatoire (Random-Forest) ou un réseau de neurones est nettement plus complexe.

# Chapitre II.2

# Rappels d'apprentissage statistique

## II.2.1 Formalisme et définitions

### II.2.1.a Modèle et apprentissage

**Définition 1** *Dans ce chapitre, nous supposerons qu'il existe un ensemble d'éléments nommé* $\Omega$. *Ces* éléments *se composent de* mesures *ou* variables explicatives $\mathbf{x} = (x_1, x_2, ..., x_P)$ *connues a priori, formant un point dans un espace* $\mathcal{X}$ *à p dimensions, et d'une variable* $\mathbf{y}$ *dite* variable cible*, que nous chercherons à exprimer, à valeur dans l'espace* $\mathcal{Y}$.

Dans le cadre de ce rapport, un élément correspond à un contrat d'assurance MRH. Les mesures $\mathbf{x}$ correspondent aux données disponibles sur ce contrat (typiquement les réponses du client au formulaire de souscription) et la variable $\mathbf{y}$ correspond à la sinistralité du contrat.

**Définition 2** *Un* modèle *est une fonction de* $\varphi : \mathcal{X} \mapsto \mathcal{Y}$ *telle que les valeurs* $\varphi(\mathbf{x}) = \hat{y}$ *soient aussi proches de* $\mathbf{y}$ *que possible.*

**Définition 3** *Un* modèle de régression *(ou* régresseur*) est un modèle* $\varphi : \mathcal{X} \to \mathbb{R}$ *: l'espace de sa variable cible* $\mathcal{Y}$ *est donc l'espace des réels.*

Dans la suite de ce mémoire, nous nous intéresserons à des modèles de régression .

**Définition 4** *Un* tarif *est un modèle de régression* $\varphi : \mathcal{X} \to \mathbb{R}^+$ *où* $\mathcal{X}$ *est l'espace des informations disponibles sur les contrats et* $\mathbb{R}^+$ *est l'espace des montants de sinistres versés au client.*

*Les éléments de* $\Omega$ *représentent donc des contrats d'assurance (souscrits ou possibles) et le montant des sinistres qui y sont (ou seraient) rattachés.*

Traditionnellement, en actuariat, le modèle $\varphi$ utilisé est un Modèle Linéaire Généralisé.

**Définition 5** *Un* algorithme de modélisation*, ou* méthode de régression *est une fonction* $\phi : \mathscr{P}(\Omega) \mapsto (\mathcal{X} \mapsto \mathcal{Y})$ *(où* $\mathscr{P}(\Omega)$ *désigne l'ensemble des parties de* $\Omega$ *:* $\{\mathcal{L} | \mathcal{L} \in \Omega\}$*), qui définit, à partir d'une ensemble* $\mathcal{L}$ *d'éléments de* $\Omega$ *un modèle* $\phi(\mathcal{L})$ *noté* $\varphi_{\mathcal{L}}$.

*L'ensemble* $\mathcal{L}$ *utilisé pour construire le modèle* $\phi(\mathcal{L})$ *est nommé* ensemble d'apprentissage *de ce modèle.*

**Définition 6** *Une* méthode de tarification *est un algorithme de modélisation défini vers l'espace des tarifs.*

*Il s'agit donc d'un algorithme de modélisation* $\phi : \mathscr{P}(\Omega) \mapsto (\mathcal{X} \mapsto \mathbb{R}^+)$*, où :*

— $\Omega$ *désigne l'ensemble des contrats et des sinistres qui y sont attachés.*
— $\mathcal{X}$ *est l'espace des informations disponibles sur les contrats*
— $\mathbb{R}^+$ *est l'espace des montants de sinistres versés au client.*
— $\phi(\mathcal{L}) = \varphi_{\mathcal{L}}$ *désigne un tarif construit à partir de l'ensemble (historique) de contrats* $\mathbf{x}$ *et de sinistres* $\mathbf{y}$*, avec* $(\mathbf{x}, \mathbf{y}) \in \mathcal{L}$*.*
— $\varphi_{\mathcal{L}}(\mathbf{x}) = \hat{\mathbf{y}}$ *représente le montant* $\hat{\mathbf{y}}$ *de la prime proposée pour le contrat* $\mathbf{x}$*, dans le cadre du tarif* $\varphi_{\mathcal{L}}$*.*

Il est important de noter que, en pratique, les modèles sont tous construits sur un sous-ensemble d'éléments $\mathcal{L}$, qui n'est pas choisi par la personne en charge de sa création (et dépendent donc de cet ensemble).

Dans la suite de ce mémoire, les modèles $\varphi_{\mathcal{L}}$ créés sont définis, d'une manière ou d'une autre, par un jeu de paramètres : définir $\varphi_{\mathcal{L}}$ reviendra donc à définir ces paramètres.
De même, l'algorithme d'apprentissage utilisé $\phi$ peut lui même être défini par un jeu de paramètres. On parle alors de méta-paramètres de la modélisation.
Plusieurs méthodes de tarification (GLMs, arbres de régression ou forêts aléatoires) feront l'objet de la partie III de ce mémoire ; de même, la méthode de création du zonier (et sa définition même) est une méthode de régression.

**Définition 7** *L'*erreur *d'un modèle* $\varphi_{\mathcal{L}}$*, notée* $Err(\varphi_{\mathcal{L}})$*, est définie comme* $Err(\varphi_{\mathcal{L}}) = \mathbb{E}_{(X,Y)}(Loss(Y, \varphi_{\mathcal{L}}(X))$.

*Celle-ci dépend d'une fonction* $Loss() : \mathcal{Y} \times \mathcal{Y} \mapsto \mathbb{R}$*. Cette fonction est nommée* fonction de perte.

L'erreur d'un modèle ne peut être calculée explicitement sans connaitre tous les éléments $(\mathbf{x}, \mathbf{y})$ de $\Omega$, mais peut être estimée sur un ensemble d'éléments $\mathcal{L}'$ : $Err(\varphi_{\mathcal{L}}, \mathcal{L}') = 1/N' \times \sum_{(\mathbf{x},\mathbf{y})\in\mathcal{L}'} Loss(\mathbf{y}, \varphi_{\mathcal{L}}(\mathbf{x}))$. Lorsque $\mathcal{L} = \mathcal{L}'$, cette erreur est appelée *erreur d'apprentissage*. A l'inverse, lorsque $\mathcal{L} \cap \mathcal{L}' = 0$, on parle d'*erreur de généralisation*.

Dans le cadre d'une régression, la fonction de perte la plus communément utilisée est l'erreur quadratique : $Loss(y, y') = (y - y')^2$. Cependant, il ne s'agit pas systématiquement de la mesure d'erreur la plus pertinente (cf. section II.3).

### II.2.1.b Modèle de Bayes et décomposition de l'erreur

**Définition 8** *On peut définir un modèle* $\varphi_B$ *nommé modèle de Bayes, qui correspond à un modèle optimal (théorique) :*

$$\forall(\mathbf{x}, \mathbf{y}) \in \Omega : \varphi_B(\mathbf{x}) = \arg\min_{y\in\mathcal{Y}} \mathbb{E}_{Y|X=\mathbf{x}}\{Loss(Y, y)\}$$

Dans le cadre d'un problème de régression utilisant la fonction de perte quadratique, le modèle de Bayes correspond à $\varphi_B(\mathbf{x}) = \mathbb{E}_{Y|X=\mathbf{x}}\{Y\}$.

L'erreur de ce modèle, nommée erreur résiduelle, est la plus faible possible, et correspond à un bruit existant dans les observations des éléments.

Par exemple, cette erreur correspond au fait que deux éléments puissent être définis par des variables explicatives $\mathbf{x}$ et $\mathbf{x}'$ de même valeur mais des variables $\mathbf{y}$ et $\mathbf{y}'$ cibles de deux valeurs différentes [11].

**Théorème 1** *Dans le cas où la fonction de perte correspond à une erreur quadratique, on peut montrer que :*

$$\mathbb{E}_{\mathcal{L}}\{Err(\varphi_{\mathcal{L}}(\mathbf{x}))\} = bruit(\mathbf{x}) + biais^2(\mathbf{x}) + var(\mathbf{x})$$

— *La composante Bruit : $Err(\varphi_B(\mathbf{x}))$ du modèle ne peut être évitée.*
— *La composante Biais : $(\varphi_B(\mathbf{x}) - \mathbb{E}_{\mathcal{L}}\{\varphi_{\mathcal{L}}(\mathbf{x})\})^2$ reflète le fait que le modèle ne suit pas le même processus de génération des Y que la nature (si c'est le cas, elle peut être annulée). Cette composante correspond à la précision du modèle.*
— *La composante Variance : $\mathbb{E}_{\mathcal{L}}\left((\mathbb{E}_{\mathcal{L}'}\{\varphi_{\mathcal{L}'}(\mathbf{x})\} - \varphi_{\mathcal{L}}(\mathbf{x}))^2\right)$ correspond au fait que l'ensemble des observations n'est pas disponible lors de la création du modèle. Elle correspond à la robustesse du modèle.*

**Démonstration 1** *On dispose d'un modèle $\varphi_{\mathcal{L}}$, dont l'erreur est mesurée à l'aide d'une fonction de perte quadratique $Loss(y_1, y_2) = (y_1 - y_2)^2$.*

$$\begin{aligned} Err(\varphi_{\mathcal{L}}(\mathbf{x})) &= \mathbb{E}_{Y|X=\mathbf{x}}\{(Y - \varphi_{\mathcal{L}}(\mathbf{x}))^2\} \\ &= \mathbb{E}_{Y|X=\mathbf{x}}\{(Y - \varphi_B(\mathbf{x}) + \varphi_B(\mathbf{x}) - \varphi_{\mathcal{L}}(\mathbf{x}))^2\} \\ &= \mathbb{E}_{Y|X=\mathbf{x}}\{(Y - \varphi_B(\mathbf{x}))^2\} + \mathbb{E}_{Y|X=\mathbf{x}}\{(\varphi_B(\mathbf{x}) - \varphi_{\mathcal{L}}(\mathbf{x}))^2\} \\ &\quad \hookrightarrow + \mathbb{E}_{Y|X=\mathbf{x}}\{2(Y - \varphi_B(\mathbf{x}))(\varphi_B(\mathbf{x}) - \varphi_{\mathcal{L}}(\mathbf{x}))\} \\ &= \mathbb{E}_{Y|X=\mathbf{x}}\{(Y - \varphi_B(\mathbf{x}))^2\} + \mathbb{E}_{Y|X=\mathbf{x}}\{(\varphi_B(\mathbf{x}) - \varphi_{\mathcal{L}}(\mathbf{x}))^2\} \\ &= \underbrace{Err(\varphi_B(\mathbf{x}))}_{Bruit} + \underbrace{(\varphi_B(\mathbf{x}) - \varphi_{\mathcal{L}}(\mathbf{x}))^2}_{Déviation} \end{aligned} \quad \text{(II.2.1)}$$

*(car $\mathbb{E}_{Y|X=\mathbf{x}}\{Y - \varphi_B(\mathbf{x})\} = 0$).*
*Le premier terme correspond à l'erreur (minimale) du modèle de Bayes, le second à la sous-performance du modèle $\varphi_{\mathcal{L}}$ créé à partir de l'ensemble $\mathcal{L}$.*

*Nous pouvons maintenant décomposer le terme de déviation par rapport au modèle de Bayes en Biais et variance :*

*En effet, on peut supposer que l'ensemble d'apprentissage $\mathcal{L}$ est une variable aléatoire sur les ensembles de $\Omega$, il est donc possible de calculer l'erreur du modèle $\varphi_{\mathcal{L}}$ entrainé sur l'ensemble $\mathcal{L}$ relativement au modèle de Bayes. Celle-ci peut être exprimée comme l'espérance $\mathbb{E}_{\mathcal{L}}\{\varphi_{\mathcal{L}}(\mathbf{x})\}$ du modèle entrainé sur les parties $\mathcal{L}$ de l'ensemble $\Omega$ :*

$$\begin{aligned} \mathbb{E}_{\mathcal{L}}\{(\varphi_B(\mathbf{x}) - \varphi_{\mathcal{L}}(\mathbf{x}))^2\} &= \mathbb{E}_{\mathcal{L}}\{(\varphi_B(\mathbf{x}) - \mathbb{E}_{\mathcal{L}}\{\varphi_{\mathcal{L}}(\mathbf{x})\} + \mathbb{E}_{\mathcal{L}}\{\varphi_{\mathcal{L}}(\mathbf{x})\} - \varphi_{\mathcal{L}}(\mathbf{x}))^2\} \\ &= \mathbb{E}_{\mathcal{L}}\{(\varphi_B(\mathbf{x}) - \mathbb{E}_{\mathcal{L}}\{\varphi_{\mathcal{L}}(\mathbf{x})\})^2\} + \mathbb{E}_{\mathcal{L}}\{(\mathbb{E}_{\mathcal{L}}\{\varphi_{\mathcal{L}}(\mathbf{x})\} - \varphi_{\mathcal{L}}(\mathbf{x}))^2\}\} \\ &\quad \hookrightarrow + \mathbb{E}_{\mathcal{L}}\{2(\varphi_B(\mathbf{x}) - \mathbb{E}_{\mathcal{L}}\{\varphi_{\mathcal{L}}(\mathbf{x})\})(\mathbb{E}_{\mathcal{L}}\{\varphi_{\mathcal{L}}(\mathbf{x})\} - \varphi_{\mathcal{L}}(\mathbf{x}))\} \\ &= \mathbb{E}_{\mathcal{L}}\{(\varphi_B(\mathbf{x}) - \mathbb{E}_{\mathcal{L}}\{\varphi_{\mathcal{L}}(\mathbf{x})\})^2\} + \mathbb{E}_{\mathcal{L}}\{(\mathbb{E}_{\mathcal{L}}\{\varphi_{\mathcal{L}}(\mathbf{x})\} - \varphi_{\mathcal{L}}(\mathbf{x}))^2\}\} \\ &= \underbrace{(\varphi_B(\mathbf{x}) - \mathbb{E}_{\mathcal{L}}\{\varphi_{\mathcal{L}}(\mathbf{x})\})^2}_{Biais^2} + \underbrace{\mathbb{E}_{\mathcal{L}}\{(\mathbb{E}_{\mathcal{L}}\{\varphi_{\mathcal{L}}(\mathbf{x})\} - \varphi_{\mathcal{L}}(\mathbf{x}))^2\}}_{variance} \end{aligned} \quad \text{(II.2.2)}$$

*On retrouve donc bien, grâce aux équations* (II.2.1) *et* (II.2.2) *:*

$$\mathbb{E}_{\mathcal{L}}\{Err(\varphi_{\mathcal{L}}(\mathbf{x}))\} = \underbrace{Err(\varphi_B(\mathbf{x}))}_{Bruit} + \underbrace{(\varphi_B(\mathbf{x}) - \mathbb{E}_{\mathcal{L}}\{\varphi_{\mathcal{L}}(\mathbf{x})\})^2}_{Biais^2} + \underbrace{\mathbb{E}_{\mathcal{L}}\{(\mathbb{E}_{\mathcal{L}}\{\varphi_{\mathcal{L}}(\mathbf{x})\} - \varphi_{\mathcal{L}}(\mathbf{x}))^2\}}_{variance}$$

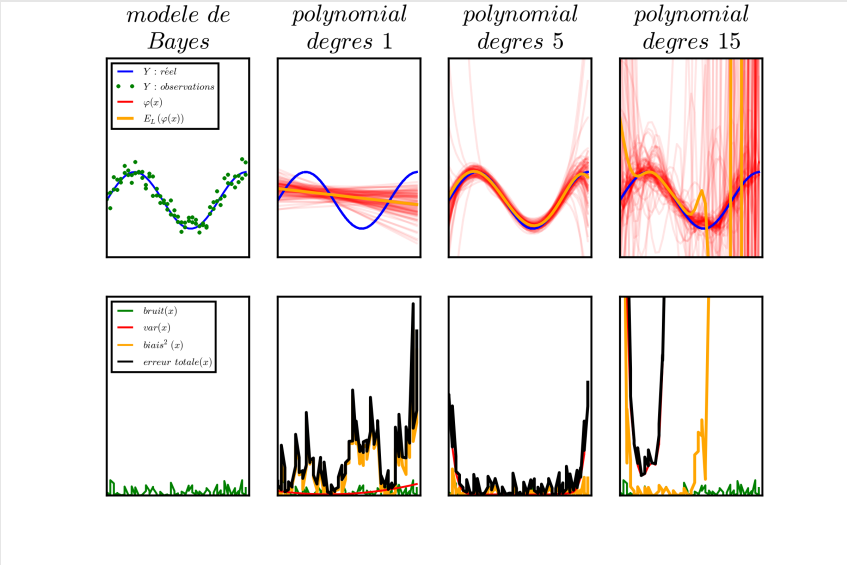



FIGURE II.2.1 – Exemple de décomposition de l'erreur

Dans cet exemple, $\Omega$ est constitué de 100 éléments définis par leur variable explicative X (axe des abscisses), des valeurs de Y (en vert) étant de la forme $\mathbf{y} = \sin(\mathbf{x}) + \epsilon$ (la courbe $\mathbf{y} = \sin(\mathbf{x})$ est représentée en bleu).
A une valeur de $\mathbf{x}$ correspondent deux points $(\mathbf{x}, \mathbf{y})$ et $(\mathbf{x}, \mathbf{y}')$.
4 modèles sont représentés :

- modèle de Bayes (à gauche), qui suit parfaitement les éléments observés
- 3 modèles polynomiaux, de complexités différentes (degrés 1, 5 et 15)

Pour construire un modèle polynomial, on ne dispose que d'un sous-ensemble $\mathcal{L}$ de 20 éléments. Suivant le sous-ensemble $\mathcal{L}$ utilisé, on construit un grand nombre de modèles $\varphi_{\mathcal{L}}$ différents (en rouge). La courbe orange représente l'espérance de ces modèles.

Sur la ligne du bas, les 3 types d'erreurs sont représentés :

- Le Bruit (en vert) : cette erreur affecte également les 3 modèles. C'est la seule erreur à affecter le modèle de Bayes.
- Le Biais (en orange) : c'est la distance entre l'espérance (sur les sous-ensembles d'éléments utilisés) des modèles produits et les observations.
- La Variance (en rouge) : c'est l'espérance de la distance quadratique entre les modèles et leurs espérances.
- Enfin, la somme de ces trois erreurs, qui correspond à l'erreur totale des modèles, est représentée en noir.

En comparant les modèles, on peut faire plusieurs remarques :

- Par définition, le modèle de Bayes est le plus performant des quatre.
- On remarque cependant qu'avec des informations supplémentaires sur le modèle réel (par exemple " il existe une relation : $Y = a + \sin(b * X) + \epsilon$ ") cela peut permettre d'annuler le biais sans faire croître la variance, et donc de converger vers le modèle de Bayes.
- L'augmentation de la complexité peut diminuer le biais, mais augmente toujours la variance. Les modèles complexes sont donc sensibles au choix (aléatoire) des observations utilisées pour les construire.

### II.2.2 Évaluation de la significativité des résultats vs. tests « out of sample »

#### II.2.2.a Test in-sample

Comme décrit ci-dessus, il est possible de mesurer l'erreur d'un modèle sur l'ensemble des éléments ayant servi à le calibrer : $Err(\varphi_{\mathcal{L}}, \mathcal{L})$. Ce type de tests est appelé *in-sample*

Dans certains cas, ce type de mesure permet d'estimer le bénéfice qu'il est possible de retirer de l'ajout d'une nouvelle variable au modèle. Ainsi, lorsque le cadre théorique du modèle permet une estimation explicite de l'impact d'une nouvelle variable, il peut être possible d'estimer si celle-ci permet d'améliorer la prédiction de manière significative (en estimant si celle-ci est liée à la variable cible, par un calcul de p-value entre deux modalités de cette variable explicative).

Cependant, cette approche, classique dans le cas de l'utilisation d'un GLM, n'est pas permise pour tous les types de modèles, et n'est certainement pas envisageable pour comparer deux modèles structurellement différents.

Il est important de noter que, le test sur les données d'apprentissage (in-sample) se réalisant sur un ensemble d'éléments $\mathcal{L}$ fixe, il ne permet pas d'estimer directement l'erreur de type Variance du modèle créé. Afin de pouvoir estimer celle-ci, il faut faire varier les ensembles sur lesquels les paramètres du modèle sont définis et testés, comme c'est le cas dans un test sur un nouvel ensemble de données (out-of-sample).

#### II.2.2.b Tests out-of-sample

**Métode** Un test out-of-sample consiste à séparer l'ensemble des éléments connus en 2 parties :

- Ensemble d'apprentissage $\mathcal{L}$ : il contient un ensemble de données à partir desquelles seront estimés les paramètres du modèle $\varphi_{\mathcal{L}}$.
- Ensemble de validation $\mathcal{L}'$ : ces données ne sont pas utilisées lors de l'estimation des paramètres du modèle. Cependant, elles sont utilisées pour estimer la qualité d'un modèle construit (sur les données de l'ensemble d'apprentissage), et donc sont exploitées pour l'estimation des méta-paramètres optimaux du modèle.

L'erreur mesurée est donc : $Err(\varphi_{\mathcal{L}}, \mathcal{L}')$.

**Principe** L'utilisation d'un ensemble de test et d'un ensemble d'apprentissage permet de mettre en évidence le phénomène de sur-apprentissage (overfitting). Ce phénomène désigne le fait de créer un modèle dont la complexité est trop grande. Il minimisera bien l'erreur d'apprentissage ($Err(\varphi_{\mathcal{L}}, \mathcal{L})$), mais devient très sensible à l'ensemble d'apprentissage $\mathcal{L}$ choisi (et, indirectement, au bruit). Son erreur de variance va donc être trop importante, ce qui le rendra inefficace sur l'ensemble des éléments de $\Omega$ (et donc, entre autres, sur un nouvel ensemble d'élément $\mathcal{L}'$ indépendant de $\mathcal{L}$).

**Utilisation d'un K-Fold** En pratique, afin de bien estimer l'impact des différents types d'erreurs, il est souhaitable de réaliser un « K-Fold » (ou validation croisée) : on décompose l'ensemble des données disponibles en $K$ ensembles, et, successivement, chacun de ces ensembles sera utilisé comme ensemble de généralisation alors que le reste ($K-1$ ensembles) des données sera utilisé en ensemble d'apprentissage.

Il est important de noter que chaque ensemble doit contenir des données indépendantes des autres (si des données sont présentes simultanément dans deux ensembles, l'usage d'une validation croisée ne permettra pas d'estimer les performance du modèle construit sur un ensemble de

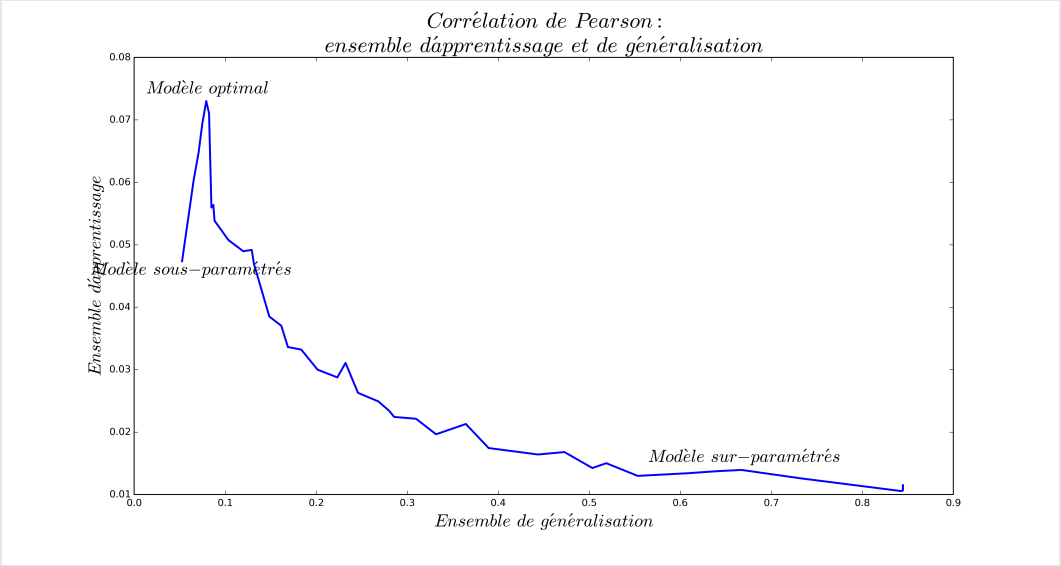



FIGURE II.2.2 – Exemple de sur-apprentissage

Le graphe ci-dessus illustre le phénomène de sur-apprentissage dans une situation réelle : on mesure l'erreur de prédiction – représentée par la corrélation entre les prédictions et les observations – de la sinistralité d'assurés MRH pour un portefeuille de 1 000 000 de contrats. Pour cela, on construit plusieurs modèles de complexité variable (les modèles utilisés sont des arbres de régression : cf. chapitre III.2.4.a ). L'erreur est mesurée sur l'ensemble $\mathcal{L}$ des points servant à calibrer le modèle ainsi que sur un autre ensemble disjoint.

On observe que, lorsque la complexité du modèle croît, l'erreur $Err(\varphi_{\mathcal{L}}, \mathcal{L})$ décroît (la corrélation sur le set d'apprentissage augmente). Cette décroissance est due au fait que l'erreur de biais du modèle décroît lorsque sa complexité augmente.
En revanche, l'erreur de généralisation $Err(\varphi_{\mathcal{L}}, \mathcal{L}')$ atteint rapidement un maximum avant de se dégrader. La première partie de ce comportement (amélioration de l'erreur) est due à une diminution de l'erreur de biais. La seconde reflète l'augmentation de l'erreur de variance.

| Type d'erreur | Evolution lorsque la complexité du modèle croît |
|---|---|
| Bruit | $\rightarrow$ |
| Biais | $\searrow$ |
| Variance | $\nearrow$ |
| Total | $\searrow$ puis $\nearrow$ |

validation). Ce type de difficultés est par exemple clairement rappelé (parmi d'autres) par P. Domingos sur les difficultés pratique de l'apprentissage automatique ([6]).

### II.2.2.c Ensemble de validation

Il est souvent conseillé de conserver un ensemble de données, dit « *ensemble de validation* », pour pouvoir tester a posteriori les choix des méta-paramètres réalisés. En effet, les méta-

paramètres du modèle créé sont optimisés sur l'ensemble des sets d'apprentissage et de généralisation (normalement dans le but de minimiser l'erreur sur ce dernier). Lors d'une mesure d'erreur de généralisation $Err(\varphi_{\mathcal{L}}, \mathcal{L}')$, le modèle $\varphi_{\mathcal{L}}$ n'est donc plus indépendant de $\mathcal{L}'$, ce qui invalide le concept d'erreur de généralisation. On peut donc créer un troisième ensemble d'éléments $\mathcal{L}''$ qui ne sera pas utilisé pour optimiser les méta-paramètres du modèle, afin de pouvoir mesurer sa qualité réelle. En pratique, il est relativement rare de voir des personnes communiquer sur des résultats obtenus sur un set de validation, dans la mesure où il est impossible de reproduire les expériences réalisées et de s'assurer que l'ensemble de validation n'a jamais été utilisé pour la construction du modèle.

# Chapitre II.3

# Fonctions de perte

Dans cette section, nous allons examiner les différentes fonction $Loss(Y_1, Y_2) : \mathcal{Y} \times \mathcal{Y} \mapsto \mathbb{R}$ permettant d'évaluer la pertinence d'un modèle $\varphi$, avec une attention spécifique sur le cas où $\varphi$ est un tarif actuariel.

La fonction *Loss* permettra de définir une erreur $Err(\varphi)$ , qui permettra de comparer deux tarifs et de déterminer le meilleur.

## II.3.1 Moindres Carrés

### II.3.1.a Mesure des moindres carrés

La méthode d'estimation des erreurs la plus répandue consiste à mesurer les carrés des erreurs commises lors de l'estimation : la qualité d'une estimation d'une valeur $\mathbf{y}$ par $\mathbf{y}'$ est alors $Loss(\mathbf{y}, \mathbf{y}') = (\mathbf{y} - \mathbf{y}')^2$.

C'est par exemple cette erreur qui a été utilisée pour la démonstration de la décomposition de l'erreur (cf. Section 1).

Minimiser cette mesure correspond à maximiser la vraisemblance des observations $\mathbf{y}$ dans le cadre d'un modèle de la forme $\mathbf{y} = \hat{\mathbf{y}} + \varepsilon$, où le bruit $\varepsilon$ suit une loi gaussienne. Cette hypothèse (selon laquelle $\hat{\mathbf{y}}$ représente la valeur "réelle" de la variable d'intérêt, et que les observations de cette variable subissent une erreur normale) est souvent contestable dans le cadre de l'estimation d'un processus non-normal (par exemple, dans le cadre du décompte d'un nombre d'évènements).

### II.3.1.b Corrélation de Pearson

La *corrélation de Pearson* (ou *coefficient de corrélation linéaire*) est définie comme le quotient de la covariance entre deux variables et la racine du produit de leurs variances :

$$\mathbf{r} = \frac{Cov(\mathbf{Y}, \mathbf{Y}')}{\sigma_{\mathbf{Y}} \times \sigma_{\mathbf{Y}'}}$$

Afin de garder une mesure d'erreur (que l'on souhaitera minimiser), on peut définir une *erreur de Pearson* : $Err_{Pearson}(\varphi) = -\mathbf{r_{Pearson}}(\mathbf{Y}, \varphi(\mathbf{X}))$, à valeur dans $[-1, 1]$.

De manière empirique, la corrélation de Pearson d'un modèle $\varphi$ avec les observations d'un

ensemble $\mathcal{L}'$ correspondra à :

$$\hat{\mathbf{r}}_{\mathbf{Pearson}}(\varphi(\mathbf{X}), \mathcal{L}') = \frac{\sum_{(x,y)\in\mathcal{L}'}(\varphi(\mathbf{x}) - \overline{\varphi(\mathbf{x})}) \times (\mathbf{y} - \overline{\mathbf{y}})}{\sqrt{\sum_{(x,y)\in\mathcal{L}'}(\varphi(\mathbf{x}) - \overline{\varphi(\mathbf{x})})^2 \times \sum_{(x,y)\in\mathcal{L}'}(\mathbf{y} - \overline{\mathbf{y}})^2}}$$

La corrélation linéaire est directement liée au modèle de régression linéaire, lui même reposant sur une minimisation des carrés des erreurs observées (dans ce cadre, on a $\mathbf{y} = \varphi(\mathbf{x}) + \varepsilon = \mathbf{A} \times \mathbf{x} + \varepsilon$, avec $\varepsilon$ l'erreur qui suit une loi Gaussienne). De même, le carré de la corrélation linéaire $\mathbf{r^2}$ représente la proportion de la variance de la variable $\mathbf{y}$ expliquée par le modèle.

L'usage d'une corrélation linéaire est donc fortement lié au modèle gaussien. Lorsque l'utilisation de celui-ci n'est pas justifiée, elle est susceptible de fournir une mauvaise représentation de la pertinence du modèle proposé.

En particulier, on peut remarquer (cf. paragraphe II.3.2) que l'utilisation de la corrélation linéaire comme mesure d'erreur est très fortement impactée par la modélisation des valeurs extrêmes, ce qui n'est pas nécessairement pertinent dans le cadre de la création d'un tarif d'assurance.

## II.3.2 Correlation de Spearman

Afin de pouvoir se dégager du modèle paramétrique gaussien, qui présuppose que le bruit présent dans les observation de la variable cible suit une loi normale $\mathcal{N}(0, \sigma)$, il peut être envisagé de comparer les rang des variables cibles prédites par le modèle $\hat{\mathbf{y}} = \varphi(\mathbf{x})$ et $\mathbf{y}$.

La *corrélation de Spearman* mesure la corrélation (de Pearson) entre les rangs de deux variables : $\mathbf{r_{Spearman}}(\mathbf{Y}, \hat{\mathbf{Y}}) = \mathbf{r_{Pearson}}(Rang(\mathbf{Y}), Rang(\hat{\mathbf{Y}}))$.

De manière empirique, la corrélation de Spearman d'un modèle $\varphi$ avec les observations d'un ensemble $\mathcal{L}'$ correspondra à :

$$\hat{\mathbf{r}}_{\mathbf{Spearman}}(\varphi(\mathbf{X}), \mathcal{L}') = \frac{\sum_{(x,y)\in\mathcal{L}'}(Rang(\varphi(\mathbf{x})) - \frac{Card(\mathcal{L}')}{2}) \times (Rang(\mathbf{y}) - \frac{Card(\mathcal{L}')}{2})}{\sum_{i=1}^{Card(\mathcal{L}')}(i - \frac{Card(\mathcal{L}')}{2})^2}$$

De même que pour la corrélation de Pearson, on peut définir une *erreur de Spearman* : $Err_{Spearman}(\varphi) = -\mathbf{r_{Spearman}}(\mathbf{Y}, \varphi(\mathbf{X}))$, à valeur dans $[-1, 1]$.

Cette corrélation permet de comparer deux variables de manière non-paramétrique, mais présente le défaut de ne pas être reliée directement à la variable que l'on cherche à estimer. En particulier, cette mesure de corrélation va permettre de distinguer un tarif qui propose des primes correctement ordonnées relativement au risque pour un grand nombre de clients, mais (à l'opposé de la corrélation de Pearson), ne sera pas sensible à la proposition de primes "absurdes" à un faible nombre de clients.

Ce comportement va tendre à favoriser les modèles sur-paramétrés au détriment des modèles simples. En effet, la dégradation de le la qualité du modèle due à l'erreur de variance va commencer à se manifester sur les valeurs extrêmes de $\hat{\mathbf{Y}}$. Sur une certaine plage de paramètres, l'augmentation de la complexité d'un modèle va donc correspondre à une diminution de l'erreur de Spearman alors que le modèle apparaitra comme sur-paramétré (ses prédictions dépendront fortement de l'ensemble d'éléments $\mathcal{L}$ utilisés pour sa création). Cette tendance au sur-paramétrage est illustrée dans la figure II.3.1.

Enfin, on rappelle que la corrélation de Spearman est conceptuellement très proche du

$\hat{\tau}_{Kendall}(\varphi)_{calL'}$ de Kendall, défini comme suit :

$$\begin{aligned}\hat{\tau}_{Kendall}(\varphi)_{calL'} =& \frac{1}{\frac{1}{2} \times Card(\mathcal{L}') \times (Card(\mathcal{L}') - 1)} \\ \hookrightarrow & \times \left(Card\left(\{((\mathbf{x},\mathbf{y}),(\mathbf{x}',\mathbf{y}'))\}_{(x,y),(x',y')\in\mathcal{L}' tq. y<y' et \varphi(x)<\varphi(x')}\right)\right. \\ \hookrightarrow & \left. -Card\left(\{((\mathbf{x},\mathbf{y}),(\mathbf{x}',\mathbf{y}'))\}_{(x,y),(x',y')\in\mathcal{L}' tq. y<y' et \varphi(x)>\varphi(x')}\right)\right)\end{aligned} \tag{II.3.1}$$

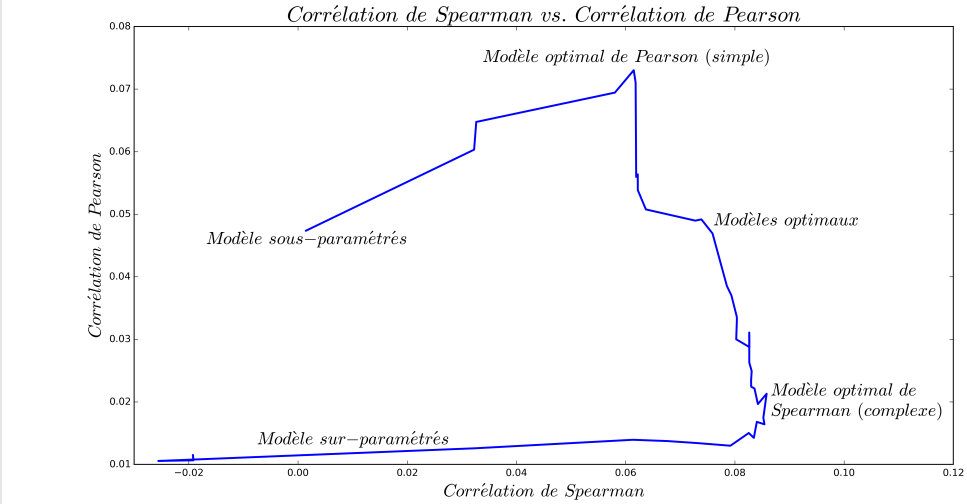



FIGURE II.3.1 – Corrélations de Spearman et Pearson

Corrélations de Spearman et Pearson, pour différents modèles dont on fait varier la complexité (données et modèles identiques à la figure II.2.2).
Ces mesures sont réalisées sur un ensemble de validation. On observe clairement une plage de modèles sur laquelle la corrélation de Spearman s'améliore alors que celle de Pearson se dégrade : celle ci correspond à des modèles de plus en plus paramétrés, et pour lesquels les prédictions des valeurs extrêmes se dégradent (augmentation de l'erreur de variance) alors que les prédictions des valeurs "normales" s'améliorent (diminution de l'erreur de biais).

### II.3.3 Vraisemblance

Une approche permettant de prendre en compte la distribution des données dans l'estimation de la qualité du modèle est de mesurer la vraisemblance d'observations, d'après celui ci.

Certains modèles supposent une distribution de l'erreur de Bruit : on suppose alors que $P((y)|\varphi(\mathbf{x}))$ suit une loi déterminée. On peut donc mesurer la vraisemblance des observations $\mathbf{y}$ :

$$Vraisemblance\,(\varphi,(\mathbf{x},\mathbf{y})) = P\,(\mathbf{y}|\varphi,\mathbf{x})\,.$$

De même, on définit la vraisemblance d'un modèle $\varphi$ sur un ensemble $\mathcal{L}'$ comme :

$$Vraisemblance_{\mathcal{L}'}(\varphi, (\mathbf{x}, \mathbf{y})) = \prod_{(x,y)\in\mathcal{L}'} P(\mathbf{y}|\varphi, \mathbf{x}).$$

Une comparaison de deux modèles est alors rendue possible, en comparant la vraisemblance d'un ensemble d'observations dans le cadre de ces deux modèles.

Il est bien entendu nécessaire, pour réaliser ce genre de comparaisons, de bénéficier d'une expression de la probabilité d'une estimation donnée pour les modèles considérés. Cette condition est remplie pour certains modèles – GLM, arbres de régression... – mais est complexe ou impossible pour d'autres – Forêts aléatoires ou autres méthodes de combinaisons de modèles, Réseaux de neurones...

### II.3.3.a Comparaison de la vraisemblance in-sample

S'il est possible d'exprimer la vraisemblance d'un ensemble d'observations données, il devient envisageable d'estimer la pertinence d'un modèle en se basant exclusivement sur un modèle d'apprentissage.

Les critères utilisés pour ce type d'estimation (Bayesian Information Criterion ou Akaike Information Criterion) reposent sur une mesure de la vraisemblance du modèle proposé pénalisée par le nombre de paramètres utilisés pour ce modèle :

$$BIC(modele) = 2 \times ln(\prod_{(\mathbf{x}_i,\mathbf{y}_i)\in\mathcal{L}} P(\mathbf{y}_i|\varphi_{\mathcal{L}}(\mathbf{x}_i)) - NbParamtres \times ln(\mathbf{card}(\mathcal{L}))$$

Cette approche permet d'estimer un nombre de paramètres optimal au sein d'une famille de modèles (par exemple d'un GLM à fonction de lien / distribution données), mais n'est pas exploitable en pratique dès que l'on souhaite comparer deux modèles de familles différentes.

### II.3.3.b Comparaison de la vraisemblance out-of-sample

Afin de comparer deux méthodes de tarification différentes, la technique la plus intuitive est sans doute la réalisation d'une mesure de la vraisemblance des observations d'un ensemble $\mathcal{L}'$ différent de l'ensemble $\mathcal{L}$ ayant servi à estimer les paramètres du modèle (ensemble de généralisation ou "out of sample") :

$$Vraisemblance_{\varphi_{\mathcal{L}},\mathcal{L}'} = \prod_{(\mathbf{x}_i,\mathbf{y}_i)\in\mathcal{L}'} P(\mathbf{y}_i|\varphi_{\mathcal{L}}(\mathbf{x}_i) \tag{II.3.2}$$

Ce type de test permet de trouver le niveau optimal de complexité du modèle de manière empirique. Cependant, les résultats trouvés dépendent du modèle choisi a priori pour estimer $P(\mathbf{y}_i|\varphi_{\mathcal{L}}(\mathbf{x}_i))$.

Ainsi, si on compare deux modèles, leurs vraisemblances (et donc leurs qualités) respectives seront différentes dans le cas d'une hypothèse de distribution d'erreur Gaussienne ou Poisson. Les conclusions tirées de ce type d'études dépendent donc directement des hypothèses choisies.

Dans le cadre d'une étude actuarielle, il serait idéalement préférable d'utiliser un estimateur non-paramétrique de la qualité de la régression proposée.

## II.3.4 Coefficient de Gini

L'utilisation d'un *critère de Gini* devient actuellement populaire grâce à ses propriétés qui pallient les limites des mesures proposées ci-dessus.

### II.3.4.a Courbe de Lorentz

L'index de Gini dérive de la *courbe de Lorentz* des sinistres.

La courbe de Lorentz est tracée en ordonnant les polices de la moins à la plus sinistrée, puis en calculant le coût cumulé des sinistres jusqu'à la $i^{me}$ police. La courbe représente la proportion du coût cumulé (axe des ordonnées) en fonction de la proportion du portefeuille nécessaire pour atteindre ce coût (axe des abscisses).

Elle relie donc les points suivants :

$$\left( \frac{i}{\mathbf{Card}(\mathcal{L})}, \frac{\sum_{jtq.\mathbf{y_j} \leqslant \mathbf{y_i}} \mathbf{y_j}}{\sum_{j \in \mathcal{L}} \mathbf{y_j}} \right)_{i \in \mathcal{L}} \tag{II.3.3}$$

avec les indexs $i$ ordonnés tels que la suite des $\mathbf{y_i}$ soit croissante.

La courbe passe donc par les points (0,0) (0% des polices et 0% des sinistres) et (1, 1) (100% des polices et 100% des sinistres). Entre les deux, elle est nécessairement convexe (la dérivée de la courbe en un point étant la sinistralité du contrat correspondant, croissant par définition – le portefeuille ayant été ordonné).

Dans une situation où tous les assurés ont subi exactement les mêmes sinistres, la courbe de Lorentz suit la diagonale $Y = X$. Dans le cas extrême inverse, ou un seul assuré représente l'ensemble des sinistres, la courbe relie les points (0,0), (1,0) puis (1,1).

De manière générale, une courbe éloignée de la diagonale (et donc proche du point (1,0) ) représente une distribution des sinistres concentrée sur quelques contrats (fréquence faible ou loi de coût à queue épaisse) quand une distribution proche de la diagonale représente une sinistralité répartie sur un grand nombre de contrats (fréquence élevée et loi de coûts à kurtosis faible).

L'aire entre la diagonale et la courbe représente donc le niveau d'inégalité dans la sinistralité : elle est comprise entre 0 et 0.5. Celle-ci est divisée par 0.5 pour obtenir le *coefficient de Gini de la sinistralité*, compris entre 0 et 1.

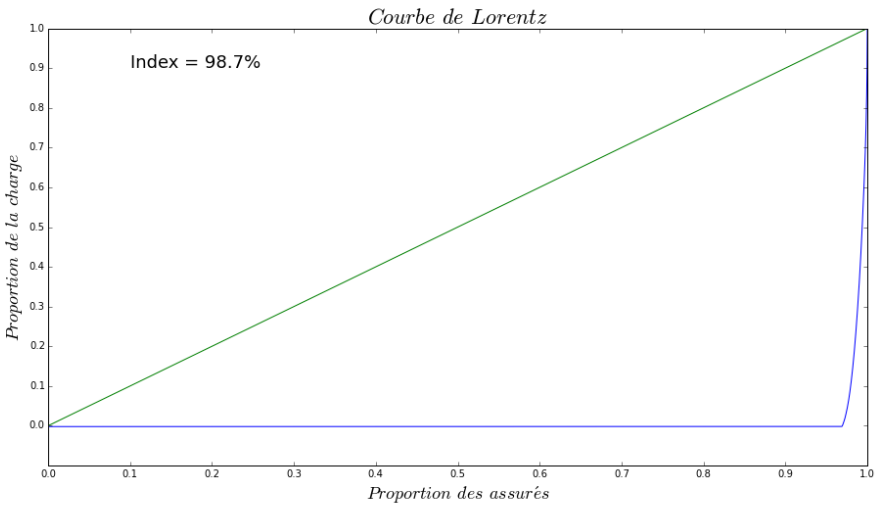



FIGURE II.3.2 – Courbe de Lorentz pour un portefeuille MRH

Courbe de Gini sur un portefeuille MRH :
95% du portefeuille n'est pas sinistré (la courbe passe par le point $(0.95, 0)$).
L'éloignement de cette courbe à la diagonale illustre le déséquilibre dans le paiement des sinistres (skewness de la distribution $\mathbf{Y}$ des montants versés pour les règlements des sinistres).

### II.3.4.b Coefficient de Gini d'un tarif

Le *coefficient de Gini* est construit de la même manière que la courbe de Lorentz, mais en ordonnant les contrats $(\mathbf{x}_i, \mathbf{y}_j)$ de la prime la moins chère à la prime la plus élevée :

$$\left( \frac{Card\left(\{j\}_{jtq.\varphi(\mathbf{x_j})\leqslant\varphi(\mathbf{x_i})}\right)}{Card(\mathcal{L}')}, \frac{\sum_{jtq.\varphi(\mathbf{x_j})\leqslant\varphi(\mathbf{x_i})} \mathbf{y_j}}{\sum_{j\in\mathcal{L}'} \mathbf{y_j}} \right)_{i\in\mathcal{L}'} \tag{II.3.4}$$

Cette courbe représente donc la quantité de sinistres correspondant aux X% des primes les plus basses du portefeuille. Plus celle-ci est loin de la diagonale, mieux les polices à risque sont identifiées lors de la tarification.

Dans le cas optimal, la courbe suit exactement la courbe de Lorentz de la sinistralité : les primes sont exactement ordonnées selon la sinistralité. A l'inverse, une courbe suivant la droite Y=X représente un tarif aléatoire : la sinistralité d'un contrat n'est pas liée au rang de sa prime.

Afin de résumer la qualité d'un tarif, on peut, de la même manière que pour la courbe de Lorentz, mesurer l'aire située entre la courbe et la diagonale. Celle-ci est comprise entre -0.5 et 0.5 (une aire négative représentant un pricing moins performant que le hasard : les plus mauvais risques sont sensiblement moins chers que la normale). Cette aire peut être divisée par la surface correspondant au tarif optimal (l'aire correspondant à la courbe de Lorentz) pour obtenir le *coefficient de Gini du tarif*, compris entre 0 et 1.

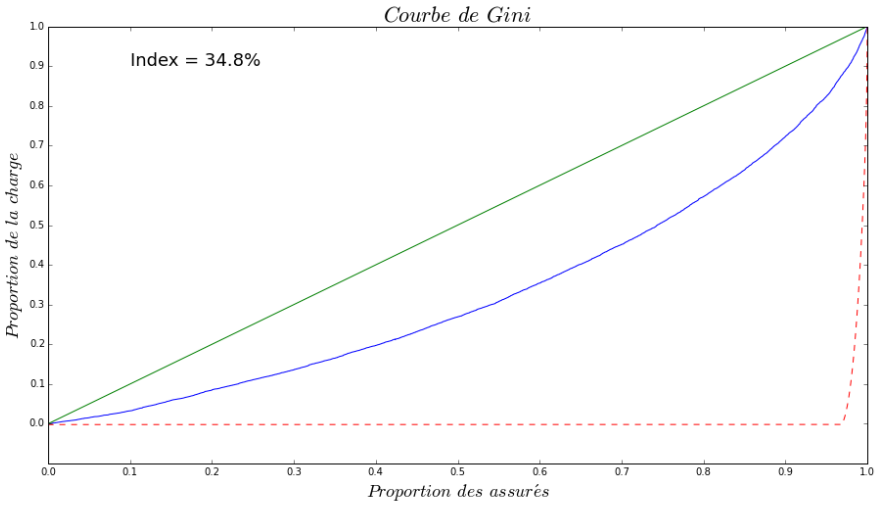



FIGURE II.3.3 – Courbe de Gini pour un tarif

Courbe de Gini pour un tarif sur un portefeuille MRH (en bleu).
La courbe d'un tarif aléatoire passerait par la diagonale (verte). Par contre le tarif optimal suivrait la courbe de Lorentz de ce portefeuille (en pointillés rouge), qui correspondrait à une prédiction parfaite de la sinistralité.

### II.3.4.c Coefficient de Gini : comparaison de tarifs

Le coefficient de Gini représentant la qualité d'un tarif, il peut être utilisé pour comparer deux stratégies [8]. Il est intéressant de noter que l'ensemble des assurés de ces deux stratégies n'est pas nécessairement le même, les coefficients ayant étés normalisés par le coefficient de Gini de leurs distributions de sinistres respectifs.

Cependant, lorsque l'on souhaite comparer deux stratégies appliquées au même ensemble de clients, il est possible de définir une courbe combinant les deux tarifs, représentant plus explicitement les différences entre ceux ci.

On dispose, pour un ensemble de contrats définis pour les éléments $(\mathbf{x}_i, \mathbf{y}_i)$ de l'ensemble $\mathcal{L}'$, de deux modèles de tarif $\varphi$ et $\varphi'$. On peut calculer, pour chaque client, une valeur nommée *r*elativité $\mathbf{R}_i = \frac{\varphi'(\mathbf{x}_i)}{\varphi(\mathbf{x}_i)}$.

La courbe de Gini sur $\mathcal{L}'$ de $\varphi'$ relativement à $\varphi$ se calcule, de manière similaire à la courbe de Gini définie sur un seul tarif au paragraphe précédent, en ordonnant les contrats $(\mathbf{x}_i, \mathbf{y}_j)$ de la relativité la plus faible à la plus élevée, et passe par les points :

$$\left( \frac{\sum_{jtq.\mathbf{R}_j \leqslant \mathbf{R}_i} \varphi(\mathbf{x_j})}{\sum_{j \in \mathcal{L}'} \varphi(\mathbf{y_j})}, \frac{\sum_{jtq.\mathbf{R}_j \leqslant \mathbf{R}_i} \mathbf{y_j}}{\sum_{j \in \mathcal{L}'} \mathbf{y_j}} \right)_{i \in \mathcal{L}'} \tag{II.3.5}$$

Pour le point $i$, cette courbe représente, sur l'axe des abscisses, la proportion des primes $\varphi(\mathbf{x}_j)$ pour tous les clients j tels que $\frac{\varphi'(\mathbf{x}_j))}{\varphi(\mathbf{x}_j))} \leqslant \frac{\varphi'(\mathbf{x}_i))}{\varphi(\mathbf{x}_i))}$, et sur l'axe des ordonnées l'ensemble des sinistres pour tous ces clients.

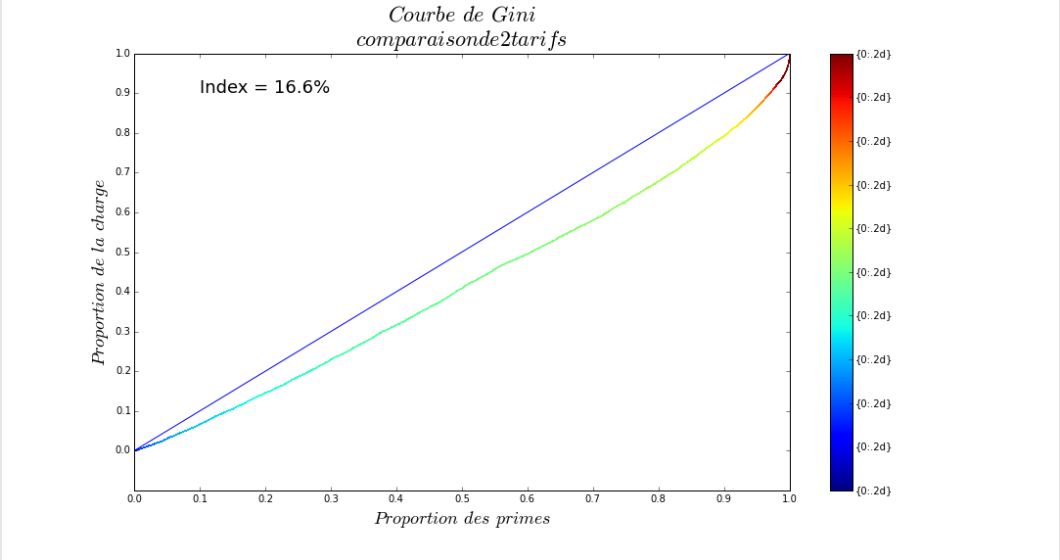



FIGURE II.3.4 – Courbe de Gini pour deux tarifs

Courbe de Gini comparant deux tarifs $\varphi$ (tarif de référence) et $\varphi'$ sur un portefeuille MRH (en bleu).
Si les deux tarifs étaient équivalents, la courbe passerait par la diagonale (bleue).
Le fait de suivre une courbe convexe indique que le tarif $\varphi'$ sur-performe le tarif de référence $\varphi$.

- Si $P_i = 1$ pour tous i, on se retrouve dans le cas d'une courbe de Gini classique, décrite plus haut. En effet, le rang des $\mathbf{R}_i$ est exactement égal à celui des $\varphi(\mathbf{x}_i)$, et la proportion des primes $\frac{\sum_{jtq.\mathbf{R}_j \leqslant \mathbf{R}_i} \varphi(\mathbf{x_j})}{\sum_{j \in \mathcal{L}'} \varphi(\mathbf{y_j})}$ correspond exactement à la proportion des clients $\frac{\sum_{jtq.\varphi(\mathbf{x_j}) \leqslant \varphi(\mathbf{x_i})} \varphi(\mathbf{x_j})}{\sum_{j \in \mathcal{L}'} \varphi(\mathbf{y_j})}$.
- Si $\varphi'$ est plus performant que $\varphi$, les clients disposant du rabais le plus important dans $\varphi'$ relativement à $\varphi$ (ie. dont la relativité est la plus faible) seront moins sinistrés, et la courbe de Gini sera située au dessous de la diagonale. Dans le cas inverse, elle sera située au dessus de la diagonale.
- Si les deux tarifs sont aussi performants l'un que l'autre, la relativité $\mathbf{R}$ sera indépendante de la sinistralité, et donc la courbe suivra la diagonale.

Cette manière de comparer deux tarifs peut sembler abstraite, mais elle peut être vue comme découlant directement d'un modèle simple de clients rationnels (cf. figure II.3.5).

Modèle simple de prospects rationnels

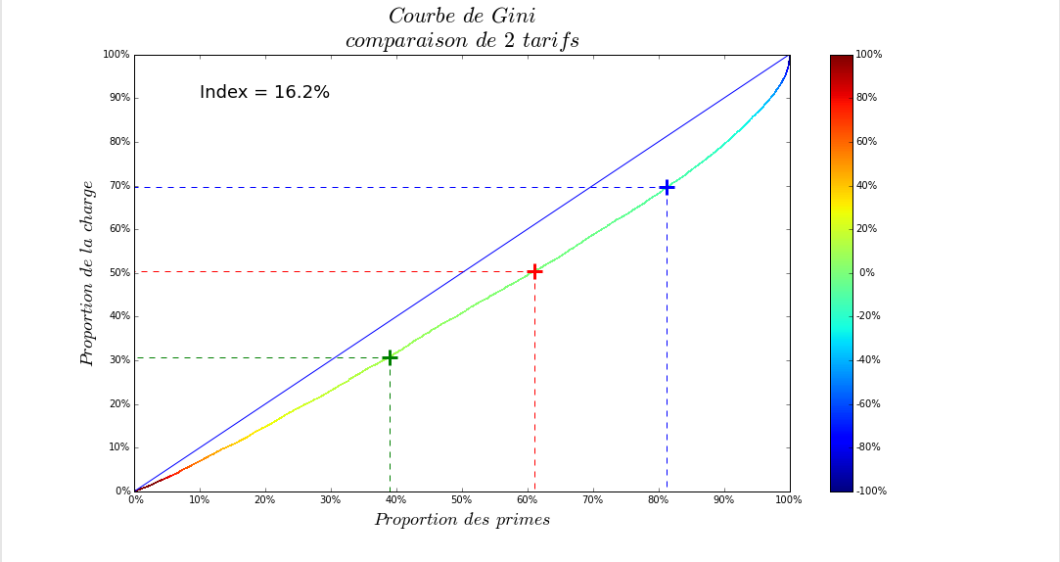



FIGURE II.3.5 – Modèle simple de prospects rationnels

Le modèle proposé repose sur les hypothèses suivantes :

- Les clients achètent toujours l'assurance la moins chère.
- L'offre est exclusivement composée de deux assureurs, $A$ et $A'$.
- L'assureur $A'$, qui propose le tarif $\varphi'$, dispose de toutes les informations sur le tarif $\varphi$ proposé par $A$.
- Les tarifs $\varphi$ et $\varphi'$ proposés aux clients sont proportionnels au risque estimé (la marge relative $M$ est identique pour tous les clients).

L'assureur $A'$ peut alors fixer le niveau de marge $M$ qu'il désire : le marché se décomposera en deux catégories :

$$\begin{cases} \{x\} \ tq. \ \varphi(x) < M \times \varphi'(x) \ : \ clients \ de \ A \\ \{x\} \ tq. \ \varphi(x) > M \times \varphi'(x) \ : \ clients \ de \ A' \end{cases}$$

Si on pose $R$ tel que $M = \frac{1}{R} - 1$, on constate que la courbe de Gini comparant $\varphi$ et $\varphi'$ correspond exactement à la part des primes (dans le tarif $\varphi$) et des sinistres acquis par l'assureur $A'$ en fonction de la marge $M$ réalisée.

Dans le graphe ci-dessus, la marge de l'assureur $A'$ est représentée par l'échelle de couleur.

- Dans les cas extrêmes, si l'assureur réalise une marge très grande (respectivement, nulle), il acquiert zéro clients (resp. l'ensemble du marché), et ne paie pas (resp. paie tous) les sinistres.
- Le point rouge indique le point d'équilibre : en appliquant directement son tarif $\varphi'$, l'assureur $A'$ conquiert 61% du marché, et doit payer 50% des sinistres.
- Le point vert correspond à une marge de 10% : l'assureur $A'$ n'obtient plus que 39% des primes, et ne règle que 31% des sinistres.
- Le point bleu correspond à une marge de $-10\%$ par rapport au tarif $\varphi'$ : l'assureur obtient donc 81% du marché et 70% des sinistres.

### II.3.5 Maximisation des profits

Comme décrit en introduction, l'une des finalités de la prime pure construite par les actuaires tarification est le calcul d'une prime commerciale.
La prime commerciale $P$ d'un contrat sera choisie pour maximiser le profit. Par exemple, dans le cas simplifié d'une maximisation du profit à un an, celui-ci s'exprime comme :

$$B_i(P) = d_i(P) \times (P - \mathbb{E}(R_i) - F_i)$$

(avec $\mathbb{E}(R_i)$ la prime pure réelle du contrat, $F_i$ les frais liés au contrat, et $d_i(P)$ la demande (probabilité de souscription) du contrat au prix $P$.
La détermination de la prime commerciale optimale revient donc à maximiser la valeur du profit $B_i(P)$ : la prime optimale ${P^*}_i$ est donc :

$${P^*}_i = Argmax_P B_i(P) = Argmax_P d_i(P) \times (P - \mathbb{E}(R_i) + F_i) \tag{II.3.6}$$

La fonction $d_i$ et la valeur de $\mathbb{E}(R_i)$ n'étant pas connues, on les remplacera par des estimations, pour obtenir une estimation du profit obtenu à un prix $P$ :

$$\hat{B}_i(P) = \hat{d}(P, X_i) \times (P - \hat{R}(X_i) - F_i) \tag{II.3.7}$$

(où $\hat{d}(P, X_i)$ est une estimation de la demande et $\hat{R}(X_i)$ est une estimation de la prime pure, en fonction des caractéristiques $X_i$ du contrat).
C'est en optimisant cette estimation $\hat{B}_i(P)$ du profit que l'on peut en pratique obtenir un prix optimal ${\hat{P}^*}_i$ :

$${P^*}_i = Argmax_P \hat{B}_i(P) = Argmax_P \hat{d}(P, X_i) \times (P - \hat{R}(X_i) - F_i) \tag{II.3.8}$$

Ce profit optimum dépend donc des caractéristiques $X_i$ du contrat, de l'estimateur $\hat{d}$ de la demande et de l'estimateur $\hat{R}$ de la prime pure utilisées.

Dans ce contexte, on peut définir une fonction d'erreur pour l'estimation $\hat{R}(X_i)$ de $\mathbb{E}(R_i)$, qui serait la baisse de profit (ou même la perte) générée par cette approximation.

Le prix et le profit optimal d'un contrat étant directement liés aux modèles de la demande $\hat{d}$ et de la prime pure $\hat{R}$, il est possible d'exprimer la performance $M$ du modèle de prime pure (fonction de perte) comme :

$$M(\hat{R}) = B_i({P^*}_i) - B_i({\hat{P}^*}_i) \tag{II.3.9}$$

Ce qui correspond à la différence de profit réalisé, entre une offre du contrat au vrai prix optimal et une offre à un prix construit à partir de notre modèle de prime pure.

L'exploitation de ce type de fonction de perte sort largement du cadre de ce mémoire, mais il est intéressant de constater qu'elle diffère largement du maximum de vraisemblance. Par exemple, elle est asymétrique : une sous-estimation de la prime pure risque de générer des pertes pour l'assureur alors qu'une sur-estimation ne lui créera "que" un manque à gagner. L'utilisation de ce type de fonctions de perte engendrerait donc une sur-estimation de la prime pure, d'autant plus grande que celle ci est incertaine, ce qui semble être un comportement sain.
Il est également notable qu'un certain nombre de phénomènes décrits dans la littérature actuarielle comme du sur-apprentissage ont probablement comme origine les limites de l'utilisation du maximum de vraisemblance comme fonction de perte.

# Troisième partie

# Algorithmes de régression

Le problème de tarification de contrats d'assurance peut être vu comme un problème de régression. On cherche donc une fonction $\varphi$, pour un assuré $i$ dont on connait les caractéristiques $\mathbf{x}_i$, une estimation $\hat{\mathbf{y}}_i = \varphi(\mathbf{x}_i)$ approximant le mieux que possible sa sinistralité.

En reprenant les définitions posées paragraphe II.2.1, on nomme $\varphi_{\mathcal{L}}$ le tarif réalisé sur l'ensemble d'apprentissage $\mathcal{L}$, et $Err(\varphi_{\mathcal{L}})$ l'erreur de ce tarif. Comme on l'a vu plus haut, l'erreur peut être mesurée de plusieurs manières (carré ou valeur absolue des erreurs, régression de Pearson ou Spearman, vraisemblance, coefficient de Gini...)

Sauf indication contraire, dans la suite de ce mémoire, la mesure d'erreur que nous chercherons à optimiser sera le coefficient de Gini (que nous chercherons donc à maximiser).

L'objectif de ce chapitre est de décrire trois algorithmes de régression, et de comparer leur performance pour la création d'un tarif sur une base de données habitation.

Il existe bien entendu un grand nombre d'algorithmes de régression. Notre choix s'est porté sur les modèles linéaires, les arbres de régression et les forêts aléatoires parce qu'ils représentaient un ensemble relativement large d'approches, permettant d'illustrer un grand nombre de concepts centraux pour l'apprentissage automatique. Cependant, d'autres méthodes existent, toutes aussi intéressantes en actuariat : par exemple les méthodes de boosting (cf. [10]), les Machines à Vecteurs de Support (ou Support Vector Machine, SVM, cf. [1]). Enfin, d'autres méthodes d'apprentissage automatiques, telles que les réseaux de neurones (dont récemment les méthodes dites d'Apprentissage Profond ou Deep-Learning, cf. [19]), ont démontré des performances remarquables sur certains problèmes, mais ne sont sans doute pas adaptées aux contraintes propres à l'estimation du risque : volumes de données relativement faibles, besoin de stabilité et de transparence des modèles...)

Le lecteur qui le souhaite pourra se reporter à des ouvrages de référence sur l'apprentissage statistique, tels que les Elements of Statistical Learning [20], ou Pattern Recognition and Machine Learning [2], dont les lectures sont vivement conseillées.

# Chapitre III.1

# Modèles Linéaires Généralisés

Les *Modèles Linéaires Généralisés* (Generalized Linear Models ou *GLM*) sont, actuellement, le type de modèle le plus répandu au sein de la communauté actuarielle pour la création de tarifs.

## III.1.1 Principe

Nous allons tout d'abord rappeler rapidement le fonctionnement des modèles linéaires généralisés.

Ces modèles étant probablement connus du lecteur, nous nous contenterons d'une description rapide de leurs hypothèses sous-jacentes. Pour plus d'information sur ce sujet, le lecteur pourra se reporter à une description approfondie des modèles par [16, 12].

### III.1.1.a Définitions

En reprenant les terminologies de la partie II, un GLM correspond à une fonction $\varphi$ permettant de relier des éléments (contrats) $\mathbf{x} = \{\mathbf{x}_1, ..., \mathbf{x}_p\} \in \mathcal{X} = \mathbb{R}^{\mathcal{P}}$ à une prime pure $\varphi(\mathbf{x}) \in \mathcal{Y} = \mathbb{R}^+$. La prime pure doit être aussi proche que possible de la sinistralité observée $\mathbf{y}$ de l'élément correspondant.

Les modèles linéaires généralisés reposent sur deux hypothèses principales :

— qu'il existe, pour chaque élément (indexé $i$), une valeur $\theta_i$ telle que la sinistralité $\mathbf{y}_i$ d'une observation suive une distribution de la forme :

$$f_{\mathbf{y}_i}(y|\theta_i) = a(\theta_i)b(y)exp\left(yQ(\theta_i)\right) \tag{III.1.1}$$

où les fonctions $a$, $b$ et $Q$ sont fixées à priori, et le réel $\theta_i$ s'exprime en fonction des variables explicatives $\mathbf{x}$ de manière linéaire : $\mu_i = \mathbf{x}_i\beta$. $\beta$, vecteur de même dimension $p$ que $\mathbf{x}_i$, est appelé paramètre du modèle.

— La moyenne $\mu_i$ de la distribution des $\mathbf{y}_i$ est liée à $\eta_i$ et donc au vecteur $\mathbf{x}_i$ des variables du contrat du client $i$ par une relation de la forme :

$$\mu_i = g^{-1}(\eta_i) = g^{-1}(\mathbf{x}_i\beta) \tag{III.1.2}$$

Où la fonction g, monotone et différentiable, est nommée fonction de lien.

**Définition 9** *Dans le cadre d'un Modèle Linéaire Généralisé, la fonction $\varphi$ prend la forme suivante :*

$$\varphi(\mathbf{x}_i) = \mathbb{E}(\mathbf{y}) = \mu_i = g^{-1}(\mathbf{x}_i\beta) \tag{III.1.3}$$

*où le paramètre $\beta$ est estimé pour maximiser, sur l'ensemble des observations, la vraisemblance de la distribution définie équation III.1.1.*

### III.1.1.b Exemples

Afin d'illustrer rapidement les concepts rappelés ci dessus, voici deux exemples simples de modèles linéaires généralisés, pour des données suivant une distribution normale (modèle linéaire simple) ou une distribution de Poisson.

**Modèle Normal** Dans le cadre d'un modèle prédictif à erreur normale, les éléments $(\mathbf{x}_i, \mathbf{y}_i)$ sont tels que la valeur $\mathbf{y}_i$ suit une loi de densité normale de moyenne $\mu_i$, fonction des variables explicatives de l'élément $\mathbf{x}_i$, et de variance $\sigma$ :

$$f_{\mathbf{y}_i}(y|\mu_i) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{-\frac{(y-\mu_i)^2}{2\sigma^2}\right\} \tag{III.1.4}$$

$$= exp\left(-\frac{\mu_i^2}{2\sigma^2}\right) exp\left(-\frac{1}{2}\left(\frac{y}{\sigma^2} - \ln\left(2\pi\sigma^2\right)\right)\right) exp\left(y\frac{\mu_i}{\sigma^2}\right) \tag{III.1.5}$$

En posant :

$$\begin{aligned} Q(\theta_i) &= \frac{\theta_i}{\sigma^2} = \frac{\mu_i}{\sigma^2} \\ a(\theta_i) &= exp\left(-\frac{\mu_i^2}{2\sigma^2}\right) \\ b(y) &= exp\left(-\frac{1}{2}\left(\frac{y}{\sigma^2} - \ln\left(2\pi\sigma^2\right)\right)\right) \end{aligned} \tag{III.1.6}$$

on retrouve bien une distribution de $y_i$ qui suit la forme décrite par l'équation III.1.1.

Dans le cas d'un modèle linéaire classique, la fonction de lien $g$, liant $\theta_i = \mathbf{x}_i\beta$ à l'espérance $\mu_i$ est donc la fonction identité.

**Modèle de Poisson** Dans ce cas, $\mathbf{y}_i$ est modélisée comme suivant une loi de Poisson de paramètre $\mu_i$ :

$$\begin{aligned} f_{\mathbf{y}_i}(y|\mu_i) &= \frac{\mu_i^y e^{-\mu_i}}{y!} \\ &= \frac{1}{exp(\mu_i)}\frac{1}{y!} exp\left(y\ln(\mu_i)\right) \end{aligned} \tag{III.1.7}$$

En posant :

$$\begin{aligned} Q(\theta_i) &= \theta_i = \ln(\mu_i) \\ a(\theta_i) &= exp\left(-exp(\theta_i)\right) = exp(-\mu_i) \\ b(y) &= \frac{1}{y!} \end{aligned} \tag{III.1.8}$$

on retrouve encore une distribution de $y_i$ qui suit la forme décrite par l'équation III.1.1.

Dans ce cas, la fonction de lien $g$, liant $\theta_i = \mathbf{x}_i\beta$ à l'espérance $\mu_i$ est donc la fonction logarithme : $\theta_i = \ln(\mu_i)$.

## III.1.2 Application

### III.1.2.a Cadre d'application

Les modèles linéaires généralisés sont employés pour l'estimation des primes pures d'assurance. Les modèles les plus fréquents sont les modèles de Poisson (pour l'estimation de fréquences) et les modèles de loi Gamma (pour les estimations de coûts).
En particulier, l'utilisation d'une fonction de lien $g$ logarithmique permet d'obtenir, pour un client $i$, un estimateur de la forme :

$$\varphi(\mathbf{x}_i) = \mu_i = e^{\mathbf{x}_i \beta} = e^{\sum_{j=1}^{p} \mathbf{x}_{ij} \beta_j} = \prod_{j=1}^{p} e^{\mathbf{x}_{ij} \times \beta_j}$$

Cette structure du modèle permet de dissocier les effets de chacune des variables, et donc de construire un modèle directement interprétable.

L'une des principales limites de ce type de modèle est le fait que l'impact des variables est linéaire (ou plus précisément log-linéaire). Afin de contourner cette limite, les variables explicatives continues ne sont pas exploitées directement dans les modèles, mais par le biais de variables indicatrices.
Par exemple, l'age du client pourrait constituer une variable (pour un client donné, $\mathbf{x}_{ij} = 40\ ans$ par exemple). Cependant, afin de ne pas se restreindre à une relation linéaire entre l'age et le risque, l'age sera représenté sous la forme d'un grand nombre de variables binaires, $\{\mathbf{x}_{age=18}, \mathbf{x}_{age=19}, ...\mathbf{x}_{age=98}, \mathbf{x}_{age=99}\}$ (si on se limite à des ages entre 18 et 99 ans). Chacune de ces variables est binaire, telle que :

$$\mathbf{x}_{i,\ age=n} = \begin{cases} 0 & \text{si} \quad age_{client\ i} \neq n \\ 1 & \text{si} \quad age_{client\ i} = n \end{cases} \tag{III.1.9}$$

Cette représentation permet de créer des modèles explicites (l'impact de chaque variable peut être représenté avec précision) et pour lesquels les relations entre les différentes variables et l'estimateur créé ne sont pas linéaires.

Les modèles linéaires généralisés sont aujourd'hui extrêmement populaires au sein de la communauté actuarielle, qui dispose de logiciels permettant de construire manuellement ce type de modèles. De plus, la plupart des infrastructures de tarification sont conçues pour exploiter ce type de modèles.

## III.1.3 Développements possibles

Il est intéressant de noter que, malgré le succès actuel de méthodes de modélisations alternatives (telles que les arbres de régression ou les forêts aléatoires, décrites ci dessous), les modèles linéaires restent un sujet de recherche très actif.

En particulier, différents algorithmes permettant un apprentissage ligne à ligne (sans avoir à manipuler simultanément l'ensemble des observations disponibles pour créer un modèle) rend ce type d'algorithmes extrêmement efficace lors de la création de modèles sur un très grand ensemble de données (par exemple, l'algorithme Vowpal Wabbit, créé par Yahoo ! Research puis Microsoft Research, ou les algorithmes implémentés dans le cadre de la librairie Lightning [7], sont remarquablement efficaces pour la création de modèles sur de très grands volumes de

données).

Un autre sujet d'étude actif sur les modèles linéaires est l'utilisation de pénalisations : au lieu de créer un modèle linéaire maximisant la vraisemblance des observations, il est possible d'introduire des critères additionnels dans l'optimisation des coefficients, ce qui permet d'obtenir des modèles bénéficiant de qualités choisies par le modélisateur (par exemple la sparcité - les coefficients non-significatifs doivent être fixés à 0, comme proposé par R. Tibshirani dans [21]) ou la régularité, pénalisant les plus grands coefficients des modèles (introduite pour la première fois par A. Tikhonov dans [22]).

# Chapitre III.2

# Arbres de régression

Dans ce chapitre, nous proposons de définir brièvement le principe des arbres de régression (ou de segmentation) [3, 9], et de présenter un algorithme d'optimisation lié à ce modèle, avant de décrire rapidement les propriétés de ce modèles et d'illustrer celui ci dans le cadre de la création d'un tarif.

## III.2.1 Principe

Un arbre de régression binaire se représente sous forme d'un arbre (informatique) : un graphe orienté dont un nœud (racine) est lié à deux nœuds fils, eux même potentiellement liés à deux nœuds fils, etc...

Un nœud ayant deux enfants est nommé nœud interne, un nœud n'ayant pas d'enfant est appelé nœud terminal ou feuille.

— Chaque nœud $N_i$ représente un sous-ensemble $\mathcal{X}_i$ de $\mathcal{X}$.
— Le nœud racine représente l'ensemble $\mathcal{X}$ de toutes les observations possibles.
— A chaque nœud interne $N_i$ est associée une "question", notée $Q_i$, fonction de $\mathcal{X}_i \to 0, 1$.
— Les sous-ensembles $\mathcal{X}_j$ et $\mathcal{X}_k$ associés aux noeuds $N_j$ et $N_k$ fils d'un noeud $N_i$ forment une bipartition de l'ensemble $\mathcal{X}_i$. Tous les éléments de $x_j$ de $\mathcal{X}_j$ vérifient $Q_i(x_j) = 0$ et tous les éléments $x_k$ de $\mathcal{X}_k$ vérifient $Q_i(x_k) = 1$.

L'ensemble des feuilles d'un arbre construisent donc une partition $\{\mathcal{X}_i\}_{i\ tq\ N_i\ terminal}$ de $\mathcal{X}$.

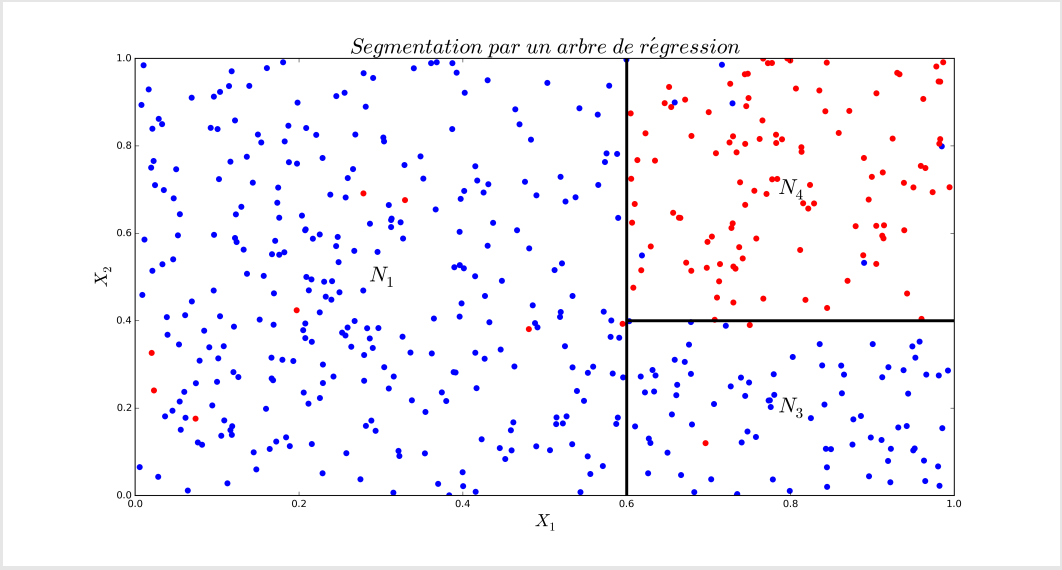
A chaque $\mathcal{X}_i$ est associée une estimation de $\hat{y}_i$ de la variable cible, à valeur dans $\mathcal{Y}$. L'estimateur $\varphi$ lié à un arbre est donc la fonction :

$$\begin{array}{rccc} \varphi & : & \mathcal{X} & \to & \mathcal{Y} \\ & & \mathbf{x} & \mapsto & \hat{y}_i \text{ avec } i \text{ tq. } \mathbf{x} \in \mathcal{X}_i \end{array}$$

Afin de rendre plus concrètes les notations proposées ci dessus, un exemple d'arbre de classification est proposé avant de proposer un algorithme permettant de créer des arbres de classification ou de régression.

Dans cet exemple, nous cherchons à classer des points à deux dimensions (nous avons $\mathcal{X} = [0,1]^2$ et $\mathcal{Y} = \{0;1\}$ ).

FIGURE III.2.1 – Exemple d'Arbre de classification



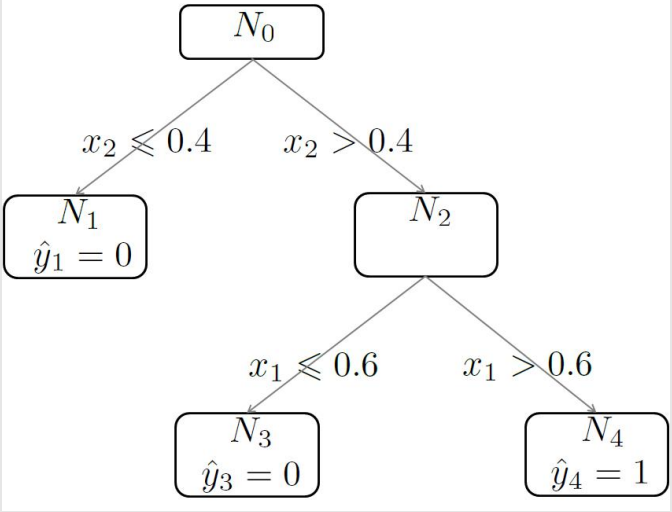

Les observations représentées sont des points de $\mathcal{X} = [0,1]^2 \times \mathcal{Y} = \{0,1\}$ ($\mathcal{Y}$ est représenté par la couleur des points, rouge pour 1 ou bleu pour 0).

Les points sont segmentés par l'arbre ci dessous (la partition créée par les feuilles est indiquées sur le graphe ci dessus). On a donc :

- $N_0$ correspond à l'ensemble des points. $Q_0$ est définie par :
  $$\forall x \in \mathcal{X} : Q_0(x) = \begin{cases} 0 & si\ x_1 \leqslant 0.6 \\ 1 & si\ x_1 > 0.6 \end{cases}$$
- $N_1$ correspond à un nœud terminal (c'est la région gauche de l'espace). Sur cette région une majorité des observations de la variable cible $y$ est à 0 (points bleus). On a donc $\hat{y}_1 = 0$.
- $N_2$ est un nœud interne. $Q_2$ est définie par : $\forall x \in \mathcal{X}_\in : Q_2(x) = \begin{cases} 0 & si\ x_2 \leqslant 0.4 \\ 1 & si\ x_2 > 0.4 \end{cases}$
- $N_3$ est un noeud terminal. On a $\hat{y}_3 = 0$
- $N_4$ est un noeud terminal. On a $\hat{y}_4 = 1$

L'arbre de classification peut être simplement représenté de la manière suivante :

FIGURE III.2.2 – Exemple d'Arbre



## III.2.2 Construction des arbres de régression

Les arbres de régression sont définis par plusieurs paramètres :

— La structure de l'arbre (combien de nœuds ont combien d'enfants). En particulier, le nombre $N$ de nœuds internes et $M$ de feuilles.

— Les $M$ estimateurs de la cible attachés à chacune des feuilles (à valeur dans $\mathcal{Y}$).

— Les $N$ "questions", fonctions de sous-ensembles de $\mathcal{X}$ vers $0, 1$.

Ces 3 points sont examinés rapidement dans cette section (avec une emphase particulière sur le choix des questions découpant l'espace défini par chaque nœud entre ses enfants), après une description brève de la manière de construire les arbres.

### III.2.2.a Création des nœuds

Les arbres de régression sont construits de manière récursive :

— On définit une fonction $\phi_0$ de l'ensemble des parties de $\Omega$ vers l'ensemble des nœuds (un nœud étant défini, comme indiqué en section III.2.1, par une question pour les nœuds internes ou par un estimateur de la variable cible pour les nœuds terminaux).

— La fonction question $Q$ associée à un nœud définit une bipartition de l'espace $\Omega_i$ associé à ce nœud. Les deux ensembles de cette bipartition peuvent être associés à deux nouveaux nœuds, construits par la fonction $\phi_0$.

L'algorithme créant un nœud et ses descendant peut donc être résumé de la manière suivante :

**function** $\phi_{noeud\ interne}(\omega_{in})$
  $Q := creeQuestion(\omega_{in})$
  $Fils_0 := \phi_0(x \in \omega_{in}\ tq\ Q(x) == 0)$
  $Fils_1 := \phi_0(x \in \omega_{in}\ tq\ Q(x) == 1)$
  **return** $\{NoeudInterne(Q), Fils_0, Fils_1\}$
**end function**

**function** $\phi_{noeud\ terminal}(\omega_{in})$
  $\hat{y}_i := creeEstimation(\omega_{in})$
  **return** $NoeudTerminal(\hat{y}_i)$
**end function**

**function** $\phi_0(\omega_{in})$
  **if** $noeudTerminal(\omega_{in})$ **then**
    **return** $\phi_{noeud\ interne}(\omega_{in})$
  **else**
    **return** $\phi_{noeud\ terminal}(\omega_{in})$
  **end if**
**end function**

Cet algorithme repose donc sur 3 fonctions :

— *creeQuestion* qui construit, à partir d'un ensemble d'éléments de $\Omega_i$, une question ( $Q : \mathcal{X}_i \to \{0, 1\}$) :

$$\begin{array}{rccc} creeQuestion & : & \mathscr{P}(\Omega) & \to & (\mathcal{X}_i \to \{0,1\}) \\ & & \mathcal{L}_i \subset \Omega_i & \mapsto & Q \end{array}$$

— *creeEstimation* qui construit, à partir d'un ensemble d'éléments de $\Omega_i$, une estimation de la variable cible (à valeur dans $\mathcal{Y}$).

$$\begin{array}{rccc} creeEstimation & : & \mathscr{P}(\Omega) & \rightarrow & \mathcal{Y} \\ & & \mathcal{L}_i \subset \Omega_i & \mapsto & \hat{y} \end{array}$$

— *noeudTerminal* qui définit, pour un ensemble d'éléments de $\Omega_i$, si le nœud construit sur cet ensemble doit être terminal ou non :

$$\begin{array}{rccc} noeudTerminal & : & \mathscr{P}(\Omega) & \rightarrow & \{0,1\} \\ & & \mathcal{L}_i \subset \Omega_i & \mapsto & estTerminal? \end{array}$$

Ces 3 fonctions, qui permettent de construire un arbre à partir d'un ensemble d'observations, vont être détaillées dans les sections suivantes.

### III.2.2.b Construction des questions

La construction des questions correspond à la définition de la fonction *creeQuestion* présentée ci dessus (chapitre III.2.2.a).

Les questions considérées dans ce mémoire sont de la forme suivante :

$$\begin{array}{rccc} Q & : & \Omega_i & \rightarrow & \{0,1\} \\ & & x & \mapsto & \begin{cases} 0 & \text{si } x_i \leq s \\ 1 & \text{si } x_i > s \end{cases} \end{array}$$

où $x_i$ représente la $i^{\text{ème}}$ dimension de $x$ ($x$ étant un vecteur à $p$ dimensions), et $s$ est une valeur seuil fixée par x.

$Q$ dépend donc des deux paramètres $i$ et $s$, et sera notée $Q_{i,s}$.

La fonction *creeQuestion* consiste donc à déterminer ces deux paramètres, à partir d'un ensemble $\mathcal{L}$ d'observations.

On remarque que l'on peut choisir, sans perte de généralité, $s$ comme étant une l'une des valeurs prises par la $i^{\text{ème}}$ dimension des observations incluses dans $\mathcal{L}$.

Le nombre de valeurs possibles de $Q_{i,s}$ est donc relativement restreint : en effet, $\mathcal{X}$ ne compte que $p$ dimensions ; $Q_{i,s}$ n'a donc au maximum que $p \times Card(\mathcal{L}')$ valeurs possibles. Il est donc possible de les examiner toutes afin de choisir la fonction $Q$ à associer aux nœuds que l'on souhaite créer.

L'objectif de l'arbre de décision est de produire (de manière itérative) un régresseur de la variable cible. Afin d'obtenir un bon régresseur, on peut réaliser, à la création de chaque nœud, une optimisation locale en optimisant la qualité du modèle simple associé au nœud. On affecte une valeur de la cible à chacune des deux parties de la bi-partition définie par $Q$, et mesure la qualité du modèle créé.

La qualité de celui ci peut être estimée de plusieurs manières, décrites dans la partie II.3 de ce rapport.

En particulier, les métriques suivantes peuvent être considérées :

— L'erreur quadratique

— Le maximum de vraisemblance (qui peut être vue comme une généralisation de l'erreur quadratique).

— L'indice de Gini

Ces 3 critères sont examinés ci dessous.

**Erreur Quadratique** L'optimisation de l'erreur quadratique (aussi appelée variance intra-classe) consiste à construire une question $Q$ divisant l'ensemble d'observations $\mathcal{L}$ liées au nœud considéré en 2 catégories telles que :

$$Variance_{intra\ classe}(Q,\mathcal{L}) = \sum_{\substack{(x_i,y_i)\in\mathcal{L}\\ tq.\ Q(x_i)=0}} (y_i-\bar{y_0})^2 + \sum_{\substack{(x_i,y_i)\in\mathcal{L}\\ tq.\ Q(x_i)=1}} (y_i-\bar{y_1})^2 \quad \text{(III.2.1)}$$

soit minimum (avec $\bar{y_0}$ et $\bar{y_1}$ les valeurs moyennes de la variable cible $y$ dans chacun des deux sous-ensembles de $\mathcal{L}$ créés par la question $Q$). La valeur de $V$ peut être calculée relativement efficacement ; il est possible de calculer celle ci pour toutes les questions $Q_{i,s}$ possibles afin de trouver le couple $(i,s)$ (et donc la fonction $Q$) minimisant V.

**Maximum de vraisemblance** Dans ce cas, on pré-suppose que les valeurs $y_i$ suivent une distribution donnée, dont le paramètre dépend de $x_i$. On cherchera donc à maximiser la vraisemblance des observations de $\mathcal{L}$ (en fixant une valeur à ce paramètre pour chacune des deux parties de la bi-partition créée par $Q$) :

$$Vraisemblance(Q,\mathcal{L}) = \prod_{(x_i,y_i)\ tq.\ Q(x_i)=0} \mathbb{P}(y_i|dist._0) \times \prod_{(x_i,y_i)\ tq.\ Q(x_i)=1} \mathbb{P}(y_i|dist._1) \quad \text{(III.2.2)}$$

où $\mathbb{P}(.|dist._0)$ correspond à la probabilité d'une observation dans la distribution associée à l'ensemble $(x,y) \in \mathcal{L}\ tq.\ Q(x) = 0)$ et $\mathbb{P}(.|dist._1)$ la probabilité d'une observation dans la distribution associée à l'ensemble $(x,y) \in \mathcal{L}\ tq.\ Q(x) = 1$.

Dans le cas d'une distribution gaussienne (dont on fixe la variance), on peut observer que les distributions sont définies par leurs moyennes, $\hat{y_0}$ et $\hat{y_1}$, et que la log-vraisemblance des observations $\log\left(Vraisemblance\left(Q,\mathcal{L}\right)\right)$ correspond à la variance intra-classe $Variance_{intra\ classe}(Q,\mathcal{L})$. Le problème du maximum de vraisemblance est donc équivalent à un problème de minimisation de variance.

D'autres distributions sont envisageables. En particulier, une loi Poisson-gamma (dont le paramètre de variance est fixé a priori) est particulièrement intéressante pour modéliser des fréquences.

**Coefficient de Gini** Dans ce cas, on cherche à maximiser le coefficient de Gini du modèle créé (cf. chapitre II.3.4).

On peut montrer (cf. démonstration 2) que, dans le cas d'une bi-partition, ce coefficient correspond à :

$$Gini(Q,\mathcal{L}) = N_0 N_1 \times \frac{\bar{y_0}-\bar{y_1}}{\bar{y}} \quad \text{(III.2.3)}$$

où :
- $N_0 = Card(\{x_i\}_{x_i\in\mathcal{L}\ tq.\ Q(x_i)=0})$
- $N_1 = Card(\{x_i\}_{x_i\in\mathcal{L}\ tq.\ Q(x_i)=1})$
- $\bar{y} = \frac{1}{N} \times \sum_{(x_i,y_i)\in calL} y_i$ (moyenne des $y_i$ de $\mathcal{L}$)
- $\bar{y_0} = \frac{1}{N_0} \times \sum_{(x_i,y_i)\in\mathcal{L}\ tq.\ Q(x_i)=0} y_i$ (moyenne des $y_i$ de $\mathcal{L}$ tels que $Q(x_i)=0$)
- $\bar{y_1} = \frac{1}{N_1} \times \sum_{(x_i,y_i)\in\mathcal{L}\ tq.\ Q(x_i)=1} y_i$ (moyenne des $y_i$ de $\mathcal{L}$ tels que $Q(x_i)=1$)

**Démonstration 2** *En plus des notations ci-dessus, on pose :*
- $N = Card(\mathcal{L})$
- $Y = \sum_{(x_i,y_i)\in calL} y_i$

— $Y_1 = \sum_{(x_i,y_i)\in\mathcal{L} \ tq. \ Q(x_i)=0} y_i = \bar{y_0} \times N_0$
— $Y_1 = \sum_{(x_i,y_i)\in\mathcal{L} \ tq. \ Q(x_i)=1} y_i = \bar{y_1} \times N_1$

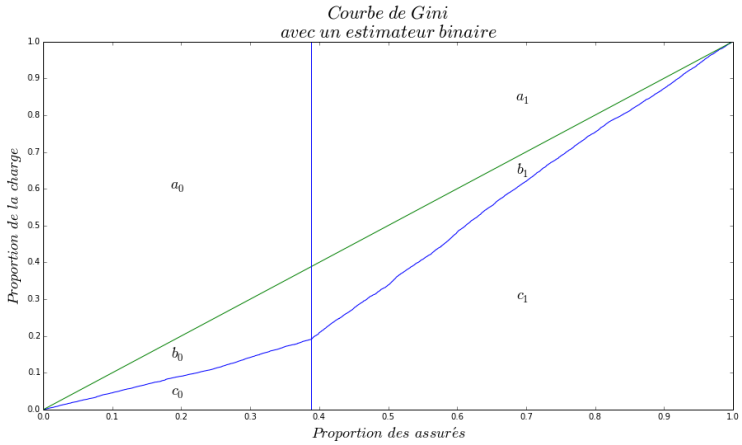



FIGURE III.2.3 – Représentation graphique du coefficient de Gini ; le coefficient de Gini (non normalisé) correspond à la somme des aires $b_0$ et $b_1$.

*Dans le graphe ci-dessus :*
— *Le coefficient de Gini (non normalisé) correspond à la somme des aires* $b_0$ *et* $b_1$.
— *L'aire* $c_0$ *vaut* $\frac{N_0 \times Y_0}{2} \times \frac{1}{N \times Y}$.
— *L'aire* $a_0$ *vaut* $(\frac{N \times Y}{2} - N_1 \times \frac{N_1}{N} \times Y) \times \frac{1}{N \times Y}$
— *Et* $a_0 + b_0 + c_0 = d_0 = \frac{Y \times N_0}{N \times Y}$.

*On a donc :*

$$\begin{aligned} 2 \times (N \times Y) \times b_0 &= 2 \times (N \times Y) \times ((Y \times N_0) - a_0 - c_0) \\ &= 2N_0(Y_0 + Y_1) - N_0 Y_0 - (N_0 + N1)(Y_0 + Y_1) + \frac{N_1^2}{N} Y \\ &= N_0 Y_1 - N_1 Y_0 - N_1 Y_1 + \frac{N_1^2}{N} Y \end{aligned}$$

*De même :*
— *L'aire* $a_1$ *vaut* $\frac{1}{2}\frac{N_1^2}{N} \times Y \times \frac{1}{N \times Y}$
— *L'aire* $c_1$ *vaut* $(Y_0 N_1 + \frac{Y_1 N_1}{2}) \times \frac{1}{N \times Y}$
— *Et* $a_1 + b_1 + c_1 = d_1 = \frac{Y \times N_1}{N \times Y}$.

*Et donc :*

$$\begin{aligned} 2 \times (N \times Y) \times b_1 &= 2 \times (N \times Y) \times ((Y \times N_1) - a_1 - c_1) \\ &= 2N_1(Y_0 + Y_1) - \frac{N_1^2}{N} Y - 2(Y_0 N_1) - Y_1 N_1 \\ &= N_1 Y_1 - \frac{N_1^2}{N} Y \end{aligned}$$

*On a donc :*

$$\begin{aligned} 2 \times (N \times Y) \times (b_1 + b_2) &= N_0 Y_1 - N_1 Y_0 - N_1 Y_1 + \frac{N_1^2}{N} Y + N_1 Y_1 - \frac{N_1^2}{N} Y \\ &= N_0 Y_1 - N_1 Y_0 \end{aligned}$$

*et donc :*

$$\begin{aligned} Gini &= 2 \times (b_0 + b_1) \\ &= \frac{N_0 Y_1 - N_1 Y_0}{N \times Y} \\ &= \frac{N_0 N_1 (\bar{y_1} - \bar{y_0})}{\bar{y}} \end{aligned}$$

### III.2.2.c Construction des estimateurs

La construction des estimateurs correspond à la définition de la fonction *creeEstimation* présentée ci dessus (chapitre III.2.2.a)

En règle générale, la moyenne de la cible sur l'ensemble d'observations attachées au nœud terminal correspond à un estimateur satisfaisant :

$\varphi(x) = \hat{y}_i = \bar{y_i}$ *avec* $i$ *tq.* $Noeud_i$ *terminal et* $x \in \mathcal{L}_i$

L'estimateur peut également être défini comme un estimateur de maximum de vraisemblance, ce qui généralise cette définition.

### III.2.2.d Structure des arbres

La structure des arbres est définie récursivement par la fonction $noeudTerminal$, qui permet de décider si un nœud est interne ou terminal (cf. chapitre III.2.2.a)

Plusieurs types de fonctions sont possibles. Par exemple :

— $noeudTerminal(calL_i) = \begin{cases} Faux & \text{si } Card(\mathcal{L}_i > N_{min\ obs} \\ Vrai & \text{si } Card(\mathcal{L}_i \leq N_{min\ obs} \end{cases}$ avec $N_{min\ obs}$ le nombre minimum d'observations nécessaire pour qu'un nœud soit interne ; ce paramètre est choisi a priori.

— $noeudTerminal(calL_i) = \begin{cases} Faux & \text{si } Prof(Noeud_i) < Profondeur\_Minimum \\ Vrai & \text{si } Prof(Noeud_i) \geq Profondeur\_Minimum \end{cases}$ avec $Prof(Noeud_i)$ la profondeur du noeud $i$ (le nombre de noeuds $Noeud_j$ de l'arbre tels que $\omega_i \subset \omega_j$) et $Profondeur\_Minimum$ la profondeur souhaitée de l'arbre.

## III.2.3 Impact des méta-paramètres

### III.2.3.a Construction des questions et métrique d'erreur

Comme présenté chapitre III.2.2.b, les arbres sont construits de manière récursive, en définissant, à la création de chaque nœud, qui optimise un critère choisi a priori.

Les 3 critères présentés correspondent à 3 mesures d'erreur (fonctions de perte) proposées dans la section II.3 (l'erreur quadratique, la vraisemblance et le coefficient de Gini).

Il est important de noter que la pertinence du choix du critère est extrêmement dure à vérifier. En effet, comme illustré ci-dessous, un critère apparaitra comme plus performant si on mesure la qualité du modèle produit avec la fonction de perte correspondante, et moins performant avec une autre fonction de perte.

Le tableau III.2.4 illustre cette difficulté.

Dans le cadre de cette étude, le critère utilisé est celui du minimum de variance. Ce choix est motivé par plusieurs raisons. En premier lieu, il est le plus fréquemment implémenté dans les librairies de machine-learning. Les résultats présentés dans ce document ont été générés grâce à des algorithmes réalisée à des fins de démonstration pour ce mémoire, mais sont moins efficaces sur de gros volumes de données que les versions fortement optimisées, disponibles par exemple dans la librairie Sk-learn.

| Critère d'optimisation | Score de Gini | Erreur Quadratique |
|---|---|---|
| Critère de Gini | 39.8 | 110 973 |
| Critère du minimum de variance | 35.1 | 107 193 |

FIGURE III.2.4 – Erreur de modèle pour l'estimation de la prime pure de contrats MRH

Dans cet exemple, deux modèles sont construits pour estimer la sinistralité d'un ensemble de contrats MRH.
Ces mesures sont réalisées sur des ensembles de validation. L'un est un arbre optimisé sur le critère du minimum de variance, l'autre sur le maximum du coefficient de Gini.
La performance de ces deux modèles est estimée en utilisant les 2 critères correspondants.
Chacun des deux arbres se révèle être efficace sur la mesure correspondant à son critère d'optimisation.

### III.2.3.b Structure de l'arbre et complexité

La structure de l'arbre peut directement être reliée à la complexité de celui ci.

En effet, le nombre de feuilles définit directement le nombre de paramètres du modèle (ou sa dimension de Vapnik-Chervonenkis).

Il est donc possible, en fixant la fonction *noeudTerminal* lors de la création d'un arbre, de définir la complexité du résultat attendu :

— nombre d'observations par feuille faible ⇔ profondeur de l'arbre élevée ⇔ grand nombre de feuilles ⇔ grande complexité
— nombre d'observations par feuille élevé ⇔ profondeur de l'arbre faible ⇔ petit nombre de feuilles ⇔ faible complexité

Dans les deux cas extrêmes :

— L'arbre n'est constitué que d'un seul nœud : prédiction constante :
$\varphi(x) = \hat{y}_0 = moyenne_{x_i, y_i \ tq. \ (x_i, y_i) \in \mathcal{L}}(y_i)$
— L'arbre est constitué d'autant de feuilles qu'il existe d'éléments $xtq.(x, y) \in \mathcal{L}$ (il s'agit du modèle de Bayes) :
$\varphi(x) = \hat{y} = moyenne_{y_i \ tq. \ (x, y_i) \in \mathcal{L}}(y_i)$

## III.2.4 Performance des arbres de régression

### III.2.4.a Biais et variance

Un modèle constitué d'un arbre de régression permet de contrôler simplement le biais, par l'intermédiaire de la fonction *noeudTerminal* définie pour sa création (cf. paragraphe III.2.3.b ) :

— Faible complexité → Faible variance et Fort biais
— Forte complexité → Forte variance et Faible biais

Une illustration de ce critère est donné figure II.2.2) : dans cet exemple, des arbres de complexité variable sont générés (en contrôlant le nombre d'éléments par feuille lors de leurs créations). Une augmentation de la complexité entraine naturellement une baisse de la variance (baisse de l'erreur d'apprentissage) ainsi qu'une hausse du biais des modèles, qui se manifeste par le phénomène de sur-apprentissage.

Cependant, malgré la simplicité du contrôle de la balance entre le biais et la variance du modèle créé, les arbres de régression restent, dans ce domaine, relativement peu performants : pour atteindre un biais donné, la variance d'un modèle basé sur un arbre de régression sera sensiblement plus importante que celle de modèles basés sur d'autres stratégies (GLM par exemple). Le chapitre suivant (III.3) propose des solutions pour remédier à cette limite.

### III.2.4.b Données d'entrées

Il est à noter que les arbres de régression sont extrêmement robustes relativement au type de données traitées par le modèle (ensemble $\mathcal{X}$ des variables explicatives). En effet, seul l'ordre des observations sur chacune de ces variables est exploité par le modèle (les questions étant de la forme $x_i \leq s$ ? pour une dimension $i$ et un seuil $s$ donné).

Ainsi, toute variable ordonnée peut être utilisée directement.

Cette simplicité permet une mise en œuvre remarquablement rapide des arbres de régression, en particulier relativement aux GLM, qui ne peuvent permettre la construction de modèles que sur un espace de variables explicatives binaires (représentant des variables catégorielles) et demandent donc un pré-traitement fastidieux des données employées.

En revanche, l'algorithme décrit ci dessus ne permet pas la prise en compte de variable catégorielle.

Celles ci peuvent être représentées par des variables binaires (de la même manière que pour la réalisation d'un modèle GLM classique), ou être prises en compte par des variations de l'algorithme présenté ci dessus, dont la présentation dépasse le cadre de ce rapport (mais on pourra se renseigner par exemple sur l'implémentation utilisée pour la librairie *rpart* du langage R).

# Chapitre III.3

# Forêts aléatoires

Comme indiqué ci dessus (III.2.4.a), les arbres de régression ne constituent pas une classe de modèles extrêmement performants.

En particulier, le fait d'obtenir un biais acceptable entrainera bien souvent la création d'un modèle extrêmement variable, dépendant fortement de l'ensemble d'observations $\mathcal{L}$ sur lequel ses paramètres ont été déterminés.

Les forêts aléatoires, présentées dans ce chapitre, représentent une solution pour palier cette faiblesse et produire des modèles à faible biais et faible variance [4].

## III.3.1 Variance des arbres de régression

La variance d'un modèle représente la sensibilité à l'ensemble d'apprentissage des prédictions qu'il produit ; formellement :

$var(\Phi) = \mathbb{E}_{\mathbf{x}\in\mathcal{X}}\left(var_{\mathcal{L}}(\varphi_{\mathcal{L}}(\mathbf{x})) = \mathbb{E}_{\mathbf{x}\in\mathcal{X}}\left(\left(\mathbb{E}_{\mathcal{L}'}\left(\varphi_{\mathcal{L}'}(\mathbf{x})\right) - \varphi_{\mathcal{L}}(\mathbf{x})\right)^2\right)\right.$

(avec le modèle $\varphi_{\mathcal{L}} = \Phi(\mathcal{L})$ le résultat de l'algorithme d'apprentissage sur l'ensemble d'observations $\mathcal{L}$).

La variances des prédictions d'un modèle, pour un $\mathbf{x}$ donné (en fonction des différents ensembles d'observations $\mathcal{L}$ sur lesquels le modèle a été entrainé) est illustré figure III.3.1.

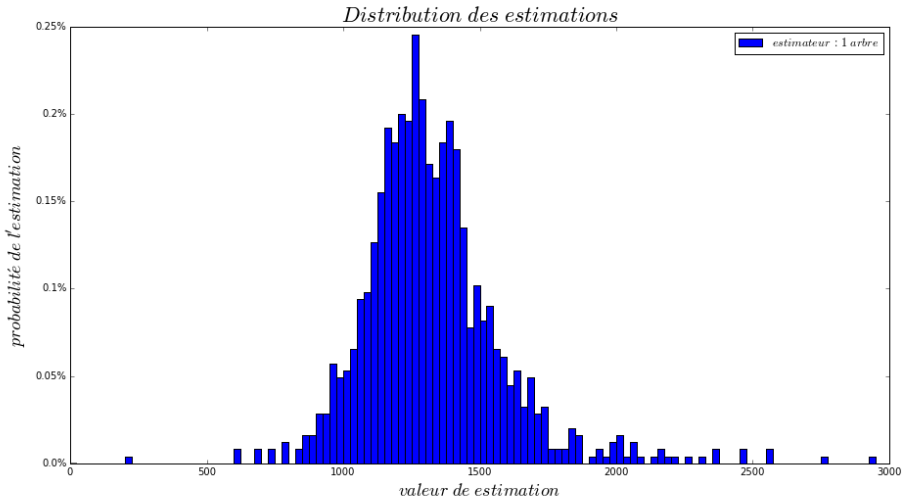



FIGURE III.3.1 – Variance des estimations d'un coût de sinistre en MRH estimé par un arbre

Illustration de la variance de modèles pour l'estimation d'un coût de sinistre en MRH.

Le graphe ci-dessus a été créé en générant, à partir de 1000 ensembles $\mathcal{L}$ d'observations différents, autant de modèles d'estimation de coût des sinistres.
Les modèles créés ont été ensuite appliqués à une observation $\mathbf{x}$. La distribution des valeurs $\hat{y} = \varphi_{\mathcal{L}}(\mathbf{x})$ est représentée par un histogramme.

On observe bien une forte variance de $\varphi_{\mathcal{L}}(\mathbf{x})$ lorsque $\mathcal{L}$ varie (ce qui représente la variance de $\Phi$).

## III.3.2 Bagging d'arbres

On peut simplement remarquer qu'il est possible (comme dans l'exemple ci-dessus), de générer plusieurs arbres visant à régresser la même variable cible $\mathbf{y}$ afin d'estimer la variance des estimateurs créés, et la valeur moyenne de ceux ci.

Cette approche correspond à une stratégie dite de *bagging* : à partir d'un ensemble $\mathcal{L}$ d'observations :

— on génère, par sampling, $n$ ensembles d'observations $\mathcal{L}_i i \in [1, n]$
— $n$ modèles $\varphi_{\mathcal{L}_i}$ sont construits à partir des observations $\mathcal{L}_i$.
— le modèle créé sera défini comme la moyenne des modèles $\varphi_{\mathcal{L}_i}$ : $\varphi_{\mathcal{L}}(\mathbf{x}) = \frac{1}{n} \sum_{i \in [1,n]} \varphi_{\mathcal{L}_i}(\mathbf{x})$

D'après le théorème central limite, si on parvient à créer $n$ modèles $\varphi_{\mathcal{L}_i}$ indépendants et identiquement distribués, la moyenne de ceux ci converge vers le modèle de Bayes plus le biais associé à l'algorithme de production $\phi$ des modèles $\varphi_{\mathcal{L}_i}$, et la variance du modèle créé tend vers 0.

La convergence de ces ensembles de modèles est illustrée figure III.3.2.

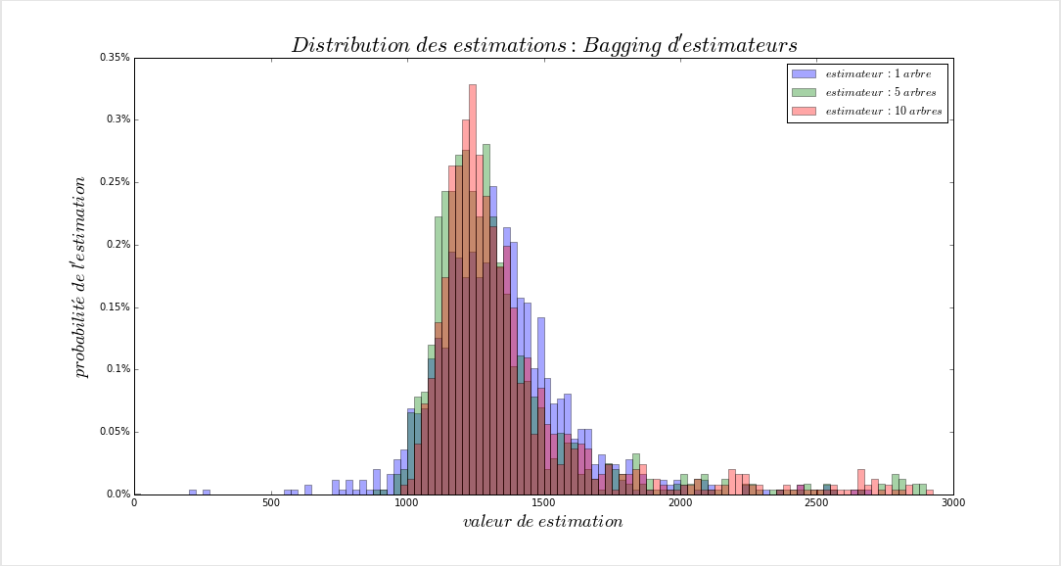



FIGURE III.3.2 – Variance des estimations d'un coût de sinistre en MRH estimé par des ensembles de 1, 5 et 10 arbres

Illustration de la variance de modèles basés sur des ensembles d'arbres pour l'estimation d'un coût de sinistre en MRH.

Le graphe ci-dessus a été créé en générant, à partir de 1 000 ensembles $\mathcal{L}$ d'observations différents, autant de modèles d'estimation de coût des sinistres.
Les modèles sont des ensembles d'arbres, créés sur des ensembles d'observations $\mathcal{L}_i$ obtenus par échantillonnage d'un ensemble d'observations $\mathcal{L}$. Les modèles créés ont été ensuite appliqués à une observation $\mathbf{x}$. La distribution des valeurs $\hat{y} = \varphi_{\mathcal{L}}(\mathbf{x})$ est représentée par un histogramme.

On observe que la variance de $\varphi_{\mathcal{L}}(\mathbf{x})$ lorsque $\mathcal{L}$ varie (ce qui représente la variance de $\Phi$) décroît lorsque le nombre d'arbres augmente.

## III.3.3 forêts aléatoires

Le concept de bagging peut être encore amélioré en renforçant l'indépendance des estimateurs créés.

En effet, le théorème central limite, qui indique que la moyenne d'un ensemble d'estimateurs sera plus stable que ceux ci pris indépendamment, présuppose une indépendance des estimateurs.

Dans le paragraphe ci-dessus, les estimateurs étaient créés en échantillonnant l'ensemble $\mathcal{L}$ des données disponibles, pour créer une série d'ensemble $\mathcal{L}_i$, à partir desquels on générait l'ensemble des estimateurs $\varphi_i$.

Afin de renforcer l'indépendance des estimateurs $\varphi_i$, il est possible de construire ceux ci, non seulement sur un sous-ensemble $\mathcal{L}_i$ des données disponibles, mais également sur un sous-ensemble des dimensions des variables explicatives $\mathcal{X}$.

L'aléa ainsi créé permet de renforcer l'indépendance entre les $\varphi_i$, et donc d'accélérer la convergence de leur moyenne (ce qui revient à diminuer la variance de celle ci).

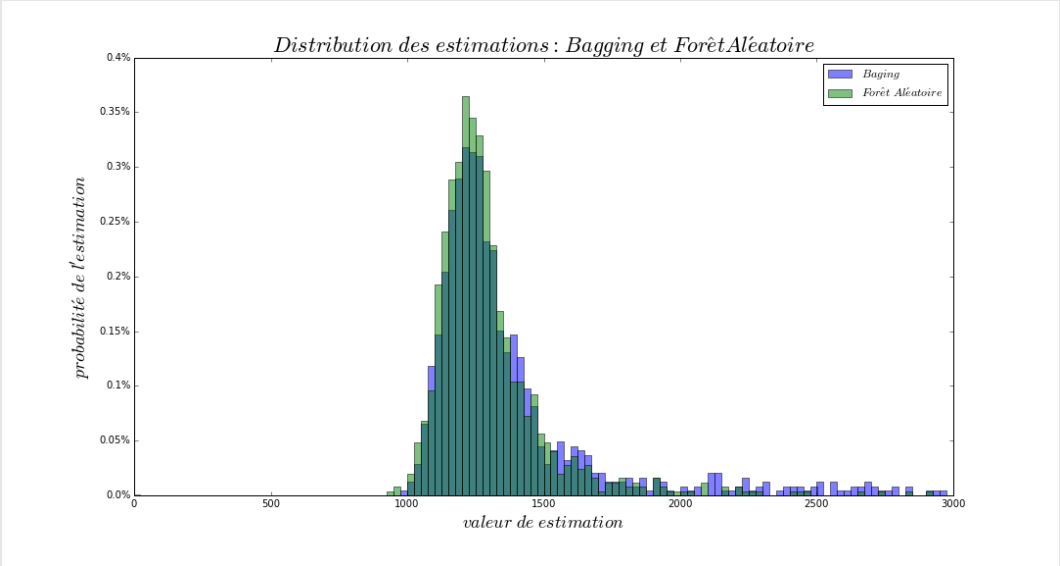



FIGURE III.3.3 – Variance des estimations d'un coût de sinistre en MRH estimé par bagging d'arbres ou forêts aléatoires

Illustration de la variance de modèles basés sur des ensembles d'arbres pour l'estimation d'un coût de sinistre en MRH.

Le graphe ci-dessus a été créé en générant, à partir de 1000 ensembles $\mathcal{L}$ d'observations différents, autant de modèles d'estimation de coût des sinistres.
Les modèles sont des ensembles d'arbres, créés sur 10 ensembles d'observations $\mathcal{L}_i$ obtenus par échantillonnage d'un ensemble d'observations $\mathcal{L}$.

- Bagging : 10 modèles sont générés à partir d'ensembles d'observations $\mathcal{L}_i$ obtenus par échantillonnage de $\mathcal{L}$.
- Random-Forest : 10 modèles sont générés à partir d'ensembles d'observations $\mathcal{L}_i$, obtenus par échantillonnage de $\mathcal{L}$ et par projection sur un sous-espace, de dimension 5, tirées aléatoirement (l'espace des variables explicatives est initialement de dimension 16 dans cet exemple).

Les modèles créés ont été ensuite appliqués à une observation $\mathbf{x}$. La distribution des valeurs $\hat{y} = \varphi_{\mathcal{L}}(\mathbf{x})$ est représentée par un histogramme.

On observe que la variance des modèles de type Forêt Aléatoire est moins élevée que la variance des ensembles d'arbres obtenus par simple échantillonnage.

# Chapitre III.4

# Comparaison des modèles

## III.4.1 Performances

Comme on peut l'imaginer à la lecture de la première partie de ce rapport, la comparaison de modèles tarifaires n'est pas aisée.

En particulier, on peut noter qu'il faut être très prudent sur la mesure d'erreur utilisée pour comparer deux modèles : par exemple, une mesure basée sur la vraisemblance des résultats obtenus favoriserait un modèle maximisant celle ci (comme un GLM), quand une mesure basée sur l'erreur quadratique favoriserait un modèle visant à la minimiser (comme la plupart des implémentations des arbres de régression).

Dans ce chapitre, les arbres de régression utilisés visent à minimiser l'erreur quadratique (comme expliqué chapitre III.2.3.a, et les GLM utiliseront une distribution précisée (loi Gamma pour les régressions de coûts ou loi de Poisson pour les régressions de fréquences).

Les erreurs des modèles produits seront mesurées à l'aide d'un coefficient de Gini, afin de ne pas suivre les distributions minimisées par les modèles présentés.

## III.4.2 Mise en œuvre et implémentation

Les données utilisées lors de la création d'un tarif peuvent être catégorisées en deux grands ensembles :

**Les données déclaratives** :
ces données sont :

- peu nombreuses : elles représentent un coût élevé d'acquisition, dans la mesure où elles doivent être explicitement demandées au client avant sa souscription.
- relativement fiables : elles engagent généralement contractuellement le client.
- bien maitrisées et retraitées : elles font l'objet d'études actuarielles depuis de nombreuses années.
- directement liées au client : elles sont fournies par un client dans le cadre d'une souscription, et sont donc directement liées à cette dernière.

**Les données externes** :
ces données sont :

— nombreuses : elles peuvent être acquises ou construites en grand nombre (nous verrons ce point plus en détail par la suite, paragraphe IV.2).
— peu fiables : ces variables peuvent être acquises gratuitement (par exemple par le biais de basses de données ouvertes, type OpenStreetMap), et, quand bien même les données en elles mêmes sont fiables, leur représentativité du client n'est pas garantie.
— mal maitrisées : ces données étant externes, les actuaires tarification ne sont pas habitués à les manipuler ; de plus, les pré-traitements subis par les données externes avant d'être acquises par une société d'assurance sont inconnus.
— indirectement liées au client : ces données sont acquises indépendamment du client et doivent donc être liées à celui ci a posteriori. Cette jointure peut se révéler extrêmement complexe.

En parallèle, une application pratique des GLM nécessite :
— Une quantification des données utilisées (si le modèle linéaire n'est pas pertinent, ce qui est souvent le cas)
— La création de variables explicites décrivant les effets croisés à prendre en compte (le modèle ne pouvant intégrer implicitement ces derniers).

Ces opération doivent aujourd'hui être réalisées manuellement (à l'aide d'outils tels que Emblem, produit par Towers-Watson).

Les modèles basés sur des modèles d'arbres, en revanche, ne nécessitent pas de réaliser de pré-traitements lourds des données utilisées.

Ces différentes qualités nous poussent donc logiquement à utiliser les GLM pour réaliser un premier modèle, basé sur les données déclaratives fournies par les clients, avant d'employer un modèle de type forêt aléatoire pour estimer les résidus de ce premier modèle à l'aide des variables géographiques collectées.

L'une des conséquences de ce choix est que nous préservons la transparence du modèle (les tarifs dépendront de manière explicite des données déclaratives et d'une carte de risque). Cette transparence est obtenue au dépens de la possibilité de créer des interactions entre les caractéristiques du client et son lieu de résidence.

La création d'un indice de risque à partir de données externes est décrite en détail dans le chapitre suivant. Cependant, avant d'exploiter des données externes, nous pouvons comparer les performances des modèles présentés sur des ensembles de données maitrisées.

### III.4.2.a Comparaison quantitative

**Base utilisée** :
Afin de valider la performance des modèles créés, nous avons testé des modèles de type linéaire généralisé ou de forêt aléatoire sur un portefeuille de clients Multi-Risques Habitation, en se concentrant sur un ensemble restreint de variables explicatives, qui correspondent aux réponses de clients à un questionnaire de souscription.

La base utilisée ainsi que les principaux traitements réalisés sont décrits en introduction (cf. I.2.3.d).

Afin de permettre une modélisation par GLM, les variables explicatives ont été retraitées manuellement :
— Les variables réelles ont été découpées en modalité, de manière à représenter les observations par des variables catégorielles.

— Des variables représentant les effets croisés (multiplication de variables binaires) ont été ajoutées

L'ensemble des données représente 900 000 contrats MRH portant sur des appartements. Environ 5% des contrats ont été sinistrés.
Pour des raisons de confidentialité, des biais ont été introduits dans la base utilisée (en réalisant un échantillonnage des polices non sinistrées et en redimensionnant la valeur des sinistres à l'aide d'une règle de trois).
Les bases en outre été légèrement retraitées afin de légèrement biaiser la performance des modèles créés pour des raisons de confidentialité.
Les coûts des sinistres ont en outre été pré-traités (écrêtage et mutualisation des sinistres graves).

**Modèles et performances** :
Quatre modèles ont été entrainés sur ces données :

— Un GLM basé sur une distribution de Poisson, pour modéliser la fréquence des sinistres.
— Un GLM basé sur une distribution Gamma, pour modéliser le coût des sinistres.
— Un GLM basé sur une loi de tweedie, pour modéliser directement le coût des sinistres (qui correspond donc à un modèle fréquence/coût).
— Une forêt aléatoire, basée sur le minimum de variance, pour modéliser la fréquence des sinistres. La complexité de ces arbres est limitée par le nombre d'éléments par feuille (1000).
— Une forêt aléatoire, basée elle aussi sur le minimum de variance, pour modéliser le coût des sinistres. La complexité de ces arbres est limitée par le nombre d'éléments par feuille (1000).

Les résultats des modèles de prédiction du coût et de la fréquence ont été multipliés afin de créer deux modèles estimant la prime pure, l'un basé sur deux forêts aléatoires, l'autre sur deux GLM.

Pour chacun des 7 modèles créés, un coefficient de Gini a été calculé sur un ensemble de généralisation (à l'aide d'un 5-fold).

| Type de modèle | Gini Fréquence | Gini coût | Gini Total |
|---|---|---|---|
| Random Forest | 31.2 | 22.0 | 38.7 |
| GLM (Poisson / Gamma) | 28.7 | 22.7 | 36.6 |
| GLM (Tweeddie) | | | 37.3 |

Les résultats du tableau présenté ci-dessus ne permettent pas clairement de classer les performances des modèles GLM et forêts aléatoires (les uns étant plus performants pour déterminer une loi de fréquence, les autres une loi de coût ; de plus, les tests ont été réalisés sur des ensembles de validation différents, ce qui ajoute du bruit à ces mesures de performance).

Cela conforte l'idée, proposée plus haut, selon laquelle les variables déclaratives doivent être exploitées à l'aide de GLMs, quand les variables externes doivent être traitées avec des forêts aléatoires.
En effet, il est important de garder à l'esprit que, si les GLM permettent de disposer de modèles aisément interprétables, le travail de pré-traitement réalisé (semi-manuellement) sur les variables utilisées pour ce test est long, et a permis de faire émerger les effets croisés adéquats.
Ce pré-traitement, économiquement pertinent sur un petit nombre de variables, ne l'est plus

lorsqu'il devient nécessaire de traiter des données nombreuses et mal connues. C'est pour cette raison que, dans la dernière section de ce mémoire, nous utiliserons des forêts aléatoires pour exploiter les données géographiques associées aux clients.

### III.4.2.b Interprétation des modèles créés

Contrairement aux GLM, les modèles forêts aléatoires (et, de manière générale, les modèles ensemblistes) ne permettent pas une interprétation directe des effets des variables explicatives.

Certaines méthodes d'analyse, comme les graphes de dépendance partielle (*Partial Dépendency Plots* en anglais) permettent d'isoler, pour une observation donnée, l'effet d'une variable, mais celui-ci sera toujours conditionné par les autres caractéristiques de l'observation considérée. En effet, l'effet d'un nœud (et de la variable qui lui est associé) dépend directement des nœuds (et donc des variables) situés en dessus et en dessous de lui.

Cependant, il est possible d'estimer l'importance d'une variable dans un arbre. Tout d'abord, il est possible de déterminer le gain de variance causé par un nœud $N$, muni d'une question $Q$ et d'un ensemble d'observations $\mathcal{L}$ (avec des notations identiques à celles de l'équation III.2.1 ) :

$$GainVariance(N, \mathcal{L}) = \sum_{(x_i,y_i)\in\mathcal{L}} (y_i - \bar{y})^2 - \Big( \sum_{\substack{(x_i,y_i)\in\mathcal{L} \\ tq.\ Q(x_i)=0}} (y_i - \bar{y_0})^2 + \sum_{\substack{(x_i,y_i)\in\mathcal{L} \\ tq.\ Q(x_i)=1}} (y_i - \bar{y_1})^2 \Big) \tag{III.4.1}$$

(avec $\bar{y}$ la moyenne de la variable cible $y$ sur les observations de $\mathcal{L}$), et $\bar{y_0}$ (resp. $\bar{y_1}$) la moyenne de la variable cible $y$ sur les observations $(x, y)$ telles que $Q(x) = 0$ (resp. $Q(x) = 1$).
Cette définition peut être associée à la variable attachée au nœud $N$, et il devient donc possible d'estimer le gain de variance généré par cette variable (en sommant les gains de variance des nœuds auxquels elle est attachée), soit au sein d'un arbre, soit dans l'ensemble d'une forêt aléatoire.
Une analyse approfondie de l'importance de variables dans des modèles basés sur des arbres de régression (ainsi que des intéractions entre variables) peut être trouvée dans [15]. Bien entendu, cette définition peut être étendue à d'autres mesures de vraisemblance.

Cette méthode d'évaluation de l'importance des variables ne permet pas d'estimer l'impact exact d'une variable (par exemple, il n'est pas possible de déterminer si une variable a un effet positif ou négatif sur la valeur de la cible). Cependant, elle permet d'estimer quelles sont les variables utilisées par le modèle, et quelles variables sont "non-significatives".
Cette mesure d'importance des variables explicatives sera utilisée dans la partie suivante de ce mémoire.

Il est important de noter que, comme toute tentative d'expliciter un modèle prédictif, les mesures d'importance des variables souffrent de ce que L. Breiman appelle "l'effet Rashomon" (cf. [5]). Comme il existe plusieurs arbres produisant des prédictions de qualité équivalente, il existe différentes importances de variables liées à un problème, et donc différentes explications possibles sur le modèle créé. Ce constat pousse la personne construisant les modèles à interpréter les résultats de ceux ci avec prudence, mais n'annule pas - bien au contraire - l'intérêt d'outils d'analyse des modèles créés.

## III.4.3 Impact tarifaire de la méthode proposée

### III.4.3.a Impact sur le loss-ratio

**Evolution de la prime moyenne proposée** :

On peut observer sur la figure III.4.1, que les trois modèles (GLM, forêt aléatoire sur la prime pure, ou forêts aléatoires estimant la fréquence et le coût des sinistres) prédisent le même risque moyen sur nos clients. Il est intéressant de noter que l'estimateur du maximum de vraisemblance peut produire des modèles en moyenne (faiblement) biaisés, mais que ce biais n'est pas sensible en pratique. Un estimateur utilisant un maximum de vraisemblance sur une hypothèse de distribution gaussienne fournit, en revanche, des prédictions en moyenne non-biaisées.
Une autre source de biais dans la moyenne des prédictions est l'utilisation d'un modèle de type fréquence-coût. En effet, comme rappelé section 0.2.2, ces modèles reposent sur une indépendance entre le nombre de sinistres subis par un client et les coûts de ceux-ci. Si cette indépendance n'est pas respectée, un modèle de fréquence-coût produira des prédictions biaisées.
Si l'actuaire souhaite, en moyenne, obtenir des informations non-biaisées – comme c'est souvent le cas - il est probablement souhaitable de redimensionner les prédictions en appliquant une règle de trois. L'impact de ce retraitement sera mineur dans le cas des données que nous considérons ici, l'effet de la méthode choisie sur la moyenne prédite étant très faible.

| Type de modèle | Forêt Aléatoire Frequence coût | Tweedie | Forêt Aléatoire |
|---|---|---|---|
| Charge moyenne | 34.37 | 34.29 | 34.66 |

FIGURE III.4.1 – Prime pure moyenne - dégât des eaux

Exemple de l'impact du modèle choisi sur la prime moyenne prédite, sur une base de validation.
La charge moyenne observée sur la base est de 34.52 euro.
La base a été biaisée (en retirant des police non sinistrées) pour des raisons de confidentialité.
Bien que les données sur lesquelles ces prédictions moyennes sont mesurées soient différentes de celles utilisées pour réaliser l'apprentissage des modèles, les prédictions sont remarquablement stables.

**Utilisation du coefficient de Gini relatif** :

L'impact de la méthodologie choisie aura en revanche un fort impact sur l'ordre des prédictions réalisées (ce qui est souvent appelé, de manière abusive, la segmentation des clients).
Les nouvelles prédictions peuvent donc avoir un très fort impact sur le loss-ratio enregistré par la société d'assurance les exploitant. La traduction pratique d'une amélioration d'un modèle (mesurée en terme de vraisemblance ou de coefficient de Gini) en loss-ratio ou en profit est un sujet ouvert en actuariat, et l'objet d'enjeux importants : il est actuellement extrêmement dur d'estimer en pratique le retour sur investissement du travail des équipes de tarification des assureurs, ou de l'achat de nouvelles bases de données améliorant l'estimation du risque des clients.
La seule manière claire d'estimer le gain généré par un nouveau tarif est donc l'utilisation de

tests A-B (utilisation aléatoire, en parallèle, de deux tarifs afin de comparer celui générant le meilleur loss-ratio). Mais une telle méthode est chère (elle génère un fort coût d'opportunité en n'appliquant pas le nouveau à l'ensemble des clients potentiels) et lente (il faut attendre un temps long avant d'obtenir des résultats concrets sur les sinistres des clients.

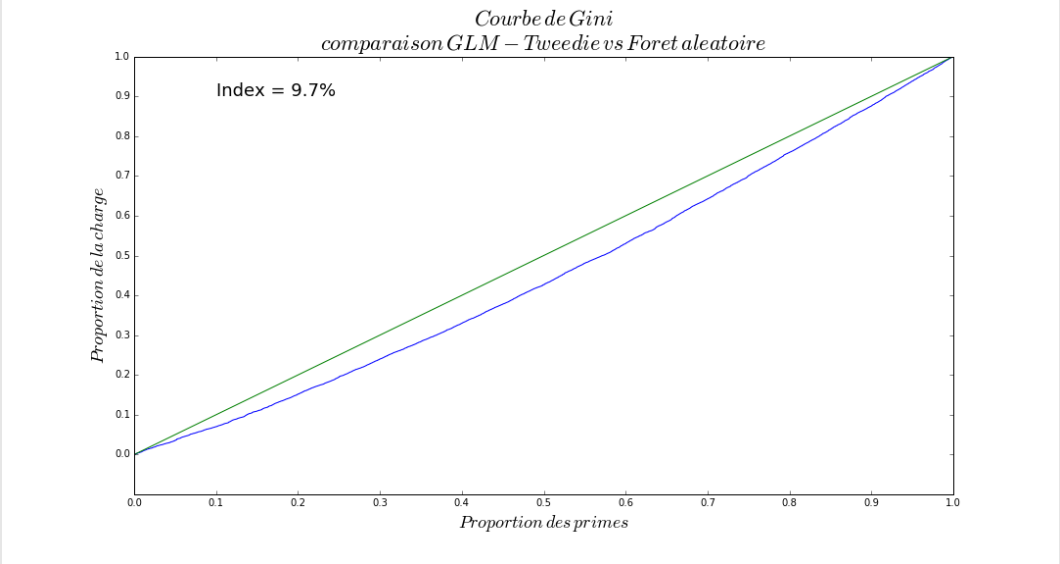



FIGURE III.4.2 – Gini relatif d'un tarif construit sur une forêt aléatoire contre un GLM utilisant une loi de Tweedie.

L'indice de Gini permet, comme décrit paragraphe II.3.4.c, de comparer deux stratégies tarifaires.
Dans notre exemple, un tarif construit sur la base d'une forêt aléatoire est comparé à un tarif construit grace à un GLM utilisant une loi de Tweedie.

On observe que, sur un marché "parfait" (dont les hypothèses sont décrites plus haut), la première stratégie permet d'identifier et de capturer en leur proposant un prix plus attractif 50% des clients qui ne représentent que 42% des primes.
Le basculement d'une stratégie à l'autre permet donc un gain théorique très sensible : sous les hypothèses décrites paragraphe IV.4.7, l'amélioration du loss-ratio serait de 16% au prix bien sûr d'une perte considérable du volume d'affaires.

La différence d'impact entre les chiffres de la table III.4.2.a (gains de gini de 1.4% et le gain de loss-ratio est frappante : aucun de ces deux indicateurs ne prédisent de manière définitive la performance du modèle dans un environnement réel.

La méthode reposant sur des bases de données historiques la plus aboutie sur ce sujet est probablement la comparaison de Gini définie paragraphe II.3.4.c. Il est important de garder en mémoire que cette méthode :

— Repose sur des hypothèses extrêmement fortes (et peu réalistes) sur la compétition et le comportement du client.
— Suppose que les marges proposées par les assureurs sont multiplicatives (proportionnelles

à la prime pure) : il n'y a donc aucune stratégie d'optimisation des marges envisagée dans cette comparaison, alors que de telles stratégies sont devenues la norme sur les marchés matures.
— Se concentre sur des profits à court terme, en n'étudiant pas l'impact long terme des décisions tarifaires.

Malgré ces limites, l'analyse permet de justifier au moins un majorant du gain de loss-ratio. Comme illustré dans la figure IV.4.7, dans le cas de l'utilisation d'une forêt aléatoire au lieu d'une loi de Tweedie, le gain serait impressionnant (16%) en particulier comparé au faible gain en terme de Gini (passant de $37,3\%$ à $38,7\%$).

### III.4.3.b Evolution des primes pures

L'estimation des risques basée sur des forêts aléatoires permet, comme illustré dans la section III.4.2.a, de prédire de manière plus précise les risques. Cependant, il est intéressant d'estimer les types de polices qui sont impactées par le choix d'une méthode ou d'une autre.

Il est tout d'abord important de noter que les différentes méthodes ne créent pas de "profils" particuliers aisément identifiables qui bénéficieraient d'une méthodologie ou d'une autre. En effet, les GLM comme les forêts aléatoires garantissent que les primes pures calculées sont, de manière univariée, bien alignées avec les risques observés. Il n'existe donc pas de "segments" ou "catégories" explicites de clients (par exemple "les jeunes conducteurs urbains") sur lesquelles les modèles ont un effet ou un autre ; les effets des variables sur la prime pure prédite sont en revanche progressifs et non tranchés.
Il est par contre possible de comparer, pour l'ensemble de notre portefeuille, les primes pures de nos clients d'après les différentes méthodologies.
Dans l'ensemble, les prédictions des deux modèles sont (heureusement) relativement cohérentes ; la corrélation entre les deux prédictions est de 87%. Une incohérence entre ces deux modèles aurait été un signe d'alarme fort sur la fiabilité de l'une ou l'autre des méthodes.
Cependant, comme illustré figure IV.4.8, la corrélation forte sur la majorité des clients ne doit pas masquer de fortes disparités sur certaines polices. En particulier sur les risques élevés, les différences de prime pure estimée peuvent être très importantes entre les deux méthodes.

Une autre manière de résumer le changement des primes pures est d'insister sur le changement de « mix » du portefeuille généré par le changement de tarification : certain clients vont bénéficier du nouveau tarif, et affluer, d'autres vont en pâtir et cesser de souscrire. Il y aura donc des "gagnants" et des "perdants" parmi les clients. Ce changement de mix est important à deux titres :
— Comme indiqué plus haut, il peut refléter une anti-sélection et donc, indirectement, un sur-apprentissage des modèles (les nouveaux profils bénéficiant de prix exagérément attractifs)
— Dans le cas d'un marché intermédié, il peut signifier que certains agents vont voir leurs chiffres d'affaires augmenter et d'autres baisser.

Le second point est moins directement "quantitatif" mais ne doit pas être négligé : les gains et pertes parmi les clients – qui ne sont pas captifs – peuvent résulter en une meilleur sélection de ceux-ci. Mais les gains et les pertes parmi les agents – qui sont captifs – créeront bien entendu de graves problèmes de relations avec les agents lésés, qui seraient probablement évités en empêchant la mise en place directe d'un tarif modifié. Si, de plus, le tarif est difficilement interprétable, comme c'est le cas d'une forêt aléatoire - cf. paragraphe III.4.2.b - cet élément devient réellement rédhibitoire.
Nous reviendrons sur ce point en conclusion de ce mémoire.

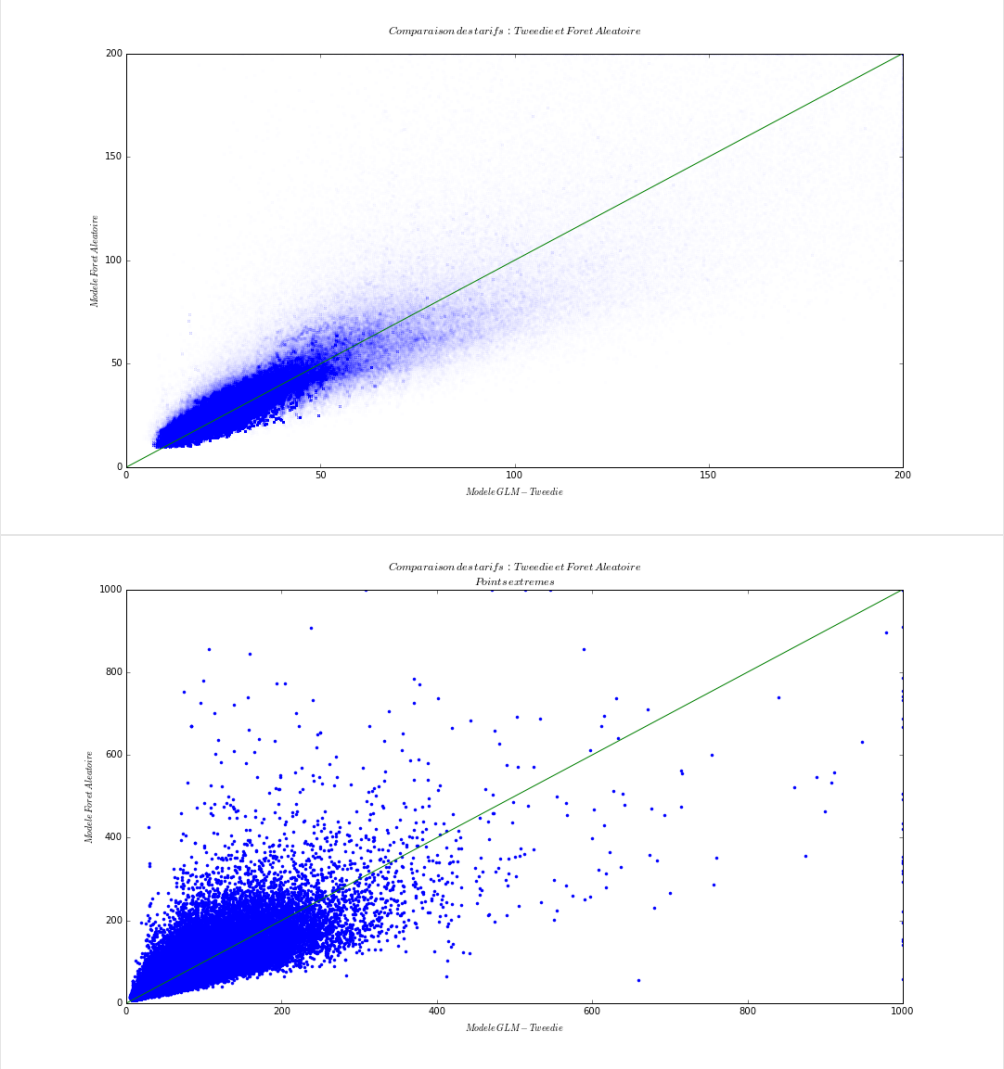



FIGURE III.4.3 – Comparaison des tarifs produits par une forêt aléatoire et par un GLM (utilisant une loi de Tweedie).

Chaque point correspond à une police ; l'axe des abscisses correspondant à la prime pure modélisée à l'aide d'un GLM, l'axe des ordonnées son estimation à l'aide d'une forêt aléatoire. Le graphe du haut couvre la très grande majorité des observations, quand le graphe du bas met en valeur les primes les plus extrêmes (noter la différence d'échelle entre les deux graphes).

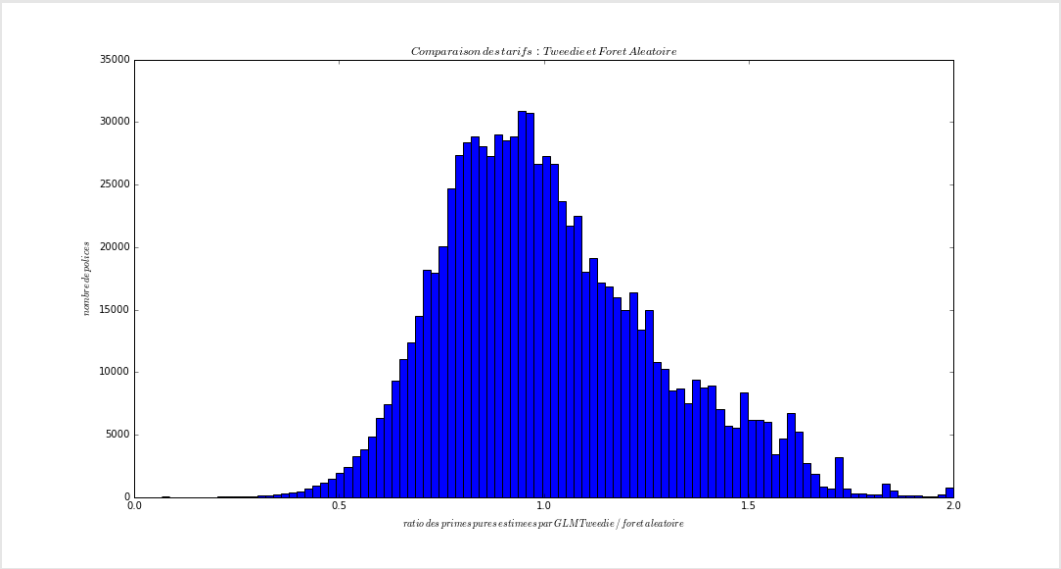



FIGURE III.4.4 – Ratio entre les primes des deux modèles créés.

Illustration de la variance de modèles pour l'estimation d'un coût de sinistre en MRH.

Différences de primes pures proposées, en se basant sur un modèle de Tweedie ou une forêt aléatoire. L'axe des abscisses représente le changement de tarif entre les deux modèles (une valeur de 1 représente une prime pure inchangée), l'axe des ordonnées représente le nombre de polices correspondant à ce changement.
Des tarifs ont été créés sur une base d'apprentissage et appliqués à un ensemble de polices. Ces illustrations démontrent la variabilité des primes pures créées, en fonction de la méthode utilisée : 20% des primes pures varient de plus d'un quart de leur valeur, alors que la variation médiane correspond à une - légère - baisse et que la prime pure moyenne est inchangée.
On observe que les risques estimés sont globalement comparables (comme illustré dans la figure de gauche IV.4.8). Cependant certaines polices voient leurs risques estimés de manière très différente selon la méthode employée.
Cette étude a été réalisée sur une base de sinistres dégât des eaux (décrite paragraphe III.4.2.a).

Quatrième partie

# Utilisation de variables géographiques

# Chapitre IV.1

# Création d'un micro-zonier

## IV.1.1 Introduction

Les principales variables tarifaires sont, historiquement, les réponses à un questionnaire posé à l'assuré lors de sa souscription.
Les questions sont conçues pour que les données (déclaratives) obtenues puissent discriminer aussi bien que possible le risque représenté par l'assuré. Ce questionnaire, qui permet d'obtenir simplement un grand nombre d'informations sur les assurés, a un coût élevé. En effet, lors de l'acquisition de clients sur Internet, la complétion d'un formulaire (souvent long et fastidieux) représente un frein important à l'acquisition de nouveaux clients (une proportion très importante de ceux ci se décourageant et quittant le processus d'acquisition avant d'obtenir un devis).
Dans ce contexte, l'obtention de données sur l'assuré à partir de sources extérieures qui peuvent lui être rattachées devient un enjeu majeur pour les assureurs.
Les données issues de bases géographiques, qui peuvent être naturellement attachées à un assuré par le biais de son adresse, représentent donc actuellement une opportunité importante pour les actuaires en tarification.
Nous décrirons dans ce chapitre une méthode possible pour exploiter ce type de données.

## IV.1.2 Principe du micro-zonier

La méthodologie de prédiction de la sinistralité proposée se décompose en plusieurs étapes :

- Création d'un modèle déclaratif : avec les données internes issues du formulaire de souscription, on prédit le coût et la fréquence de sinistre pour la garantie étudiée, en utilisant deux GLMs. Cela permet d'estimer les facteurs de risque non spatiaux. Les résidus des GLM obtenus vont constituer la cible du modèle géographique.
- Géo-localisation des clients et création des variables explicatives géographiques : les anciens clients sont géo-localisés (on calcule, pour chacun d'eux, à partir de son adresse, la position géographique exacte de son risque). Grace à cette information, il est possible d'associer à chaque client de nouvelles variables, liées à sa position géographique.
- Prédiction de la composante géographique du risque : les résidus des GLM construits lors de la première étape sont estimés à l'aide de variables explicatives liées à la position géographique de la police. Cette étape permet d'obtenir un modèle géographique du risque, associant une estimation de celui ci à chaque point de l'espace.
- Construction du micro-zonier : a partir des prédictions réalisées, on construit une carte du

risque sur l'ensemble du territoire. Cette carte est une partition de l'espace en un ensemble de polygones, auxquels est associé un niveau de risque. C'est cette carte qui sera utilisée lors du processus de tarification (pour des raisons de performance décrites plus bas).

— Intégration du zonier dans la structure tarifaire : le risque associé à une police est estimé en combinant l'estimation créée sur la base de ses données déclaratives et l'estimation géographique de son risque. Le processus (IT) de souscription doit donc intégrer le GLM créé lors de la première étape de la création du tarif, ainsi que la carte créée lors de la dernière étape.

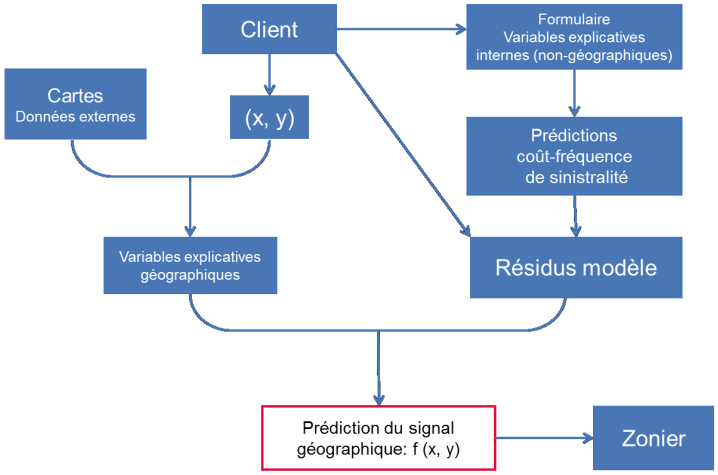



FIGURE IV.1.1 – Processus de création du zonier.

La suite du mémoire suit le processus décrit ci dessus :

— La méthodologie d'estimation du risque basée sur les informations déclaratives du client, ainsi que l'extraction des résidus de celle-ci sont classiques et ne seront pas détaillées dans ce mémoire.

— De même, les technologies de géo-localisation sortent du cadre d'un travail actuariel et ne seront décrites que très brièvement.

— Dans le chapitre suivant, nous nous concentrerons sur la création de variables explicatives à partir d'information géographique. La construction de ces variables représente la plus grande partie du travail de création du zonier.

— Nous étudierons ensuite les différentes propriétés des variables créées, et l'impact de celles ci sur la performance des prédictions réalisées à l'aide des modèles proposés lors de la seconde partie de ce mémoire, avant d'estimer la qualité effective des résultats obtenus.

— Enfin, nous proposerons une implémentation pratique efficace du modèle réalisé.

# Chapitre IV.2

# Création de variables géographiques

## IV.2.1 Sources de données

Différentes sources de données géographiques sont envisageables pour la création d'un tarif. Celles ci peuvent être internes à l'entreprise, ou externes. Dans le cas des données externes, elles peuvent être libres (et gratuites) ou propriétaires (et payantes).

### IV.2.1.a Données externes libres

#### OpenStreetMap

La plus grande source d'informations géographiques disponible gratuitement est le projet Open Street Map. Le projet communautaire, fondé en 2004 à UCL, vise à créer une base de données ouverte de l'ensemble de la planète, en se basant sur les contributions de volontaires bénévoles.

L'objectif principal de ce projet est de créer des cartes disponibles gratuitement. Celles-ci sont accessibles librement sur le site `http://www.openstreetmap.org`.

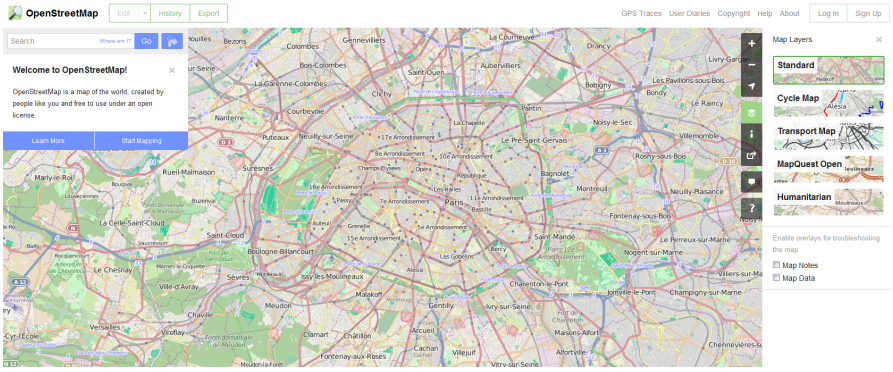



FIGURE IV.2.1 – Exemple de carte consultable sur le site de OpenStreetMap.

L'intégralité des données du projet est disponible. L'ensemble de la base Française a été

téléchargée pour ce projet. Un ensemble d'outils a été créé pour extraire les données désirées (par exemple, les données se rapportant à de la voirie, ou un type de commerce, dans une zone donnée), et les importer dans un serveur de base de données. Les outils informatiques utilisés sont brièvement décrits dans le paragraphe IV.2.3.

Par exemple, la "carte" ci-dessous a été créée en extrayant de la base OpenStreetMap les restaurants et les rues se trouvant dans le quartier de la rue Mouffetard :

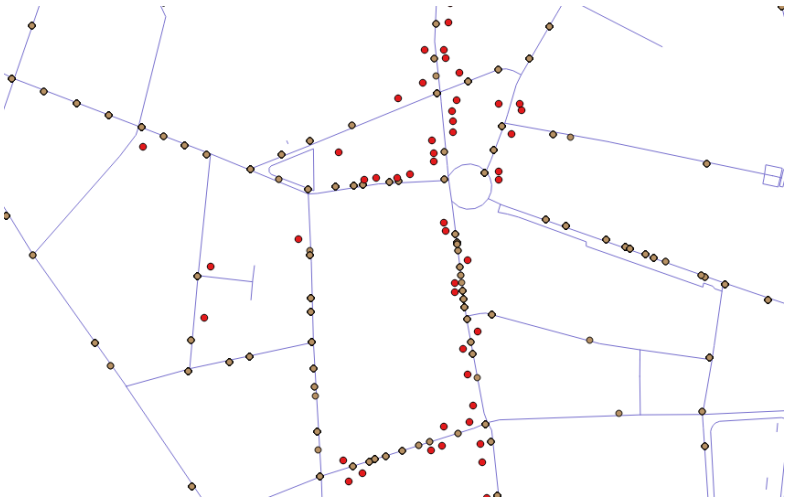


FIGURE IV.2.2 – Illustration de données extraites de OpenStreetMap : les restaurants sont affichés en rouge, et les contrats MRH AXA en jaune.

### Autres sources de données ouvertes

D'autres sources de données sont disponibles gratuitement en ligne.
Celles ci sont souvent spécialisées pour un besoin donné, et agrégées par zones plus ou moins grandes.

**INSEE** En particulier, l'INSEE propose un grand nombre d'informations, agrégées par code INSEE (grossièrement équivalent au code postal).

**GASPAR** La base GASPAR (Gestion Assistée des Procédures Administratives relatives aux Risques naturels et technologiques), créée par la Direction Générale de la Prévention des Risques, est concentrée sur la cartographie des risques naturels et technologiques. Ces données sont accessibles sur `http://macommune.prim.net/gaspar/`.

**CORINE** La base CORINE est un projet européen, dirigé par l'Agence Européenne de l'Environnement, cartographiant l'occupation des sols. La première version de la base date de 1990, et la dernière mise à jour a été réalisée en 2012. L'ensemble des données de la base CORINE est téléchargeable sur le site `http://www.stats.environnement.developpement-durable.gouv.fr/clc/CORINE_Land_Cover_-_Saisie_Demande.jsp`.

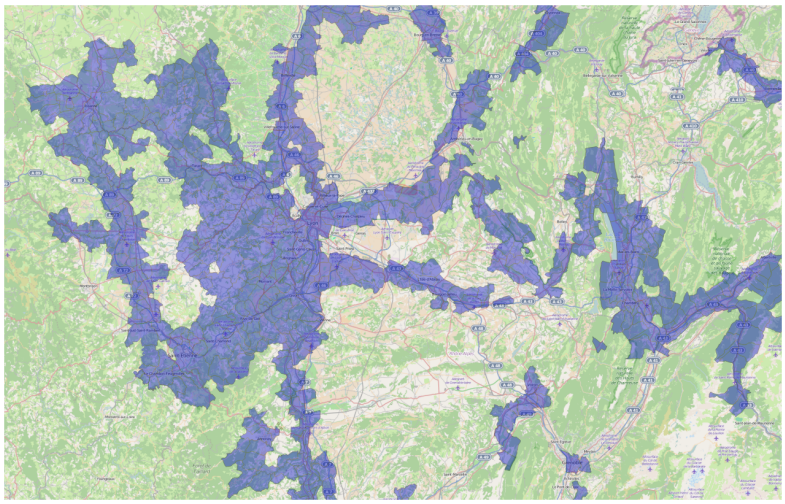


FIGURE IV.2.3 – Exemple d'extrait de la base GASPAR : risque d'inondation, par commune, dans les environs de Lyon.

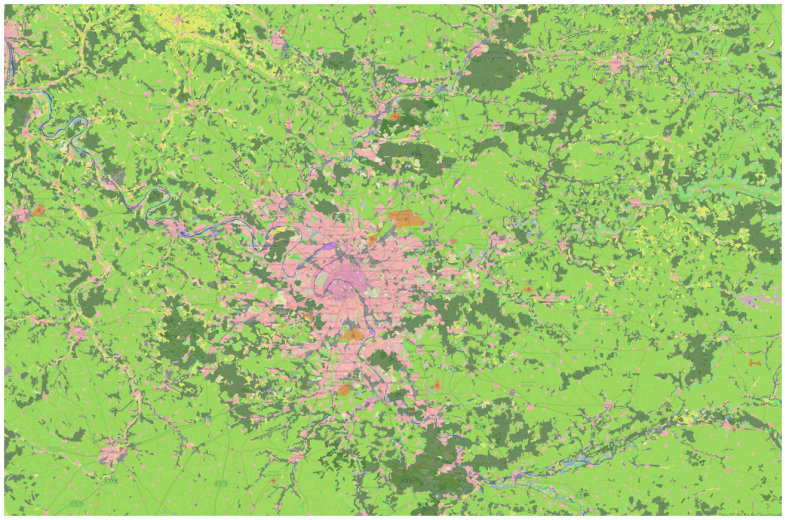


FIGURE IV.2.4 – Exemple d'extrait de la base CORINE en région parisienne. Les terrains agricoles sont représentés en vert, les zones urbaines en rose, les zones commerciales en gris, les aéroports en orange et les forêts en vert foncé.

Bien entendu, un grand nombre d'autres bases ouvertes sont disponibles (dont, par exemple, des bases de données météorologiques...)

### IV.2.1.b Données externes propriétaires

Il est aussi possible d'acheter des données à des sociétés spécialisées dans l'acquisition, la fiabilisation et la revente de données. Parmi celles ci, on peut citer Experian, ESRI, Pitney Bose, KelQuartier...
Tous ces fournisseurs proposent des informations liées à des adresses, rues, ou codes postaux, sur une région, la France entière (ou une couverture mondiale).

Enfin, certains acteurs spécialisés dans d'autres domaines ont un comportement opportuniste et tirent profit des données en leur possession, comme par exemple la chambre des notaires de Paris qui propose un accès payant à la base BIEN (Base d'Information Économiques Notariales), recensant l'adresse, le montant et d'autres informations sur les transactions immobilières réalisées dans la capitale.

### IV.2.1.c Données internes

Comme toujours, les sources d'information les plus pertinentes pour établir un tarif d'assurance sont bien entendu d'origine interne. En effet, il n'est pas envisageable de créer un tarif en se privant d'un historique de sinistres.
La disposition d'un ensemble de polices localisées, pour lesquelles nous disposons d'un historique de sinistralité suffisamment important, est bien entendu nécessaire pour construire la variable cible du modèle que nous souhaitons développer.

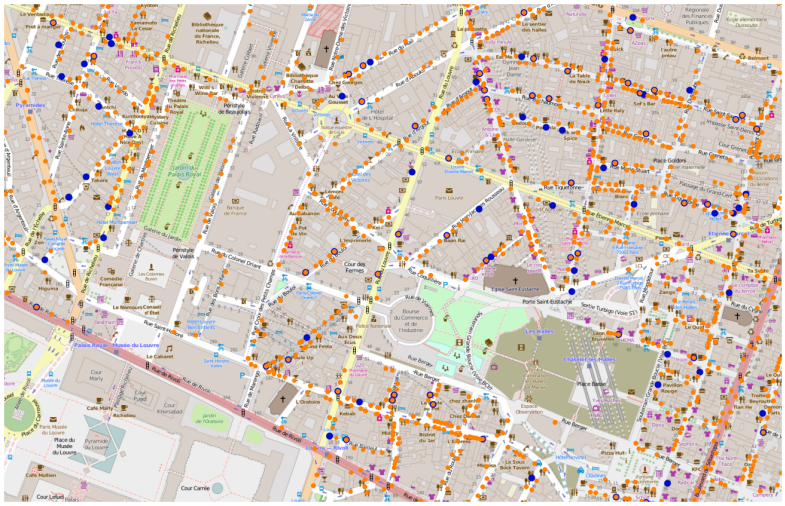


FIGURE IV.2.5 – Exemple de visualisation de la base clients d'AXA France. En orange, les clients non-sinistrés présents dans la base. En bleu, les clients sinistrés par un dégât des eaux.

Mais d'autres informations internes à l'entreprise peuvent être exploitées pour estimer le risque lié à une position géographique. Par exemple, dans le cadre de cette étude, nous avons pu

tirer profit du vaste portefeuille d'immeubles assurés par AXA France : en effet, pour chacun de ceux ci, nous disposions d'informations sur la structure (année de construction, nombre d'étages), ou l'occupation (taux de commerce et d'habitation).

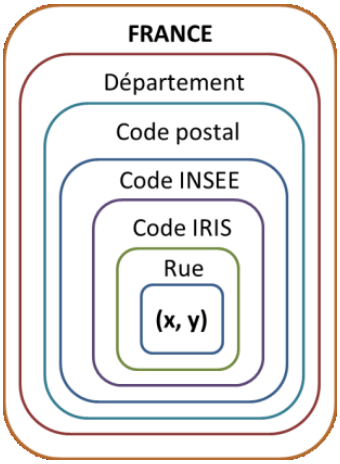



FIGURE IV.2.6 – Précisions des sources d'information disponibles ; une plus grande précision est, bien sûr, toujours souhaitable.

### IV.2.2 Géo-codage des polices

La première étape nécessaire à l'estimation du risque d'une police à partir de sa position géographique est... de déterminer sa position géographique.
Cette opération n'est pas triviale. En effet, si on peut accéder directement au code postal de l'assuré (qui l'indique dans son adresse), l'obtention d'une meilleure précision est une opération complexe.
Il est nécessaire de formater correctement l'adresse fournie par le client (redressage d'adresse), et de disposer d'une base de données permettant d'obtenir, à partir d'une adresse suivant un format pré-défini, des coordonnées géographiques (x,y) - cette opération étant nommée géo-codage dans la suite de ce mémoire. Ces opérations ne se déroulent pas toujours correctement ; il devient alors nécessaire de proposer une vision dégradée du risque (par exemple, à l'échelle du code postal ou de la rue), au cas où le géo-codage d'un contrat ne peut être réalisé
Dans le cadre de ce projet, le redressage d'adresse ainsi que le géo-codage ont été réalisés par des fournisseurs de services extérieurs.

Il est cependant intéressant de noter que des projets libres ont récemment émergé pour fournir ce type de service. En particulier, le projet BANO, mené conjointement par l'IGN, la Poste, l'État et OpenStreetMap France, propose aujourd'hui un système de redressage d'adresse et de géo-codage ouvert et gratuit. Plus d'informations sur ce projet sont disponibles sur `http://openstreetmap.fr/bano`.

## IV.2.3 Technologies utilisées

Une estimation géographique fine du risque d'assurance habitation requiert l'utilisation d'un certain nombre d'outils et de formats de fichiers qui ne sont pas nécessairement familiers aux actuaires tarification. Nous les décrivons donc rapidement dans cette section.
L'ensemble des outils permettant le stockage, le traitement et la visualisation de données géographiques s'appelle un Système d'Information Géographique, ou GIS (Geographic Information System en anglais). Cet ensemble est constitué de plusieurs parties indépendantes (décrites ci-dessous), prévues pour s'interfacer entre elles.

### IV.2.3.a Outils et logiciels

#### Moteurs de Bases de données Géographiques

Les volumes de données à traiter sont importants, et la réalisation d'opérations liées à des données géographiques lourde. L'utilisation d'un système de gestion de bases de données bien optimisé est donc nécessaire.

**PostGres** est l'un des principaux moteurs de bases de données ouvert. Fondé en 1985, il est réputé pour son haut niveau d'optimisation, sa fiabilité et sa richesse d'utilisation.
Les bases de données sont manipulées via l'interface d'administration de PostGres, PGAdmin, ou via des commandes émises par différents programmes (cf. ci dessous).

**PostGis** est une extension de PostGres, permettant le stockage et la réalisation d'opérations sur des données géographiques. Ces données sont représentées comme des points, des courbes ou des polygones, munies d'un système de projection permettant de les associer à des points physiques. Les opérations possibles sont, entre autres, le calcul de la distance entre deux points, l'appartenance d'un point à un polygone (et des jointures basées sur cette appartenance), les calculs d'intersection ou de différences entre deux polygones, et cætera...
L'utilisation de PostGres et PostGis permet de réaliser en quelques secondes ou minutes des opérations qui, sans bénéficier des optimisations développées pour ce système, prendraient un temps prohibitif. En revanche, afin de pouvoir bénéficier des optimisations prévues lors du développement de ces systèmes, il est nécessaire d'être attentif au type et aux propriétés des objets manipulés.

**QGIS** est un logiciel libre de visualisation de données géographiques. Il permet de charger différents types de données (bases de données PostGis, fichier ShapeFile -décrits plus bas- ou CSV), et d'afficher les données contenues à des fins de visualisation ou d'export.
Il est possible d'enrichir QGis à l'aide d'extensions réalisées en Python.

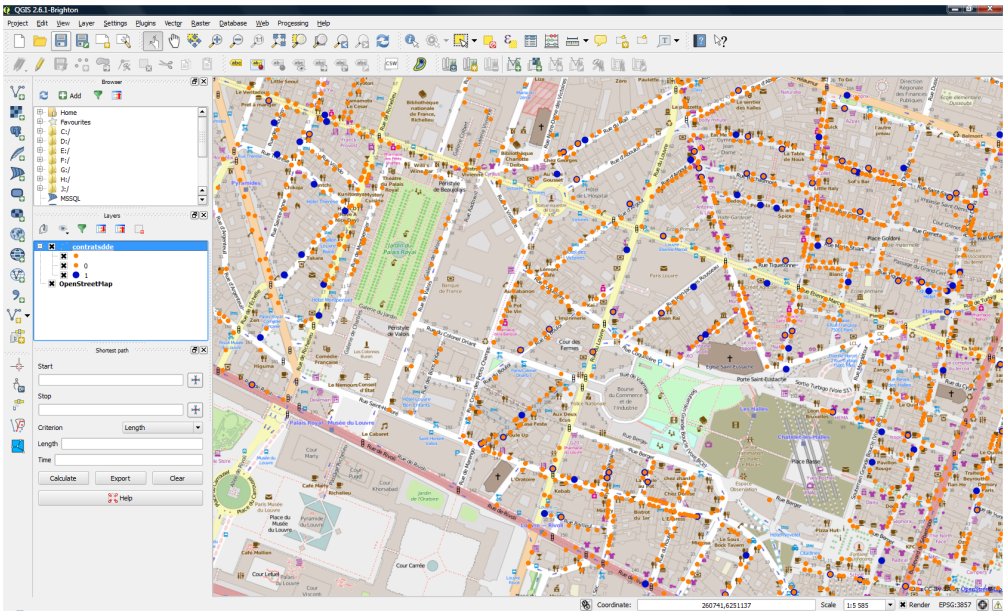


FIGURE IV.2.7 – Le logiciel QGIS permet de visualiser tout type de données géographiques : ici, une visualisation de données OpenStreetMap -directement via une connection internet- et de la base de données clients -stockée dans une base PostGis- afin de créer la figure IV.2.5.

### R et Python

**R** Le langage de programmation R est probablement le plus populaire au sein de la communauté actuarielle. En effet, ce langage dispose d'un grand nombre de librairies de calcul scientifique et d'outils statistiques, ainsi que d'un système de visualisation remarquablement puissant.

**Python** Cependant, l'interfaçage de R avec d'autres systèmes de gestion de données (en particulier PostGres) est complexe, et la manipulation, au sein du langage, d'objets géographiques est limitée. Pour ces raisons, le projet, après avoir été lancé sous R, a été recodé en Python.
Python offre un ensemble de librairies de machine-learning très large et bien optimisées (la plus connue étant SKLearn), un ensemble de librairies d'interfacage avec les systèmes de bases de données géographiques (SQLAlchemy, GeoAlchemy), des librairies de manipulation d'objets spatiaux (Shapely) et, enfin, un ensemble d'outils de visualisation performant (Matplotlib).
L'ensemble des librairies mathématiques de Python, ainsi qu'un environnement de développement et d'autres outils (mais pas de librairies ou d'outils liés à la manipulation de données spatiales) est disponible dans Anaconda, une distribution (gratuite mais propriétaire) proposée par Continuum Analytics.

### IV.2.3.b Formats

#### ShapeFile

Le format ShapeFile est l'un des principaux formats de stockage de données géographiques. Développé par ESRI, il s'agit d'un format de stockage de données, composé de plusieurs fichiers (au minimum 3 fichiers dont les extensions sont .shp, ainsi que .dbf et .shx ).
Les données étant disponibles dans un petit nombre de fichiers au format connu et lisible par un grand nombre d'outils, elles sont facilement échangeable. Ce format est donc le plus fréquemment utilisé pour partager les données, et un grand nombre de nos sources de données externes, ainsi que les cartes de risque créées dans ce projet, sont stockées sous ce format.

#### Formats ad-hoc

**Bases PostGres** Bien entendu, les bases de données PostGres sont stockées dans un format ad-hoc, auquel il est possible d'accéder via le moteur, sous forme de bases de données relationnelles.

**Format OSM** Les données partagées par OpenStreetMap sont disponibles dans un format particulier (OSM), dérivé de XML, prévu pour pouvoir extraire efficacement un type d'information et être sous forme de texte lisible.
Un outil développé dans le cadre de OpenStreetMap permet d'extraire des informations d'une base OSM et d'enregistrer celles ci sous forme de fichiers Shapefile ou dans une base PostGis.

## IV.2.4 Données disponibles

Les données disponibles sont donc :
— Les informations sur les polices d'AXA : coordonnées géographiques (x,y) et sinistralité observée.
— Un ensemble de cartes, contenant des données socio-démographiques ou physiques, qui doivent être transformées pour créer des variables associées à nos polices.

Les informations issues des cartes collectées peuvent être divisées en deux types : informations zonées, qui correspondent à une zone d'une carte à laquelle sont associées une ou plusieurs valeurs, ou des informations ponctuelles, qui représentent un point de l'espace auquel sont, de nouveau, associées une ou plusieurs valeurs.

## IV.2.5 Exploitation de données zonées

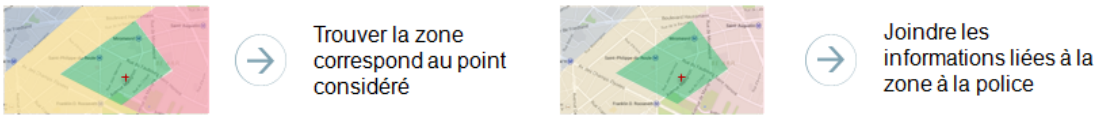



FIGURE IV.2.8 – Jointure d'une police à une information zonée ; on trouve quelle zone englobe la police considérée, et associe à la police les informations liées à la zone

L'association d'informations géographiques à une police par le biais d'un ensemble de zones est sans doute la méthodologie la plus directe qui puisse être envisagée.
On considère un ensemble de zones, qui peut former une partition de l'espace (comme, par exemple, les zones correspondant aux codes postaux) ou non (comme par exemple les zones

délimitant les forêts sur une carte).
A chaque police, on associe la zone qui l'entoure. A cette zone peuvent être associées plusieurs informations (comme par exemple l'ensemble des données INSEE liées à un code postal) ou une seule (par exemple le type d'environnement propre à cette zone : forêt, zone urbaine...).
Ces informations liées à la zone sont directement attachées à la police, et constituent des variables explicatives de son risque.

## IV.2.6 Exploitation de données ponctuelles

### IV.2.6.a Distance à des points d'intérêt

Un type de données présent sur les cartes collectées nommé Point d'Intérêt (Point of Interest, ou PoI). Il s'agit d'une information ponctuelle, signalant la position d'un service, d'un commerce, ou d'une curiosité... Par exemple, les bars, restaurants, stations de police ou de pompiers, gares ou médecins sont indiqués sous cette forme dans la base OpenStreetMap.
Afin d'exploiter ces informations dans les modèles créés, nous pouvons calculer, pour chaque police, la distance au point d'intérêt le plus proche. Nous créons donc, de cette manière, des variables représentant, pour chaque observation, sa distance au poste de police, bar ou médecin le plus proche.

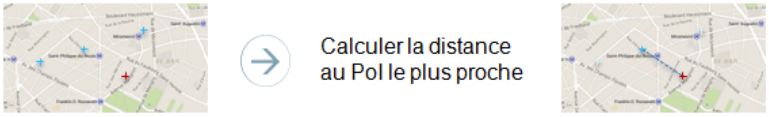



FIGURE IV.2.9 – Jointure d'une police à une information ponctuelle ; on trouve, pour un type de point d'intérêt (PoI) donné, la distance entre la police et le point le plus proche ; cette distance sert à générer une variable explicative.

Cependant, cette utilisation directe des distances aux points d'intérêt n'est sans doute pas optimale. En effet, toutes les distances à ces points d'intérêt seront fortement corrélées, et reflèteront probablement la densité d'urbanisation à un point donné : si celle ci est élevée, la distance à toute sorte de points d'intérêt sera faible et si, en revanche, la zone est peu dense, tous les types de points d'intérêt (bar, médecins, commerces...) sera grande.
Il est donc nécessaire, pour exploiter cette information, de la normaliser par une mesure de densité.
La mesure employée est la distance au 1 $000^{eme}$ plus proche client d'AXA présent dans notre base. Cette mesure de densité est, bien entendu, arbitraire. Elle a l'intérêt d'être disponible en tout point de l'espace et d'être relativement précise - contrairement à la densité moyenne d'un code INSEE, par exemple, elle reflète bien le passage d'une zone urbaine à une zone rurale au sein de ce code. En revanche, elle ne crée pas de distinction entre une zone peu densément habitée et une zone dans laquelle AXA est peu présent.
On peut voir cette normalisation comme "le nombre de clients entre la police X et le médecin/restaurant/commissariat le plus proche".
Ce traitement est illustré figure IV.2.10. On remarque que :

— Sur le $1^{er}$ graphique, les vétérinaires sont nombreux dans certaines régions, (sud-est de Paris par exemple) mais peux nombreux dans d'autres (nord-est). Cette répartition suspecte des PoI laisse penser que les données n'ont pas été correctement reportées dans

la base OpenStreetMap. En effet, cette base est créée par un travail communautaire. Si certaines zones ne comprennent pas de contributeurs actifs, la quantité d'informations qui y sont recensées sera faible.

— Sur le $2^{nd}$ graphique, on constate que la distance au vétérinaire le plus proche est plus faible dans Paris (couleurs chaudes) qu'en banlieue (couleurs froides).

— Enfin, le $3^{éme}$ graphique montre que, après normalisation, l'effet s'inverse et la distance au vétérinaire le plus proche devient plus élevée dans Paris qu'en banlieue.

Plusieurs dizaines de types de points d'intérêt ont été utilisées, permettant de créer autant de variables définissant les observations utilisées pour créer le modèle de risque.

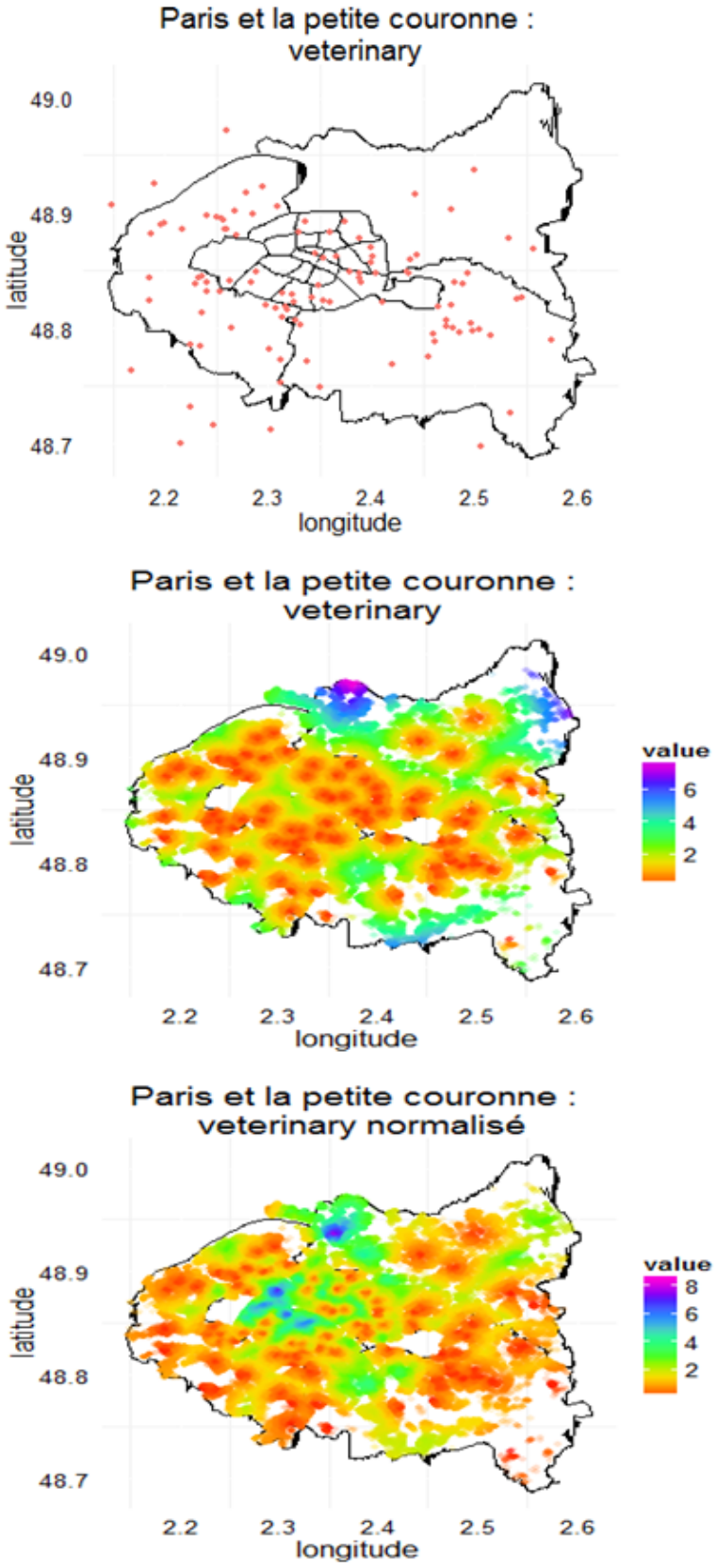



FIGURE IV.2.10 – Exemple de distance à un point d'intérêt : sur le premier graphe, les vétérinaires référencés sur OpenStreetMap sont représentés en rouge. Au milieu est représentée la distance au vétérinaire le plus proche, pour tous les clients de notre base de données (échelle logarithmique). En bas, cette distance est normalisée par la densité de clients -distance au 1 000$^{eme}$ client).

### IV.2.6.b K plus proches voisins

La distance au point d'intérêt plus proche permet d'exploiter la situation géographique de celui-ci, mais pas ses propriétés (variables quantitatives ou qualitatives attachées au point).
Pour exploiter les variables attachées à des points, nous pouvons estimer les valeurs de moyennes de celles ci sur les $K$ points les plus proches de l'observation considérée.
Par exemple, une des sources d'information disponible est la base des immeubles assurés par AXA France.
Cette base comprend l'année de construction de l'immeuble, son nombre d'étages, la proportion de sa surface utilisée à des fins commerciales et sa position géographique. Il serait donc possible de prendre, pour chaque police, l'immeuble le plus proche présent dans la base et d'associer ses informations à la police. Cependant, le résultat obtenu serait fortement bruité (deux immeubles proches ne partagent pas nécessairement la même valeur pour ces variables, et la valeur associée à la police caractérisée comportera donc une grande part d'aléa).
Il est donc pertinent de moyenner la valeur utilisée sur un certain nombre d'immeubles.
Ce nombre ne pouvant être estimé simplement (et étant probablement variable en fonction de l'environnement), on peut créer, pour chaque source d'information, une "famille" de variables : nombre d'étages de l'immeuble le plus proche, et moyenne des 3, 5 ou 10 immeubles les plus proches.

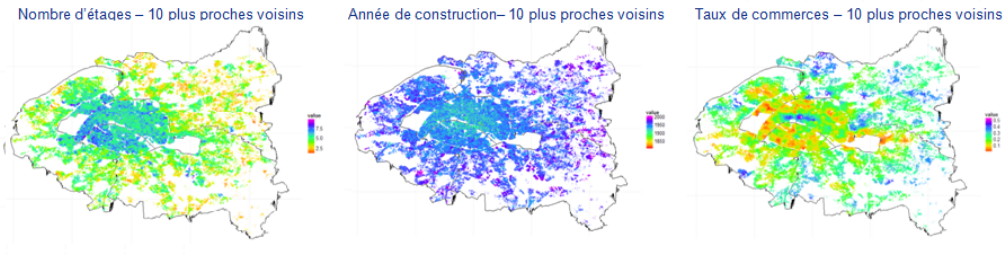



FIGURE IV.2.11 – Exemple de jointures à des variables liées à des données ponctuelles : nombre d'étages, année de construction et taux de commerce des immeubles assurés par AXA France. Les données sont moyennées sur les 10 plus proches voisins.

Une variable particulièrement intéressante est la sinistralité observée des polices présentes autour d'un point (ou plus précisément les résidus de cette sinistralité, en fréquence, ou en coût, sur les voisins sinistrés). Celle ci peut être attachée à ce point comme les autres variables liées à des informations ponctuelles : pour un point donné, on identifie les $K$ polices les plus proches dans notre base historique, et calcule la sinistralité moyenne observée sur celles-ci.
Le choix d'un $K$ élevé créera une variable moyennant un grand nombre d'observations : elle sera donc plus fiable, mais représentera une grande surface autour du point caractérisé. A l'inverse, un $K$ petit moyennera peu d'observations, proche du point d'observation ; la variable créée sera donc fortement bruitée, mais retranscrira des informations précises.

Une réserve importante doit néanmoins être apportée à ce type de variables. Si les évènements que l'on cherche à prédire ne sont pas parfaitement indépendants (ce qui est toujours susceptible d'être le cas), l'utilisation de la sinistralité chez les voisins d'une police devient une variable a posteriori.
Ce phénomène peut être illustré à l'aide d'un exemple concret. Si l'on souhaite prédire la fréquence de vols, la variable créée sera le nombre de vols parmi les K plus proches voisins (par exemple les

100 plus proches voisins). Si une série de vols a eu lieu dans le même immeuble (les voleurs s'étant procuré les clés), cette variable créée devient un prédicteur très fort, mais a posteriori (et donc non valable), de la cible. Un exemple extrême d'évènements non-indépendants est la survenance d'évènements climatiques : si mes voisins sont touchés par une inondation, il est probable que je sois touché par la même inondation... et que l'information sur mes voisins soit extrêmement prédictive au sein d'un test sur des données historiques, mais inutile pour une utilisation réelle. C'est pour cette raison que les risques liés aux évènements naturels ne sont pas couverts par ce mémoire (comme indiqué chapitre I.2.3.c).
Afin d'éviter cet effet indésirable (et dur à quantifier) il devient nécessaire d'exclure les sinistres suspects d'être liés au même évènement que ceux que l'on cherche à prédire lors de la construction des nouvelles variables. 2 manières de réaliser cette exclusion peuvent être envisagées :

- Pour une observation donnée, ne considérer que les informations plus anciennes lors de la construction de l'estimateur des K plus proches voisins. Il s'agit bien entendu de la manière rigoureusement correcte de procéder. Cependant, elle est relativement ardue à mettre en œuvre pour deux raisons. La première est algorithmique (une telle exclusion requière beaucoup de traitement des données, et est lourde à réaliser), la seconde est statistique : le nombre d'observation disponible autour d'une police change en fonction de son ancienneté (les polices anciennes ont peu de voisins exploitables, quand les polices plus récentes en ont beaucoup), et la signification de la variable créée change donc également en fonction de son ancienneté.
- Pour une observation donnée, ne considérer que les informations issues d'autres années (plus anciennes ou récentes). Cette méthode est moins rigoureuse (des informations futures peuvent être exploitées) mais résout les problèmes liés aux sinistres non-indépendants, liés à un évènement commun (comme une vague de vols dans un immeuble ou une inondation).

La seconde méthode est donc recommandée pour la création pratique de variables de type K-plus-proches voisins. L'expérience prouve que ce type de traitement des données est nécessaire, et que l'utilisation de l'ensemble des voisins sans exclusion amène à la création de variables excessivement efficaces sur des tests historiques mais très faiblement prédictives en pratique.

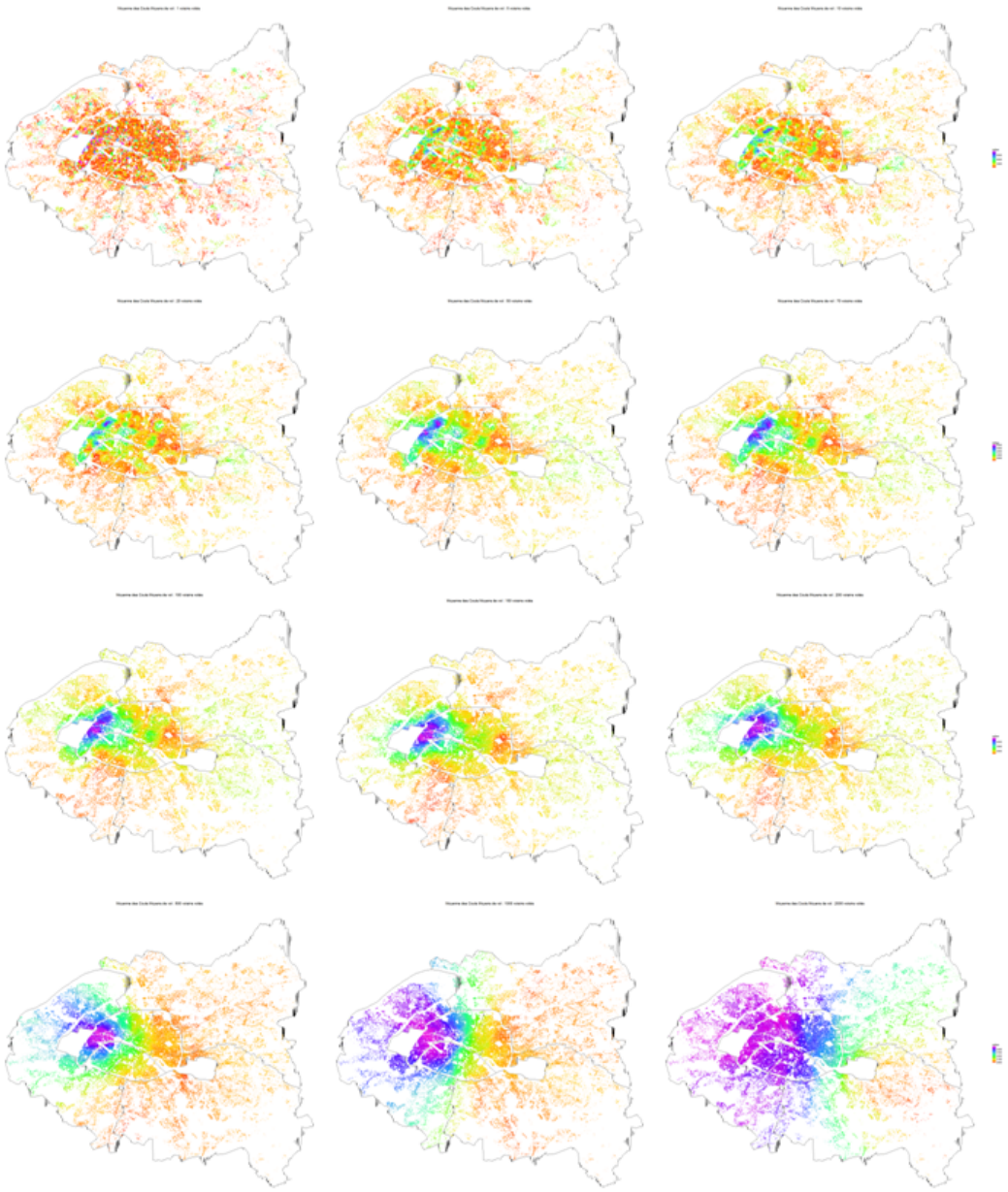


FIGURE IV.2.12 – Exemple de moyenne du coût des vols observés, sur les $K$ sinistres les plus proches, avec $K$ variant de 1 à 2000, en région parisienne. Le coût va du orange - montants les plus faibles - au violet -montant les plus élevés.

On peut considérer les variables créées comme des estimateurs directs de la variable cible (il s'agit d'un estimateur classique des K plus proches voisins - *K Nearest Neighbors* en anglais, ou *KNN*). Dans ce cadre, un $K$ petit créera un estimateur à faible biais mais forte variance, quand un grand $K$ créera un estimateur à fort biais mais faible variance, en reprenant la décomposition de l'erreur définie par l'équation II.2.2.
Cet équilibre entre biais et variance peut être illustré par le niveau de corrélation entre la sinistralité moyenne autour d'une police et la sinistralité de celle ci. Comme le nombre optimal de voisins est susceptible de varier d'une région à l'autre, et que les variables créées ne seront qu'une partie des variables explicatives des modèles créés, nous créons, de nouveau, une famille de variables représentant la sinistralité d'un nombre variable de voisins autour du point considéré.

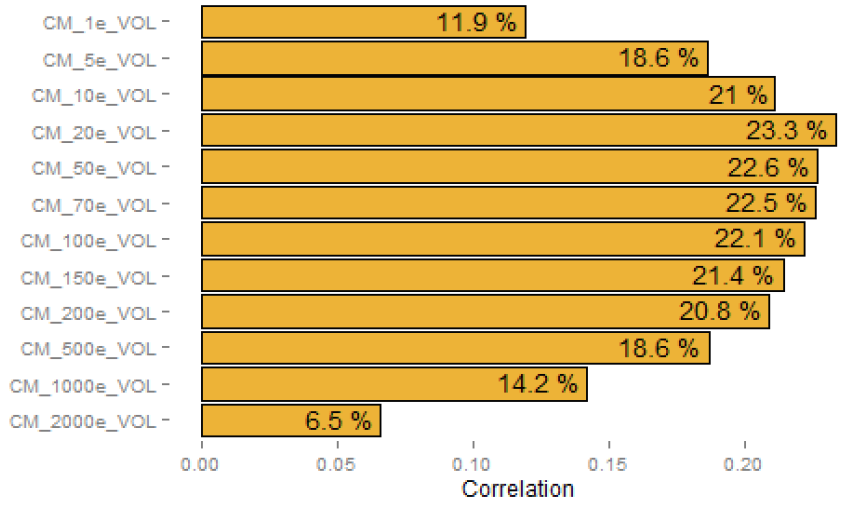



FIGURE IV.2.13 – Illustration du compromis biais-variance
Corrélation entre le coût moyen d'un sinistre vol, et le coût moyen des $K$ vols les plus proches, pour $K$ variant de 1 à 2 000. Pour un nombre de voisins $K$ faible (en haut), on observe un corrélation faible (faible biais, mais forte variance). Pour $K$ élevé (en bas), la corrélation décroit de nouveau (faible variance, mais fort biais). Le maximum est atteint à $K = 20$.

## IV.2.7 Propriétés des variables créées

Les différents procédés décrits dans le chapitre précédent permettent de créer un certain nombre de variables explicatives. Certaines méthodes, en particulier celles décrites dans la section IV.2.6.b, créant des "familles" de variables, le nombre de variables explicatives produites devient rapidement très élevé.
Avant d'utiliser ces variables pour prédire le risque de nos clients, nous nous proposons de décrire rapidement leurs propriétés.

### IV.2.7.a Colinéarité

Les variables créées sont fortement colinéaires.
Cette caractéristique est relativement intuitive. Dans le cas des familles de variables de type K plus proches voisins, celle ci est bien théorisée (et peut, sous certaines hypothèses, être exprimée

explicitement).

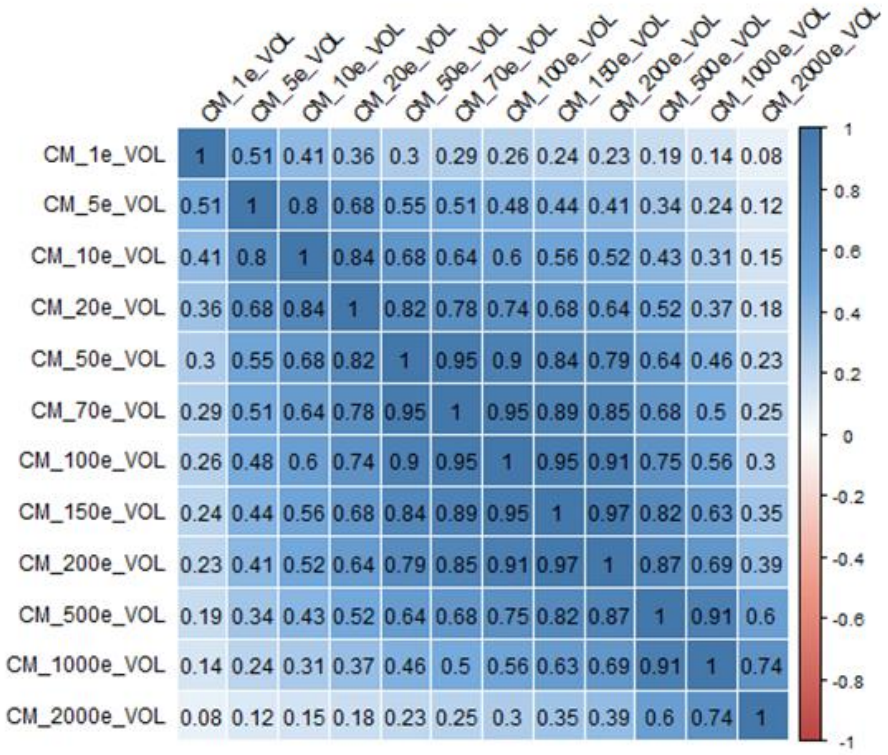



FIGURE IV.2.14 – Corrélation des variables de type KNN
Corrélation entre l'ensemble des coûts moyens des $K$ plus proches dégâts des eaux, avec $K$ variant de 1 à 2000.

Dans le cas des distances au POI, les variables reflètent souvent des réalités "physiques" comparables. Par exemple, la distance à un certain nombre de points d'intérêt sera fortement dépendante de la densité urbaine des observations. Un exemple de corrélations est donné ci-dessous.

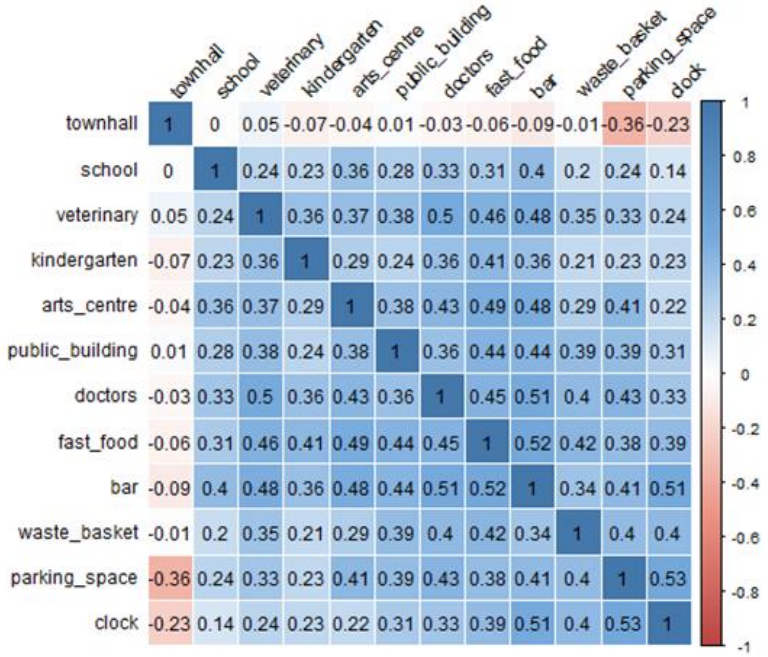



FIGURE IV.2.15 – Corrélation des variables de type distance à des POI
Corrélation entre différentes variables représentant la distance à des points d'intérêt. Ces corrélations sont presque toujours positives (à l'exception des distances à la mairie -townhall- et au parking ou au dock le plus proche ; on note que très peu de docks sont présents dans la zone étudiée).

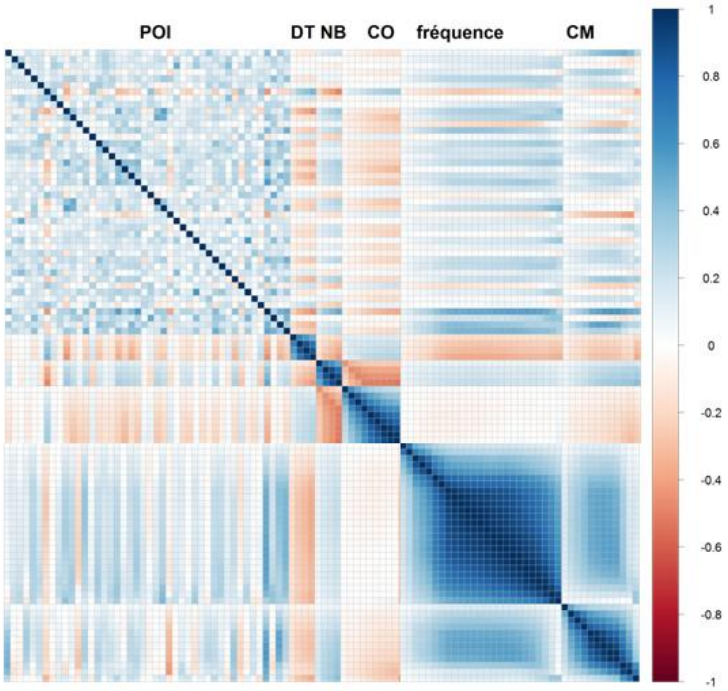



FIGURE IV.2.16 – Corrélation des variables
Corrélation entre les différentes variables utilisées pour estimer le coût de sinistres dégât des eaux.
Chaque ligne ou colonne correspond à une variable ; le nom des familles de variables est indiqué : de gauche à droite (haut en bas) : distance à des POI ($POI$), date de construction ($DT$), nombre d'étages ($NB$), taux de commerce ($CO$), fréquence de sinistres des plus proches voisins *fréquence*() et coût des plus proches sinistres ($CM$).

Une forte corrélation des variables explicatives rend, bien entendu, l'apprentissage statistique plus complexe (cf. **??**).

## IV.2.8 Non-indépendance des observations

Une autre source de difficulté, plus inhabituelle, liée à l'utilisation de variables géographiques est la corrélation entre les différentes observations (auto-corrélation) [17].
En effet, les modèles statistiques supposent que les observations sont indépendantes les unes des autres. Cette hypothèse est fortement violée dans le cas de données géographiques, ce qui peut être extrêmement gênant [13], et conduire, si les données ne sont pas retraitées, à des conclusions surprenantes [23].

Comme nous le verrons, une forte auto-corrélation des variables explicatives ainsi que de la variable à expliquer crée un fort risque de corrélation entre ces deux variables, ce qui crée un risque d'interprétation trompeuse des modèles créés, et d'extrapolation erronée dans les zones faiblement exposées.

### IV.2.8.a Illustration : auto-corrélations en 1 dimension

Le phénomène d'auto-corrélation est bien connu des divers champs traitant des séries temporelles, dont les mathématiques financières.
Supposons que nous disposons de deux séries temporelles $A$ et $B$, indexées par le temps $t$ (considéré discret dans cet exemple). Les deux séries sont générées comme suit :

$$\begin{aligned} A(t) &= A(t-1) + \Delta_A(t) \\ B(t) &= B(t-1) + \Delta_B(t) \end{aligned}$$

Avec $\Delta_A$ et $\Delta_B$ deux séries stationnaires indépendantes l'une de l'autre.

Bien que $\Delta_A$ et $\Delta_B$ soient indépendantes, les deux séries $A$ et $B$ apparaissent corrélées si on définit $\hat{\mathbf{Corr}}(\mathbf{A},\mathbf{B}) = \frac{\sum_{t\in[0,T]}(A(t)-\overline{A})\times(B(t)-\overline{B})}{\sqrt{\sum_{t\in[0,T]}(A(t)-\overline{A})^2\times\sum_{t\in[0,T]}(B(t)-\overline{B})^2}}$ leur corrélation empirique sur la période $[0,T]$ (on notera que l'usage de la définition de la corrélation utilisée ici est erronée, celle ci supposant justement l'indépendance des observations entre elles...)

Par exemple, cette corrélation est illustrée dans le test ci-dessous :
Deux séries $A$ et $B$ sont générées aléatoirement pour $t \in [0, 10000]$. Les corrélations entre les deux séries sont mesurées, ainsi que celles entre leurs variations.

Comme attendu, les variations $\Delta_A$ et $\Delta_B$ sont indépendantes les unes des autres (corrélations nulles), mais les séries $A$ et $B$ sont fortement corrélées.

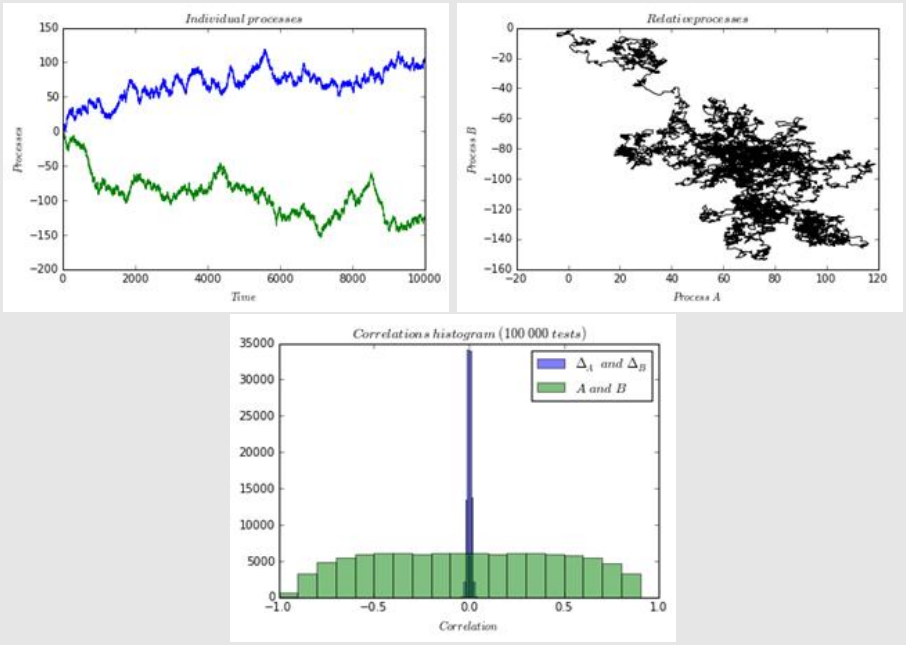



FIGURE IV.2.17 – Exemple de séries temporelles

Deux séries temporelles $A$ et $B$ sont générées à partir de variations indépendantes $\Delta_A$ et $\Delta_B$ (en haut, à gauche : la série $A$ est représentée en bleu, la série $B$ en vert, en fonction du temps). Les deux séries sont fortement corrélées (en haut, à droite : la série $B$ en fonction de la série $A$), chaque observation $(A(t), B(t))$ étant dépendante de l'observation précédente $(A(t-1), B(t-1))$. Cette intuition peut être mesurée en générant un grand nombre de processus aléatoires et en mesurant leurs corrélations (en bas, distribution en vert) ainsi que les corrélations (nulles) entre leurs incréments (distribution en bleu).

### IV.2.8.b Variables spatiales et auto-correlation

En mesurant directement les corrélations entre deux variables spatiales, des phénomènes similaires risquent de se manifester.
Il est donc nécessaire, en intégrant des données spatiales dans un modèle de risque, de prendre en compte la nature de ces données (en particulier lorsqu'on réalise des tests statistiques).

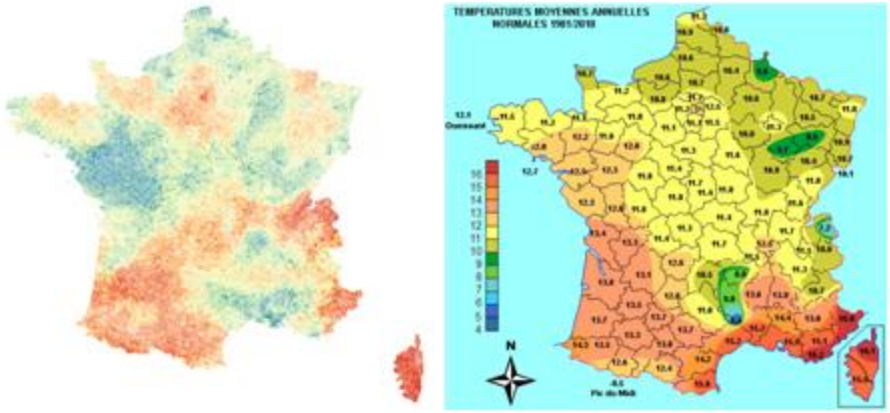



FIGURE IV.2.18 – Exemple de processus spatiaux
Exemple de processus spatiaux : fréquence de vol en France (à gauche) et température (à droite).
Bien que ces deux processus soient fortement corrélés, cette corrélation n'est pas nécessairement significative, et la connaissance de l'un n'apporte pas nécessairement d'information sur l'autre.

### IV.2.8.c Solution possible aux auto-correlations spatiales

La modélisation directe d'une variable fortement auto-corrélée (par des variables tout aussi fortement auto-corrélées) crée, comme nous l'avons vu, des corrélations artificielles entre ces variables (et donc une illusion de modèle fortement explicatif et prédictif).
Ainsi, de même que, pour le traitement de données temporelles, il est nécessaire de stationariser le signal avant de le modéliser, il semble nécessaire, dans le cas de données spatiales, de stationariser les signaux (variable cible et variables explicatives) avant de créer le modèle. Cette approche, comme nous le verrons plus bas (chapitre IV.4.1.b), est implicitement celle suivie par le modèle proposé dans le cadre de ce mémoire.

# Chapitre IV.3

# Création et utilisation d'un zonier

Le principe d'un zonier est décrit paragraphe IV.1.2 :
— Prédiction du risque à l'aide des variables déclaratives
— Géo-localisation des clients et création des variables explicatives géographiques
— Prédiction de la composante géographique du risque
— Construction du micro-zonier

Après avoir réalisé ces étapes, il est possible de produire, pour chaque individu, une estimation de la composante géographique. Cette composante peut simplement être considérée comme une variable du GLM créé.

Dans ce chapitre, nous présenterons une approche générale permettant de réaliser un tel zonier, puis nous intéresserons particulièrement à la $1^{re}$ étape du processus (en particulier, comment créer un résidu à partir des variables créées), et à la dernière étape (comment exploiter le zonier créé dans les prédictions).

## IV.3.1 Approche globale : création d'un zonier

L'objectif du zonier est de prédire les composantes du risque liées à la situation géographique du client.
Afin d'isoler celles-ci, un premier modèle de risque, basé sur les informations déclaratives fournies par le client, est construit.
Ce modèle permet de créer, sur un ensemble d'observations (base d'apprentissage), des prédictions sur les risques des clients. Ces prédictions, ainsi que les réalisations de risque, sont exploitées pour créer des résidus (processus décrit ci dessous, paragraphe IV.3.2).
Afin de pouvoir prédire ces résidus, des variables géographiques sont construites (suivant les processus décrits chapitre IV.2). Cette démarche permet de constituer une représentation des observations en fonction d'un certain nombre de variables géographiques (en plus des résidus que l'on cherche à prédire).
Ces variables explicatives sont exploitées pour réaliser un estimateur des résidus, et pour obtenir des estimations en un grand nombre de points (par exemple pour toutes les observations présentes dans notre base de données).
Une partition de l'espace est créée. Cette partition compte une observation (et donc une estimation de la valeur des résidus) par zone. Chaque point de l'espace appartenant à une et une seule zone, il dispose d'une unique estimation des résidus du modèle basé sur les variables déclaratives.

Les estimateurs des résidus ainsi créés sont intégrés parmi les variables déclaratives, et permettent d'obtenir une nouvelle estimation du risque de chaque observation.

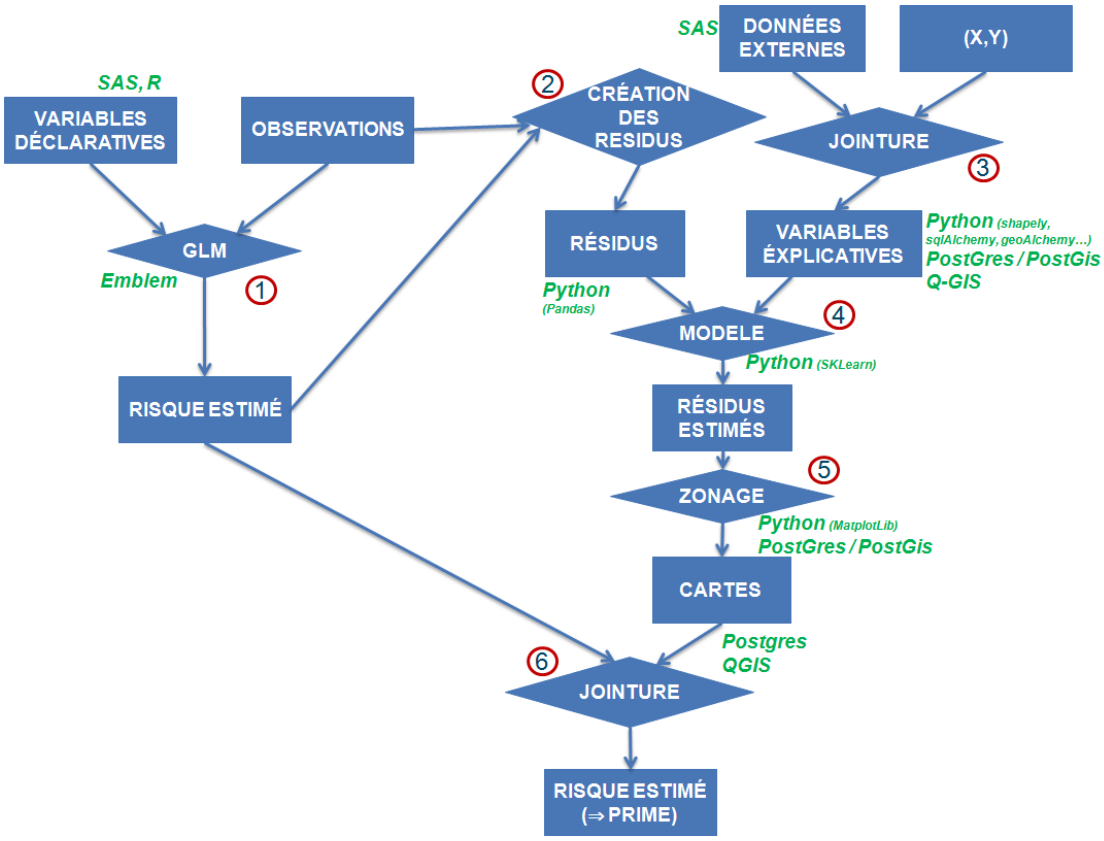



FIGURE IV.3.1 – Construction de création d'un micro-zonier
Etapes principales de la construction du zonier. Ce processus est à rapprocher de celui décrit figure IV.1.1.

La première étape est la réalisation d'un GLM (1) modélisant le risque à partir des variables déclaratives du client. Puis (2) les résidus de ce modèle sont calculés ; ils deviendront la variable cible que nous chercherons à prédire. Ensuite,(3) les variables explicatives géographiques sont créées (principalement à l'aide de jointures spatiales et de calculs de K plus proches voisins). Ces nouvelles variables sont utilisées (4) pour créer un modèle prédictif des résidus construits en (2). Les prédictions du modèle créé permettent de créer (5) une partition de l'espace - chaque zone étant attachée à une valeur du risque prédit. Enfin, les nouvelles polices sont jointes aux nouvelles observations (6) et de nouvelles prédictions, incluant les composantes géographiques du risque, peuvent être créées.

### IV.3.2 Définition de la variable cible

Afin de construire une variable indépendante à l'ensemble déjà disponible et représentant le risque du client, il est nécessaire de produire le résidu du risque $y$ des clients par les prédictions $\hat{y}_{GLM}$ construites à partir des variables déclaratives. 5 types de résidus peuvent être proposées dans ce but [12] :

- Les résidus additifs : il s'agit du type de résidus les plus courants : $R = y - \hat{y}_{GLM}$.
  Si ces résidus sont simples à exprimer, ils ne correspondent pas à la structure multiplicative des GLMs utilisés pour les modèles actuariels (fonction de lien logarithmique), et ont le défaut d'avoir une variance dépendant fortement de $y$ (et $\hat{y}_{GLM}$).
- Les résidus multiplicatifs : il s'agit d'une adaptation des résidus présentés ci-dessus aux GLMs multiplicatifs : $R = \frac{y}{\hat{y}_{GLM}}$. Cependant, ces résidus ont (comme les résidus additifs) une variance qui dépend de $\hat{y}_{GLM}$. Il est aussi important de noter que leur distribution est également fortement déséquilibrée (ces résidus sont multiplicatifs, et, si ils sont construits à partir d'un modèle non-biaisé, leur médiane vaut typiquement 1). Cette propriété peut avoir des conséquences très importantes pour construire un modèle prédictif.
- Les résidus de Pearson : $R = \frac{y - \hat{y}_{GLM}}{std(y)}$, où $std(y)$ est l'écart-type de $y$. Dans le cas présent, $std(y)$ peut simplement être estimé à partir de $\hat{y}_{GLM}$ (la relation entre ces deux valeurs étant une hypothèse du GLM construit). Dans le cas d'une loi de Poisson, le résidu est donc de la forme $R = \frac{y - \hat{y}_{GLM}}{\sqrt{\hat{y}_{GLM}}}$
  Cette approche permet d'obtenir des résidus d'une même variance (homoscédastiques), pouvant être prédits par des modèles simples (mais ne disposent pas de la structure multiplicative des GLMs).
- Les résidus de déviance : Ces résidus sont sensiblement plus complexes à exprimer que les résidus de Pearson. Ils sont basés sur l'expression de la vraisemblance des données à prédire. La prédiction du résidu dont l'erreur quadratique est la plus proche du résidu observé (en terme d'erreur quadratique) correspond à la prédiction du risque dont la vraisemblance est maximale.
  Ils permettent donc d'utiliser simplement des modèles standard (visant à minimiser l'erreur quadratique entre les prédictions et les observations). Cependant, ils ne sont pas inversibles : si il est possible, à partir d'une estimation $\hat{y}_{GLM}$ et d'une valeur de $y$, de calculer une valeur de résidu $R$, il n'est pas possible, à partir d'une estimation $\hat{y}_{GLM}$ et d'un résidu $R$, de retrouver la valeur de $y$ correspondante de manière explicite.
  Le résidu de déviance d'une observation est de la forme : $R = signe(y - \hat{y}) \times \sqrt{|d(y, \hat{y}|}$ avec $d(y, \hat{y}) = 2 \times \sum_i (y_i log(\frac{y_i}{\hat{y}_i}) - y_i + \hat{y}_i)$.
- Les résidus de Anscombe : Les résidus de Anscombe, enfin, sont une approximation des résidus de déviance, dont la formule est inversible (il est donc possible, à partir d'une estimation $\hat{y}_{GLM}$ et d'un résidu $\hat{R}$, de produire une nouvelle estimation $\hat{\hat{y}}$ de $y$.
  Les résidus de Anscombe sont de la forme $R = \frac{3}{2} \frac{(y^{\frac{3}{2}} - \hat{y}^{\frac{3}{2}})}{\hat{y}^{\frac{1}{6}}}$

Afin de réaliser un zonier (ou pour tout travail de modélisation exploitant des résidus, comme par exemple la création d'un véhiculier), le choix des résidus de Anscombe semble donc le plus adapté. Il permet simplement d'utiliser un grand nombre de méthodes de régression basées sur des erreurs quadratiques.

## IV.3.3 Création d'un modèle de résidus

Paradoxalement, la création du zonier proprement dit (la création d'un estimateur des résidus produits lors de l'étape précédente) n'est pas la partie la plus complexe de ce processus.
En effet, nous disposons, pour chaque observation, d'un ensemble de variables (obtenues à partir des données externes, à l'aide de méthodes décrites chapitre IV.2) $\mathbf{x} = (x_1, x_2, ..., x_P)$, ainsi que du résidu $R$ du modèle basé sur les données déclaratives (décrites ci-dessus).
Il est donc relativement simple de construire un modèle prédictif $\varphi$, permettant de prédire la valeur du résidu $R$ pour une observation donnée : $\hat{R} = \varphi(\mathbf{x})$.
Comme décrit ci dessous, les prédictions $\hat{R}$ ne sont pas réalisées en temps réel pour produire la prime d'une police. Il n'existe donc pas de contrainte opérationnelle, en terme d'outils utilisables ou de complexité des prédictions. Cette flexibilité a un fort impact sur la méthode de régression choisie pour estimer les résidus.
Dans ces conditions, la solution la plus simple est sans doute d'utiliser une forêt aléatoire comme modèle $\varphi$.

### IV.3.3.a Prise en compte des effets croisés

L'approche proposée par ce mémoire ne permet pas la prise en compte des effets croisés entre les variables déclaratives et les informations relatives à la localisation de la police.
Si certains effets croisés peuvent être pertinents et améliorer les pouvoirs prédictifs et explicatifs du modèle créé, l'inclusion de ceux-ci risque de mener à des conclusions fallacieuses (en particulier du fait de l'auto-corrélation des variables géographiques, cf. IV.2.8).
En revanche, l'utilisation de forêts aléatoires pour prédire les résidus $R$ implique qu'un grand nombre d'effets croisés entre les variables géographiques sont générés.

### IV.3.3.b Cannibalisation des variables

Il peut arriver que des variables déclaratives soient fortement corrélées à des variables liées à la localisation des polices. Par exemple, le nombre de pièces déclarées dans un appartement est fortement corrélée à la taille moyenne des appartements (aisément disponible, agrégée par code postal).
Si c'est le cas, l'utilisation directe de ces deux variables dans un GLM décomposerait leur impact en deux coefficients. Cette décomposition risque de générer des problèmes d'anti-sélection : tous les clients potentiels disposant d'un appartement plus grand que l'appartement typique sur leur code postal se verraient proposer un prix inférieur à leur risque réel (si l'impact de la taille de l'appartement est positif). Rapidement, les coefficients appliqués ne correspondraient plus au portefeuille (qui risque, du fait de ces problème d'anti-sélection, de devenir déficitaire).
Ce problème de cannibalisation de variables déclaratives par leurs équivalents obtenus par jointure spatiale encourage la création d'un processus de modélisation en deux étapes, utilisant des résidus pour modéliser les composantes géographiques du risque (au lieu de réaliser un modèle direct, intégrant simultanément des variables déclaratives et des variables obtenues par jointure spatiale).

## IV.3.4 Contour des zones

Afin de réaliser une carte (comme décrite plus haut - cf. IV.3.3), il est nécessaire de construire, à partir d'estimations de $\hat{R}$ en un certain nombre de points, un partitionnement du territoire en zones le couvrant (par définition, un partitionnement contient des zones non-recouvrantes, et

couvre l'ensemble du territoire - ici la France).
Les zones choisies pour réaliser ce partitionnement sont une partition de Voronoi. Cette partition associe chaque point du territoire au point d'estimation de $\hat{R}$ le plus proche. Ce choix est motivé par plusieurs arguments :

— La définition (ainsi que l'aspect...) des zones est naturelle : le territoire devant être divisé en zones de risque homogène, il est logique que celles ci soient centrées sur des zones dont le risque est connu.

— La création de ces zones est aisée (il existe des algorithmes permettant de construire leurs frontières en $n \times log(n)$ opérations (avec $n$ le nombre de zones). Des implémentations efficaces de l'algorithme de construction d'une partition de Voronoi sont disponibles simplement dans plusieurs langages de programmation (dont Python).

### IV.3.5 Exploitation des résidus prédits

Une fois l'estimation $\hat{R}_{ansc}$ des résidus de Anscombe $R$ réalisée, il est simple d'obtenir, à partir des premières estimations $\hat{y}_{GLM}$ (fonction des variables déclaratives du client) une nouvelle estimation $\hat{\hat{y}}$ de $y$. A partir de celles ci, on peut obtenir un nouveau résidu, multiplicatif, $\hat{R}_{mul} = \frac{\hat{\hat{y}}}{\hat{y}_{GLM}}$. Ce résidu peut être intégré comme une nouvelle variable explicative, et on obtient une estimation finale du risque : $\hat{y}_{GLM_{inc.Geo}} = \hat{y}_{GLM} \times \hat{R}_{mul}$.

D'un point de vue pratique, deux points importants doivent être notés :

— Les résidus liés à un point de l'espace ne doivent pas être recalculés en temps réel : leur calcul nécessite la jointure d'un grand nombre de données extérieures, ce qui peut être particulièrement lent - cf. IV.2.4 - particulièrement en ce qui concerne les données liées aux K plus proches voisins.
La valeur de l'estimateur des résidus $\hat{R}$ (puis $\hat{R}_{mul}$) doit donc être calculée en un grand nombre de points, et stockée dans une base de données (fichier shapefile) qui pourra être consultée lors de nouvelles demandes de prix.

— Le logiciel actuellement utilisé pour construire les GLM ne permet que l'utilisation de variables discrètes (ordonnées ou non). La nouvelle variable représentant l'estimation des résidus géographiques étant continue, celle ci devra être discrétisée pour être intégrée dans les logiciels. Cette discrétisation devrait logiquement être réalisée en un maximum de modalités (pour perdre un minimum d'informations), et son intégration dans les GLMs doit, naturellement respecter la relation linéaire (attendue) entre cette variable et les observations.

Bien entendu, les résidus stockés dans les cartes produites (cf. IV.3.4) sont des résidus multiplicatifs, qui peuvent être simplement intégrés aux modèles de risque.

# Chapitre IV.4

# Apports de l'approche proposée

L'approche proposée se différencie de la méthode actuellement employée sur 2 points principaux :

- L'utilisation de méthodes de régression originales (Random Forest et Gradient Boosting), qui tendent aujourd'hui à devenir populaires dans le monde actuariel. Cette approche permet d'éviter les étapes de lissage explicite utilisées habituellement.
- La création d'un micro-zonier, plus précis que les zoniers actuellement utilisés aujourd'hui. L'amélioration du niveau de précision du zonier est permise par l'utilisation de méthodes de régression non-standards ainsi que les possibilités (récentes) de géo-localisation précises des polices assurées.

## IV.4.1 Approche actuelle

### IV.4.1.a Description de l'approche

L'approche actuellement utilisée repose sur les étapes suivantes :

- Création d'un modèle à partir des variables déclaratives du client, ainsi que de variables liées à son code postal. Ainsi on produit, pour un client donné (indexé $i$), une estimation $\hat{y}_{GLM} = \prod_{j \in var\ declaratives} \beta_j X_{ij} \times \prod_{j \in var\ code\ postal} \beta_j Xij$.
- On calcule les résidus (multiplicatifs) du modèle créé : $R = \frac{y}{\hat{y}_{GLM}}$.
- Ces résidus sont lissés avec leurs voisins, afin de produire une nouvelle variable $\hat{R}$. Cette variable peut être perçue comme un estimateur des résidus $R$.
- Enfin, les valeurs $\hat{R}$ liées aux différents codes postaux sont rattachées aux polices utilisées pour créer le modèle de risque, qui devient de la forme : $\hat{y}_{GLM} = \prod_{j \in var\ declaratives} \beta_j X_{ij} \times \prod_{j \in var\ code\ postal} \beta_j Xij \times \beta_{lissage} R$

### IV.4.1.b Limites et améliorations proposées par ce mémoire

Cette approche souffre d'un certain nombre de limites, auxquelles la méthode proposée tente de remédier.

**Utilisation de forêts aléatoires**

L'utilisation de modèles de forêts aléatoires (ou de Gradient Boosting) apporte un fort avantage sur une modélisation directement basée sur des modèles linéaires.
En effet, ceux ci permettent de simultanément réaliser une stationnarisation du signal géographique à prédire (par le biais des variables de type K plus proches voisins) et une prédiction des résidus stationnarisés (par le biais des autres variables du modèle).

Cette méthode permet donc implicitement de renverser l'ordre de la modélisation : les données sont stationnarisées puis modélisées, à l'opposé des méthodes couramment employées en actuariat (modélisation par des variables géographiques puis lissage).
L'approche utilisée ici est donc plus proche de celle généralement admise dans les différents domaines exploitant des séries temporelles (par exemple, en finance, la modélisation des rendements et non des valeurs absolues des cours).
La figure IV.4.1 illustre ce procédé en deux étapes.

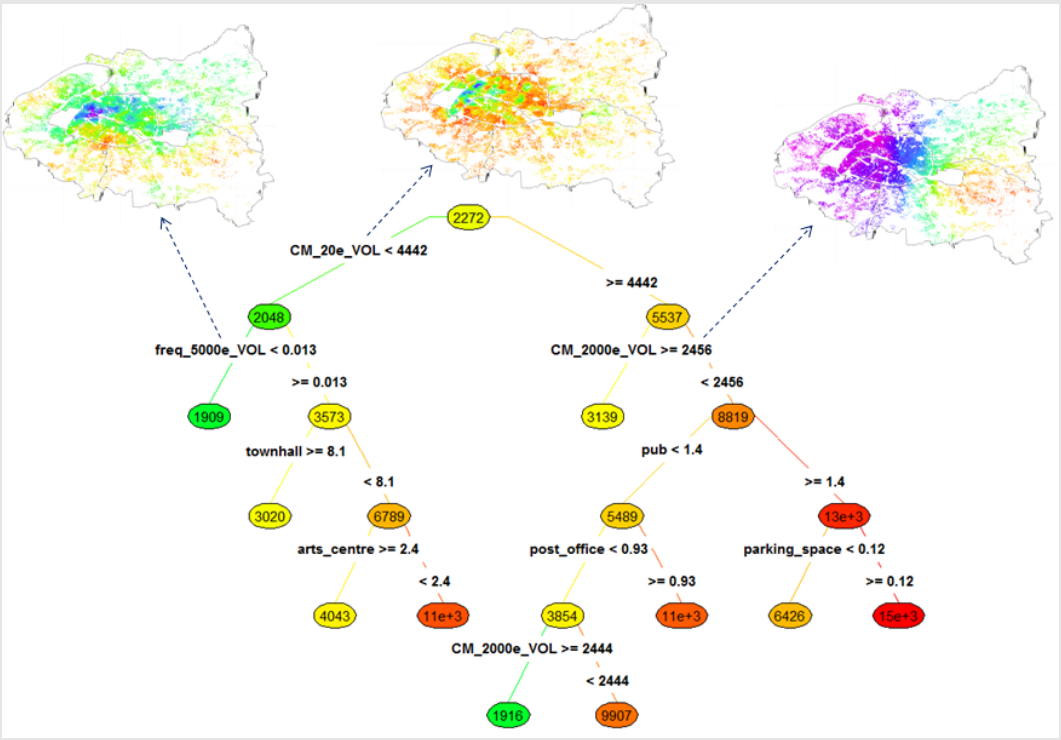



FIGURE IV.4.1 – Exemple d'arbre de régression appliqué à un signal géographique : stationnarisation du signal puis estimation des résidus.

Cet arbre est extrait d'une forêt estimant le coût moyen de vols.
Comme souhaité, les premiers niveaux de l'arbre sont basés sur des estimateurs du coût moyen des K plus proches voisins (avec $K = 20$ puis 2 000. Ils estiment donc la composante stationnaire du signal.
Les niveaux suivants fournissent des estimateurs du signal résiduel, à partir de variables géographiques.
On notera cependant que, si le procédé décrit paragraphe IV.2.8 est globalement bien suivi, celui-ci n'est pas garanti : par exemple, une variable issue de la fréquence des vols sur les 5 000 voisins les plus proche (fortement corrélée avec les coûts) est utilisée pour stationnariser le signal.

### Utilisation de codes postaux

Bien entendu, l'utilisation de code postaux limite à la fois la précision des prédictions de risques, mais aussi le nombre de variables utilisables. En particulier, il n'est plus possible d'exploiter les variables plus précises que les codes postaux, dont les distances aux différents points d'intérêt.

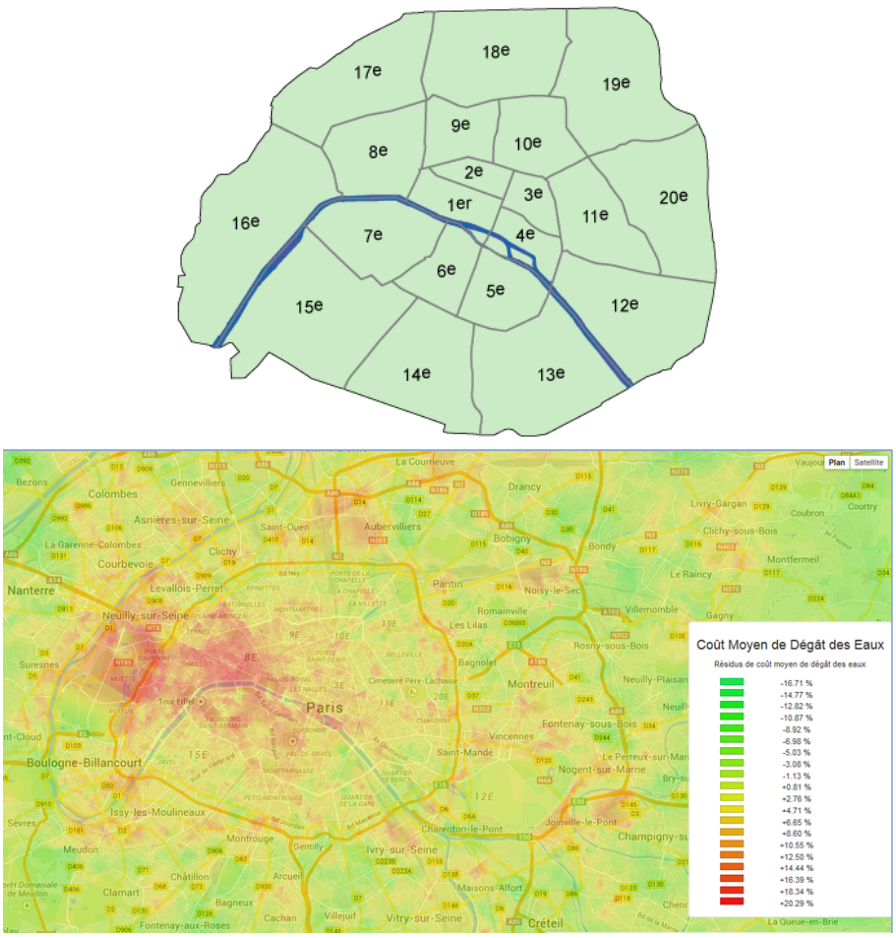



FIGURE IV.4.2 – Précision du zonier
Exemple de gain de précision du zonier créé. L'utilisation d'une géo-localisation précise de l'adresse des polices permet de passer d'une segmentation du risque par arrondissement (en haut) à une segmentation continue (en bas), améliorant ainsi sensiblement la précision des prédictions.
La carte du bas représente une estimation du coût de dégâts des eaux en région parisienne - comme l'indique la carte en sur-impression...

On peut aussi noter que, du fait de l'utilisation de codes postaux, le risque modélisé peut être fortement discontinu (à la frontière entre deux codes postaux). Cette discontinuité, peu importante en zone rurale (où la frontière entre deux codes postaux est généralement peu dense) peut être fortement contre-intuitive pour les clients en zone urbaine.

## IV.4.2 Modèles créés

### IV.4.2.a Définition des meta-paramètres

Les modèles utilisés reposent sur un certain nombre de meta-paramètres définis a priori (nombre d'arbres, complexité de ceux ci...).
Cependant, la création de modèles étant automatisée, il est relativement simple de tester un grand nombre de meta-paramètres (à l'aide de validations croisées, comme décrit paragraphe II.2.2.b) et de sélectionner les plus performants. Cette méthode d'action - par force brute, couramment appelée grid-search, a été utilisée pour la sélection des modèles construits.
Les modèles sélectionnés ont pu être, ensuite, testés sur des nouvelles données.
Afin de réaliser efficacement les tests permettant de trouver les paramètres souhaités, nous avons pu directement mesurer les performances des modèles sur les prédictions des résidus de risques, sur une validation croisée, et retenir les paramètres optimaux.

### IV.4.2.b Importance des variables

Comme décrit chapitre III.4.2.b, il est possible d'estimer l'impact de chacune des variables (grâce au gain de variance obtenu sur les nœuds où elles apparaissent).
L'importance des variables des modèles dont les performances ont été présentées ci dessus permet de confirmer l'intuition décrite ci dessus sur la stationarisation du signal (chapitre IV.4.1.b). Le signal semble bien être principalement décrit par les variables de type K plus proches voisins, avant que ses résidus (stationaires) ne soient expliqués par les variables attachées aux différentes polices.

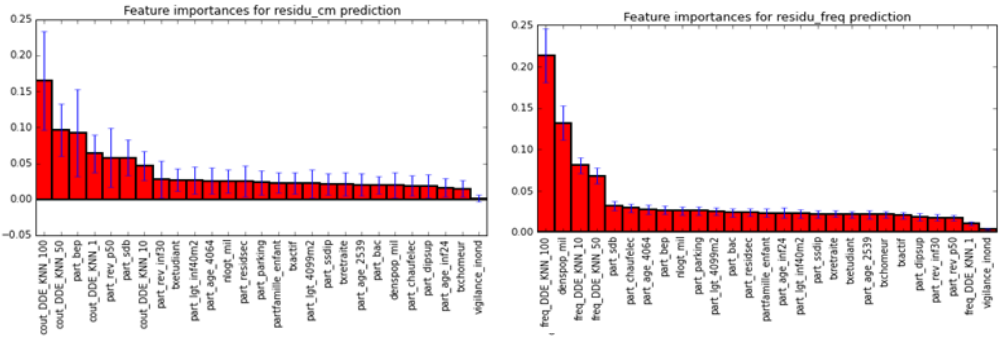



FIGURE IV.4.3 – Importance des variables
L'importance des variables de modèles de coût ou fréquence de dégâts des eaux met clairement en évidence l'importance des variables de type K-plus-proches-voisins (variables nommées freq_DDE_KNN_k ou CM_DDE_KNN_k). Ensuite, des variables liées à la richesse de l'environnement (niveau d'étude, part de revenus élevés...) puis les propriétés du logement ont un fort impact dans les modèles.

## IV.4.3 Impact sur le loss-ratio

### IV.4.3.a Evolution de la prime moyenne proposée

Un zonier repose sur un modèle prédisant les résidus de modèles linéaires généralisés, basés sur des variables déclaratives (comme décrit section IV.3.1. Si ces GLM sont non-biaisés (ce qui est normalement le cas), il "suffit" de ne pas introduire de biais lors de la création du zonier (comme décrit chapitre IV.3.3).
Cette tâche étant relativement aisée (mais nécessitant une attention particulière aux pondérations utilisées lors de la construction du modèle), les modèles intégrant les différents zoniers sont non biaisés et reflètent effectivement les charges observées.
De même que pour la création de modèles de risques en utilisant différentes méthodes (comme décrit chapitre III.4.3.a), l'utilisation correcte d'une méthodologie ou d'une autre n'impacte, par design, pas les primes moyennes estimées.

#### Coefficient de Gini

**Performances globales du modèle** Contrairement à la prime pure moyenne estimée, la performance du modèle dépend fortement de la méthode employée.
Dans l'étude que nous avons réalisée, la méthode utilisée avait un impact significatif sur les performances observées. Ces performances sont présentées table IV.4.4. Il est clair que la méthode proposée permet une amélioration sensible de la précision des modèles de risques.

Bien entendu, il est important de noter que ce type d'approche n'a de sens que si les évènements modélisés sont indépendants les uns des autres (ce qui exclut les catastrophes naturelles, où un même évènement impacte plusieurs polices).
La prédiction de ce type d'évènement à l'aide des méthodologies décrites dans ce mémoire risque, en effet, de non seulement créer des modèles peu performants, mais aussi de créer des problèmes de sur-apprentissages (les bases d'apprentissage et de test n'étant plus indépendantes), ce qui amènerait à une forte sur-estimation des performances des zoniers, et à une anti-sélection désastreuse lors de la mise en production effective des nouveaux tarifs créés.

| Type de zonier | Pas de zonier | Ancienne méthode | Méthode proposée |
|---|---|---|---|
| Coefficient de Gini | 33% | 45% | 49% |

FIGURE IV.4.4 – Prime pure moyenne - dégât des eaux

Impact de la méthodologie de construction du zonier sur les performances observées.
De même que pour la section III.4.3, les données utilisées ont été modifiées afin de bruiter les résultats obtenus - les performances exactes des modèles étant confidentielles. De plus, ces résultats ne s'appliquent qu'à une seule de nos couvertures.

Malgré ces précautions, la comparaison des différents types de zoniers reste valable, et indique un gain clair à utiliser la méthode de micro-zonier décrite dans ce mémoire.

**Impact des différents types de variables** En plus des modèles créés à l'aide de l'ancien zonier, 3 différents types de modèles ont étés comparés dans le cadre de notre expérience :

— Un modèle sans aucune variable liée à la géographie (performance sans zonier).
— Un modèle incluant un zonier, basé sur les variables descriptives liées à la géographie (variables zonées ou distances à des points d'intérêt).
— Un modèle incluant un zonier, basé sur toutes les variables disponibles (y compris les variables de types K plus proches voisins, ainsi que les mesures locales de densité).

Cette décomposition permet de constater clairement l'effet des types de variables géographiques une par une.

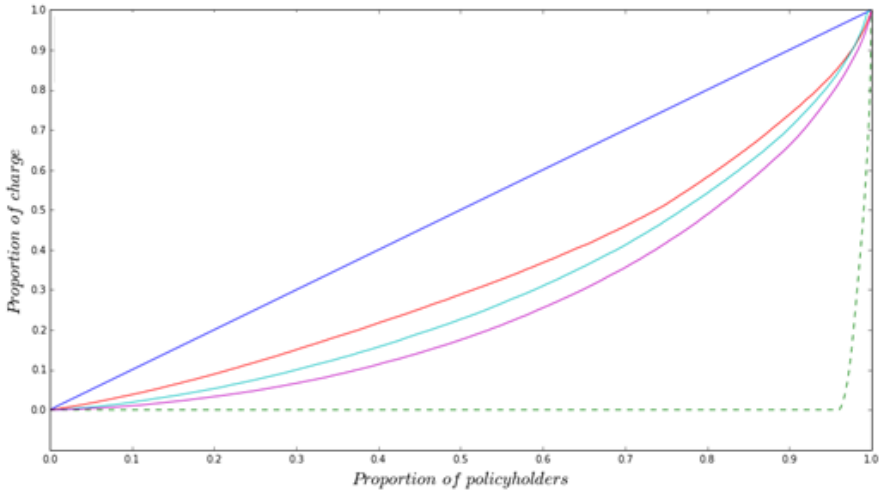



FIGURE IV.4.5 – Performances du zonier
Apport du zonier pour l'ensemble du portefeuille appartement.
L'indice de Gini des modèles construits passe de 33% sans zonier (en rouge), à 41% avec un zonier ne comprenant que les variables géographiques (en bleu clair) et 49% avec un zonier comprenant les variables géographiques et les variables de type K plus proches voisins.
Ces résultats ont bien entendu été obtenus sur des ensembles de validation ("out of sample"), indépendants des ensembles d'apprentissage des modèles ou de l'ensemble sur lequel a été réalisé la validation croisée nécessaire à la définition des paramètres optimaux.

**Performances dans divers types d'habitats** Il est intéressant de noter que le zonier construit reste performant quelque soit la densité de polices présentes dans la zone considérée (cf. figure IV.4.6). Cette robustesse permet d'exploiter la méthodologie proposée dans ce mémoire sur des zones fortement hétérogènes.

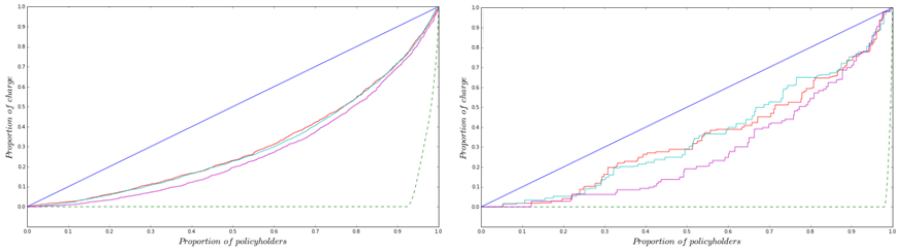



FIGURE IV.4.6 – Performances du zonier - forte et faible densité
Le zonier permet d'améliorer la performance des modèles dans les zones à forte densité d'habitation (à gauche, dans le 17ème arrondissement de Paris, gain de 7% d'indice de Gini), ou les zones à très faible densité (à droite, dans le département de l'Ariège, gain de 13% d'indice de Gini).
Les courbes correspondent à celles décrites figure IV.4.5.

**Coefficient de Gini relatif et gain de loss-ratio** :

Comme pour tout changement tarifaire, il est possible d'estimer l'impact d'une modification du zonier à l'aide du coefficient de Gini relatif.

Dans le cas du changement proposé ici, l'impact en terme de coefficient de Gini est déjà extrêmement sensible. Comme on peut le voir figure IV.4.7, cet impact se traduit par un Gini relatif important, reflétant une forte amélioration potentielle du loss-ratio.

Bien entendu (comme rappelé section III.4.3.a), ces performances doivent être prises avec précaution :

— Ils reposent sur des hypothèses très fortes.
— Ils sont issus de back-tests : une erreur dans ceux ci (par exemple, la non-indépendance des sinistres enregistrés dans la base, comme c'est par exemple le cas lors de catastrophe naturelle ou d'évènements climatiques) entraine une sur-estimation de la qualité des modèles et l'invalidité des résultats mesurés.

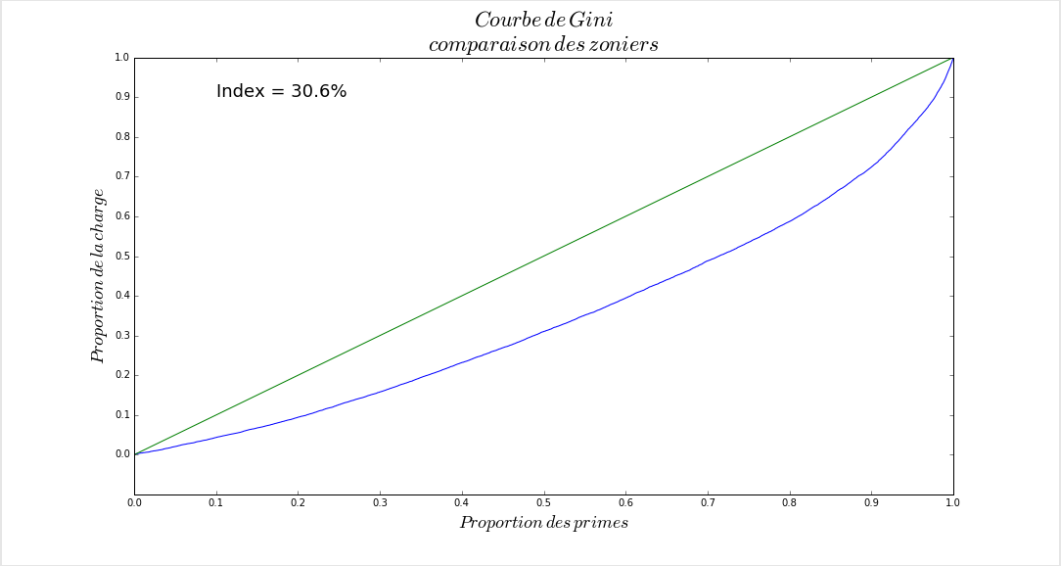



FIGURE IV.4.7 – Gini relatif d'un tarif construit en utilisant un micro-zonier développé en suivant la méthode décrite dans ce rapport, contre un zonier simple.

L'indice de Gini permet, comme décrit paragraphe II.3.4.c, de comparer deux stratégies tarifaires.

Dans cet exemple, un tarif exploitant un zonier construit suivant la méthode présentée dans ce mémoire permet de capturer 50% des primes en ne faisant face qu'à 31% des sinistres.
Selon les hypothèses fortes de comportement des clients décrites chapitre II.3.4.c, le gain de loss-ratio serait dans ce cas réduit de 38%. Ce gain représente une forte sur-estimation du gain réel, et a lieu au prix d'une forte réduction de la taille du portefeuille, mais la comparaison reste fortement en faveur de la méthodologie proposée.

Dans le cas des modèles considérés ici, les gains de loss-ratio attendus restent tout de même très sensibles.

### IV.4.3.b Evolution des primes pures

L'analyse d'impact de la méthode sur les primes pures proposées est primordiale lors de la mise en place d'un nouveau zonier.
En effet, AXA France propose un système de distribution reposant sur des agents ; chaque agent dispose d'une (ou plusieurs) boutiques, et profite donc d'une zone de chalandise naturelle autour de celle ci. La modification du zonier, en modifiant les primes pures et les tarifs commerciaux pour une zone géographique, modifiera donc fortement la compétitivité des produits AXA autour des différentes boutiques, et donc aura un impact direct et explicite sur les profits des différents agents.
Certains agents bénéficieront donc de la nouvelle tarification, d'autres en pâtiront fortement : la mise en place d'une telle évolution du tarif ne peut donc se faire sans les consulter.

Avant de lancer une telle consultation, nous pouvons estimer l'importance pour les différents clients des changements tarifaires proposés.
Comme lors du choix d'une méthode de régression ou une autre, présenté section III.4.3.b, ces changements touchent diversement les différents clients : certains clients voient leurs prix peu varier, d'autres subissent des changements très importants (avantageux ou non). Mais, contrairement au cas où ces changements sont dus aux profils propres des clients - hors leurs lieux de résidences - l'impact d'un zonier se reportera de plein fouet sur le portefeuille des agents, ceux ci ne pouvant "diversifier" leur exposition à ces changements de tarifs.
Ce problème de diversification est commun à tous les travaux liés à la géographie, pour des sociétés d'assurance intermédiées : il n'est donc pas propre à la création d'un micro-zonier, mais, en particulier dans les cas où la zone de chalandise d'une agence est petite devant le code-postal dans lequel il réside (donc dans les zones fortement peuplées) le problème est exacerbé.

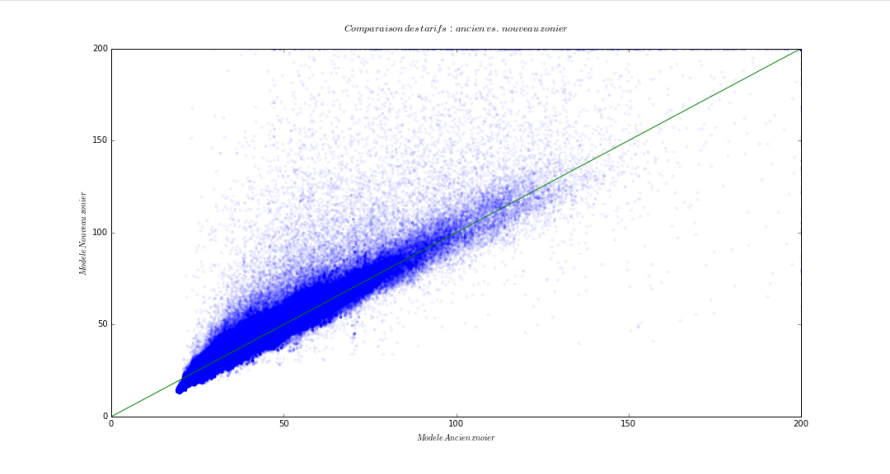



FIGURE IV.4.8 – Comparaison des tarifs produits à l'aide de deux zoniers (l'ancien zonier étant sur l'axe des abscisses, le nouveau sur celui des ordonnées.

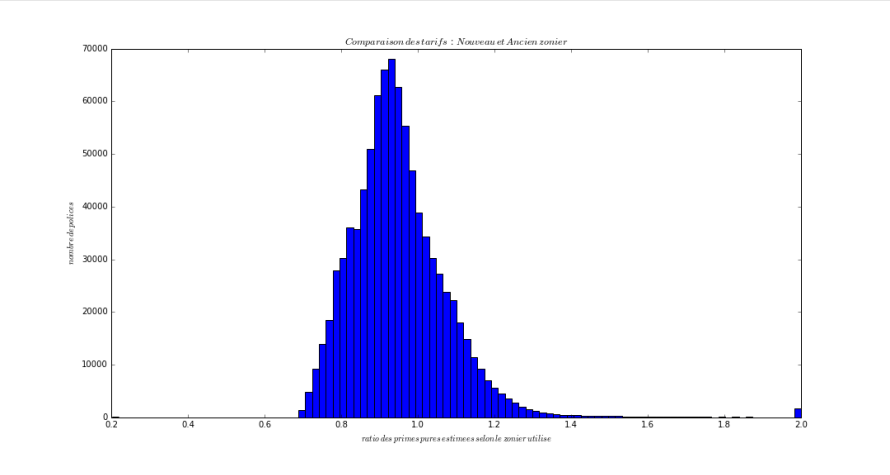



FIGURE IV.4.9 – Histogramme de l'évolution des tarifs entre les deux zoniers - une valeur de 1 dénote un prix constant.

On peut remarquer que les nouvelles primes pures ont une asymétrie (skewness) légèrement plus importante que les anciennes : le mode de leur distribution est légèrement en dessous de la diagonale, et le nombre de valeurs sensiblement plus élevées est plus important : moins de 5% des primes pures baissent de 20%, alors que 5% des primes pures augmentent de plus d'un quart ; l'évolution médiane est une baisse de 6%, et la prime pure moyenne est, bien entendu, inchangée.
Ce changement de distribution reflète une plus grande segmentation des prix (une plus grande variabilité des primes pures) et est donc relativement attendu, dans la mesure où le nouveau zonier est nettement plus fin que l'ancien.
Certaines couvertures sont beaucoup plus impactées par le changement de zonier que d'autres.

## IV.4.4 Limites de l'approche proposée

L'approche décrite dans ce mémoire propose un certain nombre d'innovations, permettant de résoudre des problèmes majeurs liés à l'utilisation de données géographiques et de réaliser un zonier d'une granularité aussi fine que souhaitée.
Cependant, un certain nombre de faiblesses sont apparues lors de la création d'un tel zonier.

### IV.4.4.a Complexité de la mise en œuvre

La réalisation d'un micro-zonier suivant la méthode proposée est relativement lourde.
En particulier, cette réalisation requiert un bon niveau en informatique (création de code long, optimisation d'algorithmes, manipulation de plusieurs langages au sein d'un projet...)
Ce type de compétences étant relativement rare au sein de la communauté actuarielle (les actuaires étant plus habitués à utiliser des outils graphiques pour réaliser leurs modèles), il est difficile de mettre en œuvre un projet de cette complexité sans créer un risque important de voir apparaitre des bugs informatiques impactant la qualité du modèle proposé (potentiellement sans que les actuaires en charge du projet ne s'en rendent compte, si les méthodes de test mises en place ne sont pas fiables).
De ce fait, la création d'un micro-zonier a nécessité des ressources importantes (et la question de la rentabilité d'un tel projet se pose donc). Ce projet a été l'occasion d'une montée en compétence des équipes de tarification d'AXA France sur les problématiques liées à l'informatique et à l'intelligence artificielle.

De plus, l'intégration des cartes créées dans les systèmes opérationnels de la société a nécessité un projet sensiblement plus complexe que prévu (par exemple, les moteurs SQL utilisés ne permettaient pas l'envoi de requêtes GIS nécessaires à la réalisation de ce projet - cf. IV.2.3.a).

### IV.4.4.b Opacité du modèle créé

#### Compréhension et validation des modèles

Bien qu'il soit possible d'estimer l'impact des différentes variables exploitées dans le zonier créé (cf. III.4.2.b), le rôle de chacune d'entre elle n'est pas explicite.
Cette faiblesse, inhérente aux modèles d'arbres de régression, est réellement gênante dans le cadre de la création d'un tarif d'habitation : les performances du tarif étant d'une importance capitale pour l'entreprise d'assurance, et leur confirmation requérant une longue durée d'exposition, il est vivement souhaitable de pouvoir vérifier a priori la pertinence des modèles créés.

De plus, comme indiqué plus haut, (cf. IV.4.1.b), les modèles créés peuvent être décomposés en 2 parties, l'une stationarisant le signal, et l'autre estimant les résidus du signal stationnarisé à l'aide des variables géographiques.
Cette décomposition, si elle semble bien avoir lieu, n'est pas explicite. Il serait donc souhaitable de disposer d'un modèle décomposant clairement ces deux actions, afin de garantir sa justesse théorique.

Cette incapacité à clairement comprendre le modèle créé est un fort handicap pour les différentes techniques d'apprentissage statistique disponibles sur le marché. La société d'assurance étant privée de moyen de vérification direct de la pertinence des modèles créés, elle doit être extrêmement attentive lors de la mise sur le marché de nouveaux tarifs.

#### Communication avec les réseaux de distribution

Comme indiqué plus haut (IV.4.3.b), l'évolution du tarif - et particulièrement du zonier utilisé - a un impact majeur sur les revenus des distributeurs des contrats d'assurance.
Afin de faire accepter les nouvelles grilles tarifaires, il est extrêmement utile de pouvoir motiver les changements réalisés. Dans ce cadre, l'utilisation de modèles non-explicites est un handicap majeur, et un long travail de visualisation des modèles créés est nécessaire.

### IV.4.4.c Impact éthique des techniques employées

Les méthodes décrites dans ce mémoire soulèvent deux questions majeures d'un point de vue éthique, sur lesquelles nous conclurons notre étude.

#### Utilisation implicite de critères non-éthiques

L'une des premières difficultés soulevées sur le plan éthique par les technologies décrites dans ce mémoire concerne l'utilisation, parfois implicite, de critères prohibés pour des raisons éthiques (sexe, origine, religion, apparence...)
Un certain nombre d'informations liées à la géographie peuvent recouper de tels critères, et il est possible de créer, sans s'en rendre compte - à cause de la non-transparence des techniques employées - des tarifs discriminant sur des motivations discutables voir illégales.
Dans le cadre de ce mémoire, un zonier "trop" précis peut aisément isoler un quartier ou une rue défavorisée par exemple pour des raisons communautaires, et potentiellement renforcer un isolement déjà en place. Ce point d'attention, présent dès que l'actuaire réalise une segmentation géographique, devient plus important à mesure que cette segmentation s'améliore.
De manière générale, les données posant des problèmes éthiques sont les propriétés du client dont celui-ci n'est pas responsable (le sexe, l'origine ou l'apparence par exemple ; le client n'est pas responsable non plus de son age, mais comme tous les clients passent par les différents ages, cette variable n'offre pas d'avantage à une catégorie particulière de personnes...) Le lieu de résidence peut entrer dans ces critères, dans la mesure où la mobilité géographique est limitée (et où cette absence de mobilité reflète souvent une absence de mobilité sociale).
D'autres méthodes de tarification fondées sur l'exploitation de données externes ou non-conventionnelles peuvent s'avérer encore nettement plus gênantes, des techniques exploitant l'analyse textuelle ou audio (hors du cadre de ce mémoire, mais exploitant des technologies comparables) de documents fournis par les clients peuvent rapidement "découvrir" des motifs liés aux origines des personnes considérées.

Afin d'éviter de tomber dans ce type de travers il est primordial d'étudier en détail les modèles créés, et d'utiliser, autant que possible, des méthodes permettant de contrôler le tarif produit.

#### Sur-segmentation des clients

Un autre aspect du travail de l'actuaire qui soulève, particulièrement depuis l'apparition de nouvelles sources de données (par exemple, l'exploitation de bases de données ouvertes, comme dans ce mémoire, ou bien d'informations issues d'objets connectés, en assurance auto), et l'appropriation des différentes techniques récentes d'apprentissage statistique, est la sur-segmentation du risque.
En effet, comme rappelé en début de mémoire, l'assurance repose sur une mutualisation des risques (et donc un transfert, a priori aléatoire, de ressources).
Chaque amélioration de la segmentation des clients amène inévitablement une réduction de ces

transferts. Cette diminution peut être perçue comme une disparition du rôle de l'assurance.

Cette question peut faire l'objet d'un mémoire entier - qui serait sans doute passionnant ! Il est cependant possible de rappeler que ce mémoire ne travaille qu'à l'estimation de la prime pure, là où les questions éthiques s'appliquent en général aux primes commerciales. La connaissance du risque par l'assureur n'implique pas nécessairement une exploitation de cette connaissance. Il appartient donc au régulateur de garantir que la segmentation des assureurs continue à ne pas reposer sur des critères illicites, et qu'elle permette à tous les clients potentiels un accès abordable à une couverture.

# Chapitre IV.5

# Conclusion

Dans ce mémoire, nous avons pu présenter différents éléments d'apprentissage statistique et les mettre en pratique dans le cadre de la construction d'un zonier en assurance habitation.

Tout d'abord, nous sommes revenus sur les bases de l'apprentissage statistique. Ce retour aux sources a été l'occasion de rappeler des éléments théoriques importants dans le cadre d'un travail de modélisation : le compromis biais-variance et manières de le résoudre, en évitant du sur-apprentissage, et les différentes méthodes de mesure de la performance de modèles.

Ensuite, nous avons pu présenter divers modèles utiles en actuariat (GLMs, arbres de régression et forêts aléatoires). Les fondements théoriques de ces modèles ont pu être présentés, et leur performance mesurée à l'aide des concepts définis dans le premier chapitre. Nous avons aussi pu tenter d'estimer l'impact pratique, en terme de loss-ratio, de traçabilité ou d'implémentation, de l'utilisation d'un modèle par rapport à un autre.

Enfin, nous avons pu appliquer les modèles présentés à la construction d'un zonier. Pour cela, nous avons présenté l'approche globale de construction d'un zonier, les difficultés spécifiques à cet exercice, et proposé une méthode permettant de construire des zoniers pour les différents risques attritionnels couverts par la garantie MRH. Enfin, nous avons pu vérifier la performance des modèles créés, et de nouveau tenter d'estimer les conséquences de l'utilisation des méthodologies décrites dans ce mémoire, tant en terme de loss-ratio que de déploiement pratique.

Les méthodes décrites dans ce mémoire ont été développées il y a déjà plusieurs années pour les équipes de tarification MRH d'AXA France. Depuis, ces équipes se sont approprié certaines parties des technologies présentées, en ont modifié d'autres, et ont largement exploité le travail réalisé pour la création de leur micro-zonier. Outre les impacts pratiques (les nouveaux zoniers sont actuellement en production chez AXA France), plusieurs mémoires d'actuariat ([14], [18]) ont également été écrits sur ce sujet. Leur lecture donne un recul intéressant sur l'appropriation par les équipes de tarification des méthodes présentées dans ce document.

# Bibliographie


[1] V. Vapnik BE. Boser, I. Guyon. A training algorithm for optimal margin classier. *Proceedings of the fifth annual workshop on Computational learning theory*, 1992.

[2] C. Bishop. *Pattern Recognition and Machine Learning*. 2007.

[3] L. Breiman. Parsimonious binary classification tree. *Technology Service Corporation*, 1978.

[4] L. Breiman. Bagging predictors. *Machine Learning*, 1996.

[5] L. Breiman. Statistical modeling : Two cultures. *Statistical Science*, 2001.

[6] P. Domingos. A few useful things to know about machine learning. *Communications of the ACM*, 2012.

[7] C. Havas-Dauphine F. Pedregosa, M. Blondel. Lightning, a library for large scale machine learning in python. 2016.

[8] E. Frees, G. Meyers, and A. Cummings. Insurance ratemaking and a gini index. *Journal of Risk and Insurance*, 2012.

[9] JH. Friedman. A recursive partitioning decision rule for nonparametric classification. *Computers, IEEE Transactions*, 1977.

[10] JH. Friedman. Greedy function approximation : a gradient boosting machine. *The annals of Statistics*, 2001.

[11] S. Geman, E. Bienenstock, and R. Doursat. Neural networks and the bias/variance dilemma. *Neural computation*, 1992.

[12] J. Gill. Generalized linear models : A unified approach. *Sage University*, 2000.

[13] P. Legendre. Spatial autocorrelation : trouble or new paradigm ? *Ecology*, 1993.

[14] C. Loiret. *Refonte du tarif multirisque habitation : construction de micro-zoniers et intégration de la sinistralité passée à l'adresse*. 2016.

[15] G. Louppes. *Understanding Random Forests*. PhD thesis, Université de Liège, 2009.

[16] P. McCullagh and JA. Nelder. Generalized linear models. *Chapman and Hall/CRC*, 1989.

[17] P. Moran. Notes on continuous stochastic phenomena. *Biometrika*, 1950.

[18] J. Pariente. *Modélisatoin du risque géographique en assurance habitation*. 2017.

[19] D. Silver. Mastering the game of go with deep neural networks and tree search. *Science*, 2016.

[20] J. Friedman T. Hastie, R. Tibshirani. *the Elements of Statistical Learning*. 2007.

[21] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, 1996.

[22] A. Tikhonov. On the stability of inverse problems. *Proceedings of the USSR Academy of Sciences*, 1943.

[23] T. Westling. Male organ and economic growth : Does size matter ? *University of Helsinki*, 2011.