Mémoire présenté le :

pour l'obtention du Diplôme Universitaire d'actuariat de l'ISFA et l'admission à l'Institut des Actuaires

Par : MARLIER AURELIA

Titre : Zonage d'un risque à événements rares : l'inondation

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

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

Membre présents du jury de l'Institut des Actuaires signature

Membres présents du jury de l'ISFA

Entreprise :

GMF ASSURANCES 148 rue Anatole-France 92597 LEVALLOIS-PERRET Cedex SOCIÉTÉ D'ASSURANCES

Nom : GMF

Signature :

Directeur de mémoire en entreprise :

Nom : TISSERAND ANGELIQUE

Signature :

Invité :

Nom :

Signature :

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

Signature du responsable entreprise

Signature du candidat

Résumé

Le mémoire a été réalisé dans le cadre d'une étude sur l'inondation pour la compagnie d'assurance GMF. L'étude présente la construction d'un zonier prime pure inondation à la maille code postal. Elle ne s'appuie pas sur une approche physique de l'événement climatique mais sur une approche statistique du risque. Il faut bien comprendre qu'elle n'anticipe pas la possibilité d'événements non encore survenus sur notre période, contrairement à l'approche physique.

La méthodologie pour le zonier est classique. Tout d'abord, l'effet géographique est contenu dans les résidus d'un modèle de prime pure estimé sans variables géographiques. Puis, cet effet est lissé sur le territoire métropolitain entre les différents codes postaux. Enfin le zonier final est déterminé par classification des codes postaux en classes de risque.

L'inondation étant un risque climatique à événements rares, le nombre de sinistres observés est faible entre 2006 et 2016. L'intérêt de ce mémoire est la mise en place de méthodes permettant la meilleure utilisation de cette information sinistre restreinte et atypique.

Mots-clés : Risque inondation, zonier, modèles linéaires généralisés, *response-based-sampling*, forêts aléatoires, lissage prédictif, classification ascendante hiérarchique

Abstract

This thesis has been carried out as part of a study for the insurance company GMF. The study shows the building of a pure premium zoning for flood risk. It is not based on a physical approach of climatic event but on a statistical one of the risk. It must be understood that it does not allow to model an event not occured in our historical data yet, on the contrary to the physical approach.

The methodology is classic for zoning. First, the geographical effect is residuals of a pure premium model estimated without geographical variables. Then, the effect is smoothed with the other postcodes for the entire metropolitan territory. Lastly, the final zoning is a classification of postcodes in risk classes.

Flooding is a climatic risk with few events occured so there a few number of claims between 2006 and 2016. The thesis' interest is the application of methodologies that allow the best utilisation of limited and unconventional claim information.

Keywords : flood risk, zoning, generalized linear models, response-based-sampling, random forests, predictive smoothing, hierarchical clustering

Remerciements

Je tiens à remercier la GMF de m'avoir permis de réaliser mon mémoire d'actuariat dans le service Étude statistiques produits. Mes remerciements s'adressent tout particulièrement à mon encadrante Angélique Tisserand et à mon manager Alexandre Daubas ainsi que le pôle MRH pour leurs conseils dans la réalisation de cette étude.

J'adresse également mes remerciements à Didier Ruillière et Xavier Milhaud pour l'encadrement théorique et leurs réponses à mes différentes questions.

Je voudrais enfin remercier mon entourage pour leurs conseils de rédaction et la relecture de ce mémoire d'actuariat.

Table des matières

Introduction

Depuis 50 ans, les inondations en France métropolitaine sont de plus en plus fréquentes. Les dommages causés par cet événement climatique n'ont cessé d'augmenter. Les installations sont plus sophistiquées et l'urbanisation est plus importante, ce qui entraîne un accroissement des indemnisations. En France, ces sinistres sont couverts sous le régime légal des catastrophes naturelles. Or, il arrive que certaines personnes affectées ne puissent pas être remboursées à ce titre. Ainsi, certains assureurs comme la GMF (Garantie Mutuelle des Fonctionnaires) propose une garantie inondation à ses assurés dans son contrat multirisque habitation (MRH) afin de les dédommager dans ce cas de figure. Pour éviter une forte surexposition, il est important de connaître ce risque et de proposer une tarification adaptée. Le risque inondation est très disparate sur le territoire. L'un des éléments majeurs dans la tarification du risque inondation est la segmentation géographique. Afin d'obtenir un tarif concurrentiel et d'éviter la surexposition, l'assureur doit connaître ces zones d'exposition et les intégrer à son tarif par l'intermédiaire d'un zonier.

Il existe deux approches pour la modélisation d'un événement climatique. La première s'appuie sur des modèles physiques. Elle se décompose en 3 modules : l'aléa, la vulnérabilité et les dommages. L'aléa est un catalogue d'événements climatiques survenus ou fictifs. Ils sont déterminés à l'aide de données météorologiques. La vulnérabilité est l'ensemble des risques du portefeuille étudié, leurs caractéristiques croisées avec les données aléa par la géolocalisation. Les dommages permettent d'estimer la sinistralité de l'assureur à l'aide de la vulnérabilité, et d'y appliquer les conditions contractuelles pour conserver ce qui est à la charge de l'assureur. Cette approche nécessite des connaissances physiques et des données fiables. Ceci est généralement difficile à mettre en place par un actuaire. C'est pour cela que les assureurs se tournent vers les logiciels AIR, RMS et EQUECAT qui leur offrent cette modélisation. Ces logiciels étant trop coûteux, nous avons choisi la deuxième approche possible, par statistiques inférentielles, qui est une décomposition fréquence $\times$ sévérité, plus communément appelé en tarification fréquence $\times$ coût moyen. De plus, l'inondation est un risque d'assurance à événements rares, ainsi il existe peu de sinistres observés sur l'intervalle d'observation. L'objectif de ce mémoire est de construire un zonier inondation en exploitant au mieux la faible information à notre disposition. Nous n'aborderons pas l'éventualité de l'intégration de ce zonier dans la tarification, ainsi aucune étude de l'impact du zonier sur les cotisations ne sera menée.

Nous présenterons dans un premier temps les spécificités du risque inondation

dans le cadre de l'étude. Dans le second chapitre, nous aborderons la modélisation de l'effet géographique. L'objectif est d'utiliser au mieux les sinistres recensés à l'aide d'une méthodologie adaptée aux événements rares. Puis le dernier chapitre contient la construction finale du zonier. L'effet géographique est lissé par forêts aléatoires utilisant des variables externes géographiques. Les classes de risque sont ensuite déterminées par classification.

Pour des raisons de confidentialité, les variables internes utilisées pour l'estimation de l'effet géographique ne sont pas dévoilées pour le respect des variables tarifaires du contrat MRH GMF. De plus, les indicateurs de sinistralité sont exprimés en indice par rapport à la moyenne afin de ne pas révéler les valeurs de sinistralité de l'entreprise.

Chapitre I

Le contexte général de l'étude

Dans ce premier chapitre, nous présenterons les choix et les hypothèses faits pour le périmètre des données utilisées et du zonier inondation construit.

1 L'inondation et son processus d'indemnisation

L'inondation est "une submersion temporaire, par l'eau, de terres qui ne sont pas submergées en temps normal, quelle qu'en soit l'origine", comme définie sur le site de l'État français [1]. L'événement climatique est provoqué par différentes origines qui permettent de le classifier en quatre catégories :

- montée lente des eaux en région de plaine (remontées de nappes ou sortie de la rivière de son lit moyen ou de son lit majeur) ;
- crues torrentielles rapides consécutives à des averses violentes (concentration rapide dans le cours d'eau) ;
- ruissellement en zone urbaine (l'imperméabilisation du sol empêche l'eau de s'infiltrer) ;
- submersion marine en bord de côte.

Ne possédant pas l'information sur le type d'inondation, la sinistralité retenue dans le cadre de l'étude comprend tous les types d'événements même si leur modélisation physique est différente. L'événement déclencheur d'une inondation est en général soit une période de pluies violentes sur un sol sec ou urbain, soit des périodes répétées à faible intervalle de temps qui provoquent une saturation en eau des sols. La moitié des catastrophes naturelles mondiales recensées sont des inondations. La France est elle-même un pays exposé du fait de son réseau hydraulique important, de son profil météorologique avec une possibilité d'orages et de fortes pluies. En témoignent ces dernières années où la présence de crues a augmenté avec des épisodes pluvieux dans le Var, dans l'Ouest à la suite de la tempête Xynthia ou encore plus récemment la crue de la Seine en mai-juin 2016. La gravité de ces épisodes est ac-

centuée par le développement urbain et économique. Les biens endommagés sont de plus en plus coûteux. Les surfaces recouvertes de bitume absorbent moins les fortes pluies et entraînent un ruissellement de l'eau. Les dommages causés ne sont pas que des dommages aux biens mobiliers et immobiliers. Ce type de catastrophe affecte également les professionnels dans leur activité quotidienne par une perte d'activité ou un chômage technique. Du fait de l'augmentation de la fréquence de ces catastrophes ainsi que du coût d'une inondation, il faut maîtriser et couvrir ce risque. L'inondation est, en conséquence, importante de considération pour un pays exposé comme la France.

En réponse à ces enjeux, le gouvernement met en place des plans de prévention pour les personnes se trouvant en zones reconnues inondables. Sous l'impulsion de l'Union Européenne, l'État français a reconnu en son sein des TRI (Territoires à Risques d'Inondation). Ce sont des parties du territoire ayant des prédispositions à subir des inondations. Il en existe également dans les DROM-COM. L'une des actions majeures pour faire face aux catastrophes naturelles, est la mise en place par la loi du 13 juillet 1982 du régime d'indemnisation des catastrophes naturelles pour les compagnies d'assurance. Ceci implique que tout assuré bénéficiant d'un contrat assurance dommages est automatiquement couvert contre les dégâts dus aux catastrophes naturelles. On définit, dans le Code des Assurances [2], les impacts d'une catastrophe naturelle comme "les dommages matériels directs non assurables ayant eu pour cause déterminante l'intensité anormale d'un agent naturel lorsque les mesures habituelles à prendre pour prévenir ces dommages n'ont pu empêcher leur survenance ou n'ont pu être prises". Il existe néanmoins quelques exclusions. Par exemple, le régime couvre très peu les pertes pour une activité professionnelle. Pour assurer cette couverture, la cotisation au régime est réglementée. Ce montant s'ajoute à la cotisation de chaque contrat ayant une garantie dommage. Elle est égale, aujourd'hui, à 12% du montant de la cotisation principale. Une franchise de 380 euros existe pour les biens à usage d'habitation et non professionnel. Pour bénéficier d'un remboursement au titre de ce régime, le processus doit suivre les étapes suivantes :

- Le maire a 18 mois à partir de la date de l'événement pour déposer un dossier de demande de reconnaissance de catastrophe naturelle au préfet.
- Le préfet dépose ensuite un dossier auprès de la Commission interministérielle.
- Le journal officiel publie la liste des communes faisant l'objet d'un arrêté à la suite de l'étude des dossiers par la Commission.
- L'assuré déclare son sinistre, au plus tard, 10 jours après la publication au journal officiel.
- L'expert évalue le coût des dommages.
- L'assureur indemnise l'assuré sous la garantie Cat-Nat.

Ce système d'indemnisation n'est pas infaillible. La reconnaissance d'une catastrophe naturelle au Journal officiel dépend de la demande du maire et de l'action du préfet. Une équipe spécialisée sur l'analyse des risques climatiques au sein de Covéa (entreprise regroupant MAAF, MMA et GMF) a mis en évidence la différence entre les arrêtés Cat-Nat et l'impact réel de l'événement. En effet, certains maires ne vont pas demander l'arrêté car très peu de résidents sont touchés alors que d'autres en font la demande même si aucun sinistre n'est recensé. L'arrêté Cat-Nat ne garantit donc pas l'indemnisation de toutes les victimes d'une inondation.

Depuis 2012, il existe la garantie inondation dans les contrats MRH de la GMF pour compléter le régime légal des Cat-Nat Elle prend en charge l'indemnisation des assurés, victimes d'une inondation qui sont dans l'impossibilité de faire fonctionner la garantie Cat-Nat car la commune n'a pas bénéficié d'un arrêté publié au Journal officiel. Cette garantie s'applique pour les maisons et appartements en résidences principales ou secondaires. Elle assure les dommages aux biens mais aussi les frais de relogement, si l'assuré ne l'est pas gratuitement.

Dans le cadre de son contrat MRH, l'indemnisation d'un sinistre inondation peut s'effectuer soit au titre du régime Cat-Nat soit au titre de la garantie inondation.

2 Présentation des données et traitements préliminaires

Le périmètre

En conséquence des garanties considérées, le périmètre de l'étude se restreint aux maisons et appartements en résidences principales ou secondaires. Géographiquement, on se limite à la France métropolitaine (Corse incluse). Les DROM-COM sont exclus car nous possédons moins d'informations et le profil climatique dans ces territoires n'est pas le même que celui de la métropole.

Le zonier inondation se doit d'être le meilleur reflet du risque supporté par l'entreprise au vu de la composition de son portefeuille d'assurés et du risque national. Il se pose la question de savoir si les observations sinistres utilisées seront issues du régime légal (que l'on appellera dans la suite du mémoire sinistres Cat-Nat), ou de la garantie Inondation (sinistres hors Cat-Nat), ou bien des deux. Le risque inondation étant un risque climatique, la survenance d'un sinistre est rare. Ainsi plus nous possédons de sinistres, plus nous arriverons à connaître l'impact des inondations en

France pour ensuite créer un zonier segmentant au mieux les zones à risques. Les graphiques I.1 et I.2 représentent la fréquence inondation des deux types de sinistres en indice base 100, où la référence est la fréquence moyenne sur le portefeuille en inondation (Cat-Nat et hors Cat-Nat). La géographie des deux garanties "inondation" est différente. Pour exemple, le département du Loiret (45), au sud de la région parisienne, possède une forte fréquence Cat-Nat alors que la fréquence hors Cat-Nat n'est pas aussi importante. Cette disparité géographique est logique car si l'on est couvert par la garantie Cat-Nat sur une commune, la garantie Inondation ne sera pas employée. En prenant donc en compte les deux sinistralités pour notre étude, l'ensemble du risque géographique sera évalué car l'historique de sinistre considère l'impact total de l'événement climatique.

En prenant les deux sinistralités, on peut remarquer que certaines zones ont tendance à recenser plus de sinistres que d'autres. On peut citer les codes postaux le long du Rhône, le sud, le nord-est de la France et le bassin parisien.

Pour appréhender au mieux les deux types de sinistralité et comprendre les sinistres manipulés, le coût d'un sinistre pour les deux indemnisations a été analysé. Le graphique I.3 montre que le coût d'un sinistre Cat-Nat est plus élevé que dans le cadre de la garantie Inondation. Cette observation est d'autant plus vraie pour les sinistres chers. En effet, les valeurs des derniers quantiles Cat-Nat explosent pour les grands sinistres. La variabilité de la sinistralité inondation globale du portefeuille (Cat-Nat et hors Cat-Nat) depuis 5 ans est importante car les coûts Cat-Nat sont élevés. Pour éviter une trop forte volatilité des coûts, les sinistres extrêmes seront écartés de la base de données. Cette analyse est développée dans la partie 3 présente à la suite du chapitre.

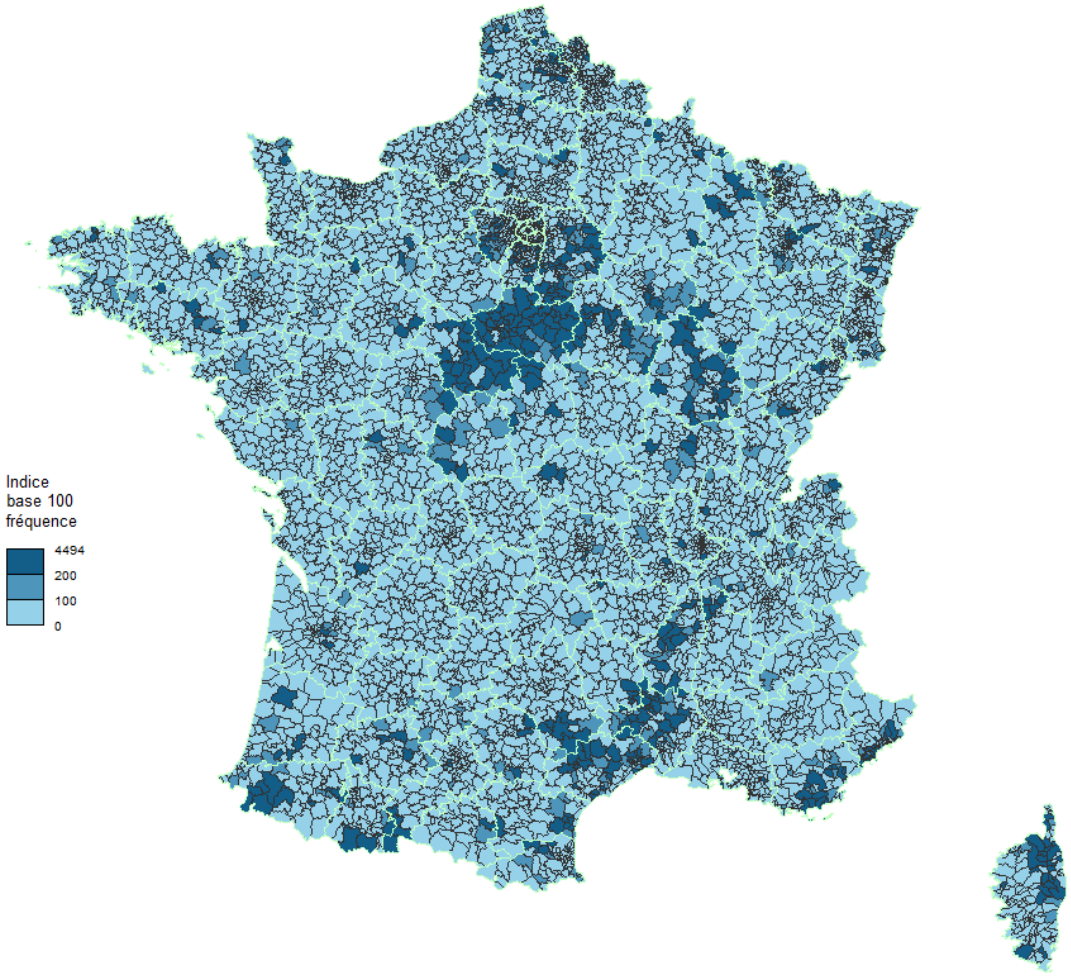


FIGURE I.1 – Indice de fréquence inondation Cat-Nat par codes postaux, de 2012 à 2016

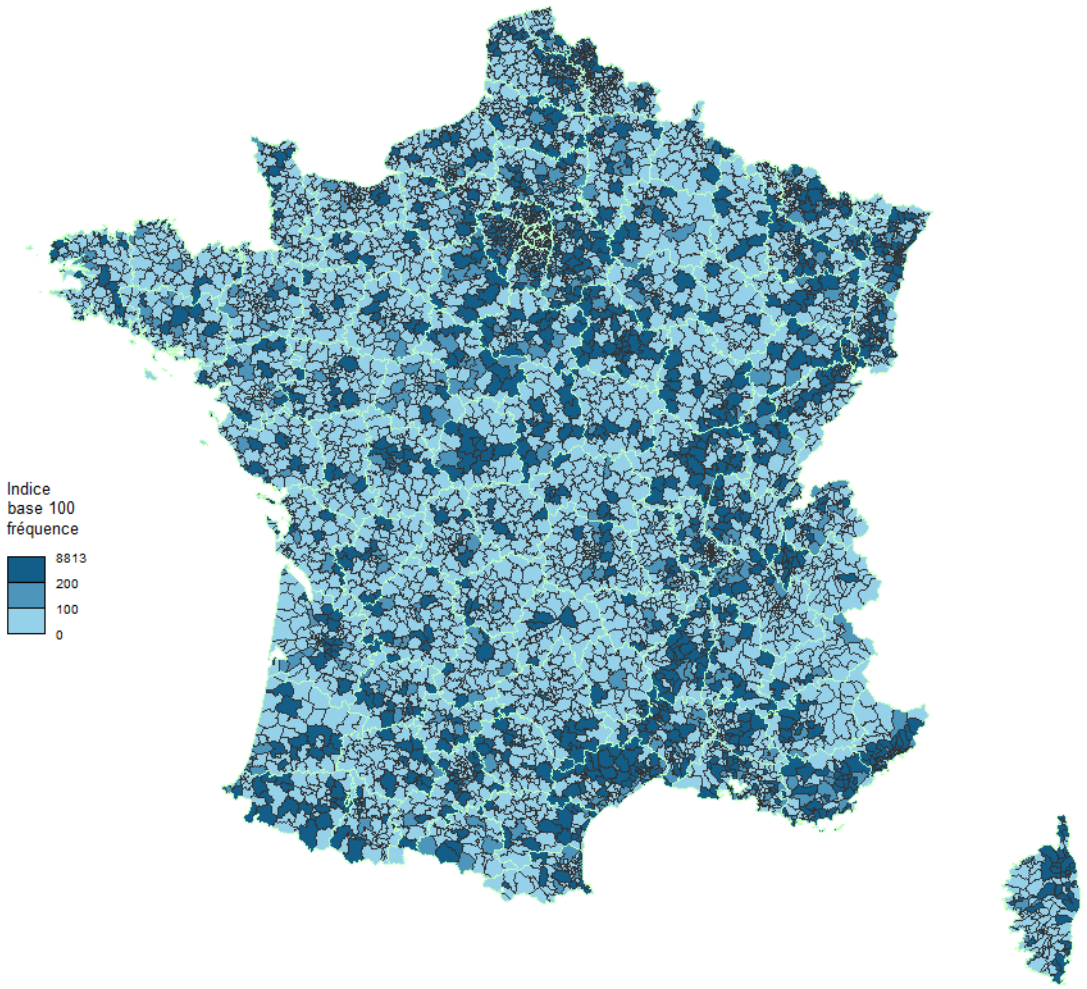


FIGURE I.2 – Indice de fréquence inondation hors Cat-Nat par codes postaux, de 2012 à 2016

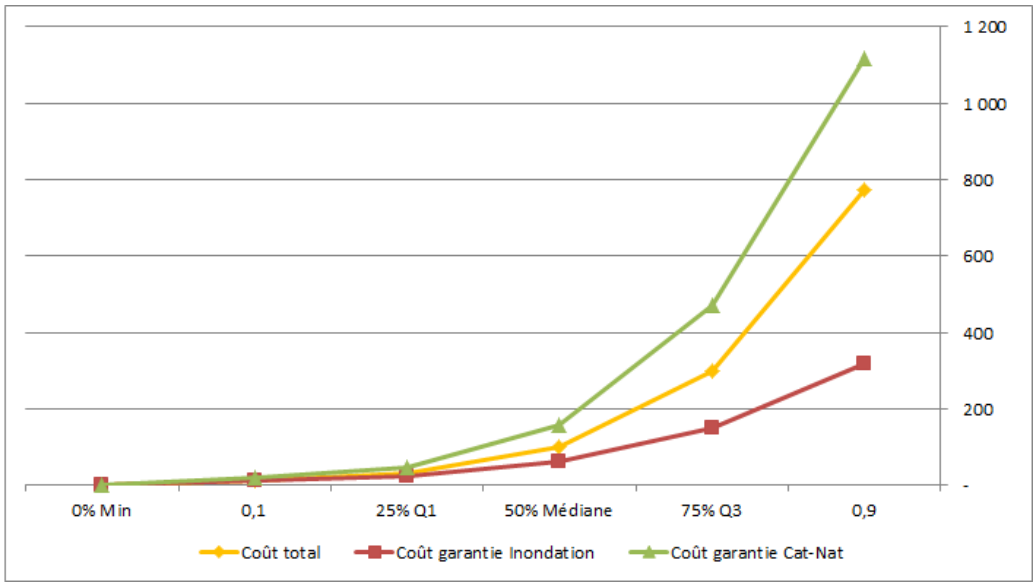


FIGURE I.3 – Comparaison des indices du coût d'un sinistre inondation en Cat-Nat, hors Cat-Nat et au global, entre 2012 et 2016

Pour estimer l'ensemble de l'effet géographique du risque inondation, nous conservons la sinistralité Cat-Nat et hors Cat-Nat. Cependant, il faut prêter attention au fait que le coût d'un événement Cat-Nat est en moyenne plus élevé que celui relevant de la garantie Inondation. De plus, l'inondation étant un risque à événements rares, il est important de pouvoir collecter le maximum de sinistres possibles pour estimer au mieux les zones à risque. L'historique le plus profond reconstitué est de 10 ans, de 2006 à 2016.

La présentation des variables et les traitements préliminaires effectués

L'entreprise a créé un nouveau produit d'assurance habitation en 2012. Les variables tarifaires entre le nouveau et l'ancien produit ne sont pas les mêmes, certaines ne sont collectées que depuis 2012, d'autres portent la même appellation mais leur définition est différente à cause des changements de garanties. Prenons un exemple illustratif avec le nombre de pièces. Pour l'un des produits, les dépendances habitables sont comptabilisées directement dans le nombre de pièces de l'habitation alors que pour l'autre, il existe deux variables, le nombre de pièces et le nombre de dépendances habitables. L'historique sélectionné s'étendant de 2006 à 2016, il prend en compte des risques habitations couverts par les deux produits. Les données disponibles ne sont donc pas uniformes pour chaque risque habitation. On va résoudre ce problème en modifiant certaines variables dans le respect des définitions et des

conditions d'applications des garanties de chacun des deux produits. Le point de départ de ce processus est de retenir comme variables toutes celles du nouveau produit, puis d'uniformiser les risques habitation couverts par l'ancien produit en récréant les garanties qu'il possède mais dans la définition juridique du nouveau produit. Voici les principales règles :

1. discrétisation des variables de capitaux et du nombre de pièces ;
2. introduction de la modalité "Indéterminé" pour toute information d'un risque que nous n'avons pu reconstituer ;
3. introduction de la modalité "Non concernée" pour tout risque ne pouvant prétendre à une caractéristique suite à la définition du produit MRH. Par exemple, un appartement ne peut pas avoir de piscine puisque que cette caractéristique n'est pas couverte par les contrats classiques MRH.

En définitif, la base de données comptabilise 25 000 000 de risques habitation sur 10 ans décrits par 28 variables internes dont :

- des variables relatives au contrat d'assurance (la formule de garantie, les garanties optionnelles...) ;
- des variables assurés (la catégorie socio-professionnelle...) ;
- des variables relatives au risque (le nombre de pièces, le type d'habitation...) ;
- une variable géographique : code postal de l'habitation ;
- le temps de présence d'un risque habitation en portefeuille sur une année (aussi appelé exposition au risque).

Cependant, pour l'étude de l'inondation, il aurait été intéressant d'avoir l'information sur l'étage de l'habitation, une information manquante dans notre base de données. A ces 28 variables caractéristiques du risque, on possède, pour chaque habitation, le nombre de sinistres sur l'année observée, le coût de chaque sinistre, la fréquence, la charge totale annuelle, la charge totale annuelle sans les sinistres extrêmes et la prime pure.

En plus de cette étape d'uniformisation des variables, certaines modalités ont été regroupées car elles ne possédaient pas un effectif suffisant pour une bonne application des modèles linéaires généralisés et l'étude des corrélations. L'analyse des corrélations par la statistique du khi-deux préconise que chaque modalité des variables qualitatives doit représenter au minimum 5% de l'effectif total. Le choix des regroupements a été effectué en étudiant la prime pure pour chaque modalité. Les modalités regroupées sont celles avec la prime pure la plus proche. Par cette démarche nous supposons que deux modalités avec des résultats de sinistralité similaires ont un même profil de risque.

L'étude des corrélations

L'analyse de la dépendance entre variables est importante afin d'éviter la redondance d'informations dans un modèle statistique. Dans notre base de données, après discrétisation, toutes les variables descriptives du risque, à l'exception de l'exposition au risque, sont qualitatives. La corrélation entre ce type de variable est réalisée à l'aide d'une table de contingence et une statistique du khi-deux. Le test d'indépendance du khi-deux est défini par la statistique D^2 :

$$D^2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{\left(n_{ij} - \frac{n_{i.}n_{j.}}{n}\right)^2}{\frac{n_{i.}n_{j.}}{n}}$$

pour deux variables qualitatives X et Y, r et c sont respectivement le nombre de modalités. n_{ij} est le nombre d'observations que possède le couple de modalités (r_i, s_j). $n_{i.}$ est le nombre d'individus ayant la modalité r_i de Y quelque soit la modalité de X. Sous H_0 : X et Y sont indépendants, D^2 suit une loi du χ^2 à $(r-1)(c-1)$ degrés de liberté. Ce test indique seulement la présence de dépendance entre variables mais ne permet pas d'ordonner les différentes corrélations 2 à 2. Pour cette raison, le V de Cramer a été choisi. Il se base sur le test d'indépendance du khi-deux. Son deuxième avantage est qu'il ne dépend pas de la taille de l'échantillon. Il se calcule par :

$$V_{cramer} = \sqrt{\frac{D^2}{n \times (min(r,c) - 1)}}$$

avec n le nombre total d'observations.

Les valeurs du V de Cramer sont comprises entre 0 et 1, où 1 indique une dépendance totale et 0 une absence de corrélation. Il reste à déterminer le seuil à partir duquel nous considérerons une corrélation entre deux variables. Nous fixons une corrélation modérée entre 0,3 et 0,5. Au dessus de 0,5 la corrélation est jugée forte. Le résultat des corrélations entre les variables se trouve en Annexes A.2. De nombreuses variables sont corrélées, ce qui limite les combinaisons possibles dans un modèle statistique. Lors des modélisations, il sera intéressant de tester les interactions entre les variables afin d'apporter plus d'informations, et de limiter les problèmes de multicolinéarité.

3 Une sinistralité non classique

L'inondation est un risque d'assurance à part en MRH. Tout d'abord, l'inondation est un risque climatique. Sa sinistralité a une structure spécifique. Par comparaison sur les années 2014 à 2016, la fréquence inondation (Cat-Nat et hors Cat-Nat) est 456 fois inférieure à la fréquence Vol qui est un sinistre classique en MRH, alors que son coût est 3,74 fois supérieur. Le risque inondation est un risque à événements rares. Le nombre de sinistres survenus est d'environ 17 000 pour 25 000 000 de risques habitation sur 10 ans, ce qui est très faible pour un risque d'assurance.

Comme précisé dans la partie 2 de ce chapitre, le coût d'un sinistre inondation est très volatile du fait de la sinistralité Cat-Nat. Or, le zonier construit doit être un outil tarifaire pérenne dans le temps. Une trop forte volatilité au sein des observations sinistres peut provoquer une volatilité du zonier. La volatilité des observations est provoquée par les sinistres extrêmes (aussi appelés sinistres graves). En assurance, on distingue les sinistres attritionnels et les sinistres extrêmes par leurs caractéristiques. Les sinistres extrêmes ont une plus faible fréquence mais un coût plus élevé que les sinistres attritionnels. Ces coûts élevés entraînent la volatilité des observations. Il est donc nécessaire de définir les sinistres extrêmes en fixant un seuil pour le montant maximal d'un sinistre attritionnel. La théorie des valeurs extrêmes (TVE) offre des outils permettant de déterminer ce seuil. La fonction moyenne des excès a été sélectionnée parmi ces techniques. C'est une méthode graphique de détermination du seuil. La séparation des sinistres graves n'étant pas le cœur du sujet, l'outil et le contexte théorique seront abordés de manière non exhaustive pour la mise en application. Pour cela, on doit postuler les sinistres graves comme des dépassements de seuils. On s'intéresse en conséquence aux observations $(x_i - u)_+$ où u est le seuil de dépassement. Cette distribution, selon la TVE, est caractérisée par la distribution de Pareto généralisée $GPD(\beta, \xi)$ qui s'exprime par sa fonction de répartition $G_{\xi,\beta}$:

$$G_{\xi,\beta}(x) = \begin{cases} 1 - \left[1 + \xi(\frac{x}{\beta})\right]_+^{-\frac{1}{\xi}} & \text{si } \xi \neq 0 \\ 1 - \exp^{-\frac{x}{\beta}} & \text{si } \xi = 0 \end{cases}$$

où

$$\begin{array}{rl} x \geq 0 & \text{si } \xi \geq 0 \\ 0 \leq x \leq -\frac{\beta}{\xi} & \text{si } \xi < 0 \end{array}$$

L'objectif est donc de déterminer le seuil u à partir duquel la distribution est une GPD. La fonction moyenne des excès permet de le fixer. On l'exprime, pour chaque

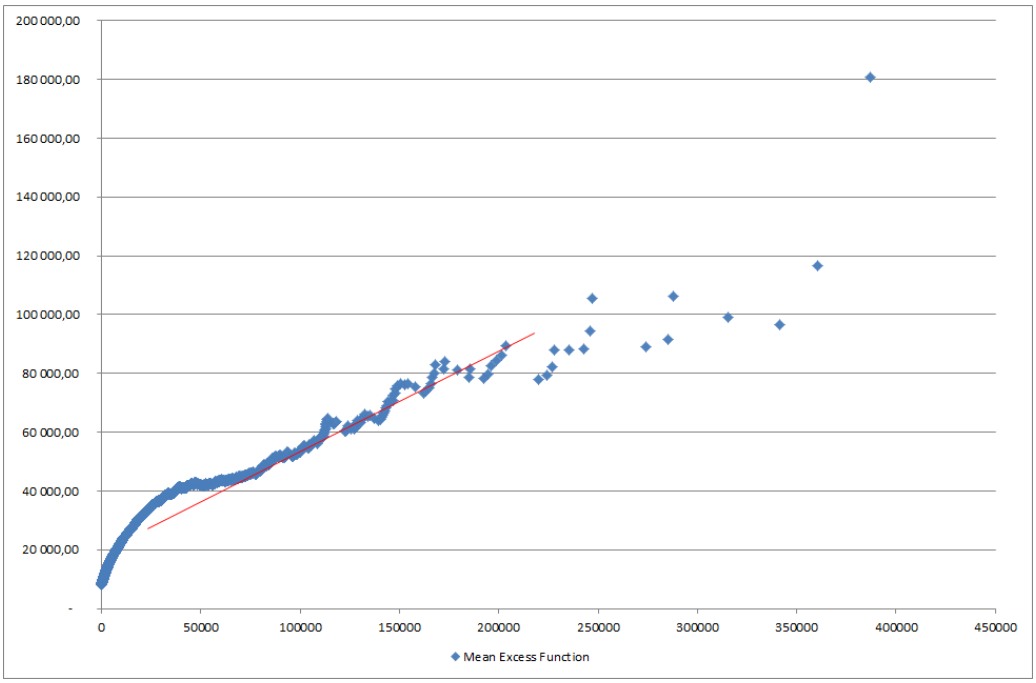


FIGURE I.4 – Fonction moyenne des excès pour le coût d'un sinistre inondation (Cat-Nat et hors Cat-Nat)

seuil u et le vecteur des coût observés $x = (x_1, .., x_n)$, par :

$$e_n(u) = \frac{\sum_{i=1}^{n}(x_i - u)_+}{\sum_{i=1}^{n} \mathbb{1}_{[x_i > u]}}$$

On représente ensuite le graphique $\{u, e_n(u)\}$. En TVE, si l'on suppose que les dépassements de seuils suivent une GPD, on peut prouver que la fonction moyenne des excès est linéaire à partir du seuil u de séparation entre sinistres attritionnels et sinistres graves.

Sur le graphique I.4, on observe une tendance linéaire pour un seuil à 70 000 (droite rouge). Le seuil $u = 70000$ retire 26,5% de la charge totale, pour 263 sinistres.

Au total de 2006 à 2016, on recense 16 828 sinistres inondation (Cat-Nat et hors Cat-Nat). Il y a 13 689 sinistres avec un coût non nul et inférieur au seuil de 70 000. Les huit sinistres concernant des recours où l'assureur a gagné de l'argent ont été retirés de la base. De manière générale pour l'évaluation d'une prime pure, les sinistres extrêmes font soit l'objet d'une étude approfondie, soit leur charge est répartie sur l'ensemble des assurés qui supportent une partie de ce risque. En tarification, il

existe plusieurs utilisations de ces sinistres extrêmes. Ils font soit l'objet d'une étude spécifique (fréquence et coût moyen), soit la charge supérieure au seuil est mutualisée sur chaque sinistre pour l'analyse du coût moyen. Ici par simplification, nous décidons de retirer les sinistres extrêmes de l'analyse du coût moyen mais de les conserver pour l'analyse de la fréquence. La faible survenance de sinistre préconise de conserver le maximum de sinistres possible. Pour rappel, notre approche statistique n'anticipe pas un sinistre non survenu dans notre historique.

4 Le type de zonier

Le premier point à aborder est l'intérêt de construire un zonier pour cette garantie. La survenance d'un sinistre inondation est particulièrement influencée par la géographie et les caractéristiques de l'environnement, telles que la présence d'un cours d'eau ou de fortes pluies. Il est facile de mettre en évidence les disparités géographiques. Sur la carte I.5, le nombre d'arrêtés CAT-NAT est représenté pour chaque code postal depuis 1982. Précédemment, il a été mis en évidence que les arrêtés Cat-Nat inondation ne couvrent pas l'ensemble de la sinistralité. Il n'en reste pas moins un bon indicateur des disparités géographiques en France métropolitaine. Les arrêtés Cat-Nat sont recensés dans la base GASPAR que l'on trouve en accès libre sur le site "georisques.gouv.fr", un site officiel gouvernemental. Ils sont publiés par code Insee (code de référence géographique). Pour information, il existe des arrêtés Cat-Nat pour d'autres risques climatiques que l'inondation comme la tempête. Cette carte met en évidence de manière déterministe les régions les plus touchées par des inondations depuis 1982. Le Nord-Est, le Sud et Sud-Ouest recensent la majorité des zones fortement sinistrées. Depuis plus de 30 ans, il existe bien des disparités géographiques quant à la survenance d'inondation sur le territoire d'où l'intérêt de segmenter géographiquement dans un tarif pour un assureur.

Le second point d'intérêt est la maille du zonier. Il est important de se demander laquelle est la plus adaptée. Les plus utilisées sont le code postal, le code Insee ou le code Iris qui est une décomposition en quartier des codes Insee. La maille détermine les zones pour lesquelles l'effet géographique sera estimé. Selon des spécialistes rencontrés, l'inondation est connue pour être un risque "ponctuel". Il est possible qu'un quartier d'une commune soit en zone inondable et pas son voisin. Il suffit d'être proche d'un cours d'eau et que le terrain facilite la propagation d'une crue, alors que le quartier voisin est légèrement plus en altitude et ne sera donc jamais atteint par une montée des eaux. De plus, la zone inondable ne suit évidemment pas le contour géographique. Ainsi, le zonier le plus juste serait un zonier à petite échelle,

à l'adresse ou au code Iris. Cependant, nous allons mettre en place une approche par statistiques inférentielles, qui comme expliqué en Introduction n'est pas totalement exacte pour appréhender le risque, puisque notre historique n'est pas assez important pour capter l'ensemble des événements inondations possibles en France métropolitaine. Créer un zonier sur une segmentation fine nécessite des informations les plus précises et fiables possibles. Cette information étant incomplète avec seulement 10 ans d'historique, la maille choisie est le code postal. Elle est plus grossière que le code Insee. On compte environ 6 000 codes postaux en France métropolitaine contre 36 000 codes Insee. Opter pour une maille plus grande privilégie une mutualisation des risques plus importante. Dans notre cas, cette maille semble appropriée car une segmentation trop fine serait risquée par le manque d'observations sinistres.

La dernière chose à définir est le type de zonier. Il est commun de choisir le type en fonction de l'indicateur le plus pérenne et le moins volatile dans le temps. Or, l'inondation étant un risque à événements rares, les trois indicateurs que sont la prime pure, la fréquence et le coût moyen d'un sinistre sont tous les trois volatiles. Les disparités géographiques sont présentes aussi bien pour la fréquence, en se référant au graphique I.1, I.2 et pour le coût, soit en conséquence pour la prime pure. La carte des disparités du coût moyen d'un sinistre par codes postaux se trouve en annexe A.1. Il existe un effet géographique quant à la survenance des sinistres mais également un effet pour le coût d'un sinistre. Au regard des deux cartes, les disparités géographiques ne sont pas présentes au même endroit. Les deux effets géographiques ne sont donc pas similaires. En effet, la survenance d'un sinistre dépend plus du caractère météorologique alors que le coût sera plus influencé par l'urbanisation de la zone. Si l'événement climatique se produit sur une ville plutôt que sur des terres non exploitées, le coût sera plus élevé. Il est donc intéressant de connaître ces deux effets. Dans le cadre du mémoire, nous supposons que l'effet géographique issu de la prime pure résume ces deux impacts géographiques.

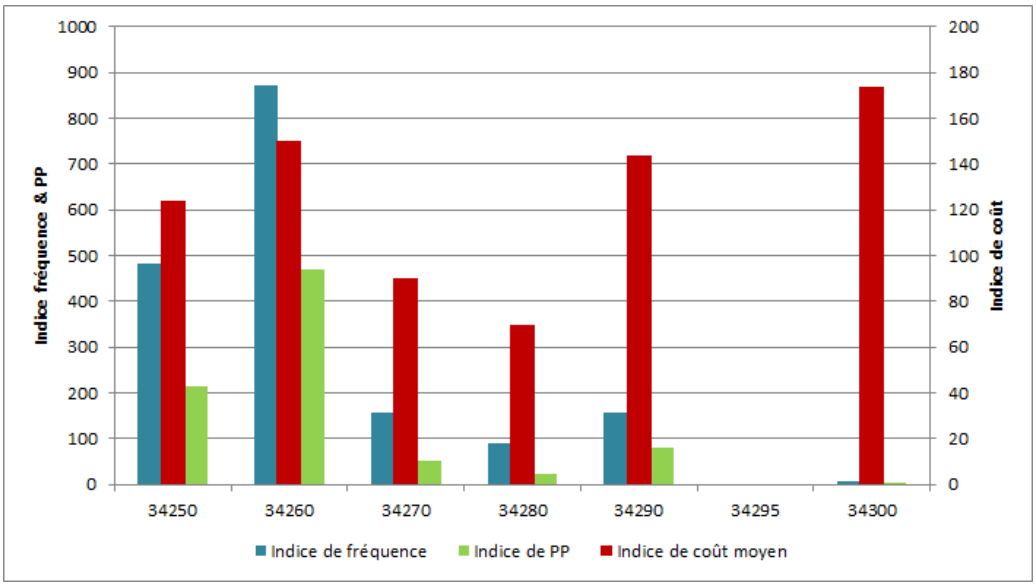


FIGURE I.6 – Indicateurs de sinistralité sur différents codes postaux de l'Hérault, de 2012 à 2016

Le graphique I.6 représente les trois indicateurs selon quelques codes postaux du département 34. Les disparités géographiques sont présentes aussi bien pour la fréquence, le coût moyen et la prime pure. Ceci confirme que la prime pure est discriminante au niveau géographique et donc confirme notre choix de type de zonier.

Le zonier inondation sera un zonier prime pure au code postal. La sinistralité utilisée est Cat-Nat et hors Cat-Nat. Le périmètre de sinistralité choisi permet d'obtenir une information globale sur la sinistralité de l'événement climatique en France. Il n'en reste pas moins que celle-ci est restreinte vis-à-vis de la structure du risque. Certains types d'événement ont une période de retour de plus de 40 ans. Par exemple, on parle d'une crue de la Seine tous les cent ans environ. Il va donc falloir utiliser au mieux ces informations sinistres pour modéliser l'effet géographique.

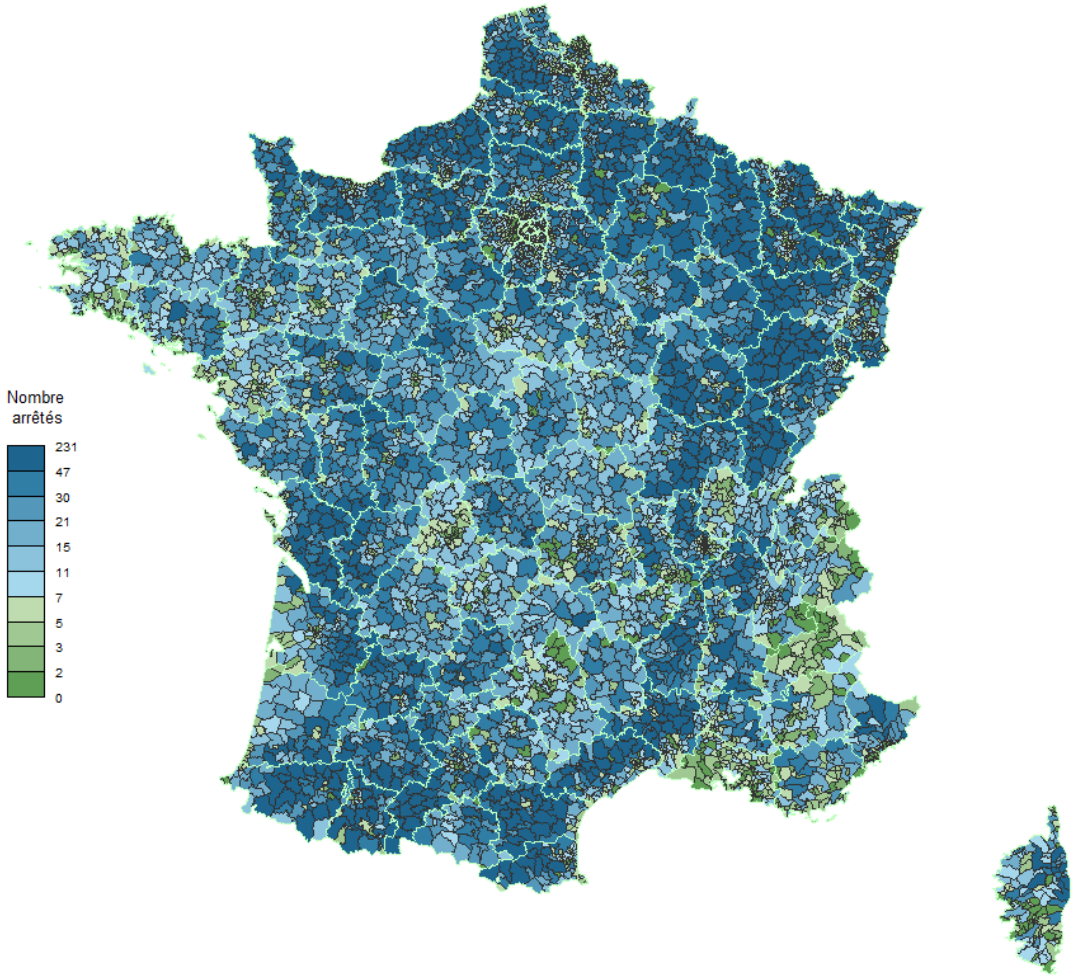


FIGURE I.5 – Disparités géographiques des arrêtés CAT-NAT inondation depuis 1982 pour chaque code postal

Chapitre II

La modélisation de l'effet géographique

La première étape dans la construction d'un zonier est de modéliser l'effet géographique pour chaque zone. Ici les zones choisies sont les codes postaux. Pour ce faire, la méthode classique est d'isoler l'effet géographique dans les résidus d'un modèle linéaire généralisé de prime pure sans les variables géographiques. Ces résidus (que l'on appellera résidus spatiaux) sont calculés pour chaque risque habitation sur 10 ans puis agrégés par code postal.

Cependant, les informations sinistres étant peu nombreuses, la modélisation de la prime pure classique fréquence × coût moyen ne convient pas. Le nombre de risques habitation non sinistrés est trop important pour une bonne estimation par processus de comptage. Ainsi, nous procédons à une approche alternative qui permet d'utiliser des méthodes spécifiques à la survenance d'événements rares.

1 Formalisation de l'approche

En tarification IARD, la prime pure pour un assuré est la base de la cotisation de son contrat. La prime pure est la charge annuelle moyenne pour un assuré. Elle peut s'écrire comme :

$$\textit{Prime Pure} = \textit{Fréquence} \times \textit{Coût Moyen}$$

En assurance, la prime pure est estimée par profil de risque homogène (ensemble d'assurés ayant les mêmes caractéristiques risques). La prime pure correspond à l'estimation de la charge supportée par l'assureur pour un profil de risque. On l'exprime par :

$$\Pi = E(Z|X)$$

où X est le vecteur de variables représentant le profil de risque. Z est la charge totale calculée à l'aide du nombre de sinistres N et du coût de chaque sinistre C_i :

$$Z = \begin{cases} \sum_{i=1}^{N} C_i & \text{si } N > 0 \\ 0 & \text{si } N = 0 \end{cases}$$

Sous hypothèse d'indépendance entre N et C, on trouve la formule classique d'estimation de la prime :

$$\Pi = \mathbb{E}(N|X)\ \mathbb{E}(C|X') \tag{II.1}$$

où X est le vecteur des variables explicatives retenues pour la fréquence et X' pour le coût moyen, $\mathbb{E}(N|X)$ la fréquence et $\mathbb{E}(C|X)$ le coût moyen d'un sinistre.

Cette approche classique de la fréquence, généralement par une loi de Poisson, a été testée mais ce modèle n'est pas ajusté, car il ne respecte pas le critère de la déviance évoqué dans la partie suivante des modèles linéaires généralisés 2. Ainsi, nous optons pour une modélisation moins classique de la prime pure. Pour cela, on utilise la formule des espérances totales, où pour une partition de l'univers $(B_i)_{i=1..n}$, on a :

$$\mathbb{E}(X) = \sum_{i=1}^{n} \mathbb{P}(B_i) E(X|B_i) \tag{II.2}$$

Ainsi, dans notre cas on cherche à connaître $E(Z|X)$. On pose l'événement $I \in \{0, 1\}$ représentant le fait d'avoir été sinistré dans l'année ou non, qui est une partition de l'univers. Par application de l'équation (II.2), on exprime la prime pure inondation par :

$$\Pi = \mathbb{E}(Z|X, I = 1)\mathbb{P}\left[I = 1\right] + \mathbb{E}(Z|X, I = 0)\mathbb{P}\left[I = 0\right] \tag{II.3}$$

Le deuxième terme de cette équation vaut 0 car si aucun sinistre n'est survenu $[I = 0]$, la charge annuelle Z vaut 0. Cette modélisation de la prime pure est intéressante car il existe des méthodes spécifiques quant à la modélisation d'événements rares par régression logistique. La survenance de sinistres est extrêmement faible. En effet, on possède seulement 0.07 % de risques habitation sinistrés sur une année, dont 16 427 risques habitations ayant un seul sinistre, 193 risques ayant 2 sinistres et 5 avec 3 sinistres. Ainsi, il existe peu de situations où un risque habitation a plus d'un sinistre sur une année, ce qui conforte l'idée de la modélisation de la prime pure.

Classiquement en tarification, la modélisation de la prime pure d'un risque d'assurance dépend de plusieurs variables explicatives :

- variable risque : décrivant le risque, ici l'habitation ;
- variable contrat : relatif au profil du souscripteur ;
- variable géographique : la localisation du risque ou des variables socio-économiques.

La prime pure inondation Π^I peut se décomposer comme :

$$\Pi^I = \textit{Effet non géographique} + \textit{Effet géographique} + \textit{Résidus}$$

Le zonier représente la segmentation géographique du risque d'assurance étudié. Pour estimer cet effet, on peut isoler l'effet géographique par une différence :

$$\Pi^I - \textit{Effet non géographique} = \textit{Effet géographique} + \textit{Résidus}$$

Boskov et Verrall [3] font partie des premiers auteurs à envisager la modélisation de l'effet géographique ainsi. Pour récupérer l'effet géographique du risque inondation, il faut donc connaître l'effet non géographique et la prime pure Π^I. Cette dernière est déterministe, elle est égale à la prime pure inondation observée dans notre historique. Puis il faut extraire l'effet non géographique. Dans l'article de Brouhns, Denuit et Masuy [4], la méthode de Boskov et Verall est mise en application dans le cas de l'assurance automobile. Il y est proposé de modéliser la partie non géographique par un modèle linéaire généralisé sans les variables géographiques. L'hypothèse est qu'en retirant les variables à tendance géographique, nous ne modéliserons que la partie non géographique de la prime pure. Cette approche est viable si le modèle linéaire généralisé sans ces variables est correctement calibré et n'omet aucune variable explicative potentielle hormis les variables géographiques.

Par cette méthode, nous obtenons donc l'effet géographique individuel (ou résidu spatial) de chaque risque habitation de l'historique considéré. On note cet impact géographique par R_j où $j \in \{1, n\}$ avec n le nombre de risques habitation, soit ici environ 25 000 000 pour l'étude. Formellement, on définit les résidus spatiaux R_j comme :

$$R_j = \Pi_j^I - \hat{\Pi}_j^I \tag{II.4}$$

où Π_j^I est la prime pure inondation observée pour un risque habitation et $\hat{\Pi}_j^I$ la prime pure estimée sans les variables géographiques. Si le résidu spatial est positif, ceci implique que la géographie de ce risque habitation favorise la sinistralité de l'assuré. Alors qu'un risque spatial négatif signifie que sa zone géographique diminue la sinistralité, ainsi l'habitation est peu exposée au risque inondation.

Or, pour le zonier inondation, nous cherchons à connaître l'effet géographique de chaque code postal. On agrège donc les effets géographiques de chaque risque

habitation pour un même code postal C_k.

$$C_k = \frac{\sum_{j=1}^{l} e_j R_j}{\sum_{j=1}^{l} e_j} \tag{II.5}$$

où l est le nombre de risque habitation dans le code postal k et e_j le temps d'exposition au risque sur une année, i.e le temps de présence du risque habitation en portefeuille sur une année.

L'effet géographique du risque inondation est donc contenu dans les résidus d'un modèle linéaire généralisé de prime pure sans intégrer les variables géographiques. La prime pure sera modélisée à l'aide d'une régression logistique pour la survenance de sinistres dans l'année. Puis la charge annuelle sera modélisée par un GLM plus classique en tarification. L'important dans ces modélisations est d'utiliser au mieux l'information disponible. Ainsi, la régression logistique sera utilisée par *response-based-sampling*, une méthode spécifique à la survenance d'événements rares. Quant à la modélisation de la charge annuelle, la validation croisée permettra d'utiliser l'ensemble des observations disponibles pour l'apprentissage du modèle tout en évaluant sa qualité prédictive.

2 Les modèles linéaires généralisés

Les modèles linéaires généralisés (GLM) sont mis en place pour la modélisation de la survenance de sinistres et de la charge annuelle. Les GLM sont très utilisés en tarification d'assurance non vie. Ils permettent de modéliser la prime pure d'une garantie et d'en tirer des conclusions sur l'influence des critères tarifaires.

Présentation

Trois composantes caractérisent un GLM :
- Y la variable réponse i.e que l'on cherche à modéliser
- $X_1, .., X_p$ p covariables ou variables explicatives
- La fonction de lien g

Il est à noter que les GLM ne fonctionnent que sous certaines hypothèses. Il faut une indépendance entre les individus 2 à 2 et entre les covariables $X_j, j \in \{1, p\}$ pour éviter les problèmes de multicolinéarité. Ce dernier est évité en réalisant au préalable une étude des corrélations entre les variables 2 à 2. Cette analyse a été

réalisée au sein des traitements préliminaires de la base de données au chapitre 1 partie 2. Une condition fondamentale est l'appartenance de la variable réponse Y à la famille exponentielle. Une variable Y appartient à cette famille si sa densité peut s'écrire sous la forme d'une exponentielle à deux paramètres θ et Φ tel que :

$$f(y, \theta, \Phi) = \exp\left(\frac{y\theta - b(\theta)}{a(\Phi)} + c(y, \Phi)\right), y \in S$$

avec

- S support de la loi de Y.
- $\theta \in \mathbb{R}$ paramètre de la moyenne
- $\Phi \in \mathbb{R}$ paramètre de dispersion
- a : fonction définie sur $\mathbb{R}$ et non nulle
- b : fonction définie sur $\mathbb{R}$ et deux fois dérivable
- c : fonction définie sur $\mathbb{R}^2$

A noter que si Y appartient à la famille exponentielle, on peut montrer que $\mathbb{E}(Y) = b'(\theta)$ et $Var(Y) = b''(\theta)\ a(\Phi)$.

Si ces hypothèses sont respectées, on peut exprimer les GLM sous forme matricielle par :

$$g(\mathbb{E}(Y|X)) = \beta X \tag{II.6}$$

La fonction de lien g est construite de manière à assurer la compatibilité entre l'espace de définition de la fonction g et les valeurs possibles de $\mathbb{E}(Y)$. On définit pour chaque loi une fonction de lien canonique noté g_c, qui est telle que :

$$g_c(\mathbb{E}(Y|X)) = \theta$$

Il faut noter que l'on n'utilise pas toujours la fonction de lien canonique. En assurance non vie, le logarithme est apprécié, car on obtient un modèle multiplicatif pour les coefficients β, ce qui rend plus facile leur interprétation vis-à-vis de la variable réponse Y.

Le tableau II.1 est un récapitulatif des lois classiques d'utilisation d'un modèle linéaire généralisé et leurs caractéristiques. A cette liste, on peut également rajouter la loi binomiale négative et l'inverse gaussienne

Estimation des paramètres

On recherche donc à estimer dans l'équation (II.6), le vecteur des coefficients $\beta = (\beta_1, .., \beta p)$ à partir du vecteur des covariables X et des observations de la variable

Loi	$g_c(x)$	θ	Φ	$a(\Phi)$	$b(\theta)$	$c(y,\Phi)$
$\mathcal{N}(\mu,\sigma^2)$	x	μ	σ^2	Φ	$\frac{\theta^2}{2}$	$-\frac{1}{2}\left(\frac{y^2}{\sigma^2}+log(2\pi\sigma^2)\right)$
$\mathcal{P}(\lambda)$	$ln(x)$	$log(\lambda)$	1	1	$exp(\theta)$	$-log(y!)$
$\Gamma(\mu,\gamma^{-1})$	$\frac{1}{x}$	$-\mu^{-1}$	γ^{-1}	Φ	$-log(-\theta)$	$\left(\frac{1}{\Phi}-1\right)log(y)-log\left(\Gamma(\frac{1}{\Phi})\right)$
$\mathcal{B}(n,p)$	$ln(\frac{x}{1-x})$	$log(\frac{p}{(1-p)}$	1	1	$nlog\left(1+exp(\theta)\right)$	$log\binom{n}{y}$

TABLE II.1 – Récapitulatif des différentes lois usuelles et de leurs paramètres pour un GLM

réponse $(y_1,..,y_n)$. Ces coefficients de régression sont estimés par la méthode du maximum de vraisemblance. La log-vraisemblance du GLM est définit comme :

$$l(y,\theta(\beta),\phi)=ln\left(\prod_{i=1}^{n}f(y_i,\theta_i,\phi_i)\right)$$
$$l(y,\theta(\beta),\phi)=\sum_{i=1}^{n}\frac{y_i\theta_i-b(\theta_i)}{a(\phi_i)}+c(y_i,\phi_i)$$

Ainsi l'estimateur du maximum de vraisemblance respecte les conditions d'optimisation suivantes :

$$\frac{\partial l(y,\theta(\beta),\phi)}{\partial\beta}=0$$
$$\frac{\partial^2 l(y,\theta(\beta),\phi)}{\partial\beta^2}<0$$

Il est usuellement approché par l'algorithme de Newton-Raphson dans les logiciels de calculs.

Validation du modèle

Comme expliqué dans les séances de cours de Xavier Milhaud [5], la validation d'un GLM doit respecter certaines conditions :

1. création d'un échantillon d'apprentissage et un échantillon de test, indépendants et créés aléatoirement ;
2. significativité globale du modèle à l'aide du critère de la déviance i.e la déviance observée $D=2\left(ln(L_{sat})-ln(L)\right)$ est inférieure au khi-deux à 95 % de degrés de libertés $n-p-1$, où n le nombre d'observations, p le nombre de variables explicatives, L_{sat} est la vraisemblance du modèle saturé (modèle possédant autant de paramètres que d'observations) ;

3. significativité locale, les coefficients de régression $\hat{\beta}$ doivent être significatifs au titre du test de Wald i.e $H_0 : \hat{\beta}_j = 0$ contre $H_1 : \hat{\beta}_j \neq 0$;
4. ajustement du modèle par l'analyse des résidus. Ils doivent être homoscédastiques et centrés autour de 0 ;
5. évaluation de la qualité prédictive en appliquant le modèle à la base de test.

En pratique le critère de significativité selon le test de Wald est difficile à respecter pour l'ensemble des coefficients de régression. En tarification, il est toujours préférable de regrouper à chaque étape de la modélisation deux modalités ayant un sens à être regroupées. Dans la majeur partie des cas, si une modalité n'est pas significative selon le test de Wald mais qu'il n'est pas logique selon le sens métier de la regrouper, alors cette modalité est laissée tel quelle dans le GLM. Il faut cependant prêter attention au fait qu'elle ne sera pas interprétable vis-à-vis de la modalité de référence.

De plus, un modèle est jugé de meilleure qualité si il dispose d'un nombre de variables assez restreint, ce que l'on appelle le critère de parcimonie. Ce principe est respecté en minimisant les critères tel que celui d'Akaike (AIC) ou le critère bayésien (BIC). Ils ne fonctionnent que pour des modèles emboîtés c'est-à-dire la comparaison de deux modèles dont l'un possède une variable en moins que l'autre mais toutes les variables explicatives sont identiques.

Pour rappel, on souhaite estimer la prime pure inondation à l'aide d'un modèle GLM en omettant les variables géographiques. Les résidus vont contenir l'effet géographique et la perturbation aléatoire. Ainsi, certaines conditions sont susceptibles de ne pas être significatives, ni interprétables. Les résidus ont la possibilité de ne pas être concentrés autour de 0, si l'effet géographique est une grande part de la survenance de sinistres ou de la charge annuelle. Si leur analyse n'est pas totalement concluante, il ne faut pas juger que le modèle est faux puisqu'il manque les variables géographiques. L'analyse nécessite donc une attention particulière. Ce raisonnement est également valable pour la qualité prédictive, qui peut être importante mais qui n'indique pas un mauvais ajustement de la partie non géographique.

Les GLM sont donc utilisés dans notre étude pour la modélisation de la prime pure par la décomposition survenance $\times$ charge annuelle sans variables géographiques pour en extraire l'effet géographique de l'inondation.

3 Modélisation de la survenance de sinistres inondation

La modélisation de la survenance de sinistres utilise une régression logistique qui nécessite d'être adaptée. La forte présence de 0 au sein des observations de la base empêche une bonne estimation de la probabilité de survenance de sinistres dans l'année. Les méthodes employées sont spécifiques aux événements rares.

3.1 Rappels sur la régression logistique

La survenance d'un sinistre inondation se résume par une variable binaire $I \in \{0, 1\}$ où 1 représente un risque habitation sinistré dans l'année sans détail du nombre de sinistres, et 0 un risque habitation non sinistré. En conséquence, on pose I comme :

$$I = \begin{cases} 0 & \text{si } N = 0 \\ 1 & \text{si } N > 0 \end{cases}$$

Par une loi de Bernouilli, on a $\mathbb{P}(I = 1) = 1 - \mathbb{P}(I = 1) = p$. Le modèle GLM qui permet l'estimation de cette distribution de probabilité est une régression logistique. Il existe trois fonctions de lien usuelles, pour ce modèle GLM :

- la fonction logit $x \to log(\frac{x}{1-x})$
- la fonction probit $x \to \Phi^{-1}(x)$, soit l'inverse de fonction normale
- la fonction log-log complémentaire $x \to log(-log(1-x))$

Elles possèdent quelques propriétés remarquables. La fonction complémentaire log-log n'est pas symétrique contrairement aux deux autres. De plus sur l'intervalle de définition $x \in [0, 1]$, pour les petites valeurs, la fonction log-log complémentaire est proche de la fonction logit. La fonction de lien la plus souvent utilisée est le logit car elle permet une interprétation plus facile des résultats à l'aide des odds-ratio ce que les autres ne permettent pas.

Par application de l'équation des GLM (II.6), on trouve la formule suivante :

$$g(p_i) = \beta X \tag{II.7}$$

où $p_i = \mathbb{P}(I = 1|X)$ représente la probabilité de survenance d'un sinistre inondation sachant X, le vecteur des variables explicatives. Les coefficients β sont estimés par

Observations \Prédictions	1	0
1	a = Vrais positifs	b = Faux négatifs
0	c = Faux positifs	d = Vrais négatifs

TABLE II.2 – Matrice de confusion d'une régression logistique

maximum de vraisemblance. La log-vraisemblance du modèle s'écrit comme :

$$l(\beta|i,X) = \sum_{j=1}^{n} i_j \; log(g^{-1}(X_j\beta)) + \sum_{i=j}^{n}(1-i_j) \; log(1-g^{-1}(X_j\beta))$$

Dans le cadre de l'étude, la régression logistique est utilisée pour estimer la probabilité $\mathbb{P}(I=1|X)$. Dans d'autres secteurs tels que le marketing, elle est souvent utilisée en tant que classifieur $\{0,1\}$. On cherche, dans ce cas, non pas à prédire $\hat{p}_i$ mais à connaître la prédiction 0 ou 1. Cette approche permet de comparer la qualité prédictive des différentes modélisations.

Puisque $\hat{p}_i$ est prédit, il faut réussir à déterminer un seuil s pour lequel :

$$\hat{I} = \begin{cases} 0 & \text{si } \hat{p}_i \leq s \\ 1 & \text{si } \hat{p}_i > s \end{cases}$$

Ce seuil est déterminé à l'aide de la matrice de confusion et de la courbe ROC (*Receiver Operating Characteristic*), respectivement présentées par le tableau II.2 et le graphique II.1. A l'aide de la matrice de confusion, on définit plusieurs indicateurs utiles à l'évaluation de l'ajustement et de la qualité prédictive d'une régression logistique.

- Le taux d'erreur = $\frac{b+c}{n}$, n l'effectif total
- La sensibilité ou taux de vrais positifs = $\frac{a}{a+b}$
- La spécificité ou taux de vrais négatifs = $\frac{d}{d+c}$
- Le taux de faux positifs = 1 - spécificité

A l'aide de ces indicateurs, on peut construire la courbe ROC qui est usuellement utilisée pour définir l'ajustement du modèle ou déterminer le seuil s. Elle représente le taux de vrais positifs en fonction du taux de faux positifs pour chaque seuil $s \in [0,1]$. Le point optimal de la courbe ROC est $(0,1)$ car cela indique qu'il n'y a aucun faux positifs et seulement des vrais positifs. Ainsi le seuil optimal est le point de la courbe le plus proche de (0,1), du point de vue de la distance euclidienne. Seulement, les

probabilités de survenance estimées seront très faibles. Ainsi, observer la courbe ROC n'a pas de sens car la probabilité maximale estimée sera inférieure à 0,01.

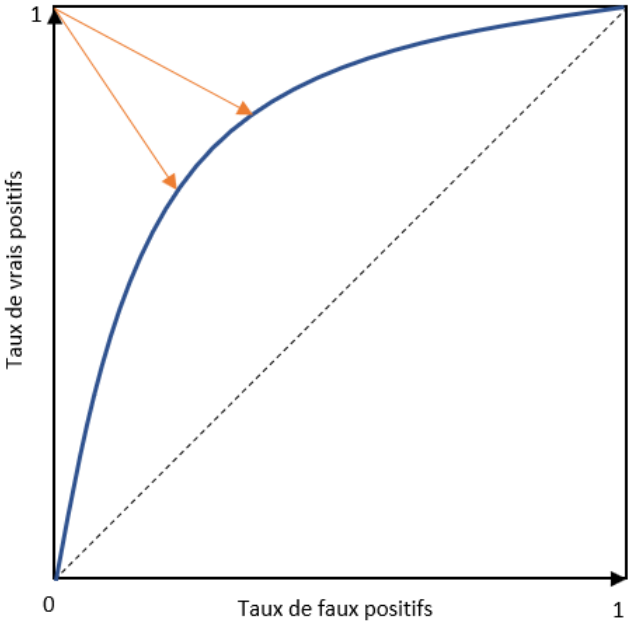


FIGURE II.1 – Exemple d'une courbe ROC (en bleu) pour une régression logistique

Pour déterminer un seuil s et pouvoir calculer les indicateurs de la qualité prédictive cités précédemment nous avons employé une autre technique. Pour chaque seuil entre 0 et 0.01, la sensibilité et 1-spécificité sont calculés comme pour la courbe ROC. Puis nous réalisons le graphique II.2 de la sensibilité et la spécificité en fonction des seuils s. Le croisement des deux courbes correspond à la valeur du seuil optimal, car il correspond à un compromis entre le taux de vrais positifs et le taux vrais négatifs. Ainsi, on minimise l'erreur de classification du modèle.

Une fois le seuil optimal déterminé pour un modèle, les indicateurs de la qualité prédictive sont estimés, puisque l'on peut désormais classer chaque risque habitation comme sinistré ou non dans l'année.

Cependant, la régression logistique ne peut pas s'appliquer telle quelle dans l'estimation d'une probabilité d'événements rares, c'est ce qui est expliqué dans la suite de cette partie.

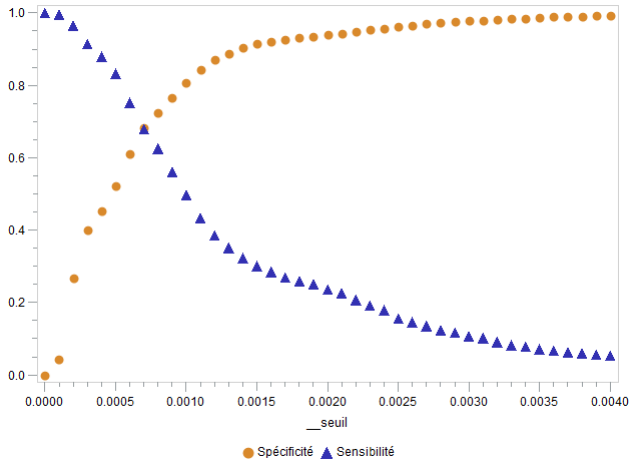


FIGURE II.2 – Détermination du seuil s en fonction de la spécificité et de la sensibilité

3.2 Les écueils de la régression dans le cadre d'événements rares

La variable réponse provient d'un risque à événements rares. Elle présente, en conséquence, un taux de réponse très faible, c'est-à-dire peu d'événements $[I = 1]$ sont observés dans la base de données. Pour le risque inondation, ce taux est de l'ordre de 0,07 %, soit environ 16 000 risques habitations sinistrés pour 25 000 000 risques observés. Cette structure est appelée dans la littérature par *unbalanced dataset*, que l'on traduira par base de données déséquilibrée. Cette configuration de base empêche l'obtention de bons résultats par une régression logistique. En théorie, nous chercherons à avoir une estimation au moins asymptotiquement sans biais et de variance la plus petite possible. Ces caractéristiques sont difficiles à obtenir pour des événements rares. Ces difficultés d'estimation sont détaillées dans le livre de MacCullagh et Nelder [6] et l'article de King et Zeng [7].

Dans un premier temps, le faible taux de réponse introduit un biais dans l'estimation des paramètres $\hat{\beta}$ de la régression logistique. On définit le biais d'un estimateur $\hat{\beta}$ par $E(\hat{\beta}) - \beta$. MacCullagh et Nelder ont démontré que le biais pour une régression logistique sur une base de données déséquilibrée s'exprime par :

$$biais(\hat{\beta}) = (X'WX)^{-1}X'W\xi$$

avec $\xi_i = 0.5Q_{ii}\left[(1 + w_i)\hat{p}_i - w_i\right]$, w_i est la pondération de l'observation i dans l'estimation par maximum de vraisemblance. Q_{ii} sont les éléments diagonaux de la matrice $Q = X(X'WX)^{-1}X'$ et $W = diag\{\hat{p}_i(1 - \hat{p}_i)w_i\}$.

King et Zeng ont démontré que le coefficient de régression β_0 est directement impacté dans le cadre d'une base de données déséquilibrée. Il a été prouvé que le biais $\hat{\beta}_0$ est toujours négatif, ce qui implique $\mathbb{E}(\hat{\beta}_0) < \beta_0$. Donc β_0 sera toujours sous-estimé. Or, le coefficient de régression $\hat{\beta}_0$ influence directement l'estimation des probabilités des événements rares $\hat{p_i}$, ainsi elle est aussi inexacte.

Le deuxième écueil démontré par King et Zeng s'explique à l'aide de la matrice de variance-covariance de $\hat{\beta}$. Dans le cadre d'une régression logistique, elle s'exprime comme :

$$Var(\hat{\beta}) = \frac{1}{\sum_{i=n}^{n} p_i(1-p_i)x_i'x_i} \tag{II.8}$$

La partie de cette équation (II.8) affectée par la présence d'événements rares est $p_i(1-p_i)$. Si le taux de réponse est faible alors l'estimation de la probabilité d'occurrence sera elle aussi faible, soit p_i proche de 0. Or, la fonction définie par $x \to x(1-x)$ pour $x \in [0,1]$ atteint son maximum en 0,5 et son minimum en 0 ou 1. Plus la valeur de la fonction $x(1-x)$ sera grande, plus la variance de l'estimateur va diminuer. En conséquence, l'estimation des paramètres de régression β s'améliorera. Ainsi, dans le cadre d'événements rares, une bonne estimation des paramètres semble compromise. On peut déduire de cette dernière observation que la proportion d'événements $[I = 1]$ est plus significative dans une base de données pour une bonne estimation que son événement contraire $[I = 0]$. Ainsi, augmenter le nombre d'événements réponses $[I = 1]$ permettrait d'améliorer la modélisation.

3.3 Résolution par *response-based-sampling*

En réponse à ces deux écueils, on trouve différentes méthodes applicables à la régression logistique. Nous privilégions ici la technique du *response-based-sampling* aussi nommé en économétrie *choice-based-sampling* ou encore *endogenous startified sampling*. La méthode du *response-based-sampling* est expliquée dans l'article de mise en application de King et Zeng [7].

Cette dernière est un type d'échantillonnage spécifique à une base de données déséquilibrée. En pratique, on distingue trois types d'échantillonnages de base de données. Si on décrit une base de données D composée de variables explicatives X et d'une variable réponse Y. Le premier échantillonnage est la sélection aléatoire de lignes, peu importe la configuration de (X, Y). Cette technique se nomme le *Random Sampling*. L'*Exogene Stratified Sampling* consiste à tirer aléatoirement des individus. Mais cette fois-ci, on garde la structure de départ de la base D pour les

variables X, c'est-à-dire que l'on conserve les proportions des sous-groupes. Enfin, le *response-based-sampling* est l'échantillonnage qui permet d'augmenter la proportion d'événements $[Y = 1]$ et de conserver la structure de chaque modalité des variables explicatives X. Cette démarche a été présentée dans la partie 3.2 précédente comme bénéfique pour la régression logistique d'événements rares puisque l'on augmente la proportion de $[Y = 1]$ dans la base. Cette technique est une réponse aux difficultés énoncées. Elle se déroule selon les étapes suivantes :

- sélection de toutes les lignes d'événements $[Y = 1]$;
- sélection aléatoire d'événements $[Y = 0]$ pour obtenir un pourcentage τ_e de réponse fixé au préalable, en conservant les proportions des variables X.

On obtient donc deux bases de données, celle d'origine ayant un taux de réponse faible que l'on note τ_o et la base échantillonnée avec le nouveau taux de réponse τ_e. Pour rappel, le taux de réponse est le pourcentage d'événements $[I = 1]$ dans la base de données. Dans la littérature, aucun chercheur n'a prouvé l'existence d'un taux optimal. Le choix est laissé libre à l'utilisateur du *response-based-sampling*. Selon les praticiens, un taux supérieur à 20% donne de bons résultats. Dans la table, notre taux de réponse de départ étant extrêmement faible, on a décidé de se fixer un taux échantillonné τ_e légèrement au dessus de 20%. Après échantillonnage, on obtient un taux de 24,81%.

Cette technique d'échantillonnage nous permet donc d'augmenter le nombre d'événements $[I = 1]$ et ainsi obtenir une estimation de la probabilité d'occurrence d'un sinistre inondation $\hat{p}_i$ plus précise. Mais ce changement de structure pour la variable réponse I a un impact logique sur la valeur des coefficients de régression. Comme expliqué précédemment, $\hat{\beta}_0$ est le seul impacté par cette modification. L'étape suivante de la méthode consiste à ajuster ce coefficient pour retrouver la valeur de la base d'origine et non plus celle de l'échantillon créé. Cette correction peut être réalisée de deux manières : soit la *weighting method*, soit la *prior correction*.

Weighting method

Manski et Lerman en 1977 [8] sont les deux auteurs à l'origine de cette correction. L'objectif est de compenser la différence des taux de réponse τ_o et τ_e. Ils préconisent d'utiliser la log-vraisemblance pondérée à la place de la log-vraisemblance pour l'estimation des paramètres de la régression logistique :

$$ln(L(\beta|y)) = w_1 \sum_{Y_i=1} ln(p_i) + w_0 \sum_{Y_i=0} ln(1 - p_i)$$

où les poids de pondération sont égaux à :

$$w_1 = \frac{\tau_o}{\tau_e}$$
$$w_0 = \frac{1-\tau_o}{1-\tau_e}$$

Prior correction

La *prior correction*, quant à elle, est effectuée à l'aide de l'information a priori que l'on a du taux de réponse dans chacune des deux bases. Cette correction s'effectue après avoir calculé l'estimateur du maximum de vraisemblance contrairement à la *weighting method*. En reprenant les notations introduites ci-dessus, la correction s'opère sur le coefficient $\hat{\beta}_0$ par :

$$\hat{\beta}_0 - ln\left[\left(\frac{1-\tau_o}{\tau_o}\right)\left(\frac{\tau_e}{1-\tau_e}\right)\right]$$

L'avantage de ces deux méthodes consiste en leur simplicité d'utilisation, soit par l'introduction de poids, soit par l'introduction d'un offset. Les deux méthodes seront mises en application et nous conserverons le meilleur modèle au sens de la qualité prédictive et du critère de parcimonie.

La validation et la comparaison des différents modèles s'effectuera à l'aide du calcul des indicateurs présentés auparavant : la spécificité, la sensibilité et le taux d'erreur. Nous respectons le principe de division des données en deux bases différentes car le modèle est appris sur la base échantillonnée et sa qualité prédictive est calculée sur l'ensemble de données de l'historique.

3.4 L'exposition au risque au sein de la régression logistique

Dans un modèle de fréquence, la prise en compte de l'exposition au risque permet d'introduire le temps d'observation du risque habitation. En effet, il s'exprime généralement par un intervalle de temps $[0, 1]$. 1 correspond à un risque habitation présent en portefeuille une année entière sans modification du profil de risque. Par exemple, si le profil de risque change ou que le risque n'est plus assuré après six mois, alors le temps d'exposition sera de 0,5. Dans un modèle d'estimation d'une prime pure, ce facteur est une constante qui modifie le risque de base propre à chacun, non lié à un profil de risque en particulier. Cette constante est introduite dans un modèle

log-poisson par le logarithme de cette exposition.

Cependant, pour une régression logistique, la prise en compte de l'exposition au risque n'est pas si simple. Il n'existe quasiment aucun article abordant ce sujet. Arthur Charpentier a abordé ce sujet dans ses cours [9] ainsi que dans l'un de ces posts sur son blog [10]. Il propose une approche en s'appuyant sur un processus de Poisson. L'une des solutions qu'il a trouvée est de modéliser l'événement $[Y = 0]$ au lieu de l'habituel $[Y = 1]$, par la fonction de lien log-log complémentaire et d'introduire l'exposition au risque par un offset égal au logarithme de celle-ci. Cependant dans le cadre de notre étude, le *response-based-sampling* est présenté par les auteurs pour une régression logistique modélisant l'événement $[Y = 1]$ avec une fonction de lien logit. Ainsi, la solution proposée par Arthur Charpentier semble difficile à mettre en place car nous ne savons pas si les corrections du *response-based-sampling* présentés dans le cadre d'une régression logistique classique sont valables pour un régression logistique avec une fonction de lien différente.

Nous avons donc décidé pour cette modélisation de procéder sans introduire l'exposition au risque. Cette simplification est acceptable. En effet, ce sont les risques ayant une exposition inférieure à 1, qui possèdent une plus forte chance de mauvaise estimation de leur prime pure, et donc une mauvaise estimation de l'effet géographique. Pour rappel, l'effet géographique est la différence entre la prime pure observée et la prime pure estimée. Or, la prime observée est égale au quotient entre la charge annuelle et l'exposition au risque totale (ou temps d'exposition). Cependant, l'effet géographique est ensuite agrégé par code postal en pondérant par le temps de présence en portefeuille par risque habitation. Ainsi, des risques habitation peu présents auront une faible influence au sein de l'effet géographique du code postal. De plus, si l'on regarde l'exposition au risque au sein de notre base de données, on remarque que plus 50 % sont présents toute l'année, et pour les risques habitations sinistrés plus de 75 % ont une exposition au risque de 1. Nous considérons donc ce modèle sans introduire l'exposition au risque par simplification.

3.5 Résultats

Le modèle de survenance de sinistres inondation sur une année est une régression logistique. Elle s'effectue sur une base échantillonnée, car la trop forte présence de risques habitation non sinistrés empêche le bon fonctionnement du modèle. Elle est corrigée par la méthode du *response-based-sampling*. Il faut retenir que par simplification, l'exposition au risque n'est pas prise en compte.

Les deux types de corrections du *response-based-sampling* ont été testées afin de

déterminer le meilleur modèle. Pour cela, il faut déterminer les poids et le facteur de correction. Après échantillonnage, les poids w_1 et w_0 valent respectivement 0,002658 et 1,329083. Alors que l'introduction du facteur de correction pour la *prior correction* vaut 6,214595.

Tous les modèles suivants vérifient le critère de significativité locale et globale (test de Wald et déviance). Ils ont été construits par respect également du sens métier. Les variables sont anonymisées pour le respect du déroulé tarifaire de l'entreprise. De plus, n'étant pas le cœur du sujet de ce mémoire, nous ne présentons pas l'évaluation des coefficients de régression et leur interprétation mais seulement le choix du meilleur modèle au vue du critère de la qualité prédictive. Ces résultats pour le modèle sélectionné se trouvent néanmoins en Annexe A.3. Le symbole $*$ indique l'introduction d'une interaction entre les variables. Elle est utilisée pour les variables corrélées entre-elles. Un bon modèle de régression logistique se définit par un taux d'erreur le plus faible possible, une sensibilité et une spécificité les plus élevées.

Dans le tableau II.3 sont référencés les résultats pour la *weighting method*.

Modèle	Variables	Seuil s	Sensibilité	Spécificité	Taux erreur
1	Var5, Var25	0,0006	62%	69%	36%
2	Var21, Var25	0,0009	15%	96%	5%
3	Var5, Var25∗Var22	0,0007	60%	71%	29%
4	Var7, Var21, Var22	0,00013	19%	95%	6%
5	Var11, Var15	0,0005	85%	44%	55%

TABLE II.3 – Résultats de la régression logistique corrigée par *weighting method*

Dans un premier temps, on peut remarquer que les modèles comptent très peu de variables finales. Un faible nombre de variables respecte le critère de parcimonie pour s'assurer d'une meilleure qualité prédictive. Le modèle 2 présente un taux d'erreur très faible mais une sensibilité très faible. Pour rappel, la sensibilité est le taux de vrais positifs. Ainsi, le modèle a du mal à classer les risques habitations sinistrés. Le taux d'erreur est faible car le nombre de sinistrés est de l'ordre de 0,07 % de la base d'origine et donc l'erreur sur les vrais positifs est négligeable dans le taux d'erreur. Nous privilégions les modèles avec une sensibilité élevée. Ainsi les modèles 1, 3 et 5 présentent une bonne sensibilité. Seul le modèle 5 a un taux erreur trop élevé. Les modèles 1 et 3 sont similaires. Nous avons décidé de conserver le modèle 3 comme modèle final pour la *weighting method* car il présente un taux d'erreur plus faible que le modèle 1.

Le tableau II.4 contient les résultats de la seconde correction, la *prior correction*. Les modèles présentent un nombre de variables conséquent, ce qui est moins bon

pour le respect du critère de parcimonie d'un modèle. De plus, on remarque que la sensibilité est globalement meilleure que précédemment mais le taux d'erreur et la spécificité ne montrent pas d'améliorations significatives.

Modèle	Variables	Seuil s	Sensibilité	Spécificité	Taux erreur
1	Var3, Var4, Var5, Var12, Var14, Var19, Var21, Var25, Var26	0,0007	68%	68%	32%
2	Var3, Var4, Var5, Var12, Var13, Var19, Var21, Var26	0,00065	67%	68%	34%
3	Var3, Var4, Var7, Var10, Var12, Var19, Var21	0,0006	70%	70%	30%

TABLE II.4 – Résultats de la régression logistique corrigée par la *prior correction*

Pour le respect du critère de parcimonie, nous choisirons pour modèle de survenance de sinistres sur une année, le modèle numéro 3 corrigé par *weighting method*.

4 Modélisation de la charge annuelle

Afin de terminer la modélisation de la prime pure sans effet géographique, il reste à modéliser la charge annuelle pour chaque risque habitation.

4.1 Choix du GLM

Pour rappel, les sinistres jugés extrêmes ont été retirés de la base afin de capter une tendance et non des événements climatiques exceptionnels. La première étape consiste à choisir la loi adéquate pour la modélisation de la charge annuelle inondation. En tarification non vie, le coût d'un sinistre est modélisé généralement par une loi gamma ou une loi log-normale. Or, dans notre cas, un risque habitation a une très faible probabilité d'avoir plus d'un sinistre dans l'année. Ainsi, la distribution du coût moyen d'un sinistre inondation est similaire à celle de la charge annuelle. Sur le graphique II.3, on retrouve les QQ-plot des deux distributions citées au regard de la distribution de la charge annuelle inondation.

Pour rappel, le QQ-plot, aussi appelé diagramme quantile-quantile, est un outil statistique permettant de connaître l'ajustement d'une distribution observée à une

distribution théorique. Pour cela, on calcule les quantiles x_i au sein de la distribution observée. Puis, on fait de même avec ceux de la distribution théorique que l'on souhaite comparer. On note ces quantiles x_i^*. Le graphique représente le nuage de points (x_i^*, x_i). Si les points sont alignés alors $x_i = ax_i^* + b$. Donc les deux distributions suivent le même modèle théorique à une transformation affine près.

Par comparaison des deux QQ-plot, la distribution de la charge se rapproche plus d'une loi log-normale, puisque le QQ-plot est très proche de la droite affine contrairement à celui de la loi gamma. Cependant les dernières valeurs s'éloignent fortement de la bissectrice, ce qui nous fait penser que le choix du seuil de séparation entre sinistres attritionnels et sinistres graves n'est pas optimal. On peut également supposer qu'il existe des sinistres intermédiaires qui ne possèdent pas le même profil que les attritionnels. Il n'en reste pas moins que la loi log-normale semble plus adaptée à nos observations sinistres.

On souhaite modéliser la charge annuelle par une loi log-normale. Cependant, cette dernière ne fait pas partie de la famille exponentielle, et donc, ne peut pas être employée dans le cadre de GLM. On va donc se ramener à une structure de loi de famille exponentielle par l'équivalence :

$$Y \rightsquigarrow \mathbb{LN}(\mu, \sigma^2) \iff ln(Y) \rightsquigarrow \mathbb{N}(\mu, \sigma^2)$$

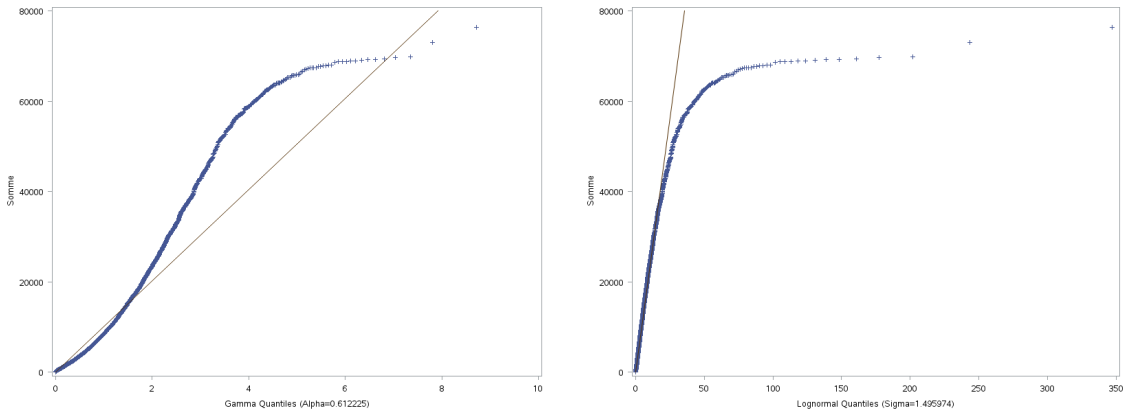


FIGURE II.3 – QQ-plot loi gamma (à gauche) et QQ-plot loi log-normale (à droite) pour la charge annuelle hors sinistres extrêmes

Par passage au logarithme, la distribution de la charge suit une loi Normale. Cette loi de probabilité, quant à elle, fait bien partie de la famille exponentielle et donc l'application des modèles linéaires généralisés est possible pour la variable ln(charge). Cependant, il y a un point d'attention pour la prédiction des valeurs,

puisque l'objectif est d'obtenir la valeur estimée de la charge et non de ln(charge). On pose C la charge annuelle inondation pour un risque habitation en portefeuille. Il faut noter que :

$$\mathbb{E}\left(ln(C)|X\right) \neq ln\left(\mathbb{E}(C|X)\right)$$

Mais on a :

$$\mathbb{E}(C|X) = e^{\mu + \frac{\sigma^2}{2}}$$

avec $\mu = E(ln(C)|X)$. Pour un GLM loi normale, on retrouve la régression linéaire multiple avec inférences statistiques, qui postule que les résidus du modèle ϵ_i sont normaux, centrés et homoscédastiques, soit $\mathbb{N}(0, \sigma^2)$. Le terme σ^2 correspond donc à la variance des résidus du modèle. L'estimateur sans biais de cette valeur est :

$$\hat{\sigma}^2 = \frac{1}{n-p}\sum_{i=1}^{n}\epsilon_i^2$$

avec n le nombre d'individus du modèle, p le nombre de covariables significatives, ϵ_i les résidus du modèle GLM.

4.2 La validation croisée

Nous ne possédons qu'environ 13 000 observations ayant une charge annuelle non nulle. Afin de favoriser la bonne estimation du modèle nous souhaitons conserver cet ensemble pour l'estimation des coefficients de régression du GLM Normale. Mais cela ne permet pas de respecter l'un des critères de validation de l'ajustement du modèle qui est la séparation des données en une base d'apprentissage et une base de test pour la qualité prédictive.

Généralement à l'aide de la base d'apprentissage, le modèle est calibré. Puis, les prédictions sont calculées sur les individus de la base de test. On applique une mesure sur cette base, communément, le *Root Mean Square Error* (RMSE) défini par :

$$RMSE = \frac{1}{n}\sqrt{\sum_{i=1}^{n}\left(Observation_i - Prediction_i\right)^2}$$

Cette métrique est calculée pour chacun des modèles retenus. Celui ayant le RMSE le plus proche de 0 sera retenu comme modèle final car il minimise la qualité prédictive.

Le problème qui se pose avec le risque inondation est l'existence dans notre base de seulement 13 000 observations sur 10 ans. Séparer cette base de données en deux

tables implique de perdre de l'information sinistre pour construire le GLM. Ceci semble préjudiciable dans le cas d'un risque à événements rares. La *k-fold cross validation* permet de résoudre cette difficulté. Elle est adaptée aux modèles calibrés sur de petites bases de données, où l'on souhaite conserver toute l'information disponible pour le calibrage.

Dans la suite de cette partie, la méthode et le processus de mise en place ne seront pas présentés en détails. Certains paramètres tels que le choix du type de validation croisée et de la métrique seront omis.

Formellement on suppose avoir plusieurs modèles GLM que l'on notera f_m avec $m \in M$, M l'ensemble des modèles, appris sur une base de données D. La métrique d'évaluation de la performance d'un modèle est notée P. La *k-fold cross validation* consiste à diviser la base de données D en k sous-échantillons. Il est préférable que ces sous-échantillons soient les plus équirépartis possibles.

Le processus de *k-fold cross validation* se décrit comme suit :

- calibrage du modèle f_m sur les $(k-1)$ sous-échantillons ;
- prédiction du modèle sur le k-ème sous-échantillon ;
- calcul de métrique P sur ce k-ème sous-échantillon, noté P_j.

On répète cette opération jusqu'à ce que chacun des sous-échantillons ait servi d'échantillon de validation. Une fois, cette métrique obtenue pour chaque sous-échantillon, on calcule la performance $P(f_m)$ du modèle f_m par :

$$P(f_m) = \frac{1}{k} \sum_{j=1}^{k} P_j(f_m)$$

Cette équation est valable si les k sous-échantillons possèdent le même nombre d'individus. Dans le cas contraire, il faut pondérer cette moyenne par l'effectif de chaque sous-échantillon.

Après avoir obtenu cette mesure pour chaque modèle $f_m, m \in M$, il reste à déterminer le plus performant de tous. Celui-ci respecte la condition de minimisation suivante :

$$\hat{m} = \arg\min_{m \in M} \left(P(f_m)\right)$$

Le RMSE est choisi comme métrique P pour la validation et la comparaison des modèles construits dans le cadre de l'estimation de la charge annuelle. Dans la littérature [11], il n'existe pas de nombre optimal k de sous-échantillons. Mais

certains l'estiment entre 5 et 10. Le temps de calcul machine augmente de manière significative avec k et le nombre de lignes de chaque sous-échantillon. La base de la charge annuelle étant petite, le nombre de sous-échantillons est fixé à 10.

4.3 Résultats

Les résultats des différents modèles pour la charge sont présents dans le tableau II.5. Le meilleur modèle est le modèle qui respecte au mieux le critère de parcimonie, évoqué précédemment et ayant la meilleure qualité prédictive par la minimisation du RMSE issu de la validation croisée. De plus, chacun des modèles du tableau respecte les critères d'ajustement des modèles : la significativité globale et locale (déviance et test de Wald). Comme pour la régression logistique, les interactions sont signalées par le symbole $*$. Parmi les différents modèle, celui qui minimise le RMSE et qui respecte le critère de parcimonie est le numéro 4. L'estimation des coefficients de régression se trouve dans l'Annexe A.4.

Modèle	Variables	RMSE
1	Var4, Var7, Var12, Var14, Var17, Var19, Var21, Var24, Var26	10046,76
2	Var4, Var8, Var12, Var13, Var21, Var24, Var26	10108,8957
3	Var4, Var8, Var12, Var13∗Var25, Var14, Var17, Var21, Var24	10046,77
4	Var8, Var13∗Var25, Var17, Var21	10041,41
5	Var7, Var17, Var21, Var24, Var25	10057,82
6	Var8, Var17∗Var26, Var25	10051,60
7	Var8, Var13∗Var25, Var17∗Var26, Var21	10043,20

TABLE II.5 – Résultats des différents modèles pour charge annuelle inondation

Le RMSE est globalement élevé, ainsi la qualité prédictive du modèle n'est pas parfaite. De plus, si l'on analyse les résidus du modèle de régression linéaire multiple (GLM de loi normale), on peut mettre en évidence l'ajustement incomplet du modèle de charge. Les graphiques II.4 présentent deux représentations des résidus classiques de régression linéaire multiple. Tout d'abord, on observe que l'histogramme des résidus est symétrique et possède l'allure d'une loi normale. De plus, l'histogramme est globalement centré autour de 0

Cependant sur le nuage de points, on observe une tendance au sein des résidus. Il ne faut pas confondre l'apparition d'une tendance avec l'hypothèse d'hétéroscédasticité. Ici, la variance des résidus reste la même le long de l'abscisse. Mais il y a bien une

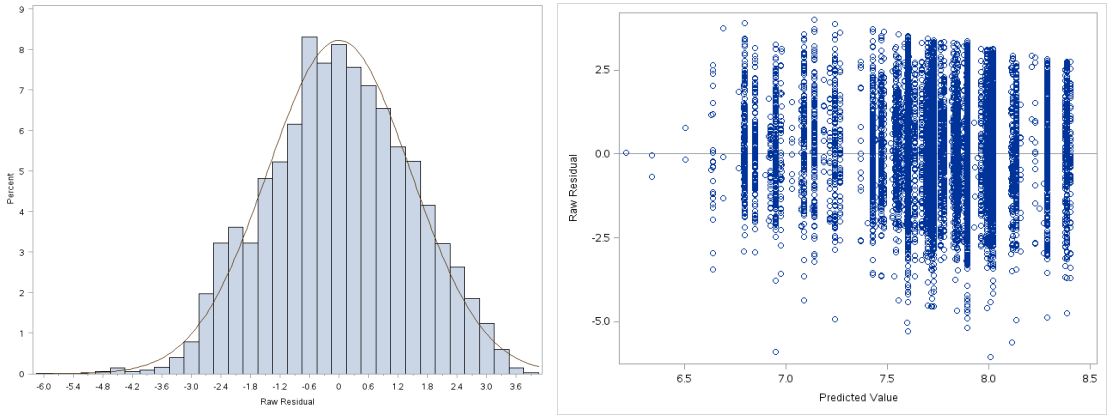


FIGURE II.4 – Histogramme des résidus (à gauche) et nuage de points entre résidus et prédictions du modèle 4

tendance décroissante. Cette structure des résidus, en régression linéaire multiple est spécifique lorsqu'il manque au modèle une variable explicative importante. L'analyse des résidus confirme bien un ajustement correct mais le manque d'information est présent. Cette observation rejoint la remarque sur le RMSE légèrement élevé. Ces phénomènes s'expliquent par l'influence du facteur géographique qui est important dans le cadre du risque inondation, une variable non présente dans la modélisation de la charge annuelle.

L'objectif de ce chapitre était de modéliser l'effet géographique issu de la prime pure pour le risque inondation pour chaque code postal du territoire métropolitain. Cependant l'approche classique a du être adaptée car l'information à notre disposition sur le risque inondation pour le portefeuille GMF est faible. Ainsi, l'approche fréquence × coût moyen est remplacée par la modélisation de la survenance de sinistres dans l'année × celle de la charge annuelle. La validation croisée est utilisée pour l'ajustement du modèle de charge afin de conserver l'ensemble des informations pour l'apprentissage du GLM. La régression logistique pour la survenance de sinistres utilise la méthode du response-based-sampling afin d'obtenir une estimation de la probabilité spécifique aux événements rares. On obtient donc un effet géographique pour les 6048 codes postaux suite à l'agrégation des résidus spatiaux R_j.

Chapitre III

La construction du zonier final

A la suite de l'agrégation, il existe plusieurs codes postaux pour lesquels l'exposition au risque est faible sur les 10 dernières années, ce qui ne permet pas d'accorder une grande fiabilité dans les résultats obtenus pour ces codes postaux. Le lissage permet de résoudre ce problème. Il estime la valeur de ces zones sous-exposées en fonction de celle des autres codes postaux, et ainsi permet de supprimer certaines valeurs aberrantes . Il existe plusieurs techniques de lissage qui seront détaillées afin de choisir la plus adaptée au contexte du risque inondation.

De plus, conserver chaque valeur estimée de l'effet géographique pour chacun des codes postaux est trop important à piloter en tarification. Il est donc important de réduire le nombre de modalités pour intégrer cet effet dans une éventuelle tarification mais également pour savoir si le zonier construit segmente correctement le risque sur le territoire. Pour cela, nous allons créer des classes de risques grâce à la classification ascendante hiérarchique (CAH).

1 Choix du type de lissage

Le lissage d'un zonier est fondamental puisqu'il permet, comme expliqué précédemment, d'obtenir une valeur fiable pour chaque zone. Son deuxième avantage est de lisser le tarif afin que deux codes postaux similaires ne possèdent pas un effet géographique différent.

La première méthode est le lissage proposé par Boskov et Verrall [3]. Il est mis en application pour l'assurance automobile par J. Mathis dans son mémoire [12]. Il faut formaliser l'étude dans un cadre bayésien hiérarchique. L'hypothèse fondamentale est la notion de voisinage, c'est-à-dire qu'une zone adjacente a plus de chances d'avoir un risque similaire qu'une région éloignée géographiquement. Ainsi ce lissage s'appuie sur le risque spatial issu des modèles GLM comme étant une variable aléatoire. L'objectif est d'optimiser sa densité conditionnelle. Ils utilisent le *Monte Carlo Markov Chain*

et l'échantillonneur de Gibbs à l'aide des codes postaux suffisamment exposés.

Le deuxième lissage est prédictif. Il a été utilisé par A. Marlet et D. Adolphe dans leur mémoire [13]. L'objectif est d'expliquer les effets géographiques par des variables externes. Elles peuvent être socio-démographiques, économiques, topographiques ou climatiques. L'hypothèse fondamentale, ici n'est plus une similarité par localisation géographique, mais est une similarité par caractéristique. Ici, deux zones ayant les mêmes caractéristiques ont plus de chances d'avoir un risque similaire. Par cette méthode, on apprend sur les zones suffisamment exposées, puis on prédit à l'aide du modèle considéré.

D'autres techniques de lissage ont été formalisées pour le lissage de l'effet géographique. C. Sepulveda [14] s'est appuyé sur la théorie de la crédibilité et le modèle de Bülmann-Straub pour en construire une autre. Le risque géographique d'une commune dépend donc de son propre risque et du risque des autres communes. L'influence des autres communes décroît avec l'augmentation de la distance entre chacune d'entre elles. L'avantage de cette méthode est qu'on ne tient pas seulement compte des régions voisines mais des communes dans leur ensemble. Ce lissage postule la même idée que celui de Boskov et Verrall : deux communes voisines ont plus de chances d'être similaires que deux communes éloignées.

Le lissage que l'on retient pour la construction du zonier inondation est le lissage prédictif. Il existe deux raisons qui portent notre choix sur celui-ci. La première est que les autres lissages s'appuient sur la notion de voisinage. Or nous avons mis en évidence précédemment que pour le risque inondation, un quartier peut être inondable alors que le voisin ne l'est pas forcément. Ces lissages par similarité de localisation géographique semblent moins adaptés que le lissage prédictif qui préconise la similarité par description de la zone géographique. La deuxième raison est l'ajout d'information sur le risque inondation et sa géographie grâce aux variables externes. Elles complètent ainsi la connaissance et l'influence des facteurs sur le risque inondation en France métropolitaine. Le lissage permet également d'estimer la valeur de l'impact géographique pour des codes postaux où aucune habitation assurée GMF n'est référencée. Cependant, son inconvénient est la sélection des variables externes géographiques. Il est possible de ne pas disposer de toutes les variables souhaitées ou de les choisir de manière erronée.

2 Lissage prédictif de l'effet géographique

2.1 Traitement des variables externes

La première étape du lissage prédictif est la sélection de variables externes décrivant les codes postaux. La plupart des informations en Open Data sur internet sont des données au code Insee. Nous avons donc décidé de récolter ces informations à cette maille puis de les agréger au code postal à l'aide d'une table de correspondance entre ces deux mailles de l'entreprise Covéa. Il y a plus de codes Insee que de codes postaux. Ainsi, il est logique d'avoir plusieurs codes Insee pour un code postal, ce qui n'est pas gênant pour l'agrégation. Par contre, il existe à l'inverse plusieurs codes postaux pour un même code Insee. Dans ce cas, arbitrairement le code postal ayant le plus de risques habitation pour le portefeuille de l'assureur a été sélectionné pour cette correspondance.

Lors de l'utilisation de variables externes, il est primordial de s'assurer que la source de nos données est fiable et reconnue. Le site de l'Insee a fourni des données socio-économiques. Les données climatiques et spécifiques au risque inondation ont été récoltées à l'aide de la base GASPAR (évoquée au premier chapitre de ce mémoire). Les dernières variables utilisées sont issues de la base GEOFLA construit par l'IGN (Institut National de l'Information Géographique et Forestière), qui est un service public certifié. Cette base est remplacée par Admin Express depuis début 2017. Notre historique d'étude allant de 2006 à 2016, nous avons utilisé les données 2016 de la base GEOFLA. Chacune de ces sources est soit gouvernementale, soit certifiée dans son domaine, ce qui nous assure une fiabilité pour les données que nous manipulons.

Ainsi au code Insee, nous avons pu récupérer plusieurs variables externes dont la liste et la description de celles-ci se trouvent dans le tableau III.1.

Le PPRN (plan de prévention des risques naturels) est un plan d'action de l'État pour l'ensemble des risques naturels sur une zone prédéfinie. Le but est de qualifier les enjeux du territoire liés à l'aléa climatique. Il est un indicateur du nombre de procédures au sein de la zone mise en place pour les événements climatiques. Parmi ces variables externes, nous aurions souhaité posséder des variables relatives à la météorologie des communes. Or ces bases sont soit payantes, soit difficilement exploitables, soit à une maille trop grossière par rapport au code postal.

La corrélation des variables géographiques, toutes quantitatives, a été réalisée à l'aide du coefficient de corrélation de Pearson. La matrice des résultats se trouve

Variable	Origine	Descriptif
Log_14	Insee	Nombre de logements en 2014
Log_RP_14	Insee	Nombre de logements en résidences principales en 2014
Log_RS_14	Insee	Nombre de logements en résidences secondaires et occasionnels en 2014
Log_autre_14	Insee	Nombre de logements vacants en 2014
Nv_vie_median_13	Insee	Médiane du niveau de vie en 2013
Population	Insee	Population recensée en 2013
Altitude	Geofla	Altitude moyenne en mètre
Statut	Geofla	Statut administratif
Superficie	Geofla	Superficie en hectares
Nb_arretes_inondation	Gaspar	Nombre d'arrêtés Cat-Nat de 1982 à 2016
PPRn	Gaspar	Nombre de plans de prévention des risques naturels
Densité	Calculé	Densité au km carré $(Population/(Superficie \times 0.01))$
Ratio_Arr	Calculé	Nombre arrêtés Cat-Nat/Nombre de code Insee dans le code postal
Ratio_PPRn	Calculé	Nombre PPRn/Nombre de code Insee dans le code postal

TABLE III.1 – Liste des variables externes pour le lissage prédictif

sur le graphique III.1. Au sein de cette matrice de corrélation, toutes les valeurs supérieures à 0.7 sont en rouge, et les corrélations modérées comprises entre 0.5 et 0.7 sont en jaune. Les variables décrivant la structure des logements, ainsi que la population sont fortement corrélées. Ce qui est plus surprenant est la corrélation entre le nombre d'arrêtés Cat-Nat inondation et le niveau de vie médian au sein du code postal.

Ces variables géographiques que l'on appelle variables externes vont nous permettent de lisser l'effet géographique. Elles sont importantes car elles complètent l'information géographique sur le risque inondation.

2.2 Lissage prédictif par forêts aléatoires

Comme expliqué le lissage prédictif consiste à l'estimation de cet effet géographique à l'aide de variables externes présentées précédemment. L'effet géographique

	Superficie	Population	Altitude	Log_14	Log_RP_14	Log_RS_14	Log_autre_14	Nv_vie_median_13	Nb_arretes_inondation	PPRn	Densite	Ratio_Arr	Ratio_PPRn
Superficie	.	-0,033	0,225	-0,017	-0,031	0,078	0,018	0,641	0,538	0,354	-0,229	-0,225	-0,281
Population	-0,033	.	-0,134	0,989	0,996	0,364	0,943	-0,045	0,024	0,004	0,434	0,421	0,180
Altitude	0,225	-0,134	.	-0,100	-0,119	0,118	-0,081	-0,006	-0,054	0,018	-0,142	-0,244	-0,079
Log_14	-0,017	0,989	-0,100	.	0,995	0,464	0,960	-0,058	0,017	-0,002	0,412	0,430	0,173
Log_RP_14	-0,031	0,996	-0,119	0,995	.	0,373	0,948	-0,055	0,014	-0,004	0,419	0,410	0,170
Log_RS_14	0,078	0,364	0,118	0,464	0,373	.	0,427	-0,063	0,010	0,006	0,143	0,356	0,120
Log_autre_14	0,018	0,943	-0,081	0,960	0,948	0,427	.	-0,034	0,038	0,010	0,360	0,409	0,144
Nv_vie_median_13	0,641	-0,045	-0,006	-0,058	-0,055	-0,063	-0,034	.	0,780	0,530	-0,188	-0,254	-0,264
Nb_arretes_inondation	0,538	0,024	-0,054	0,017	0,014	0,010	0,038	0,780	.	0,601	-0,139	0,097	-0,089
PPRn	0,354	0,004	0,018	-0,002	-0,004	0,006	0,010	0,530	0,601	.	-0,101	-0,024	0,361
Densite	-0,229	0,434	-0,142	0,412	0,419	0,143	0,360	-0,188	-0,139	-0,101	.	0,194	0,170
Ratio_Arr	-0,225	0,421	-0,244	0,430	0,410	0,356	0,409	-0,254	0,097	-0,024	0,194	.	0,400
Ratio_PPRn	-0,281	0,180	-0,079	0,173	0,170	0,120	0,144	-0,264	-0,089	0,361	0,170	0,400	.

FIGURE III.1 – Matrice de corrélation de Pearson pour les variables externes géographiques

prend ses valeurs dans $\mathbb{R}$. La première approche est d'envisager à nouveau l'utilisation des modèles linéaires généralisés. Au sein des lois classiques, citées dans la partie sur les GLM 2, seule la loi Normale peut être envisagée car son ensemble de définition sur $\mathbb{R}$ correspond. Le graphique III.2 représente le QQ-plot de la distribution de l'effet géographique par rapport à la loi Normale. On constate que cette loi n'est pas adaptée car la bissectrice n'est pas suivie.

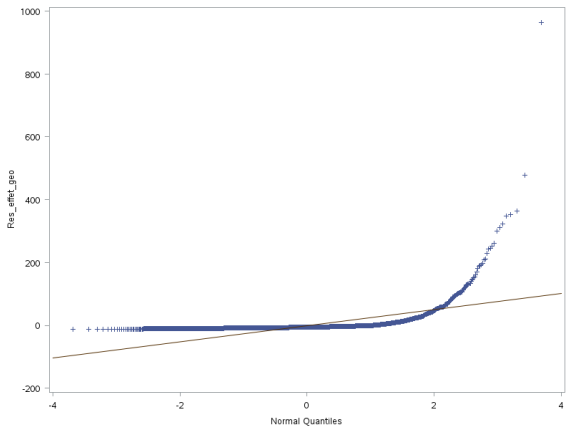


FIGURE III.2 – QQ-plot entre la distribution de l'effet géographique de la base d'apprentissage et la loi normale

Il semble donc difficile de trouver une loi qui soit adéquate à une modélisation par GLM. Ainsi, la modélisation paramétrique ne semble pas convenir. Pour cette raison, notre choix se tourne vers la modélisation non paramétrique, puisqu'elle ne nécessite pas d'hypothèse quant à la loi de distribution de la variable d'intérêt. Elle est connue en actuariat grâce aux algorithmes de data science. Ces algorithmes sont des méthodes d'apprentissage statistique. Le but est non plus d'expliquer les données mais d'apprendre des données. Parmi eux, on trouve notamment les arbres de décisions, les forêts aléatoires ou les réseaux de neurones. Il est possible grâce à eux de s'affranchir de l'hypothèse de distribution. Cependant, nous recherchons à avoir une certaine visibilité sur le fonctionnement de l'algorithme et sur l'influence des variables. C'est pourquoi nous écartons les réseaux de neurones, jugés comme des "boites noires". Ils sont difficilement interprétables et compréhensibles. Il reste donc les arbres CART et les forêts aléatoires.

Les arbres CART sont des arbres de décision qui permettent de segmenter les individus selon les variables explicatives pour construire des classes les plus homogènes possibles au vu de la variable d'intérêt. Cependant, ils sont instables dans leur estimation car ils dépendent fortement de la base d'apprentissage. Les forêts aléatoires ont été créées afin de résoudre ce problème d'instabilité. Cette méthode est une agrégation d'arbres CART afin de rendre l'estimation plus robuste et plus précise. Les avantages de l'agrégation sont de réduire la variance, le biais et le risque de modèle. De plus, on conserve une certaine lisibilité quant à l'importance des variables au sein de l'algorithme. Les forêts aléatoires représentent donc le meilleur compromis entre stabilité et lisibilité pour le lissage prédictif de l'effet géographique.

Les forêts aléatoires (ou *Random Forest*) ont été introduites par Breiman en 2001 [15]. Il convient, dans un premier temps, de rappeler l'aspect théorique des arbres CART puis de définir l'amélioration qu'apportent les forêts aléatoires ainsi que leur manipulation.

2.2.1 Les arbres de décision (algorithme CART)

Les arbres CART sont des arbres de décision qui permettent de segmenter les individus selon les variables explicatives pour construire des classes les plus homogènes possibles au vu de la variable d'intérêt. La variable d'intérêt peut être qualitative ou quantitative. On parle soit d'arbre de classification pour la première soit d'arbre de régression. Dans notre cadre, l'effet géographique que nous souhaitons lisser est quantitatif. La théorie suivante présente donc les arbres de régression.

Un arbre est une classification des individus présents dans la base d'apprentis-

sage. La lecture de l'arbre est descendante. Les classes sont déterminées selon des tests (ou questions) binaires sur les variables explicatives. On peut décrire un arbre selon le schéma suivant III.3. Cette représentation schématique illustre des variables explicatives X_1 et X_2 quantitatives mais il est tout à fait envisageable de posséder des variables explicatives qualitatives.

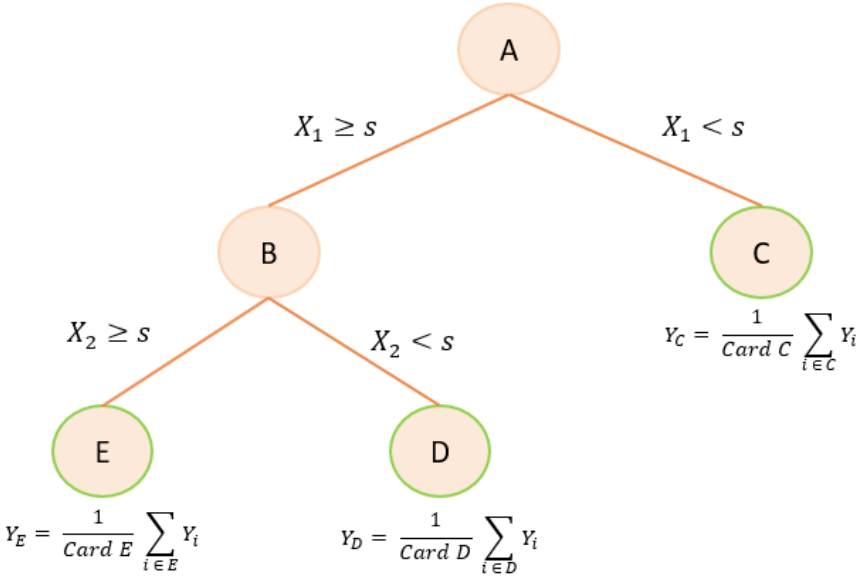


FIGURE III.3 – Représentation théorique d'un arbre de décision

Il est composé d'un nœud racine en haut de l'arbre, puis de nœuds intermédiaires et enfin de feuilles présentes aux extrémités de l'arbre (en vert). Chaque nœud correspond à une question binaire, pour laquelle la réponse détermine la lecture du nœud ou de la feuille suivante de l'arbre. Cette question s'exprime selon une variable explicative de la base d'apprentissage. De manière générale, la réponse oui est la branche de gauche et la réponse non, la branche de droite. Les individus sont donc classés selon les réponses à ces différentes questions. Une feuille est une classe d'individus supposés avoir le même effet sur la variable d'intérêt Y. Pour un arbre de régression, la valeur de cette feuille est la moyenne des valeurs Y pour tous les individus appartenant à cet ensemble. Cette valeur donne une estimation de la variable d'intérêt selon les réponses aux questions des nœuds précédant cette feuille. Prenons l'exemple de la feuille E, la valeur estimée pour cette classe est :

$$\hat{Y}_E = \frac{1}{cardE} \sum_{i \in E} Y_i$$

Les divisions à chaque nœud ne sont pas déterminées de manière aléatoire mais respectent un critère d'optimalité. Les questions respectent l'optimisation de fonctions d'hétérogénéité. Lorsque la variable d'intérêt est une variable continue, la fonction d'hétérogénéité est la variance intra-groupe. La meilleure division à un nœud est celle qui maximise la réduction de l'hétérogénéité. Ainsi, pour chaque nœud on répertorie toutes les partitions possibles à l'aide des variables explicatives présentes. Puis la variance intra-groupe est calculée pour chacun des deux groupes issus de la division. Celle qui est retenue maximise la réduction d'hétérogénéité, donc maximise l'équation suivante :

$$\delta = \text{Hétérogénéité A - (Hétérogénéité B + Hétérogénéité C)}$$

$$\delta = \sum_{i \in A} (y_i - \bar{y}_A)^2 - \left(\sum_{i \in B} (y_i - \bar{y}_B)^2 + \sum_{i \in C} (y_i - \bar{y}_C)^2 \right)$$

où $\bar{y}_A$ est la moyenne des y_i pour l'ensemble que définit le nœud A, de même pour le nœud B et le nœud C.

La croissance de l'arbre s'arrête aux feuilles. L'arrêt est effectué soit parce que la classe est homogène ou que la classe respecte un nombre d'individus optimal par feuille déterminé a priori. Par définition, on appelle l'arbre saturé (ou arbre maximal), celui qui contient un seul individu par feuille.

2.2.2 L'amélioration des arbres grâce aux forêts aléatoires

Les arbres de décision n'étant pas assez robustes, Breiman a développé une technique d'agrégation d'arbres, appelée forêt aléatoire. Le but est d'agréger les estimateurs de plusieurs arbres pour obtenir un seul estimateur final plus précis et plus robuste. Cela facilite la réduction de la variance de l'estimateur. Les forêts aléatoires sont intéressantes pour cette raison, mais également car les paramètres du modèle sont peu nombreux. On peut néanmoins déplorer une perte de lisibilité quant au sens de l'influence d'une variable explicative sur la variable d'intérêt.

Les forêts aléatoires s'optimisent suivant deux paramètres. Ils permettent d'introduire des aléas dans l'algorithme pour s'assurer du bon fonctionnement. Le premier est le *bagging* ou (*bootstrap aggregating*) et le second est le tirage aléatoire de variables explicatives pour chaque nœud.

Le *bagging* (ou *bootstrap aggregating*)

Le *bagging* est une construction de plusieurs arbres CART par *bootstrap*. En effet,

on détient au départ une base d'apprentissage de taille n. On crée par *bootstrap* B échantillons de taille n. Ces échantillons sont donc réalisés par tirage aléatoire avec remise d'individus au sein de la base d'apprentissage. A l'aide de ces B échantillons, on construit B arbres de décisions. Les arbres sont saturés, pour rappel cela signifie construits jusqu'à obtenir un seul individu dans chaque feuille.

Une fois ces B arbres disponibles, on dispose de B estimateurs notés β_j qui sont agrégés afin d'obtenir un estimateur unique pour chaque individu. La variable étant continue, l'estimateur final β_a est égal à la moyenne des estimations des B arbres soit :

$$\beta_a = \frac{1}{B} \sum_{j=1}^{B} \beta_j$$

La contrepartie de cette modélisation est la perte de lisibilité. Il est dorénavant impossible d'obtenir une représentation graphique avec les questions binaires déterminant la classe de l'individu. Nous n'obtenons que l'estimation final des arbres agrégés.

Choix du sous-ensemble de variables explicatives

Le deuxième paramètre introduit au sein des forêts aléatoires est le nombre de variables sélectionné pour réaliser la division de chaque nœud. Pour chaque nœud des arbres, un nombre fixé a priori de variables sera tiré aléatoirement. Le bon fonctionnement de cette agrégation d'arbres CART suppose l'indépendance entre les arbres. Cette indépendance est assurée par le choix de ne pas sélectionner toutes les variables explicatives à chaque nœud pour déterminer les partitions possibles. Cette sélection permet la prise en compte de toutes les variables et empêche la prépondérance d'une variable trop importante par rapport aux autres.

Ils existent des valeurs par défaut pour ce paramètre. Ces valeurs sont fixées selon le type d'arbre de décision, avec p le nombre de variables explicatives :

- dans le cadre d'un arbre de régression, la valeur est $p/3$
- dans le cadre d'un arbre de classification, la valeur est $\sqrt{p}$

Ajustement et interprétation du modèle

Comme pour chaque type de modélisation, il est important de connaître l'ajustement du modèle et pouvoir interpréter un sens final. Malgré la perte de lisibilité du modèle, évoquée précédemment, il existe une mesure de l'importance des variables explicatives qui aide à la détermination des plus influentes.

Dans un premier temps, il faut définir un critère de bon ajustement du modèle, un critère qui permet de comparer les différentes forêts aléatoires construites. L'algorithme contient en son sein ce que l'on appelle l'*out-of-bag*. L'*out-of-bag* est issu de la technique de *bagging*. Nous avons précisé que B échantillons étaient créés par *bootstrap*. Seulement, pour la construction de chacun des arbres, toutes les observations de l'échantillon ne sont pas prises en compte. Certaines ne sont pas utilisées et constituent l'*out-of-bag*. A l'aide de ces observations, l'erreur prédictive de l'arbre est calculée. Ici, puisque la variable d'intérêt est continue, l'erreur est estimée par le MSE (*Mean square error*). Cette valeur est dénommée par l'erreur *out-of-bag* dans la modélisation. On dispose donc de B erreurs *out-of-bag*, qui seront moyennées pour obtenir l'erreur *out-of-bag* finale de la forêt aléatoire. Selon Breiman, cette mesure est aussi efficace que l'utilisation d'un échantillon test, car les arbres sont indépendants.

Une fois que l'on obtient un modèle correctement ajusté, nous allons chercher à connaître une interprétation de celui-ci. Les forêts aléatoires ne permettent pas de connaître le sens de l'influence de chaque modalité pour toutes les variables. Les critères de segmentation des arbres ne sont pas connus. On peut seulement établir un classement de l'influence de chacune des variables explicatives. Cette métrique d'influence utilise l'erreur *out-of-bag*. L'importance d'une variable est évaluée comme l'augmentation marginale de l'erreur *out-of-bag* due à la modification des valeurs des observations. Si une valeur n'est pas importante pour le modèle, alors le changement de ces valeurs ne modifiera pas ou très peu l'erreur *out-of-bag*. Ainsi, en fonction de la grandeur d'augmentation de l'erreur *out-of-bag*, un classement d'importance est établi entre les différentes variables.

2.3 Résultats

Détermination des codes postaux de la base d'apprentissage

Au sein des 6048 codes postaux, il est possible d'avoir des codes postaux faiblement exposés. Si un code postal l'est il est susceptible d'apporter un effet géographique non représentatif. Il est donc important de déterminer une liste des codes postaux pour lesquels nous jugeront qu'ils sont correctement exposés sur les 10 ans d'observation. Si ils ont un exposition suffisante, alors nous considérerons que l'information géographique qu'ils apportent est significative.

Cette liste a été sélectionnée de manière déterministe. Nous possédons le nombre de logements par codes postaux sur l'année 2014 et l'exposition au risque sur l'année 2014. L'exposition au risque sur une année prend sa valeur entre 0 et 1, où 1 correspond à un risque présent toute l'année en portefeuille. Nous avons calculés le

ratio suivant : *Exposition GMF 2014 / Nombre de logements en 2014* pour chacun des codes postaux. Il est à noter que la définition d'un logement selon l'Insee ne comprend pas que les maisons et appartements contrairement à notre base de données. Il n'en reste pas moins un bon indicateur pour sélectionner les codes postaux avec une exposition suffisante.

La distribution du ratio d'exposition s'étend de 0 à 4,27. Les dix codes postaux ayant un ratio supérieur à 1 ne sont pas conservés. De plus, de manière arbitraire nous supprimons les codes postaux avec un ratio inférieur au premier décile de la distribution qui s'élève à 2,025%. La base d'apprentissage du lissage prédictif comprend 5434 codes postaux. On retire donc 614 codes postaux.

Présentation de la forêt aléatoire optimale

Les forêts aléatoires ont été réalisées à l'aide de la librairie *RandomForest* sur le logiciel R. Selon les explications précédentes, il y a deux paramètres à optimiser pour l'utilisation de forêts aléatoires : le nombre d'arbres et le nombre de variables explicatives utilisées pour la construction des nœuds de chaque arbre. Le premier paramètre optimisé est le nombre d'arbres. Pour connaître sa valeur optimale, une forêt aléatoires de 1000 arbres est construite. De manière générale, le nombre d'arbres est fixé selon un compromis entre un nombre à partir duquel la qualité du modèle ne s'améliore plus et un nombre qui n'entraîne pas un temps machine trop important. Puisque notre base est petite, on peu s'affranchir de la contrainte du temps machine pour l'algorithme. On représente sur le graphique III.4 l'erreur de la forêt en fonction du nombre d'arbres. On observe une convergence de l'erreur à partir de 400 arbres. Pour s'assurer d'une robustesse du modèle, on fixe le nombre d'arbres à 500.

Une fois le nombre d'arbres fixé, on cherche à connaître la valeur optimale du nombre de variables explicatives considérées pour chaque nœud des arbres CART. Dans notre modèle, on compte 13 variables géographiques externes. Ainsi, la valeur par défaut est 4. On prend toujours l'arrondi inférieur de $p/3$. Par comparaison des différentes erreurs *out-of-bag* en fonction du nombre de variables, la valeur optimale obtenue est 2.

Par conséquent, la forêt aléatoire permettant le lissage prédictif de l'effet géographique comprend 500 arbres et tire deux variables explicatives pour la division de chaque nœud. Le RMSE sur les 6048 codes postaux s'élève à 19,97 ce qui semble relativement élevé pour une distribution de l'effet géographique issue du GLM entre -13 et 1200 pour la valeur la plus élevée. Pour s'assurer de la bonne cohérence de notre modèle, nous avons étudié plus en détails la distribution de l'effet géographique observé vis-à-vis de l'effet géographique prédit (ou lissé). Ceci est représenté

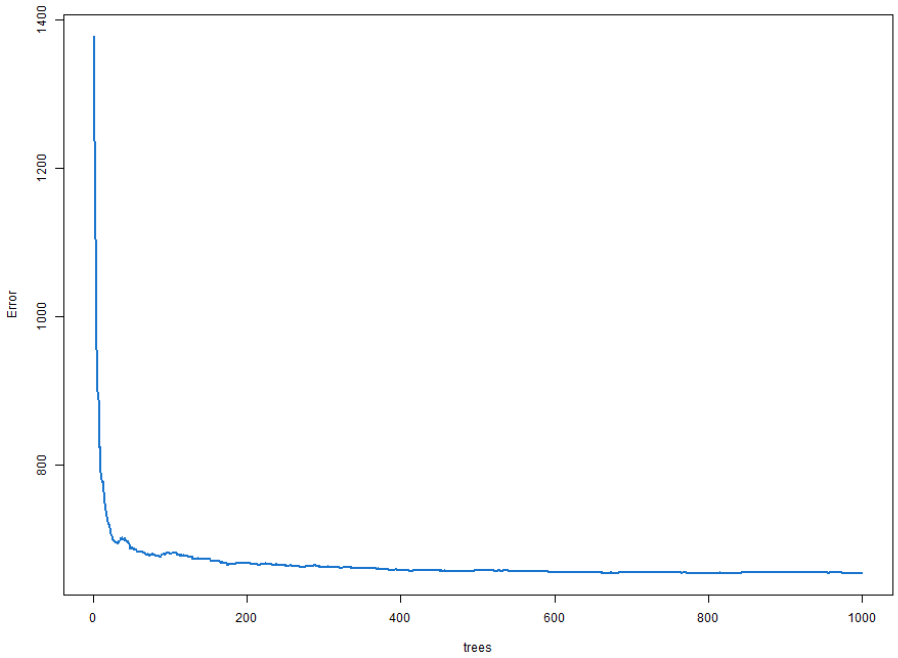


FIGURE III.4 – Erreur de la forêt aléatoire en fonction du nombre d'arbres pour le lissage des résidus spatiaux

dans le nuage de points du graphique III.5. Ce graphique est un zoom sur l'effet géographique inférieur à 100. L'ensemble des points est présent en Annexe A.5. On distingue nettement une tendance linéaire au sein du graphique pour les dernières valeurs.

Cependant, pour les premières valeurs, la tendance linéaire s'efface. Les points sont plus éparpillés. Au vue de cette remarque, on cherche à savoir combien de codes postaux changent de sens pour l'effet géographique, c'est-à-dire combien possèdent un effet observé positif et un effet lissé négatif et inversement. Sur l'ensemble des 6048 codes postaux 10,8% procèdent à ce changement. Ainsi notre modèle semble juste dans sa prédiction. De plus, on remarque que le lissage permet d'atténuer les variables extrêmes car une observation à 400 est lissée par une valeur à 200. Cette forte différence pour les valeurs extrêmes augmente forcément la valeur du RMSE de la forêt aléatoire. En conclusion, cette forêt aléatoire semble effectuer un lissage cohérent.

La dernière étape consiste à l'analyse de l'importance de chaque variable géographique dans le modèle. On peut distinguer trois paquets de variables. Le premier contenant seulement la variable altitude, puis toutes les variables socio-démographiques du code postal, du nombre de logements en résidence secondaire aux logements va-

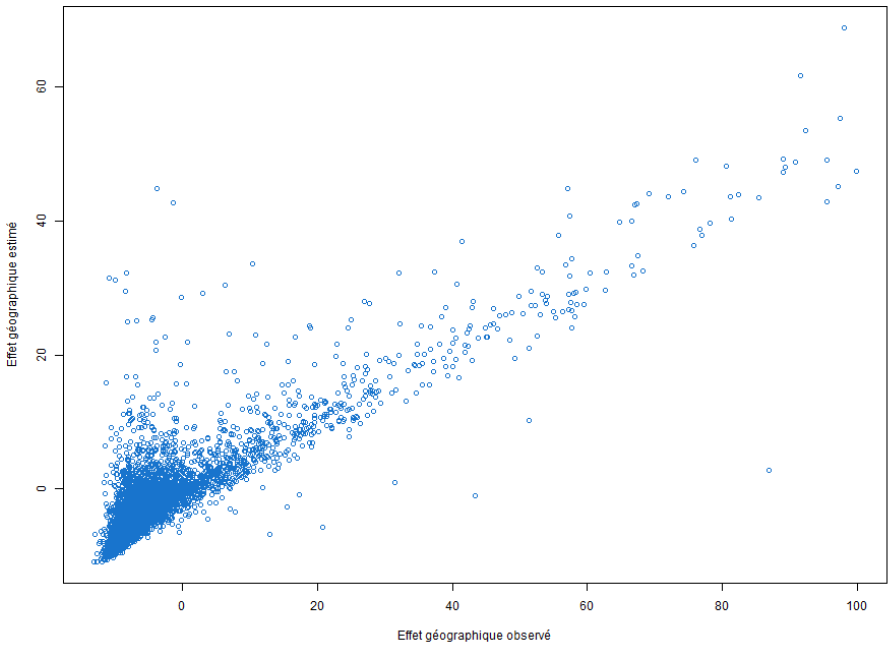


FIGURE III.5 – Représentation des observations en fonction des prédictions de l'effet géographique par les forêts aléatoires

cants. Puis on retrouve les variables décrivant le profil du risque inondation de chaque code postal, avec les arrêtés Cat-Nat et les PPRn. Chacune des variables est importante dans la construction de la forêt aléatoire.

3 Zonier inondation final

Une fois le lissage prédictif effectué, la dernière étape de la construction du zonier final est la création de classes de risque inondation par classification ascendante hiérarchique. Ces classes de risques vont nous permettent de savoir si la méthode choisie a permis d'avoir une segmentation de l'effet géographique sur le territoire et si elle est pertinente au vue des observations.

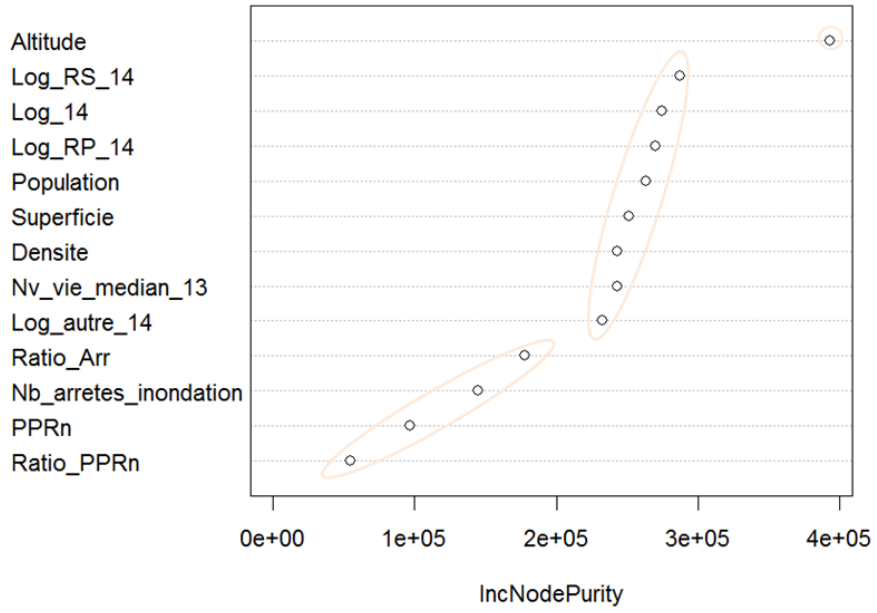


FIGURE III.6 – Importance des variables de la forêt aléatoire

3.1 Théorie de la classification

La méthode de classification permet de partitionner un échantillon de données. Le diviser au mieux, nécessite le respect de deux principes fondamentaux. Les individus d'une même classe doivent être le plus ressemblant possible, et les individus de deux classes différentes doivent être le plus dissemblable.

Parmi les méthodes de classification (*clustering*), la classification ascendante hiérarchique a été sélectionnée puisqu'elle détermine le nombre de classes "optimal" et en extrait des classes homogènes d'individus. Cette stratégie d'agrégation suit le processus suivant, pour n individus à classifier :

- partition en n classes, chaque individu représente une seule classe ;
- calcul des distances entre chaque individu deux à deux selon l'effet géographique lissé. Puis les deux les plus semblables sont regroupés dans une même classe ;
- enfin, les distances entre les individus et cette classe sont calculées et les deux plus proches sont réunis selon le critère d'agrégation.

Ce processus est répété jusqu'à l'obtention d'une seule classe contenant tous les individus.

A chaque étape, on obtient une partition différente de l'ensemble des individus en fonction de la similarité géographique du risque inondation. Chacune de ces partitions est caractérisée par l'inertie interclasse et l'inertie intraclasse. Pour une partition des données, regroupant les n individus en q classes $\{A_1, .., A_q\}$, on a :

$$\text{Inertie interclasse} : \sum_{j=1}^{q} m_j d^2\left(G_{A_j}, G_E\right)$$

$$\text{Inertie intraclasse} : \sum_{j=1}^{q} \sum_{i \in A_j} d^2\left(n_i, G_{A_i}\right)$$

où :

- m_j est le poids de la classe A_j
- G_{A_j} est le barycentre de la classe A_j
- $d^2(,)$ est la distance euclidienne calculée avec les valeurs de l'effet géographique entre deux sous-ensembles
- n_i le i-ème individu de la classe A_j

Lors de la première étape, l'algorithme regroupe les deux individus les plus proches au sens de la distance euclidienne vont former la première classe lors de la seconde itération de la CAH. Puis il faut ensuite calculer la distance entre la classe et les individus. Ceci est permis par le critère d'agrégation. En effet, il détermine la distance entre deux classes ou entre une classe et un individu. Il en existe plusieurs pour une CAH. Dans le cadre de cette étude, le critère de Ward a été préféré aux autres. Cette agrégation sélectionne le regroupement de l'étape suivante, qui provoque la plus petite augmentation d'inertie intraclasse. En effet, une bonne partition d'un espace E doit avoir une inertie intraclasse la plus faible possible et une inertie interclasse la plus élevée possible. Ainsi, on s'assure du respect des deux principes fondamentaux de la classification cités au début de partie. Le critère de Ward définit la distance entre deux classes par la distance de leurs barycentres au carré, pondérée par l'effectif de ces classes :

$$D_{Ward}\left(A_j, A_i\right) = \frac{m_i m_j}{m_i + m_j} d^2\left(G_{A_j}, G_{A_i}\right)$$

Les deux classes A_i, A_j qui se regroupent sont celles qui minimisent le critère de Ward. Cette agrégation est une optimisation à chaque étape et donc ne garantit pas une optimalité globale, mais pour ce zonier, nous ferons l'hypothèse que l'on obtient la meilleure optimisation globale par cette approche.

En pratique, les étapes de regroupement à chaque partition (des n classes à une

seule classe) sont résumées par le dendrogramme. Il représente par un arbre les différentes étapes, en partant des n individus jusqu'à la classe unique. De plus, on trouve sur ce graphique les distances de Ward égales aux branches de l'arbre. Il est notamment utile pour déterminer le nombre de classes final. Ce choix est un compromis entre un nombre de classes souhaité et ce que nous indique la classification. En effet, si la distance de Ward est importante pour le regroupement de cette étape, ceci indique que les classes ne possèdent pas vraiment de ressemblance, et donc, que la partition n'est pas adéquate. De plus, sélectionner un nombre de classes trop important retire l'intérêt de faire une classification sur les codes postaux, dont le but est de réduire le nombre de modalités.

3.2 Résultats

La classification ascendante hiérarchique a été effectuée sur l'effet géographique lissé des 6048 codes postaux. Pour déterminer le nombre de classes pour le zonier inondation, il faut étudier les sauts d'agrégation de chaque étape de la CAH et déterminer le nombre de classes à l'aide d'un compromis entre homogénéité et parcimonie.

Sur le graphique III.7, on représente le critère de Ward à chaque nœud. On cherche à déterminer les regroupements où la perte d'homogénéité est importante. On peut donc remarquer sur le graphique différentes augmentations, une lors du passage de 9 à 8 puis consécutivement de 5 à 2. Un nombre de classes envisageable est le nombre de nœuds que contient l'arbre avant cette forte perte d'homogénéité. Par conséquent, plusieurs classifications sont possibles : 9 zones, 5 zones, 4 zones ou 3 zones. On pourrait imaginer ne prendre que deux zones à la vue du graphique, seulement cela paraît être trop peu pour un zonier en assurance.

Le choix de la classification finale s'effectue selon la segmentation du risque inondation qu'elle entraîne. Pour cela, on évalue la représentativité de chaque zone selon l'exposition au risque et l'évolution des indicateurs de sinistralité dans les différentes zones. On retrouve sur les graphiques III.8, III.9, III.10, III.11 les critères évoqués pour chacune des quatre classifications.

On peut globalement observer sur l'ensemble des quatre classifications, une évolution croissante des trois indicateurs de sinistralité, prime pure, fréquence et coût moyen en fonction de la classe. Ainsi, on peut voir que les zoniers segmentent le risque inondation. Cependant, l'augmentation est forte pour chacune des dernières zones car l'inondation est un risque à événements rares, soit volatile ce qui explique l'explosion des indicateurs.

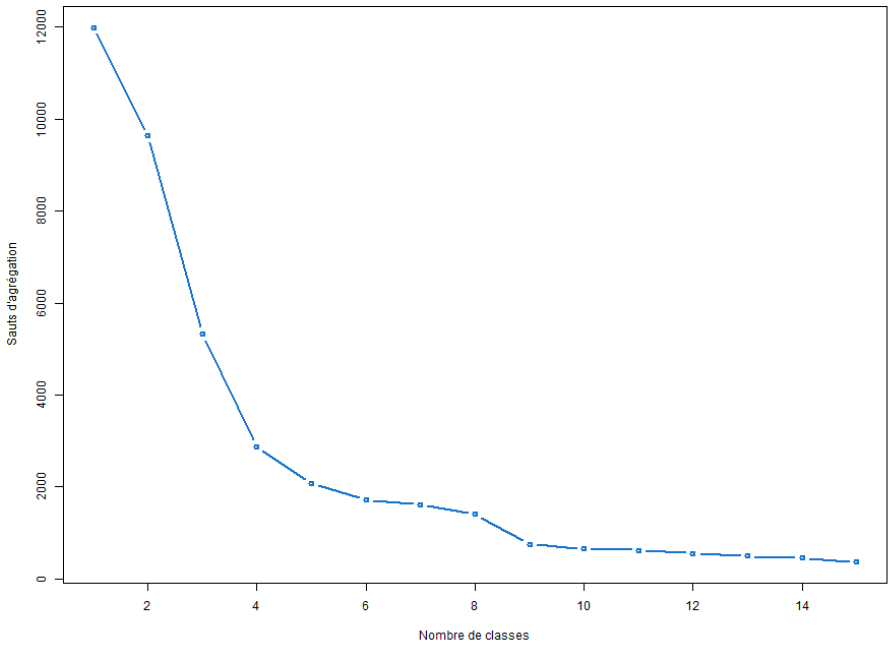


FIGURE III.7 – Choix du nombre de classes par la représentation de l'évolution du critère de Ward à chaque regroupement effectué

L'exposition au risque, représentée par l'histogramme sur chacun des quatre graphiques, est la part d'exposition des 10 ans d'observations présente au sein de la classe concernée. Les dernières zones sont globalement moins exposées et donc sont moins représentatives que les premières. Cette remarque est d'autant plus vraie pour le zonier à neuf classes où il y a moins de 5 % des risques habitations à partir de la zone 6. Une zone non représentative n'est pas viable pour la construction d'un zonier en assurance puisque cela n'assure pas une bonne fiabilité des indicateurs de suivi de sinistralité dans le temps. Les zoniers à trois et quatre zones ont des classes de risque correctement bien exposées. Enfin, la dernière zone de la classification à cinq zones est sous-exposée.

Suite à ces différentes observations, le zonier final choisi est celui à quatre zones. Chacune de ces classes de risques est représentative contrairement aux classifications à cinq et neuf zones. De plus, l'évolution des indicateurs de sinistralité est plus progressive et moins brutale entre les classes, à l'inverse du zonier à trois zones où l'augmentation est très importante entre la zone 2 et la zone 3.

Les analyses précédentes mettent en évidence que le zonier à quatre zones seg-

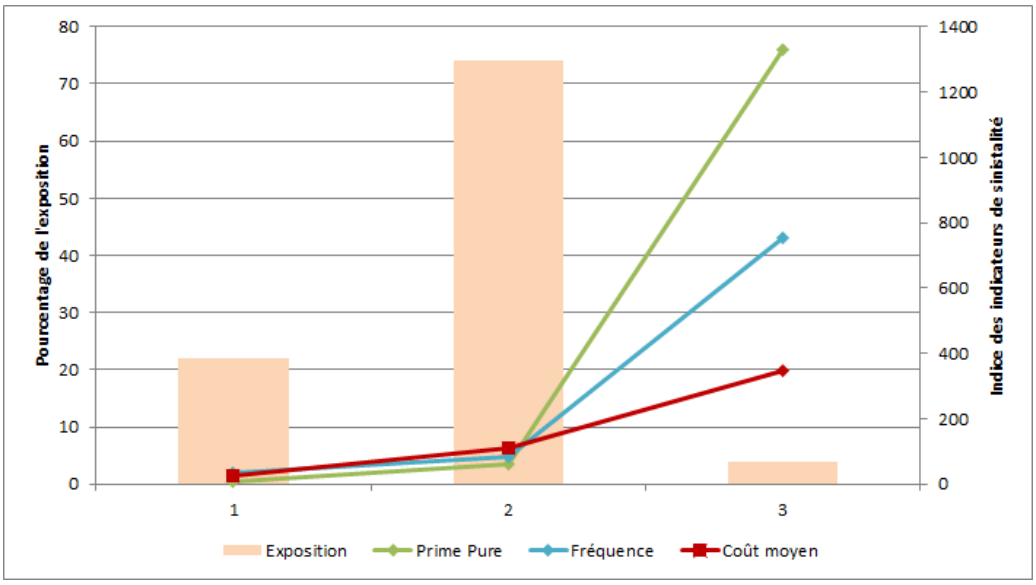


FIGURE III.8 – Segmentation de la classification à trois zones sur les sinistres Cat-Nat et hors Cat-Nat hors sinistres extrêmes de 2006 à 2016

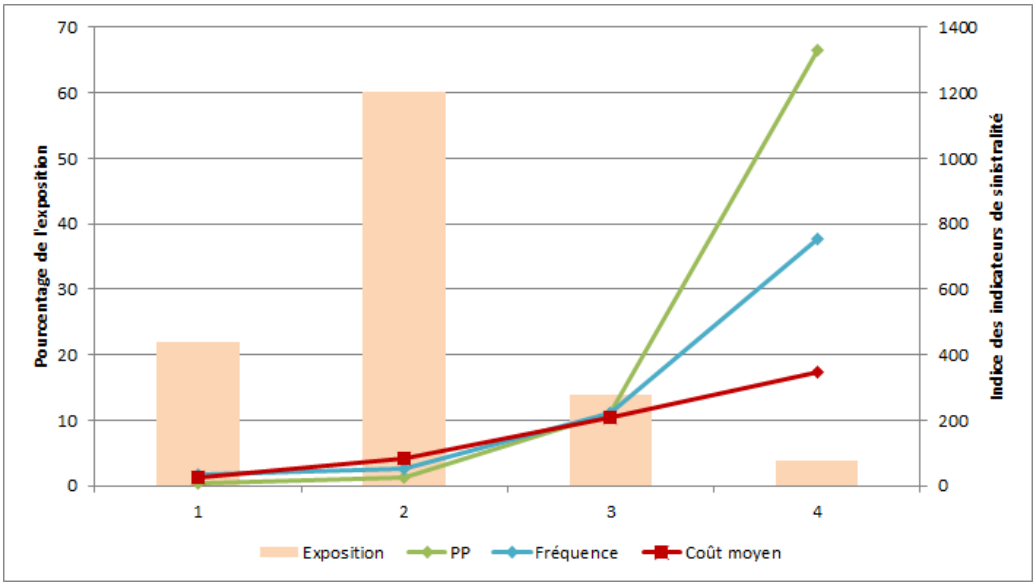


FIGURE III.9 – Segmentation de la classification à quatre zones sur les sinistres Cat-Nat et hors Cat-Nat hors sinistres extrêmes de 2006 à 2016

mentent bien la prime pure de l'inondation. Afin de valider la pertinence de notre zonier, nous allons chercher à savoir s'il représente bien la sinistralité observée sur les 10 dernières années.

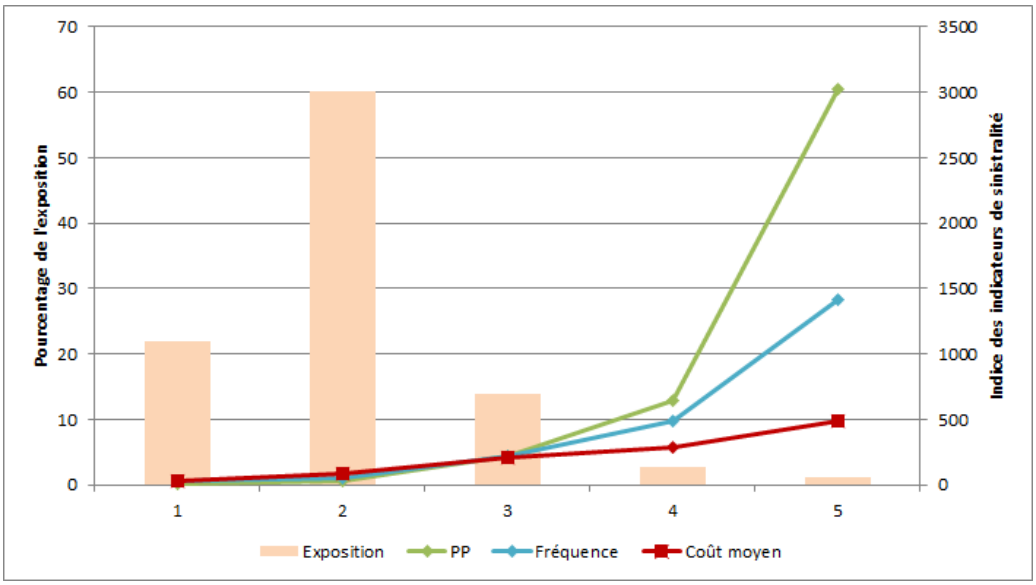


FIGURE III.10 – Segmentation de la classification à cinq zones sur les sinistres Cat-Nat et hors Cat-Nat hors sinistres extrêmes de 2006 à 2016

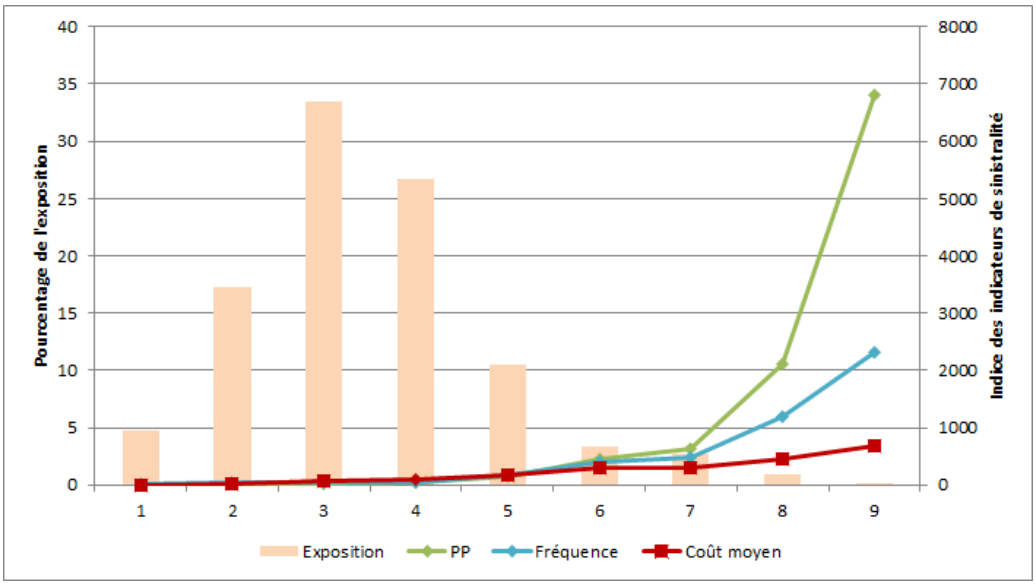


FIGURE III.11 – Segmentation de la classification à neuf zones sur les sinistres Cat-Nat et hors Cat-Nat hors sinistres extrêmes de 2006 à 2016

La carte III.12 est une représentation du zonier final à 4 zones. A l'échelle nationale, le zonier montre que le sud de la France, le long du Rhône, l'ouest de la Corse, le sud Parisien ainsi que l'est de la France sont les régions à plus fort risque

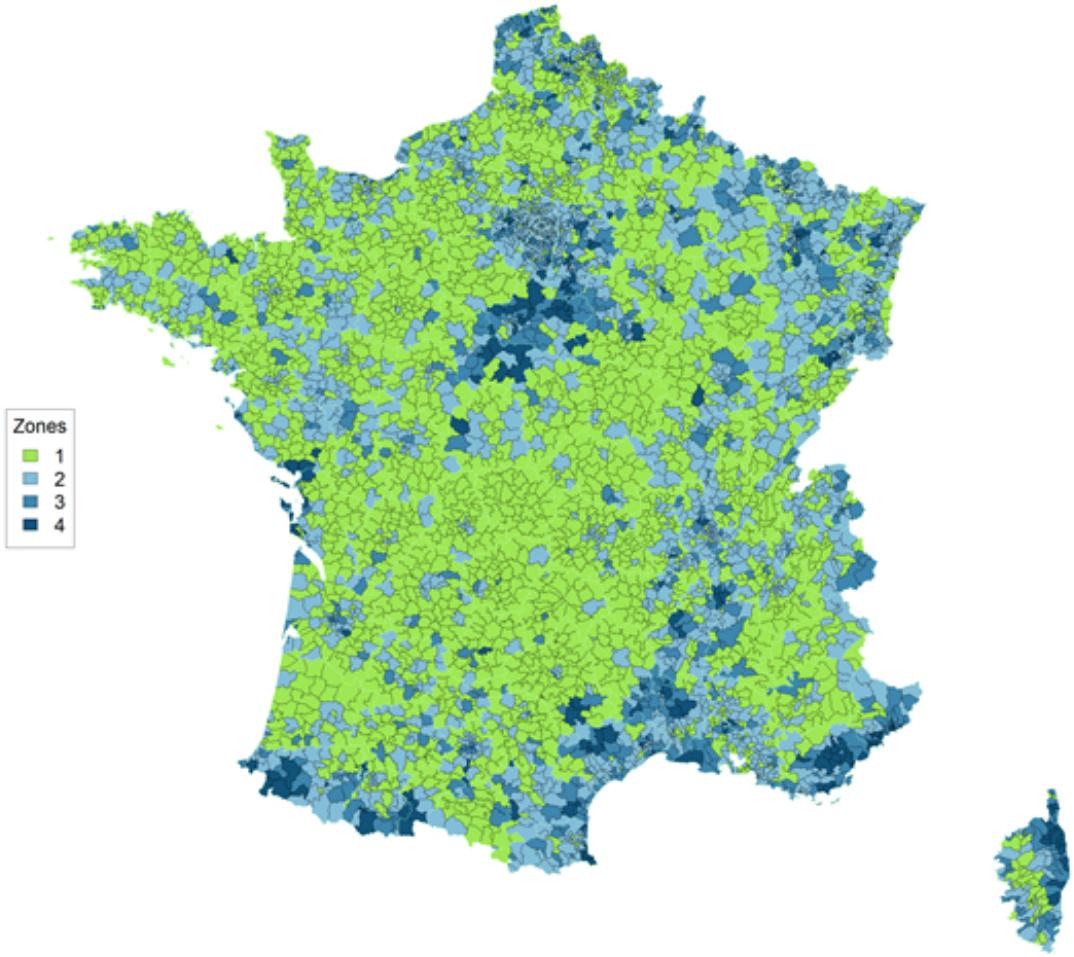


FIGURE III.12 – Cartographie du zonier inondation final à quatre zones à la maille code postal

inondation. Ces zones en concordance avec les observations des arrêtés Cat-Nat et celles issue des cartes I.1, I.2 et A.1.

Le zonier présent est donc pertinent au sens de la prime pure technique du risque inondation sur le portefeuille GMF maisons et appartements pendant 10 ans. Il semble en effet que notre approche a permis d'avoir une bonne utilisation de l'information en notre possession.

Notre méthode est cohérente au sens de l'approche statistique. Elle admet néanmoins quelques limites et quelques améliorations envisageables détaillées dans la

dernière partie du mémoire.

3.3 Limites et améliorations envisageables

Dans un premier temps, il faut avoir conscience que la méthode proposée pour construire le zonier inondation s'appuie sur l'approche statistique et non l'approche physique d'un risque climatique. En d'autres termes, on s'appuie sur un historique de données, et on n'anticipe donc pas une éventuelle sinistralité non survenue. On peut voir par exemple que notre zonier présente la Normandie en zone 1 alors qu'elle recense un nombre d'arrêtés Cat-Nat important. L'approche physique, quant à elle, envisage bien cette éventualité en s'appuyant sur des modèles plus sophistiqués, et des données métrologiques comme le proposent les logiciels évoqués en Introduction.

La méthodologie proposée est adaptée aux risques à événements rares grâce à la régression logistique corrigée par *weighting method* issue du *response-based-sampling* et à la validation croisée. Cependant, l'introduction de l'exposition au risque, habituelle en modélisation de la fréquence, ne semble pas possible si l'on s'appuie sur la proposition d'Arthur Charpentier.

De plus, il est possible d'envisager plusieurs axes d'améliorations au zonier construit. Tout d'abord en étudiant plus précisément la maille choisie. Le code postal pourrait être une maille trop grossière en inondation. Seulement pour un risque à événements rares, le principe de mutualisation est très important, en témoigne le fonctionnement du régime légal des catastrophes naturelles où tout assuré cotise pour un même pourcentage.

La deuxième amélioration serait de faire deux zoniers différents : un de fréquence et un autre de coût moyen pour ne pas mélanger les deux effets géographiques qui sont certainement différents. Dans cette nouvelle modélisation, l'étude du coût d'un sinistre pourrait être plus analysés pour la partie extrême puisque l'on a vu dans l'étude que la loi log-normale n'est pas très adaptée pour les charges annuelles intermédiaires.

De plus, on pourrait également tester la proportion choisie lors du *response-based-sampling*. Il a été fixé arbitrairement à 20 % mais on pourrait envisager d'étudier l'impact de cette proportion sur l'estimation des paramètres.

Le lissage prédictif mis en place pourrait également être mis en comparaison avec un lissage utilisant l'exposition comme facteur de fiabilité. On peut envisager une approche par crédibilité ou par krigeage par exemple.

Le dernier axe est une collecte de données externes plus complète. En effet, avoir des données externes concernant la météorologie de chaque zone compléterait l'information géographique apportée lors du lissage par forêts aléatoires.

Conclusion

L'objectif de ce mémoire était d'arriver à construire un zonier sur le risque inondation par une approche statistique afin de s'affranchir de l'utilisation de logiciels spécialisés pour les événements climatiques. La problématique de ce choix a été d'utiliser au mieux le faible nombre de sinistres présents dans notre base.

Pour ce faire, nous avons mis en place différentes méthodes. Dans un premier temps, le *response-based-sampling* a permis de s'assurer d'une bonne estimation de la probabilité de survenance de sinistres. Puis la validation croisée estime la qualité prédictive de modèles tout en conservant l'ensemble des observations pour l'apprentissage de celui-ci. L'information sur la géographie a ensuite était complétée par l'apport des variables externes lors du lissage.

Nous obtenons finalement une segmentation géographique cohérente selon des 10 dernières années d'observation. Les deux axes d'améliorations principaux seraient d'étudier plus en détails les sinistres extrêmes et de tester un autre lissage qui prendrait en compte la fiabilité de l'exposition comme facteur en utilisant la théorie de la crédibilité. Cependant, la limite de ce modèle reste la non anticipation d'événements climatiques non survenus dans notre historique.

Afin d'envisager une éventuelle insertion de ce zonier dans la tarification du contrat MRH GMF, il resterai à savoir si ce zonier est cohérent commercialement. Le zonier ne doit pas entraîner de trop fortes augmentations de la cotisation d'un assuré entre deux zones ou entre deux codes postaux adjacents.

Bibliographie

[1] Site GOUVERNEMENTAL. *Définition de l'inondation*. URL : `http://www.gouvernement.fr/risques/inondation`.

[2] Code des ASSURANCES. *L'assurance des risques de catastrophes naturelles*. URL : `https://www.legifrance.gouv.fr`.

[3] M BOSKOV et RJ VERRALL. « Premium rating by geographic area using spatial models. » In : *Insurance Mathematics and Economics* 3.16 (1995), p. 274.

[4] Natacha BROUHNS, Michel DENUIT et Bernard MASUY. « Ratemaking by geographical area in the Boskov and Verrall model : a case study using belgian car insurance data ». In : *Actu–L* 2 (2002), p. 3–28.

[5] Xavier MILHAUD. « Pratique de la tarification non vie avancée ». Support de cours. 2017.

[6] P MCCULLAGH et al. « Nelder. JA (1989), Generalized Linear Models ». In : *CRC Monographs on Statistics & Applied Probability, Springer Verlag, New York* (1973).

[7] Gary KING et Langche ZENG. « Logistic regression in rare events data ». In : *Political analysis* 9.2 (2001), p. 137–163.

[8] Charles F MANSKI et Steven R LERMAN. « The estimation of choice probabilities from choice based samples ». In : *Econometrica : Journal of the Econometric Society* (1977), p. 1977–1988.

[9] Arthur CHARPENTIER. « Actuariat de l'assurance non vie slides 2 ». Support de cours Ensae ParisTech. Octobre 2015 - Janvier 2016.

[10] Arthur CHARPENTIER. *Modelling occurence of events with some exposure*. URL : `http://freakonometrics.hypotheses.org/20133#more-20133`.

[11] Sylvain ARLOT. « Validation croisée ». In : (2017).

[12] Julien MATHIS. « Elaboration d'un zonier en assurance de véhicules par des méthodes de lissage spatial basées sur des simulations MCMC ». Mémoire DUAS. 2009.

[13] A. MARLET et D. ADOLPHE. « Leviers de création de valeur en MultiRisque Habitation ». Mémoire CEA. 2017.

[14] C. SEPULVEDA. « Modélisation du risque géographique en Santé, pour la création d'un nouveau Zonier. Comparaison de deux méthodes de lissage spatial ». Mémoire ISUP. 2016.

[15] Leo Breiman. « Random forests ». In : *Machine learning* 45.1 (2001), p. 5–32.

Table des figures

Liste des tableaux

Liste des abréviations, des sigles et des symboles

CAH Classification ascendante hiérarchique

Cat-Nat issu de la garantie réglementaire sur les catastrophes naturelles

DROM-COM Départements ou régions français d'outre-mer et les collectivités d'outre-mer

GLM Generalized linear model ou modèles linéaires généralisés

GPD Generalized pareto distribution

Hors Cat-Nat issu de la garantie inondation du contrat MRH GMF

Insee Institut national de la statistique et des études économiques

MRH Multirisque habitation

PPRn Plan de prévention des risques naturels

RMSE Root mean square error

ROC Receiver operating characteristic

TRI Territoires à risque important d'inondation

Annexe A

Annexes

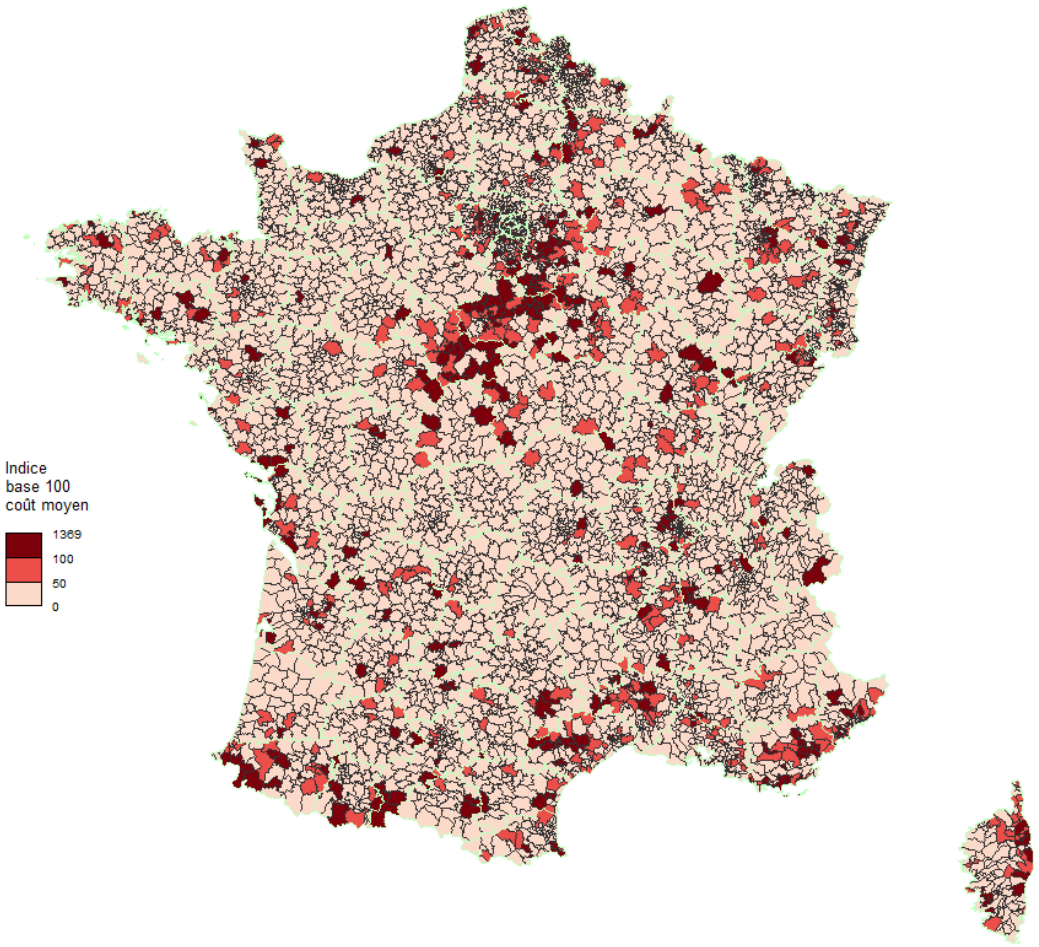
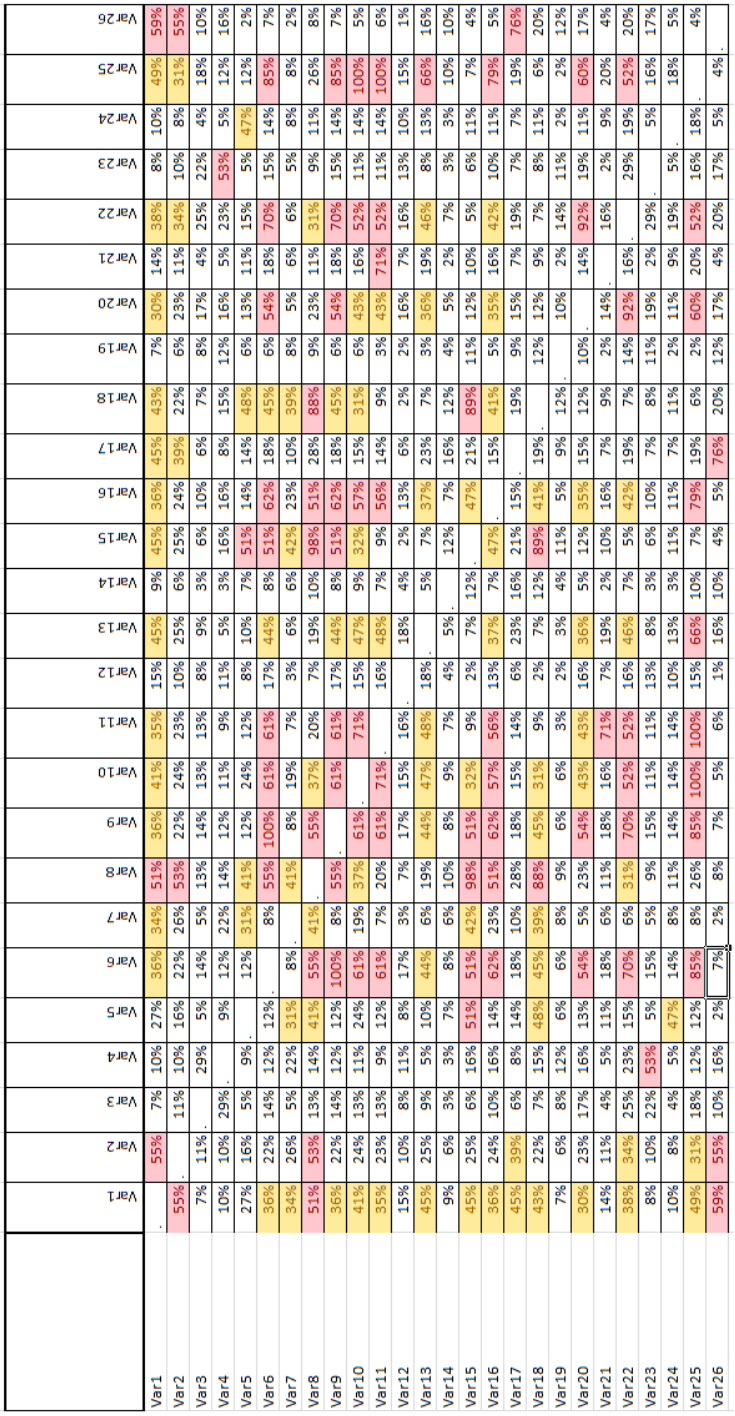


FIGURE A.1 – Indice du coût moyen d'un sinistre par codes postaux, de 2006 à 2016



	Var1	Var2	Var3	Var4	Var5	Var6	Var7	Var8	Var9	Var10	Var11	Var12	Var13	Var14	Var15	Var16	Var17	Var18	Var19	Var20	Var21	Var22	Var23	Var24	Var25	Var26
Var1	.	55%	7%	10%	27%	36%	34%	51%	36%	41%	35%	15%	45%	9%	45%	36%	45%	43%	7%	30%	14%	38%	8%	10%	49%	59%
Var2	55%		11%	10%	16%	22%	26%	53%	22%	24%	23%	10%	25%	6%	25%	24%	39%	22%	6%	23%	11%	34%	10%	8%	31%	55%
Var3	7%	11%	.	29%	5%	14%	5%	13%	14%	13%	13%	8%	9%	3%	6%	10%	6%	7%	8%	17%	4%	25%	4%	4%	18%	10%
Var4	10%	10%	29%	.	9%	12%	22%	14%	12%	11%	9%	11%	5%	3%	16%	16%	8%	15%	12%	16%	5%	23%	53%	5%	12%	16%
Var5	27%	16%	5%	9%	.	12%	31%	41%	12%	24%	12%	8%	10%	7%	51%	14%	14%	48%	6%	13%	11%	15%	5%	47%	12%	2%
Var6	36%	22%	14%	12%	12%	.	8%	55%	100%	61%	61%	17%	44%	8%	51%	62%	18%	45%	6%	54%	18%	70%	15%	14%	85%	7%
Var7	34%	26%	5%	22%	31%	8%	.	41%	8%	19%	7%	3%	6%	6%	42%	23%	10%	39%	8%	5%	6%	6%	5%	8%	8%	2%
Var8	51%	53%	13%	14%	41%	55%	41%	.	55%	37%	20%	7%	19%	10%	98%	51%	28%	88%	9%	23%	11%	31%	9%	11%	26%	8%
Var9	36%	22%	14%	12%	12%	100%	8%	55%	.	61%	61%	17%	44%	8%	51%	62%	18%	45%	6%	54%	18%	70%	15%	14%	85%	7%
Var10	41%	24%	13%	11%	24%	61%	19%	37%	61%	.	71%	15%	47%	9%	32%	57%	15%	31%	6%	43%	16%	52%	11%	14%	100%	5%
Var11	35%	23%	13%	9%	12%	61%	7%	20%	61%	71%	.	16%	48%	7%	9%	56%	14%	9%	3%	43%	71%	52%	11%	14%	100%	6%
Var12	15%	10%	8%	11%	8%	17%	3%	7%	17%	15%	16%	.	18%	4%	2%	13%	6%	2%	2%	16%	7%	16%	13%	10%	15%	1%
Var13	45%	25%	9%	5%	10%	44%	6%	19%	44%	47%	48%	18%	.	5%	7%	37%	23%	7%	3%	36%	19%	46%	8%	13%	66%	16%
Var14	9%	6%	3%	3%	7%	8%	6%	10%	8%	9%	7%	4%	5%	.	12%	7%	16%	12%	4%	5%	2%	7%	3%	3%	10%	10%
Var15	45%	25%	6%	16%	51%	51%	42%	98%	51%	32%	9%	2%	7%	12%	.	47%	21%	89%	11%	12%	10%	5%	6%	11%	7%	4%
Var16	36%	24%	10%	16%	14%	62%	23%	51%	62%	57%	56%	13%	37%	7%	47%	.	15%	41%	5%	35%	16%	42%	10%	11%	79%	5%
Var17	45%	39%	6%	8%	14%	18%	10%	28%	18%	15%	14%	6%	23%	16%	21%	15%	.	19%	9%	15%	7%	19%	7%	7%	19%	76%
Var18	43%	22%	7%	15%	48%	45%	39%	88%	45%	31%	9%	2%	7%	12%	89%	41%	19%	.	12%	12%	9%	7%	8%	8%	6%	20%
Var19	7%	6%	8%	12%	6%	6%	8%	9%	6%	6%	3%	2%	3%	4%	11%	5%	9%	12%	.	10%	2%	14%	11%	2%	2%	12%
Var20	30%	23%	17%	16%	13%	54%	5%	23%	54%	43%	43%	16%	36%	5%	12%	35%	15%	12%	10%	.	14%	92%	19%	11%	60%	17%
Var21	14%	11%	4%	5%	11%	18%	6%	11%	18%	16%	71%	7%	19%	2%	10%	16%	7%	9%	2%	14%	.	16%	2%	9%	20%	4%
Var22	38%	34%	25%	23%	15%	70%	6%	31%	70%	52%	52%	16%	46%	7%	5%	42%	19%	7%	14%	92%	16%	.	29%	19%	52%	20%
Var23	8%	10%	22%	53%	5%	15%	5%	9%	15%	11%	11%	13%	8%	3%	6%	10%	7%	8%	11%	19%	2%	29%	.	5%	16%	17%
Var24	10%	8%	4%	5%	47%	14%	8%	11%	14%	14%	14%	10%	13%	3%	11%	11%	7%	11%	2%	11%	9%	19%	5%	.	18%	5%
Var25	49%	31%	18%	12%	12%	85%	8%	26%	85%	100%	100%	15%	66%	10%	7%	79%	19%	6%	2%	60%	20%	52%	16%	18%	.	4%
Var26	59%	55%	10%	16%	2%	7%	2%	8%	7%	5%	6%	1%	16%	10%	4%	5%	76%	20%	12%	17%	4%	20%	17%	5%	4%	.

FIGURE A.2 – Résultats du V de Cramer pour les variables explicatives à l'exception de l'exposition au risque, et de la localisation géographique

Variables	DDL	Valeur estimée	Standard Error	Wald Chi-Square	Pr > Khi-2
Intercept	1	-7,4015	0,3107	567,4894	<.0001
Var22*Var25	1	-1,9428	0,6668	8,489	0,0036
Var22*Var25	1	-0,6088	0,6603	0,8501	0,3565
Var22*Var25	1	-1,1139	0,5259	4,4861	0,0342
Var22*Var25	0	0	.	.	.
Var5	1	1,1913	0,3897	9,3443	0,0022
Var5	1	0,5923	0,3893	2,3147	0,1282
Var5	0	0	.	.	.

FIGURE A.3 – Tableau d'estimation des coefficients de régression du modèle final de survenance de sinistres

Variables	DDL	Valeur estimée	Wald 95% Confidence		Wald Chi-Square	Pr > Khi-2
Intercept	1	7,8928	7,8473	7,9384	115321	<.0001
Var8	1	-0,4516	-0,7004	-0,2029	12,66	0,0004
Var8	1	-0,2932	-0,3455	-0,2408	120,5	<.0001
Var8	1	-0,2392	-0,3935	-0,0848	9,23	0,0024
Var8	0	0	0	0	.	.
Var13*Var25	1	-0,6523	-0,7758	-0,5289	107,2	<.0001
Var13*Var25	1	0,091	-0,0822	0,2642	1,06	0,3032
Var13*Var25	1	-0,8032	-0,9236	-0,6829	171,09	<.0001
Var13*Var25	1	0,0975	0,0066	0,1883	4,42	0,0356
Var13*Var25	1	-1,3902	-2,3991	-0,3813	7,29	0,0069
Var13*Var25	1	0,1148	0,0443	0,1852	10,19	0,0014
Var13*Var25	1	-0,7521	-0,888	-0,6162	117,67	<.0001
Var13*Var25	0	0	0	0	.	.
Var17	1	0,3947	0,2951	0,4943	60,33	<.0001
Var17	1	0,127	0,0677	0,1864	17,6	<.0001
Var17	0	0	0	0	.	.
Var21	1	-0,1302	-0,2446	-0,0158	4,98	0,0257
Var21	1	-0,174	-0,2564	-0,0916	17,13	<.0001
Var21	0	0	0	0	.	.
Scale	1	1,4549	1,4377	1,4724		

FIGURE A.4 – Tableau d'estimation des coefficients de régression du modèle final de charge annuelle

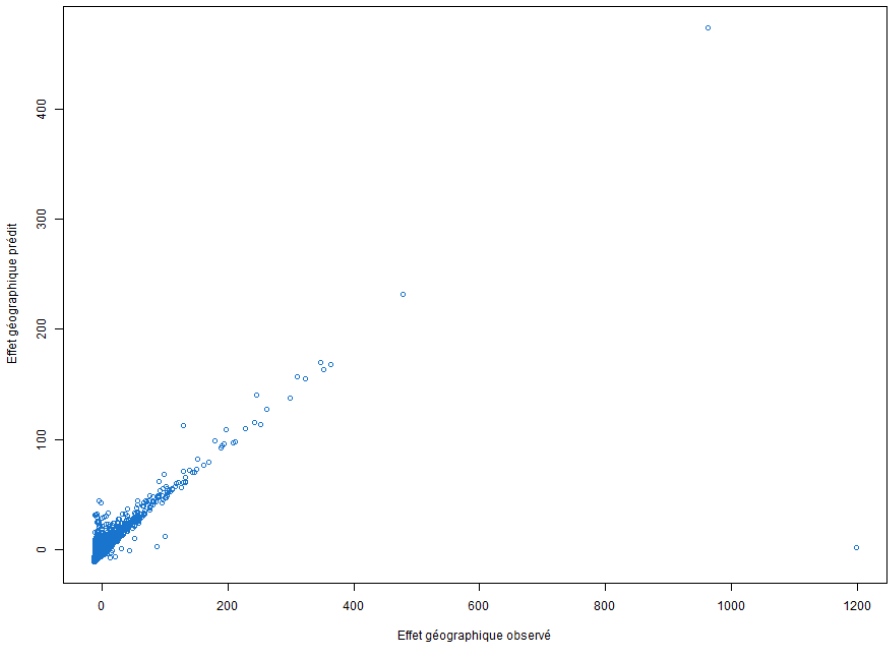


FIGURE A.5 – Nuage de points entre effet géographique observé et celui lissé par codes postaux

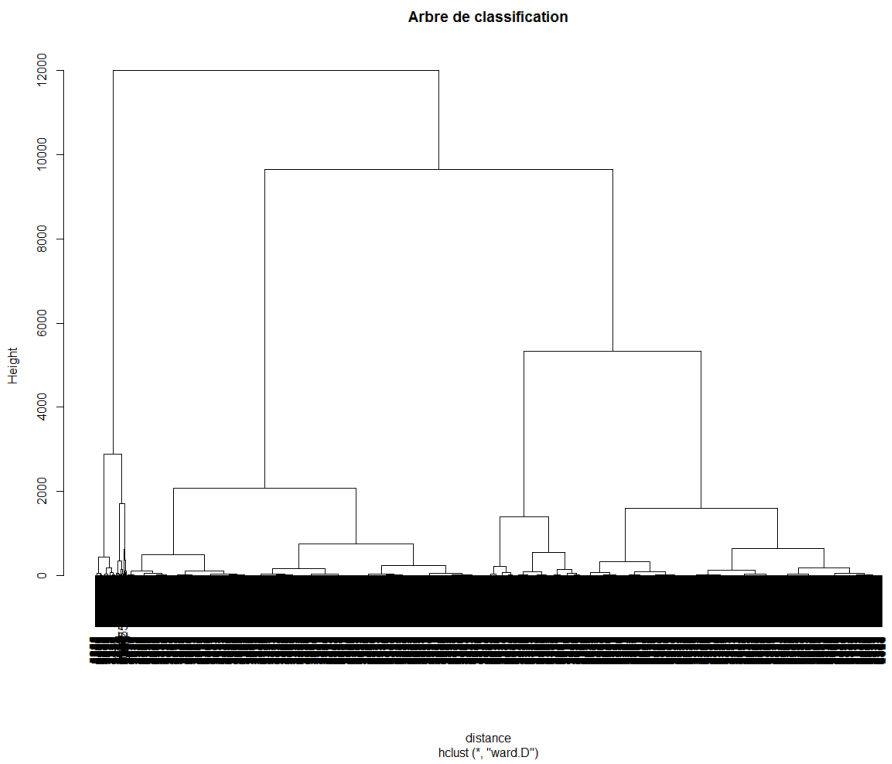


FIGURE A.6 – Dendrogramme issu de la CAH sur l'effet géographique lissé des 6048 codes postaux