Mémoire présenté le :

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

Par : **Mohamed GAHBICHE**

Titre **Estimation de la Prime Pure Catastrophes Naturelles au travers de données géographiques**

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

P. PERRET - B. BALTESAR - L. LAURENT

Membres présents du jury de l'ISFA

X. MILHAUD

Entreprise :

Nom : **AXA France**

Signature :

Directeur de mémoire en entreprise :

Nom : **Sarah BRICHET**

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

Secrétariat **Christine DRIGUZZI**

Bibliothèque : **Patricia BARTOLO**

Signature du candidat

Résumé

Mots clés : Catastrophes naturelles, Tarification, IARD, GLM, CART, Cartes, Inondations, Sécheresse, Submersion marine.

Les catastrophes naturelles sont un risque très particulier auquel les assureurs font face. Le caractère fortement aléatoire et dévastateur justifie cette appréhension face à ce risque. Afin d'analyser au mieux ce risque, plusieurs méthodes sont mises en oeuvre : cartes d'aléas, simulations informatiques sur plusieurs années, études géologiques... Dans le cadre de la multi risques habitation, le risque catastrophes naturelles représente une part non négligeable de la prime pure observée. Ceci se justifie par les coûts considérables d'un sinistre catastrophe naturelle. Dans ce mémoire, nous nous intéressons à l'analyse de trois risques de catastrophes naturelles : Inondations, Sécheresse et Submersion marine.

Des cartes issues d'études géologiques et météorologiques sont utilisées afin de déterminer les niveaux de risque. Une estimation de la prime pure catastrophes naturelles est faite en utilisant les variables internes du portefeuille habitation d'AXA France et les zones identifiées sur les cartes. La modélisation est basée sur la technique des modèles linéaires généralisés. Afin d'extraire le plus d'informations des cartes, un modèle par risque est calibré. Les trois modèles ainsi obtenus sont en suite agrégés pour donner une prédiction de la prime pure catastrophes naturelles.

La prime allouée aux catastrophes naturelles étant soumise à réglementation, le modèle obtenu dans cette étude ne peut être utilisé à des fins de tarification. Néanmoins, le modèle nous permet de calculer un indicateur d'exposition aux risques. La prime pure prédite obtenue sert à calculer un *Expected Loss Ratio* (Ratio sinistres sur primes espéré). Le suivi de cet indicateur permet de mesurer en continu l'exposition du portefeuille. Pour déterminer les profils les plus risqués, un arbre de décision basé sur l'algorithme CART (*Classification And Regression Tree*) est utilisé.

Abstract

Keywords : Natural Catastrophies, Pricing, P&C, GLM, CART, Maps, Flood, Drought, Coastal Flood.

Natural catastrophies represent a very special risk for insurers. The devastating and highly unpredictable effect justifies this apprehension. In order to analyse this risk, multiple methods are being set up : Geographical maps, Computer simulations overs multiple years, geological studies... In the housing insurance, the natural catastrophies represent a considerable part of the observed burning cost. This is justified by the fact that the average cost of a natural catastrophy claim is very high compared to other property damage. In this study, we analyse three different types of natural catastrophies : Flood, Drought and Coastal Flood.

Maps from geological and weather studies are used in order to define the risk levels. An estimation of the burning cost is done using internal pricing varaibles from the AXA France house insurance portfolio as well as identified risk areas in the external maps at our disposal. The model is based on the generalized linear models. In order to extract the most information from the maps, we consider a model for each risk. The three models are then aggregated in order to get a prediction for the burning cost for each contract.

The premuim allocated to natural catastrophies being fixed by the regulator, the model calibrated in this study cannot be used for pricing purposes. Nevertheless, the model can be used to compute an indicator of risk exposure. The estimated burning cost is used to compute an Expected Loss Ratio. Following this indicator allows us to measure continuously the portfolio exposure. In order to asses the risk profiles, a decision tree based on the CART algorithm is used.

Remerciements

Je tiens tout d'abord à remercier Mme Anne-Laure Le Gallo de m'avoir fait confiance en m'intégrant dans son équipe.

Je remercie particulièrement Mme Sarah Brichet, Mme Nguyen Tuan Doan-Trang et Mme Camille Loiret de leurs conseils et disponibilités.

Je remercie également toute l'équipe Multirisques Habitation et l'équipe de réassurance de l'entité Axa Global Property & Casualty pour leur accueil et contribution aux agréables conditions de travail.

Table des matières

Résumé **2**

Problématique **5**

1 Introduction et Contexte **6**
1.1 Les Catastrophes Naturelles . . . 7
1.2 Contexte actuel Multi Risques Habitation AXA France . . . 9
1.3 Maîtrise de la géo-localisation des contrats . . . 11
1.4 Base de modélisation . . . 14

2 Modélisation Prime pure Catastrophes Naturelles **19**
2.1 Techniques de modélisation utilisées . . . 20
2.2 Modélisation de la Prime Pure en IARD . . . 24
2.3 Modèle de Tweedie versus Poisson-Gamma séparé . . . 25
2.4 Les Cartes actuelles . . . 27
2.5 Une Distinction de risques avec les nouveaux zoniers . . . 28
2.6 Calibration du GLM pour l'Inondation . . . 36

3 Gestion des risques à la souscription **47**
3.1 Cadre légal et éthique . . . 48
3.2 Analyse de la répartition de la prime pure . . . 50
3.3 Un écart de prime impactant la marge . . . 53

4 Estimation d'un indicateur de performance **55**
4.1 Expected Loss Ratio (ELR) . . . 56
4.2 Analyse de l'ELR . . . 56
4.3 Arbre de classification et régression (CART) . . . 59
4.4 Calibration de l'Arbre des ELR . . . 65
4.5 Arbre des ELR . . . 68
4.6 Utilisation opérationnelle de l'arbre des ELR . . . 70

Conclusion **70**

Bibliographie **72**

A Annexes **75**
A.1 Lexique 75
A.2 Graphiques annexes 76
A.3 Polygones de Voronoi 76

Problématique

Avec une récurrence du risque inondation de plus en plus importante, il est nécessaire de bien estimer ce risque pour l'assurance habitation. En juin 2016 après plusieurs jours de pluie, le niveau de la Seine a commencé à augmenter jusqu'à atteindre 6.10 mètres. Cette crue de la seine a inondé plusieurs habitations et commerces. En Octobre 2015, de fortes pluies ont causé des débordements de rivières dans le Sud-Est de la France. L'urbanisation de plus en plus forte favorise les milieux inondables, par exemple la région Languedoc-Roussillon estime que ce phénomène a agrandi la surface inondable de 15% sur la région.

Dans le cas d'AXA France, les catastrophes naturelles représentent 10% [1] de la prime pure pour les maisons. Aujourd'hui, on a plus de moyens de capter et de modéliser ces risques particuliers. Une approche axée sur les reliefs, la géologie et des simulations d'évènements rares se développe au sein du groupe AXA dans le cadre des travaux de prévention et d'investissement pour la recherche.

Dans le cadre de la refonte de la gamme MRH, on souhaite maîtriser l'estimation de la prime pure catastrophes naturelles afin de mieux appréhender notre exposition au risques. Cette motivation soulève plusieurs questions quant à l'utilisation de nouvelles données géographiques : Cette donnée géographique n'étant pas forcément construite sur la base de nos sinistres, comment mesurer la validité de ces cartes et quel degré de confiance devrait-on accorder à leurs pouvoir de prédiction ? Nous avons aujourd'hui des cartes construites sur notre sinistralité passée. Sauf que dans un contexte où les catastrophes naturelles ont une très faible fréquence , ces zoniers peuvent présenter quelques instabilités. Quel est alors l'apport en stabilité d'une modélisation avec une carte construite avec des données géographiques ? Et surtout comment suivre notre exposition aux risques ?

1. Sur le périmètre des sinistres 2008 à 2014 non annulés

Chapitre 1

Introduction et Contexte

« *Le sens de l'histoire c'est le cheminement irrésistible de catastrophe en catastrophe.* »

Stéphane Hessel

Sommaire

1.1 Les Catastrophes Naturelles 7
1.1.1 Les Catastrophes Naturelles récentes en France 7
1.1.2 L'assurance MRH des Catastrophes Naturelles 8
1.2 Contexte actuel Multi Risques Habitation AXA France 9
1.2.1 Un marché concurrentiel et dynamique 9
1.2.2 Un cadre concurrentiel 9
1.2.3 Une tarification à l'adresse 10
1.3 Maîtrise de la géo-localisation des contrats 11
1.3.1 Une terre Ellipsoïdale 12
1.3.2 La projection 12
1.4 Base de modélisation 14
1.4.1 Périmètre 14
1.4.2 les bases à disposition 15
1.4.3 Structure de la base finale 15
1.4.4 Quantification du biais de géocodage 16

1.1 Les Catastrophes Naturelles

1.1.1 Les Catastrophes Naturelles récentes en France

L'été 2003, l'Europe connaît un épisode de canicule sans précédent. en France, les records de chaleurs sont battus. Cette vague de chaleur n'a pas été sans risques sur la population et les habitations. Environ 15 000 décès sont imputés à cette canicule. La catégorie de personnes la plus touchée est celle des personnes de plus de 75 ans. La déshydratation étant la cause principale de ces décès. Hormis les personnes, cette canicule a eu un effet néfaste sur un grand nombre d'habitations : plusieurs immeubles fissurés.

Plus récemment, la Tempête Xynthia frappe la France la nuit du 27 février 2010. Cet évènement a fait des dégâts matériels et humains considérables. Plusieurs maisons ont été sévèrement inondées et leurs habitants évacués par des hélicoptères de la sécurité civile. La Fédération Française des Sociétés d'Assurances - FFSA a estimé *"au moins 1,2 milliards d'euros"* le coût pour les assureurs de cette catastrophe.

FIGURE 1.1 – Maisons inondées à la suite de Xynthia

C'est l'évènement référence de submersion marine en France. Les zones touchées sont principalement la Vendée et la Charente-Maritime.

Tout dernièrement, en début juin 2016, le niveau de la Seine est montée à 6.10 mètres ce qui a occasionné quelques débordements à certains endroits. Le phénomène a beaucoup impressionné par sa rapidité et l'incapacité d'agir face à un tel évènement. Une des mesures "populaires" du niveau de la Seine a été le niveau d'inondation de la statue du Zouave sur le pont d'Alma. La moitié de cette statue de 5.2 mètres a été submergée, ce qui donnait une indication sur le niveau exceptionnel atteint.

FIGURE 1.2 – Inondation de Paris, juin 2016 [1]

Des milliers d'habitations ont été coupées d'électricité et certaines stations de métro ont été fermées. Cette crue a bien sûr occasionné des inondations d'habitations proches de la Seine.

1.1.2 L'assurance MRH des Catastrophes Naturelles

"Une catastrophe naturelle est caractérisée par l'intensité anormale d'un agent naturel (inondation, coulée de boue, tremblement de terre, avalanche, sécheresse...)" INSEE [2]

Les catastrophes naturelles de part leur nature causent des dommages matériels importants. L'habitation est un des majeurs secteurs touchés par les catastrophes naturelles. Les dommages peuvent être très divers : de la perte du mobilier à cause des inondations à la fissuration des murs à cause de la sécheresse. Il est ainsi important de pouvoir mesurer l'exposition de l'assureur face à ce risque particulier.

Le contrat multi-risques habitation garantit les dommages matériels causés par l'intensité anormale d'un agent naturel. Il peut s'agir notamment d'une inondation, d'un glissement de terrain, d'une coulée de boue, de la sécheresse ou d'un tremblement de terre. Cette garantie est enclenchée dès lors qu'une annonce constate l'état de catastrophe naturelle sur le Journal Officiel de la République Française.

Dans le cas de la survenance d'un sinistre, une franchise légale est appliquée. Cette franchise est de 380 € sauf dans le cas de mouvements de terrains dus à une sécheresse ou une ré-hydratation des sols où elle est égale à 1520 €.

En 1995, une loi a institué la mise en place de Plans de Prévention des Risques naturels. Ces plans servent à informer le citoyen des risques naturels auxquels il est soumis.

1. Source : http ://www.lemonde.fr
2. Institut National de la statistique et des Études Statistiques,http ://www.insee.fr

Ces plans sont édités par l'État et peuvent interdire les constructions sur certains sites jugées très risqués. Dans les communes non dotées de ces plans de prévention, une modulation de la franchise est appliquée. La modulation dépend du nombre de catastrophes survenues dans la commune durant les cinq dernières années. La franchise est doublée à la troisième survenance , triplée à la quatrième et quadruplée à la cinquième.

1.2 Contexte actuel Multi Risques Habitation AXA France

1.2.1 Un marché concurrentiel et dynamique

La loi du 17 Mars 2014 relative à la consommation (Loi Hamon), a bouleversé le marché de l'assurance. Cette loi donne plus de flexibilité à l'assuré. Précédemment, l'assuré avait une fenêtre de deux mois pour résilier son contrat (avant l'échéance de la tacite reconduction). Avec cette nouvelle mesure, une fois la durée d'une année de contrat passée, l'assuré peut résilier à tout moment son contrat d'assurance. Cet apport de flexibilité a beaucoup dynamisé le marché de l'assurance. En plus de cette loi, un autre phénomène vient amplifier le dynamisme du marché : la digitalisation. L'apparition des comparateurs d'assurance sur internet facilite voire oriente le choix des assurés vers des assureurs en comparant leurs garanties mais surtout leurs prix. Face à ce changement du marché de l'assurance, les modèles de tarification doivent être revus afin de suivre l'évolution du marché.

1.2.2 Un cadre concurrentiel

Avec l'arrivée sur le marché de la loi Hamon, le turnover des portefeuilles d'assurance MRH a pris de l'amplitude. En effet selon l'étude de la Fédération Française d'Assurances, le taux de résiliation a connu une hausse sensible de +0.7% alors qu'il était stable les années d'avant. En même temps le taux d'affaires nouvelles s'est apprécié de 1.4%.

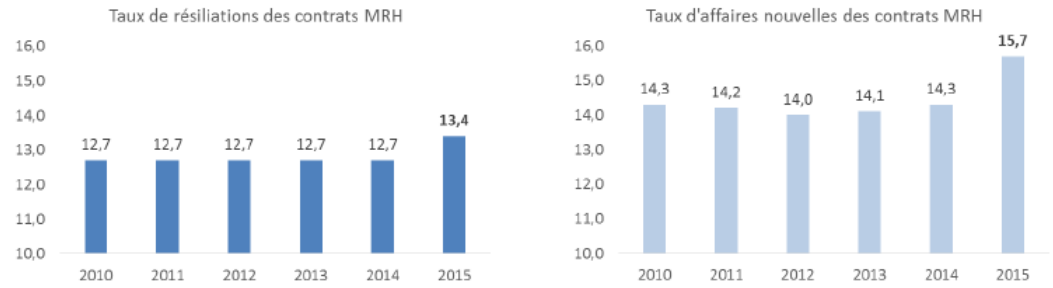

FIGURE 1.3 – Evolution du taux de résiliation MRH sur le marché français

Ce changement simultané est bien le signe d'un changement notable sur le marché de l'assurance habitation. On peut voir aussi l'effet de la loi Hamon sur les contrats habitation Axa. En effet, l'évolution de notre taux de résiliation suit le marché. Mais le motif de résiliation Hamon est la cause essentielle de ce changement.

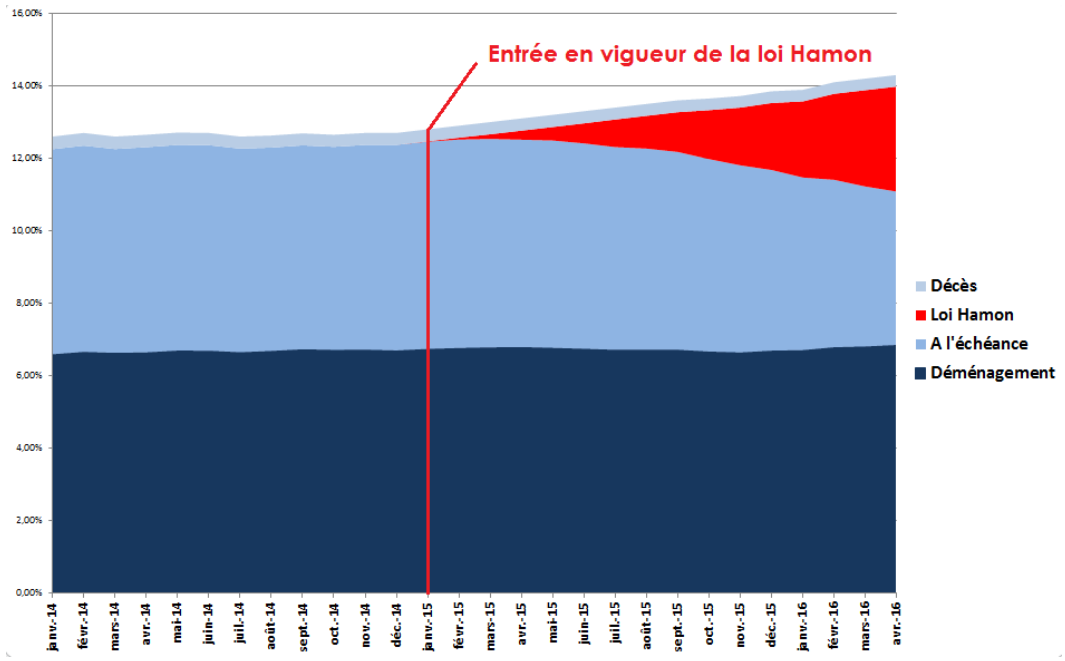

FIGURE 1.4 – Taux de résiliation en fonction du motif

Pour des raisons de confidentialité les données on été légèrement modifiées. Cependant la tendance globale est conservée. Le portefeuille Axa a aussi une tendance comparable à celle du marché au niveau du taux de résiliation. Ce graphique montre bien l'impact direct sur les résiliations qu'a engendré la loi Hamon. Il est a pondérer en effet par le fait que les résiliations Hamon prennent une partie des résiliations à l'échéance. Cependant, le saut sur les taux de résiliation traduit très bien le changement du marché avec l'entrée en vigueur de cette loi.

Dans cet environnement devenu donc plus compétitif et liquide, la segmentation du tarif devient donc primordiale pour pouvoir suivre la tendance du marché et surtout se prémunir de l'anti-sélection.Les consommateurs prennent conscience de leur niveau de primes avec les comparateurs sur internet. Cela nécessite donc de créer un tarif : bien segmenté, ajusté au niveau tarifaire des concurrents et surtout transparent pour le consommateur. La modélisation de notre prime pure Catastrophes Naturelles, nous permet d'estimer si le 12% de prime collectée nous permet de couvrir totalement notre besoin.

1.2.3 Une tarification à l'adresse

L'objectif que s'est fixé Axa France pour la revue des modèles est de segmenter le prix le plus finement possible.Cela passe par la création de modèles plus transparents pour le consommateur : possibilité de modules de garanties ainsi que plusieurs options claires. Derrière ce prix, une géolocalisation est faite afin d'estimer le niveau de risques de chacune des garanties. Pour chaque garantie nous avons un zonier à la maille du quartier. Avant on avait donc un coefficient par zone INSEE, alors que maintenant on a un coefficient par zone contrat. Bien sur la construction des zoniers a nécessité des techniques de

lissages (plus proches voisins) pour avoir une cohérence des zoniers Voronoi (Définition en Annexe A.3).

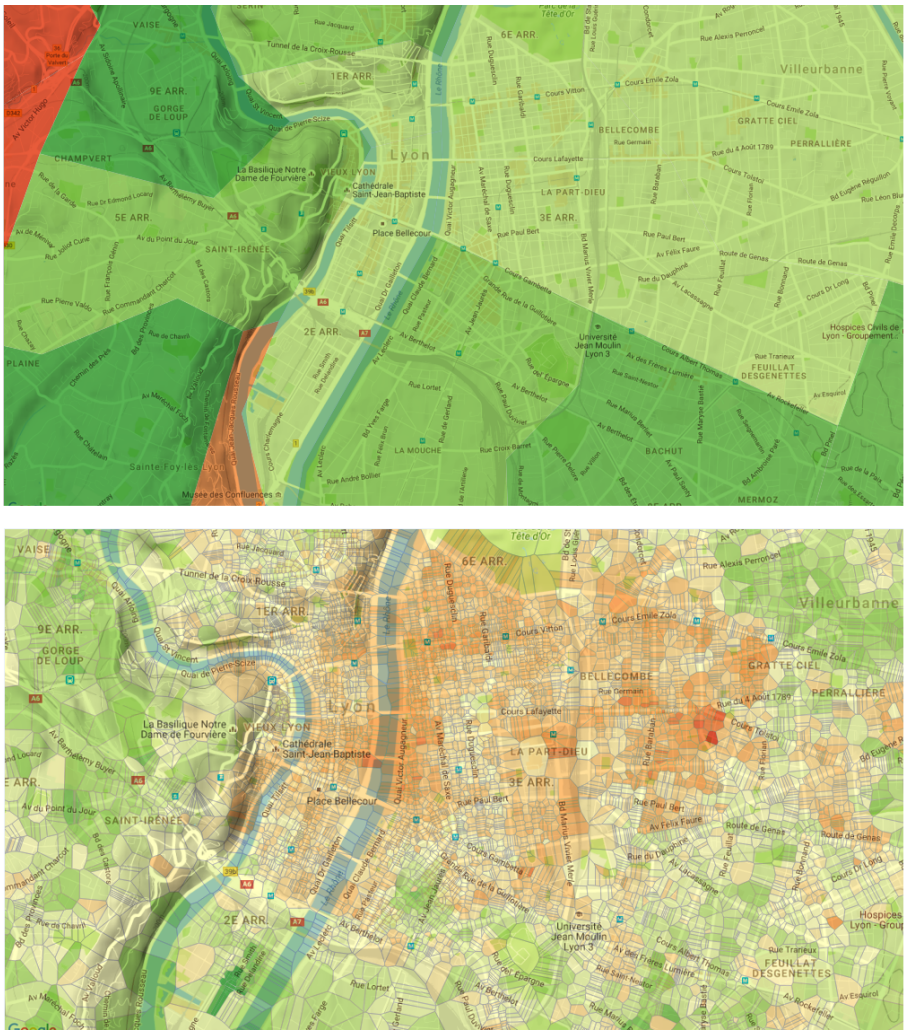

FIGURE 1.5 – L'apport d'un maillage très fin- Garantie Dégâts des eaux

Ici nous avons la représentation de deux zoniers sur Lyon. Le premier zonier est un zonier fait à la maille INSEE. Le deuxième est un maillage Voronoi. On peut déjà remarquer avec un degré de précision aussi fin nous pouvons avoir une segmentation plus fine de notre tarif. Passer à une maille aussi fine est l'un des challenges actuel de l'équipe actuarielle d'Axa.

1.3 Maîtrise de la géo-localisation des contrats

La figure précédente montre la précision à laquelle on peut segmenter notre modèle. Cependant elle soulève surtout une problématique de géolocalisation. Afin de localiser une adresse sur terre, un système de coordonnées est nécessaire. Ce système attribue des coordonnées à chaque point dans un espace tridimensionnel. Plusieurs référentiels sont possibles pour cette opération. Les coordonnées sont généralement exprimées en Degrés , en mètres , ...

1.3.1 Une terre Ellipsoïdale

Un système de coordonnées largement utilisé est le système de coordonnées géographiques. La terre ici est modélisée par une ellipsoïde.

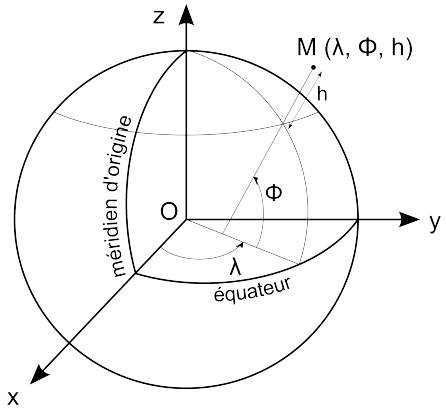

FIGURE 1.6 – Coordonnées géographiques [3]

Dans ce référentiel on a :

- λ : La longitude
- Φ : La latitude
- h : La hauteur ellipsoïdale

La longitude et la latitude sont exprimées en unité de mesures d'angle (Degrés, radian,...)

La hauteur correspond à la hauteur du point par rapport à la normale à l'ellipsoïde. Cette grandeur diffère généralement de l'altitude normale (mesurée par rapport à la hauteur au géoïde)

Les systèmes de coordonnées sont recensés par l'European Petroleum Survey Group, EPSG. On peut citer quelques uns :

- **WGS84** : un des systèmes les plus courants. il a été établi en 1984 (révisé en 2004) par le Département de la Défense des États Unis. Il est utilisé notamment pas les GPS [4].
- **RGF93** : Réseau Géodésique Français a été mis en place dès 1993.

1.3.2 La projection

La projection des coordonnées tridimensionnelles sur une surface plane permet de faciliter plusieurs calculs. Sur une surface plane les coordonnées sont exprimées en unités de métriques. La projection nous permet donc de facilement calculer les distances et

3. Source : http ://www.ades.cnrs.fr
4. Global Positioning System : Géo-positionnement par Satellite

les surfaces comparé à un système ellipsoïdal. Les projections ont toute fois une perte d'information.

En comparaison, on peut avoir des projections qui conservent la surface, ou les distances ou les angles. il est à noter qu'on ne peut pas avoir une projection conservant toutes les propriétés.

Il existe plusieurs façons de projeter les coordonnées géographiques en surface plane : Projections Coniques, Cylindriques, Azimutales...

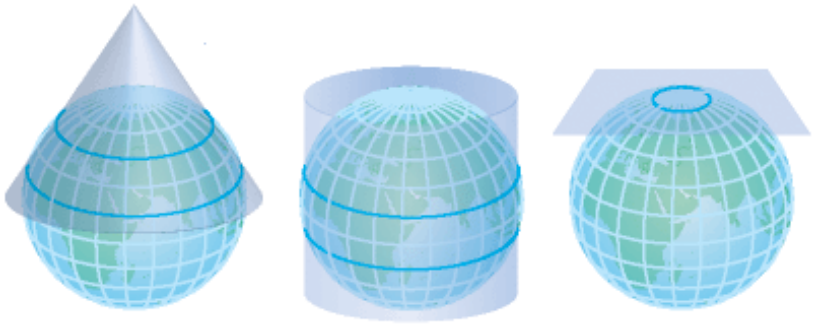

FIGURE 1.7 – Principaux types de projection

Nous baserons notre étude sur la projection Conique Conforme Lambert 93. Cette projection réduit les déformations pour la carte de la France.

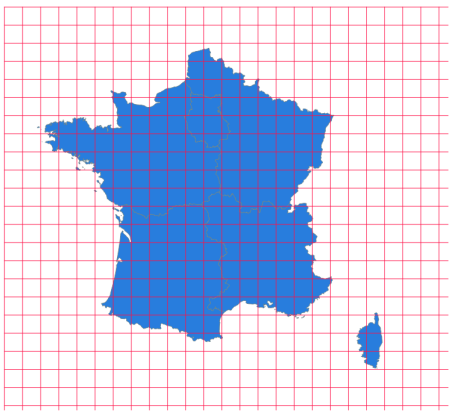

FIGURE 1.8 – Carte de France en RGF 93

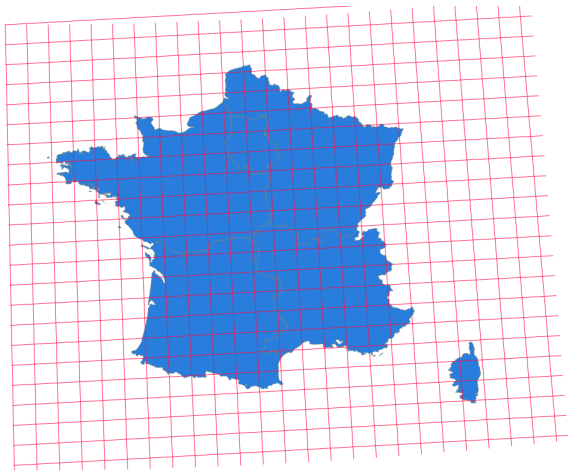

FIGURE 1.9 – Carte de France en projection Azimutale Nord

Ci-dessus, deux projections de la France. La première est celle à laquelle on est le plus habitués. Il s'agit de la projection officielle de la France. la deuxième est une projection Azimutale Pôle Nord. Nous avons rajouté un quadrillage à la carte pour mettre en avant la difficulté qu'on peut rencontrer si on veut changer de projection tout en restant sous un même quadrillage (c'est le cas si le format de la carte est un Raster [5]).

5. Raster : carte sous forme de grille. L'information est stockée sous forme de matrice

1.4 Base de modélisation

1.4.1 Périmètre

Le travail de modélisation passe d'abord par une étape incontournable : la construction de la base de données. Il est impératif de s'assurer de la validité des informations contenues et aussi de les nettoyer. Dans notre périmètre de modélisation, nous avons décidé de se fixer aux contrats appartenant aux formules suivantes :

— **N5AA :** la formule la plus répandue. Elle est connue sous le nom "Confort". Cette formule couvre toutes les garanties classiques de la MRH.

— **N5AB :** la formule "Confort étendu". Cette formule est une extension de la formule "Confort". Les garanties sont plus larges.

— **N5A2 :** la formule qui garantit exclusivement les habitations secondaires.

— **N5AM :** la formule pour les locataires d'appartements. Cette formule est moins onéreuse que la formule confort et par conséquent présente aussi moins de garanties.

— **N5AK :** la formule pour les propriétaires d'appartements.

Sur ce périmètre, les contrats ont presque les mêmes garanties. Vis à vis de la garantie Catastrophes Naturelles, tous les contrats y sont exposés de manière équivalente. On peut donc justifier l'utilisation de ce périmètre. De plus, l'information sur les variables clés y est renseignée. Les variables primordiales qu'on cherche à récupérer sont :

— **L'année :**il s'agit ici de récupérer l'année de vision du contrat.

— **Le numéro de contrat :** C'est la clé unique attribuée à chaque contrat. c'est une clé de jointure commune à toutes les bases de stockage.

— **La qualité de l'occupant :** c'est l'information sur le fait d'être locataire ou propriétaire. Cette différence permet de déceler un comportement différent.

— **Le type d'habitation :** Elle renseigne l'information sur le fait que le risque soit une maison, un appartement en étage ou en rez-de-chaussée.

— **L'inhabitation :** cette variable recense le fait que le risque soit une habitation principale ou secondaire

— **Le capital :** variable continue qui représente le capital mobilier déclaré par l'assuré.

— **Le taux d'objets de valeur :** représente la part en pourcent des objets de valeurs du capital.

— **La géolocalisation du risque :** les coordonnées (X, Y) en Lambert 93 du risque.

— **Le nombre de sinistres :** pour chacune des garanties on récupère le nombre de sinistres par année.

— **La charge des sinistres :** on récupère la charge totale pour les sinistres par garantie

1.4.2 les bases à disposition

Afin de construire notre base de modélisation, nous avons à notre disposition les bases suivantes :

— **Bases ACN :** Ces bases recensent tous les contrats MRH avec les variables descriptives du risque (habitation,capital,qualité,...). Ces bases couvrent le périmètre à partir de l'année 2003

— **Bases sinistres :** Ces bases recensent les sinistres du périmètre IARD d'Axa France. Ces bases couvrent le périmètre à partir de 2003.

— **Bases géocodage :** Ces bases recensent les coordonnées (X, Y) de chacun des risques. Une note est attribuée quant à la qualité du géocodage. Le géocodage est renseigné pour les contrats à partir de 2009.

1.4.3 Structure de la base finale

Afin d'avoir la plus grande exposition, nous décidons de faire une base de modélisation sur 13 années : de 2003 à 2015. Nous avons cependant une contrainte quant au géocodage. En effet, nous n'avons pas les coordonnées des contrats antérieurs à 2009. Pour compléter notre base, on a fait l'hypothèse que les contrats de 2003 à 2008 qui sont encore en portefeuille en 2009 n'ont pas changé d'adresse. Ainsi, on peut leur attribuer des coordonnées. Cette hypothèse va bien sur induire un certain biais dans la base dû au fait que certains ont effectivement changé d'adresse mais qui sont restés en portefeuille.

Pour le géocodage, nous ne gardons que les contrats dont l'adresse renseignée a permis de très bien les géolocaliser.

Nous construisons les bases années par années. On extrait d'abord les contrats des bases ACN pour une année. Ensuite on joint à cette base l'information du géocodage et enfin on joint les sinistres. On obtient ainsi une base complète pour chacune des années. On joint alors ces bases les unes au autres pour obtenir notre base finale de modélisation. La structure finale est de cette allure :

Année	Numéro	Habitation	Qualité	...	X	Y	Nb Inond	...	Ch Séch
2003	00034325	Maison	Locataire	...	153099.43	6152964.66	0	...	1200
2003	00034665	Appartement	Propriétaire	...	154654.63	6147924.74	0	...	0
...	...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...	...
2004	00034325	Maison	Locataire	...	153099.43	6152964.66	0	...	0
2004	00034665	Appartement	Propriétaire	...	154654.63	6147924.74	0	...	0
...	...	...	...	...	...	...	...	...	...
2015	00034665	Appartement	Propriétaire	...	154654.63	6147924.74	0	...	0

TABLEAU 1.1 – Table de modélisation finale

1.4.4 Quantification du biais de géocodage

Le géocodage de nos contrats a commencé en 2009. Nous avons donc pour chaque contrat les coordonnées de position. Le référentiel choisi pour ces coordonnées est le RGF 93. Ce référentiel est d'ailleurs le référentiel officiel de la France. En plus des coordonnées nous avons une note attribuée à la qualité du géocodage. Cette note est entre 0 et 4 :

— 4 : Adresse exacte

— 3 : Numéro de rue approché

— 2 : Centroïde de la voie

— 1 : Centroïde ville

— 0 : erreur

Dans notre base de modélisation, nous ne gardons que les contrats dont on a une précision de géocodage égale à 4. Nous prenons cette précision car c'est celle qui apportera le moins de biais. Cependant, se restreindre à ce périmètre a un coût sur le nombre de contrats. En effet, en nous restreignant à ce périmètre nous réduisons considérablement notre base de modélisation.

Comme on l'a précisé sur la construction de la base nous avons le géocodage pour les adresses à partir de 2009. Les adresses de 2003 à 2008 ont eu les coordonnées de 2009 s'ils sont restés.

	Nombre de contrats	**Pourcentage restant**
2009-2015	22 812 142	55%
2003-2009	15 147 606	39%
Total	37 959 748	48%

TABLEAU 1.2 – Pertes liées au géocodage

La contrainte qu'on impose nous fait perdre un peu plus de la moitié de notre base de données. On regarde aussi la perte par département pour juger de l'uniformité de la perte sur le territoire. Nous représentons la perte de contrats par département.

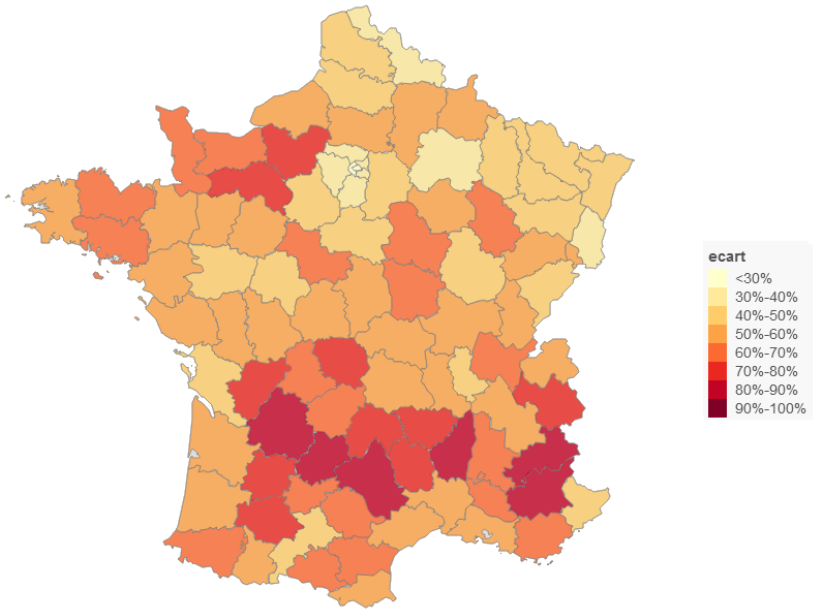

FIGURE 1.10 – Perte par département par la contrainte de géocodage

La perte en précision est clairement non uniforme sur la France. Nous avons une meilleure qualité de géocodage pour l'île de France et la région Nord-Est. Par contre pour certains départements du Sud, la perte de données est considérable. La contrainte qu'on a utilisée sur la qualité de géocodage introduit un biais géographique sur notre base de données. Ce biais peut se traduire potentiellement par un biais sur la fréquence ou le coût moyen. En effet, si les départements mal géocodés sont particulièrement touchés par les catastrophes naturelles nous aurons une sinistralité observée moindre sur la base. Nous vérifions donc la fréquence ainsi que le coût moyen sur les deux bases.

	Sécheresse		**Inondation**		**Submersion Marine**	
	Fréquence	*Coût moyen*	*Fréquence*	*Coût moyen*	*Fréquence*	*Coût moyen*
Base totale	0.41‰	15 600	0.69‰	7 850	0.45‰	3 660
Base Contrainte	0.29‰	18 350	0.54‰	7 100	0.45‰	3 660

TABLEAU 1.3 – Effet de la contrainte spatiale sur les sinistres

La restriction qu'on a fait sur notre base a modifié la structure de notre base. Nous avons une modification de la fréquence et du coût moyen. Cette modification introduit bien un biais sur nos données. La submersion marine n'est pas touchée par la contrainte géographique. Cela s'explique par le fait que la France n'a connu qu'un seul évènement de submersion marine conséquent durant la période considérée : Xynthia en 2010. De plus, la perte de géocodage sur la zone touchée par Xynthia est presque uniforme (Côte Ouest).

Il nous est donc impossible de travailler avec un biais aussi fort. Nous proposons donc de travailler avec la base totale en attribuant à chacune des adresses une zone de risque approchée. Les bases de données contiennent une information moisn précise mais fiable

sur la localisation du risque. Cette variable est le code INSEE. Cette information reste grossière par rapport à la finesse de la géolocalisation à l'adresse mais elle a le mérite d'être renseignée.

Nous allons donc créer pour toutes les cartes sur lesquelles on va travailler, une version à l'INSEE. On affecte à chaque code INSEE la zone moyenne. La zone moyenne est pondérée par la surface de chaque zone à l'intérieur de l'INSEE.

Chapitre 2

Modélisation Prime pure Catastrophes Naturelles

« *Tous les modèles sont faux, mais certains sont utiles* »

George Box

Sommaire

2.1 Techniques de modélisation utilisées 20
2.1.1 Les modèles linéaires généralisés (GLM) 20
2.1.2 L'indice de Gini : un critère de performance des modèles 23
2.2 Modélisation de la Prime Pure en IARD 24
2.3 Modèle de Tweedie versus Poisson-Gamma séparé 25
2.3.1 La famille des distributions Tweedie 25
2.3.2 L'approche Poisson-Gamma séparés 26
2.3.3 L'approche Tweedie 27
2.4 Les Cartes actuelles 27
2.5 Une Distinction de risques avec les nouveaux zoniers 28
2.5.1 Carte inondation 29
2.5.2 Carte de Sécheresse 34
2.5.3 Carte de submersion marine 35
2.6 Calibration du GLM pour l'Inondation 36
2.6.1 Analyse préliminaire 37
2.6.2 Calibration du GLM 42

2.1 Techniques de modélisation utilisées

2.1.1 Les modèles linéaires généralisés (GLM)

Le modèle linéaire généralisé est la technique qui permet d'avoir un large choix pour la distribution de la variable à expliquer. Les GLM ont pris une grande importance dans le domaine de la tarification IARD. En effet, la plupart des modèles de tarification aujourd'hui sont basés sur une estimation par GLM. Comme on le verra par la suite, ces modèles offrent des estimations faciles à mettre en production notamment par un calcul multiplicatif.

Le but du GLM est d'expliquer la variable Y par les variables $X_1...X_p$. Dans notre cas la variable Y représente le coût ou la fréquence du sinistre. Les $X_1...X_p$ représentent les variables internes explicatives telles que (le nombre de pièces, le capital, l'âge de l'assuré , ...) ainsi que les variables géographiques (Zone sécheresse, Zone submersion marine,...).

Techniquement un GLM est défini par trois composantes :

— **La Distribution :** Choix de la loi modélisant les observations Y_i

— **Les prédicteurs :** Les variables explicatives du risque $X_1...X_p$

— **La fonction de lien :** permet de lier les deux premières composantes.

La Distribution :

Dans le cadre des GLM, la distribution à choisir pour la variable à prédire fait partie de la famille des lois exponentielles. On suppose que l'échantillon de $Y_i, i = 1...n$ indépendantes et tirées de loi dont la distribution suit la forme suivante :

$$f(y,\theta,\phi) = exp(\frac{y\theta - \nu(\theta)}{u(\phi)} + w(y,\phi))$$

avec $a(.)$, $b(.)$ et $c(.)$ sont des fonctions , et θ appelé le paramètre naturel. ϕ est supposé connu.

On peut citer des lois usuelles qui sont majoritairement utilisées et qui font partie de la famille des lois exponentielles (ex : Normale, Poisson , Gamma , Inverse gaussienne, Binomiale).

Les prédicteurs linéaires

Les observations liées aux $Y_i, i = 1...n$ sont les composantes déterministes du modèle et représentées par les $(X_{i,j})_{\substack{i=1,...,n \\ j=1,...,p}}$. Cela se traduit dans notre cas par les variables liées aux contrats (Nombre de pièces, Capital déclaré, Age,...).

On définit le prédicteur linéaire par :

$$\eta = X\beta$$

avec β les paramètres à estimer du GLM.

La fonction lien

La fonction de lien est celle qui lie Y à η. Notons cette fonction g.

$$g(Y) = \eta$$

Dans le cadre de l'assurance IARD. La fonction de lien logarithme est largement utilisée. La raison pour ce choix vient du fait que le modèle induit est multiplicatif. Ce type de modèle est facile à mettre en production. Il permet aussi d'avoir la relativité entre les classes d'une variable facilement par exemple si les appartements ont un coefficient à 1 et les maisons un coefficient à 1,3. Cela signifierais que les maisons sont 30% plus "risquées".

$$\begin{aligned} ln(E(Y_i)) &= X\beta \\ E(Y_i) &= exp(X\beta) \\ &= exp(\beta_0)\, exp(x_1^i\beta_1)\, exp(x_2^i\beta_2)... exp(x_p^i\beta_p) \\ &= \gamma_0\, \gamma_{x_1^i}\, \gamma_{x_2^i} \cdots \gamma_{x_p^i} \end{aligned}$$

Le modèle

Sachant une certaines réalisation $(Y_i, x_1, ..., x_p)$, le modèle linéaire généralisé estime Y_i selon la formule suivante :
On note μ_i l'espérance de Y_i, $x_1, ..x_p$ les i^{eme} observations de $X_1 ... X_p$ en suit alors

On note μ_i l'espérance de Y_i, $x_1, ..x_p$ les i^{eme} observations de $X_1 ... X_p$ en suit alors

$$g(\mu_i) = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p$$

Avec :
$g(.)$ une fonction monotone et dérivable appelée la fonction lien
β_i : les estimateurs du MLG

La fonction $g(.)$ peut être choisie de façon à répondre à un certain critère. On note $g(.)$ fonction de lien canonique si $g(\mu_i) = \theta_i$. On peut donner les fonctions de lien canoniques correspondant aux lois usuelles de famille exponentielle.

Estimation des paramètres β_j

Notons $l_i(\theta_i, \Phi; y_i) = ln(f(y_i; \theta_i, \Phi))$ la log-vraisemblance de la i^eme observation. on a alors :

$$\frac{\partial l}{\partial \theta_i} = \frac{y_i - v'(\theta_i)}{u(\Phi)}$$
$$\frac{\partial^2 l}{\partial \theta_i^2} = \frac{-v''(\theta_i)}{u(\Phi)}$$

Pour une loi appartenant à la famille des exponentielles on a le résultat suivant :

$$\mathrm{E}\left(\frac{\partial l}{\partial \theta_i}\right) = 0 \ \ et \ \ \mathrm{E}\left(\frac{\partial^2 l}{\partial \theta_i^2}\right) = 0$$

alors en utilisant les deux équations, nous obtenons :

$$\mathrm{E}(Y_i) = v'(\theta_i)$$
$$Var(Y_i) = -v''(\theta_i)\, u(\Phi)$$

On s'intéresse maintenant à la log-vraisemblance de l'échantillon de calibration. Cette quantité s'écrit :

$$\mathscr{L}(\beta) = \sum_{i=1}^{n} ln(f(y_i; \theta_i, \Phi))$$
$$= \sum_{i=1}^{n} l(\theta_i, \Phi; y_i)$$

Afin d'estimer le maximum de vraisemblance on résout l'équation $\frac{\partial \mathrm{L}(\beta)}{\partial \beta_j} = 0$ pour tout $j = 1...p$
On regarde alors pour chacun des l_i la dérivée par rapport à β_j. Afin de mener les calculs on décompose la dérivée selon des quantités qu'on peut exprimer :

$$\frac{\partial l_i}{\partial \beta_j} = \frac{\partial l_i}{\partial \theta_i} * \frac{\partial \theta_i}{\partial \mu_i} * \frac{\partial \mu_i}{\partial \eta_i} * \frac{\partial \eta_i}{\partial \beta_j}$$

Chacune de ses dérivées peut etre exprimée de la façon suivante :

$$\frac{\partial l_i}{\partial \theta_i} = \frac{y_i - \mu_i}{u(\Phi)}$$
$$\frac{\partial \theta_i}{\partial \mu_i} = \frac{Var(y_i)}{u(\Phi)}$$
$$\frac{\partial \eta_i}{\partial \beta_j} = x_{i,j}$$
$$\frac{\partial \mu_i}{\partial \eta_i} \text{ donné par la fonction de lien g(.)}$$

Il en suit donc que les équations à résoudre se réécrivent :

$$\sum_{i=1}^{n} \frac{y_i - \mu_i}{Var(y_i)} \frac{\partial \mu_i}{\partial \eta_i} = 0 \text{ pour } j = 0, ..., p$$

Ces équations sont généralement résolues à l'aide des techniques de résolution numériques comme Newton-Raphson.

Ajustement du modèle

La déviance est un critère utilisé pour mesurer la qualité d'ajustement des modèles aux données. Cette mesure représente la différence entre la log-vraisemblance du modèle et celle du modèle saturé. Le modèle saturé est défini comme étant le modèle qui estime exactement les données.

$$\mathscr{D} = -2(\mathscr{L} - \mathscr{L}_{\text{sat}})$$

La déviance suit asymptotiquement une loi de Chi-2 avec $n - p$ degrés de libertés.

Les résidus du modèle

Pour les modèles linéaires (fonction de lien identité), il est naturel de considérer comme résidus du modèle les quantités :

$$r_i = y_i - \widehat{y}_i$$

Or dans le cas des GLM cette définition est inappropriée. Les principaux résidus utilisés pour les GLM sont :

— Résidus de Pearson : $r_i = \frac{y_i - \widehat{y}_i}{s_i}$ avec s_i l'écart type de $\widehat{y}_i$.

— Résidus de Déviance : $r_i = (y_i - \widehat{y}_i)\sqrt{d_i}$.

2.1.2 L'indice de Gini : un critère de performance des modèles

L'indice de Gini procure une mesure de la qualité de segmentation du modèle. Cet indice couplé avec la courbe de Lorenz est aussi utile lorsque l'on a plusieurs modèles à la suite d'addition de variables explicatives ou croisements de variables par exemple. La combinaison des deux outils permet de choisir le plus segmentant des modèles et aussi peut donner une interprétation de la répartition des primes selon le degré de risque (un exemple sera donné lors de la modélisation de la prime pure inondation).

L'indice de Gini est calculé à l'aide de la courbe de gain (appelée aussi courbe de Lorenz). La courbe de gain a été développée par Lorenz en 1905 dans son papier *Methods of Measuring the Concentration of Wealth.*

Afin de tracer la courbe de Lorenz, on définit :

— n : le nombre de lignes de contrats

— E_i : l'exposition de la i^{eme} police

— S_i : la charge de la i^{eme} police

— PP_i : La prime pure prédite par le modèle pour la i^{eme} police

— $Z_i = PP_i\, E_i$: L'estimateur de S_i

On trie ensuite les $((Z_i)_{i=1...n})$ pour avoir la courbe de Lorenz des prédictions du modèle :

— $Z_{i,n}$: la i^{eme} statistique d'ordre des Z_i

— E'_i : l'exposition associée à $Z_{i,n}$

— S'_i : la charge associée à $Z_{i,n}$

La courbe de Lorenz des observations est obtenue en reliant les points de coordonnées :

$$\left(\frac{\sum_{j=1}^{n} E'_{i,n}}{E}, \frac{\sum_{j=1}^{n} S'_{i,n}}{S} \right)$$

En assurance IARD, la plupart de nos contrats sont non sinistrés ce qui donne un indice de Gini sur les observations très proche de 1.

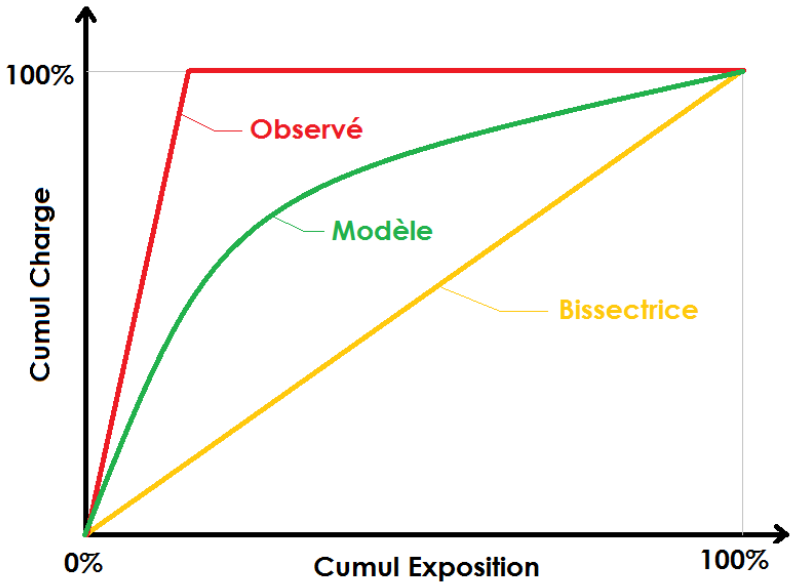

FIGURE 2.1 – représentation des courbes de Lorenz pour l'observé et un modèle

2.2 Modélisation de la Prime Pure en IARD

Dans la tarification d'assurance IARD, certaines quantités jouent un rôle particulier. La première définition à apporter est l'année police. L'année police représente la durée sur laquelle un contrat est exposé au risque. par exemple un contrat qui arrive début avril de l'année N et reste jusqu'à la fin de l'année, aura une exposition égale à 0.6667 (8 mois de présence / 12 mois de l'année=0.6667 exposition). Cette quantité fera une pondération

des observations lors de l'estimation par GLM. Trois autres quantités forment la base de l'assurance.

— La *fréquence* : elle représente le nombre de sinistres ramené à l'exposition.

— Le *coût moyen* : Il s'agit de la moyenne des charges observées

— La *prime pure* : c'est le coût espéré des sinistres par police. Elle est obtenue en multipliant la fréquence par le coût moyen

Les GLM sont capables d'estimer ces trois quantités. Par pratique, on n'utilise que des variables quantitatives pour la calibration des GLM. Les variables continues (comme par exemple le capital, l'âge, le montant d'objets de valeurs,...) sont discrétisées afin de les rendre quantitatives. L'assurance repose principalement sur la loi des grands nombres. Pour cette raison, il est nécessaire d'avoir une certaine représentativité de chacune des classes des variables explicatives.

Dans notre cas, on va considérer une prime pure modélisée par un modèle Poisson-Gamma. Notons le coût d'un sinistre est représenté par Y_i et N représente le nombre de sinistres total. On fait l'hypothèse que Y_i suit une loi Gamma et N suit une loi de Poisson. La somme totale des sinistres est donnée par :

$$S = \sum_{i=1}^{N} Y_i$$

Dans ce modèle, on suppose que le coût et la fréquence sont indépendants. Cette hypothèse nous permet d'avoir l'égalité suivante :

$$\mathbb{E}(S) = \mathbb{E}(N)\ \mathbb{E}(Y_1)$$

Dans les cas classiques de l'assurance, on modélise chacune des deux espérances à part Ensuite la prime pure est obtenue en multipliant les deux espérances. On modélise alors la fréquence avec un modèle GLM avec une loi de Poisson pour la variable à expliquer. Le coût est quant à lui supposé suivre une loi Gamma. L'approche d'un modèle avec une loi Tweedie est différente. On cherche ici à estimer directement la prime pure. On modélise ainsi la loi de S directement. La loi de Tweedie peut être vue comme un modèle Poisson-Gamma composé pour certaines valeurs de p (paramètre défini dans la section suivante).

2.3 Modèle de Tweedie versus Poisson-Gamma séparé

2.3.1 La famille des distributions Tweedie

La famille de Tweedie est une sous famille des lois exponentielles. la loi de Tweedie n'a pas de formule fermée pour la densité. Elle est cependant définie grace à la fonction de variance. Une loi appartient à la famille de Tweedie si elle est de la famille des lois exponentielles et si sa fonction de variance vérifie la condition suivante sur $]0,\infty)$:

$$\mathbb{V}(\mu) = \mu^p \text{ pour } p \in \mathbb{R}$$

La famille de Tweedie contient plusieurs lois dont certaines usuelles selon la valeur de p. On donne dans le tableau suivant les différentes lois qu'on peut retrouver avec le paramètre p.

Valeur de p	Loi
$p = 0$	Normale
$p = 1$	Poisson
$p \in]1,2[$	Poisson-Gamma Composé
$p = 2$	Gamma
$p = 3$	Normale Inverse

TABLEAU 2.1 – Les lois usuelles retrouvées dans la famille Tweedie

Pour notre cas de modélisation, nous prendrons donc le paramètre p entre 1 et 2 pour avoir une modélisation en Poisson-Gamma composé. Dans ce cas la loi Tweedie modélise :

$$S = \sum_{i=0}^{N} Y_i \text{ Avec } Y_0 = 0$$

La Tweedie a une densité continue sur $]0,\infty)$ et une valeur positive en 0. L'allure de la densité de Tweedie est la suivante.

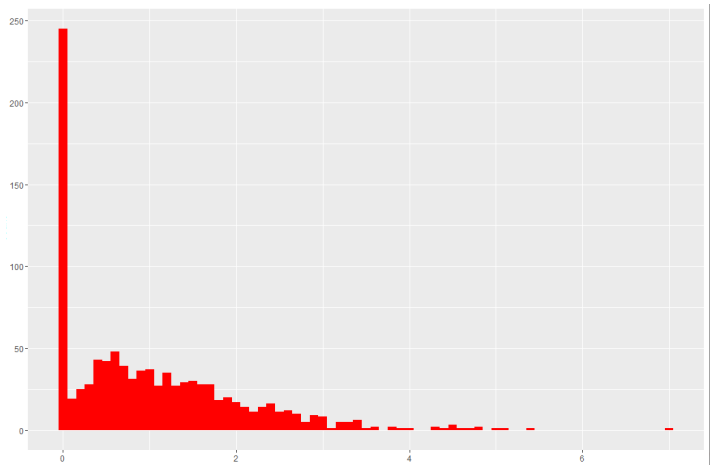

FIGURE 2.2 – Densité simulée d'une Tweedie avec $p = 1.3, \mu = 1, \Phi = 1$

Cette allure est proche de celles des charges de sinistres qu'on peut avoir. En effet, une large partie de nos contrats sont non sinistrés ainsi ils ont une charge nulle. les contrats sinistrés se retrouvent avec une charge continue et positive.

2.3.2 L'approche Poisson-Gamma séparés

Dans cette approche, on calibre deux modèles séparément. Dans un premier temps, on modélise la fréquence par une loi de Poisson. Dans le GLM, la variable à prédire est donc

le nombre de sinistres. le modèle va donc estimer une fréquence de sinistre par profil. On choisit les variables qui expliquent le mieux le risque. Une fois le modèle validé, on passe à la modélisation de la charge. Pour modéliser cette charge, la base de données pour le GLM ne doit contenir que les contrats ayant eu de sinistres. De cette façon, la charge n'a pas de poids en zéro. Dans la pratique, on commence par le modèle de fréquence car celui ci est plus stable en général.

On suppose dans notre cas qu'on a m variables prédictives. On note m_{freq} le nombre de variables retenues pour le modèle de fréquence et m_{cout} celles retenues pour le modèle de coût. Le modèle de fréquence doit donc estimer $m_{freq}+1$ paramètres (le 1 en plus est pour le coefficient de base). Donc on aura pour le modèle de prime pure par Poisson-Gamma séparé : $m_{freq}+m_{cout}+2$ coefficients à estimer.

2.3.3 L'approche Tweedie

Dans la modélisation Tweedie, le GLM cherche à expliquer la charge observée directement. Dans la base sur laquelle le GLM se calibre, les charges nulles sont présentes. La variable à expliquer est la charge qu'elle soit strictement positive (cas de la modélisation du coût par une loi Gamma) ou nulle.

L'avantage de cette approche est le fait d'avoir moins de paramètres à estimer. En effet, dans ce cas nous avons le paramètre p, le coefficient de base et les estimateurs des variables explicatives. Donc si nous utilisons les variables utilisées pour la modélisation Poisson-Gamma nous aurons moins de paramètres à estimer. En effets, supposons qu'on aient k variables communes aux modèles Poisson et Gamma alors le modèle Tweedie aura $m_{freq}+m_{cout}-k+2$ coefficients à estimer.

2.4 Les Cartes actuelles

Dans la trame de construction des zoniers de la nouvelle gamme, une carte catastrophes naturelles a été créée combinant tous les évènements catastrophes naturelles.

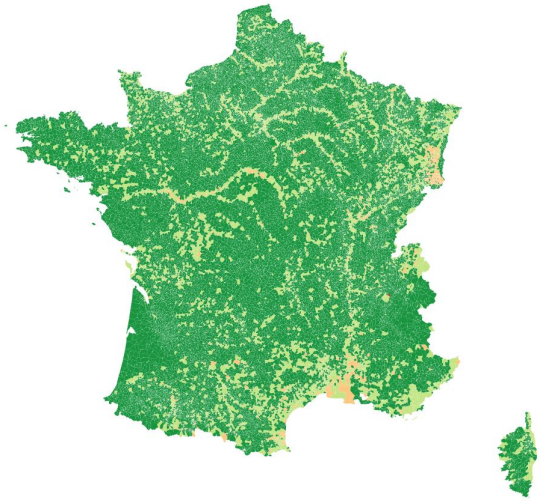

FIGURE 2.3 – Carte de suivi actuelle

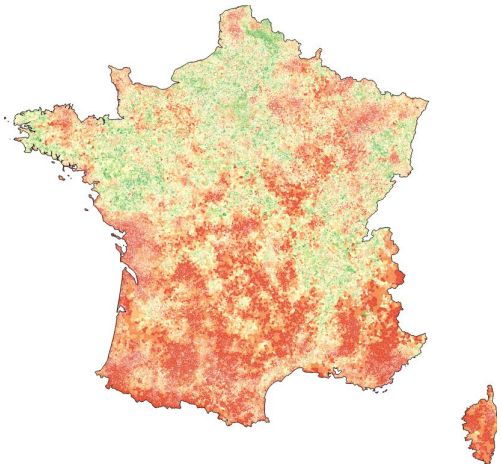

FIGURE 2.4 – Carte Voronoi

La carte de suivi actuel permet d'estimer notre exposition aux risques d'inondation. Elle est à trois niveaux de risques et faite à la maille du code administratif de l'INSEE. Elle est construite sur une base de données représentant l'historique de 10 ans s'étalant de 2002 à 2011. Sur cette carte on reconnaît directement le réseau hydrographique de la France. Le Sud-Est de la France regroupe la majeure partie de la zone la plus risquée. Cela est aussi caractéristique de cette région souvent touchée par des inondations. La raison de ce risque accru est dû à la combinaison de deux phénomènes présents dans cette région. En effet la zone Sud-Est est confrontée à des précipitations fortes sur une durée de 24 à 72 heures. Les rivières n'étant pas assez profondes et larges pour absorber cet afflux , elles débordent. Le débordement de rivières va donc occasionner des inondations.

La carte Voronoi est construite sur un historique de sinistralité de 10 ans aussi mais sur la période 2003 à 2013. La sinistralité prise en compte dans la construction est la sinistralité de l'unité de production Catastrophe Naturelles. Cette carte combine donc les sinistres liés à la sinistralité inondation qui, rappelons le, constitue la majeure partie de la sinistralité catastrophes naturelles mais aussi des sinistres liés à la sécheresse. Cette carte est décomposée en une vingtaine niveaux de risques. On a une exposition égale dans chacune des classes. Le maillage de cette carte est très fin. En effet pour les communes les plus représentées dans notre portefeuille, une zone peut représenter un seul bâtiment. Cette précision a pu être obtenue notamment par la forte présence d'Axa dans l'assurance habitation. Dans la figure 2.4 la carte ne peut donc être interprétée de façon macro car les zones sont tellement fines que l'analyse doit être faite sur une zone particulière.

2.5 Une Distinction de risques avec les nouveaux zoniers

Le risque catastrophes naturelles peut être distingué en trois risques majeurs. Nous considérons que la France est risquée en premier lieu à l'inondation. En effet ce risque est présent dû au fait d'une urbanisation très forte autour des rivières et aussi en raison de

certaines régions exposées à d'intenses pluies. Le deuxième risque est la sécheresse : les sinistres de 2003 ont montré l'exposition du territoire à ce risque et aussi l'ampleur d'un tel phénomène. En dernier, la tempête Xynthia nous a fait prendre conscience de l'importance du risque de submersion marine. Nous décidons donc de répartir notre prime de catastrophes sur ces trois risques.

$$\mathrm{PP}_{Cat} = \mathrm{PP}_{\mathrm{I}nondation} + \mathrm{PP}_{\mathrm{S\acute{e}cheresse}} + \mathrm{PP}_{\mathrm{S}submersion\mathrm{M}arine}$$

Nous avons donc pu avoir accès à trois cartes qui classent ces trois risques sur la France. Ces trois zoniers sont totalement construits sur des données géographiques et les sinistres n'ont à aucun moment été utilisés.

2.5.1 Carte inondation

Au sein du groupe Axa, l'équipe réassurance d'**Axa Global Property & Casualty** a étudié de près le risque inondation en France. Ces études ont amené à créer une cartographie de ce risque. Trois cartes de susceptibilité du risque sont crées par cette équipe. Ces trois cartes sont de même résolution : un découpage de la France est fait en carreaux de 75 mètres par 75 mètres.

— Carte de période de retour 200 ans : cette carte représente les zones très proches des rivières et à faible hauteur par rapport à la rivière. Cette carte est binaire (0 : non risqué , 1 : risqué)

— Carte de hauteur par rapport à la rivière la plus proche

— Carte de hauteur par rapport au drainage le plus proche

Afin d'avoir une cohérence et une simplicité du modèle, on décide transformer les deux dernières cartes en cartes binaires en risque. Le but sera d'arriver à une seule carte résumant le risque d'inondation.

La construction du zonier final repose sur la bonne discrétisation des deux cartes de hauteur.

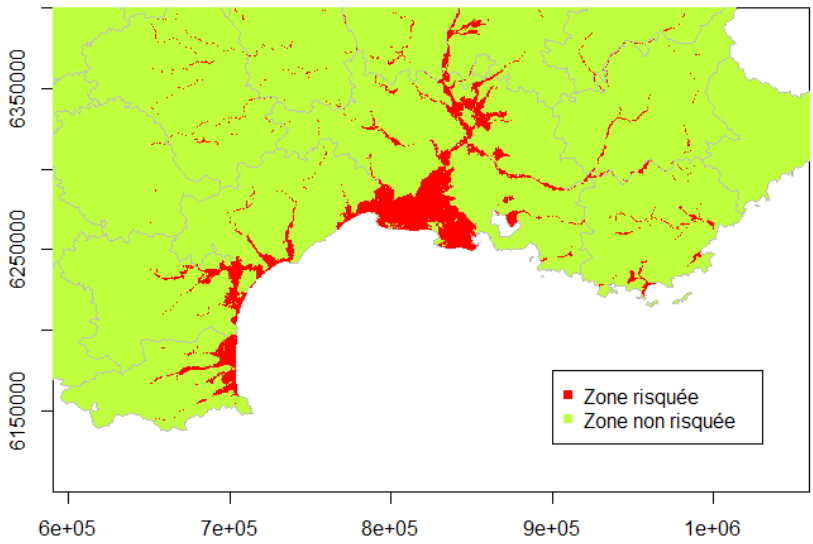

FIGURE 2.5 – Carte de modélisation de période de retour 200 ans

Cette carte va représenter la base de la détection des zones les plus exposées au risque. Ici nous avons représenté la zone géographique du Sud-Est. On reconnaît la partie du parc régional de Carmargue. Cette région est particulièrement exposée en raison du faible niveau avec la mer, la proximité des étangs et le déversement du Rhône dans la méditerranée. Ces éléments additionnés en font une partie particulièrement risquée. Le zonier va donc permettre de détecter les zones équivalentes sur tout le territoire de la France.

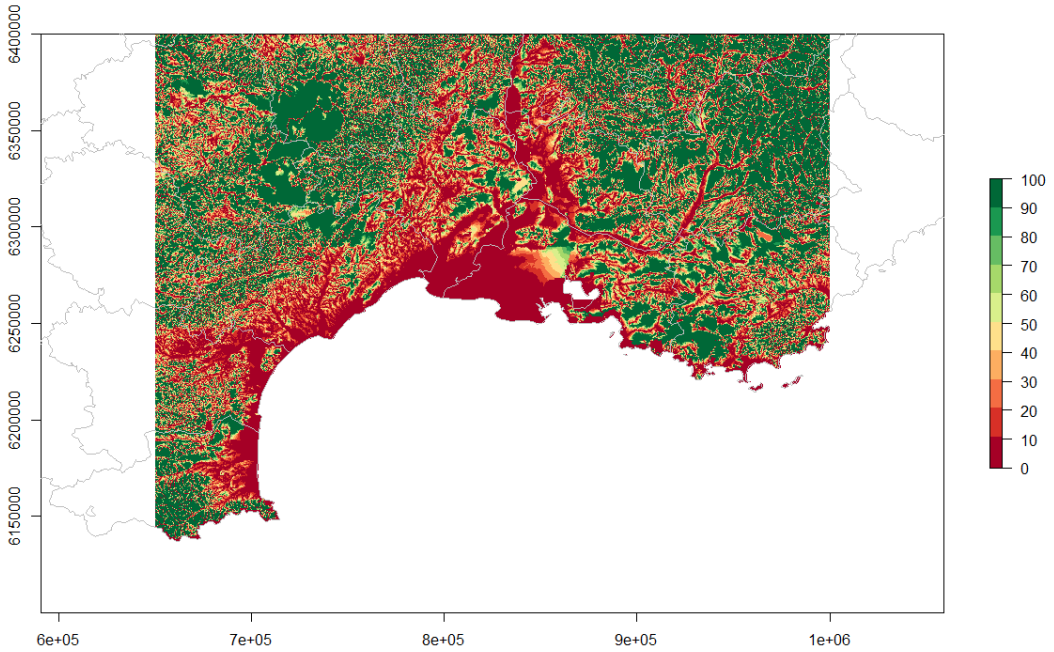

FIGURE 2.6 – Carte de hauteur par rapport à la rivière la plus proche

Cette carte est établie en mesurant la hauteur vis-à-vis de la plus proche rivière. Pour que cette carte soit lisible, dans notre représentation nous avons mis tous les niveaux en dessous du niveau de la mer à 0 mètre de hauteur et tous ceux qui dépassent 100 mètres

à 100 mètres. On peut remarquer que cette carte offre des zones complémentaires à la carte précédente . L'apport d'information portera principalement sur l'effet des rivières sur le risque inondation. Notre but sera donc d'utiliser cet apport pour essayer d'affiner le zonier inondation. Cependant pour avoir une certaine facilité d'interprétation et une robustesse du modèle, il nous faut simplifier cette carte.

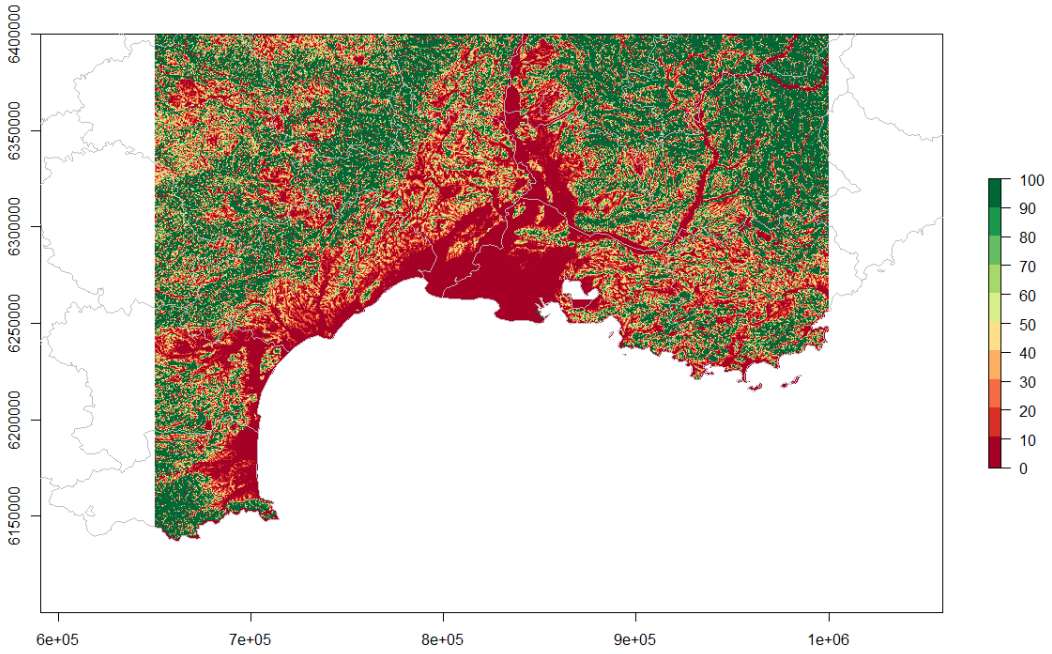

FIGURE 2.7 – Carte de hauteur par rapport au drainage le plus proche

La carte représente la hauteur par rapport au plus proche drainage. Les drainages sont les chemins naturels d'eau. Lors des fortes pluies , l'eau suivra donc ses chemins pour principalement se déverser dans les rivières. Ces chemins sont donc secs la plupart du temps. Cette carte va donc compléter les deux cartes antécédentes. D'ailleurs ici aussi on remarque une certaine inclusion des cartes antécédentes dans les zones à très basse hauteur. On voit que les zones en rouge (plus risquées) sont plus larges que celles d'avant d'où l'impression d'inclusion.

Comme précisé précédemment, on cherche à rendre les trois cartes binaires en risque. pour cela nous fixons une hauteur en dessous de laquelle le risque est considéré présent. Le choix d'une hauteur de 3 mètres pour les rivières et 5 mètres pour les réseaux de drainages nous a été suggéré par l'équipe de l'entité **AGPC**. On rend alors binaire les deux cartes en segmentant par rapport à ces deux hauteurs. On a maintenant à disposition trois cartes binaires en risques : Risque de période de retour 200 ans, Risque de rivière et Risque de réseau de drainage. On se propose alors de définir une hiérarchie entre ces trois risques d'inondation. Ainsi on se propose la mécanique suivante :

— On regarde si le contrat est dans la **Zone de période de retour 200 ans**

— Sinon on regarde s'il est en **Zone Rivières**

— Sinon on regarde s'il est en **Zone Ruissellements**

— Sinon il est **Hors zone de risque**

En utilisant cette mécanique, on arrive a une seule carte qui résume les trois niveaux du risque inondation.

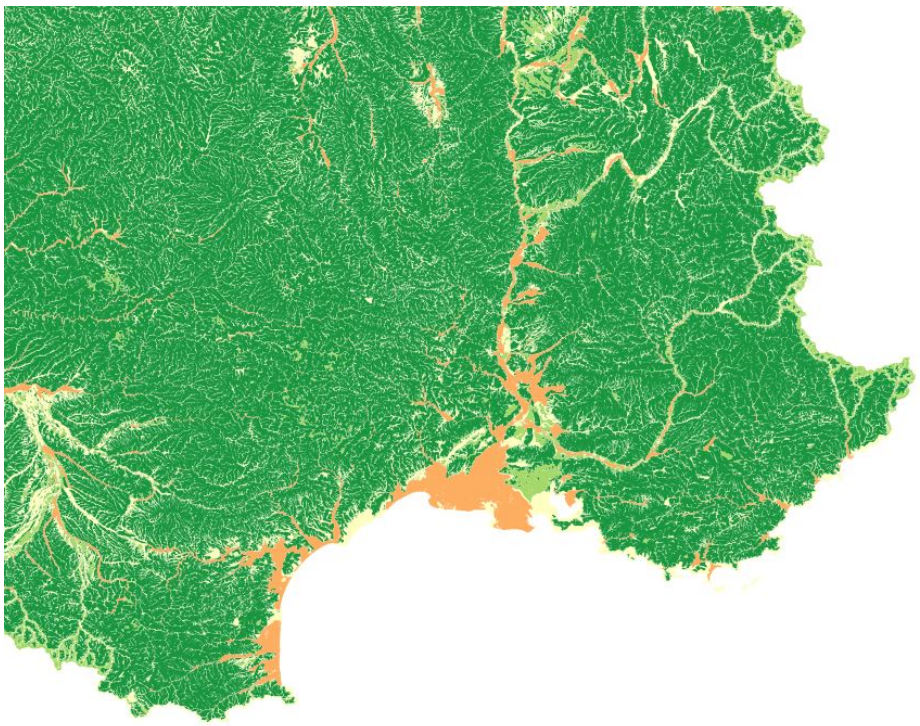

FIGURE 2.8 – Zoom sur le Sud-Est de la carte de risque inondation

On retrouve donc ici la tendance qu'on apercevait sur notre première carte de modélisation inondation. Les deux zones qu'on a rajoutées apportent un complément d'information dont on essaiera de quantifier l'apport lors de la calibration du modèle. Cependant, on remarque déjà que le fait d'avoir 4 niveaux de risques rend la carte beaucoup plus lisible. On retrouve la partie risquée du parc régional de Camargue. On distingue aussi facilement les majeures rivières. En analysant de plus près, on aperçoit aussi les réseaux de drainage qui se déversent dans les rivières.

Il est important de quantifier la couverture de ces zones sur la France. On regarde d'abord, le pourcentage de superficie de la France que chacun des risques représente.

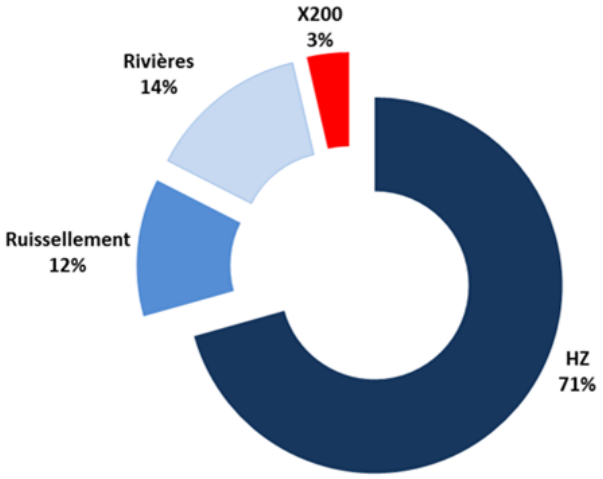

FIGURE 2.9 – Une répartition du territoire raisonnable des risques inondations

Le zonier classifie donc 29% du territoire en tant que risqué à un certain degré de niveau pour l'inondation. Ce pourcentage est rassurant, car la première carte ne représentait que 3% du territoire ce qui pouvait dégrader la performance des modèles. L'apport des zones complémentaires de rivières et ruissellements a permis d'agrandir la surface risquée. Ces zones sont intuitivement risquées. On n'utilise donc pas nos données de sinistres pour établir ce zonier. On souhaiterait donc que cette dé-corrélation du zonier des sinistres permette d'avoir une carte qui soit stable sur la durée.

Cependant une part de ses superficies est submergée par les rivières. Par exemple, le Rhône est inclus dans la zone X200 comme toutes les rivières. il est donc légitime qu'on regarde plutôt la répartition de notre portefeuille dans ce zonier. Nous décidons de faire une jointure spatiale de notre portefeuille 2015 sur ce zonier et d'analyser sa répartition. On garde en mémoire qu'un biais est fait sur cette répartition.

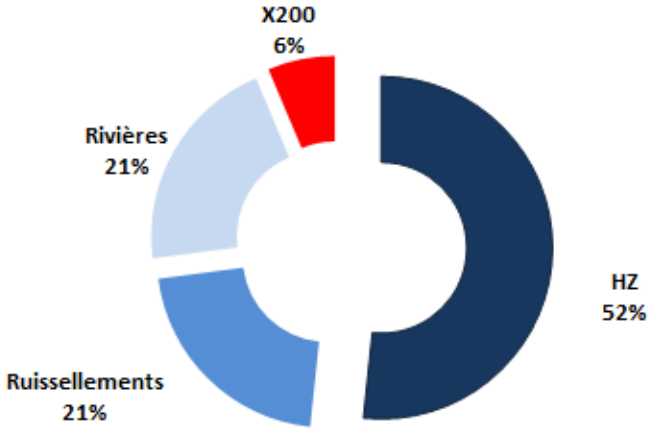

FIGURE 2.10 – La moitié des contrats est située sur des zones de risques inondation

On remarque que la répartition selon les zones n'est pas la même. En effet, le portefeuille

est plus exposé au risque. La moitié des contrats est classée comme étant risquée. Un explication intuitive à ce résultat vient du fait qu'on a une forte urbanisation près des rivières (Paris et Lyon par exemple). La répartition qu'on obtient sur les zones, nous rassure cependant sur le fait qu'on a assez de représentation pour pouvoir segmenter.

2.5.2 Carte de Sécheresse

Le retrait ou gonflement du sol argileux est un phénomène qui est potentiellement très dangereux pour le bâtiment. En effet ce phénomène peut créer des tassements différentiels. Le manifestement de ces tassements se repère par des fissures sur les murs. Deux phénomènes sont responsables de ces fissures : le retrait et le gonflement. Le retrait intervient sur le sol argileux qui a subi un dessèchement au point qu'il s'enfonce. Un gonflement est dû au remplissage par l'eau des couches argileuses desséchées à la suite d'un épisode de sécheresse.

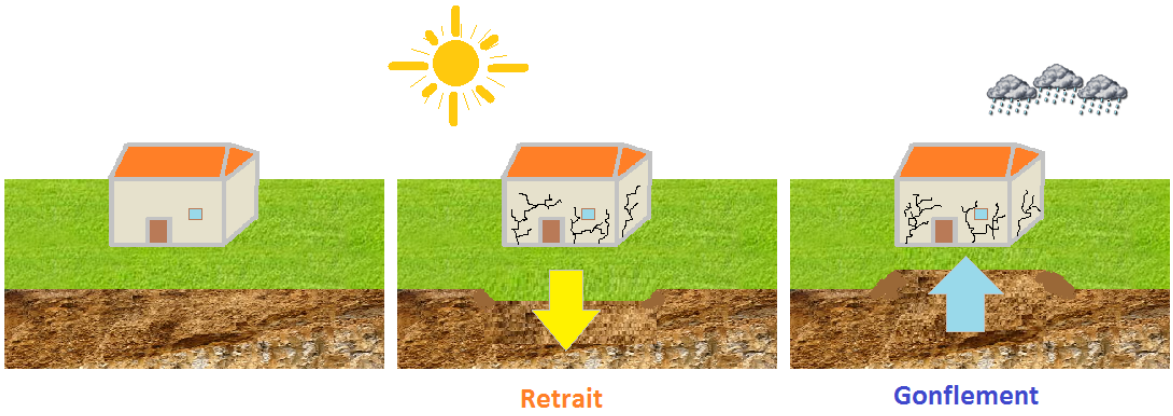

FIGURE 2.11 – Retrait et Gonflement à l'origine des fissures de murs

Dans cette partie, nous présentons brièvement l'estimation des coefficients de GLM du modèle sécheresse.

La carte est établie pour chaque département. Le financement a été fait par le Bureau de Recherches Géologiques et Minières (BRGM) et par les régions. Cette étude, réalisée par le BRGM dans le cadre de sa mission de service public sur les risques naturels, s'intègre dans un programme national de cartographie de l'aléa retrait-gonflement des sols argileux. Chaque département crée une carte de risque selon une méthodologie définie. Ensuite, la carte est coordonnées avec les départements limitrophes. Ce processus permet d'avoir une carte d'une très grande finesse et homogènes sur la France.

La méthode utilisée se divise en deux parties :

— Évaluation de l'Aléa par géologie Carte de susceptibilité

— Évaluation par densité de Sinistres

La sécheresse est le résultat d'une combinaison de deux facteurs : les facteurs de prédisposition et les facteurs de déclenchement.
Facteurs de prédisposition sont représentés par La nature du sol , la géomorphologie, la végétation,..
Facteurs de déclenchement sont principalement les phénomènes climatiques exceptionnels. Ces phénomènes modifient de façon abrupte la teneur en eau du Sol ce qui entraîne un retrait ou un gonflement soudain du sol.

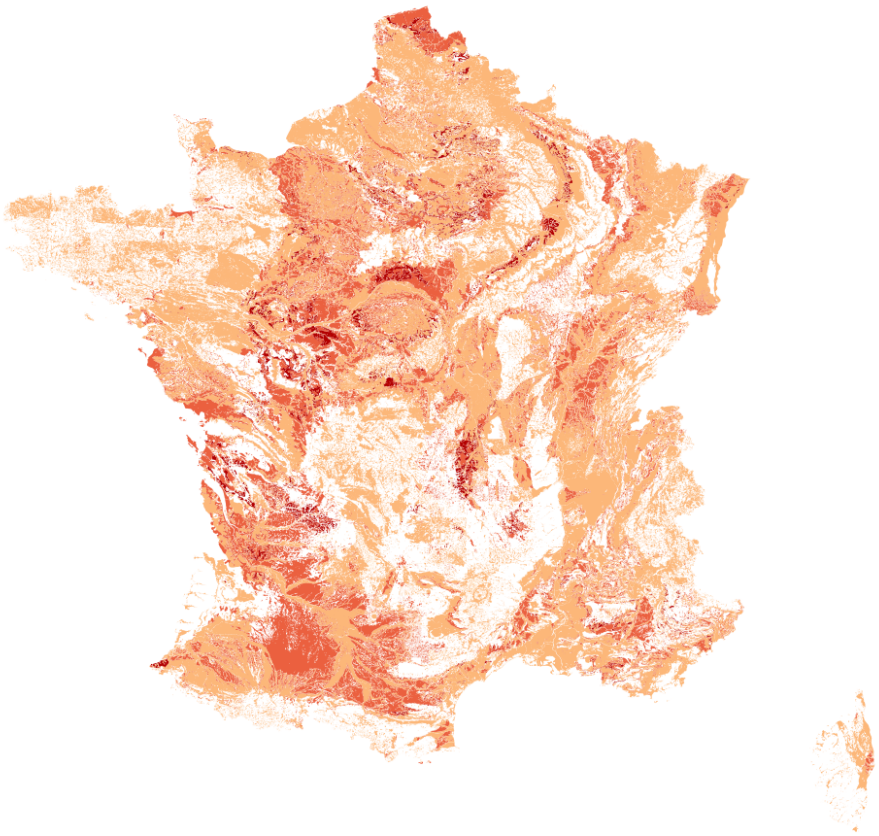

FIGURE 2.12 – Niveaux de sécheresse sur la France

La carte finale (représentée ci-dessus) est à quatre niveaux de risque : Nul, Faible, Moyen et Fort. Sur cette carte on voit qu'une large zone à fort risque se situe au Sud-Ouest de la France. Lors de la construction du zonier Voronoi, figure 2.4 - page 28, (Basé sur tous les sinistres catastrophes naturelles) cette région est aussi apparue comme zone risquée. Cela provenait principalement de l'évènement de sécheresse de 2003.

2.5.3 Carte de submersion marine

Afin de calibrer notre modèle de prime pure submersion marine, nous utilisons une carte de risques faite par Swiss Re. Cette carte classe les zones de risques de submersion marine en 5 niveaux.

— **Hors Zone** : la zone n'est pas risquée au terme de la submersion marine. cette zone représente bien évidemment la majeure partie de la France.

— **Faible** : Zone près de la côte mais la hauteur et la distance par rapport à cette dernière reste considérable.

— **Moyen** : Zone près de la côte mais la hauteur ou la distance représente un risque.

— **Fort** : la zone est proche de la côte en terme de hauteur de distance

— **Très Fort** : il s'agit des zones très proches des côtes avec une faible hauteur par rapport au niveau de mer

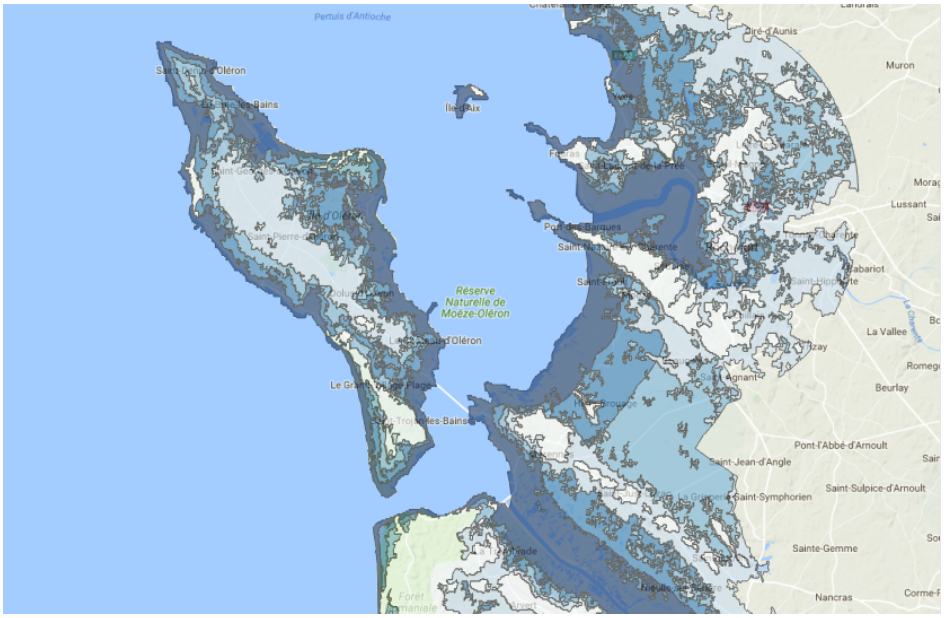

FIGURE 2.13 – Niveaux de submersion marine au niveau de l'île d'Oléron

La zone géographique concernée par ce type d'évènement est uniquement la côte. Les zones ayant un risque sont les zones situés sur une distance allant jusqu'à quarante kilomètres de la côte. Bien évidemment, cette carte va concerner une très faible partie du portefeuille. Cependant, sur les zones à fort risque, le perte potentielle est très élevée.

2.6 Calibration du GLM pour l'Inondation

Les inondations représentent la majeure partie de l'assurance Catastrophes Naturelles. En terme de charges , les inondations représentent deux tiers des charges de catastrophes naturelles en France. En partant de ce constant, la modélisation de cette garantie est primordiale pour le modèle de catastrophes naturelles entier.

Une collaboration avec l'équipe de réassurance de l'entité Axa Global Property & Casualty a été nécessaire à l'élaboration du modèle et surtout de la carte de risque.

Dans cette étude, nous présentons une modélisation de la prime pure Inondation avec l'utilisation des GLM.

Il est important d'avoir un regard critique quant à l'utilisation de ce modèle. En effet, la fréquence des sinistres est très faible. Cependant, on décide de tester ce genre de modèle tout en restant prudent dans l'interprétation de certains résultats.

2.6.1 Analyse préliminaire

Dans cette partie nous tentons d'analyser les variables qui auraient le plus de pertinence pour expliquer le risque d'inondations. Pour cela, nous analysons les tendances de la prime pure observée par modalités pour les variables clés.

L'habitation

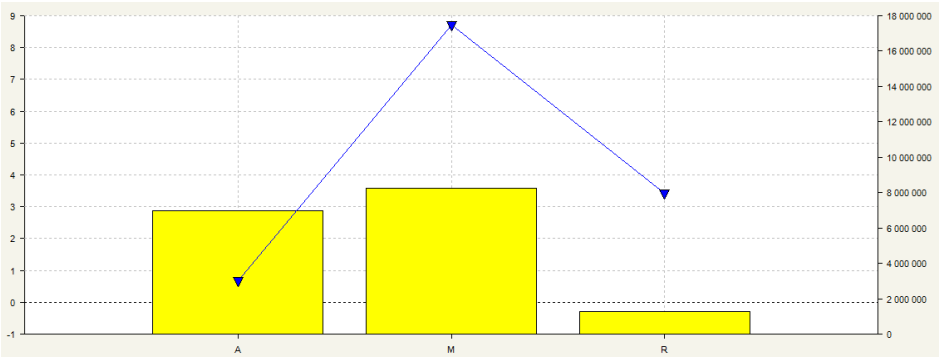

FIGURE 2.14 – Prime pure inondation par type d'habitation

Ce graphique se lit de la manière suivante : Les histogrammes en jaune représentent l'exposition de la modalité. Par exemple on voit qu'ici les maisons représentent environ la moitié. La courbe en bleu représente la prime pure observée. Ainsi ici les maisons sont les plus risquées vis à vis de l'inondation. Avec une prime pure observée de 9€. Viennent ensuite les rez-de-chaussés à 3,5€ et enfin les appartements à moins d'un euro de prime pure observée.

Ces observations viennent confirmer notre intuition. En effet, il est intuitif de se dire qu'un appartement en étage ne soit pas touché par les inondations. les rez-de-chaussés sont bien évidemment plus exposés.

Une façon de confirmer cette tendance est de tracer la tendance par année. Bien sur sur les années non sinistrées nous aurons une courbe plate.

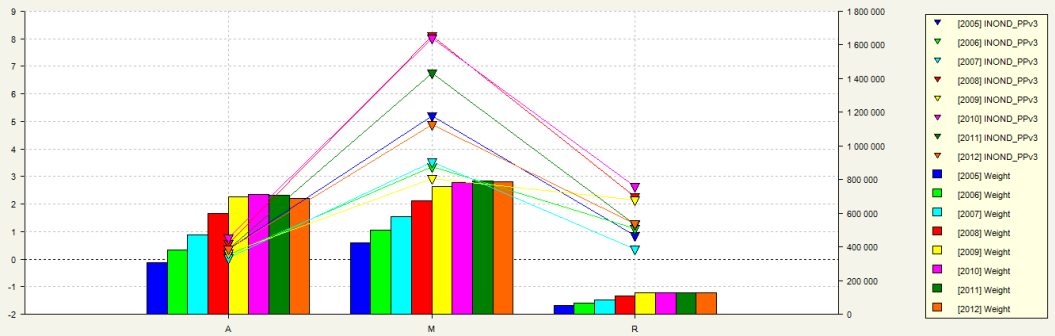

FIGURE 2.15 – Prime pure inondation par type d'habitation et par année

Ce graphique confirme la consistance de la tendance observée sur la base totale.

Objets de valeur

La part d'objet de valeur représente le pourcentage des objets de valeurs du capital mobilier déclaré. Cette variable est qualitative (0% , 10%, 20%, 30%, 40% ou 50%). Elle est demandée lors de la souscription. Nous décidons de voir la tendance sur cette variable à la place du capital car la discrétisation de la variable capital est un peu contraignante mais aussi car la tendance observée sur le capital est moins stable que celle sur la part d'objets de valeur.

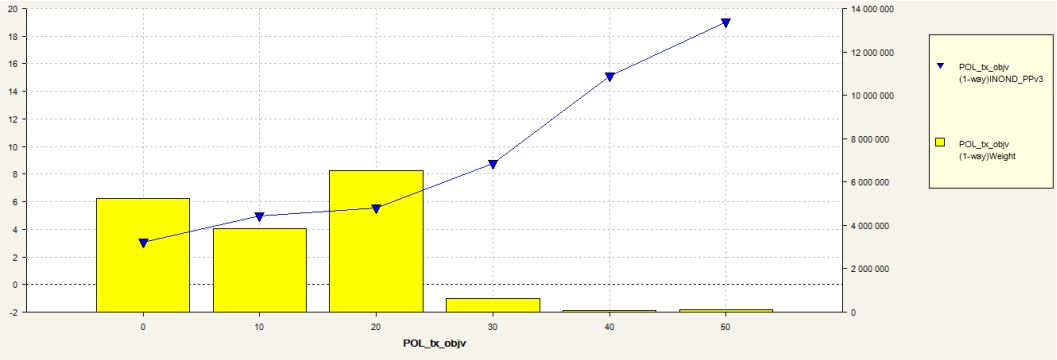

FIGURE 2.16 – Prime pure inondation par part d'objets de valeur

On distingue bien ici une tendance croissante du risque en fonction de la part d'objets de valeur. Cependant, on remarque qu'à partir de 30% de part d'objets de valeur , les modalités sont très peu représentées. On devra tenir compte de ce phénomène lors de la calibration de notre modèle. Pour valider cette tendance, on regarde aussi le croisement avec l'année.

L'explication de cette tendance vient probablement du fait que lors d'une inondation ces objets de valeurs sont endommagés. Cette variable est aussi corrélée avec le capital alors il se pourrait aussi qu'on perçoive l'effet du capital dans cette analyse. Plus on a de capital mobilier plus le coût du sinistre sera élevé .

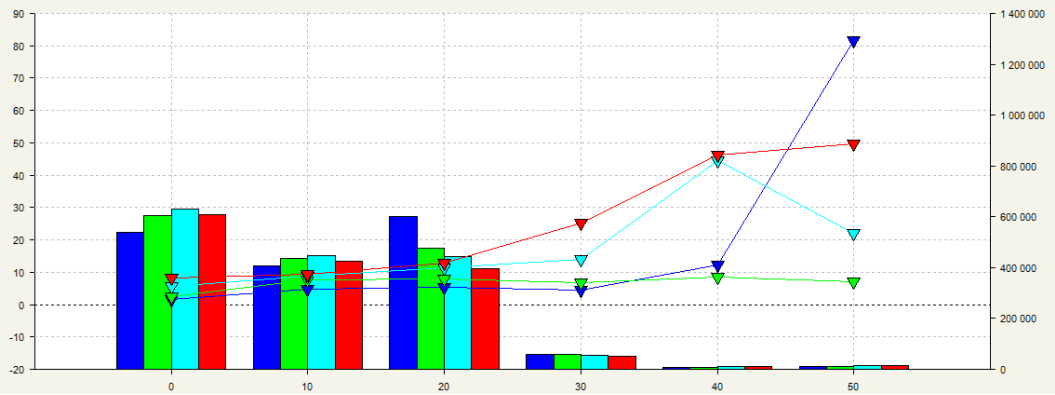

FIGURE 2.17 – Prime pure inondation par part d'objets de valeur et par année

Nous ne traçons ici que les années qui ont été sinistrées au titre de l'inondation pour une lecture plus facile du graphique. On peut confirmer cette tendance sur les modalités jusqu'à 30% . Cependant on voit un peu l'effet de sous représentation des modalités 40% et 50%. Sur ces deux dernières modalités, la tendance est beaucoup plus aléatoire.

La qualité

La qualité de l'occupant joue un rôle important dans certaines garanties. Il serait donc judicieux de regarder si cette variable a aussi une tendance pour les inondations.

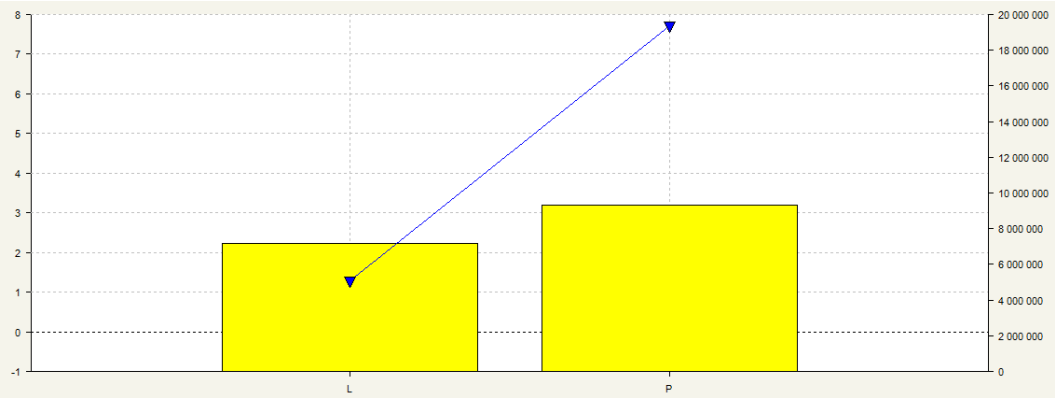

FIGURE 2.18 – Prime pure inondation par qualité d'occupant

La différence de prime pure observée entre les locataires et les propriétaires est remarquable. En effet, elle passe de 1.3€ à 7.7€. Cette tendance est aussi consistante dans le temps.

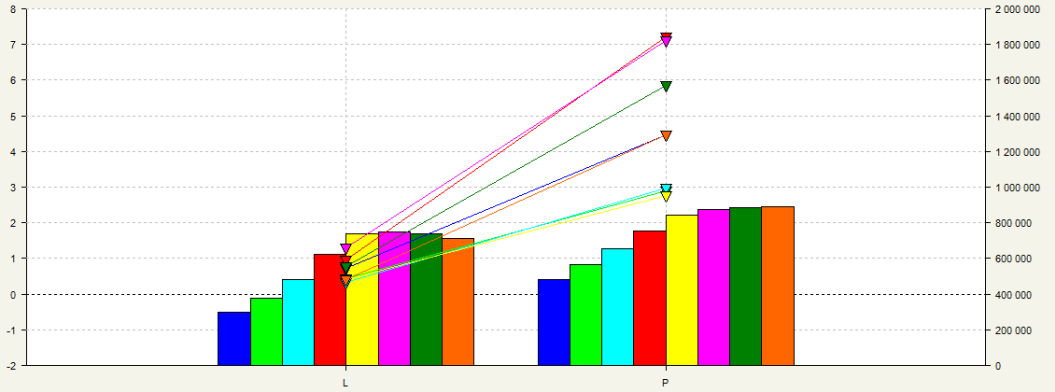

FIGURE 2.19 – Prime pure inondation par qualité d'occupant et par année

Nous devrons voir si une corrélation existe ou non entre la qualité d'occupant et le type d'habitation. Il se pourrait qu'on ait par exemple la plupart des propriétaires qui habitent en maison et la plupart des locataires en appartement. Auquel cas, il faudrait plutôt inclure la variable croisée Habitation et Qualité dans le modèle.

Le zonier

Comme précisé sur la méthode de construction du zonier inondation, les données d'Axa n'ont été utilisées à aucun moment. Le zonier est basé purement sur des constations géographiques. Cette étape va donc nous donner une idée sur la performance de ce zonier quant à sa capacité à distinguer les zones de risques.

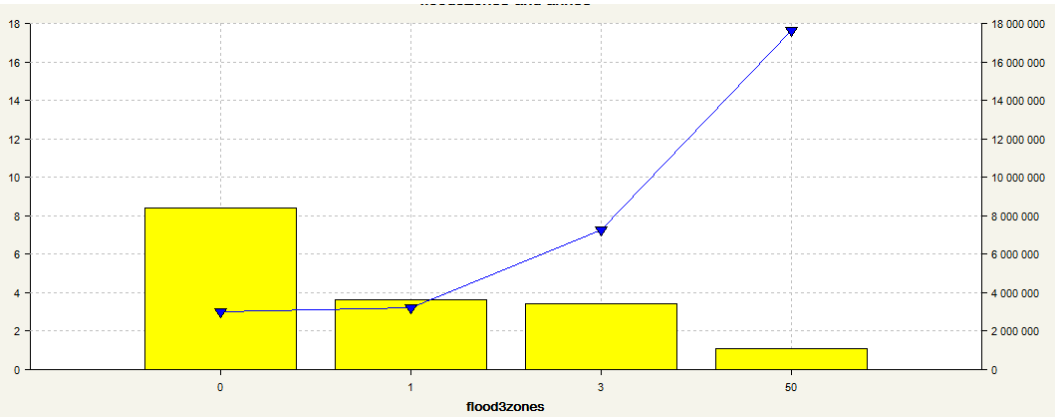

FIGURE 2.20 – Prime pure inondation par zone de risque inondation

Le résultat est très satisfaisant à première vue. la tendance est clairement croissante en fonction des zones. on aperçoit un faible apport par les zones de ruissellements mais les zones rivières et surtout la zone de période de retour 200 ans permettant de détecter les contrats les plus risqués.

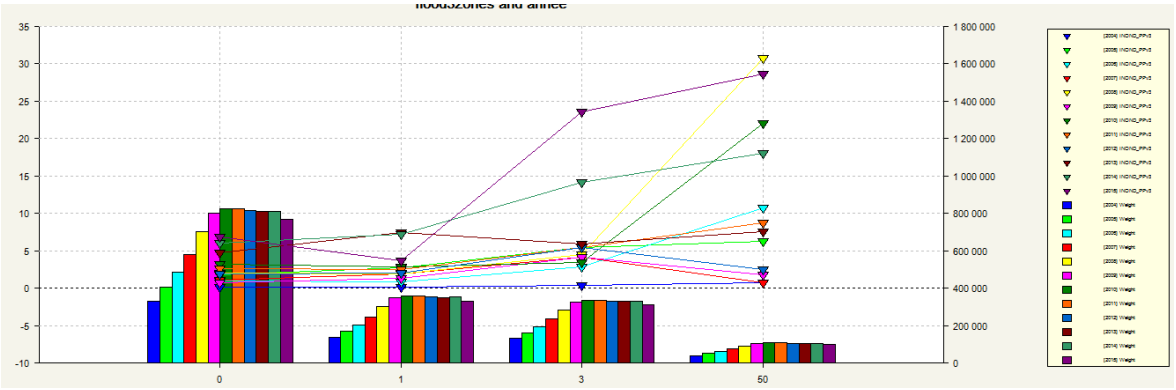

FIGURE 2.21 – Prime pure inondation par zone de risque inondation et par année

On décide ici d'illustrer toutes les années. On aperçoit une certaine stabilité temporelle de notre zonier. En effet, il est concevable que sur les années à très faible sinistralité, le zonier ne puisse pas détecter ces évènements. On peut donc estimer que notre zonier sera une variable bien segmentante pour notre modèle.

Sélection des variables

La méthode de sélection des variables choisie est la sélection Forward. Cette méthode consiste à considérer en premier lieu le modèle avec une simple moyenne. On teste ensuite le rajout de toutes les autres variables une à une. On garde celle qui permet le plus grand gain selon un critère (par exemple meilleur réduction du critère AIC : Akaike information criterion ou BIC : bayesian information criterion). On rajoute ensuite une autre variable jusqu'à obtention du modèle complet. Cependant , en général, on fixe un critère d'arrêt de l'algorithme avant l'obtention du modèle complet. Ce critère peut être par exemple que pour le rajout d'une variable en plus il faut que l'amélioration du critère soit plus grande qu'une certaine quantité.

Sur nos données nous avons testé une méthode Forward avec pour critère la réduction du BIC. le nombre de variables à tester est de 12 variables. Les résultats suivants représentant les variables qui apportent le plus d'information et qui seront inclues dans le modèle.

- **HAB_habit** : type d'habitation (Appartement, Rez-de-chaussée et Maison)
- **Precip** : Zone de précipitations
- **flood3zones** : carte inondation France avec 4 niveaux
- **New HAB_nb_pieces** : nombre de pièces capé à 7 pièces
- **HAB_qual** : la qualité de l'occupant (Locataire ou propriétaire)
- **POL_tr_objv** : taux d'objets de valeurs
- **HAB_inhab** : inhabitation (principale ou secondaire)

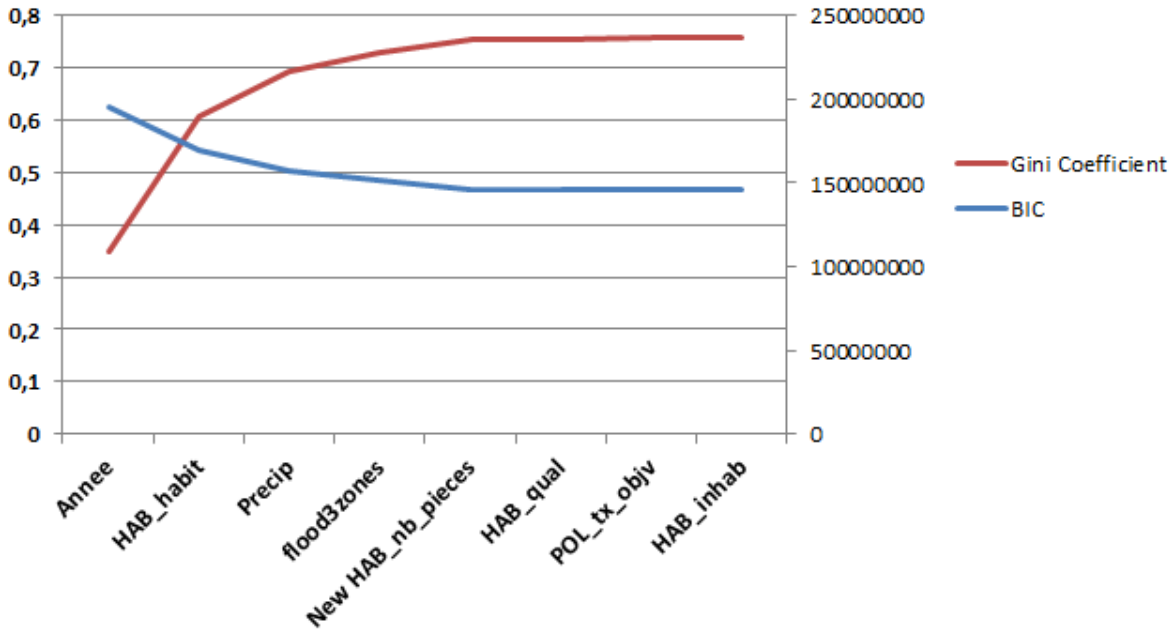

FIGURE 2.22 – Résultats de la méthode Forward sur nos données

On voit que ce qui joue le plus dans la segmentation de notre prime pure c'est en premier le type d'habitation : Maison ou Appartement. Le plus intéressant se trouve juste après en effet, nos deux zoniers sont les variables qui expliquent le plus. Sachant que ces zoniers sont construits sans prendre en compte les sinistres, cette information nous permet de valider leur pertinence.

2.6.2 Calibration du GLM

A la suite d'une sélection de variables avec la méthode Forward, nous avons retenu les variables suivantes :

- l'année : elle permet de détecter des années particulièrement sinistrées
- le type d'habitation : Maison ou Appartement
- Le taux d'objets de valeurs
- L'inhabitation : primaire ou secondaire
- La qualité : Propriétaire ou locataire
- Le nombre de pièces
- le zonier inondation
- le zonier de précipitation

Pour le nombre de pièces, nous décidons de faire une simplification. En raison de la faible exposition des contrats de plus de sept pièces, nous allons les regrouper dans une seule modalité. Nous mettons donc toutes les habitations de plus de 7 pièces dans la même classe.

La cross-validation

Pour valider le modèle, nous allons faire une cross-validation. D'abord nous excluons deux années de notre base (2014 et 2015) pour les réserver pour des tests. Sur la base restante (2003-2013), nous faisons une cross-validation . La corss-validation permet d'améliorer le modèle en réduisant le sur-apprentissage. On peut aussi l'utiliser pour vérifier la consistance de notre modèle.

La cross-validation simple consiste à couper la base de données en deux morceaux. Le premier morceau consiste en la base d'apprentissage. Les coefficients du GLM sont calibrés dessus. Ensuite le deuxième morceau sert de base de test.

La k-fold cross-validation est une amélioration de la méthode simple. Le set de données est divisé en k sous ensembles. Un sous-ensemble est utilisé comme base de test et les k-1 sous-ensembles restants constituent la base d'apprentissage. Le modèle final est obtenu en moyennant les modèles. Cette technique, nous permet par exemple de regarder la stabilité des coefficients des modèles intermédiaires.

Résultats de la cross-validation

Afin de calibrer notre modèle, nous décidons de faire une 10-fold cross-validation sur la base de 2003 à 2013. Les sous-ensembles sont constitués à 70% de base d'apprentissage et de 30% de base de test.

En premier on essaie de regarder la stabilité du Gini.

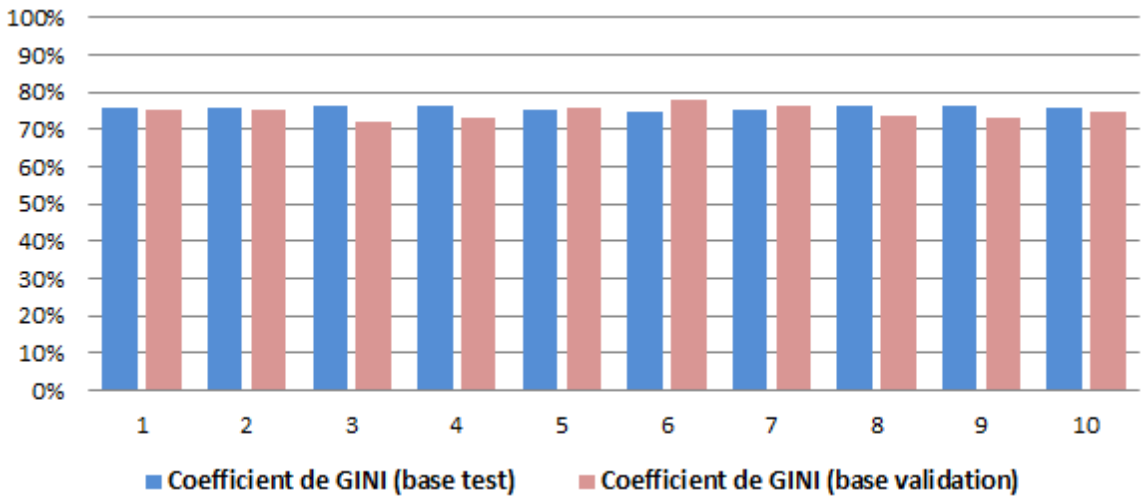

FIGURE 2.23 – Un Indice de Gini stable

Les résultats sur le Gini, sont plutôt satisfaisants, nous avons un Gini moyen à 75.81% pour les bases de modélisation et à 74.85% pour les bases de tests. Le plus grand écart se situe à la troisième itération. Cet écart est de 4 points de Gini. Cependant étant donné la

stabilité moyenne de cet indice, nous pouvons conclure une stabilité globale du modèle. Il faut noter que cet indice est utile pour juger de la qualité de classement des risques. Le choix de ce critère est un choix de la direction afin de simplement comparer des modéles.

Tendances et consistances

Grâce à la cross-validation, nous pouvons voir les tendances estimées par chacun des 10 modèles. Cela nous permet de voir si on a une certaine consistance des coefficients. Nous nous intéressons tout d'abord à la tendance observée sur les deux zoniers qu'on a rajoutés. On souhaiterait surtout avoir une consistance sur ces données externes.

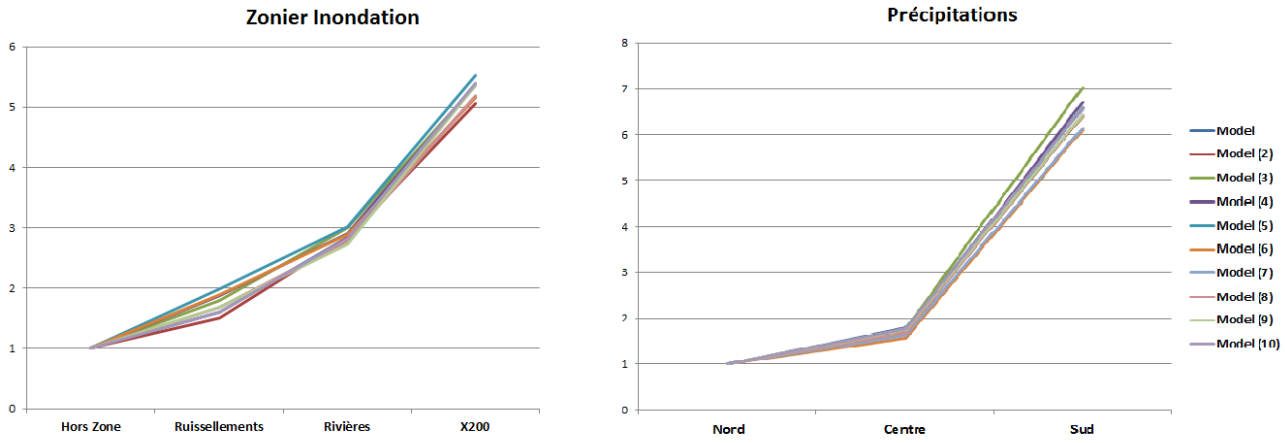

FIGURE 2.24 – Une tendance consistante sur les données géographiques

Sur les deux cartes introduites, on voit une tendance qui est consistante sur les dix modèles. Cette information nous conforte sur l'utilité et l'apport de ces données. On remarque surtout que le spread de relativité sur les deux variables est très grand (1 :6). Ces variables segmentent bien le risque.

Cette cross-validation nous a aussi permis de détecter d'autres phénomènes sur certaines variables. On présente ci-dessous la tendance sur la variable taux d'objets de valeurs.

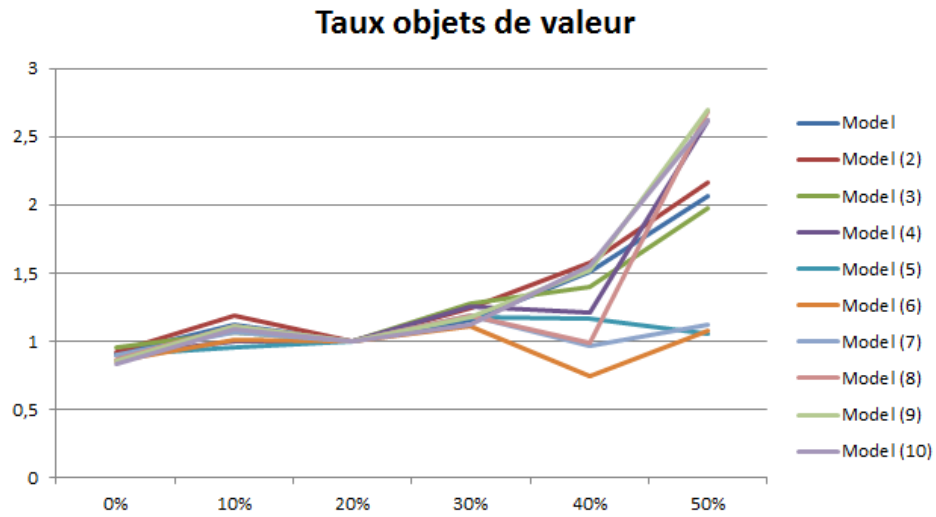

FIGURE 2.25 – Une forte volatilité des taux d'objets de valeur élevés

On remarque que sur cette variable, les deux dernières modalités représentent une forte volatilité. Le fait que ces modalités sont peu représentées explique ce phénomène observé. On pourrait en premier lieu regrouper ces deux modalités avec la modalité 30% (créer une classe Taux d'objets de valeurs >30%). Mais on peut se questionner sur l'apport de cette variable dans le modèle. En effet on aurait une courbe quasiment plate. Nous décidons donc d'exclure cette variable.

Nous avons repéré un autre phénomène sur la variable Inhabitation. Cette variable a été la dernière à être rajoutée par la sélection Forward. Nous regardons donc la stabilité de ses coefficients.

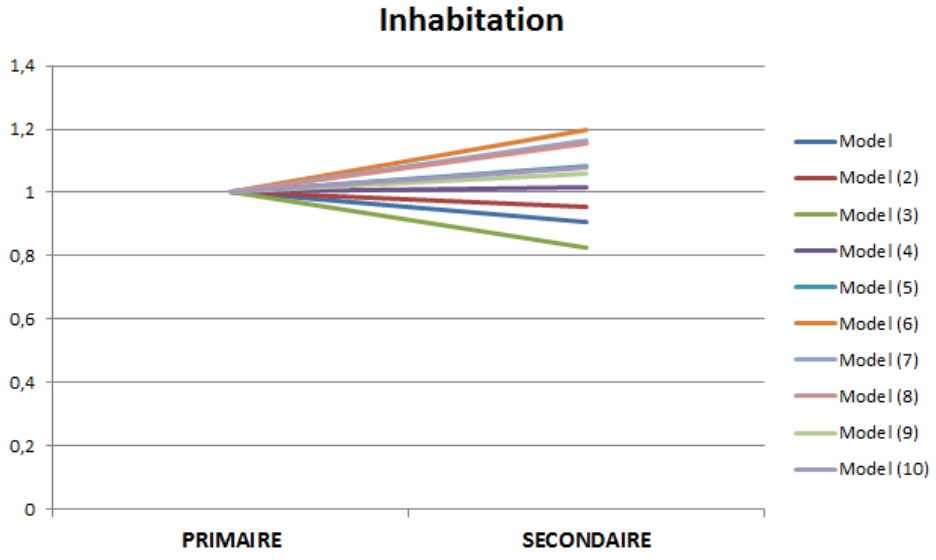

FIGURE 2.26 – Une instabilité avérée de la variable

L'instabilité des coefficients est claire sur cette variable. La tendance est même parfois dans des sens contraires selon les itérations du 10-fold. Nous retirons donc cette variable

de notre modèle.

On a aussi essayé de mesurer la stabilité temporelle de notre zonier. Pour cela nous testons le GLM obtenu sur 2003-2013 sur les deux années 2014 et 2015. Nous obtenons les résultats suivants :

	Variables Internes + Voronoi	**Variables Internes + Nouvelles Cartes**
Apprentissage	81.88%	81.40%
Test	69%	79.51%

TABLEAU 2.2 – Mesure de stabilité du Gini

On remarque ici que la nouvelle carte introduite permet de garder une certaine stabilité du modèle. De plus sur la base d'apprentissage, les deux modèles ont un indice de Gini comparable. Les nouvelles cartes permettent donc de garantir une certaine stabilité temporelle du modèle.

Chapitre 3

Gestion des risques à la souscription

« Le propre de la solidarité, c'est de ne point admettre d'exclusion. »

Victor Hugo

Sommaire

3.1 Cadre légal et éthique 48
3.1.1 Les tentatives étatiques 48
3.1.2 L'article L125 du Code de l'assurance 48
3.1.3 La réassurance des Catastrophes Naturelles 49
3.1.4 Une égalité entre les assurés ? 49
3.2 Analyse de la répartition de la prime pure 50
3.2.1 Analyse de la prime pure estimée des Maisons 51
3.3 Un écart de prime impactant la marge 53
3.3.1 Gestion de la marge sur les contrats "sur-assurés" 53
3.3.2 Gestion des contrats les plus risqués 54

3.1 Cadre légal et éthique

3.1.1 Les tentatives étatiques

La difficulté à estimer le coût et la fréquence des évènements climatiques a longtemps repoussé les assureurs à couvrir un tel risque. De plus la loi du 13 juillet 1982 selon laquelle *"Sauf convention contraire, l'assurance ne couvre pas les incendies directement occasionnés par les éruptions volcaniques, les tremblements de terre et autres cataclysmes"*. [1], laissait la main libre aux assurés de choisir ou non cette couverture. Bien-sur cela crée une demande de couverture pour les zones qui y sont principalement exposées. Cela implique une prime d'assurance élevée dûe au manque de mutualisation de ce risque.

Face à l'urbanisation de plus en plus forte, la nécessité de gérer l'indemnisation des sinistres de catastrophes naturelles prend de l'ampleur. L'Etat indemnise d'abord les victimes au travers des aides d'extrêmes urgence prélevée du budget du Ministère de l'Intérieur. L'Etat crée ensuite un fond sous le nom de Fonds de secours aux victimes de sinistres et calamités en 1956. Ce fond est alimenté par l'Etat , les collectivités et les dons publiques. Il intervient lors des évènements rares climatiques, tremblement de terre ou attentat. Toutes ces tentatives reposaient sur une indemnisation des sinistrés au travers de l'administration étatique, ce qui pouvait donner des délais conséquents et une charge en plus pour l'Etat.

3.1.2 L'article L125 du Code de l'assurance

L'Etat cède sa gestion des sinistres de type catastrophes naturelles avec la parution de la loi du 13 juillet 1982 qui laisse aux assureurs la charge de l'indemnisation de ces sinistres. Cette loi impose la garantie catastrophes naturelles aux contrats liés aux dommages aux biens. Cette extension de garantie est aussi apposée d'une prime additionnelle réglementaire. La garantie est donc obligatoire pour l'assureur qui se doit de la proposer mais aussi à l'assuré qui ne peut pas choisir de ne pas y souscrire. Cette garantie des évènements catastrophes naturelles fait donc partie du socle des garanties du produit MRH. Cet article de loi vient donc imposer d'une certaine façon la solidarité des assurés entre eux face aux risques de catastrophes naturelles. Ainsi toute personne ayant un contrat MRH participe au travers d'une part de sa prime à la couverture des risques catastrophes naturelles. L'article fixe aussi la méthode de calcul de cette prime. dans le cadre de la MRH, cette prime représente 12% de la part de la prime d'assurance destinée à couvrir les dommages causés aux biens.

Il est à noter que l'assureur ne peut refuser de souscrire ou renouveler un contrat pour cause que ce dernier a été sinistré au titre de la garantie catastrophes naturelles ou qu'il réside dans une zone particulièrement risquée. Dans le cas ou cela se réalise l'assuré

1. Article L122-6 du code des assurances

pourra se tourner vers le Bureau Central de Tarification (BCT). Le BCT va donc imposer à l'assureur la souscription du contrat. Dans le cas ou l'assureur refuse, il encourt le retrait de l'agrément administratif. Cependant une exception est faite pour les biens reconstruits sur des terrains jugés inconstructibles par les plans de préventions de risques naturels prévisibles (PPRN) et aussi aux biens immobiles construits en violation des règles administratives.

3.1.3 La réassurance des Catastrophes Naturelles

Les risques de catastrophes naturelles sont des risques de sévérité. Les assureurs obligés de couvrir les assurés pour ces risques, se tournent généralement vers un processus de réassurance. Cela permet de se couvrir contre des évènements potentiellement très coûteux et qui pourraient avoir un fort impact sur le bilan financier de l'entreprise. La Caisse Centrale de Réassurance (CCR) est l'organisme de référence en France pour la réassurance des catastrophes naturelles. Cette entreprise jouit en effet de la garantie d'Etat [2].Cette garantie d'Etat permet donc de rassurer les assureurs de la capacité de la CCR d'agir face aux sinistres élevés. Sur le marché français, le mécanisme de réassurance mis en place est un mécanisme de traité en quote-part. Une cession 50% des primes récoltés au titre de la garantie catastrophes naturelles et fait face à 50% de la charge des sinistres. De plus un mécanisme de couverture en excédent de perte annuelle est rajouté au mécanisme de réassurance en "quote-part". Ce second traité permet d'assurer la partie de rétention de l'assureur. Ce traité entre en jeu dans le cas ou la sinistralité dépasse un certain montant (en général exprimé en pourcentage des primes conservées).

3.1.4 Une égalité entre les assurés ?

Il est vrai que le fait d'imposer la garantie catastrophes naturelles à tous les contrats permet une mutualisation presque parfaite du risque entre les assurés. Cependant, est-il éthique -logique- de faire payer 12% de la prime pour un appartement au cinquième étage au 3ème arrondissement de Lyon et le même pourcentage pour une maison en bord de rivière ? Même si les risques sont certes différents, il est clair que vis à vis des risques de catastrophes naturelles, la géographie et la situation dans l'immeuble : en Rez-de-chaussée ou en étage, permettent d'avoir une différence conséquente par rapport à l'exposition au risque. Aujourd'hui comme énoncé précédemment (Article L125), il est interdit de jauger cette partie de la prime. Mais la question reste ouverte quant aux possibilités futures. Néanmoins, dé-contraindre cette solidarité pourrait avoir des effets néfastes directs sur les contrats exposés aux risques : l'augmentation de la prime totale. Il sera donc judicieux de garder cette prime fixe et d'envisager des mesures afin de limiter la sinistralité des contrats les plus risqués tout en avantageant les contrats aux différents degrés de risque.

2. Article L431-9 du code des assurances

3.2 Analyse de la répartition de la prime pure

Le but de cette analyse est surtout de faire une étude sur la prime pure modélisée. Dans cette étude on comparera aussi la prime catastrophe naturelles légale et la prime pure modélisée.

Notre prime pure totale de catastrophes Naturelles est définie comme la somme des trois risques :

$$\mathrm{PP}_{Cat} = \mathrm{PP}_{Inondation} + \mathrm{PP}_{Sécheresse} + \mathrm{PP}_{SubmersionMarine}$$

On établit l'étude sur les affaires nouvelles de 2015. Le périmètre sera un peu réduit à cause la contrainte de bon géocodage qu'on impose. Sur ce périmètre, nous avons une prime pure moyenne de 7.30€ et qui représente 2,7% de la prime dommages. En moyenne la prime est répartie de la façon suivante :

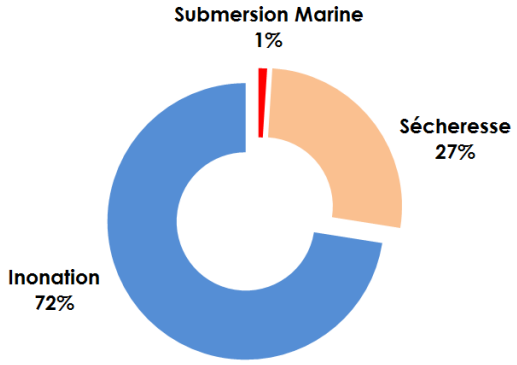

FIGURE 3.1 – Répartition moyenne de la prime pure catastrophes naturelles estimée

Comme attendu, la prime pure modélisée suit les observations sur le nombre de sinistres. La prime pure inondation porte 3/4 de la prime pure catastrophes naturelles. La submersion marine est quant à elle très faible car elle ne concerne que très peu de contrats.

L'avantage de segmenter cette prime en trois risques auquel le sol français est exposé nous permet d'avoir une estimation plus fine. En effet cette répartition varie beaucoup en fonction de la localisation du risque assuré. La part inondation va être plus grande dans le Sud-Est région qui est particulièrement risquée. La part de submersion marine peut aussi atteindre un pourcentage de la prime pure très élevé (plus de 95% pour de rares cas). Dans ces cas il s'agit principalement des maisons en bords d'océans.

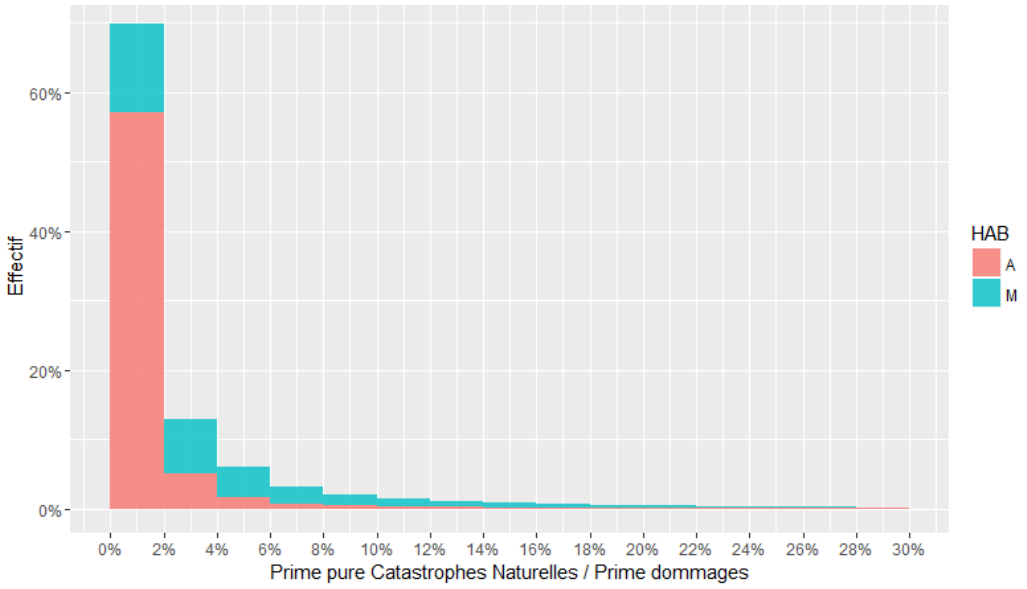

FIGURE 3.2 – Pourcentage de la prime pure Catastrophes naturelles de la prime dommages

Cependant comme on l'a vu lors de la calibration des GLM, il est primordial de distinguer entre appartement et maison. On voit que près de la moitié des appartements ont une prime pure estimée inférieure à 2% de la prime pure dommages. Alors que pour les maisons nous avons une décroissance moins marquée.

3.2.1 Analyse de la prime pure estimée des Maisons

La prime pure catastrophes naturelles représente 10% de la prime pure observée totale. Celà montre une sur exposition par rapport aux appartements pour qui elle ne représente que 2%. Nous devons donc porter une grande attention à l'analyse de cette prime pure estimée pour détecter les maisons risquées et déterminer l'origine de ce risque. Nous prenons pour référence de calcul la prime légale de catastrophes naturelles qui représente 12% de la prime dommages.

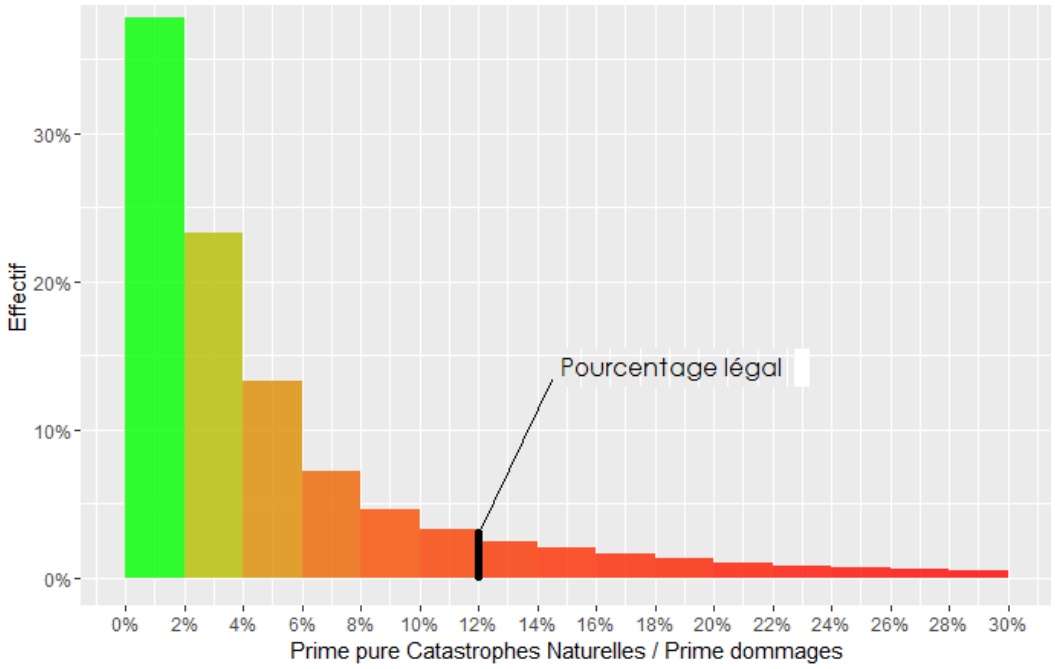

FIGURE 3.3 – Pourcentage de la prime pure Catastrophes naturelles de la prime dommages pour les maisons

La prime légale considère un seuil qui est justifié par cette répartition de la prime pure estimée. En effet, dans 90% des cas, le pourcentage estimé est inférieur à 12%. Le seuil légal est donc prudent tout en restant raisonnable.

On remarque que pour les contrats qui ont un pourcentage important (plus de 90% de la prime dommages), c'est la part de submersion marine qui en est la cause. En effet ces contrats sont littéralement en bord d'océan. On a aussi remarqué qu'une majorité de ces contrats sont placés à la limite des plages de marées basses et hautes.

Afin de repérer si certaines zones sont plus risquées de façon agrée, nous représenterons le pourcentage de la prime pure estimée totale par département. La formule est la suivante :

$$\text{Pourcentage Département} = \frac{\sum_{i \in contrats} \text{Prime pure estimée}_i}{\sum_{i \in contrats} \text{Prime Dommages}_i}$$

Si on faisait ce calcul sur la france entière, l'ideal serait d'arriver à un pourcentage inférieur de peu à 12%. De cette maniére la prime légale sera à la fois prudente et raisonnable.

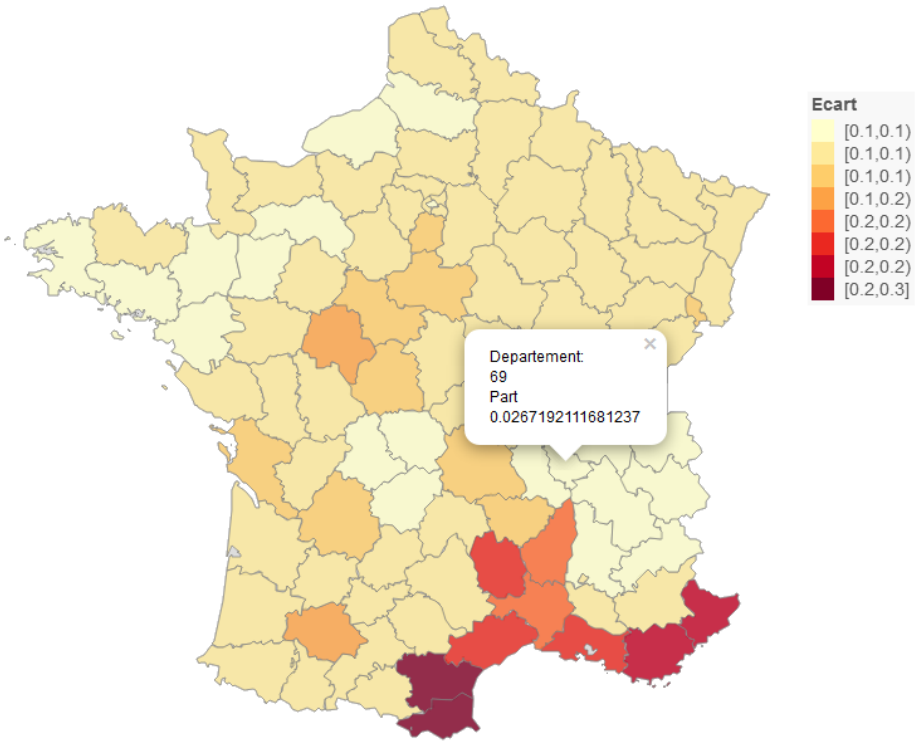

FIGURE 3.4 – Pourcentage de la prime pure Catastrophes naturelles de la prime dommages pour les maisons agrégées par département

La région sud-est est la région qui a les plus forts pourcentage. Comme on l'a vu lors de la prime pure inondation, cette région est particulièrement touchée. la prime pure inondation y est donc supérieure au reste de la France.

3.3 Un écart de prime impactant la marge

3.3.1 Gestion de la marge sur les contrats "sur-assurés"

Comme on l'a définit dans la partie 2, les contrats sur-assurés représentent x% de notre portefeuille. Nous avons donc une partie de la prime Catastrophes Naturelles qui est "non utilisée".

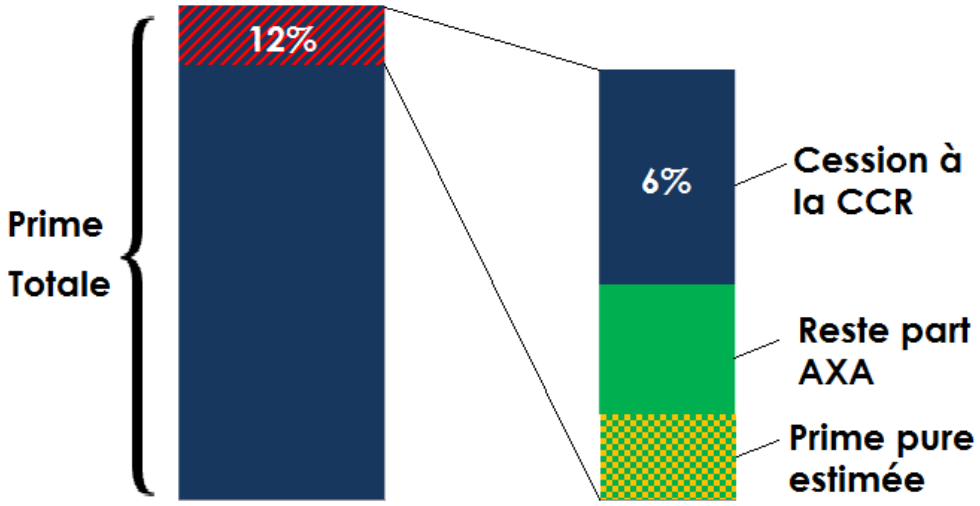

FIGURE 3.5 – Répartition de la prime pour les contrats "sur-assurés"

Afin de faire une optimisation vis à vis de la marge, cette part "non travaillante" nous permet donc de moduler notre marge. En effet, ayant détecté ce profil de contrats, nous pouvons agir sur notre marge afin d'attirer ces profils dans notre portefeuille. Ainsi avec ce mécanisme, la prime collectée pour les catastrophes naturelles reste la même mais le tarif TTC a été ajusté afin de convertir ces profils. Cependant cette partie, sera peu probablement mise en production. Le but principal de ces cartes étant le suivi et la prévention.

3.3.2 Gestion des contrats les plus risqués

L'estimation de la prime pure nous permet de détecter les contrats les plus exposés aux risques. Cette détection donne accès à la visualisation et l'appréhension du degré de risque et de la qualité de notre portefeuille en termes de catastrophes naturelles. D'autre part le fait d'avoir une géolocalisation de ces risques, permet d'avoir des mesures préventives plus poussées et surtout mieux ciblées. Ainsi, nous proposons à l'équipe chargée d'envoyer les alertes préventions de mettre en place un module de prévention des risques de catastrophes. Ce module aura donc pour cible les contrats risqués et enverra un message personnalisé.

Chapitre 4

Estimation d'un indicateur de performance

« l'actuaire doit préciser les données, les hypothèses et les méthodes utilisées. Il doit faire ressortir la sensibilité des résultats aux hypothèses et aux choix de modélisation. »

Code de déontologie de l'Institut des Actuaires

Sommaire

4.1 Expected Loss Ratio (ELR) 56
4.2 Analyse de l'ELR 56
4.3 Arbre de classification et régression (CART) 59
4.3.1 Le partitionnement et arbre maximal 60
4.3.2 Élagage de l'arbre 61
4.4 Calibration de l'Arbre des ELR 65
4.4.1 L'arbre Tronc 65
4.4.2 L'arbre maximal 65
4.4.3 L'arbre optimal 66
4.4.4 L'arbre final 67
4.5 Arbre des ELR 68
4.6 Utilisation opérationnelle de l'arbre des ELR 70

4.1 Expected Loss Ratio (ELR)

Après la création d'un modèle de prime pure, il en découle un indicateur de suivi. Cet indicateur est le EXPECTED LOSS RATIO (ELR). Cette mesure peut être utilisée dans le suivi de la profitabilité de segments de notre portefeuille.

L'ELR est un indicateur pour les 12 prochains mois. Ceci est dû au fait que nos bases d'observations sont faites année par année. Les ELR permettent de calculer la rentabilité d'un segment dès l'affaire nouvelle alors que le LR se calcule une fois le risque en portefeuille. L'analyse des ELR nous permet donc d'être plus réactif sur les nouveaux risques en portefeuille. Un autre avantage de L'ELR, par rapport au Loss Ratio (LR) observé, est que la volatilité est moindre pour des segments. Cette réduction de volatilité est due à la mutualisation faite par le modèle de prime pure.

L'ELR est défini de façon simple :

$$\text{ELR} = \frac{\text{P}rime\ \text{P}ure\ estimée}{\text{P}rime\ \text{P}ayée\ actuelle}$$

Dans notre cas, la prime payée actuelle est représentée par les 12% de la prime dommages. Cet indicateur est plus parlant qu'une comparaison du taux Prime pure par Prime dommages et du taux légal de 12%.

Notre but dans cette partie est de prédire un ELR par segment en utilisant une technique de classification non supervisée. On introduira des variables en plus dans cette étude afin de voir si certains comportements peuvent influer sur cet indicateur. Les variables supplémentaires qui sont introduites sont utiles dans la compréhension du risque mais ne peuvent être utilisées dans le calcul d'une prime pure.

- **PMptf** : c'est la prime commerciale réellement payée par l'assuré. Cette prime prend donc compte des éventuels rabais accordés
- **RatioOpt** : le pourcentage de la prime payée qui est représentée par les options telles que le rééquipement à neuf, dommages électriques, etc
- **Zone_P_C** : la zone de submersion marine
- **Zone_P_S** : la zone de sécheresse
- **Zone_P_F** : la zone d'inondation (croisement entre la carte d'inondation et la carte des précipitations)

4.2 Analyse de l'ELR

L'analyse de l'ELR, nous permettra dans cette partie de comprendre la répartition du risque en fonctions de certaines variables clés. A partir de cette analyse, nous souhaitons

Moyenne	Q25	Q50	Q75	Q90	Ecart type
0.360	0.021	0.077	0.329	0.966	0.838

TABLEAU 4.1 – Éléments statistiques sur l'ELR

identifier les profils les plus à risques. Nous faisons dans un premier temps une analyse uni-variée afin de repérer l'influence globale de certaines variables sur le risque. Cette analyse nous servira aussi à interpréter et valider l'agrégation des trois modèles.
Une première analyse de cet ELR purement catastrophes naturelles le positionne au niveau moyen d'un ELR toutes garanties confondues (en moyenne l'ELR est à 0.40 sur les garanties dommages hors Catastrophes Naturelles). Cependant, on voit que l'écart type de cette valeur est grand. Deux explications peuvent avoir lieu : la prime pure estimée est non adéquate à ce type de risque ou la prime actuellement payée est faible pour certains profils.

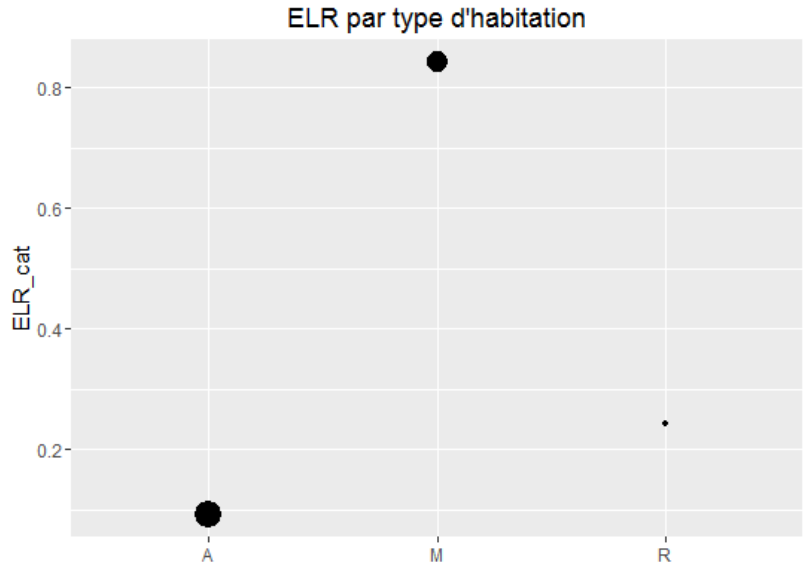

FIGURE 4.1 – Les maisons sont presque à un ELR neutre

Dans ce graphique, la taille des points est proportionelle à l'exposition dans le portefeuille. Un fort effet se distingue à partir de ce graphique. En effet, une grande partie de l'ELR est portée par les maisons et les Rez-de-chaussée. Cela s'explique notamment par le fait que ce type d'habitation est par nature plus exposé aux risques catastrophes naturelles. les appartements en étage sont par exemple moins sujets aux inondations. Le risque inondation est pratiquement nul sur ce type de bien. Par contre la vulnérabilité face à la sécheresse reste plus ou moins la même.

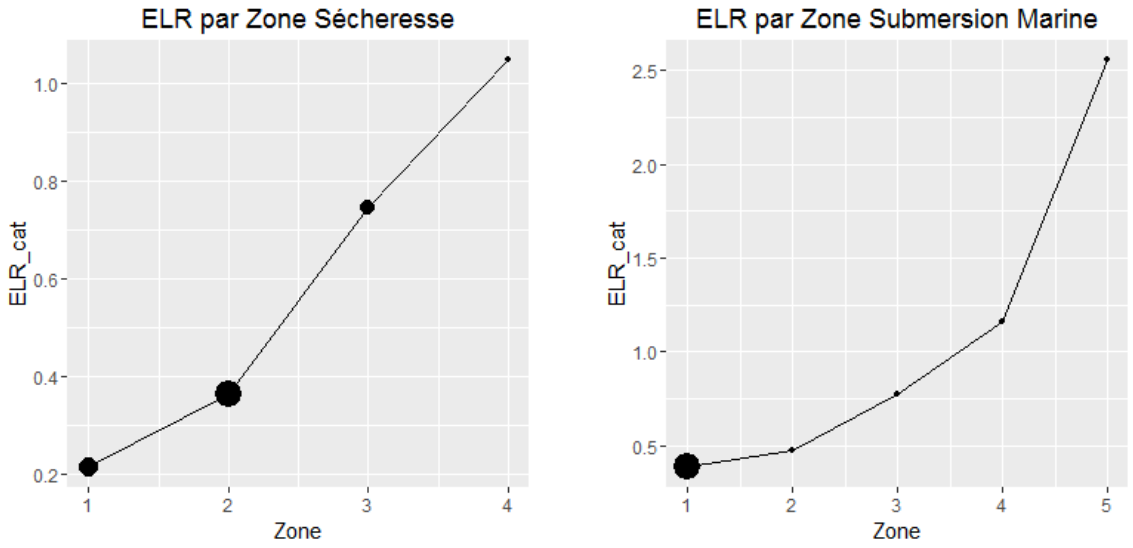

FIGURE 4.2 – La tendance par zone est respectée pour les deux risques

Les ELR par zone de sécheresse semblent bien suivre la tendance attendue. La taille des points étant relative à l'exposition dans la zone, on a aussi une cohérence vis à vis de la répartition sur le territoire. Cela nous rassure quant à la faible exposition en zone de fort risque. Le risque sécheresse pouvant causer des dommages très considérables (lourds frais d'expertise, reconstruction du bâtiment). D'autre part les ELR par zone de submersion marine est très largement concentré l'ELR à 0.36 (la moyenne). Cela est bien sur du au fait que ce risque ne peut concerner qu'une très faible partie du territoire. On observe ici la croissance de l'ELR par zone. On s'aperçoit aussi que pour ce risque l'ELR est très haut à partir de la troisième zone. On devrait donc ajuster le modèle pour réduire la prime pure prédite. Cette manipulation se justifierait par le fait que ce type d'évènement est très rare (revoir la période de retour de cet évènement).

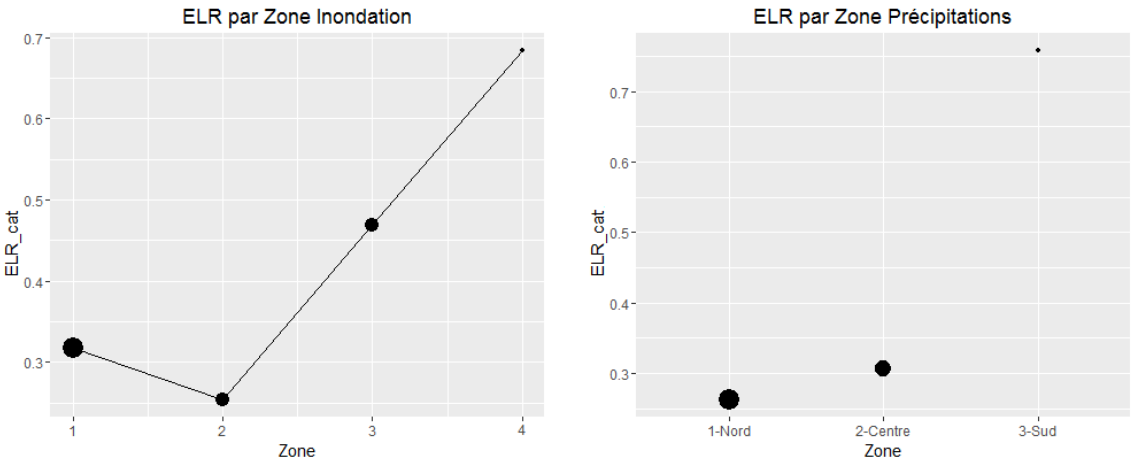

FIGURE 4.3 – Un irrégularité sur la tendance des zones d'inondations

Sur les deux cartes, on a une tendance de croissance générale. Cependant on remarque nettement une irrégularité pour les deux premières zones de la carte d'inondation. L'ELR ne suit pas la tendance attendue. Nous cherchons donc à expliquer cet écart. Nous es-

sayons de vois si à partir d'une analyse à deux variables on peut dégager l'élément qui biaise la tendance. On rajoute pour cela une différentiation par type d'habitation à la tendance par zone. Nous avons vu auparavant que cette variable avait un fort impact sur l'ELR.Un déséquilibre dans la répartition du type de bien sur les zones d'inondation entraînerait un biais sur la tendance globale par zone.

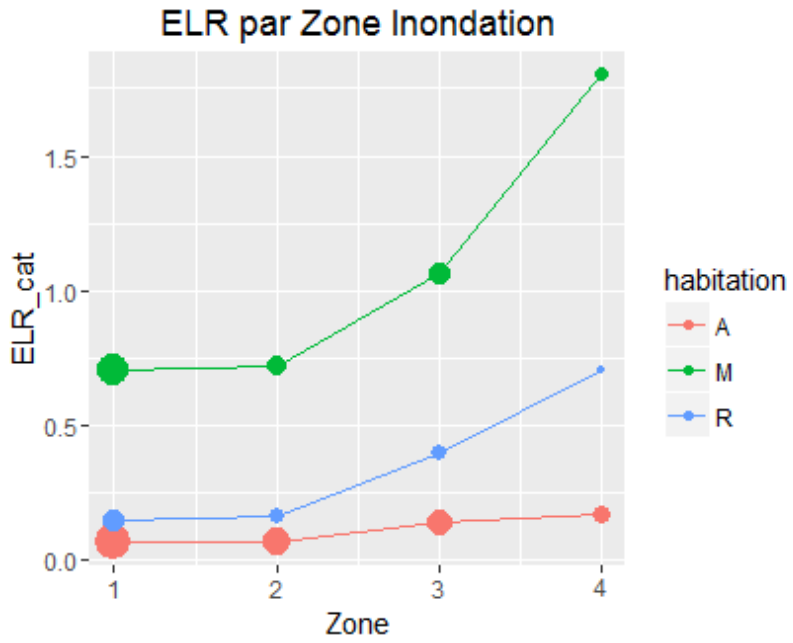

FIGURE 4.4 – L'irrégularité était principalement portée par le biais du type d'habitation

On voit qu'en séparant l'ELR par type d'habitation, la tendance est clairement identifiée. On retrouve la croissance de l'ELR en accord avec la croissance du risque par zone. On a par ailleurs des ELR supérieurs à 1 pour les maisons en zones 3 et 4. Nous pouvons expliquer le biais à partir de la taille des points. En effet on voit que pour la zone 1, les maisons et appartements ont une exposition à peu près équivalente. Pour la zone 2, l'exposition des appartements est plus grande (donc l'ELR aura tendance à baisser).

Cette analyse nous a permis de visualiser la tendance de l'ELR par variable. Nous nous intéressons maintenant à l'établissement de profils de risques. Pour cela nous avons choisi la méthode des arbres de décisions.

4.3 Arbre de classification et régression (CART)

Dans cette étude, nous utilisons les définitions de la publication de *Breiman L.*. Les arbres de classification servent à expliquer une variable catégorielle (exemple : la valeur du client, le type d'habitation, la qualité, ...). Les arbres de régression servent à expliquer une variable quantitative (par exemple : la charge, le capital, le montant d'objets de valeurs, ...).

Nous nous intéressons dans cette partie à étudier les ELR. On souhaiterait en effet détecter les personnes les plus à risque. Pour cela, nous estimons les ELR pour nos contrats et nous tentons d'expliquer cette quantité par un arbre.

4.3.1 Le partitionnement et arbre maximal

L'idée principale de la création d'un arbre de décision est le partitionnement. Á chaque étape, le partitionnement doit créer deux sous ensembles de données les plus homogènes possible. C'est un double processus. En effet, en premier on doit trouver la variable qui permet d'avoir deux sous ensembles les plus homogènes. Une fois cette variable choisie, l'algorithme doit choisir la façon de découper par rapport à cette variable. Si cette variable est continue, l'algorithme doit choisir la valeur (quantitative) de découpage. Si cette variable est qualitative, l'algorithme doit choisir l'ensemble de modalités qui sépare le mieux.
On définit :

- t_p : le noeud parent
- t_d : le noeud fils droit
- t_g : le noeud fils gauche
- t_f : les noeuds fils en général
- x_j : la variable qui sépare le mieux le jeu de données du noeud parent
- x_j^* : la valeur de la variable qui sépare le mieux

Les étapes élémentaires à chaque itération de l'algorithme peuvent être schématisées par :

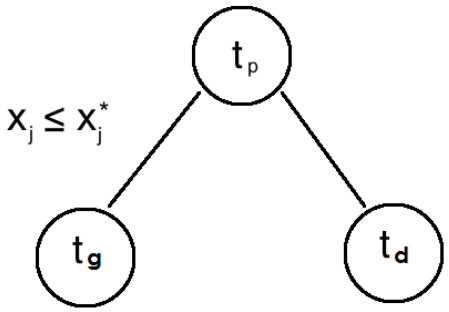

FIGURE 4.5 – Représentation d'un noeud

Afin d'obtenir la meilleure homogénéité, une fonction d'impureté est introduite. Notons cette fonction i. La valeur de cette fonction pour le noeud parent est $i(t_p)$. Nous cherchons maintenant a obtenir la variable qui minimise l'impureté totale des deux noeuds fils.

$$\min_{x_j \leq x_j^*} \mathbb{E}(i(t_e))$$

le concept de réduction d'impureté peut être appréhendé comme sur le graphique suivant.

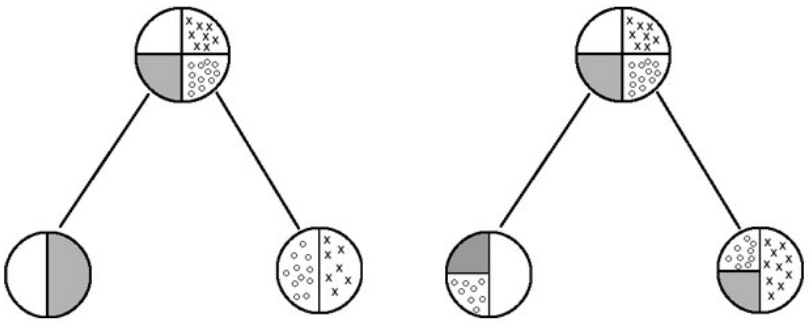

FIGURE 4.6 – Représentation de la réduction d'impureté

Nous pouvons visuellement concevoir ce principe de réduction. En effet le découpage gauche semble donner une meilleure homogénéité.

Le problème de minimisation peut se réécrire de la façon suivante

$$\min_{x_j \leq x_j^*} \mathbb{P}\, i(t_g) + \mathbb{P}\, i(t_d)$$

la solution de ce problème nous donnera le couple d'information qui permet de mieux découper les données à cette étape. Le couple est représenté par :

— x_j : la variable qui découpe le mieux la base parent

— x_j^* : la valeur qui permet de séparer au mieux

Dans notre cas, la variable Y à expliquer est une variable continue. Nous pouvons alors utiliser la variance comme fonction d'impureté. ainsi le problème se réécrit :

$$\min_{x_j \leq x_j^*} \mathrm{P}_g\, \mathrm{V}ar(\mathrm{Y}_g) + \mathrm{P}_d\, \mathrm{V}ar(\mathrm{Y}_d)$$

Pour créer notre arbre de décision, nous utilisons cette étape élémentaire autant de fois que nécessaire, jusqu'à l'obtention de l'arbre maximal. Cet arbre est celui dont les feuilles comportent des individus identiques (au sens des variables x_j)

4.3.2 Élagage de l'arbre

Importance de l'élagage

L'élagage fait référence à la taille de l'arbre qu'on souhaite utiliser. Dans la première étape, nous créons l'arbre maximal. Cet arbre est très utile pour expliquer les données. Sauf que sa taille est tellement grande que l'interprétation est impossible. Il faut donc trouver la taille adéquate pour une interprétation facile. La taille joue aussi beaucoup sur la phase prédiction :

— Une grande taille sur-apprend sur la base d'apprentissage

— Une très petite taille d'arbre impliquerait une prédiction non précise.

L'étape d'élagage joue un rôle important tant dans l'interprétation de l'arbre que l'ajustement sur les données d'apprentissage.

les sous-arbres

Afin de définir proprement le processus d'élagage, nous définissons la notion de sous-arbres. Nous donnons les définitions suivantes :

— Descendant : un noeud t' est descendant du noeud t s'il existe un chemin de l'arbre T du noeud t à t'

— Branche : une branche T_t de l'arbre T est le noeud t avec tous ses descendants.

— Sous-arbre : le sous arbre T′ est l'arbre T dont une branche a été élaguée. (exemple : $T' = T - T_t$

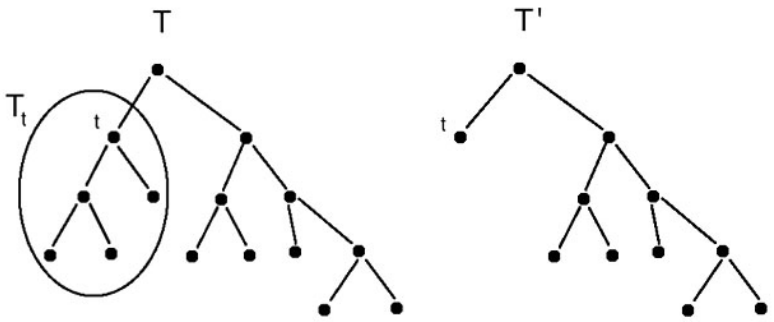

FIGURE 4.7 – Élagage de l'arbre T en enlevant la branche T′

le processus

Le processus d'élagage doit déterminer la taille optimale de l'arbre. La première étape, consiste à caractériser la taille de l'arbre. Comme le propose *Breiman L.*, l'arbre peut être caractérisé par le nombre de noeuds finaux. On sait que l'arbre maximal possède un grand nombre de noeuds finaux et chaque noeud est totalement pur (la fonction d'impureté est nulle). D'un autre côté, si on considère la racine de l'arbre (le premier noeud), nous avons un nombre de noeuds minimal et une impureté maximale. Le processus d'élagage permet de choisir l'arbre optimal en faisant un arbitrage entre le nombre de noeuds finaux et la fonction d'impureté. Cet arbitrage est géré par la fonction de complexité. Cette fonction est définie par :

$$R_\alpha(T) = \sum_{k=1}^{|T|} R(k)P(k) + \alpha \mid T \mid$$

Avec :

— $|T|$: nombre de noeuds finaux du sous arbre T

— $P(k)$: probabilité d'être dans le noeud k

— α : le paramètre de complexité

— $R(.)$: la fonction d'erreur de classification définie par

$$R(k) = \sum_{m=1}^{|T|} i(m)P(m)$$

On définit T_{max} l'arbre maximal. L'objectif est de trouver l'arbre $T(\alpha)$ qui minimise la fonction de complexité :

$$R_\alpha(T(\alpha)) = \min_{T \leq T_{max}} R_\alpha(T)$$

$T \leq T_{max}$ signifie que T est une version élaguée de T_{max}.

Le point de départ de l'algorithme est $\alpha_1 = 0$ et T_1 le plus petit sous-arbre qui vérifie $R(T_1) = R(T_{max})$. Pour obtenir l'arbre T_1, nous commençons par élaguer tous les noeuds où $R(t) = R(t_g) + R(t_d)$

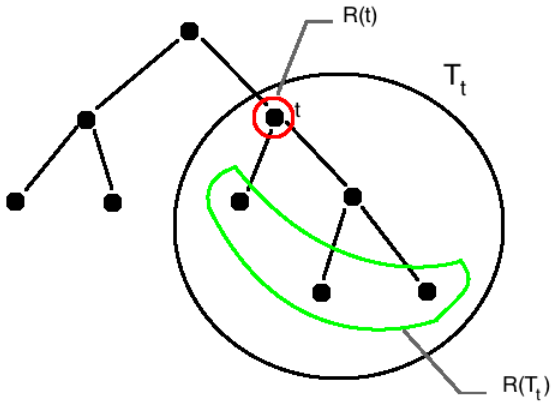

FIGURE 4.8 – Les noeuds considérés pour la fonction d'erreur de classification

Pour la T_t une branche de T_1 on définit $R(T_t)$ par :

$$R(T_t) = \sum_{t \in \widetilde{T_t}} R(T)$$

Avec $\widetilde{T_t}$: les noeuds finaux de T_t.

Nous obtenons donc les deux équations suivantes : $\begin{cases} R_\alpha(t) = R(t) + \alpha \\ R_\alpha(T_t) = R(T_t) + \alpha\,|\,\widetilde{T_t}\,| \end{cases}$

D'après la construction de la branche T_t nous avons que :

$$R_\alpha(t) > R_\alpha(T_t)$$

Alors :

$$\alpha < \frac{R_\alpha(t) - R_\alpha(T_t)}{|\,\widetilde{T_t}\,| - 1}$$

On définit la fonction $g_1(.)$ par la partie droite de cette inégalité :

$$g_1(t) = \frac{R_\alpha(t) - R_\alpha(T_t)}{|\widetilde{T_t}| - 1}$$

Dans son papier, *Breiman* définit le plus faible lien par t_1. Ce noeud est celui qui minimise la fonction g_1 sur l'arbre T_1.

$$g_1(t_1) = \min_{t \in T_1} g_1(t)$$

On définit ensuite $\alpha_2 = g_1(t_1)$ et $T_2 = T_1 - T_{t_1}$

Ce processus est répété ensuite sur l'arbre T_2 et ainsi de suite jusqu'à la racine $T_{n_p} = \{t_p\}$. Ce processus permet la création d'une série croissante $(\alpha_i)_{i=1...n_p}$ et des arbres $(T_i)_{i=1...n_p}$. La prochaine étape est donc de choisir l'arbre optimal à partir des arbres créés.

Choix de l'arbre optimal par cross-validation

Pour le choix du meilleur arbre, nous allons utiliser la cross-validation (le principe est défini à la section de calibration du GLM). Ici la base initiale $\mathscr{L}$ est coupée V sous ensemble de données indépendants de la même taille qu'on notera $\{\mathscr{L}_v\}_{v=1...V}$. On crée ensuite nos bases $\mathscr{L}^{(v)} = \mathscr{L} - \mathscr{L}_v$.

L'idée principale est de créer V arbres maximaux T^v_{max} en utilisant les base $\mathscr{L}^{(v)}$. Le reste de la base (la partie $\mathscr{L}_v$) sert à calculer l'erreur de classification. Cette erreur est définie par :

$$R^{CV}(T(\alpha)) = \frac{1}{V} \sum_{v=1}^{V} \frac{N^{(v)}_{err}}{N^{(v)}}$$

Avec :

— $N^{(v)}$: le nombre d'individus dans la base d'apprentissage $\mathscr{L}^{(v)}$

— $N^{(v)}_{err}$: le nombre d'individus de la base de test $\mathscr{L}_v$ qui ont été mal classés par l'arbre élagué $T^{(v)}(\alpha)$

Afin de trouver l'arbre qui minimise l'erreur de classification sur la cross-validation, on réutilise les arbres $(T_i)_{i=1...n_p}$ et la série croissante des paramètres de complexité $(\alpha_i)_{i=1...n_p}$. On a que pour tout α il existe k tel que $\alpha_k \leq \alpha \leq \alpha_{k+1}$. Alors l'arbre correspondant à α est $T_k = T(\alpha_k)$.
On définit $\beta_k = \sqrt{\alpha_k \alpha_{k+1}}$ et l'erreur de classification de l'arbre T_k est :

$$R^{CV}(T_k) = R^{CV}(T(\beta_k))$$

On choisit maintenant l'arbre T_k qui a une erreur de classification minimale. L'arbre qui minimise cette valeur est donc l'arbre optimal pour expliquer nos données.

4.4 Calibration de l'Arbre des ELR

Dans cette section, nous allons détailler la méthode choisie pour la sélection de l'arbre.

Tout d'abord, nous gardons un échantillon d'apprentissage et un échantillon de test. L'échantillon de test ne sera jamais utilisé lors de la calibration de l'arbre. L'échantillon de test représente 10% de la base totale.

Après sélection de nos échantillons, nous devons définir une mesure nous permettant de comparer les modèles entre eux. Dans notre cas, le calcul de l'erreur quadratique moyenne (MSE - *Mean Squared Error*) convient pour comparer les modèles entre eux. Étant donné que nos échantillons (test et validation) n'ont pas le même nombre d'observations, nous proposons de calculer un MSE ajusté.

$$\text{MSE}_{adjust} = \frac{\sum_{i=1}^{n} (y_i - \widehat{y}_i)^2}{n}$$

avec :

— n : nombre d'observations de la base

— y_i : ELR estimé par le modèle GLM

— $\widehat{y}_i$: ELR estimé par l'arbre

4.4.1 L'arbre Tronc

La première étape avant d'entamer la calibration d'un arbre est de calculer la performance de l'arbre tronc. L'arbre tronc consiste en la moyenne de l'échantillon de validation. On attribue donc à toutes les observations la moyenne et on estime le MSE. Cette mesure nous permet d'avoir une erreur référence. On pourra ainsi mesurer le gain apporté par un modèle par rapport à la moyenne.

	MSE(app)	**MSE(test)**
Arbre Tronc	0.7011342	0.7152979

TABLEAU 4.2 – Mesures de l'erreur de l'arbre tronc

Les deux mesures sont comparables. Évidemment ces deux valeurs sont dépendantes du choix de l'échantillon. Cependant, les deux valeurs devraient rester proche.

4.4.2 L'arbre maximal

Nous nous intéressons dans ce cas à l'arbre maximal. Cet arbre nous donnera l'erreur minimale qu'on peut obtenir sur un échantillon donné. Bien sur cette erreur n'est minimale que sur l'échantillon sur lequel l'arbre a été calibré.

	App	Test
MSE	0.185322	0.2296301
Gain	73.6%	67.9%

TABLEAU 4.3 – Mesures pour l'arbre maximal

Le gain apporté par l'arbre maximal est donné par le 73.6%. On s'aperçoit aussi que l'arbre maximal prédit plutôt bien sur l'échantillon test. La principale raison pour cette adéquation est que les ELR modélisé par GLM prennent en compte un nombre faible de variables. Or ces combinaisons vont être retraduites par l'arbre.
Il est a noter que lors de plusieurs tests (changement de la base d'apprentissage et test), nous avons constaté des valeurs plus élevées pour le MSE de l'échantillon test. Ainsi le gain est beaucoup plus faible sur l'échantillon test.

4.4.3 L'arbre optimal

Afin d'obtenir l'arbre Optimal, nous utilisons l'algorithme décrit à la section précédente. On affiche dans le tableau suivant l'erreur relative de cross-validation en fonction du paramètre de complexité (l'équivalent de α).

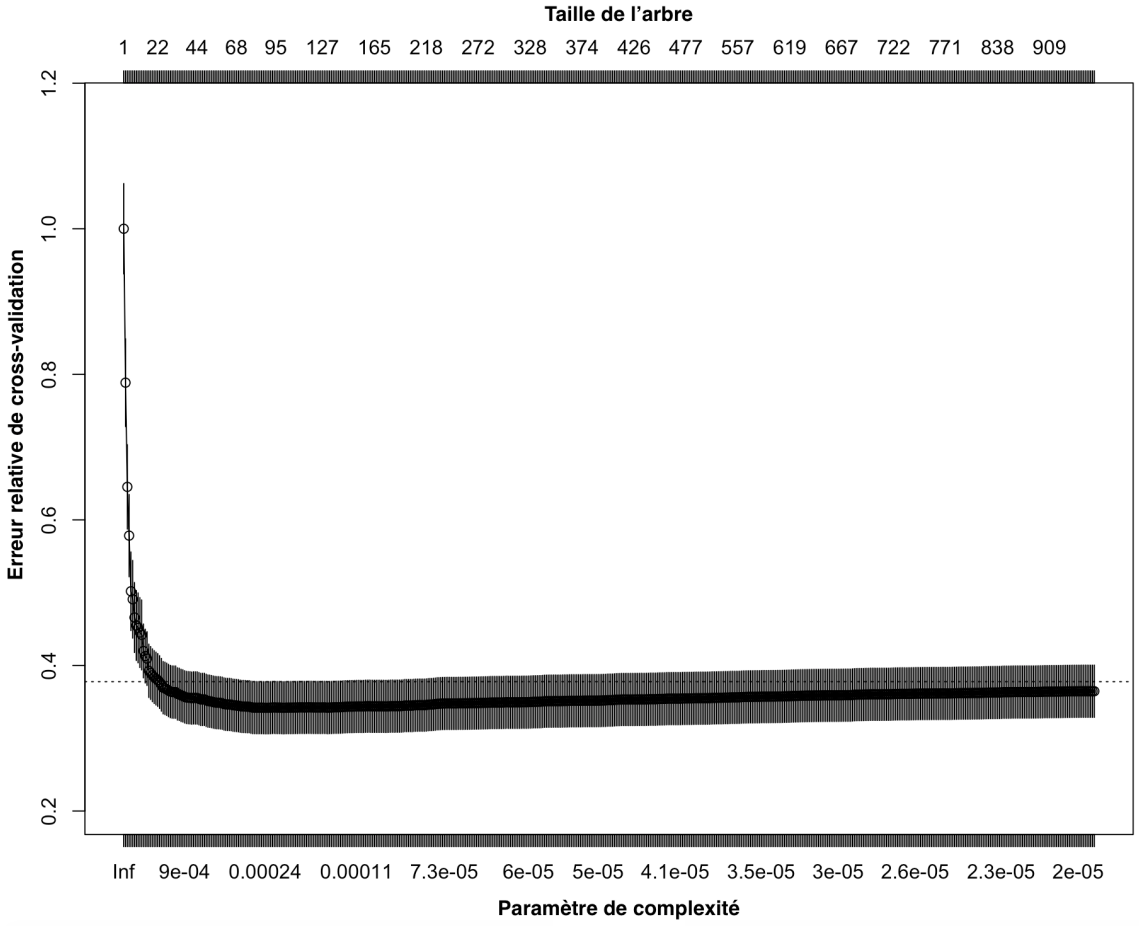

FIGURE 4.9 – Erreur relative de cross-validation en fonction du paramètre de complexité

A partir de α = 0.00012, l'erreur relative commence légèrement à remonter. Il s'agit

du minimum d'erreur relative. Cette courbe est caractéristique de la calibration des modèles en *Machine Learning*. En effet, au début plus on segmente plus le modèle prédit mieux. Ensuite, le modèle rentre dans une phase de sur-apprentissage sur la base. Ce sur-apprentissage se traduit sur la courbe par une croissance de l'erreur relative.

Ainsi, en prenant le paramètre de complexité qui minimise l'erreur relative nous obtenons l'arbre Optimal. Cependant, cet arbre possède 135 feuilles. Il est donc utile pour la prédiction mais dans notre cas il sera dur à utiliser pour comprendre le risque. En effet, notre but dans cette partie est de représenter un arbre et de voir visuellement.

En termes de performances de prédiction nous obtenons les résultats suivants :

	App	**Test**
MSE	0.2114202	0.2107911
Gain	69.8%	70.5%

TABLEAU 4.4 – Mesures pour l'arbre optimal

On voit que l'arbre optimal, permet une meilleure stabilité du modèle face à l'échantillon de test. En effet, le MSE est presque stable avec cet arbre. Certes le gain un peu moindre sur l'échantillon d'apprentissage mais il est plus stable sur l'échantillon test.

4.4.4 L'arbre final

Dans notre cas, l'arbre que l'on souhaiterait obtenir doit pouvoir se représenter. Nous cherchons a obtenir un visuel qui soit intuitif et facile à lire. En effet, nous ne cherchons pas a prédir les ELR avec l'arbre mais à les expliquer. Ce sont les modèles GLM qui nous donnent l'estimation des ELR.

Afin de choisir l'arbre final, nous décidons de nous concentrer sur les premières itérations de l'algorithme. Nous représentons l'erreur relative des premiers arbres. (Il s'agit ici d'un zoom sur la partie gauche de la figure 4.9)

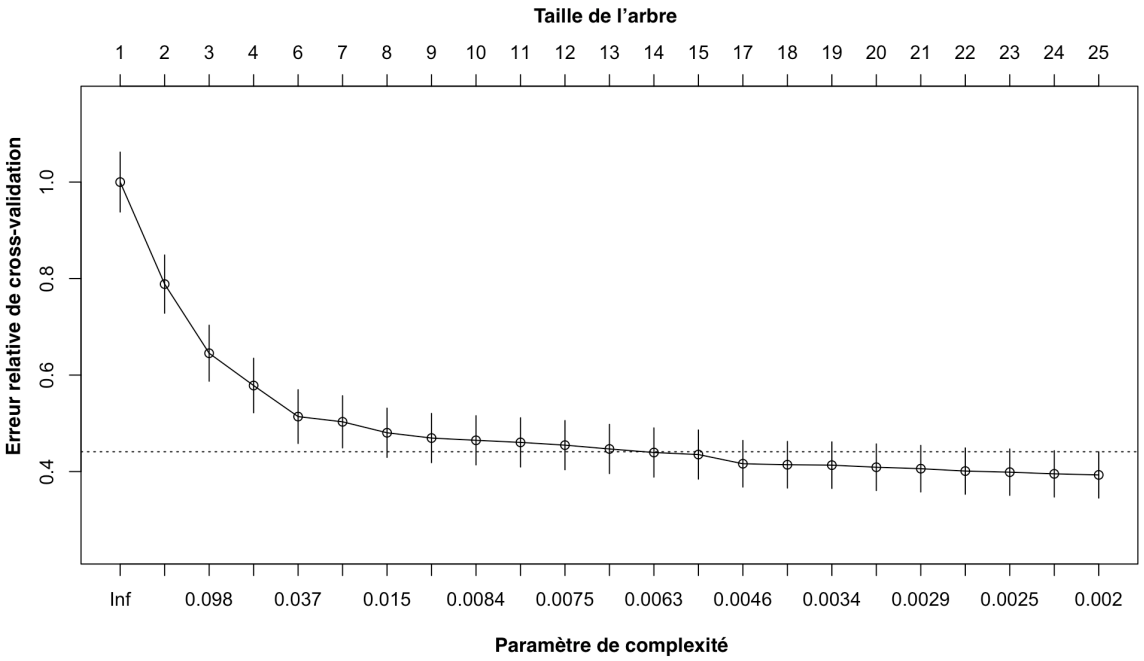

FIGURE 4.10 – Erreur relative de cross-validation en fonction du paramètre de complexité premières itérations

Nous allons utiliser l'arbre à 13 feuilles pour notre représentation. Le choix est ici arbitraire. Nous estimons alors sa performance. Dans ce cals on aura donc une perte de performance comparée à l'arbre optimal. En effet, on choisit un arbre avec une erreur relative qui est supérieure.

	App	**Test**
MSE	0.2643803	0.2788704
Gain	60.2%	63.0%

TABLEAU 4.5 – Mesures pour l'arbre final

On mesure bien ici une décroissance du pouvoir explicatif de l'arbre. En effet on passe d'un gain au alentours de 70% à un gain à %60. Cependant, on garde quand même un gain explicatif considérable et on a aussi un arbre qui est représentable visuellement. Nous représentons cet arbre par la suite.

4.5 Arbre des ELR

Nous utilisons la théorie des arbres de décisions sur les ELR calculés. La technique est utilisée dans le but d'identifier la typologie des contrats les plus risqués.

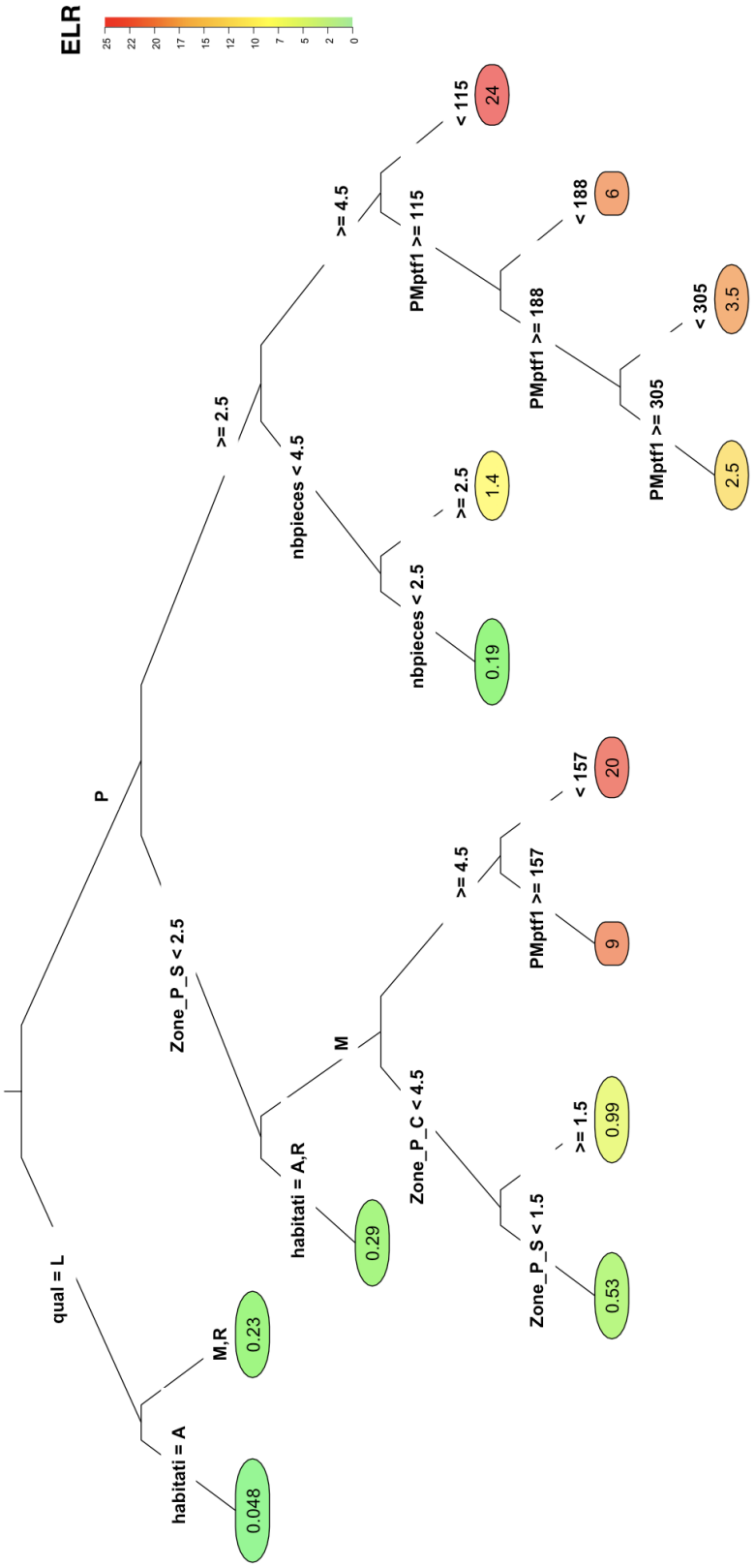

FIGURE 4.11 – Arbre final des ELR Cat

La première variable la plus discriminante est la qualité d'occupation. Les ELR dans la catégorie des locataires sont plutôt faibles (hormis les maisons). La branche des

propriétaires est, elle, plus segmentée. C'est cette branche qui est plus importante. En effet, le risque de catastrophes naturelles est beaucoup porté sur ces derniers.

Sans surprise, les contrats avec les forts ELR sont les maisons avec un grand nombre de pièces et situées dans les zones les plus risquées. Cependant, l'arbre fait apparaître quelques cas particuliers. En effet, la prime payée (PMptf) apparaît comme variable discriminante et permet de détecter des contrats avec une prime plutôt faible pour le risque qu'ils représentent. Par exemple sur la feuille la plus extrême , nous avons des contrats de maisons avec plus de 5 pièces avec un tarif annuel inférieur à 115€. Certes le nombre de contrats dans cette catégorie est très faible, mais il permet de détecter des cas de sous tarification.

Cet arbre nous donne aussi une indication sur l'importance de chacune des cartes de risques. En effet elles sont parmi les premières variables utilisées par l'algorithme. Cependant il faut relativiser ce résultat. En effet, la prime pure catastrophes naturelles est calculée à partir des trois GLM dans lesquels le spread de la variable géographique est l'un des plus importants.

4.6 Utilisation opérationnelle de l'arbre des ELR

L'arbre des ELR a pour but de proposer une technique de suivi du portefeuille. Cet arbre nous permet de visualiser l'exposition de notre portefeuille dans les différentes branches. En plus de proposer une vision ponctuelle, l'étude est amené à être relancée de manière périodique afin de suivre l'évolution du portefeuille. Ce suivi permettra d'être réactif face à une surexposition en branches très risquées. Le calcul des ELR va aussi permettre de comparer le risque catastrophes naturelles que l'on a actuellement sur le portefeuille et celui des affaires nouvelles.
D'autre part, cet arbre permet aussi de surveiller et détecter certains cas de sous tarification avérée. Même si ce n'est pas le but principal de cet outil, la détection de tel contrats est intéressante.

Conclusion

La tarification des catastrophes naturelles est un sujet de recherche large. Dans ce ce mémoire, nous avons essayé d'appliquer les techniques classiques de l'actuariat non-vie. Cette approche nous permet d'estimer une prime pure par contrat. Le principal apport de cette étude est la mise en évidence de la possibilité d'utiliser des données purement géologiques dans l'estimation du risque. La manipulation de données géologique implique une utilisation de systèmes d'information géologiques. L'apport des manipulations géographiques est utile lors de la construction des zoniers hors délimitations géographiques.

Cette étude a eu pour but d'estimer une prime pure catastrophes naturelles. Nous avons segmenté les catastrophes en trois risques majeurs ce qui a permis une modélisation plus fine des risques. L'apport de nouvelles cartes a permis d'avoir une meilleure stabilité des modèles. Cette modélisation a surtout permis de connaître l'exposition face aux catastrophes naturelles. L'étude comparative par rapport à la prime légale, nous a permis de nous rassurer sur le fait que cette prime est suffisante pour couvrir les risques.

Nous pouvons nous poser la question quant à la pertinence de l'utilisation des GLM pour la modélisation de ce risque particulier. En effet, on pourrait tester d'autres techniques de modélisation comme les arbres de décisions (CART, Random forest ou Gradient Boosting) qui pourraient être plus adaptée à un risque de si faible fréquence. Il est à noter aussi qu'en général, la tarification des catastrophes naturelles concerne la réassurance. Dans ce cas, la tarification est globale. Elle est basée sur des modèles de simulations. Certains outils du marché comme RMS ou EQECAT permettent d'avoir une modélisation de qualité. Cette tarification est globale pour le portefeuille. Elle n'était donc pas adéquate à notre problématique d'estimation à la maille contrat. Cependant, une vérification de la prime pure globale a permis de conforter notre estimation.

Bibliographie

[1] AXA BELGUIM (2013), Assurances & Actuariat IARD, Document interne.

[2] AXA FRANCE (2012),Guide technique MRH, Document interne.

[3] BESSE P., Introduction au modèle linéaire général, Université Toulouse.

[4] BREIMAN L., Machine Learning, University of California.

[5] CHARPENTIER A. (2005) Mathématiques de l'assurance non-vie - Tome II : Tarification et provisionnement, Economica.

[6] DUNCAN A. (2007), A practionner's guide to generalized linear models, Anderson , Towers Watson.

[7] GILCHRIST R. et DRINKWATER D. (1999) , The use of the Tweedie distribution in statistical modeling, University of North London, COMPSTAT 2000.

[8] LOIRET C. (2016), Refonte du tarif Multirisque Habitation : construction de microzoniers et intégration de la sinistralité passée à l'adresse, ISFA, Mémoire d'actuariat.

[9] QUIJANO O. et GARRIDO J. (2014), Generalised linear models for aggregate claims ;to Tweedie or not ?, Concordia University.

[10] YOO J-H. Introducing the generalized linear models, GLIS Consulting.

Table des figures

1.1 Maisons inondées à la suite de Xynthia . . . 7
1.2 Inondation de Paris, juin 2016 [1] . . . 8
1.3 Evolution du taux de résiliation MRH sur le marché français . . . 9
1.4 Taux de résiliation en fonction du motif . . . 10
1.5 L'apport d'un maillage très fin- Garantie Dégâts des eaux . . . 11
1.6 Coordonnées géographiques [2] . . . 12
1.7 Principaux types de projection . . . 13
1.8 Carte de France en RGF 93 . . . 13
1.9 Carte de France en projection Azimutale Nord . . . 13
1.10 Perte par département par la contrainte de géocodage . . . 17

2.1 représentation des courbes de Lorenz pour l'observé et un modèle . . . 24
2.2 Densité simulée d'une Tweedie avec $p = 1.3, \mu = 1, \Phi = 1$ . . . 26
2.3 Carte de suivi actuelle . . . 28
2.4 Carte Voronoi . . . 28
2.5 Carte de modélisation de période de retour 200 ans . . . 30
2.6 Carte de hauteur par rapport à la rivière la plus proche . . . 30
2.7 Carte de hauteur par rapport au drainage le plus proche . . . 31
2.8 Zoom sur le Sud-Est de la carte de risque inondation . . . 32
2.9 Une répartition du territoire raisonnable des risques inondations . . . 33
2.10 La moitié des contrats est située sur des zones de risques inondation . . . 33
2.11 Retrait et Gonflement à l'origine des fissures de murs . . . 34
2.12 Niveaux de sécheresse sur la France . . . 35
2.13 Niveaux de submersion marine au niveau de l'île d'Oléron . . . 36
2.14 Prime pure inondation par type d'habitation . . . 37
2.15 Prime pure inondation par type d'habitation et par année . . . 38
2.16 Prime pure inondation par part d'objets de valeur . . . 38
2.17 Prime pure inondation par part d'objets de valeur et par année . . . 39
2.18 Prime pure inondation par qualité d'occupant . . . 39
2.19 Prime pure inondation par qualité d'occupant et par année . . . 40
2.20 Prime pure inondation par zone de risque inondation . . . 40
2.21 Prime pure inondation par zone de risque inondation et par année . . . 41
2.22 Résultats de la méthode Forward sur nos données . . . 42

2.23 Un Indice de Gini stable 43
2.24 Une tendance consistante sur les données géographiques 44
2.25 Une forte volatilité des taux d'objets de valeur élevés 45
2.26 Une instabilité avérée de la variable 45

3.1 Répartition moyenne de la prime pure catastrophes naturelles estimée . . . 50
3.2 Pourcentage de la prime pure Catastrophes naturelles de la prime dommages 51
3.3 Pourcentage de la prime pure Catastrophes naturelles de la prime dommages pour les maisons 52
3.4 Pourcentage de la prime pure Catastrophes naturelles de la prime dommages pour les maisons agrégées par département 53
3.5 Répartition de la prime pour les contrats "sur-assurés" 54

4.1 Les maisons sont presque à un ELR neutre 57
4.2 La tendance par zone est respectée pour les deux risques 58
4.3 Un irrégularité sur la tendance des zones d'inondations 58
4.4 L'irrégularité était principalement portée par le biais du type d'habitation . 59
4.5 Représentation d'un noeud 60
4.6 Représentation de la réduction d'impureté 61
4.7 Élagage de l'arbre T en enlevant la branche T′ 62
4.8 Les noeuds considérés pour la fonction d'erreur de classification 63
4.9 Erreur relative de cross-validation en fonction du paramètre de complexité 66
4.10 Erreur relative de cross-validation en fonction du paramètre de complexité premières itérations 68
4.11 Arbre final des ELR Cat 69

A.1 Carte des précipitations 76
A.2 représentation d'un maillage de Voronoi 77

Annexe A

Annexes

A.1 Lexique

Coût moyen : charge totale des sinistres divisée par le nombre de sinistres
Fréquence : nombre de sinistres divisé par l'exposition totale au risque
Prime Pure : la fréquence multipliée par le coût moyen. Elle représente le coût d'une garantie pour couvrir les risques.

Zone : ensemble de délimitations géographiques présentant un même niveau de risque.
Shapefile : type de fichier géographique sous format vectoriel. Le fichier contient les limites (Département, région, zone,...). Ce format peut contenir les formes géométriques suivantes : lignes , points ou polygones.
Raster : type de fichier géographique sous format grille. le fichier peut être assimilé à une matrice. Chaque cellule contient une valeur (température moyenne , précipitations , hauteur par rapport à un certain niveau,...)
RGF 93 : Le Réseau Géodésique Français 1993. le RGF 93 est le référentiel spatial officiel de la France.
WGS 84 : World Geodetic System 1984. C'est le système géodésique le plus couramment utilisé. Le GPS est notamment basé sur ce système.
Géocodage : transformer l'adresse du risque en coordonnées.
Jointure spatiale : En utilisant les coordonnées d'un contrat, on rajoute l'information provenant de la carte (la zone).

GLM : abréviation pour modèle linéaire généralisé.
ELR : Expected Loss Ratio. c'est le ratio de perte espéré pour un certain contrat
Random Forest : technique de machine learning. La méthode utilise plusieurs arbres de régression ou classification avec pour chaque arbre un échantillon. La prédiction finale est l'agrégation des prédictions de chaque arbre de la foret.

A.2 Graphiques annexes

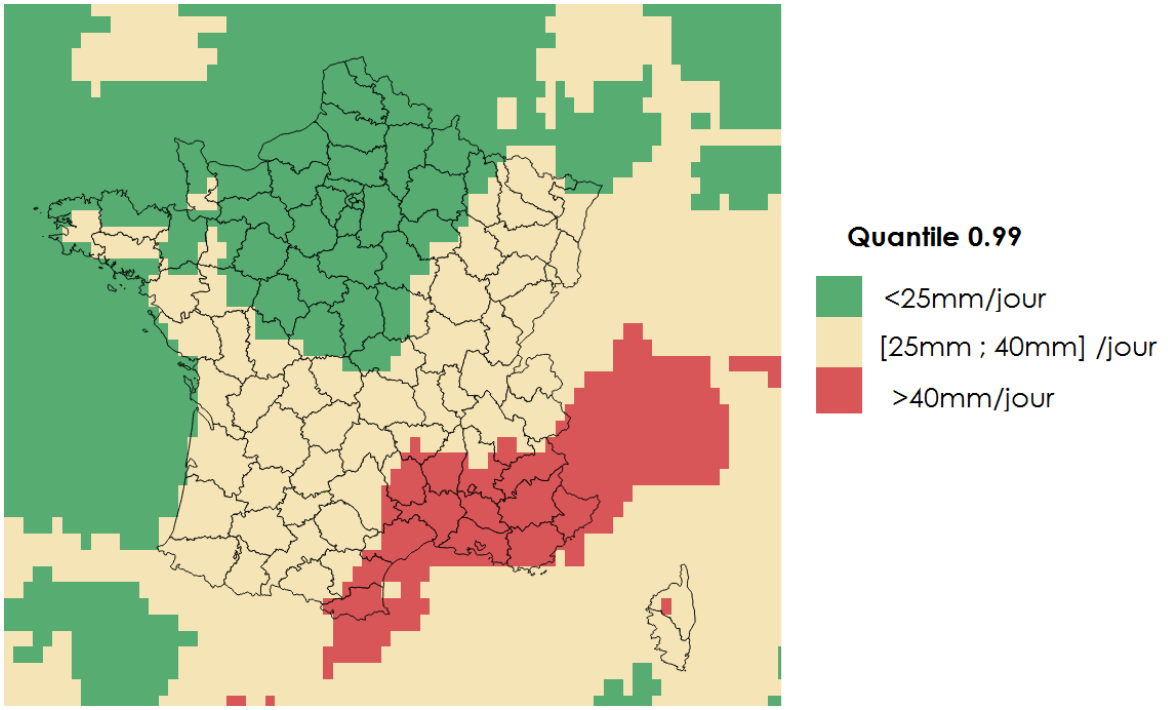

FIGURE A.1 – Carte des précipitations

A.3 Polygones de Voronoi

Les polygones de Voronoi sont utiles à l'analyse d'un ensemble de points répartis dans un espace de deux dimensions. Ces points sont appelés sites ou germes. Dans chacun des polygone, tout point est plus proche du site du polygones auquel il appartient que des sites des autres polygones.

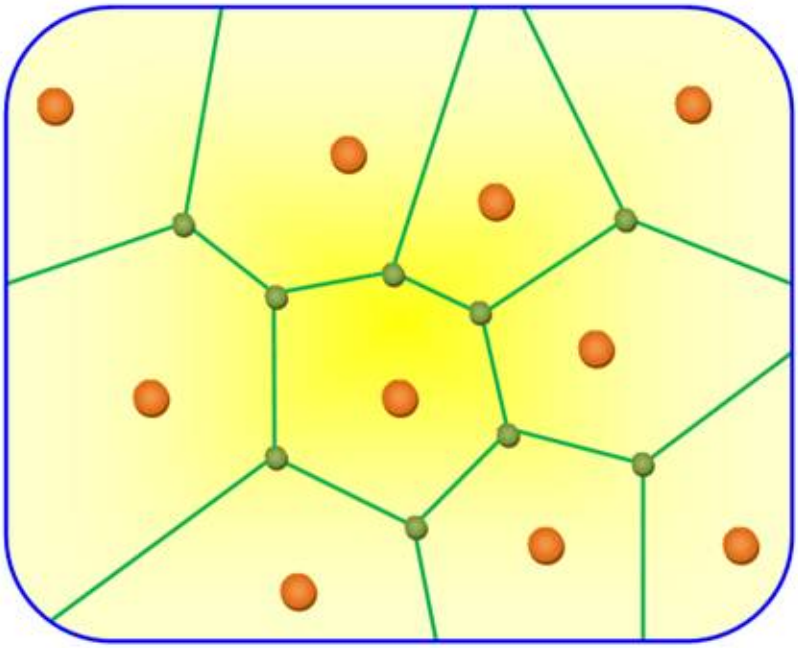

FIGURE A.2 – représentation d'un maillage de Voronoi

Mathématiquement cela se traduit par :

On dénote S un ensemble de points appelés sites ou germes dispersés sur un plan. Pour deux sites distincts $p, q \in$ S, le dominant de p sur q est défini par le plan au moins aussi proche de p que de q. En notant δ la distance euclidienne on a :

$$dom(p,q) = \{x \in \mathbb{R}^2 | \delta(x,p) \leq \delta(x,p)\}$$