Sorbonne Université
Mission Risques Naturels

École doctorale **École Doctorale Sciences Mathématiques de Paris Centre**

Unité de recherche **Laboratoire de Probabilités, Statistique et Modélisation**

Thèse présentée par **Antoine Heranval**

Soutenue le **23 septembre 2022**

En vue de l'obtention du grade de docteur de Sorbonne Université

Discipline **Mathématiques appliquées**

Spécialité **Statistique**

Contributions des données de l'assurance à l'étude des risques naturels

Application de méthodes d'apprentissage statistique pour l'évaluation de la nature et du coût des dommages assurés liés aux événements naturels en France

Thèse dirigée par Olivier Lopez directeur
Maud Thomas co-encadrant

Composition du jury

Rapporteurs	Arthur Charpentier	professeur à l'UQAM	
	Denys Pommeret	professeur à l'Institut de Mathématiques de Marseille Aix-Marseille Université	
Examinateurs	Caroline Hillairet	professeure à l'ENSAE - CREST	présidente du jury
	Olivier Wintenberger	professeur à Sorbonne Université	
	Philippe Naveau	directeur de recherche au LSCE CNRS	
Invité	Eric Petitpas	Mission Risques Naturels	
Directeurs de thèse	Olivier Lopez	professeur à Sorbonne Université	
	Maud Thomas	MCF à Sorbonne Université	

Colophon

Mémoire de thèse intitulé « Contributions des données de l'assurance à l'étude des risques naturels », écrit par Antoine Heranval, achevé le 22 décembre 2022, composé au moyen du système de préparation de document LaTeX et de la classe yathesis dédiée aux thèses préparées en France.

Mots clés : assurance, risques naturels, apprentissage statistique, valeurs extrêmes, analyse de texte, théorie de la crédibilité

Keywords: insurance, natural hazards, statistical learning, extreme values theory, text analysis, credibility theory

Cette thèse a été préparée au

Laboratoire de Probabilités, Statistique et Modélisation
Sorbonne Université
Campus Pierre et Marie Curie
4 place Jussieu
75005 Paris
France

Contributions des données de l'assurance à l'étude des risques naturels
Application de méthodes d'apprentissage statistique pour l'évaluation de la nature et du coût des dommages assurés liés aux événements naturels en France

Résumé

Dans un contexte d'augmentation du coût des dommages assurés lié aux événements climatiques, d'un niveau déjà élevé, les assureurs ont vocation à participer à l'amélioration de la connaissance et de la prévention. Dans cette thèse, nous présentons des applications de méthodes d'apprentissage statistique pour l'évaluation de la nature et du coût des dommages assurés dû aux risques naturels en France. Nous commençons par étudier la sinistralité à l'échelle fine du bâti. Pour cela, nous analysons les données textuelles des rapports d'expertise. Ensuite, nous présentons des travaux portant sur l'estimation du coût de la sécheresse en France. Enfin, nous proposons une méthode estimation du coût des événements inondations, rapidement après leurs occurrences. Nous introduisons une méthode combinant des arbres de régression et la théorie des valeurs extrêmes. Nous ajoutons une application de la théorie de la crédibilité pour compléter cette estimation.

Mots clés : assurance, risques naturels, apprentissage statistique, valeurs extrêmes, analyse de texte, théorie de la crédibilité

Abstract

In the context of increasing costs of insured damages due to climatic events, at a level already high, insurers have to contribute to the improvement of natural risk knowledge and reduction. In this thesis, we present applications of statistical learning methods for the evaluation of the cost of insured damages due to natural hazards in France. In the first place, we will begin with a study of the damage distribution at the scale of the building. For this purpose, we analyze the textual data of the expert's reports. Then, we will present work on the estimation of the cost of drought in France. Finally, we propose a method for estimating the cost of flood events, quickly after their occurrence. We will introduce a method combining regression trees and extreme value theory. We will add an application of credibility theory to this estimation.

Keywords: insurance, natural hazards, statistical learning, extreme values theory, text analysis, credibility theory

Laboratoire de Probabilités, Statistique et Modélisation
Sorbonne Université – Campus Pierre et Marie Curie – 4 place Jussieu – 75005 Paris – France

Remerciements

C'est avec plaisir que je clôture la rédaction de ce manuscrit par ces remerciements. Je les adresse dans un premier temps à mes directeurs de thèse Olivier et Maud. Je n'aurai pas pu espérer meilleur encadrement. Vous avez réussi à me laisser une grande autonomie tout en étant toujours disponible pour répondre à mes questions et me guider dans ces domaines plutôt nouveaux pour moi à l'époque. J'ai énormément appris à vos côtés et je vous remercie pour toutes les connaissances que vous m'avez transmises et la générosité et la patience dont vous avez fait preuve. Ce fut vraiment une expérience très enrichissante, merci.

Je remercie Arthur Charpentier et Denys Pommeret pour avoir accepté de rapporter cette thèse. Merci aussi à Caroline Hillairet, Philippe Naveau et Olivier Winterberger pour avoir accepté de faire partie de mon jury, en particulier merci Caroline de l'avoir présidé.

Cette thèse se déroulant en entreprise, j'ai aussi pu bénéficier d'un accompagnement riche de ce côté-là. Je remercie donc chaleureusement Sarah, Eric et Lilian qui ont tous, à leur façon, pleinement participé à cette thèse et à son bon déroulement. Merci Sarah pour ta confiance, tes conseils toujours avisés et pour tout ce que tu as fait pour que cette thèse se passe le mieux possible. Merci Lilian pour toutes les heures que nous avons passé à discuter et à essayer de bien analyser les données. Ton expérience et ton soutien ont été très précieux. Enfin merci à Eric, sans qui ce projet n'aurai jamais vu le jour. Il fait partie de tes nombreuses contributions aux domaines de l'assurance et de la bonne prise en charge des risques naturels. Je te remercie surtout d'avoir partagé avec moi ta passion, ta connaissance et ton expérience sur tant de sujets différents. C'était une chance immense pour moi de t'avoir à mes côtés.

Je remercie mes collègues de la MRN, Lidia, Lucas, Léa, Mathis et la dernière arrivée Ombeline pour tous les bons moments que nous avons partagés et l'ambiance générale toujours très agréable à la MRN. C'était un plaisir de travailler avec vous.

Avec une thèse en grande partie en télétravail et à distance je n'ai pas beaucoup profité du laboratoire mais je remercie Sébastien qui a su bien m'accueillir. J'ai beaucoup apprécié collaborer avec toi et je te remercie pour ton aide et tes explications précieuses sur de nombreux sujets. J'ai aussi eu la chance de faire une conférence malgré les conditions particulières et je remercie donc les organisateurs et participants de Valpred de m'avoir donné ce joli aperçu.

Je remercie aussi toutes les autres personnes qui ont contribué à la faisabilité et au bon déroulement de mes travaux. En particulier, toutes les sociétés d'assurance contribuant aux données de la MRN et les deux réseaux d'expertises participant à nos études. Votre appui opérationnel a été déterminant dans ces travaux.

Je finirai par remercier mes proches, mes amis et ma famille pour leur soutien. Merci en particulier à mes parents qui m'ont toujours encouragé et soutenu dans mes études et sans qui je n'en serai pas là.

Table des matières

Liste des tableaux

Table des figures

Chapitre 1

Introduction

1.1 Préambule

La France n'est pas épargnée par les événements climatiques extrêmes et elle a toujours connu des catastrophes naturelles d'ampleur, comme les tempêtes extrêmes de Lothar et Martin en 1999 ou les inondations de la Seine en 1910. L'histoire récente le confirme, elle a été, dans les 15 dernières années, soumise à des événements significatifs de tempêtes, de grêle, d'inondations, de submersions marines, mais aussi aux effets dévastateurs sur le bâti, d'épisodes répétés de sécheresse géotechnique, et même dans ces territoires d'Outre-Mer aux cyclones. Les pertes économiques et humaines qui en découlent sont désastreuses et l'assurance a un rôle majeur à jouer dans l'atténuation et dans la gestion de ces risques.

La Mission des sociétés d'assurances pour la connaissance et la prévention des risques naturels (MRN) a été créée en 2000. Comme son nom l'indique, son rôle, pour le compte de la profession de l'assurance, est de contribuer à une meilleure connaissance des risques naturels et d'apporter une contribution technique aux politiques de prévention. Cet apport technique de la MRN a pour objectif, de permettre une diminution des dommages, notamment par une réduction de la vulnérabilité des enjeux assurés.

Nous nous inscrivons dans cette mission, en apportant de la connaissance sur l'évaluation du coût et des conséquences, notamment sur le bâtiment, des risques naturels, en France, pour le marché de l'assurance dommages aux biens. Nous utilisons pleinement les données précieuses créées et collectées par le secteur de l'assurance, qui est en première ligne dans l'observation des dommages liés aux événements naturels.

Pour cela nous tirons profit des dernières méthodes d'apprentissage statistique, utilisées dans un contexte actuariel et dans un contexte d'analyse de données textuelles. Nous proposons aussi une application mêlant la théorie des valeurs extrêmes et la théorie de la crédibilité. Chaque chapitre décrit une application concrète d'un domaine d'études en mathématiques aux risques naturels. Les parties répondent à un problème de recherche appliquée et transposent ou étendent des sujets de recherches académiques récents. Cette thèse étant réalisée dans une entreprise (MRN) nos applications répondent à un besoin et à des contraintes industrielles.

Nous commençons par présenter dans le présent chapitre, le contexte de l'assurance des risques naturels en France et en particulier son régime CatNat. Nous proposons ensuite un panorama de l'impact assurantiel des risques naturels en France. Nous détaillons enfin les contributions de nos travaux et leurs apports dans ce contexte.

1.2 Assurance des risques naturels en France

1.2.1 Les différents régimes d'indemnisations

L'assurance joue un rôle majeur dans la prise en charge des risques liés aux aléas climatiques. Plusieurs stratégies sont possibles pour assurer ces risques, du fonds public à l'assurance de marché classique en passant par des systèmes mixtes comme c'est le cas en France. On trouve ces différents cas de figures dans d'autres pays, par exemple, aux États Unis ou en Allemagne des systèmes de marché sont plutôt privilégiés (Klein et S. Wang 2009 ; Surminski et Thieken 2017). En Espagne, le système repose sur le Consorcio de compensación de seguros, un organisme public d'assurance. Une étude comparative, des divers régimes d'assurance des risques naturels étrangers, peut être trouvée dans (*Mission d'enquête sur le régime d'indemnisation des victimes des catastrophes naturelles, Rapport de synthèse* 2005).

Ces différences reflètent différents modèles d'assurance, ceux basés sur la solidarité et ceux se basant sur une assurance personnalisée dépendant du risque, reposant sur la responsabilité individuelle. L'aptitude à mesurer le risque joue un rôle déterminant dans le choix des stratégies. Pour les risques naturels de plus en plus de données sont disponibles et le risque est aujourd'hui relativement bien mesuré. En 2021, Charpentier, Barry et Molly R. James, présentent une étude intéressante dans laquelle ils montrent qu'il serait possible d'introduire des primes basées sur le risque pour les catastrophes naturelles en France.

La solidarité reste historiquement le système choisi. On distingue principalement deux garanties en France pour les dommages liés aux événements climatiques :

- Une garantie assurantielle « classique » contractuelle avec une assurance de marché et une réassurance privée, pour les dommages considérés comme assurables (dommages causés par la tempête, la grêle ou le poids de la neige, communément appelé garantie TGN). Ces garanties sont contractuelles, facultatives ou obligatoires. La loi du 25 juin 1990 a permis la généralisation de la garantie tempête en rendant obligatoire la couverture des dommages pour toute personne ou entreprise détentrice d'un contrat d'assurance garantissant les dommages incendie. Pour les bâtiments une garantie grêle est incluse avec cette garantie.
- Un système mixte faisant appel à la fois à l'État et à l'assurance avec l'État réassureur de dernier ressort, dans le cadre du système de régime d'indemnisation des catastrophes naturelles (communément appelé régime CatNat) instauré par la loi du 13 juillet 1982.

Ces deux systèmes assurent un bon niveau de couverture des risques liés aux aléas climatiques et une indemnisation correcte pour les sinistrés. Ils sont tous les deux basés sur une mutualisation étendue du risque. Le montant de la garantie TGN pourrait être ajusté en fonction du risque mais on observe que les territoires les plus exposés ne sont pas pénalisés par des primes significativement plus élevées (Chneiweiss et Bardaji 2020). Le régime CatNat repose lui sur la solidarité nationale, il est au coeur de notre étude et nous le détaillons dans la section suivante.

On peut aussi mentionner que l'assurance agricole n'est pas obligatoire. Les agriculteurs peuvent assurer leurs récoltes via des contrats spécifiques grêle ou via une multirisque climatique récolte (MRC). Les agriculteurs qui ne sont pas assurés peuvent éventuellement bénéficier d'une indemnisation par le Fonds national de gestion du risque agricole (FNGRA, fonds des « calamités agricoles ») qui couvre les dommages ayant été reconnus par un arrêté du ministère de l'Agriculture. La non-assurance est particulièrement répandue, ce qui est une source de préoccupation croissante dans un contexte de dérèglement climatique. L'assurance agricole ne fait cependant pas partie du cadre de nos travaux.

Nous nous focalisons ici sur l'assurance de biens et responsabilité, et sur les dommages aux bâtiments (particuliers, entreprises, agriculteurs). Toute notre étude se concentre sur ce périmètre et nous ne prenons pas en compte la sinistralité agricole et automobile dans les chiffres et résultats

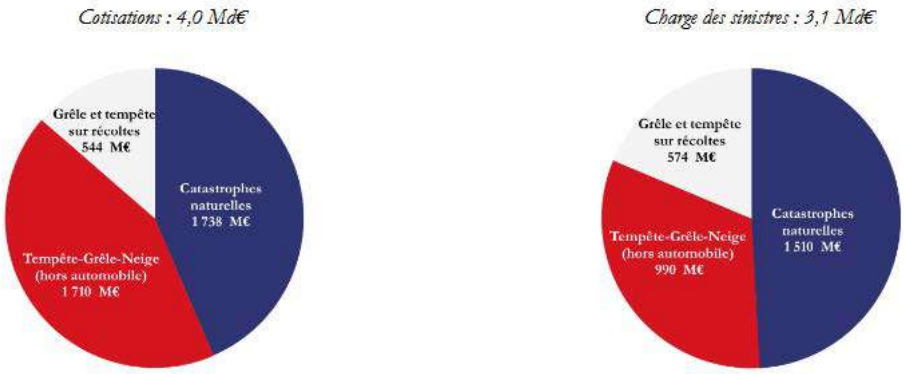


FIGURE 1.1 – Répartition des cotisations et de la charge des sinistres payés au titre des événements naturels en France en 2020 (Source : (*L'assurance des événements naturels en 2020* 2022)

que nous avançons. C'est en effet le champ d'action historique de la Mission Risques Naturels et nos données se restreignent à ce périmètre. Pour prendre en compte l'augmentation du coût de la construction nous présentons, dans la mesure du possible, les coûts en euros constants, c'est à dire actualisés avec l'indice de la Fédération Française du Bâtiment (FFB) le plus récent.

Si l'on regarde la répartition, on peut voir que la couverture des événements naturels totalise en 2020 un volume de cotisations de 4,0 milliards d'euros dont 43 % correspondent aux catastrophes naturelles, 43 % à la garantie TGN (hors automobile) et 14 % aux assurances des cultures. On trouve une répartition similaire pour les sinistres, (*L'assurance des événements naturels en 2020* 2022).

1.2.2 Le cas particulier du régime Catnat

En France, les catastrophes naturelles sont assurées via un partenariat public-privé, appelé le régime CatNat. Il repose sur une garantie, ce n'est pas une assurance obligatoire mais une extension de garantie obligatoire à tout contrat d'assurance dommages couvrant un bien situé en France métropolitaine et dans une grande partie de l'Outre-mer. Cette spécificité française influence fortement la gestion des sinistres catastrophes naturelles. Ce régime d'indemnisation des catastrophes naturelles a été créé par la loi du 13 juillet 1982, (*LOI numéro 82-600 du 13 juillet 1982 relative à l'indemnisation des victimes de catastrophes naturelles* 1982), et repose sur un principe de solidarité : pour chaque contrat, le même taux de surprime d'assurance, fixé par le gouvernement, est utilisé pour compenser les pertes des catastrophes naturelles. C'est un partenariat Public-Privé entre l'État et les sociétés d'assurance, les assureurs se chargent de la gestion des sinistres mais l'État régule les caractéristiques clé du contrat notamment, la définition des périls couverts, la tarification, la franchise et le processus de reconnaissance en état de catastrophe naturelle. L'état apporte sa garantie illimitée au régime CatNat par l'intermédiaire d'une entreprise publique, la Caisse Centrale de Réassurance (CCR). En pratique, une très grande majorité des assureurs se réassure auprès de la CCR, ce qui en fait un acteur clé et incontournable dans la gestion des risques naturels et plus spécifiquement du régime CatNat, (GRISLAIN-LETRÉMY et PEINTURIER 2010).

Liste des périls couverts Son champ d'application est large et il n'est pas limité, aujourd'hui en pratique il couvre :

- les inondations et/ou coulées de boue ;
- les mouvements de terrain ;
- les inondations par remontée de nappe ;
- les vents cycloniques ;
- les secousses sismiques ;
- les chocs mécaniques liés à l'action des vagues ;
- les avalanches ;
- les algues sargasses ;
- les laves torrentielles ;
- les éruptions volcaniques ;
- les raz de marée ;
- les effondrements et/ou affaissements ;
- les éboulements et/ou chutes de blocs ;
- les glissements de terrain ;
- les mouvements de terrain différentiels consécutifs à la sécheresse et à la réhydratation des sols.

Cette liste n'est pas exhaustive et peut être complétée. Depuis 1982, les 35 000 communes de France ont toutes fait l'objet d'au moins un arrêté catastrophe naturelle. Les inondations génèrent le plus d'arrêtés avec 52 % des arrêtés, ensuite la sécheresse avec 23 % et puis les mouvements de terrain avec 12 %.

Tarifications et franchises Dans le cadre du régime CatNat, la cotisation correspond à un taux uniforme de surprime sur tout le territoire, établi par les pouvoirs publics. Elle s'élève à 12 % sur les assurances de dommages aux biens des particuliers et des professionnels, et à 6 % sur les garanties vol et incendie d'un véhicule terrestre à moteur (ou à défaut 0,5 % sur la garantie dommages). À titre d'illustration, la prime CatNat est de 26 € en moyenne pour une habitation, (Chneiweiss et Bardaji 2020). Cette prime repose sur la solidarité nationale puisque tout le monde paye le même taux, quelle que soit son exposition. Il n'y a pas de modulation de prime selon l'exposition, le type de bien ou sa construction. Un prélèvement de 12 % est appliqué sur cette prime pour alimenter le Fonds national de prévention des risques naturels majeurs (FPRNM), souvent appelé « fonds Barnier ». Ce fonds contribue au financement des études des Plans de Prévention des Risques (PPR), des Programmes d'Actions de Prévention des Inondations (PAPI), à l'information préventive et peut aussi servir à financer l'aménagement de certains quartiers ou le déménagement des habitants les plus exposés.

Une franchise obligatoire et non-rachetable est aussi fixée par l'état. Elle s'établit à 380 € pour les habitations et les véhicules pour tous les types de CatNat à l'exception de la sécheresse où dans ce cas elle est de 1520 €. Pour les biens à usages professionnels, elle est de 10% des dommages directs avec un minimum 1140 € et de 3 jours ouvrés pour les pertes d'exploitation. En l'absence de PPR dans la commune pour le péril reconnu, la franchise pouvait être modulée en fonction du nombre d'arrêtés parus pour le même péril survenu dans les 5 années précédentes. La règle est la suivante, dans les 5 dernières années s'il y a eu 1 à 2 reconnaissances alors la franchise de base est appliquée, 3 reconnaissances alors il y a doublement de la franchise, 4 reconnaissances triplement de la franchise, 5 reconnaissances et plus quadruplement de la franchise. Cette modulation est suspendue dès la prescription d'un PPR pour le péril concerné mais elle est réactivée en cas d'absence d'approbation de ce PPR à l'issue d'un délai de quatre ans. Cette règle sera supprimée, sauf pour les collectivités locales, à partir du 1^{er} janvier 2023.

FIGURE 1.2 – Processus d'indemnisation dans le cadre du régime CatNat (Source : DGSCGC)

Processus de reconnaissance en état de catastrophe naturelle Pour recevoir une indemnisation, un arrêté du gouvernement reconnaissant la commune en état de catastrophe naturelle ou non, doit être publié au « Journal Officiel », où sont publiées toutes les lois et manifestations législatives de la République française. Lors d'un sinistre, pour recevoir une indemnité, l'assuré se déclare auprès de sa mairie. La mairie fera ensuite une demande auprès de la préfecture. La préfecture va centraliser les demandes qu'elle transmettra à la Direction Générale de la Sécurité Civile et de la Gestion des Crises (DGSCGC) qui va instruire les demandes. Ensuite les demandes sont analysées dans une commission interministérielle qui va rendre un avis. C'est un processus complexe comme décrit dans la figure 1.2.

La décision est motivée par une commission interministérielle, qui évalue le caractère exceptionnel de l'agent naturel de l'événement au niveau de la ville. Pour les inondations, la décision est fondée sur la période de retour de l'événement. Pour la sécheresse, l'évaluation prend en compte le type de sol et son humidité. Cela correspond à l'exposition au retrait et au gonflement de l'argile (type de sol) et à l'intensité météorologique de la sécheresse dans la ville (humidité). La classification de la propension au retrait-gonflement des argiles est accessible au public par le biais d'une cartographie fixe produite par le Bureau de recherches géologiques et minières (BRGM) et présentée dans la section 3.3.2. Sur la base des valeurs de l'indice d'humidité pendant plusieurs mois et selon que la ville présente ou non des zones argileuses, l'arrêté reconnaîtra la ville en état de catastrophe naturelle, voir (*Procédure de reconnaissance de l'état de catastrophe naturelle - Révision des critères permettant de caractériser l'intensité des épisodes de sécheresse-réhydratation des sols à l'origine de mouvements de terrain différentiels.* 2019 ; *Contribution de Météo-France à l'analyse de la sécheresse géotechnique à l'attention de la Commission CatNat pour l'année 2019* 2020 ; *Météo-France dans le dispositif CATNAT sécheresse.* 2020). Ce processus peut prendre du temps, le délai moyen entre la survenance de l'événement et la décision est d'environ 18 mois pour la sécheresse contre 50 jours pour les autres aléas, (*Sécheresse Géotechnique, de la connaissance de l'aléa à l'analyse de l'endommagement du bâti* 2018). De plus, si aucune demande n'est faite, ou si la commission refuse la demande, il n'y aura pas d'indemnisation de ce régime. Dans ce cas, une couverture complémentaire peut être fournie par la compagnie d'assurance mais pour la sécheresse, c'est très rare.

Réforme Le 28 décembre 2021, une loi relative à l'indemnisation des catastrophes naturelles a été promulguée. Cette loi vise à améliorer la transparence du processus décisionnel de reconnaissances à l'égard des maires et des sinistrés. La commission interministérielle de reconnaissance de l'état de catastrophe naturelle, précédemment mentionnée, est désormais inscrite dans la loi. Les délais pour déclarer un sinistre et obtenir réparation ont aussi été modifiés. Les frais de relogement d'urgence des sinistrés de catastrophes naturelles seront intégrés à l'indemnisation et la modulation de franchise précédemment mentionnée est supprimée. Enfin pour le risque sécheresse des mesures particulières vont être prises sur la base d'un rapport qui devra être remis dans un délai de six mois, (*LOI numéro 2021-1837 du 28 décembre 2021 relative à l'indemnisation des catastrophes naturelles* 2021). Ce risque a aussi fait l'objet d'un récent rapport de la Cour des comptes, (*Sols argileux et catastrophes naturelles* 2022).

C'est un régime qui permet un bon niveau d'assurance en France contre les catastrophes naturelles, même si l'on peut lui adresser certaines critiques avec notamment un caractère déresponsabilisant, voir par exemple (Gérin 2011 ; Latruffe et Picard 2005). En effet dans ce système, se basant sur la solidarité, les assurés ne payent pas le juste prix de leurs risques et sont peu incités à prendre des mesures de prévention. La récente réforme n'apporte pas de réponses à ce problème et va le traiter dans un rapport à part pour la sécheresse. Ce problème de la prévention et de l'assurance des catastrophes naturelles est complexe comme l'illustre la thèse de (Goussebaile 2016). Ces travaux recommandent que les politiques publiques s'attachent à rendre les agents conscients du risque et responsables de leur choix d'exposition. Dans le cadre du régime CatNat et de la solidarité nationale, ce n'est cependant pas la direction qui est privilégiée.

1.3 Panorama de l'impact assurantiel des risques naturels en France

Selon une nouvelle étude publiée par France Assureurs (FA) (*Etude : Changement climatique et assurance à l'horizon 2040* 2021), au total, le montant des sinistres dus aux événements naturels pourrait atteindre 143 milliards d'euros en cumulé entre 2020 et 2050, soit une augmentation de 93%, c'est-à-dire 69 milliards d'euros de plus par rapport à la période 1989 - 2019. Cette augmentation s'ajoute à un montant déjà élevé, la charge des sinistres des événements naturels étant de 2,4 milliards d'euros par an en moyenne pour 416 000 sinistres sur cette période de 1989 - 2019. On observe une augmentation sur les dernières années avec une moyenne annuelle de 3,8 milliards d'euros pour la période 2016-2019, marquée par des sécheresses importantes et des événements extrêmes comme le cyclone Irma ou les inondations de la Seine en 2016.

La France est touchée par la majorité des aléas comme la sécheresse, inondation, submersion marine, tempête, grêle, cyclone, séisme pour ne citer que les principaux. Leur assurance est donc un enjeu majeur pour la société. Dans la suite de cette section, nous détaillons les principaux aléas que nous traitons dans cette thèse.

1.3.1 Inondation

Une très grande partie du territoire français est exposée aux inondations et l'on pense souvent à cet aléa quand on parle de risques naturels. On compte 21 000 communes en zone d'exposition et 34 700 ont rapporté au moins un sinistre inondation depuis 1990 selon les bases de données de la MRN. Selon les données des enveloppes approchées d'inondations potentielles, prés d'une personne sur quatre habite en zone inondable. C'est un péril qui touche une large partie du territoire ce qui s'observe dans les montants indemnisés. Il y a eu entre 2016 et 2020 en moyenne

FIGURE 1.3 – Répartition annuelle de la charge des sinistres et principaux événements (Source : (*Etude : Changement climatique et assurance à l'horizon 2040* 2021))

700 millions d'euros indemnisés par an. Des événements extrêmes pèsent pour une grande partie de cette charge avec notamment les inondations des bassins de la Seine moyenne et de la Loire de mai-juin 2016 responsables à elles seules de 1 500 millions d'euros de dommages (*L'assurance des événements naturels en 2020* 2022). Ces inondations avaient touché un large périmètre à forts enjeux, mais il y a aussi dans les esprits des inondations plus concentrées et intenses, comme en octobre 2020 dans le sud de la France, suite à la tempête Alex aux conséquences désastreuses pour les vallées de la Roya et de la Vésubie. Ces inondations ont eu un lourd bilan avec des pertes humaines et ont aussi été responsables de près de 200 millions d'euros de dommages selon la CCR.

La France est exposée à plusieurs types d'inondations, il est parfois difficile de séparer ou de déterminer le type de phénomène et certains événements sont généralisés avec plusieurs types en même temps. On peut cependant distinguer plusieurs catégories d'inondations comme présentées dans (*Inondations, s'informer pour mieux se protéger* 2019) :

— Les inondations par débordement de cours d'eau, de type « crues lentes de plaine », se produisent lorsqu'un fleuve ou une rivière sort lentement de son lit mineur et envahit son lit moyen, voire son lit majeur. Ces inondations sont assez communes. Elles sont souvent lentes et prévisibles car elles font suite à des longues périodes de pluies dans des cours d'eau déjà hauts, elles causent peu de pertes humaines directes.
— Les inondations par débordement de cours d'eau, de type « crues rapides et torrentielles », se produisent principalement en zone montagneuse suite à des précipitations intenses. Elles peuvent provoquer des inondations éclairs aux conséquences potentiellement dévastatrices qui peuvent entraîner des pertes humaines. La montée des eaux peut être très rapide et moins prévisible.
— Les inondations par ruissellement, se produisent lorsque les précipitations ne peuvent pas ou plus s'infiltrer dans le sol. Suite à des fortes précipitations, concentrées sur plusieurs jours ou rapides sur plusieurs heures, les eaux ruissellent dans des zones habituellement sèches. Elles sont aggravées par l'artificialisation du sol qui crée des zones imperméables

empêchant l'absorption du surplus d'eau. En milieu urbain cela peut provoquer des écoulements avec des vitesses importantes à l'origine de dégâts humains et matériels potentiellement conséquents. En milieu rural cela peut se transformer en coulée de boue et provoquer des dégâts très importants. Ce type d'inondation touche tout le territoire et est difficile à prévoir.

— Les inondations par submersions marines, se produisent dans des zones côtières lorsque des conditions météorologiques et océaniques défavorables produisent des inondations par la mer. Plus spécifiquement, elles peuvent être liées à des débordements du niveau de la mer, aux franchissements de paquets de mer liés aux vagues ou à la rupture du système de protection. Avec sa façade maritime et de ses côtes basses, la France est particulièrement exposée au risque de submersion marine. Ce sont des inondations rapides et de courtes durées qui font souvent suite à des tempêtes et à des conditions de marées particulières.

— Les inondations par remontée de nappe, elles sont provoquées par la montée du niveau de la nappe phréatique jusqu'à la surface du sol. Lorsque des événements pluvieux exceptionnels se produisent et chargent de façon anormale les nappes phréatiques, déjà dans des conditions particulières, alors il peut y avoir des inondations par remontée de nappe. Ces inondations peuvent provoquer des dommages dans les sous-sols, garages semi-enterrés ou caves, des fissurations d'immeubles et des dommages au réseau routier et de chemins de fer.

Les dommages liés aux inondations sont nombreux et variés, ils dépendent du type d'inondations. Sur le bâtiment individuel, les dommages affectent principalement les embellissements et notamment les revêtements de murs ou les réseaux. Des inondations torrentielles ou intenses peuvent néanmoins aussi affecter la structure.

Selon l'étude de la fédération, (*Etude : Changement climatique et assurance à l'horizon 2040* 2021), le coût des dommages liés aux inondations va augmenter pour atteindre 50 milliards d'euros sur la période 2020-2050, principalement suite à une augmentation de la richesse, qui se traduit par de plus fortes concentrations d'entreprises et de logements, et d'avantages d'infrastructures et de biens. Les mêmes conclusions sont obtenues pour les submersions marines avec une augmentation supposée de 87% par rapport à la période passée. Ici l'impact du changement climatique est plus important et pèse pour $6,5$ milliards d'euros sur les 54 prévus.

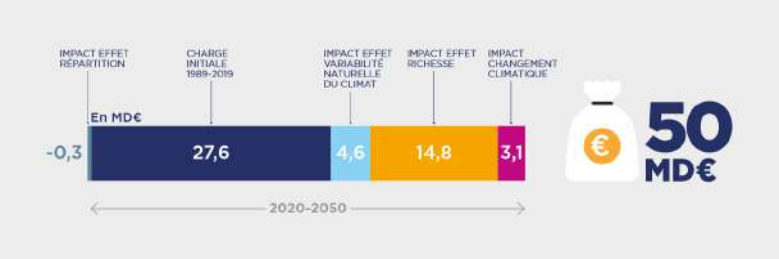


FIGURE 1.4 – Évolution supposée du coût lié aux inondations pour la période 2020-2050 (Source : (*Etude : Changement climatique et assurance à l'horizon 2040* 2021))

1.3.2 Sécheresse

Bien qu'il soit généralement peu connu, le risque de sécheresse est responsable d'environ 30 % du montant total des indemnités versées par le régime français CatNat (*Sécheresse Géotechnique,*

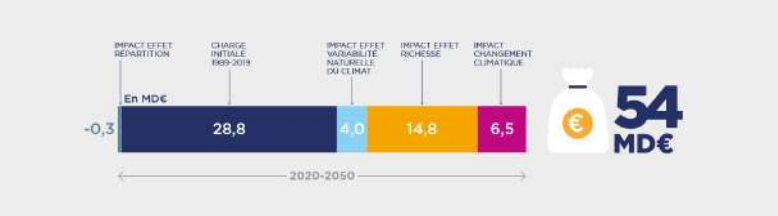


FIGURE 1.5 – Évolution supposée du coût lié aux submersions marines pour la période 2020-2050 (Source : (*Etude : Changement climatique et assurance à l'horizon 2040* 2021))

de la connaissance de l'aléa à l'analyse de l'endommagement du bâti 2018). Ici et dans tout ce document nous parlerons de sécheresse et de retrait-gonflement des sols argileux (RGA) de façon indistincte pour désigner les dommages aux bâtiments liés aux mouvements du sol, suite à des modifications des conditions météorologiques. Il ne s'agit donc pas des dommages causés aux cultures agricoles, ni des seules conséquences du déficit de la ressource en eau, mais bien de ceux qui affectent les bâtiments et principalement les maisons individuelles. Résultant de plusieurs facteurs tels que la nature géologique du sol, le contexte hydrogéologique, les conditions climatiques et les défauts de construction (*Face aux risques – Le retrait gonflement des argiles* 2007).

Les dommages liés à ce risque représentent près de 14 milliards d'euros sur la période 1989-2019. Avec beaucoup d'années extrêmes, sur les 20 événements les plus coûteux sur la période 1989-2019, 11 sont liés à des épisodes de sécheresse dont celui de 2003 avec 1,9 Md € de dommages. On peut également citer celui de 2018 à 1,3 Md € et ceux de 1990 et 2017 estimés à environ 800 M €. Les estimations pour 2019 et 2020 sont aussi très élevés avec respectivement 750 et 1 100 millions d'euros estimés. Sur la période 2016-2020 on arrive donc à une moyenne annuelle de 900 millions d'euros indemnisés.

Ces montants sont dûs à une exposition du territoire importante, 10,5 millions de maisons sont implantées dans une zone susceptibilité moyenne ou forte, (*Cartographie de l'exposition des maisons individuelles au retrait-gonflement des argiles* 2021). En effet, 48 % du territoire est en zone d'exposition moyenne et forte, concentrant 93 % de la sinistralité. Les sinistres ont un coût moyen très élevés avec 16 300 €, ce qui en fait le plus cher des garanties dommages.

La gestion des reconnaissances CatNat pour cet aléa est aussi problématique, sur les neuf dernières années 50 % des demandes de reconnaissance n'ont pas été acceptées. Les délais de reconnaissances sont aussi longs et pendant longtemps les critères ont étés assez opaques. Une récente amélioration est à noter, notamment avec la loi, évoquée plus haut, de décembre 2021. Ce processus de demande semble cependant inadapté à cet aléa comme évoqué par le rapport de la cour des comptes, (*Sols argileux et catastrophes naturelles* 2022) et de nombreux travaux se sont penchés sur ce problème et ont conclu à la nécessité de réviser ce dispositif. Ce qui, comme prévu par la loi de décembre 2021, (*LOI numéro 2021-1837 du 28 décembre 2021 relative à l'indemnisation des catastrophes naturelles* 2021), devrait être fait à la suite de la publication d'un rapport sur les pistes à envisager, dont la publication devrait être imminente.

Les dommages liés aux RGA sont en effet multifactoriels et il est difficile de les conditionner à seulement des critères météorologiques et géologiques comme nous le verrons au chapitre 5. De nombreux facteurs, comme des défauts de conception, de fondations et de structures ou un environnement défavorable comme la présence de végétation à proximité peuvent aggraver les dommages, (*Avant de construire – Prendre en compte les risques du terrain* 2014). Ainsi de nombreuses dispositions préventives peuvent être prisent pour atténuer les effets du RGA comme

illustré en 1.6. Cependant en pratique de nombreuses maisons ne sont pas construites selon ces dispositions et la sinistralité est donc fréquente.

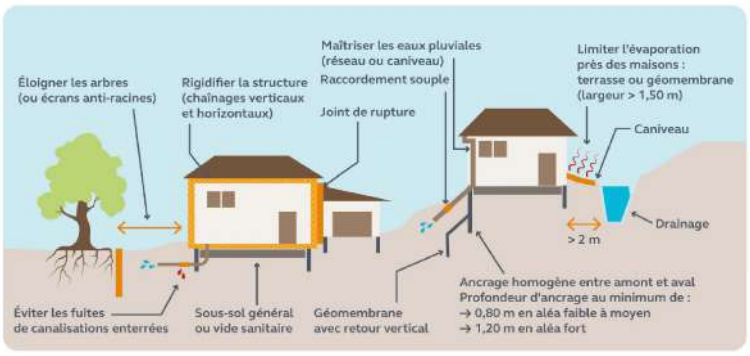


FIGURE 1.6 – Exemples de dispositions préventives pour construire en zones argileuses (Source : (*Sols argileux et catastrophes naturelles* 2022))

Les dommages observés sont importants et les modes de réparations lourds. Les dommages les plus usuels sont des fissures sur le bâtiment. Pour réparer ces fissures, il faut souvent reprendre les fondations et les revêtements de façade, ce qui peut nécessiter aussi de défaire puis de refaire les embellissements. Le coût d'un sinistre sécheresse, dont une analyse est faite dans le rapport, (*Sécheresse Géotechnique, de la connaissance de l'aléa à l'analyse de l'endommagement du bâti* 2018), peut être décomposé selon trois composantes. Premièrement, des dommages de reprises en sous-oeuvre des fondations, qui représentent plus de la moitié de la charge et présente un coût moyen élevé avec 24 000 €. La reprise en sous-oeuvre total d'une maison est très onéreuse, son coût peut atteindre plusieurs centaines de milliers d'euros. Ensuite les réparations de façades, qui sont présentes dans 72% des dossiers, cela correspond au traitement des fissures et la reprise des enduits de façade avec un coût moyen de 8 800 €. Enfin des travaux sur les embellissements peuvent être nécessaires, ils présentent des coûts moyens plus faibles.

La sécheresse est fortement soumise au changement climatique et sur la période 2020-2050 le montant total pourrait atteindre 43 milliards d'euros dont 17 liés au changement climatique. Les réflexions et reformes en cours sont d'autant plus nécessaire du fait de cette augmentation à prévoir.

FIGURE 1.7 – Évolution supposée du coût lié à la sécheresse pour la période 2020-2050 (Source : (*Etude : Changement climatique et assurance à l'horizon 2040* 2021))

1.3.3 Tempête

La tempête est le premier enjeu climatique en terme de fréquence et de charge annuelle sur le territoire français. Sur la période 2016-2020 il y a en moyenne 1 milliard d'euros indemnisés par an au titre de la tempête. Les tempêtes peuvent toucher tout le territoire et être généralisés et extrêmes. Sur les vingt dernières années, trois tempêtes majeures ont touché la France, les tempêtes Lothar et Martin de 1999 qui restent encore à ce jour, et de loin, l'événement climatique le plus coûteux avec un montant de plus de 7 milliards d'euros constants. Ensuite deux autres tempêtes majeures ont frappé la France, Klaus et Quinten en 2009 et Xynthia en 2010 avoisinant toutes deux les 2 milliards d'euros de dommages assurés. En touchant une grande partie du territoire les tempêtes atteignent des coûts très importants, bien que le coût moyen des sinistres soit moins élevé que pour les autres aléas.

Nous considérons ici les effets du vent suite à des perturbations atmosphériques. Ces vents violents s'accompagnent de fortes précipitations et parfois d'orages, ce qui peut faire que l'on observe aussi des inondations lors de grandes tempêtes, et donc des reconnaissances CatNat. Les tempêtes peuvent avoir un impact considérable aussi bien pour les personnes que pour leurs activités ou leur environnement. C'est un risque assez bien étudié et ne relevant pas du régime CatNat, sa gestion est propre à chaque compagnie d'assurance. On peut ainsi trouver une littérature de ce mode de gestion avec par exemple la thèse (Mornet 2015) qui présente un modèle de tarification. C'est un phénomène qui reste difficile à prédire, notamment l'estimation de leur intensité et de la zone géographique touchée. En effet, Météo France est capable de prédire la survenance des tempêtes deux ou trois jours à l'avance mais il est difficile de déterminer avec précision la trajectoire de la tempête.

Les dommages des effets du vent sont principalement concentrés sur la charpente et la couverture. Cependant tout ouvrage extérieur est soumis à ces effets et est donc vulnérable. On peut aussi observer une influence des matériaux employés et de la date de construction sur la sinsitralité, (*Lettre d'information de la Mission Risques Naturels 36* 2021), ce qui ajoute une échelle locale difficile à prendre en compte pour la prédiction de la charge sinistre.

Il n'y a pas de tendance climatique sur l'évolution de l'intensité des tempêtes au cours des prochaines décennies. Dans l'évolution prévue par la fédération, on n'observe pas d'effet du changement climatique sur la charge sinistre mais l'augmentation du nombre et de la valeur de biens reste un facteur aggravant. Il est donc attendu que le coût lié aux tempêtes augmente encore pour atteindre 46 milliards d'euros sur la période 2020-2050.

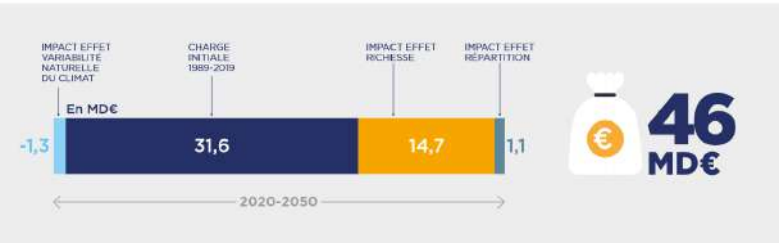


Figure 1.8 – Évolution supposée du coût lié aux tempêtes pour la période 2020-2050 (Source : (*Etude : Changement climatique et assurance à l'horizon 2040* 2021))

1.3.4 Grêle

Dans les périls que nous étudions la grêle est le moins coûteux pour les dommages aux bâtiments avec en moyenne 300 millions d'euros par an sur la période 2014-2018. On observe moins d'événements extrêmes que pour les autres mais il y en a quand même, les orages de grêle liés à l'événement Ela de la pentecôte 2014, responsable de plus de 900 millions d'euros de dommages, en sont un bon exemple.

La grêle se manifeste sous la forme de précipitations de grains de glace appelés grêlons d'un diamètre moyen de quelques centimètres. Les averses de grêle, affectent tout type de biens et peuvent être parfois dévastatrices que ce soit pour les habitations, mais aussi pour les récoltes et les véhicules. Il n'y a pas de cartographie officielle de la grêle en France, les travaux de (Vinet 2002), semblent observer que certaines zones semblent plus exposées au péril grêle. Ce que la MRN dans une étude semble confirmer (*Lettre d'information de la Mission Risques Naturels 34* 2020).

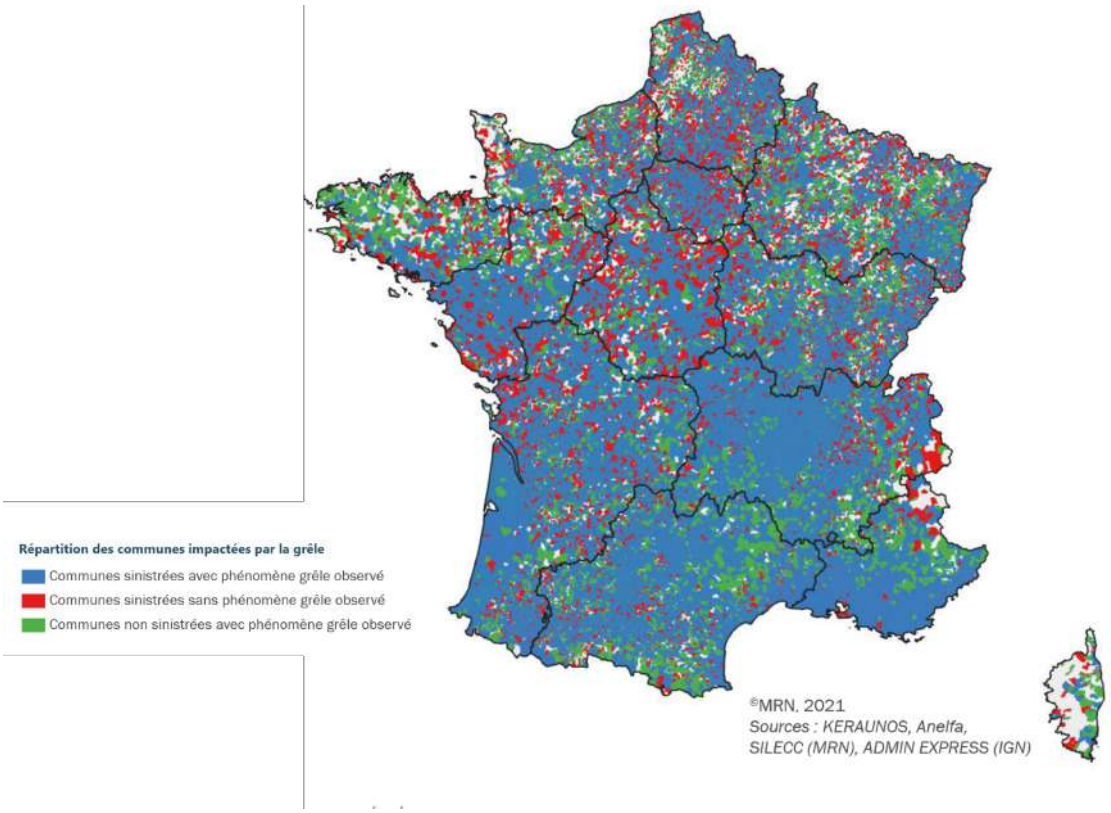


Figure 1.9 – Cartographie de la répartition des communes impactées par des sinistres grêles entre 2006 et 2020 (Source : MRN)

L'évolution du coût lié au changement climatique n'est pas estimée mais des études semblent indiquer une augmentation non pas de la fréquence de la grêle mais plutôt de l'intensité, fortement corrélée à la température de la nuit, (Claude Berthet, Jean Dessens et Sánchez 2011). Les projections prévues par les modèles d'évolution du climat donneraient une augmentation de 40% de l'intensité des chutes de grêle à l'horizon 2040 sans évolution de la fréquence, (Dessens,

Berthet et Sanchez 2015).

Nous ne les étudions pas ici mais on peut aussi mentionner les séismes et les cyclones qui ont des conséquences potentiellement désastreuses. Deux événements récents le rappellent, notamment le cyclone Irma très important et responsable de près de 2 milliards de dommages assurés. Les cyclones sont des phénomènes météorologiques violents qui se forment dans les régions tropicales. Les départements, français d'outre-mer sont concernés en particulier Saint-Martin et Saint-Barthélemy pour le cyclone Irma. Pour les séismes récemment en novembre 2019, un séisme de magnitude Mw de 4,8 – 4,9, a occasionné 17 850 sinistres pour un coût global estimé à 261 millions d'euros de dégâts dans le sud de la France. Ces deux risques peuvent avoir des conséquences importantes et une partie du territoire est exposée.

1.4 Apport de la statistique

C'est dans ce contexte d'augmentation du coût des dommages assurés, d'un niveau déjà élevé, que nos travaux s'inscrivent. En faisant partie d'une thèse CIFRE avec la Mission Risques Natuels, ils ont vocation à participer à l'amélioration de la connaissance et de la prévention des risques naturels en France. Pour cela la statistique présente un ensemble d'outils très utiles et son usage est particulièrement indiquée. Elle nous offre des techniques permettant d'explorer et d'analyser les données précieuses, récoltées par le secteur de l'assurance. En effet nous reposons sur les données recueillies par les sociétés d'assurance et par les experts d'assurance, pour nous permettre d'améliorer la connaissance des risques naturels en France. Nous nous appuyons en particulier sur la théorie des valeurs extrêmes qui permet l'étude des événements extrêmes, ce qui est un des cadres usuel d'étude pour les risques naturels. La théorie de la crédibilité nous permet de prendre en compte l'expérience passée sur les événements étudiés et de lui attribuer un niveau de confiance. Nous reposons aussi sur des modèles d'apprentissage statistique et des méthodes classiques et usuelles d'actuariat. Enfin les méthodes d'analyses de texte nous permettent d'obtenir des informations précises sur la nature des dommages provoquée par les risques naturels. Ces méthodes nous fournissent des cas d'applications intéressantes, utiles au marché de l'assurance et à l'intérêt général que nous présentons dans la section suivante.

1.5 Nos contributions

Nos contributions sont organisées en quatre chapitres qui bien que reliés, peuvent être lus indépendamment et nous présentons au préalable dans deux chapitres distincts le contexte théorique et industriel de nos travaux. Le chapitre 4 présente une étude de la sinistralité à l'échelle fine du bâti. Pour cela nous analysons les données textuelles des rapports d'expertise. Ensuite le chapitre 5 présente une application de méthodes d'apprentissage automatique pour l'estimation du coût de la sécheresse en France. Cette contribution a été soumise à une revue. Enfin le chapitre 6 propose une méthode estimation du coût des événements inondations rapidement après leurs occurrences. Ce chapitre repose en partie la théorie de la crédibilité et pour trouver l'a priori sur des arbres de régression avec une loi Pareto généralisée. Ce chapitre fait aussi l'objet d'un article en cours de finalisation. Nous finissons par présenter des résultats sur la consistance de la procédure des arbres de régression avec une loi de Pareto généralisée en chapitre 7. Ils ont aussi été étudiés dans un article soumis.

Analyse de la sinistralité à l'échelle fine du bâti Dans ce travail, nous comparons différentes méthodes de classification de textes, appliquées aux données récoltées dans les rapports

d'expertise, de constatation des dommages, suite à un événement naturel. L'objectif de cette étude est d'améliorer la connaissance des conséquences des événements naturels, au niveau des bâtiments, afin de donner aux gestionnaires de risques les moyens d'en réduire le coût. Pour cela nous cherchons à mieux connaître les dommages occasionnés et améliorer la connaissance de la nature et du coût de l'endommagement à l'échelle fine du bâti. En effet, le coût total d'un événement naturel est généralement bien rapporté par les compagnies d'assurance, mais pas la décomposition exacte des dommages. Notre objectif est de créer une base de données rapportant les dommages indemnisés selon des composantes du bâti prédéfinies. Pour ce faire, nous nous appuyons sur les rapports d'expertise et les informations textuelles présentes dedans. Nous comparons deux méthodes de classification de texte avec des réseaux de neurones, pour rapprocher le texte de nos composantes. Nous présentons aussi une étude prospective sur des données non structurées, les rapports d'expertise en entier, pour voir ce que l'on est en mesure de récupérer automatiquement pour enrichir nos analyses.

Estimation du coût d'un épisode de sécheresse Cette partie traite de la prédiction du montant total des dommages causés par un épisode de sécheresse dans le contexte particulier du régime CatNat. Du fait de la spécificité de ce régime, une estimation précoce du montant final des sinistres est particulièrement stratégique. Grâce à ce partenariat avec la Mission Risques Naturels, nous avons eu accès à une base de données de sinistres catastrophes naturelles alimentée par les principales compagnies d'assurance françaises. En combinant les sinistres liés aux événements de sécheresse contenus dans la base de données avec des données météorologiques et socio-économiques, nous avons pu avoir une meilleure connaissance de l'exposition. Notre approche de prédiction repose sur la comparaison de différents modèles statistiques et algorithmes d'apprentissage automatique. Notamment les modèles linéaires généralisés combinés avec des pénalités Lasso et Elastic-Net, les forêts aléatoires ou l'Extreme Gradient Boosting. Pour améliorer les performances, nous proposons une agrégation des différents modèles. La principale difficulté vient du fait que les données sont déséquilibrées puisqu'une grande majorité des villes ne sont pas touchées par un épisode de sécheresse. Les prédictions obtenues à partir des différents modèles sont ainsi évaluées à l'aide des courbes Precision et Recall, des F_1-scores et des matrices de confusion.

Estimation du coût des inondations Dans ce chapitre nous proposons deux méthodes permettant d'estimer le coût des événements inondations. Nous nous inscrivons dans le cadre d'une mission d'appui à France assureurs pour dimensionner les réponses en cas de gestion de crise liée aux événements naturels. Nous essayons d'estimer le coût d'un événement inondation après son occurrence. L'estimation du coût des inondations est un enjeu majeur pour le secteur de l'assurance qui est étudié en particulier, pour évaluer leurs expositions. Nous tirons profit des informations disponibles à la MRN sur les événements passés pour caractériser les événements en cours. Pour cela on repose sur deux méthodes, une méthode d'arbre de régression en fonction du comportement de la queue de distribution et une méthode de type sévérité fréquence se basant sur une comparaison d'événements similaires. La première méthode repose sur une régression, fondée sur des arbres de régressions couplées à des distributions de Pareto généralisées permettant d'étudier les comportements extrêmes. Cette méthode est complétée par la théorie de la crédibilité pour donner une estimation du coût à l'échelle de la commune.

Arbre de régression avec une loi de Pareto généralisée Dans ce chapitre nous étudions des résultats théoriques de la procédure d'arbre de régression avec une loi de Pareto généralisée (CART GPD) utilisée dans le chapitre précédent. Nous montrons que cette procédure est consistante. On commence par introduire les notations et hypothèses faites avant de montrer la

consistance d'un arbre fixé à K feuilles, en séparant la partie stochastique de l'erreur et la partie de mauvaise spécification causée par l'approximation de la GPD. On étudie ensuite la consistance de l'élagage de l'arbre. Des preuves sont disponibles en annexe.

Chapitre 2

Contexte théorique

Nous présentons dans ce chapitre les principales méthodes statistiques utilisées dans cette thèse. Nous nous appuyons en particulier sur la théorie des valeurs extrêmes, la théorie de la crédibilité, des modèles d'apprentissage statistique et les dernières méthodes d'analyses de texte. Nous utilisons des méthodes classiques de statistique et d'actuariat que nous appliquons sur des données réelles, que nous décrivons dans les chapitres suivants.

2.1 Théorie des valeurs extrêmes

2.1.1 Introduction

D'un point de vue statistique, les événements climatiques, en particulier les catastrophes naturelles, sont souvent des événements dits « extrêmes », c'est-à-dire des événements qui, lorsqu'ils se produisent, prennent de très petites ou de très grandes valeurs et peuvent avoir des lourdes conséquences.

L'étude de ces événements est motivée par la gestion des risques et vise à répondre à des problèmes d'inférence en dehors du support de l'échantillon, comment estimer la probabilité d'occurrence ou l'ampleur d'un événement lorsque celui n'a pas été observé.

La théorie des valeurs extrêmes propose le cadre statistique permettant de répondre à ces problèmes. Historiquement, sa création remonte aux travaux de (Fréchet 1927 ; Fisher et Tippett 1928 ; Gnedenko 1943 ; Gumbel 1958). Ils identifient les lois limites permettant de décrire le comportement, sous certaines hypothèses, des données extrêmes, c'est-à-dire celles dépassant un certain seuil.

Aujourd'hui, les domaines d'application sont nombreux et notamment pour l'étude des événements naturels (Bousquet et Bernardara 2021), elle est aussi particulièrement utilisée en hydrologie, (Katz, Parlange et Naveau 2002 ; Guillou et Willems 2006 ; J. A. Smith 1987), ou en actuariat (Brodin et Rootzén 2009 ; Embrechts, Klüppelberg et Mikosch 2013 ; Resnick 1997 ; Rootzén et Tajvidi 1997 ; Farkas, Lopez et Thomas 2021).

La théorie des valeurs extrêmes apparaît donc comme particulièrement indiquée pour notre étude. Dans la suite de cette section, nous introduisons les notations et les outils de cette théorie nécessaires pour la suite de cette thèse.

2.1.2 Méthode Peaks over threshold

Dans le cadre de cette thèse, nous nous appuyons sur l'approche par dépassements de seuil, traduction de Peaks-Over-Threshold notée PoT, dont le résultat fondamental a été prouvé par (Balkema et De Haan 1974). Elle repose sur l'utilisation des observations ayant dépassé un certain seuil.

Considérons des variables aléatoires Y_1, Y_2, ...,Y_n, indépendantes et identiquement distribuées (i.i.d) de fonction de répartition F inconnue. On note dans toute la suite, $\overline{F}$ la fonction de survie associée définie par $\overline{F}(y) = \mathbb{P}(Y_i > y)$ pour tout y.

Dans l'approche PoT, une observation est dite extrême si elle dépasse un certain seuil u préalablement choisi (le choix de ce seuil sera discuté plus tard). Sachant qu'une observation est extrême, on définit l'excès correspondant comme la différence entre cette observation et le seuil u. La figure 2.1 illustre cette méthode. La ligne rouge correspond au seuil u, et les points rouges aux observations extrêmes, à savoir celles qui ont dépassé le seuil u.

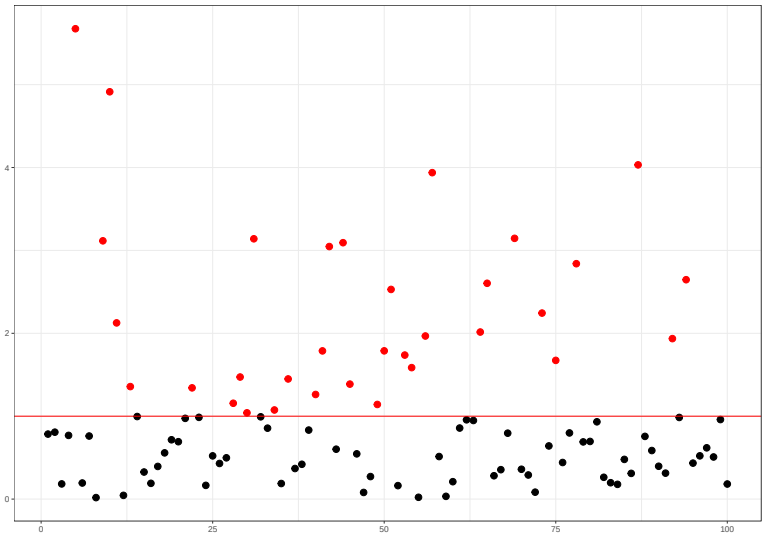


Figure 2.1 – Illustration de la méthode « Peaks-over-Threshold ». La ligne rouge représente le seuil u et les points rouges les observations extrêmes.

La loi des excès s'obtient facilement à partir de la fonction de répartition F :

$$\overline{F}_u(z) = \mathbb{P}[Y_i - u > z \mid Y_i > u] = \frac{\overline{F}(u+z)}{\overline{F}(u)}, z > 0.$$

En 1975, Pickands III démontre que si $\overline{F}$ satisfait la propriété suivante :

$$\lim_{t \to +\infty} \frac{\overline{F}(ty)}{\overline{F}(y)} = y^{-1/\gamma_0}, \forall y > 0,$$

avec $\gamma_0 > 0$, alors

$$\lim_{u \to +\infty} \sup_{z>0} |\overline{F}_u(z) - \overline{H}_{\sigma_{0u},\gamma_0}(z)| = 0,$$

où $\sigma_{0u} > 0$ et $\overline{H}_{\sigma_{0u},\gamma_0}$ est la fonction de survie d'une loi non dégénérée qui appartient nécessairement à la famille des lois de Pareto généralisée (GP) avec

$$\overline{H}_{\sigma_{0u},\gamma_0}(z) = \left(1 + \gamma_0 \frac{z}{\sigma_{0u}}\right)^{-1/\gamma_0}, z > 0,$$

σ_{0u} est un paramètre d'échelle et $\gamma_0 > 0$ un paramètre de forme, appelé indice de queue, reflétant l'épaisseur de la queue de F. Plus γ_0 est grand, plus la queue de distribution est lourde. On peut noter que si $\gamma_0 \in]0\,;1[$ alors l'espérance est finie, alors que si $\gamma_0 > 1$ l'espérance de Y_i est infinie, (Coles et al. 2001).

En pratique, une fois le seuil u choisi, une loi GP est ajustée aux excès associés au seuil. L'estimation des paramètres σ_{0u} et γ_0 peut être faite par maximum de vraisemblance.

Le choix du seuil u est complexe et reflète d'un compromis biais-variance : un seuil trop petit affaiblirait les estimations asymptotiques et amènerait des biais dans l'analyse, alors qu'un seuil trop haut, laissant que peu de données au dessus du seuil mènerait à une grande variance. Le choix optimal de ce seuil dépend de paramètres inconnus en pratique. Actuellement, les méthodes existantes sont majoritairement graphiques (Davison et R. L. Smith 1990 ; Coles et al. 2001). Il n'existe pas à notre connaissance, de méthode automatique. On peut ainsi aussi se reposer sur l'expertise métier pour choisir un seuil tout en assurant une adéquation raisonnable d'une loi GP sur les excès.

Un des avantages de cette méthode est qu'elle permet de comprendre le comportement des événements situés dans la queue de la distribution, à savoir ceux qui ont dépassé un certain seuil. Dans le contexte des risques naturels, une majorité des dommages rapportés est due un petit nombre d'événements très intenses. La théorie des valeurs extrêmes, et en particulier la méthode du Peaks over Threshold, permet donc d'étudier ces événements.

Nous verrons dans le chapitre 6, comment cette méthode peut être couplée à des arbres de régression et à la théorie de la crédibilité pour estimer le coût d'un événement inondation en fonction de ses caractéristiques.

Nous introduisons dans la partie suivante les notations de la théorie de la crédibilité.

2.2 Théorie de la crédibilité

La théorie de la crédibilité est assez largement utilisée en assurance, on peut trouver une description détaillée dans le cours de (Bühlmann et Gisler 2005). Elle trouve ses origines dans les travaux de (Mowbray 1914). C'est aujourd'hui une méthode de référence pour la tarification des primes qui permet de prendre en compte l'expérience et le profil de risque. Nous proposons ici une brève introduction des concepts et notations utilisées et nécessaires pour notre étude.

Notons :

— I, classiquement le nombre d'assurés ;
— n_i, le nombre d'années d'observations ;
— $Y_{i,j}$, le montant des sinistres pour l'événement j pour l'assuré i ;
— $\mathbb{Y} = (Y_{i,j})_{j=1,....n_i,i=1.....I}$;
— $\theta_i \in \Theta$, le profil de risque de l'assuré.

En reprenant la terminologie bayésienne, on peut écrire que, θ, notre profil de risque, est une variable aléatoire distribuée de loi $\mathbb{T}$ que l'on appelle loi a priori. Les $Y_{i,j}$ sont conditionnellement à θ_i des variables aléatoire i.i.d de loi F_{θ_i}. La loi de θ_i sachant les $Y_{i,j}$ est, elle, la loi a posteriori. La prime individuelle π_i est égale à l'espérance de la loi conditionnelle des $Y_{i,j}$ sachant θ_i. La prime collective, à l'espérance des $Y_{i,j}$, cela correspond à l'espérance de tous les assurés. La prime

de crédibilité est égale à l'espérance conditionnelle de $Y_{i,n+1}$ sachant $Y_{i,1}, ..., Y_{i,n}$.

$$\begin{aligned}\pi_{ind,i} &= \mathbb{E}[Y_{i,n_i+1} \mid \theta_i] \\ \pi_{col} &= \mathbb{E}[\mathbb{Y}] \\ \pi_{cred,i} &= \mathbb{E}[Y_{i,n_i+1} \mid Y_{i,1}, ...Y_{i,n_i}]\end{aligned}$$

Considérons maintenant un assuré i avec $(Y_{i,1}, ..., Y_{i,n_i})$ le montant des sinistres pour l'événement. Soit $g = g_t$ la densité d'une loi de paramètre t. On a $\theta_i \sim \mathbb{T}$ avec $\mathbb{T}$ la loi a piori, une loi sur le paramètre de la loi de g. Les $(Y_{i,j})_j$ sont conditionnellement à $\theta_i = t$, i.i.d de loi g_t.

Pour trouver $\pi_{cred,i}$ la prime de crédibilité, on doit connaître alors la loi a posteriori de θ_i soit la loi de θ_i sachant les Y_{ij}.

$$\begin{aligned}\pi_{cred,i} &= \mathbb{E}[Y_{i,n_{i+1}} \mid Y_{i,1}, ...Y_{i,n_i}] \\ &= \int_y y f_{Y_{i,n_i+1}|Y_{i,1},....Y_{i,n_i}}(y)\mathrm{d}y_{n+1} \\ &= \int_y y \int_t f_{Y_{i,n_i+1},\theta_i|Y_{i,1},....Y_{i,n_i}}(y,t)\mathrm{d}t\mathrm{d}y \\ &= \int_y \int_t y g_t(y) f_{\theta_i|Y_{i,1}...Y_{i,n_i}}\mathrm{d}t\mathrm{d}y.\end{aligned}$$

On peut remarquer que

$$\pi_{cred,i} = \int_t \mathbb{E}[Y_{i,n_{i+1}}|\theta_i = t] f_{\theta_i|Y_{i,1}...Y_{i,n_i}}(t)\mathrm{d}t.$$

Par définition des densités conditionnelles, on a

$$f_{\theta_i|Y_{i,1}=y_1,...,Y_{i,n_i}=y_{n_i}}(t) = \frac{f_{Y_{i,1},....,Y_{i,n_i},\theta_i}(y_1, ..., y_{n_i}, t)}{f_{Y_{i,1},....,Y_{i,n_i}}(y_1, ..., y_{n_i})}.$$

On a pour le numérateur

$$\begin{aligned}f_{Y_{i,1},....,Y_{i,n_i},\theta_i}(y_1, ..., y_{n_i}, t) &= f_{Y_{i,1},....,Y_{i,n_i}|\theta_i=t}(y_1, ..., y_{n_i}) f_{\theta_i}(t) \\ &= g_t(y_1)....g_t(y_{n_i}) f_{\theta_i(t)},\end{aligned}$$

car les $(Y_{i,j})_j$ sont i.i.d de loi g_t conditionnellement à $\theta_i = t$ et pour le dénominateur,

$$\begin{aligned}f_{Y_{i,1},....,Y_{i,n_i}}(y_1, ..., y_{n_i}) &= \int_t f_{Y_{i,1},....,Y_{i,n_i},\theta_i}(y_1, ..., y_{n_i}, t)\mathrm{d}t \\ &= \int_t g_t(y_1)....g_t(y_{n_i}) f_{\theta_i(t)}\mathrm{d}t,\end{aligned}$$

d'où

$$f_{\theta_i|Y_{i,1}=y_1,...Y_{i,n_i}=y_{n_i}}(t) = \frac{g_t(y_1)....g_t(y_{ni}) f_{\theta_i}(t)}{\int_s g_s(y_1)....g_s(y_{n_i}) f_{\theta_i}(s)\mathrm{d}s}.$$

Ici $\int_s g_s(y_1)....g_s(y_{n_i}) f_{\theta_i}(s)\mathrm{d}s$ joue le rôle de constante de normalisation pour assurer que $g_t(y_1)....g_t(y_{n_i}) f_{\theta_i}(t)$ soit bien une densité de probabilité. En pratique on calcule le numérateur et on essaye de reconnaître une loi usuelle pour ne pas calculer l'intégrale.

Une des difficultés est d'estimer la loi a priori, on verra en chapitre 6 une méthode reposant

sur les lois GP et les arbres de régression. Nous présentons dans la partie suivant ces arbres ainsi que toutes les méthodes d'apprentissage statistique utilisées dans le chapitre 5.

2.3 Modèles d'apprentissage statistique utilisés

2.3.1 Introduction

L'apprentissage statistique s'est largement développé ces dernières années et est maintenant utilisé dans tous les domaines. On trouve de nombreuses applications en assurance et plusieurs récentes thèses traitent de ce sujet comme par exemple (Piette 2019 ; Ly 2019 ; Baudry 2020). Cela fait en partie suite à l'augmentation croissante de la quantité de données disponibles et l'assurance est pleinement concernée. En plus des bases de données usuelles sur la sinistralité dont le volume croit avec les années et les changements technologiques, les actuaires et « data scientists » des sociétés d'assurance ont à leurs dispositions de plus en plus de données géographiques et satellites notamment, permettant d'affiner la connaissance du risque, mais aussi des données textuelles de plus en plus fines et nombreuses. Ces travaux de thèse s'inscrivent pleinement dans cet environnement, comme décrit dans le chapitre suivant, nous avons accès à des bases de données volumineuses relatives à la sinistralité pour nos analyses et grâce à l'analyse de données textuelles, nous cherchons aussi à obtenir de nouvelles données. Face à ces données, les méthodes d'apprentissage statistique sont particulièrement efficaces et permettent de nombreuses applications. Comme décrit dans la thèse (Ly 2019), ce changement de paradigme a imposé un modèle d'actuaire data scientist reposant de plus en plus sur l'apprentissage statistique, dans lequel nous nous inscrivons.

Dans cette section, nous décrivons les différents modèles d'apprentissage statistique que nous avons utilisés dans cette thèse dans un cadre général, avant de décrire dans leur chapitre respectif, leurs applications dans un cadre plus spécifique. Nous commençons par les modèles paramétriques classiques avant de regarder des modèles tels que les arbres de régression permettant d'introduire des non-linéarités puis nous présentons des modèles plus complexe dit de « boîtes noires ».

2.3.2 Modèle linéaire généralisé pénalisé

Le modèle linéaire généralisé (GLM), voir par exemple (Nelder et Wedderburn 1972 ; Denuit et Charpentier 2005), est une manière générique de considérer les problèmes de régression qui est largement utilisée dans le domaine de l'assurance. Cette classe de modèles stipule que, pour une variable réponse Y et des covariables X dans $\mathbb{R}^p$,

$$g\left(\mathbb{E}[Y|X]\right) = X\beta,$$

avec $\beta \in \mathbb{R}^p$ est le vecteur des paramètres inconnus, et g une certaine fonction monotone connue, appelée fonction de lien. De plus, la distribution conditionnelle de Y sachant X est supposée appartenir à une certaine famille de lois exponentielle.

Par exemple, dans un problème de classification binaire, la distribution de $Y \mid X$ est supposée être une distribution de Bernoulli avec un paramètre inconnu $p(X) = \mathbb{E}[Y|X] = \mathbb{P}(Y \mid X)$. En ce qui concerne la fonction de lien g, un choix standard consiste à prendre $g(y) = \text{logit}(y) = \log\left(y/(1-y)\right)$. Cela correspond à la fonction de lien canonique, la fonction de lien conduisant aux meilleures propriétés théoriques pour les GLM. Il s'agit également d'une fonction simple qui fait correspondre $[0, 1]$ à $\mathbb{R}$.

L'estimation peut être effectuée par maximum de vraisemblance. Cependant lorsque la dimension p des covariables est relativement élevée, ce qui sera notre cas, cela pose un problème

puisque la précision statistique diminue avec le nombre de coefficients à estimer. De nombreux problèmes numériques peuvent aussi survenir. D'autre part, la plupart des variables sont susceptibles d'être non pertinentes, sans savoir lesquelles à l'avance. Cela nuit à l'interprétabilité du modèle, qui est un des avantages du GLM. Ainsi, l'estimateur GLM elastic-net (GLMNET) est un bon moyen de réduire la dimension en résolvant l'instabilité numérique grâce à la pénalisation (Zou et Hastie 2005).

Soit $f_\beta(y, x)$ la vraisemblance du modèle. L'estimateur GLMNET est défini comme suit

$$\widehat{\beta} = \arg\max_{\beta} \sum_{i=1}^{n} \log(f_\beta(Y_i, X_i)) - \lambda \left\{ \alpha \|\beta\|_1 + (1-\alpha)\|\beta\|_2 \right\},$$

avec $\|\beta\|_1$ (resp. $\|\beta\|_2$) désignant la norme l^1- (resp. l^2-) du vecteur β, les hyper-paramètres λ et α étant des constantes positives. La pénalisation de la log-vraisemblance par $\|\beta\|_2$ correspond à une pénalisation de Ridge, voir (Marquardt et Snee 1975), qui stabilise le résultat de l'estimation en réduisant certains problèmes numériques qui peuvent survenir en haute dimension. La pénalisation par $\|\beta\|_1$ correspond à une pénalité Lasso, voir (Robert Tibshirani 1996), conçue pour produire un modèle plus parcimonieux, c'est-à-dire un modèle dans lequel la plupart des coefficients de $\widehat{\beta}$ sont égaux à zéro. Elle permet donc de réduire la dimension effective des covariables. La pénalisation elastic-net est un bon moyen de prendre en compte les différences et avantages des deux formes de pénalisations. Elle permet aussi de mieux traiter les variables explicatives corrélées entre elles. La pénalisation lasso a tendance à ne sélectionner qu'une seule variable dans ce cas et la pénalisation ridge à estimer les coefficients avec la même valeur. En faisant varier α, on choisit le type de pénalisation que l'on veut prendre en compte. La constante de pénalisation, λ est usuellement choisie par k-fold-cross-Validation.

Cela consiste à sélectionner aléatoirement k sous-échantillons dans nos données. On va ensuite fixer λ_k et pour chaque k, entraîner le modèle sur les données en enlevant le k sous-échantillon. Ensuite on applique ce modèle au k sous-échantillon, ce qui nous donne des prédictions y_k. En faisant varier pour chaque k, λ_k, on obtient plusieurs y_k et l'on peut sélectionner le meilleur selon un critère adapté, par exemple la précision ou l'aire sous la courbe ROC (AUC ROC) dans le cas d'une classification. Cela permet d'identifier les paramètres permettant d'avoir les meilleures prédictions, cependant le préalable est d'avoir assez de données pour pouvoir faire des sous-échantillons.

L'avantage de GLMNET est de produire un modèle intelligible et facilement interprétable. D'autre part, le fait de pouvoir sélectionner automatiquement les variables qui ont un effet sur Y nous permet de considérer un modèle suffisamment complexe pour espérer un bon ajustement. Néanmoins, l'hypothèse paramétrique sous-jacente peut être trop forte en pratique. C'est pourquoi en pratique nous avons souvent recours à des techniques « boîtes noires » de l'apprentissage statistiques.

2.3.3 Arbres de régression

Les arbres de régression, introduits par (Breiman et al. 1984), font eux aussi partie des outils simples et interprétables largement utilisés dans le secteur de l'assurance. Ils présentent de nombreuses propriétés intéressantes, comme la possibilité d'introduire des non-linéarités tout en produisant un modèle facilement compréhensible. Le but est de constituer des classes d'observations qui ont un comportement similaire relativement à une variable réponse Y. Pour cela, via l'algorithme CART (Classification And Regression Tree), on définit des « règles » qui affectent les observations dans des classes selon les valeurs de ces covariables X. Cette procédure à deux phases, une première phase de construction de l'arbre maximal puis une seconde phase d'élagage

de l'arbre.

Construction de l'arbre maximal Dans l'algorithme CART on détermine par itération des « règles » $\mathbf{x} = (x^{(1)}, \ldots, x^{(d)}) \to R_j(\mathbf{x})$ pour partitionner les observations, selon des critères. Pour chaque valeur des covariables X, $R_j(\mathbf{x}) = 1$ ou 0 selon si certaines conditions sont satisfaites ou non, avec $R_j(\mathbf{x})R_{j'}(\mathbf{x}) = 0$ pour $j \neq j'$ et $\sum_j R_j(\mathbf{x}) = 1$. On peut se représenter ces règles comme des arbres de décision, chaque règle R_j à l'étape k génère deux règles R_{j1} et R_{j2} à l'étape $k+1$, avec $R_{j1}(\mathbf{x}) + R_{j2}(\mathbf{x}) = 0$ si $R_{j1}(\mathbf{x}) = 0$. On cherche à extraire une fonction de régression dans chaque classe, $\widehat{\theta}(\mathbf{X})$, appartenant à une certaine classes, tel que, $\sum_{i=1}^{n} \phi(Y_i - u(\mathbf{X}_i), \widehat{\theta}(\mathbf{X}_i))\mathbf{1}_{Y_i \geqslant u(\mathbf{X}_i)}$ soit maximal. Pour alléger les notations on note, $\varphi(Y_i, \theta) = \phi(Y_i - u(\mathbf{X}_i), \theta)\mathbf{1}_{Y_i \geqslant u(\mathbf{X}_i)}$.

Détaillons les étapes :

Étape 1 : $R_1(\mathbf{x}) = 1$ pour tous les $\mathbf{x}$,, et $n_1 = 1$, c'est la racine de l'arbre.

Étape k+1 : Avec (R_1, ...,R_{n_k}) les règles obtenues à l'étape k. Pour $j = 1, \ldots n_k$,

— si toutes les observations telles que $R_j(\mathbf{X}_i) = 1$ ont les mêmes caractéristiques alors on garde la règle R_j, on ne peut pas faire plus de partition,

— sinon, la règle R_j est remplacée par deux nouvelles règles, R_{j1} et R_{j2} que l'on détermine en définissant pour chaque $X^{(\ell)}$ de $\mathbf{X} = (X^{(1)}, \ldots, X^{(d)})$ le meilleur seuil $x_{j\star}^{(\ell)}$ permettant de séparer les données, de telle sorte que $x_{j\star}^{(\ell)} = \arg\max_{x^{(\ell)}} \Phi(R_j, x^{(\ell)})$, avec

$$\begin{aligned}\Phi(R_j, x^{(\ell)}) &= \sum_{i=1}^{n} \varphi(Y_i, \theta_{\ell-}(\mathbf{X}_i, R_j))\mathbf{1}_{X_i^{(\ell)} \leqslant x^{(\ell)}} R_j(\mathbf{x}) \\ &+ \sum_{i=1}^{n} \varphi(Y_i, \theta_{\ell+}(\mathbf{X}_i, R_j))\mathbf{1}_{X_i^{(\ell)} > x^{(\ell)}} R_j(\mathbf{x}),\end{aligned}$$

où

$$\begin{aligned}\widehat{\theta}(R_j) &= \arg\max_{\theta \in \Theta} \sum_{i=1}^{n} \varphi(Y_i, \theta) R_j(\mathbf{X}_i), \\ \theta_{\ell-}(x, R_j) &= \arg\max_{\theta \in \Theta} \sum_{i=1}^{n} \varphi(Y_i, \theta)\mathbf{1}_{X_i^{(\ell)} \leqslant x} R_j(\mathbf{X}_i), \\ \theta_{\ell+}(x, R_j) &= \arg\max_{\theta \in \Theta} \sum_{i=1}^{n} \varphi(Y_i, \theta)\mathbf{1}_{X_i^{(\ell)} > x} R_j(\mathbf{X}_i).\end{aligned}$$

On sélectionne ensuite le meilleur indice $\widehat{\ell} = \arg\max_\ell \Phi(R_j, x_{j\star}^{(\ell)})$.

On définit deux nouvelles règles, $R_{j1}(\mathbf{x}) = R_j(\mathbf{x})\mathbf{1}_{x^{(\widehat{\ell})} \leqslant x_{j\star}^{(\widehat{\ell})}}$, and $R_{j2}(\mathbf{x}) = R_j(\mathbf{x})\mathbf{1}_{x^{(\widehat{\ell})} > x_{j\star}^{(\widehat{\ell})}}$.

— On incrémente à n_{k+1} règles

Condition d'arrêt : on s'arrête si $n_{k+1} = n_k$

Cette procédure à bien une structure d'arbre de décision, les règles $(R_j)_{1 \leq j \leq n_k}$ sont identifiées aux feuilles de l'arbre à l'étape k. Le nombre de feuilles croît en passant de l'étape k à $k+1$. A chaque étape les deux règles forment une décision possible.

Dans cette version toutes les covariables sont continues ou binaires (0,1). Les variables catégorielles doivent être codées en variables binaires au préalable. On peut aussi modifier l'algorithme pour que les critères de séparations de chaque R_j trouvent quelle modalité de la variable catégorielle minimise la fonction de perte. Il est aussi possible de modifier la condition d'arrêt pour

assurer un minimum d'observations dans chaque feuille de l'arbre.

Fonction de régression Avec notre ensemble de règles $\mathcal{R} = (R_j)_{j=1,\dots,s}$, soit $\mathcal{T}_j = \{\mathbf{x} : R_j(\mathbf{x}) = 1\}$, la $j$ème feuilles de l'arbre correspondant. L'estimateur $\widehat{\theta}$ associé avec un arbre $\mathcal{T} = (\mathcal{T}_\ell)_{\ell=1,\dots,K}$ (où K est le nombre total de feuilles) est obtenu par

$$\widehat{\theta}(\mathbf{x}) = \sum_{\ell=1}^{K} \widehat{\theta}(R_j) R_j(\mathbf{x}) = \sum_{\ell=1}^{K} \widehat{\theta}_\ell \mathbf{1}_{\mathbf{x}\in\mathcal{T}_\ell}.$$

L'arbre maximal, $T_{\max}$ est obtenu quand l'algorithme précédent s'arrête. Il correspond à un estimateur trivial puisque soit le nombre d'observation est de 1 soit elles ont toutes les mêmes caractéristiques $\mathbf{x}$. On passe ensuite passer à l'élagage de l'arbre qui consiste à choisir un sous-arbre optimal en faisant un compromis entre la simplicité de l'arbre et son ajustement.

Élagage de l'arbre Pour l'élagage, une approche usuelle est d'utiliser de la pénalisation pour trouver un sous-arbre adapté, (Gey et Nedelec 2005). Pour un arbre T_K avec K feuilles $(\mathcal{T}_\ell)_{\ell=1,\dots,K}$, associé à l'estimateur $\widehat{\theta}$, la performance de l'arbre est mesuré par

$$\frac{1}{k_n} \sum_{\ell=1}^{K} \sum_{i=1}^{n} \varphi(Y_i - u, \widehat{\theta}(\mathbf{X}_i)) \mathbf{1}_{\mathbf{X}_i\in\mathcal{T}_\ell} - \lambda K. \tag{2.3.1}$$

On peut interpréter λ comme un coefficient de pénalisation car il va imposer de privilégier les arbres avec un petit nombre de feuilles pour minimiser 2.3.1.

Pour déterminer ce sous arbre, il n'est pas nécessaire de calculer chaque sous-arbre. Il suffit de déterminer pour $K \geq 0$ le sous-arbre T_K qui minimise ce critère parmi tous les autres sous-arbres à K feuilles et ensuite choisir l'arbre final parmi une liste de $K_{\max}$ arbres. Selon (Breiman et al. 1984), cet arbre T_K est simple à déterminer, on obtient chaque T_K en enlevant une feuille à T_{K+1}

Le critère de pénalisation λ est ensuite choisi par k-fold-cross-validation.

Le principal inconvénient des arbres de régression est leur instabilité, de nouvelles données entrantes peuvent perturber la structure de la partition.

2.3.4 Forets aléatoires

Les forêts aléatoires (Random Forets, RF) sont un moyen de stabiliser les arbres de régression. Ils constituent une procédure d'apprentissage automatique reposant sur l'agrégation d'arbres de régression (Breiman 2001). Ils sont obtenus en faisant la moyenne d'arbres de régression avec certaines spécificités :

- chaque arbre est de petite taille ;
- chaque arbre se développe sur un échantillon bootstrap séparé ;
- les règles d'un arbre donné ne sont fondées que sur un petit sous-ensemble de covariables (sous-ensemble sélectionné au hasard).

On construit donc une forêt en prenant l'agrégation de petits arbres construit sur un échantillon boostrap de données. Cette méthode est appelée Bagging (Boostrap-aggregating) et a été introduite par (Breiman 1996).

Dans un premier temps on tire des échantillons boostrap avec remise, on construit des arbres de régression sur ces sous-échantillons et l'on agrège ensuite les modèles obtenus en prenant la moyenne. Cela conduit expérimentalement à de meilleurs résultats et Breiman a proposé la combinaison des arbres de régression CART et du bagging dans (Breiman 1996). Il a ensuite

proposé le concept de forêts aléatoires dans (Breiman 2001) en introduisant de l'aléatoire dans le choix des covariables à chaque noeud. On ne cherche plus le critère optimal à chaque noeud parmi l'ensemble des critères mais seulement parmi un sous-ensemble de critères se basant sur certaines covariables tirées au hasard. Chaque prédicteur est moins précis mais l'estimateur agrégé lui devient plus précis et moins sensible aux variations. Une autre différence par rapport à l'algorithme CART décrit dans la partie précédente est que dans le cadre des forêts aléatoires on ne cherche pas à élaguer l'arbre obtenu pour trouver un sous-arbre optimal, mais on fait plusieurs arbres de petites tailles en contraignant le nombre de feuilles maximal. En reprenant les notations de la partie précédente :

$$\hat{\theta}_{RF} = \frac{1}{\Theta}\sum_{j=1}^{\Theta}\hat{\theta}_j,$$

avec $\hat{m}_j$ obtenue pour chaque sous-arbre en contraignant sa taille et sa construction sur un échantillon bootsrap.

Cet algorithme permet d'améliorer grandement les prédictions et du fait de sa construction c'est un algorithme rapide à calculer. C'est particulièrement intéressant en grande dimension, ce qui sera notre cas. Cependant on perd le caractère interprétable et simple des arbres de régressions et cette méthode peut être qualifiée de « boite noire ».

2.3.5 Extreme Gradient Boosting

L'Extreme Gradient Boosting (XGBOOST), voir (Chen et Guestrin 2016), est une méthode alternative aux RF qui repose également sur des arbres de régression, mais au lieu d'ajuster ces arbres simultanément, ils sont ajustés de manière itérative. On utilise ici le boosting (Schapire 1990) et non plus le bagging. Les arbres sont entraînés sur les résidus des arbres précédents.

Le prédicteur $\widehat{m}^{(t)}(x)$ à la $t-$ième étape de l'algorithme est obtenu à partir du prédicteur $\widehat{m}^{(t-1)}(x)$ à la $(t-1)-$ième étape par $\widehat{m}^{(t-1)}(x) + \pi_t(x)$, où $\pi_t(x)$ est un arbre de régression sélectionné de manière à faire diminuer la fonction de perte autant que possible, c'est-à-dire à maximiser la log-vraisemblance dans le cas de Bernoulli avec une pénalité de régularisation.

Soit $\ell(y_i, \widehat{m}^{(t-1)}(x_i))$ la log-vraisemblance négative pour l'observation i (y_i, x_i) à l'étape $t-1$ (cette fonction est également appelée entropie croisée « cross-entropy ») dans la littérature sur le « machine learning »).

À l'étape t, l'algorithme tente de trouver π_t qui minimise

$$\sum_{i=1}^{n}\partial_2\ell(y_i, \widehat{m}^{(t-1)}(x_i)) \times \pi_t(x_i) + \frac{1}{2}\partial_2^2\ell(y_i, \widehat{m}^{(t-1)}(x_i)) \times \pi_t^2(x_i) + pen(\pi_t),$$

où *pen* désigne la pénalité de régularisation, et ∂_2 (resp. ∂_2^2) désigne la dérivée partielle (resp. du second ordre) d'une fonction par rapport à son second argument.

Cet algorithme est devenu aussi très populaire dans l'apprentissage statistique en donnant de très bonnes prédictions. Il est rapide et l'on peut contrôler de nombreux paramètres. Les mêmes reproches que pour les forêts aléatoires s'appliquent néanmoins, le caractère simple et interprétable est perdu dans l'agrégation d'arbres.

2.4 Analyse de texte

2.4.1 Introduction

Nous nous intéressons ici à l'analyse de texte de manière automatique via des algorithmes, qui est désignée par les termes « Text Mining » ou « Natural Language Processing » (NLP) dans l'industrie de l'assurance, malgré des différences de concept entre les deux termes. Le Text Mining se référant plutôt à l'extraction des termes et le NLP à son interprétation. C'est un champ qui s'est beaucoup développé ces dernières années, et qui devient de plus en plus accessible.

Le secteur de l'assurance est directement concerné, dans les processus de gestion beaucoup de données textuelles sont créées. On peut citer les contrats, les constats, les pièces justificatives et dans notre étude les rapports d'expertise. Ces données sont stockées mais ne sont pas exploitées systématiquement. L'extraction et l'analyse de ces données, « Text Mining », voir (Feldman, Sanger et al. 2007), est un enjeu important pour le secteur de l'assurance et l'on peut trouver des études sur ce sujet dès le début du champ, voir par exemple, (Francis 2006 ; Ellingsworth et Sullivan 2003 ; Kolyshkina et Rooyen 2006). Les progrès récents dans le NLP ont augmenté les possibilités en permettant de fournir du contexte et d'une certaine façon du sens dans l'analyse du texte. Une analyse des progrès des 20 dernières années dans le contexte de l'assurance peut être trouvée dans (Zappa et al. 2021) et une application de ces techniques dans (Xu 2021). Ce sont des perspectives intéressantes pour l'exploitation industrielles de ces données et nous présentons au chapitre 4 un cas d'étude. Dans cette section nous présentons les principaux outils mathématiques que nous allons utiliser et aussi les avancés majeures récentes du domaine dans un cadre général.

2.4.2 Réseaux de neurones

Comme évoqués précédemment, les réseaux de neurones sont devenus incontournables dans l'état de l'art du machine learning. Pour l'analyse de texte aussi ils se sont imposés comme des outils très puissants. Les réseaux de neurones sont apparus il y a relativement longtemps, par exemple (Rosenblatt 1961) sur le perceptron, mais depuis les années 2010 ils ont pris un essor particulier avec le « Deep Learning » (G. James et al. 2021). Ils sont maintenant largement utilisés et des bonnes descriptions peuvent être trouvées dans (Goodfellow, Bengio et Courville 2016 ; Hastie et al. 2009), ou plus récemment mise à jour dans (G. James et al. 2021).

Un réseau de neurones va prendre en entrée un vecteur de p variables $X = (X_1, ...X_p)$, chaque élément de ce vecteur est ensuite pondéré et en fonction de l'information appris, le signal est transmis ou pas. Cela est fait par le biais d'une fonction d'activation. Le signal de sortie est ensuite calculé via l'application de cette fonction d'activation au vecteur d'entrée pondéré. Appelons $f(X)$ la fonction non-linéaire permettant de prédire Y à partir de X. Le réseau de neurones est de la forme suivante :

$$f(X) = \beta_0 + \sum_{k=1}^{n} \beta_k h_k(X),$$

$$f(X) = \beta_0 + \sum_{k=1}^{n} \beta_k g(w_{k0} + \sum_{j=1}^{p} w_{kj} X_j).$$

On construit le réseau en deux temps, dans un premier temps les Activations, $A_k = h_k(X) = g(w_{k0} + \sum_{j=1}^{p} w_{kj} X_j)$, sont construits comme des fonctions du vecteur d'entrée. Avec g une

fonction non-linéaire, appelé fonction d'activation, que l'on spécifie avant. Ensuite ces Activations de la couche cachée alimentent la couche de sortie :

$$f(X) = \beta_0 + \sum_{k=1}^{n} \beta_k A_k.$$

Ce processus est schématiquement résumé en 2.2. Les paramètres, $\beta_0, ...\beta_k$ et w_{k0} sont estimé par apprentissage sur les données. On cherche classiquement à minimiser l'erreur via l'optimisation d'une fonction de coût. L'algorithme classique de correction des erreurs est la rétropropagation du gradient, qui cherche de manière itérative une configuration optimale des poids.

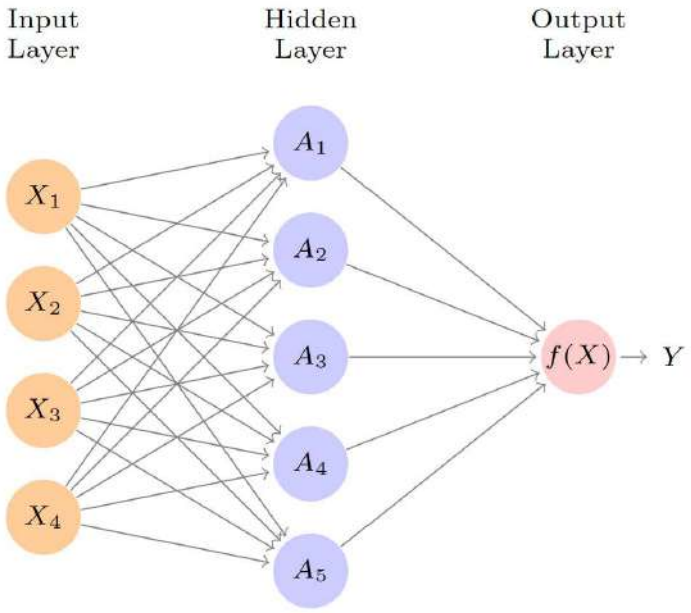


FIGURE 2.2 – Réseau de neurone à une couche cachée (Source (G. JAMES et al. 2021))

Le choix de la fonction d'activation permet d'ajuster la représentation du vecteur d'entrée. On peut citer deux fonctions usuelles, la fonction ReLu, qui est très largement utilisée :

$$g(z) = \max(0, z),$$

et la fonction Sigmoid, qui donne une valeur continue entre 0 et 1 ce qui la rend très utilisée pour la classification binaire :

$$g(z) = \frac{1}{1 + e^{-z}}.$$

Ici on a décrit un réseau de neurones avec une seule couche cachée mais en pratique les réseaux sont multicouches, ce qui rend vite le nombre de paramètres des réseaux très grand. Cela contribue en partie au caractère « boite noire » des réseaux de neurones.

2.4.3 Représentation des mots

One-hot encoding Pour pouvoir utiliser des réseaux de neurones, ou tout autre modèle pour faire de l'analyse de texte, il faut au préalable représenter le texte sous forme de covariable. Pour cela une première approche est de créer une matrice représentant tous les mots de notre corpus de mots. Cette représentation est appelée « sac de mots » ou « Bag of Words » (BoW). Il existe plusieurs méthodes permettant de faire cela, une première méthode est de faire du « One-hot encoding ». Si notre corpus est de taille C avec t mots ou termes alors on va représenter le corpus par une matrice de C lignes et t colonnes, $M = (m_{ij})$, avec $1 \leq i \leq C, 1 \leq j \leq t$. Le terme m_{ij} est soit égale à 1 si le mot j correspond au i^{me} mot du corpus et 0 sinon. On obtient alors une matrice avec seulement un 1 sur chaque ligne.

On peut prendre en compte la récurrence d'un mot en considérant la fréquence au lieu d'une valeur binaire. Pour cela on peut calculer la fréquence du terme ou mots, « Term frequency » (TF) avec :

$$tf(t, C) = \frac{n_{t,C}}{\sum_{t' \in C} n_{t',C}}.$$

Le numérateur est le nombre de mots dans le document et le dénominateur correspond au nombre total de mots dans le document. Pour pondérer les mots trop fréquemment utilisés n'apportant pas de sens on peut utiliser la pondération par « Term frequency–inverse document frequency » (TF-IDF). Pour cela on considère le nombre de documents d dans notre corpus C et l'on calcule la fréquence inverse du document, « inverse document frequency » (IDF) comme :

$$idf(t, C) = log \frac{|C|}{d \in C : t \in d},$$

avec $|C|$ le nombre de documents dans le corpus, $d \in C : t \in d$ le nombre de documents où apparaît le mot t dans le corpus, pour éviter de diviser par 0 il est courant d'ajouter 1. On calcule ensuite le « Term frequency–inverse document frequency »(TF-IDF) comme :

$$tfidf(t, d, C) = tf(t, d).idf(t, C).$$

Ces méthodes sont décrites dans (Silge et Robinson 2017) par exemple. Elles sont simples d'implémentation et peuvent donner des résultats intéressants, cependant elles ne prennent pas en compte le contexte des mots seulement leurs occurrences. Pour cela on peut aussi considérer des n-grams au lieu des mots, la méthode reste similaire mais on considère plusieurs mots ensemble dans la matrice. Le terme de la matrice ne correspond plus à un mot mais à une séquence consécutive de n mots. On augmente alors la taille du corpus par le n choisi pour le n-grams. Avec ces méthodes on transforme un texte dans un format exploitable par un réseau de neurones et des modèles de machine learning. Une première limite de cette méthode est la taille de la matrice. Si le corpus est trop grand on se retrouve avec une matrice très volumineuse contenant pourtant peu d'informations. Cette méthode n'est pas parcimonieuse et cela peut nuire aux performances de certains modèles. De plus on n'utilise les relations entre les mots que quand ils sont exactement égaux, la notion de distance lexicale n'est pas du tout prise en compte.

Word Embedding Pour dépasser ces limites on peut utiliser le plongement de mots ou « Word Embedding ». Cela consiste à représenter les mots par des vecteurs denses de valeur numérique, comme illustrée en 2.3. Cela permet de réduire la dimension et de faire en sorte que des mots similaires ont un encodage proche, voir par exemple, (Kusner et al. 2015 ; Bojanowski et al. 2017). On représente les mots dans un espace de dimension inférieure, adaptée pour établir une distance appropriée entre les mots. Dans notre corpus de taille C avec t mots, la matrice

d'embedding va être une matrice W de dimension $t \times E$, avec E la taille de l'embedding tel que $E \leq C$. On applique la matrice à chaque mot codé, x, et l'on produit une couche cachée, telle que :

$$h_i = W_x = \sum_{j=1}^{t} w_{ij} x_j.$$

La différence majeure est qu'ici le vecteur est déterminé par apprentissage alors que dans le cas de la matrice elle est codée en fonction des mots. Le but ici, est de représenter la sémantique des mots dans un espace géométrique où chaque mot correspond à un vecteur.

Un exemple connu de cette représentation est le $\vec{Roi} - \vec{Homme} + \vec{Femme} \approx \vec{Reine}$, illustré en 2.4, (Vylomova et al. 2015). On peut faire des opérations sur les vecteurs de mots qui ont un sens au niveau de la sémantique des mots. Les vecteurs n'ont pas vraiment de sens mais ils arrivent à extraire la structure statistique du corpus d'un texte et cela peut mener à des opérations intéressantes.

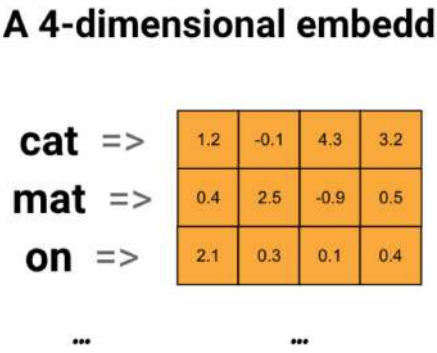


Figure 2.3 – Illustration du word embedding, (Source : https ://www.tensorflow.org)

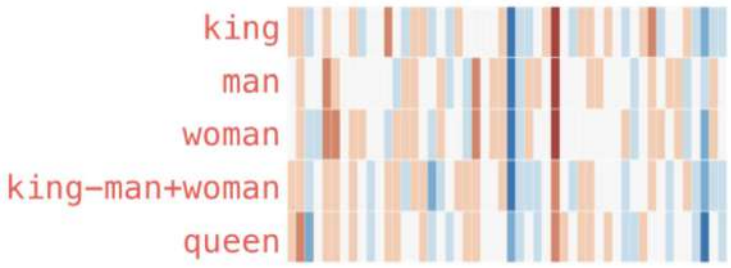


Figure 2.4 – Illustration de l'opération $\vec{Roi} - \vec{Homme} + \vec{Femme} \approx \vec{Reine}$, (Source : https ://-jalammar.github.io)

On peut réaliser le Word Embedding dans la phase d'apprentissage du réseau de neurones, les poids sont alors appris de la même manière qu'un réseau apprend des poids pour une couche cachée. La dimension E permet de définir le niveau de finesse de l'embedding et c'est un paramètre que l'on peut contrôler de la même manière que l'on contrôle la dimension d'une couche.

D'autres méthodes existent pour faire ce Word Embeddings comme par exemple Word2vec (Mikolov et al. 2013), GloVe (Pennington, Socher et Manning 2014) ou fastText (Joulin et al. 2016). On peut trouver des versions pré-entraîneés dans plusieurs langages, ce qui permet d'avoir déjà une structure représentant les liens entre les mots. Une description détaillée de ces méthodes peut être trouvée dans (Lavrač, Podpečan et Robnik-Šikonja 2021).

L'embedding permet de réduire grandement la dimension des vecteurs tout en prenant en compte la sémantique, deux mots assez proches en sens auront des vecteurs proches dans l'espace. C'est aussi très utile pour prendre en compte les fautes d'orthographe dans un corpus, dans une matrice cela va créer deux mots distincts alors qu'ici il va y avoir deux vecteurs mais si la sémantique est bien apprise ils seront proches.

2.4.4 Architectures des réseaux de neurones pour l'analyse de texte

L'architecture du réseau de neurones utilisée est une composante importante qui va impacter les performances. Une revue des architectures utilisées peut être trouvée dans (Minaee et al. 2021) et une description plus détaillée dans (Goldberg 2017). Pour l'analyse de texte deux architectures ressortent dans la littérature : les Convolutional Neural Networks (CNN) et les Recurrent neural network (RNN) plus particulièrement les Long short-term memory (LSTM).

Convolutional Neural Networks (CNN) Célèbre en analyse d'image avec notamment l'article de (Krizhevsky, Sutskever et Hinton 2012) les Convolutional Neural Networks (CNN), aussi appelées convnets ou en Français Réseaux de Neurones Convolutionnels sont également utilisés pour l'analyse de texte, (Collobert et al. 2011). Introduit par (LeCun, Bengio et al. 1995) ils sont fondés sur une structure particulière qui permet de reconnaître des motifs spatiaux dans les données d'entrée. Cela est particulièrement utile pour la classification textuelle, quand par exemple la position et l'enchaînement des mots sont importants.

Sans rentrer dans le détail de leurs constructions, on peut retenir que les CNN vont combiner deux types de couches spécialisées, qui vont permettre d'extraire des caractéristiques des données d'entrées, les couches de convolutions et les couches de pooling. L'idée est d'appliquer une fonction non-linéaire, entraînée, sur chaque terme d'une fenêtre de k-mots glissante pour chaque phrase. Cette fonction, aussi appelée filtre, transforme cette fenêtre en une valeur scalaire numérique. Plusieurs filtres sont appliqués, ce qui donne un vecteur de dimension du nombre de filtres, l. Cela permet de capturer les caractéristiques des mots dans la fenêtre. C'est la phase de convolution. Ensuite pour la phase de pooling on combine les vecteurs de toutes les fenêtres en un vecteur de dimension l. Pour cela en fonction des méthodes on peut prendre la moyenne ou la valeur max. Le but est de se concentrer sur les caractéristiques les plus importantes dans une phrase, sans regarder la position globale. Chaque filtre extrait une caractéristique différente sur la fenêtre et le pooling zoom sur la plus importante. Ces étapes sont décrites en 2.5

Dans cet exemple une couche de convolution d'une dimension (1D) et une couche de pooling sont appliquées sur une phrase. La convolution est faite avec une fenêtre de 3 mots, chaque mot est transformé en un vecteur de dimension 2 (cela correspond au word embeding de la section précédente). Les embeddings sont ensuite concaténés pour former les fenêtres de 6 mots visibles sur le schéma. Chaque fenêtre est ensuite transférée dans un filtre 6×3 qui applique une combinaison linéaire, l'opération de convolution, pour former les 7 fenêtres de dimension 3 qui servent à alimenter la couche de pooling. Dans cette phase on prend le max de chaque dimension pour former le vecteur final.

Le vecteur résultant sert ensuite de données d'entrées d'un réseau de neurones classique, pour effectuer la prédiction voulue. Le rétropropagation du gradient de la phase d'entraînement est

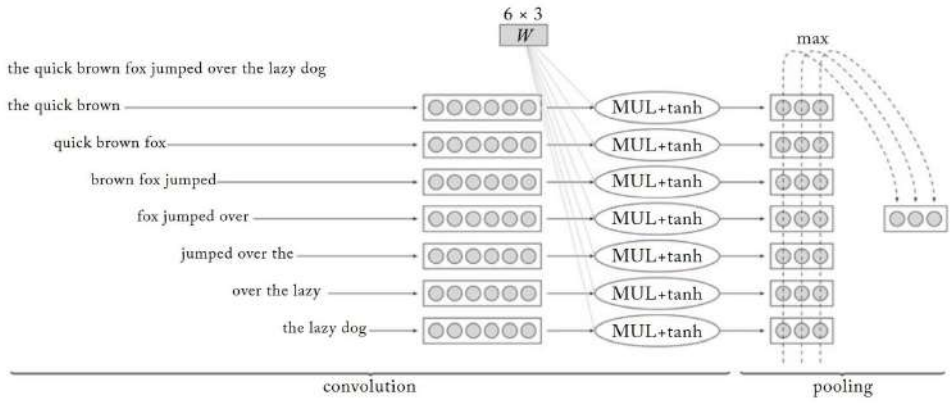


FIGURE 2.5 – Illustration d'un CNN, (Source : (GOLDBERG 2017))

utilisée pour ajuster les paramètres des filtres. Ils permettent donc de souligner les caractéristiques des données ayant un bon pouvoir prédictif pour la tâche à accomplir.

Il existe plusieurs configurations possibles pour la convolution et pour la phase de pooling et donc plusieurs CNN. Nous décrivons ici une architecture simple. On trouve de nombreuses applications pour l'analyse de texte en assurances, voir par exemple (SABBAN, LOPEZ et MERCUZOT 2020 ; LY 2019).

Recurrent neural network (RNN) Les réseaux de neurones récurrents, Recurrent neural network (RNN) introduit par (ELMAN 1990), permettent de prendre en compte la structure temporelle dans les données d'entrées et donc de gérer des données dépendantes. Pour l'analyse de texte et surtout pour la compréhension de ce texte, c'est très important car les mots ne sont pas indépendants. Le sens d'une phrase dépend souvent de la position des mots. C'est pour cela que pour le NLP, les RNN et plus particulièrement les Long short-term memory (LSTM) (HOCHREITER et SCHMIDHUBER 1997) font partie des contributions majeures des dernières années, (GOLDBERG 2017).

Pour cela les RNN utilisent des boucles qui permettent de garder l'information en mémoire. On peut voir cela comme plusieurs réseaux de neurones en série qui se fournissent des informations comme illustrés en 2.6.

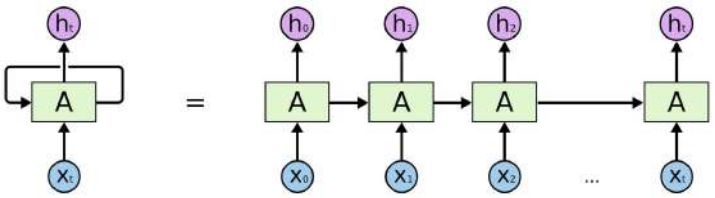


FIGURE 2.6 – Illustration d'un RNN déroulé, (Source : https ://colah.github.io/posts/2015-08-Understanding-LSTMs/)

Chaque donnée d'entrées, des mots ou des fenêtres de mots, passent successivement dans le réseau A. Le réseau produit une sortie h_t mais aussi une cellule mémoire qui est envoyée au réseau suivant. Chaque neurone utilise donc un input x_t, mais aussi l'information transmise par le réseau précédent. C'est une façon de prendre en compte l'information des mots précédents. En pratique dans les RNN simples les contributions des informations perdent en intensité à mesure que le nombre de mots augmente. Les Long Short Term Memory networks (LTSM) ont permis d'améliorer cela en gardant en mémoire les informations pertinentes à mesure qu'ils progressent. Pour cela ils utilisent une structure plus complexe décrite en 2.7.

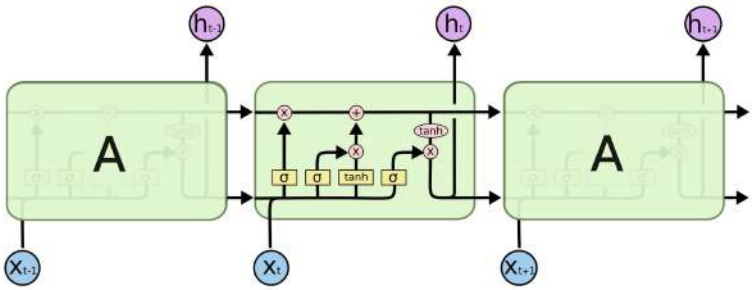


FIGURE 2.7 – Illustration d'un LTSM, (Source : https ://colah.github.io/posts/2015-08-Understanding-LSTMs/)

La flèche du bas représente la sortie h_t alors que celle d'en haut représente une sortie différente, appelée cellule c_t. C'est cette cellule qui fait la spécificité des LTSM, elle permet de transmettre l'information qui est pertinente au neurone suivant h_{t+1}. Elle transmet l'information plus directement même si le réseau à la possibilité d'enlever ou de rajouter de l'information à chaque étape via des portes (gates).

Les RNN et les LTSM sont largement utilisées dans la littérature et fournissent de très bons résultats voir par exemple en assurance (BAILLARGEON, LAMONTAGNE et MARCEAU 2021 ; SABBAN, LOPEZ et MERCUZOT 2020).

Transformers Il est difficile de parler d'analyse de texte et de réseaux de neurones sans aborder les Transformers (VASWANI et al. 2017) et notamment le modèle qui repose sur cette architecture le Bidirectional Encoder Representations from Transformers (BERT) (DEVLIN et al. 2018).

Les transformers se basent sur un mécanisme d'attention qui permet de traiter les données en même temps et non plus de manière séquentielle comme c'est le cas pour les RNN. L'attention va s'intéresser aux relations entre les mots et comment chaque mot influence les autres. Elle va calculer des poids dynamiques pour chaque mot représentant l'influence relative qu'ils ont dans la phrase. En combinant ces mécanismes d'attentions et des réseaux de neurones on peut construire les modèles transformers dont la structure complète est présentée en 2.8.

Le modèle BERT repose sur ces modèles de transformer pour être pré-entraîné sur des grands corpus de texte. Il apprend ainsi une structure syntaxique et sémantique des mots. Une différence majeure avec les autres modèles précédemment décrits est que ce modèle est bidirectionnel, il est entraîné pour prendre en compte le contexte avant et après le mot. Pour cela des mots sont masqués pendant la phase d'entraînement que le modèle essaie ensuite de prédire. Ce modèle a défini un nouvel état de l'art et l'on trouve des modèles entraînés sur des corpus de texte français comme CamemBERT (L. MARTIN et al. 2019) ou FlauBERT (LE et al. 2019) qui donne aussi de

très bon résultats. L'entraînement de ces modèles est possible car l'architecture du transformer le rend plus parallélisable que dans le cas d'un modèle classique à base de réseaux récurrents.

Une description détaillée de ces modèles et leurs nombreuses applications en assurance peut être trouvée dans (Ly, Uthayasooriyar et T. Wang 2020). Toutes ces méthodes font partie d'un ensemble d'outils en analyse de texte donnant de très bons résultats, ce qui présentent de nombreuses opportunités. Cependant, quelle que soit la performance des méthodes, on est toujours borné par la disponibilité et la qualité des données. Ce sont les facteurs les plus déterminants et nous sommes pleinement confrontés à ce problème. Dans le chapitre suivant nous décrivons le contexte industriel et les données à notre disposition, qui peuvent présenter certaines limites.

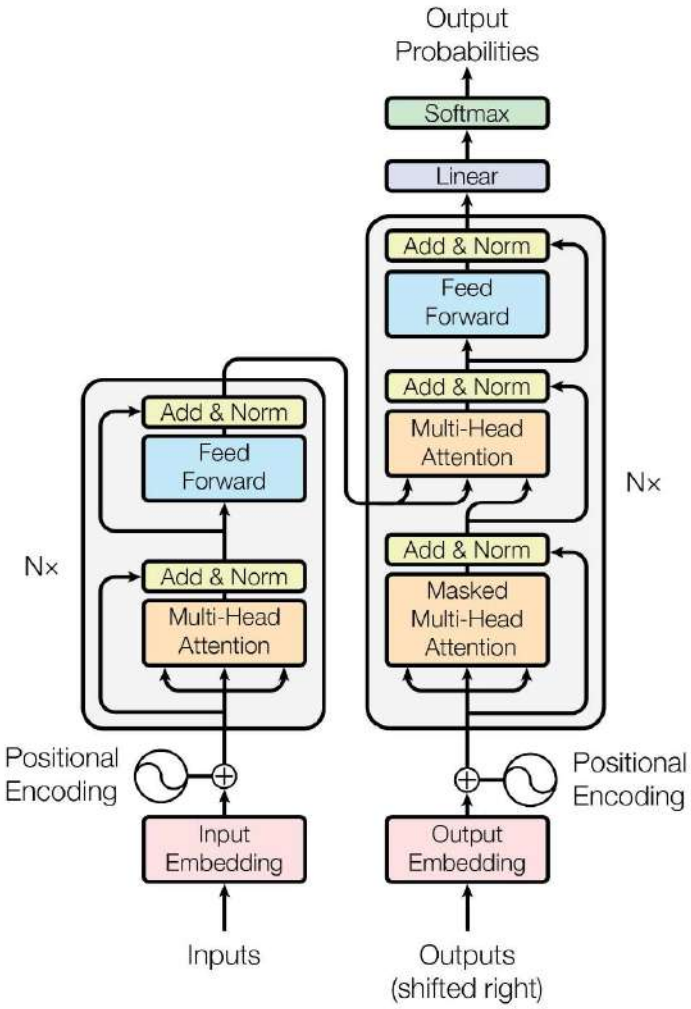


FIGURE 2.8 – Illustration du modèle Transformer, (Source : (VASWANI et al. 2017))

Chapitre 3

Contexte Industriel

Nous présentons dans ce chapitre le contexte industriel et les données utilisées pour cette thèse. Ces données serviront de base pour toutes nos applications et le contexte est important pour comprendre certaines contraintes, notamment au niveau des acteurs en jeu et des missions de l'entreprise d'accueil.

3.1 Acteurs clés

3.1.1 Mission Risques Naturels

L'association, Mission des sociétés d'assurances pour la connaissance et la prévention des Risques Naturels, abrégé en Mission Risques Naturels, a été créée en mars 2000, entre la Fédération Française des Sociétés d'Assurances (FFSA) et le Groupement des Entreprises Mutuelles d'Assurance (GEMA), aujourd'hui regroupés dans France Assureurs. Sa création fait suite aux tempêtes dévastatrices Lothar et Martin de 1999, qui sont encore à ce jour les événements climatiques les plus coûteux en France avec 13,9 Md € constants 2020, (*Etude : Changement climatique et assurance à l'horizon 2040* 2021). Il s'agit pour la profession de l'assurance de contribuer à une meilleure connaissance des risques naturels et d'apporter une contribution technique aux politiques de prévention.

La gouvernance de l'association est assurée par le conseil d'administration qui est constitué de représentants des grands groupes d'assurance en France. Cependant l'ensemble des sociétés adhérentes à France Assureurs, qui opèrent sur le marché français dans la branche « dommages aux biens des particuliers et des professionnels », sont associées au financement de son programme d'activités. C'est un groupement technique adhérent du GIE « Gestion Professionnelle des Services de l'Assurance » (GPSA) et son budget annuel est voté d'un commun accord entre ses membres.

Ses activités s'organisent, selon quatre dimensions interdépendantes, comme illustrées en 3.1 :

— Connaissance ;
— Co-construction ;
— Innovation ;
— Prévention.

Connaissance Pour pouvoir informer et sensibiliser, la MRN a développé plusieurs outils qui contribuent à une meilleure connaissance des risques naturels en France.

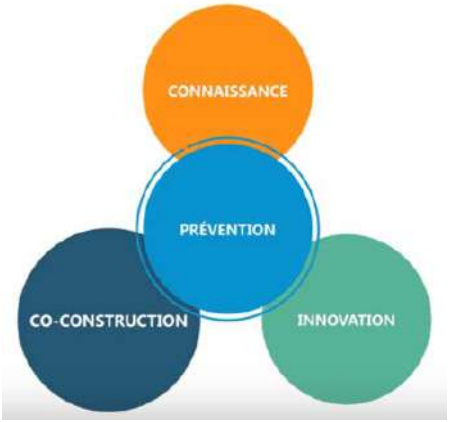


FIGURE 3.1 – Activités de la Mission Risques Naturels (Source : MRN)

Le premier outil historique pour la Mission Risques Naturels est le SIG MRN. C'est un outil d'analyse de l'exposition à l'adresse des biens assurés aux aléas naturels et climatiques, développé par (CHEMITTE 2008) et enrichi depuis. L'indice d'exposition est construit avec les assureurs et permet d'identifier les zones vulnérables pour chaque aléa. Comme son nom l'indique il repose sur les Systèmes d'informations géographiques (SIG), incontournables dans l'étude des risques naturels.

Un deuxième outil phare est la base de données des sinistres liés aux catastrophes climatiques et naturelles en France (SILECC), introduite par (BOURGUIGNON 2014). Cette base est centrale dans nos travaux et sera donc détaillée par la suite de ce chapitre. Elle permet de mieux connaître le coût des événements naturels ce qui est un enjeu majeur pour améliorer la gestion et la prévention des risques. Cette base vise à être la plus exhaustive possible et ne se concentre pas que sur les événements de grande ampleur mais aussi sur les événements de fréquence. Cette base est profondément liée à une autre base de données constituée à la MRN, la base de données des événements CatNat et climatique, aussi détaillée dans la suite. Elle repose sur la constitution d'événements naturels en fonction de leurs périmètres spatio-temporels, comme introduit par (BOURGUIGNON 2014) et ensuite élargi aux autres aléas par la MRN. La BD SILECC permet l'agrégation de la sinistralité par événement ce qui facilite son étude. Elle est aussi très utile pour étudier les territoires impactés. Les retours d'expérience permis par l'exploitation conjointe de ces bases de données sont très précieux pour la profession mais aussi pour l'intérêt général. Comme nous le verrons par la suite ils ont permis entre autres d'améliorer les cartographies d'expositions aux risques naturels.

Grâce à ces bases de données et outils la MRN suit les événements naturels et en propose des bilans à ses adhérents. Un autre aspect central dans cette thèse est le suivi en temps réel des événements naturels pour le compte de la profession et plus particulièrement de la fédération. Cet aspect s'inscrit pleinement dans l'activité d'amélioration de la connaissance des événements naturels de la MRN et mobilise toute son expertise métiers. La MRN a aussi participé à la mise à jour de l'étude climat de France assureurs, précédemment mentionnée en introduction, (*Etude : Changement climatique et assurance à l'horizon 2040* 2021).

Co-construction La MRN travaille étroitement avec les autres acteurs de la gestion des risques naturels en France pour co-construire des actions de préventions. Une bonne illustration est l'ac-

tualisation de la carte de susceptibilité au retrait gonflement des argiles faites conjointement avec le Bureau de recherches géologiques et minières (BRGM) en 2020, (*Lettre d'information de la Mission Risques Naturels 30* 2019). Cette nouvelle carte prends en compte la sinistralité passée et permet de mieux rapporter l'exposition à cet aléa. Cette cartographie est déterminante pour la reconnaissance CatNat comme décrit précédemment. Il est donc d'autant plus important que cette carte soit la plus précise possible. De plus dans la loi pour l'évolution du logement, de l'aménagement et du numérique (ELAN) de 2018, l'article 68 met en place un dispositif permettant de favoriser la prévention pour les maisons individuelles construites dans les zones dont l'exposition au phénomène retrait-gonflement des argiles (RGA) est identifiée comme moyenne ou forte. Ce qui donne à cette carte une importance réglementaire.

Un autre projet est de co-construire avec les experts d'assurance un observatoire de l'endommagement du bâti. Ce projet est au coeur de cette thèse, notamment l'analyse des données alimentant cet observatoire. Il sera donc détaillé en chapitre 4.

Pour mieux identifier et valoriser les référentiels techniques de conception du bâti la MRN anime un groupe de travail qui publie un répertoire de référentiels de résilience du bâti, (*Référentiels de résilience du bâti aux aléas naturels* 2022). Ce répertoire a pour but d'améliorer à terme la résilience des bâtiments aux aléas naturels, enjeu majeur pour la maîtrise de leurs coûts.

Innovation L'innovation est aussi au coeur des préoccupations de la Mission Risques Naturels avec en particulier un lien étroit avec la recherche. Six thèses CIFRE ont été réalisées ou sont en cours de réalisation en comptant celle-ci. Deux ont permis d'avoir des outils innovant pour la MRN et la profession comme précédemment mentionnée, (Chemitte 2008 ; Bourguignon 2014), et deux ont permis d'améliorer l'évaluation des mesures de prévention en France (Gérin 2011 ; Guillier 2017). La présente thèse contribue en utilisant des méthodes d'apprentissages statistiques pour l'évaluation du coût et des conséquences des événements naturels. Une autre thèse en cours à la MRN développe un Diagnostic de Performance de Résilience du bâti (DPR) afin d'obtenir une cotation de performance à la résilience.

La MRN produit aussi des vidéos publiques pour promouvoir des actions innovantes de prévention portées par les sociétés d'assurance et par les acteurs des risques naturels.

Prévention Avec l'augmentation à prévoir du coût des catastrophes naturelles la prévention est un levier d'action majeur. C'est la mission principale de la MRN et toutes les activités précédemment mentionnées oeuvrent directement ou indirectement à une meilleure prévention. Elle participe aussi en informant et sensibilisant largement, c'est pour cela que la MRN publie des études publiques se basant sur ses nombreux outils. Les travaux de cette thèse contribuent directement à ces communications. On peut citer par exemple (*Lettre d'information de la Mission Risques Naturels 34* 2020 ; *Lettre d'information de la Mission Risques Naturels 36* 2021 ; *Sécheresse Géotechnique, de la connaissance de l'aléa à l'analyse de l'endommagement du bâti* 2018).

La contribution à la prévention de la MRN passe aussi par l'évaluation de l'efficacité des mesures nationales. L'évaluation de la pertinence de la couverture des Plan de Prévention des Risques est faite dans (Gérin 2011). Une approche expérimentale de l'efficacité des Programmes d'Action de Prévention des Inondations (PAPI) est développée dans (Guillier 2017). Les deux mesures phare de préventions des inondations au niveau national ont été abordées dans des travaux de la MRN.

C'est donc au sein de cette association de huit collaborateurs que cette thèse CIFRE s'est effectuée. La nature de cette structure présente une particularité intéressante : elle a la réactivité et la polyvalence d'une petite équipe, tout en reposant sur un GIE de moyens importants.

Elle peut aussi s'appuyer sur le soutien de la fédération et des sociétés d'assurances. C'est une structure active qui possède un large spectre de compétences. Son réseau et ses moyens lui permettent de mener une activité unique en son genre et pionnière dans son activité en France. Cet environnement privilégié a permis de réaliser des travaux riches et variés comme illustrés dans cette thèse.

3.1.2 France Assureurs

La MRN est un groupement technique dépendant de la Fédération Française de l'Assurance, qui depuis 2022 a pris pour nom d'usage France Assureurs. La Fédération a été créée en juillet 2016 avec la fusion de la Fédération française des sociétés d'assurance (FFSA) et du Groupement des entreprises mutuelles d'assurance (GEMA). Elle rassemble ainsi l'ensemble des entreprises d'assurance et de réassurance opérant en France, soit 247 sociétés représentant plus de 99 % du marché global de l'assurance.

Les principales missions de la Fédération sont les suivantes :

— représenter l'assurance auprès des pouvoirs publics nationaux et internationaux, des institutions et des autorités administratives ;
— fournir les données statistiques essentielles de la profession ;
— informer le public et les médias ;
— promouvoir les actions de prévention.

Une partie de cette thèse est d'assister la fédération dans ses missions et en particulier l'estimation en temps réel de l'ampleur et le coût potentiel des événements naturels. Elle utilise ces informations pour pouvoir répondre aux sollicitations des pouvoirs publics. Elle est en effet à l'interface entre ces derniers et la profession et peut être amenée en temps de crise à intervenir rapidement. Ces informations permettent aussi de suivre les événements et de dimensionner les réponses nécessaires. Ce lien étroit avec France Assureurs permet à ces travaux d'être bénéfique à tout le marché de l'assurance.

3.1.3 Sociétés d'assurances

Nous travaillons cependant aussi en lien avec les sociétés d'assurances. La gouvernance de l'association est assurée par le conseil d'administration qui est constitué de représentants des grands groupes d'assurance en France, comme illustré 3.2. Ils participent donc à l'orientation des projets et assurent leurs bons développements. Les bases de données sont aussi construites grâce à leurs contributions. L'ensemble des sociétés adhérentes de France Assureurs, qui opèrent sur le marché français dans la branche « dommages aux biens des particuliers et des professionnels », sont associées au financement de son programme d'activités.

3.1.4 Réseaux d'expertise

Un autre acteur clé dans la gestion des risques naturels et d'importance ici est l'expert d'assurance.

Lors d'un sinistre, l'assureur missionne un expert d'assurance chargé d'évaluer les dommages. Suite à cet ordre de mission, l'expert se rend chez l'assuré et produit un rapport d'expertise, rapport comprenant l'évaluation des dommages et les chiffrages pour les réparations. Ce rapport est ensuite transmis à la société d'assurance. Sur la base de ce rapport, l'assureur indemnise l'assuré. Ces rapports contiennent des informations très utiles pour comprendre un sinistre et pour analyser les dommages. Pour son chiffrage, l'expert se base sur les prix du marché de la construction et des biens. Il peut s'appuyer sur des devis d'artisans pour la réparation de biens

FIGURE 3.2 – Gouvernance de la Mission Risques Naturels (Source : MRN)

spécifiques. Pour que le client puisse faire face aux travaux les plus urgents, l'expert définit un montant de versement immédiat correspondant à la valeur des biens assurés, décoté en fonction de leur vétusté. Dans certains cas, le remboursement de la différence avec la valeur à neuf des biens peut être demandé en fonction du contrat souscrit et sur présentation des factures d'achat.

Les experts peuvent être regroupés dans des réseaux ou sociétés d'expertise et dans le cadre de cette thèse nous avons travaillé avec deux sociétés en particulier, Saretec et Elex. Les sociétés d'expertise sont regroupées dans la Fédération des sociétés d'expertise (FSE) et les experts dans la compagnie des experts (CEA). Ce sont des acteurs clés car ils sont impliqués dans la gestion directement auprès de l'assuré et ont donc une connaissance très précieuse sur le déroulement d'un sinistre. Ils produisent et traitent ces données en récoltants les informations à la source. Dans les informations qu'ils récoltent on peut trouver l'état initial du bien avant sinistres, pouvant par exemple être utilisé pour identifier des facteurs potentiellement aggravants. On retrouve aussi les informations relatives aux dommages qui sont exploitées pour alimenter l'observatoire de l'endommagement du bâti.

C'est une source clé pour analyser les dommages et cette démarche a été initiée par (ANDRÉ 2013) qui pour la submersion marine a analysé les données d'expertise pour mieux comprendre les différents type de dommages. Cependant nous nous sommes heurté au même problème de manque de données que dans son étude. En effet nous n'avons pas réussi à créer une base de données comprenant les informations pour un grand nombre de réseaux d'expertise, permettant de s'affranchir de caractéristiques locales spécifiques. La généralisation de cette démarche reste complexe dans un contexte concurrentiel et commercial, avec de nombreux acteurs, aux intérêts pas toujours convergents.

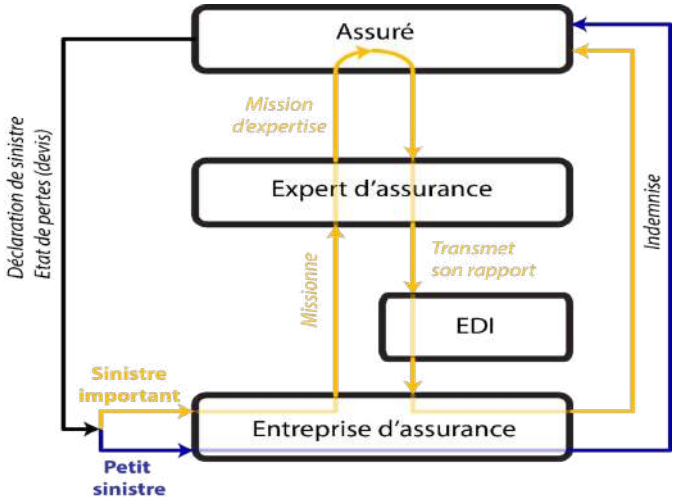


FIGURE 3.3 – Schéma du processus de règlement des sinistres (Source : MRN)

3.2 Bases de données

La place centrale de la Mission Risques Naturels lui a permis de développer de nombreuses bases de données utiles au marché de l'assurance. Ces bases de données sont au coeur des travaux de cette thèse et nous les détaillons dans les sections suivantes.

3.2.1 Base de données événements

Pour étudier les risques naturels la première étape est de les regrouper par événement. Cette démarche se base sur le première définition faite par (BOURGUIGNON 2014) mais a depuis été enrichie et élargie. Les données de sinistralité sont reçues à l'échelle communale et pour une date données mais il est intéressant de créer des événements pour en faciliter l'analyse. Plus particulièrement dans notre cas le regroupement par événement permet d'avoir une base d'apprentissage pour ensuite estimer le coût lorsqu'un événement se produit. On peut aussi grâce à ce regroupement fournir des indicateurs intéressants sur les territoires impactés.

Aujourd'hui il y a des bases pour les inondations, pour la grêle et pour la tempête. Nous ne décrirons ici que la base pour les inondations que nous utilisons dans le chapitre 6.

Dans notre cas un événement est défini par une date de début, une date de fin et un ensemble de communes impactées. Les territoires impactés sont recensés a posteriori selon les demandes de reconnaissances CatNat et peuvent aussi être complété par la sinistralité. Les demandes CatNat sont regroupées afin de définir les événements selon un périmètre spatio-temporel cohérent. Pour cela un arbre de décision est mis en place, présenté en 3.4. Pour créer un événement à partir d'une liste de communes, on regarde si les communes ont des dates cohérentes et ensuite si elles sont situées dans un même secteur hydrographique ou dans des secteurs hydrographiques adjacents. Si ces conditions sont remplis alors on forme un même événement, sinon on forme plusieurs événements.

Cet arbre se base sur des critères dépendant de deux variables, les secteurs hydrographiques touchés, et les dates. Le secteur hydrographique provient de la BD Carthage de l'IGN. Ce découpage est défini par la (*Circulaire numéro 91-50 du 12 février 1991* 1991). Les limites s'appuient

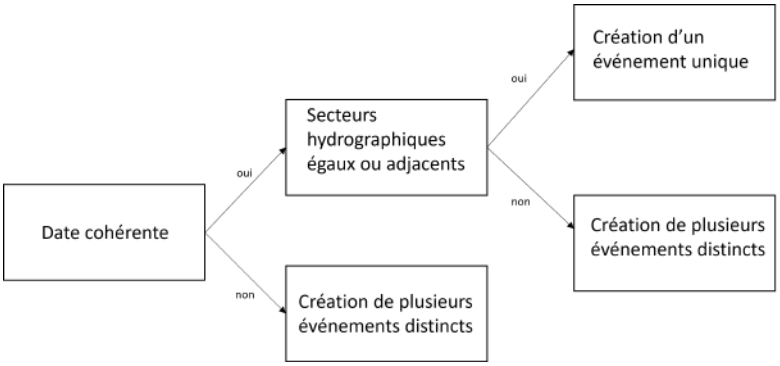


FIGURE 3.4 – Construction de la Base de données événements

sur celles des bassins-versants topographiques, ces secteurs permettent de séparer et d'identifier les différents territoires hydrographiques. Les dates sont spécifiées par l'arrêté CatNat et correspondent aux dates de début d'événement. Pour rapprocher deux dates, il faut qu'elle soit assez proches et le seuil dépend du régime de crue. En effet lors d'une crue lente un écart de trois jours peut être lié à un même événement mais pas forcement dans un crue rapide. Une table définissant les limites pour chaque régime de crue est donc déterminée en amont.

Un événement correspond ainsi à un périmètre de dommages indemnisés au titre du régime CatNat, dans une période de temps restreinte, à la maille communale. L'ensemble des arrêtés CatNat inondation depuis 1982 sont exploités et regroupés pour construire cette base de données.

La base de données événement est constituée de près de 140 000 arrêtés CatNat inondation regroupés en plus de 4 300 événements distincts entre 1982 et 2021. Cette base est très déséquilibrée, comme beaucoup de jeu de données en assurance, il y a des événements majeurs qui concentrent une grande partie des communes. Les 10 événements de plus grande ampleur représentent 35% de la base. Il existe aussi des événements extrêmes en terme de nombre de communes comme l'atteste la figure 3.5.

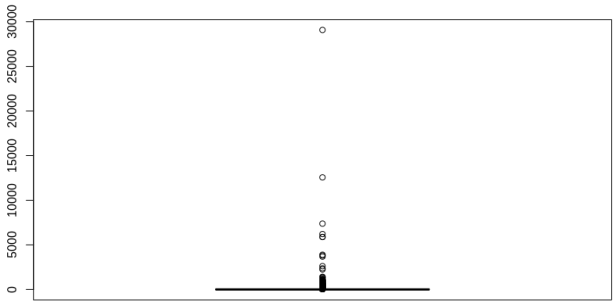


FIGURE 3.5 – Nombre de communes par événement inondation

Cette base couvre tout le territoire, 99% des communes sont touchés par au moins un événe-

ment inondation, on recense 225 000 lignes dans cette base, une ligne étant un couple commune événement. On prend en compte tous les périls relatifs aux inondations, en table 3.1 on peut retrouver les différents libellés des arrêtés CatNat pris en compte. On retrouve aussi une part non-négligeable de commune hors CatNat qui sont rapportés par la sinistralité.

Péril	Part de communes rapportées
Chocs Mécaniques lies a l'action des Vagues	3%
Coulée de Boue	0%
Inondations et/ou Coulées de Boue	61 %
Inondations Remontée Nappe	1%
Lave Torrentielle	0%
Raz de Marée	0%
Hors CatNat	35 %

Tableau 3.1 – Nombre de communes par péril CatNat

3.2.2 Base de données des sinistres

Pour étudier la sinistralité la MRN récolte les sinistres de tous les périls CatNat et climatique auprès de 12 grandes compagnies d'assurance françaises. Cela représente 70% du marché dommages aux biens en France. Ces sinistres sont ensuite harmonisés et agrégés pour former une base cohérente. Ils sont aussi localisés avec différents niveaux de précision pour pouvoir les croiser avec les événements et avec les cartes d'exposition. Certains sinistres sont localisés à la commune et d'autres au bâti. Le même travail est réalisé pour les portefeuilles ce qui permet de connaître la distribution des primes, ce qui sera utile dans le suite de notre étude.

Le coût des sinistres est actualisé, selon l'indice de la Fédération Française du Bâtiment (FFB), calculé par rapport au coût de la construction d'un immeuble en France. Cela permet d'être sur le même référentiel de coût lorsque l'on parle de sinistres éloignés dans le temps en prenant en compte l'augmentation des coûts liés à la construction.

Après la phase préliminaire de mise en forme nous obtenons une base renseignant pour chaque sinistre, la date de survenance, le péril, le segment de risque, la localisation et le coût.

Du fait de sa profondeur temporelle il peut y avoir des incertitudes sur le coût réel des sinistres. Ce problème est bien fréquent dans la littérature actuarielle du provisionnement (« micro-level reserving »), voir par exemple (Norberg 1993 ; Norberg 1999 ; Antonio et R. Plat 2014 ; Pigeon, Antonio et Denuit 2014) : pour certains types de dommages, plusieurs années peuvent s'écouler entre l'instant de la survenance et celui de la clôture, où le montant du sinistre devient connu, ce qui relie l'étude de ces problématiques à certains développements d'analyse de survie. Il peut y avoir des incertitudes aussi liées à l'adresse, elle peut ne pas correspondre à l'adresse exacte du sinistre ou peut aussi être mal geolocalisée. Ces incertitudes sont inhérentes à une étude de la sinistralité à une échelle si fine. Étant dans une logique de retour d'expérience nous n'appliquons aucun correctif et analysons la base en l'état. Nous essayons de réduire au maximum les sources d'incertitude en amont, grâce à un nettoyage et un traitement minutieux des données, mais certaines font partie de la base et se répercutent sur nos études.

On obtient une base de données volumineuse, qui atteint un bon niveau de représentativité global de la sinistralité en France.

3.2.3 La base de données SILECC

Ces deux bases de données sont ensuite jointes pour former la base de données des sinistres liés aux catastrophes climatiques et naturelles en France (SILECC).

FIGURE 3.6 – Construction de la BD SILECC (Source : MRN)

Cette base permet de suivre la sinistralité en fonction des événements. On rajoute à la liste des informations de la base de donnée des sinistres les informations relatives à l'événement. On peut par exemple représenter pour les inondations la répartition des événements les plus coûteux, comme en 3.7. C'est une base très précieuse pour suivre la répartition du coût des événements naturels.

On peut aussi observer des différences dans les coûts en fonction du type d'événement et du segment de risque, on observe des disparités en fonction de la reconnaissance CatNat et du segment.

Sinistralité	Coût moyen particuliers	Coût moyen professionnels
Reconnue CatNat	8 000 €	28 000
Non reconnue CatNat	3 000 €	19 000
Hors CatNat	5 000 €	14 500

TABLEAU 3.2 – Répartition du coût moyen des sinistres en fonction du segment de risque et de la reconnaissance CatNat

Ce type d'étude et cette base sont très utiles pour la profession, notamment pour le positionnement des assureurs sur des sujets de place, comme la réforme du régime CatNat ou bien l'étude sur l'impact du changement climatique précédemment mentionnée. La BD SILECC a été utilisée pour rendre compte de la sinistralité passé dans cette étude.

Elle peut aussi être utile à l'intérêt général avec par exemple la mise à jour de la carte retrait gonflement des argiles ou bien l'étude du ruissellement comme évoqué plus bas.

3.2.4 Données d'expertises

Pour compléter les informations de la BD SILECC à une échelle plus fine nous avons exploité les données provenant des réseaux d'expertise. Comme évoqué précédemment, en étant au plus proche des sinistres ils récupèrent des données très précises sur l'endommagement. Nous avons

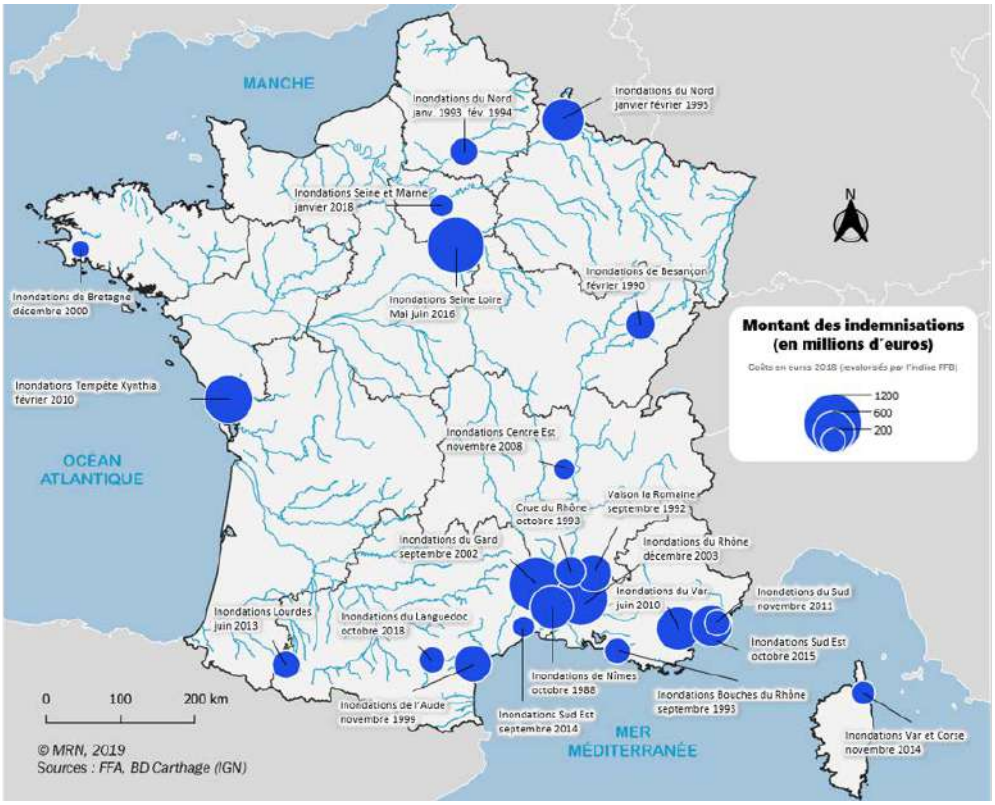


FIGURE 3.7 – Montants indemnisés pour les inondations CatNat les plus coûteuses selon la BD SILECC (Source : MRN)

récupéré pour deux réseaux d'expertise des données sur la grêle et la tempête. Ces données sont structurées et proviennent de leur système d'information, elles traitent seulement des dommages mais à une échelle de décomposition fine. Les données se présentent sous la forme d'un fichier texte présentant une ligne par partie du bâtiment impactée. Le fichier comporte donc, plusieurs lignes par dossier de sinistre, en fonction du niveau de détail sur les dommages dans la base de données. Un corps d'état correspond à un niveau de dommage pour le bâti utilisé par les experts, ces corps d'état sont ensuite rapprochés aux composantes prédéfinies, cette démarche sera développée en chapitre 4. Cependant les réseaux d'expertise n'avaient pas les données sous cette forme pour la sécheresse, nous avons donc dû nous tourner vers les rapports d'expertise en entier. L'expert remplit et transmet à la fin de sa mission un rapport reprenant toutes les informations nécessaires à la gestion, dont le chiffrage. Ce rapport est un document majoritairement au format PDF. L'analyse est cependant plus complexe car les données doivent être extraites et ensuite classées. Pour la sécheresse il y a une relative harmonisation entre les rapports sécheresse utilisés par les réseaux d'expertises. En analysant les rapports d'expertise nous pouvons aussi récupérer des informations sur l'état initial de la maison et sur son environnement. Ce sont des informations nécessaires pour comprendre la complexe sinistralité liée à la sécheresse.

Avec toutes ses informations nous cherchons à créer une base de données des SInistres Liés aux événements EXpertisés catnat et climatiques (SILEX). Cette base s'inspire de la BD SILECC

et permet d'étudier aussi la sinistalité mais avec un niveau de précision supérieur. Cette base est cependant aujourd'hui restreinte dans sa représentativité. En ayant accès qu'à deux réseaux d'expertise l'analyse par événement est limitée car nous avons une vision partielle à l'échelle locale.

3.3 Carte d'expositions

Un aspect important pour étudier les risques naturels est l'exposition. La Mission Risques Naturels, grâce à son expertise, et à ces données a participé à l'élaboration et à la mise à jour de deux cartes décrites dans cette section. Ces cartes vont ensuite être utilisées pour les bases d'apprentissages des modèles d'estimations des coûts des chapitres suivants.

3.3.1 Carte exposition ruissellement

Plusieurs cartes existent pour mesurer l'exposition en France, cependant ces cartes prennent surtout en compte les inondations pas débordement et non pas par ruissellement. Avec la BD SILECC la MRN a observé qu'une grande partie de la sinistralité se trouvait en dehors de toute cartographie. En effet 56 % du nombre de sinistres inondations se trouvent en dehors des Enveloppes Approchées des Inondations Potentielles (EAIP), cartographie des inondations développée par le Ministère de la Transition écologique.

A partir de ce constat la MRN a entrepris le développement d'une cartographie d'exposition prenant aussi en compte l'accumulation des eaux de ruissellement. Pour cela elle se base sur la BD ALTI®de l'IGN. Dans cette base se trouve un modèle numérique de terrain (MNT) maillé qui décrit le relief du territoire français à moyenne échelle. En utilisant le MNT de 25 mètre préalablement corrigé, la MRN calcule ensuite le Compound Topographic Index (Moore, Grayson et Ladson 1991). Cet indice combine l'information sur les zones de concentration et sur les pentes pour déterminer les zones de forte accumulation des eaux. Ensuite cet indice est discrétisé selon des critères dépendant des hydro-ecorégions, cette étape permet de prendre en compte les différences entre les territoires. Ce nouvel indice est ensuite agrégé par cercle de 75m et classé selon 5 niveaux de risques. Après intégration des zones EAIP pour prendre en compte le débordement, nous obtenons une carte 3.8 rendant compte de 5 niveaux d'exposition aux inondations. Cette carte permet de mieux prendre en compte la sinistralité dans sa globalité, en effet dans cette carte tout le territoire est couvert par une zone d'exposition. On retrouve une majorité de sinistres dans les zones moyennes, fortes et très fortes. Ces trois zones concentrent 80 % du nombre de sinistres. Cela conforte la validité et l'utilité de cette cartographie qui permet donc de mieux rendre compte de la sinistralité rapportée.

Zone d'exposition	en nombre	en charge
Très forte	38 %	47 %
Forte	27%	27%
Moyenne	15 %	13%
Faible	5 %	4 %
Très faible	15%	9%

Tableau 3.3 – Répartition des sinistres inondations par zone d'exposition

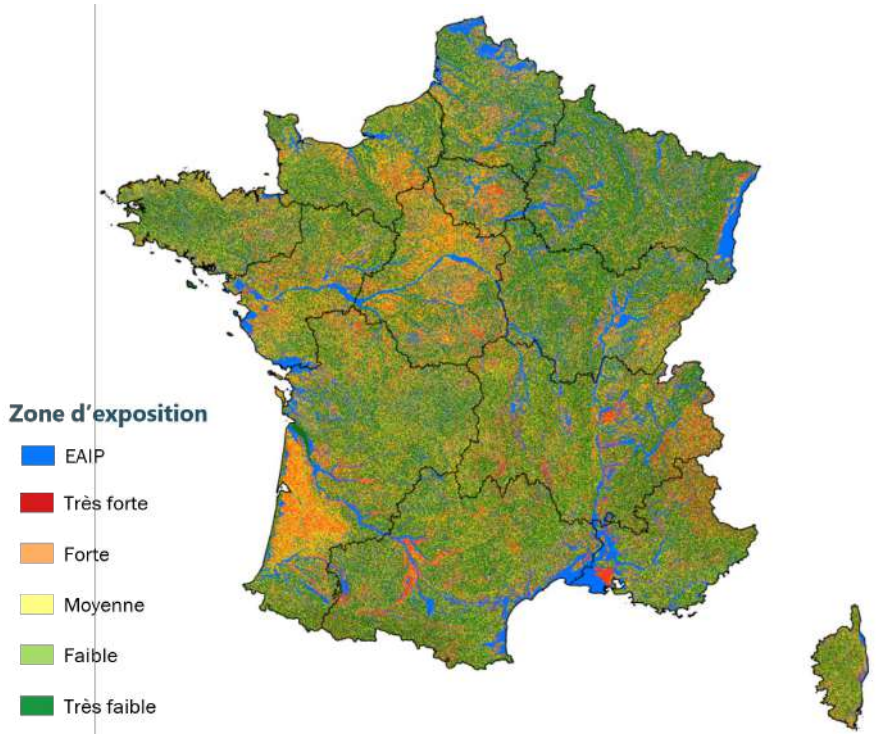


FIGURE 3.8 – Cartographie MRN d'exposition aux inondations (Source : MRN)

3.3.2 Carte exposition RGA

Pour caractériser la susceptibilité au retrait-gonflement des argiles dans le sol, nous nous appuyons sur un indicateur utilisé dans la cartographie publiée par le BRGM. C'est une carte de susceptibilité au phénomène RGA, qui a été établie lors du programme de cartographie départementale de cet aléa, entre 2001 et 2010, pour toute la Métropole, (VINCENT, E. PLAT et LE ROY 2007). L'indice est formé par agrégation complexe de trois critères propres à la qualité du sous-sol : caractéristiques liées à la lithologie (mesure de la formation et de la proportion d'argile), à la composition minéralogique de la zone argileuse (indication la présence et la proportion des minéraux argileux les plus sensibles), et à la géotechnique (comportement mécanique de la couche argileuse). L'historique des sinistres est également utilisé pour refléter de la fréquence des dégâts dans la région. La mise à jour de cette carte a été réalisée avec l'aide de la MRN en 2019, comme décrit dans (*Lettre d'information de la Mission Risques Naturels 30* 2019). C'était un travail très attendu, il a permis après mise à jour, de passer d'un taux de couverture de la sinistralité par une exposition moyenne ou forte au RGA de 59 % à 93 %. En effet, comme pour les inondations, cette carte ne captait pas toute la sinistralité, ce que cette mise à jour a permis d'améliorer grandement. C'est une bonne illustration du caractère d'intérêt général de cette base de données.

Une fois calculé, cet indice peut être compris comme un facteur de risque, et un classement des différentes zones est effectué, définissant trois classes hiérarchiques, fournissant une cartographie nationale qui décrit la propension au gonflement de l'argile. Cette cartographie est présentée dans la figure 3.9.

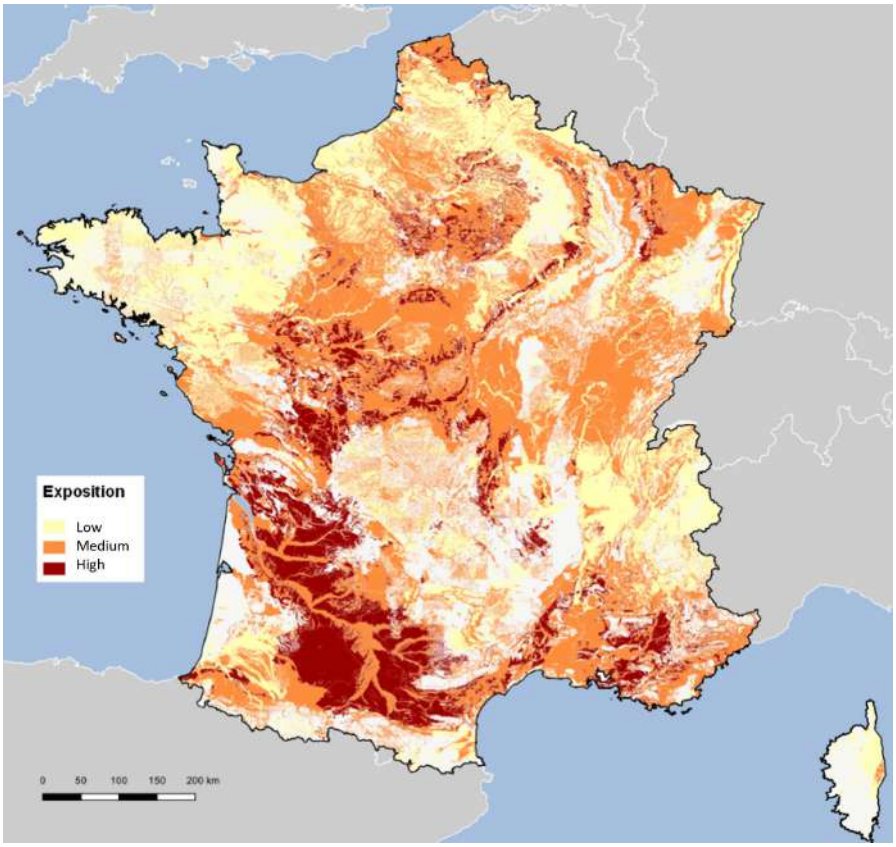


FIGURE 3.9 – Carte de sensibilité au retrait gonflement des argiles (Source : Géorisque)

Chapitre 4

Analyse de la sinistralité à l'échelle fine du bâti

Dans ce chapitre nous proposons une application des méthodes d'analyses textuelles aux données d'expertise d'assurance. Nous présentons deux applications reposant sur les réseaux de neurones avec word embedding.

4.1 Introduction

Dans le cadre de ses activités d'études sur la connaissance et la prévention des risques naturels, la MRN conduit depuis 2016 des retours d'expérience sur la sinistralité du bâti à partir de l'exploitation des données technico-économiques provenant des rapports d'expertise d'assurance. L'objectif est de mieux connaître les dommages occasionnés par des événements naturels et d'améliorer la connaissance de la nature et du coût de l'endommagement à l'échelle fine du bâti, ce qui permet d'identifier :

— des pistes d'amélioration de la résilience du bâti (Build Back Better) ;
— des leviers d'actions professionnelles possibles visant à réduire le coût économique d'un événement climatique.

Le but de ces retours d'expérience est ainsi de créer une base de données de sinistres expertisés, qui alimenterait un observatoire de l'endommagement du bâti suite à des événements naturels. L'objectif est d'être en capacité de produire des chiffres précis et des statistiques sur la sinistralité suite à des événements naturels à l'échelle du bâtiment. Cet observatoire doit permettre d'améliorer la prévention des risques naturels en connaissant mieux leurs conséquences, afin de réduire le coût économique d'un événement naturel. Les effets des événements naturels sur le bâtiment sont par exemple étudiés dans, (Salagnac et al. 2014 ; Salagnac 2015), mais on manque de données chiffrées et systématiques permettant d'évaluer les mesures de protections et préventions possibles. Cette base pourrait aussi permettre d'améliorer les modèles d'estimation du coût des événements, on va le voir dans le chapitre suivant, pour la sécheresse on manque d'informations fines sur le bâti.

Comme évoqué précédemment, l'idée d'utiliser les rapports d'expertise pour mieux comprendre la sinistralité à l'échelle du bâti a déjà été expérimentée par (André 2013). Cette analyse était cependant menée seulement pour la submersion marine et pour les événements suite aux tempêtes Johanna (2008) et Xynthia (2010). Notre démarche est similaire mais pas restreinte à la submersion marine. Nous nous intéressons à tous les aléas et tous les événements. Nous voulons

créer une base de données pour faire un reporting systématique donc nous avons attaché une importance particulière à faire une méthode qui permettrait d'analyser plus de données si nécessaire. En effet, il est toujours possible, en y mettant le temps nécessaire, d'analyser les rapports mais ici notre démarche est de créer une méthode simple et automatisable permettant d'analyser un grand nombre de données. Le but est donc de créer une méthode qui peut potentiellement être utilisée pour tout le marché. Cela contraint les méthodes d'analyses disponibles. Nous avons développé et testé plusieurs méthodes et profité des progrès récents en analyse de texte tout en gardant à l'esprit cet objectif.

A l'étranger on peut trouver des démarches similaires avec par exemple la base HOWAS 21 en Allemagne, (Kellermann et al. 2020). Elle est constituée de données fines sur les dommages liés aux inondations. Cette base de données est gérée par l'Université de Postdam et alimentée par des sources diverses. Une partie provient de campagnes téléphoniques mais une partie provient aussi de données collectées par les experts d'assurances lors d'inondations de grandes ampleurs. On peut aussi mentionner les Établissements Cantonaux d'Assurance (ECA) suisses qui analysent les données d'expertises et en font une analyse statistique (*La Fondation de prévention des établissements cantonaux d'assurance en Suisse, Cahiers spéciaux de la MRN* 2019). En se basant sur un grand nombre de sinistres et de données ils peuvent faire une analyse très fine des dommages, des liens avec l'aléa et l'état initial du bâtiment comme on peut le voir dans (Nicolet, Voumard et al. 2014 ; Nicolet, Choffet et al. 2015). On peut toutefois noter le contexte particulier des ECA qui sont en quasi-monopole pour le marché de l'assurance habitations. En France, nous sommes dans un contexte concurrentiel avec un grand nombre d'acteurs ce qui rend la coordination plus complexe.

4.2 Contexte et données disponibles

4.2.1 Données et acteurs

Le coeur de notre étude est d'analyser les données provenant des rapports d'expertise. Les rapports d'expertise contiennent de nombreuses informations permettant d'améliorer la connaissance de l'endommagement. Leur analyse permet de descendre au niveau des composantes du bâti. Ils comportent une description plus ou moins détaillée, selon l'expert et le déroulement de l'expertise, de l'étendue des dommages. Ces dommages sont habituellement classés en trois catégories :

— Les dommages au bâti ;
— Les dommages aux embellissements ;
— Les dommages au mobilier.

L'essentiel de cette donnée source reste pourtant inexploité à l'heure actuelle, seule la sinistralité à l'échelle de la France ou de la commune est connue et exploitée avec par exemple la BD SILECC. La première étape était de faire un état des lieux des sources disponibles pour les données d'expertise. Nous avons alors exploré deux pistes, les réseaux d'expertises et les sociétés d'échanges de données. Il y a pour le marché de l'assurance française de nombreux échanges entre de nombreux acteurs et pour faciliter ces échanges, des solutions d'Échanges de Données Informatisées (EDI) ont été développées. On peut citer DARVA qui est le prestataire d'EDI majoritairement utilisé dans le secteur de l'assurance. Toutes les informations transitent donc par DARVA qui a pour but de stocker ces données.

On distingue pour notre étude deux types de données :

— les données structurées qui contiennent des informations déjà mises en base de données ;
— les rapports aux formats PDF qui sont des données non structurées.

Dans un premier temps nous avons utilisé les données provenant de DARVA car elles étaient assez structurées. Cependant pour notre étude nous cherchons à décomposer les dommages selon des composantes prédéfinis et avec les données de DARVA nous avions beaucoup de perte durant cette étape. Nous nous sommes rapproché des réseaux d'expertise pour récupérer les données structurées directement à la source. Nous avons aussi expérimenté sur des données non structurées pour voir ce que l'on était en mesure de faire pour récupérer des informations dans les rapports. En travaillant directement avec les réseaux nous avons amélioré la qualité des données et la disponibilité. Cependant nous perdons en représentativité générale du marché car dans le cadre de notre étude nous avons réussi à avoir les données pour seulement deux réseaux d'expertises.

4.2.2 Données structurées

Des données sur la tempête et la grêle ont été récupérées pour deux réseaux d'expertise. Ces données étaient structurées et provenaient de leurs systèmes d'information. En effet les réseaux d'expertise ont en base de données une partie des informations présentes dans les rapports d'expertise, en particulier la partie traitant des dommages.

Les données se présentent sous la forme d'un fichier texte avec une ligne par poste de dommage. Le fichier comporte, plusieurs lignes par dossier de sinistre, en fonction du niveau de détail sur les dommages dans la base de données. Un poste de dommage correspond au niveau de dommage pour le bâti utilisé par les experts, ces postes sont ensuite rapprochés à des composantes que nous avons définies. Pour cela on utilise un champ commentaire qui décrit les dommages. En colonne, on trouve les différents champs d'informations nécessaires à l'analyse, tels que le coût pour ce poste de dommage ou la localisation.

L'objectif principal de l'étude est de déterminer le coût des dommages associés à chaque composante du bâti. Cette répartition en composantes du bâti est la problématique principale pour ces données. L'enjeu majeur est de produire un fichier contenant pour chaque ligne, la composante principale, et secondaire si la précision est suffisante, impactée et le coût associé. On peut, ci-dessous, voir des exemples des données reçues et les champs associés pour chaque réseau d'expertise.

Champ 1	Champ 2
Tempête	Couverture
Aménagements extérieurs	Risque assuré
NA	VERANDA
Dommages consécutifs : dégât des eaux	Plâtrerie
plancher combles	combles plancher

Tableau 4.1 – Exemple des deux champs décrivant les dommages pour un des réseaux d'expertise.

Nous pouvons déjà remarquer que les données sont différentes en fonction de la source. En effet nous pouvons constater que ce n'est pas le même niveau de détail ni les mêmes informations. C'est une difficulté supplémentaire mais pour s'assurer que notre méthode permet de traiter l'ajout de nouvelles données sans développement supplémentaire, nous avons regroupé les données et nous les traitons ensemble. Nous formons un champ commentaire comprenant toutes les informations, comme en 4.2. En 4.3.1 nous décrivons les modèles de classification textuelle utilisés pour faire correspondre ces commentaires à des composantes prédéfinis.

Nous sommes ici face à des données textuelles avec une structure particulière, mélangeant des mots-clés et des fragments de phrase. Chaque expert a sa façon particulier de décrire les dommages et nous avons des données très hétérogènes même au sein d'une même compagnie.

Champ 1	Champ 2	Champ 3	Champ 4
Deplacement et evacuation	Réparation clôture	les dommages concernent Immobilier : Clôture 32ml arraché par les vents.	les dommages sont localisés dans la pièce suivante: -jardin
-	-	les dommages concernent Embellissements : et Immobilier : .	0
fourniture portail aluminium 3m25	portail coulissant aluminium 3m25	Les dommages concernent Immobilier : Portail coulissant en aluminium 3m25	Les dommages sont localisés dans la pièce suivante: -Entrée
Remise en état des peintures endommagées	Peinture	Les dommages concernent Embellissements : -Peinture	Les dommages sont localisés dans les pièces suivantes: - Chambre 1 -Séjour -Salle de bain
Broyeur de branche	Arbre	Les dommages concernent Immobilier : Porte de garage en bois et fenêtre + 3 chênes + branches.	les dommages sont localisés : - Façade -Jardin

FIGURE 4.1 – Exemple des quatre champs décrivant les dommages pour un des réseaux d'expertise.

Nous sommes aussi en présence d'un vocabulaire très spécifique, propre aux rapports d'expertise. Comme dans toutes les données réelles, dans ces données nous rencontrons des fautes et un gros travail de nettoyage est nécessaire.

Nous obtenons en tout 41 000 lignes de textes, pour la tempête et la grêle, couvrant une période de 2014 à 2019. Pour la sécheresse les réseaux d'expertise n'avaient pas les données sous cette forme au moment de notre étude, nous avons donc essayé de nous tourner vers les rapports d'expertise en entier.

4.2.3 Données non-structurées

L'expert remplit et transmet à la fin de sa mission un rapport reprenant toutes les informations nécessaires à la gestion dont le chiffrage. Ce rapport est sous la forme d'un document Word ou pdf. L'analyse est cependant plus complexe car les données doivent être extraites et ensuite classées. Pour la sécheresse il y a une relative harmonisation dans les rapports sécheresse présents dans les réseaux d'expertises. Ils s'inspirent tous largement d'un rapport type « sécheresse » produit en 2000 par les assureurs à destination des experts.

Ce rapport type a été réalisé dans le cadre d'un groupe de travail piloté par la FFSA et le GEMA en 1999, juste avant la création de la MRN. Il contient un plan qui décrit les informations que les experts doivent récolter sur place lors de leurs visites de reconnaissance. Ce rapport contient donc de nombreuses informations qui peuvent être très utiles dans notre étude. Grâce à ce rapport on a une structure commune qui nous permet de faire un modèle d'apprentissage.

En analysant ces rapports on peut aussi récupérer des informations sur l'état initial de la maison et sur son environnement. Ce sont des informations nécessaires pour comprendre la complexe sinistralité liée au retrait gonflement des argiles. Ces données sont utiles pour affiner

Commentaire final
Abri piscine Piscine
les dommages concernent Embellissements : et Immobilier :
Tempête Couverture
Tabliers de porte-fenêtres doubles Remplacement de tabliers de volets roulants et de brises-soleil Volet roulants alu et brises-soleil Façade Ouest et Sud-Ouest
Serre de jardin 240 x 240 CONTENU Dommages suites aux chocs des grelons sur: -Toiture en polycarbonate de la veranda servant de serre aux végétaux.Veranda d'environ 12m^2 au sol installée en 1996. -Anémometre de fermeture d'urgence du storre . -Serre exterieure de jardin 240x240.
TOITURE DEPENDANCES TOITURE
VEGETAUX : SELON FACTURE L'ARBRE ET LA CIME
Tempete sur brise vue (installation extérieure mobilière) mur de cloture
Revêtements intérieurs touchés par les infiltrations Cuisine

FIGURE 4.2 – Illustration de la concaténation des champs décrivant les dommages pour les deux réseaux d'expertise.

l'estimation du coût d'un événement en permettant de déterminer des facteurs potentiellement aggravants par exemple.

Ici nous étions dans une démarche exploratoire pour identifier quelles informations nous arrivions à récupérer facilement et avec quelle fiabilité. L'exploitation des rapports dans ce format est nécessaire pour un rattrapage du passé mais pas pour les futures études. Nous voulions donc, en investissant peu de temps et en utilisant les derniers développements en traitement de texte, tester les possibilités. Une contrainte étant aussi de vérifier que la solution identifiée pouvait s'adapter sans développement aux autres réseaux d'expertises. Nous avons donc fait des tests avec deux échantillons, un premier de 900 dossiers, comprenant aussi des dossiers sans-suites (après visite de l'expert la sécheresse n'est pas établie comme responsable du sinistre) où nous avons cherché à récupérer les informations concernant l'état initial du bâtiment. Un autre échantillon plus conséquent de 4000 dossiers a été utilisé pour les informations sur le chiffrage.

4.2.4 Catégories créées

Pour que toutes les informations puissent être analysées et comparées, nous avons au préalable crée des catégories d'intérêt, au sein d'un groupe de travail comprenant des représentants de sociétés d'assurance spécialistes en gestion de sinistres. Pour chaque champ on définit plusieurs modalités prédéterminées. Pour les dommages nous avons 15 composantes principales et 75 composantes secondaires. Ces composantes permettent d'harmoniser les dommages et de faire des analyses communes à tous les réseaux. Par souci de cohérence avec les standards appliqués dans le secteur de la construction, la typologie développée par L'Agence Qualité Construction (AQC), dans le cadre de la collecte d'informations sur les pathologies des constructions neuves (SYCODES) a servi de base. L'Agence Qualité Construction (AQC) est une association qui voit

le jour en 1982, au lendemain de la promulgation de la « loi Spinetta » relative à la responsabilité et à l'assurance dans le domaine de la construction. L'AQC regroupe les acteurs du secteur et a pour mission notamment de recenser et d'analyser les désordres décennaux du bâti, notamment sur la base des rapports d'expertise des sinistres constructions. Dans ses missions d'amélioration de la qualité de la construction, l'AQC gère une base de données de sinistralité du bâti, le SYstème de COllecte des DESordres (SYCODES) qui collecte les données issues des rapports d'experts construction et répertorie les origines techniques des désordres liés à la construction, principalement sur des bâtiments neufs. Cette base de données fait partie de l'Observatoire des pathologies. Depuis 2007, elle permet aussi de mesurer l'impact des mesures de prévention adoptées pour éviter les désordres précédemment révélés.

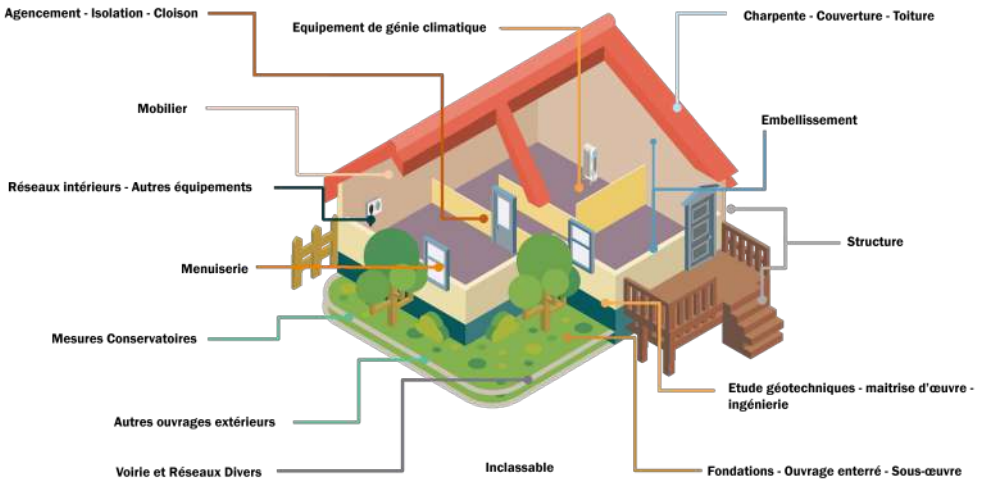


FIGURE 4.3 – Composantes principales du bâtiment retenues pour l'étude

Ces composantes ont été adaptées à l'étude des événements climatiques et cela fait partie des valeurs ajoutées de notre base de données. La liste complète se trouve en annexe B. Ces composantes permettent d'analyser de façon statistiques les données de tous les réseaux et pour tous les aléas. Une partie importante de notre travail consiste à classer au préalable les dommages selon cette typologie. Pour cela nous faisons de la classification textuelle et nous avons comparé les diverses méthodes décrites en section 2.4. En reprenant notre exemple, on cherche à faire correspondre à chaque commentaire une composante principale et une secondaire comme dans l'exemple 4.4.

4.3 Méthodes d'analyses

4.3.1 Classification de texte

Nous décrivons ici les méthodes de classification de texte pour faire correspondre à chaque ligne de commentaire les composantes mentionnées ci-dessus. Ces méthodes ont évolué au cours de nos travaux et des avancées du domaine. Dans tous les cas nous nous utilisons des méthodes d'apprentissages statistiques et la première étape est de faire un échantillon d'apprentissage. Pour cela nous nous sommes reposés sur l'expérience métier de la MRN, pour bien interpréter les

Commentaire final	Composante principale	Composante secondaire
Abri piscine Piscine	Viabilité Réseau Extérieur Jardin Piscine	Piscine
les dommages concernent Embellissements : et Immobilier :	Embellissement	Non spécifié
Tempête Couverture	Charpente Couverture	Couverture
Tabliers de porte-fenêtres doubles Remplacement de tabliers de volets roulants et de brises-soleil Volet roulants alu et brises-soleil Façade Ouest et Sud-Ouest	Menuiserie	Stores et fermetures
Serre de jardin 240 x 240 CONTENU Dommages suites aux chocs des grelons sur: -Toiture en polycarbonate de la veranda servant de serre aux végétaux.Veranda d'environ 12m^2 au sol installée en 1996. -Anémometre de fermeture d'urgence du storre . -Serre exterieure de jardin 240x240.	Autres ouvrages extérieurs	Jardin
TOITURE DEPENDANCES TOITURE	Charpente Couverture	Couverture
VEGETAUX : SELON FACTURE L'ARBRE ET LA CIME	Autres ouvrages extérieurs	Jardin
Tempete sur brise vue (installation extérieure mobilière) mur de cloture	Autres ouvrages extérieurs	Clôture
Revêtements intérieurs touchés par les infiltrations Cuisine	Embellissement	Non spécifié

FIGURE 4.4 – Illustration de la concaténation des champs décrivant les dommages pour les deux réseaux d'expertise et des composantes correspondantes.

termes techniques et le fonctionnement des processus de réparation. Pour chaque commentaire on faisait correspondre une composante du bâti. Cette étape était déjà difficile, un commentaire peut correspondre à plusieurs composantes. Il peut aussi être incomplet ou pas assez précis pour le classer. Même avec l'expertise métier cette étape présente une part d'interprétation et d'incertitude. Il est donc attendu que les modèles de classification fassent des erreurs.

Dans une première approche nous avions représenté notre corpus de texte comme un « Bag of Words ». Chaque mot correspondait à une colonne et pour chaque ligne de texte, en colonne un 0 ou un 1 indiquait si les mots sont présents ou pas. Nous avons aussi fait cela avec une pondération TF-IDF. Cette matrice peut constituer un échantillon d'apprentissage pour des algorithmes classiques de classification. Cependant pour utiliser cette méthode nous devions au préalable faire un travail très conséquent de nettoyage des données :

- La première étape était l'élimination des accents, de la ponctuation et des chiffres, le passage en minuscules et la suppression des espaces en trop.
- La deuxième étape consistait à supprimer les phrases inutiles n'apportant pas d'information, on avait identifié prés de 300 phrases types qui ont été enlevées.
- Ensuite, nous appliquons des « Replace rules », le but est de corriger les fautes d'orthographe et de normaliser les mots (féminin/masculin, pluriel/singulier). Nous avions écrit 1 500 replace rules.
- La quatrième étape était d'agréger des groupes de mots afin de créer des catégories de mots. Par exemple « étude de sol » devient « etudedesol », cela permet de donner un sens différent aux mots « étude » en fonction du contexte.
- Enfin la dernière étape de nettoyage consistait à supprimer les mots inutiles qui polluent les données et qui ne serviront pas à la classification. Pour cela, nous avions identifié près de 4 000 « stopwords », c'est-à-dire des mots que nous avons supprimés.

Toutes ces règles étaient très spécifiques aux corpus et donc à l'aléa et aux réseaux d'expertise participant, cette méthode était donc difficilement transposable et à chaque fois il fallait refaire un lourd travail de nettoyage.

Pour palier à cela nous nous sommes tournés vers les méthodes utilisant des réseaux de neurones et notamment le word embedding. Avec ces méthodes nous faisions un nettoyage minimum des données et ensuite on comptait sur l'embedding pour prendre en compte les fautes d'orthographe et les mots proches. Le principe est que :

- chaque mot devient, après un apprentissage sur l'ensemble des commentaires, un vecteur. Les coefficients de ce vecteur sont appris à partir du texte et plus spécifiquement des interactions entre les mots.
- Ici l'embedding est entraîné en même temps que le réseau de neurones, c'est la première couche du réseau de neurones. Dans ce cas l'embedding consiste simplement à un ensemble de vecteurs pour chaque mot. Les poids sont initialisés aléatoirement et ensuite ajustés par rétropropagation pour répondre le mieux à notre problème.
- Nous avons utilisé la bibliothèques Keras (Chollet et al. 2015) pour faire l'apprentissage de nos réseaux de neurones. L'embedding est une des couches fournis par Keras. La représentation est alors ajustée à notre problème de classification.

Dans le cas de l'analyse de nos données sur la grêle, il était possible d'observer que pour le mot « couverture » les mots les plus proches sont « zinguerie » ou « toiture ». Nous pouvions aussi vérifier que les mots avec des orthographes différentes sont proches. Ce qui illustre bien l'intérêt d'une telle représentation des mots et son apport pour la classification. Nous ne sommes pas vraiment dans du NLP car notre corpus de texte est très spécifique et nous n'avons pas de phrases complètes mais un mélange de mot-clés et d'expressions. Il est donc plus pertinent d'entraîner directement l'embedding sur notre corpus que d'utiliser des embeddings pré-entraînés.

Ensuite nous avons comparé pour la tempête et les deux réseaux d'expertise, les deux architectures de réseaux de neurones présentées dans la section 2.4.4. Nous cherchons ici avec une ligne de texte à prédire la composante secondaire. Nous avions 42 classes distinctes et 7 000 lignes dans l'échantillon d'apprentissage. Nous gardons 25% de cette base, choisie aléatoirement, pour faire un échantillon de test. Cet échantillon est commun pour les deux méthodes et les hyperparamètres sont choisis par cross validation.

CNN Nous avons essayé un réseau de neurones convolutionnel (CNN) avec trois couches, la couche d'Embedding puis une couche de convolution et une couche de pooling. Nous obtenons une précision, sur l'échantillon d'apprentissage, de 0.97 et de 0.75 sur l'échantillon de test. Nous pouvons voir l'évolution selon les « epochs » en figure 4.5. Les epochs correspondent au nombre de fois où le modèle s'est entraîné sur les données. Une epoch est un apprentissage sur le jeu de données complet. Il semblerait que nous soyons face à une limite et entraîner les réseaux une nouvelle fois sur les données ne semblent pas améliorer les résultats.

LTSM Nous avons fait de même pour un réseau de neurones récurent de la forme « Long short-term memory » (LTSM). Il y a ici trois couches, la couche d'embedding, une couche de SpatialDropout1D qui évite les problèmes d'overfitting et la couche du LTSM. Nous utilisons ici aussi les couches de keras. Les résultats sont similaires, légèrement inférieur sur l'échantillon d'apprentissage avec une précision 0.94 mais légèrement meilleur sur le test avec une précision de 0.77. De même nous pouvons observer qu'augmenter le nombre d'epoch n'améliore plus significativement les performances à partir de 2.

Avec ces deux méthodes on obtient des bons résultats et la phase de pré-traitement est considérablement allégée. De plus ici on travaille sur les données des deux réseaux mélangées, c'est assez encourageant car on a vu qu'elles étaient très différentes. On peut donc penser que

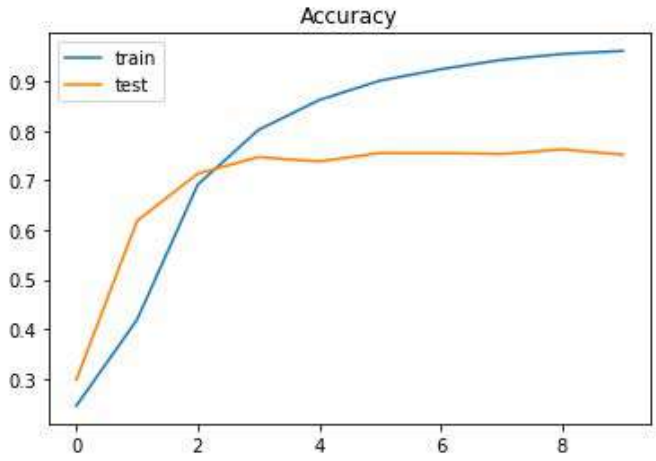


FIGURE 4.5 – Évolution de la précision pour les échantillons de train et de test, en fonction du nombre d'epochs pour le CNN

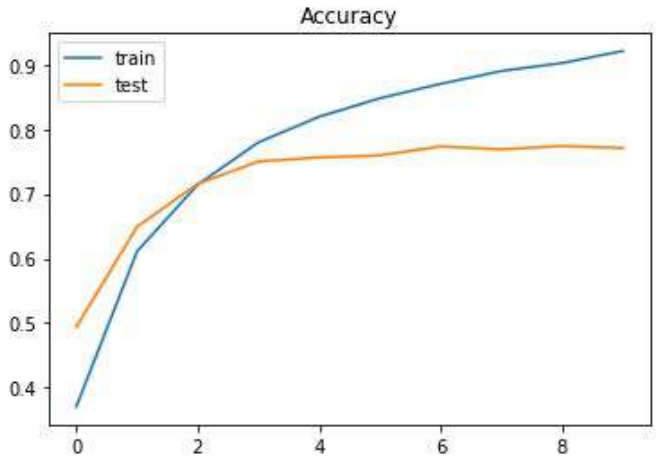


FIGURE 4.6 – Évolution de la précision pour les échantillons de train et de test, en fonction du nombre d'epochs pour le LTSM

l'ajout d'autres réseaux ne devrait pas trop nuire aux performances de nos modèles. Les réseaux de neurones et l'embeddings sont une bonne solution pour cette tache de classification textuelle. Ces deux méthodes sont aussi assez rapides contrairement aux modèles pré-entraînés qui peuvent être plus longs. Nous devons cependant pour faire cette classification, avoir une ligne de texte correspondant à un poste de dommage, ce qui n'est pas toujours le cas.

4.3.2 Reconnaissance d'entités nommées

Dans cette section nous expérimentons la reconnaissance d'entités nommées ou Named-Entity Recognition (NER) pour extraire automatiquement des entités dans un document texte. Nous sommes dans un contexte industriel précis et nous cherchions une solution simple à mettre en oeuvre et utilisable facilement pour vérifier la possibilité d'extraire des informations. Nous avons choisi le logiciel Prodigy qui permet d'annoter facilement des documents et d'utiliser la librairie open source de NLP, spaCy.

Une manière simple d'extraire des entités dans un document est d'utiliser les expressions régulières ou regex. Cela permet de délimiter par des expressions ce que l'on veut récupérer. Par exemple on peut récupérer les chiffres après « Date du sinistre », ce qui nous donnera la date au format numérique si elle existe. C'est très efficace et très utilisés mais le problème est que les règles doivent être écrites pour chaque cas de figure. C'est un ensemble de règles précises et on ne peut pas traiter des cas qui sortent de ces règles. Dans notre cas on devrait écrire des règles différentes pour chaque réseau d'expertise, car même si la structure est la même, les expressions et le titre des sections varient entre chaque rapport. Ce qui est aussi valable au sein d'un même réseau où l'on peut voir des différences dans le nom des sections ou les termes précédant les champs recherchés.

Nous nous sommes tournés vers des méthodes d'apprentissage statistiques qui sont plus souples pour reconnaître des expressions et entités malgré des petites variations. Dans le cas de l'apprentissage statistiques la NER peut s'apparenter à de la classification textuelle. On étudie chaque mot ou groupe de mots et on regarde la probabilité qu'il appartienne à la classe recherchée en prenant en compte le contexte, les autres mots du document. Ce problème n'est pas nouveau et on peut citer les travaux de (Bikel et al. 1998 ; Rau 1991 ; Rabiner 1989). Aujourd'hui cette tache est principalement faite avec des réseaux de neurones, voir (J. Li et al. 2020).

On se base sur un échantillon d'apprentissage pour entraîner le modèle et c'est cette étape que le logiciel Prodigy permet de faciliter. Grâce à ce logiciel on peut créer un échantillon d'apprentissage rapidement, ce qui dans un contexte industriel est un atout de taille, l'interface est illustrée en 4.7. Nous avons essayé de récupérer deux types d'informations, des données textuelles sur la description de l'environnement et de la construction, présentes dans l'exemple, et les données sur le chiffrage des dommages. Pour cela nous avons travaillé avec un réseau d'expertise qui nous a mis à disposition des rapports. Nous avions un petit nombre de rapports et le but de l'étude était de tester la faisabilité du projet. Nous avons deux échantillons distincts, dans un premier temps un de 900 dossiers sur lequel nous avons testé les données de description, cependant cet échantillon comprenait beaucoup de dossiers sans suite. Nous n'avions donc pas assez de dossiers pour travailler sur les dommages. Nous avons reçu un autre échantillon comprenant que des dossiers garanties, permettant de traiter les dommages. Cet échantillon est plus conséquent, il comporte 4000 dossiers.

Les modèles utilisés pour le NER sont les modèles de la librairie spaCy. C'est une librairie largement utilisée qui fournit des bons résultats. On peut trouver une description et comparaison des différentes librairies couramment utilisées dans (Schmitt et al. 2019 ; Nasiboglu et Gencer 2021). Dans notre cas nous utilisons le modèle pré-codé dans le logiciel prodigy qui repose sur des Réseaux de Neurones Convolutionnels avec un Word Embedding préalable. Nos catégories ne correspondent pas à des catégories pré-entraînables et le corpus étant très spécifique nous utilisons un modèle vide que nous entraînons sur nos données. Cet algorithme est décrit dans (Honnibal 2017). C'est assez similaire au CNN utilisé pour la classification textuelle, avec cependant des différences dans la façon dont l'embedding est fait.

Dans le premier échantillon de 900 dossiers nous avons essayé de récupérer les informations

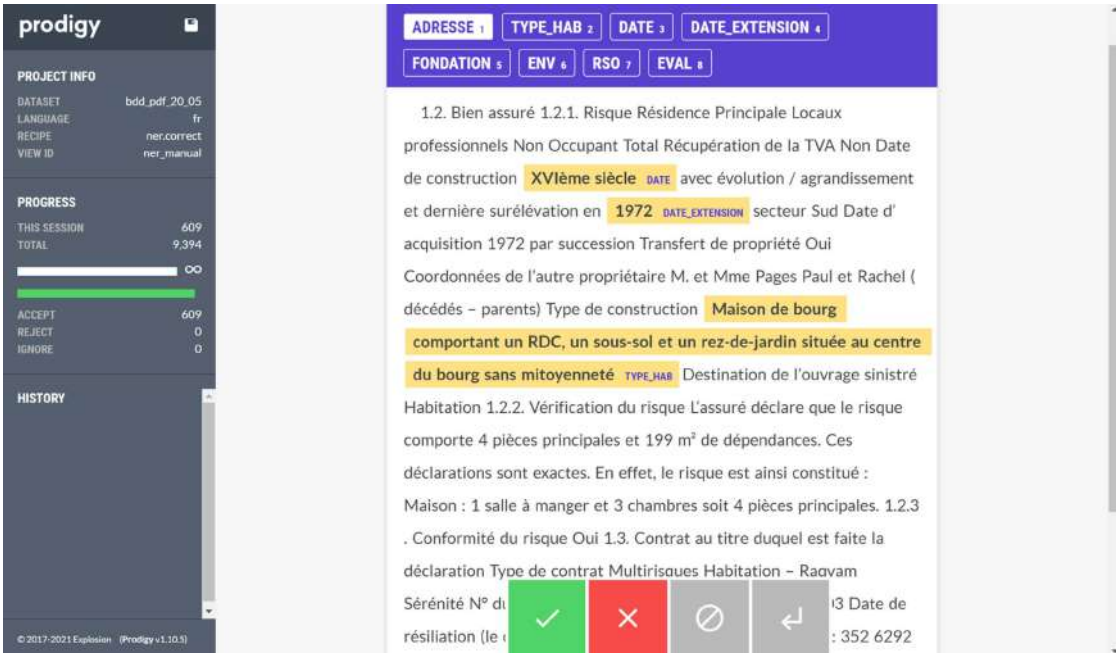


FIGURE 4.7 – Interface du Logiciel Prodigy, utilisé pour l'annotation des documents, sur un exemple de rapport d'expertise

suivantes :

- Adresse, adresse complète du bien.
- Date de construction, année de la construction, présente dans le texte décrivant le bien.
- Date d'extension, année de construction l'extension s'il y en a une.
- Environnement, description de l'environnement, avec notamment la pente qui est facteur important pour les dommages liés à la sécheresse.
- Évaluation, coût total du sinistre quand il y en a un, nous avions dans cet échantillon un certain nombre de dossiers sans-suite.
- Type de fondations, description des fondations de la maison, autre facteur d'importance pour les sinistres sécheresses.
- Type d'habitation, type de construction et présence de sous-sol ou de vide sanitaire.

Dans le deuxième échantillon de 4000 dossiers qui ont tous un coût nous avons cherché :

- Numéro de dossier, un numéro permettant d'identifier le sinistre.
- Adresse, adresse complète du bien.
- Date, année de début de la reconnaissance CatNat associée.
- Coût total, coût total du sinistre.
- Coût de la reprise en sous-oeuvre, coût associé à la reprise en sous-oeuvre des fondations.
- Coût des dommages sur les façades, coût associé aux travaux de réparations de la façade.
- Coût de l'étude de sol, coût de l'étude de sol quand il y en a une.

Dans les deux cas nous avons évalué les modèles sur un échantillon d'apprentissage conséquent, 700 dossiers pour le premier et près de 600 pour le deuxième. En effet dans cette démarche exploratoire nous cherchons avant tout à évaluer les fiabilités de cette méthode. Nous avons obtenu des résultats très encourageants pour les données sur la description et un peu moins sur

les chiffrages. Nous regardons à chaque fois la Précision, le Rappel et le F_1-score précédemment introduit en section 5.3.1.

Informations	Précision	Rappel	F_1-score
Adresse	0.89	0.89	0.89
Date de construction	0.83	0.80	0.82
Date d'extension	0.56	0.45	0.50
Environnement	0.95	0.93	0.94
Évaluation	0.89	0.95	0.92
Type de fondations	0.92	0.93	0.92
Type d'habitation	0.97	0.97	0.97

Tableau 4.2 – Résultats des prédictions pour la description de la maison

Informations	Précision	Rappel	F_1-score
Numéro de dossier	0.99	0.97	0.98
Adresse	0.91	0.90	0.91
Date	0.93	0.92	0.92
Coût total	0.89	0.61	0.72
Coût de la reprise en sous-œuvre	0.94	0.52	0.66
Coût des dommages sur les façades	0.87	0.45	0.67
Coût de l'étude de sol	0.75	0.24	0.37

Tableau 4.3 – Résultats des prédictions pour les dommages

En regardant les résultats on voit effectivement que le rappel pour les dommages est relativement bas. Le modèle n'arrive pas à bien retrouver les informations. En effet les chiffrages sont souvent dans des tableaux et la mise en page rentre alors beaucoup en jeu. En prenant que le texte du PDF on perd cette information. C'est un problème de formatage de la donnée. Sans investir beaucoup de temps dans le pré-traitement cette méthode ne semblent pas adaptée pour récupérer le chiffrage. Le test est néanmoins très concluant pour les données textuelles où les modèles de NER donnent des résultats tout à fait acceptables. Les résultats sont bons pour ce type de problème avec un F_1-score autour de 0.9.

4.4 Discussion des résultats

Ces études permettent donc d'avoir pour chaque aléa étudié les coûts moyens et la part total du coût par composante. Ces résultats sont valorisés dans les productions MRN, (*Lettre d'information de la Mission Risques Naturels 34* 2020 ; *Lettre d'information de la Mission Risques Naturels 36* 2021), aussi reproduits en 4.9 et 4.8.

Pour la tempête on peut remarquer que les trois composantes les plus impactées sont la Charpente, les autres ouvrages extérieurs, et les embellissements, avec respectivement 37 %, 27 % et 12 % de la charge des sinistres. Lors de tempêtes, les dommages sont attendus sur les charpentes, du fait des effets du vent. Le coût moyen est assez élevé autour de 4 500 €. Pour les autres ouvrages extérieurs cela correspond aux dommages sur les abris extérieurs, les clôtures et les portails, le coût moyen est inférieur à la charpente mais reste important autour de 4 000 €. La dernière composante la plus impactée correspond aux embellissements. Cela fait souvent suite aux dommages de mouille, les dommages entraînent un coût moyen plus faible autour de 2 300 €.

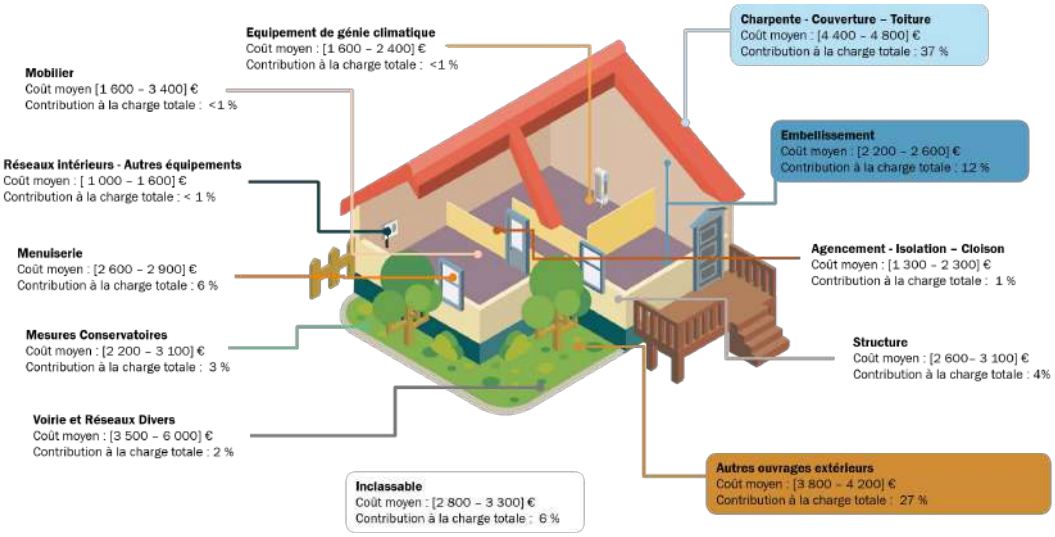


FIGURE 4.8 – Intervalle de confiance bootstrap des coûts moyens des composantes du bâti et part dans la charge pour la tempête.

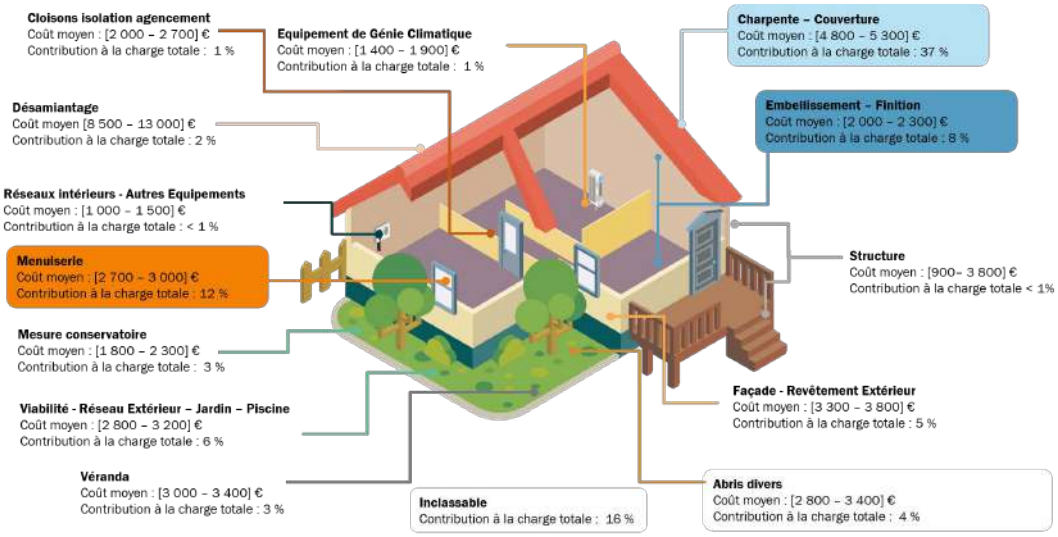


FIGURE 4.9 – Intervalle de confiance bootstrap des coûts moyens des composantes du bâti et part dans la charge pour la grêle.

Pour la grêle on obtient des résultats similaires, les composantes Charpente et embellissements sont aussi impactées et ont des coûts moyens analogues. Ici la composante menuiserie remplace la composante autres ouvrages extérieurs dans les composantes les plus impactées et a un coût moyen qui avoisine les 3 000 €.

En appliquant des réseaux de neurones avec du word embedding pour faire de la classification

textuelle des données d'expertise, on arrive à obtenir des résultats très intéressants. Cependant, il s'agit d'échantillons restreints et ces analyses gagneraient à être renforcées. On peut grâce aux développements en analyse de texte bien traiter les données provenant des rapports d'expertise avec une bonne précision. Il est également possible, avec une bonne fiabilité, de récupérer les informations textuelles décrivant le bien, dans les rapports en entier. C'est une très bonne source de données qui peut aider à comprendre la sinistralité à l'échelle du bâti. On peut par exemple identifier les zones du bâtiment les plus vulnérables et faire des actions de préventions pour réduire le coût sur ces composantes. Pour pouvoir pleinement faire un observatoire de la sinistralité il faudrait néanmoins étendre l'analyse à plus de réseaux d'expertises.

Chapitre 5

Estimation du coût d'un épisode de sécheresse

Ce chapitre se base sur l'article (Heranval, Lopez et Thomas 2021) et il présente une application de méthodes d'apprentissage statistique pour l'estimation du coût de la sécheresse en France. Nous commençons par des modèles linéaires généralisés simples, combinés avec des pénalités Lasso et Elastic-Net et nous nous tournons ensuite vers des algorithmes d'apprentissage automatique, tels que les forêts aléatoires ou l'Extreme Gradient Boosting. Nous étudions comment ils permettent de déterminer les communes potentiellement sinistrées et nous lions cette information aux données d'exposition pour prédire un coût.

5.1 Introduction

Comme évoqué en introduction le coût des dommages causés par les catastrophes naturelles en France, notamment la sécheresse, devrait augmenter dans les années à venir en France. Le changement climatique a un impact important sur la sécheresse et ses effets, notamment sur les maisons individuelles en raison des mouvements du sol liés à l'argile. Ces mouvements sont causés par le retrait et le gonflement de l'argile en réponse à l'humidification et à l'assèchement du sol. Le retrait et le gonflement de l'argile provoquent des mouvements verticaux et horizontaux, qui peuvent entraîner des dommages significatifs surtout aux maisons individuelles (*Avant de construire – Prendre en compte les risques du terrain* 2014 ; Assadollahi 2019). Ces dommages ont été observés dans d'autres pays et le coût associé est également très important, jusqu'à 500 millions de livres sterling par an au Royaume-Uni, voir (Pritchard, Hallett et Farewell 2015).

Nous proposons ici une méthode pour estimer le montant total des dommages d'un événement de sécheresse peu après sa survenance pour l'ensemble du marché français. L'objectif principal est de fournir des outils à la fédération et aux compagnies d'assurance pour évaluer la sévérité des événements de sécheresse. En raison de l'importance des montants en jeu l'évaluation de l'ordre de grandeur du coût d'un tel épisode est un enjeu majeur. Comme nous l'avons sur les dernières années, en moyenne, 900 millions d'euros sont indemnisés chaque année. C'est donc un poste important pour les sociétés d'assurance et un enjeu national de plus en plus préoccupant. En étudiant ce risque et son lien avec le phénomène météorologique nous pouvons apporter de la connaissance utile à l'intérêt général.

La spécificité du système français et du régime CatNat, décrit précédemment rend cette

estimation d'autant plus difficile. Notamment car les reconnaissances CatNat conditionnent l'indemnisation. Cela alourdit la gestion des sinistres puisque l'on observe un délai moyen de 18 mois avant la reconnaissance CatNat et un taux de non-reconnaissance élevé (*Sécheresse Géotechnique, de la connaissance de l'aléa à l'analyse de l'endommagement du bâti* 2018). Le phénomène de sécheresse étant complexe et multifactoriel on peut aussi observer des maisons dans des communes reconnues CatNat, se voir refuser l'indemnisation suite à la visite de l'expert. Celui-ci peut en effet établir que la sécheresse n'est pas le facteur déterminant. Les sinistres étant graves et souvent complexes l'indemnisation peut prendre beaucoup de temps avec des lourds travaux de reprise s'étalant dans le temps. Il peut s'écouler plus de cinq ans entre l'occurrence d'un événement de sécheresse, c'est à dire communément une année, et sa clôture par une société d'assurance. Dans ce contexte l'estimation précoce de la sécheresse est très importante pour le provisionnement des sociétés d'assurance.

Notre méthode pour prédire le coût des épisodes de sécheresse repose sur la comparaison de différents modèles statistiques telles que les modèles linéaires généralisés combinés avec des pénalités Lasso et Elastic-Net (Friedman, Hastie et Rob Tibshirani 2010) avec des algorithmes d'apprentissage automatique, tels que les forêts aléatoires (Breiman 2001) ou l'Extreme Gradient Boosting (Chen et Guestrin 2016). La calibration de ces méthodes est effectuée sur une importante base de données couvrant environ 70 % du marché français de l'assurance dommage. Une difficulté importante réside dans le fait que cette base de données est très déséquilibrée. En effet, les sinistres liés à la sécheresse sont relativement rares à l'échelle de la commune, et la plupart des communes ne sont pas affectées par aucun sinistre. Pour améliorer les performances et bénéficier de tous les modèles considérés, nous proposons une agrégation des résultats sur laquelle nous pouvons baser de nouvelles prédictions. Les prédictions obtenues à partir des différents modèles sont ainsi évaluées à l'aide des courbes Precision et Recall, des F_1-scores et des matrices de confusion.

Il est intéressant de mentionner que deux travaux très récents traitent aussi de l'évaluation de l'impact de la sécheresse en France. Le premier, (Ecoto, Bibaut et Chambaz 2021), utilise des Super Learner pour prédire le coût d'un sinistre au niveau de la commune. Le problème de prédiction est légèrement différent puisque la reconnaissance en l'état de catastrophe naturelle des communes qui sont étudiées dans (Ecoto, Bibaut et Chambaz 2021) est déjà connu au moment de la prédiction. Dans notre cas, cette information n'est pas disponible et nous cherchons à estimer le coût avant les reconnaissances. Le deuxième article, (Charpentier, Molly Rose James et Ali 2021), considèrent un problème similaire au nôtre, mais en se basant sur des données et des indices différents pour mesurer la sévérité de la sécheresse. Toutes ces approches, y compris celle que nous proposons dans la présente thèse, contribuent à une évaluation de l'impact de la sécheresse qui manque aujourd'hui d'indicateurs officiels précis.

Le reste de ce chapitre est organisé comme suit. Dans la section 5.2, nous décrivons le cadre de ce problème et la variable utilisée pour prédire le coût. La section 5.3 est consacrée à la description générale des modèles statistiques utilisés pour la prédiction et des résultats de ces prédictions. La section 5.4 présente les résultats de la prédiction des coûts. Le chapitre se termine par une discussion dans la section 5.5.

5.2 Description du problème et des données

Dans cette partie nous décrivons le problème que nous cherchons à résoudre et les données utilisées pour entraîner nos modèles.

5.2.1 Un problème de classification binaire

Une première étape pour prévoir le coût d'un événement de sécheresse est d'identifier les communes qui pourraient être reconnues en état de catastrophe naturelle. Malheureusement, ce n'est pas possible car, deux obstacles se présentent :

— l'incertitude liée au fait que la commune n'effectuera pas nécessairement une demande à la commission ;

— l'indisponibilité de l'indice météorologique qui sera utilisée par la commission dans sa prise de décision, qui n'est pas public immédiatement après l'événement.

Pour surmonter ce problème, nous proposons de prédire plutôt les communes susceptibles d'avoir un sinistre, ce qui est rendu possible par les bases de données de la MRN. On prédit les sinistres à l'échelle de la commune et pour une année entière.

Mathématiquement parlant, nous avons affaire à un problème de classification binaire. Soit $Y \in \{0, 1\}$ la variable réponse et $X \in \mathbb{R}^p$ les covariables, $Y = Y_{ij}$ est égal à 1 si un événement de sécheresse a eu lieu dans la commune i l'année j et 0 sinon. Notre objectif est donc d'estimer $\mathbb{P}[Y = 1 \mid X]$. Les résultats de ce problème de prédiction sont ensuite associés à un coût dans la section 5.2.4. Dans les sections suivantes, nous décrivons la base de données et les covariables utilisées pour traiter ce problème de classification binaire.

5.2.2 La base de données SILECC RGA

Bien que la base de données de SILECC couvre plusieurs risques naturels, nous nous concentrons dans ce document sur les événements de sécheresse, c'est-à-dire les sinistres liés au retrait et au gonflement de l'argile. Nous nous sommes concentrés sur la période de 2003 à 2018 pour laquelle nous disposons d'un nombre suffisant de sinistres. Cette période offre une forte représentativité des événements de sécheresse en France et couvre les épisodes majeurs tels que ceux observés en 2003 et en 2011. Nous avons utilisé les données de 2003 à 2017 pour l'estimation et l'année 2018 a été conservée comme validation de nos méthodes de prédiction.

Dans la base de données, le nombre de communes concernées par un sinistre représente 6% du nombre total de communes en France métropolitaine. Cela correspond en moyenne à 1 948 communes avec un sinistre par an, sur 34 840 communes en France métropolitaine. Ce ratio varie au cours de notre période. Comme écrit précédemment, pour certaines années, un grand nombre de sinistres a été observé, comme en 2003 où 25 % du nombre total de communes ont été touchées par un sinistre, 2017 et 2011 avec 12 % et 10 % sont également importants. La figure 5.1 montre le pourcentage annuel de communes touchées par un sinistre par rapport au nombre total de communes touchées par un sinistre entre 2003 et 2017.

5.2.3 Covariables

Pour caractériser la susceptibilité au retrait-gonflement des argiles dans le sol, nous nous appuyons sur l'indicateur résultant de la cartographie publiée par le BRGM décrite en 3.3.2. Cet indice peut être compris comme un facteur de risque, et un classement des différentes zones est effectué, définissant trois classes hiérarchiques, fournissant une cartographie nationale qui décrit la propension au gonflement de l'argile. Cette cartographie nous permet de calculer la surface et la proportion de chaque zone (faible, moyenne et forte) au niveau de la commune. Nous avons ensuite estimé le nombre de maisons individuelles dans chaque zone en utilisant les données de l'INSEE de 2015. Pour tenir compte de l'évolution du nombre de maisons individuelles, nous avons appliqué une augmentation ou une réduction de 1% pour chaque année (Arnold 2018).

En ce qui concerne l'indice météorologique, celui utilisé par la commission est disponible depuis avril 2021 mais n'est pas publié assez tôt pour l'anticipation des coûts. Nous avons donc

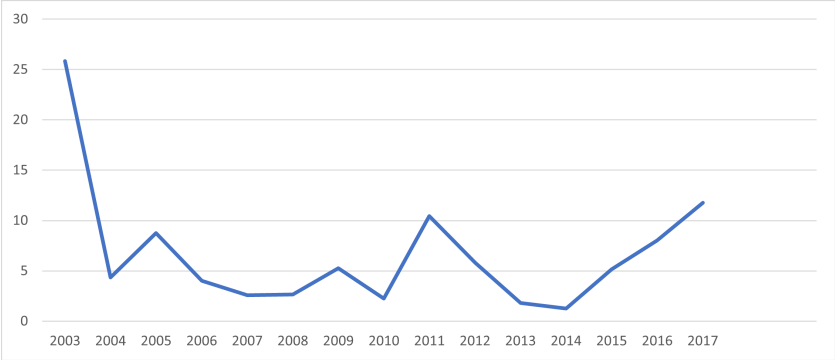


FIGURE 5.1 – Pourcentage annuel de communes touchées par un sinistre sur le nombre total de communes touchées par un sinistre entre 2003 et 2017

utilisé un autre indice météorologique spatio-temporel, produit par Météo-France, l'indice standardisé d'humidité des sols (SSWI), comme indicateur de la sévérité d'un événement de sécheresse. Cet indice est issu d'un projet de recherche de Météo-France appelé Climsec, décrit dans (Jean-Philippe VIDAL et MOISSELIN 2011). Le calcul de l'indice SSWI est effectué par l'analyse des précipitations, de l'humidité du sol et des débits des suites hydrométéorologiques Safran-Isba-Modcou (SIM) (HABETS et al. 2008), et s'inspire des procédures de calcul de l'indice de précipitation standardisé (SPI) (MCKEE, DOESKEN, KLEIST et al. 1993). Une description plus détaillée peut être trouvée dans (J.-P. VIDAL et al. 2012). Quatre séries temporelles sont ensuite obtenues à partir des séries temporelles du SSWI sous forme de moyennes mobiles sur un, trois, six et douze mois. Cela nous donne quatre indices pour chaque mois représentant l'humidité du sol.

L'indice SSWI est un indice normalisé et prend donc des valeurs centrées autour de 0. Une valeur négative suggère une sécheresse alors qu'une valeur positive suggère l'humidité. La figure 5.2 illustre la distribution géographique de l'indice SSWI pour 2018. Il est très variable et 2018 a été une année avec une sécheresse importante en France.

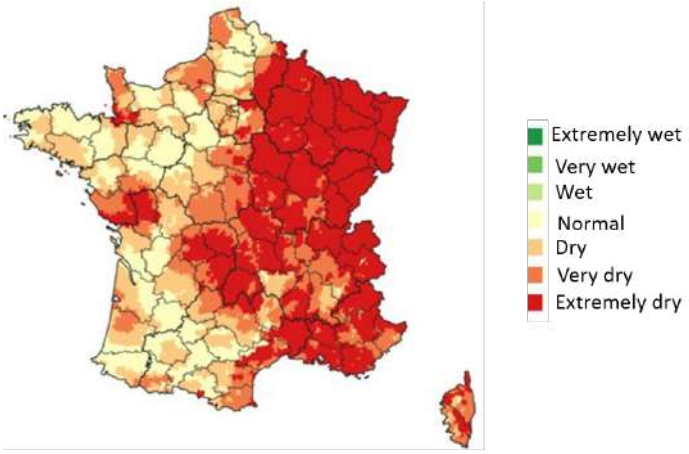


FIGURE 5.2 – Cartographie du SSWI pour l'année 2018 (Source : MRN)

Nous avons également calculé quatre autres indices pour caractériser l'épisode de sécheresse lui-même, tel que définis dans (J.-P. Vidal et al. 2012) :

— la durée : le nombre de mois consécutifs pendant lesquels l'indice SSWI est négatif ;
— la gravité : la valeur absolue de la valeur minimale du SSWI atteinte pendant l'événement ;
— la magnitude : la valeur absolue de la somme des SSWI pendant l'événement ;
— la rareté : une classification de la sévérité en 7 classes (extrêmement humide, très humide, humide, normal, sec, très sec et extrêmement sec) comme le montre la figure 5.3.

Ces indices sont calculés pour chaque commune et pour chaque année. Dans le cas d'événements multiples, (dans notre cas, le maximum est de quatre événements au cours d'une même année,) nous utilisons la valeur des indices pour tous les événements survenus au cours de l'année. La figure 5.3 illustre la définition de ces indices sur un exemple.

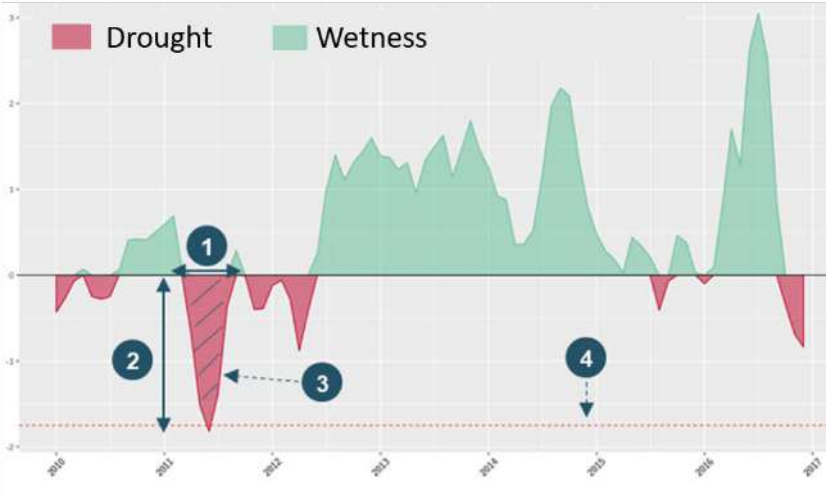


Figure 5.3 – Description des quatre indices basés sur le SSWI que nous utilisons à partir d'un exemple. 1 représente la durée de l'événement, 2 sa sévérité, 3 son ampleur et 4 sa rareté.

Nous avons également utilisé des indications sur l'état de catastrophe naturelle. Les critères conduisant à l'arrêté de catastrophe naturelle ont changé six fois au cours des 20 dernières années. Par conséquent, les mêmes effets peuvent ne pas entraîner les mêmes conséquences dans notre base de données, selon la période. L'indemnisation est en effet conditionnée par la reconnaissance CatNat, même s'il existe, s'il n'y a pas de reconnaissance un sinistre n'apparaîtra pas dans notre base. Toutes ces contraintes rendent la prédiction délicate. Nous avons donc mis en place une variable qui indique les critères utilisés par la commission au moment de la demande. C'est une variable catégorielle qui dit quel critère été utilisé en fonction des années. Nous avons également considéré les résultats de la décision de la commission comme si elle avait utilisé notre indice météorologique SSWI au lieu du sien. Nous avons reproduit le calcul décrit dans (*Météo-France dans le dispositif CATNAT sécheresse.* 2020), mais en utilisant le SSWI à notre disposition pour calculer les rangs. En effet pour être reconnue, on considère que la durée de retour de la sécheresse géotechnique doit être supérieure ou égale à 25 ans et pour cela on regarde si l'indicateur du trimestre considéré se classe au premier ou deuxième rang parmi les indicateurs calculés sur les 50 dernières années pour ce trimestre. Les indices prennent des valeurs différentes mais en calculant ce critère avec nos valeurs, nous supposons que les distributions sont assez similaires. Cette variable nous donne une information permettant de dire si la commune est susceptible d'être reconnue.

Dans l'ensemble, notre base de données contient 154 variables et 522 600 observations, toutes les variables sont numériques et les variables catégorielles ont été codées comme des variables binaires.

Catégorie de données	Nombre
Covariables relatives au SSWI	96
Description des événements de sécheresse	37
Critère utilisé par la commission	4
Susceptibilité au retrait et au gonflement de l'argile	11
Population dans la commune	1
Déclarations de catastrophe naturelle passées	4

TABLEAU 5.1 – Synthèse des données utilisées pour notre base de données d'apprentissage

Ces variables constituent notre base de données d'apprentissage, sur laquelle les modèles seront entraînés. La variable à prédire étant l'occurrence d'un sinistre rapporté dans la BD SILECC pour la commune dans l'année. En effet on agrège les sinistres à la commune et à l'année et la variable devient binaire indiquant seulement si un sinistre a été enregistré ou non.

5.2.4 Méthode générale

La première étape de notre méthode consiste à prédire les communes qui auront un sinistre pendant un épisode de sécheresse. Pour cela, nous avons utilisé différents modèles d'apprentissage usuels, décrits dans la Section 2.3. Nous avons utilisé des modèles linéaires généralisés pénalisés avec le `R`-package `GLMNET`, (FRIEDMAN, HASTIE et Rob TIBSHIRANI 2010), décrit en 2.3.2, des forêts aléatoires avec le `R`-package `ranger`, (WRIGHT et ZIEGLER 2017), décrit en 2.3.4, et des Extreme Gradient Boosting avec le `R`-package `xgboost`, (CHEN et GUESTRIN 2016), décrit en 2.3.5.

Une fois que nous connaissons les communes qui sont susceptibles d'être affectées par un événement de sécheresse, nous avons calculé le nombre de maisons dans ces communes qui ont une susceptibilité moyenne ou forte au retrait et au gonflement de l'argile. Pour ce faire, nous avons encore utilisé la cartographie réalisée par le BRGM. Nous avons ensuite utilisé une régression linéaire pour relier le nombre de maisons au coût de l'événement. Ce modèle linéaire a été entraîné sur notre base de données de sinistres. La figure 5.4 résume cette méthodologie globale.

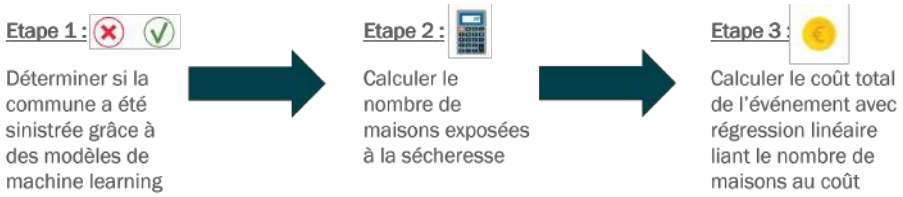


FIGURE 5.4 – Méthode générale

5.3 Résultats

5.3.1 Évaluation des performances

Pour évaluer les performances des différents modèles, nous avons divisé aléatoirement notre base de données en un ensemble d'apprentissage (80 %) et un ensemble de test (20 %). Rappelons que notre base de données est très déséquilibrée dans le sens où la proportion de communes ayant eu un sinistre est très faible.

Les méthodes courantes d'évaluation des performances des classifieurs binaires comprennent les taux de vrais positifs et de vrais négatifs, ainsi que les courbes ROC (Receiver Operating Characteristics), qui affichent le taux de vrais positifs par rapport au taux de faux positifs. Ces méthodes sont toutefois peu informatives lorsque les classes sont fortement déséquilibrées. Dans ce contexte, le F_1-score et les courbes Precision-Recall (PRC) sont plus indiquées : (Brownlee 2020 ; Saito et Rehmsmeier 2015). Ils sont tous deux calculés à partir des valeurs de :

$$\text{Precision}(p_c) = \frac{\text{vrai positifs}}{\text{vrai positifs} + \text{faux positifs}}$$

et

$$\text{Rappel}(p_c) = \frac{\text{vrai positifs}}{\text{vrai positifs} + \text{faux négatifs}},$$

où p_c est une probabilité de coupure variant entre 0 et 1. La précision quantifie le nombre de prédictions positives correctes parmi toutes les prédictions positives effectuées et le rappel (souvent appelé sensibilité) quantifie le nombre de prédictions positives correctes parmi toutes les prédictions positives qui auraient pu être effectuées. Les deux se concentrent sur la classe des positifs, minoritaire, qui sont les communes avec un sinistre et ne se préoccupent pas des négatifs, majoritaire, qui sont les communes sans sinistre.

Le F_1-score combine ces deux mesures en un seul indice en prenant la moyenne harmonique de ces deux valeurs. Il est dérivé de la mesure F introduite dans (Chinchor et Sundheim 1993 ; Rijsbergen 1979) et défini comme suit

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Rappel}}{\text{Precision} + \text{Rappel}}.$$

Plus le score F_1- est proche de 1, plus le modèle de prédiction est bon.

La courbe PRC affiche les valeurs de Précision et de Rappel lorsque la probabilité de coupure p_c varie de 0 à 1. La courbe PRC d'un bon modèle tend vers le point de coordonnées $(1, 1)$. La courbe d'un mauvais classificateur sera une ligne horizontale sur le graphique avec une coordonnée y proportionnelle au nombre de positifs dans l'ensemble de données. Pour un ensemble de données équilibré, cette valeur sera de $0,5$ (Brownlee 2020).

La PRC et le F_1-score sont complémentaires dans notre approche, la PRC est utilisée sur les prédictions probabilistes des modèles, elle donne la meilleure configuration et le meilleur modèle. Alors que le F_1-score est utilisé pour sélectionner la meilleure valeur de seuil, la probabilité de coupure p_c, utilisée pour la prédiction des deux classes, pour chaque modèle. Nous allons détailler ce processus dans la section suivante.

5.3.2 Résultats

Dans cette section, nous présentons les principaux résultats de notre analyse pour les trois modèles, GLMNET, les forêts aléatoire (RF) et Extreme Gradient Boosting (XGBOOST). Nous

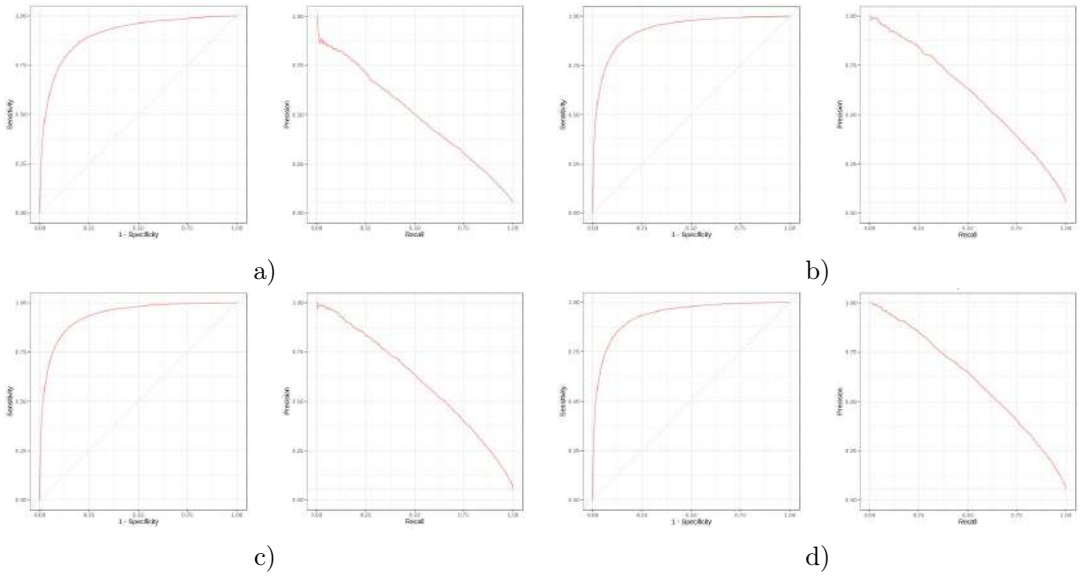


Figure 5.5 – ROC (Gauche) and PR (Droite) courbes pour a) GLMNET b) RF c) XGBOOST d) AGGREGATE, réalisées à partir de (Saito et Rehmsmeier 2015), sur l'échantillon de test contenant 5 924 communes avec sinistres et 98 596 sans sinistres.

considérons également l'agrégation de ces trois modèles. Pour l'agrégation (AGGREGATE), nous avons examiné une moyenne arithmétique simple des probabilités résultants de chacun des trois modèles, afin de faire une synthèse de toutes les prédictions. Pour chacun, la figure 5.5 montre le ROC et la PRC et le tableau 5.2 l'aire sous la courbe (AUC) pour le ROC et la PRC. Plus l'AUC est proche de 1, meilleure est la méthode de prédiction.

Modèle	AUC ROC	AUC PRC
GLMNET	0.907	0.503
RF	0.933	0.604
XGBOOST	0.936	0.609
AGGREGATE	0.936	0.615

Tableau 5.2 – AUC des différents modèles

Dans le tableau 5.2, nous pouvons remarquer que l'AUC pour le ROC est proche de 1 pour les quatre modèles alors que l'AUC pour le PRC a des valeurs autour de 0,60. Cela illustre ce qui a été expliqué dans la section 5.3.1, les courbes ROC ne sont pas pertinentes pour les ensembles de données déséquilibrés, puisque les courbes ROC se concentrent également sur les classes positives et négatives, (Saito et Rehmsmeier 2015). Par conséquent, lorsque la classe des négatifs est largement prédominante dans l'ensemble de données, un modèle prédisant toujours un négatif aura une AUC proche de 1, mais ne prédira aucun des positifs.

Les AUC PRC dans le tableau 5.2 montrent que :

— le XGBOOST et le RF semblent fournir de meilleurs résultats sur nos données que le GLMNET ;

— l'agrégation donne les meilleurs résultats, ce qui plaide fortement en faveur de l'utilisation de cette méthode.
— Les courbes PRC de la figure 5.5 semblent être convenables compte tenu de notre problème de classification, elles ne sont pas parfaites mais le compromis est acceptable.

Il est également très intéressant de voir que l'agrégation lisse les résultats et tire le meilleur de chaque modèle. Le début du graphique est moins discontinu. L'agrégation semble être le meilleur modèle à utiliser pour la prédiction.

Nous avons ensuite choisi un seuil pour faire la prédiction, c'est-à-dire la valeur de la probabilité de coupure p_c au-delà de laquelle nous considérons que la prédiction est un 1. Une valeur classique de 0,5 donnerait le même poids au 0 et au 1 et supposerait qu'ils sont symétriques. Or ce n'est pas notre cas, les prédictions que nous essayons de faire sont rares et donc une probabilité de 0,5 peut déjà être un score fort. Nous cherchons donc à affaiblir ce déséquilibre en donnant une probabilité de coupure plus faible. Pour ce faire, nous avons utilisé le F_1-score et essayé différents seuils avec un pas de 0,001 pour trouver le meilleur compromis. Les résultats sont présentés dans le tableau 5.3.

Modèle	F1-score	Seuil
GLMNET	0.503	0.221
RF	0.570	0.306
XGBOOST	0.573	0.291
AGGREGATE	0.576	0.264

Tableau 5.3 – Meilleur F_1-score et seuil associé pour chaque modèle.

Encore une fois, nous pouvons voir que le meilleur F_1-score est obtenu par l'agrégation du modèle, avec une valeur de 0.576 et un seuil de 0.264. Nous trouvons le même ordre que précédemment, le XGBOOST fonctionne mieux que le Random Forest. Les seuils sont compris entre 0,2 et 0,3, ce qui confirme l'idée qu'un rééquilibrage est utile pour améliorer les prédictions.

La figure 5.6 montre les matrices de confusion pour chaque modèle. Sur les 5 924 communes affectées par un sinistre dans l'ensemble de test, les quatre modèles ont prédit entre 3 157 et 3 526 communes avec un sinistre. Cela signifie qu'ils parviennent à prédire plus de la moitié d'entre elles, mais il y a quelques fausses prédictions. GLMNET fait plus de fausses prédictions, ce qui peut expliquer les différences observées dans le F_1-score. Lorsque nous examinons les matrices de confusion, il n'y a pas de différence notable entre les autres modèles.

5.3.3 Importance des variables

L'importance d'une variable renseigne sur la mesure dans laquelle une variable influence les prédictions, faites à partir d'un modèle donné. Plus un modèle s'appuie sur une variable pour faire des prédictions, plus celle-ci est importante pour le modèle. L'importance de la variable est également un bon outil pour interpréter les modèles « boîte noire » et leurs performances (Robert Tibshirani 1996). L'importance des variables est fondée sur différentes métriques en fonction du modèle considéré. Pour GLMNET, elle est mesurée par la valeur du coefficient associé à la variable, après normalisation des données. Pour RF, l'importance de la variable est mesurée grâce à l'indice de Gini pour la classification (Wright et Ziegler 2017). Pour XGBOOST, l'importance de la variable représente la contribution fractionnelle de chaque caractéristique au modèle, sur la base du gain total des splits de la caractéristique. Un pourcentage plus élevé signifie une caractéristique prédictive plus importante. Le tableau 5.7 rapporte les 10 variables contribuant le plus selon les métriques pertinentes pour chaque modèle.

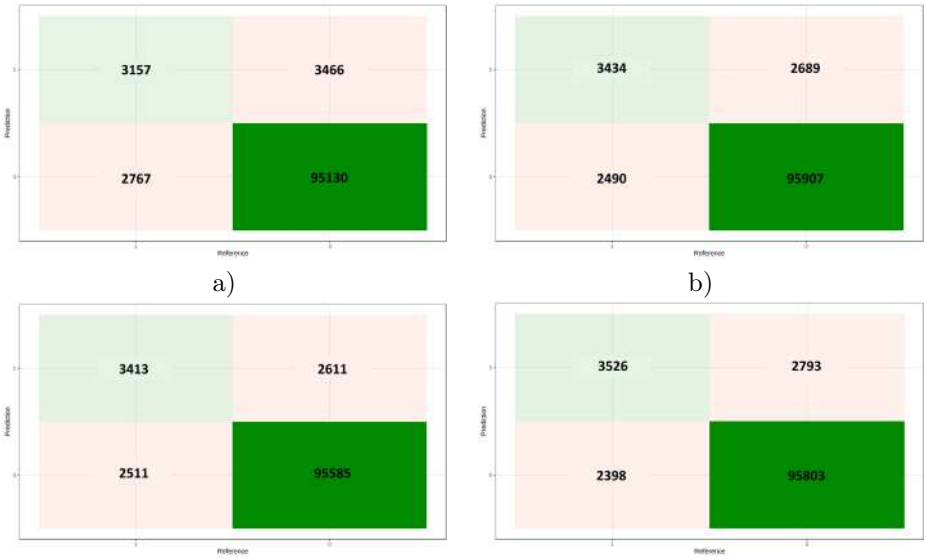


FIGURE 5.6 – Matrice de confusion pour a) GLMNET b) RF c) XGBOOST d) AGGREGATE sur l'échantillon test.

GLMNET semble s'appuyer davantage sur les données météorologiques et moins sur la description de la commune. Les deux autres méthodes, RF et XGBOOST, s'appuient davantage sur une combinaison d'exposition au retrait gonflement de l'argile et des déclarations CatNat d'événements passés que sur les données météorologiques. Nous ne montrons que les 10 premières et les données météorologiques sont toujours importantes pour RF et XGBOOST, elles viennent juste après. Néanmoins, il est intéressant de noter qu'elles ne sont pas les variables les plus importantes. Cela pourrait expliquer les différences observées dans les résultats, d'autant plus que RF et XGBOOST semblent donner des résultats très similaires, en termes de scores AUC PRC mais aussi pour le nombre de communes correctement prédites. De plus, beaucoup de variables disponibles ne sont pas utilisées, ce qui est attendu puisque nous avons choisi de donner beaucoup de covariables et d'utiliser des méthodes qui seront capables de choisir celles qui sont pertinentes.

5.4 Estimation du coût

Dans cette section, nous appliquons notre méthode pour prédire le coût d'un événement de sécheresse. Le modèle d'évaluation du coût est décrit dans la section 5.4.1 et les résultats finaux sont présentés dans la section 5.4.2.

5.4.1 Régression linéaire

Nous avons déjà mentionné que le coût d'un événement de sécheresse est susceptible d'être corrélé avec le nombre de maisons. En utilisant la base de données SILECC, nous avons ajusté un modèle pour quantifier cet impact. Comme la base de données représente 70 % du marché, nous avons multiplié chaque coût par 1,42 pour obtenir un coût pour l'ensemble du marché français. Nous avons agrégé les événements d'une même année afin de réduire la variabilité de l'estimation.

GLMNET	RF	XGBOOST
Valeur maximale de l'indice SSWI 12 pour février	Nombre de déclarations passées de catastrophe naturelle	Nombre de déclarations passées de catastrophe naturelle
Valeur maximale de l'indice SSWI 12 pour le mois d'août	Surface sans susceptibilité au retrait gonflement des argiles	Nombre d'événements pour l'année précédente
Valeur maximale de l'indice SSWI 6 pour novembre	Nombre de maisons	Nombre de maisons
Valeur maximale de l'indice SSWI 12 pour le mois de juin	Proportion de la surface ayant une faible susceptibilité au retrait gonflement des argiles	Surface ayant une faible susceptibilité au retrait gonflement des argiles
Valeur maximale de l'indice SSWI 3 pour le mois d'août	Surface en zone urbaine	Valeur minimale de l'indice SSWI 1 pour le mois d'août
Classement de la gravité des événements	Surface avec susceptibilité moyenne au retrait gonflement des argiles	Valeur minimale du SSWI 3 pour le mois d'octobre
Valeur maximale du SSWI 12 pour le mois de juin	Nombre de maisons ayant une susceptibilité moyenne au retrait gonflement des argiles	Nombre de refus passés avec le calcul effectué avec notre SSWI
Valeur minimale de l'indice SSWI 12 pour le mois de juin	Durée totale des épisodes de sécheresse	Surface avec une propension moyenne au retrait et au gonflement de l'argile
Valeur maximale de l'indice SSWI 12 pour janvier	Nombre d'événements pour l'année précédente	Nombre de maisons avec moyenne susceptibilité au retrait gonflement des argiles
Valeur maximale de l'indice SSWI 6 pour le mois de janvier	Valeur minimale de l'indice SSWI 6 pour le mois de novembre	Durée totale des épisodes de sécheresse passés

FIGURE 5.7 – Top 10 des variables selon les indicateurs pertinents pour chaque modèle

Le coût annuel le plus important a été observé en 2003 avec 2 milliards d'euros et le coût moyen entre 2003 et 2017 était de 415 millions d'euros par an. Le nombre de maisons a une distribution similaire, allant jusqu'à 4,7 millions en 2003, avec une moyenne de 1,7 million de maisons par an. En désignant par M le coût d'un événement, on peut écrire

$$\mathbb{E}[M] = \text{Nombre de maison} \times 464.4 - 4.121e + 08,$$

avec des erreurs standard de $1.5e+08$ pour l'intercept et $55,79$ pour le nombre de maisons. Nous avons trouvé une bonne corrélation entre ces deux variables dans notre base de données, avec un $R^2 = 0.84$ et une erreur standard résiduelle de $2.198e + 08$.

La figure 5.8 montre que la régression linéaire est une approximation convenable pour notre problème. Nous utilisons l'intervalle de confiance de 95%, que nous utiliserons pour notre prédiction.

Ce modèle linéaire est bien sûr très grossier, car il est ajusté sur le petit nombre d'observations dont nous disposons (seulement 15), ce qui explique le choix du modèle de régression le plus simple. Bien que le R^2 soit relativement proche de 1, il ne faut bien sûr pas être trop confiant sur cet ajustement en raison du faible nombre de points utilisés pour estimer les paramètres du modèle.

5.4.2 Résultats des prédictions pour 2018

Le modèle précédent est ensuite relié aux modèles de prédiction de la section 5.3. Une fois que nous avons le nombre de maisons, nous pouvons estimer un coût avec un intervalle de confiance.

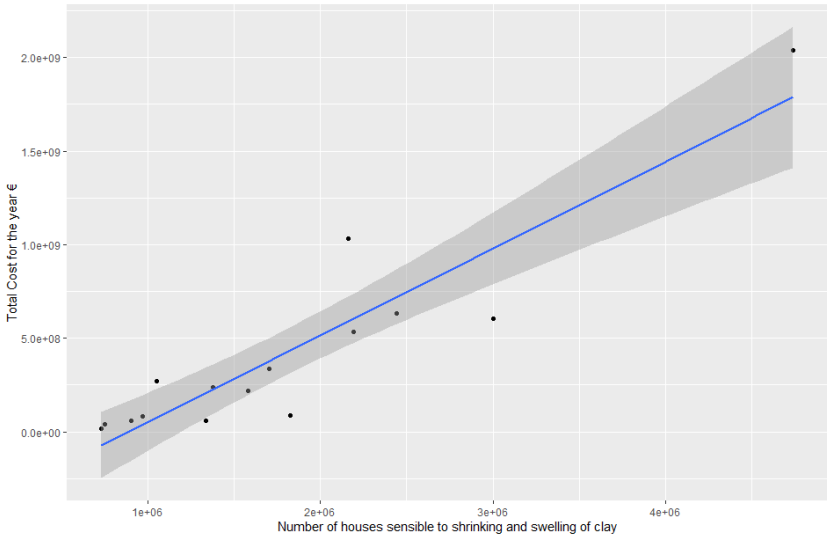


FIGURE 5.8 – Régression linéaire pour le coût des sinistres en fonction du nombre de maisons. Les points sont les observations, la ligne bleue la ligne de régression et en gris la bande de confiance.

Dans notre cas, la perte totale peut être écrite comme suit

$$L = \sum_{i=1}^{N} Y_i M_i,$$

où $Y_i = 1$ si un sinistre est survenu et 0 sinon, et M_i est le montant correspondant du sinistre (pour lequel nous connaissons le nombre de maisons n_i), et N est le nombre de communes considérées. Si Y_i et M_i sont indépendants, alors la variance est de

$$\mathrm{Var}(Y_i M_i \mid X) = \left(\mathbb{E}[M^2] p_i (1 - p_i) + p_i \mathrm{Var}(M)\right),$$

où $p_i = \mathbb{P}(Y_i = 1|X)$. Ainsi la variance σ^2 de L peut être estimée par

$$\widehat{\sigma}^2 = \sum_{i=1}^{n} \left(\hat{m}_{2,i} p_i (1 - p_i) + p_i \tilde{\sigma}^2\right),$$

où $\tilde{\sigma}$ est l'erreur standard estimée dans le modèle de régression linéaire de la section 5.4.1, et

$$\hat{m}_{2,i} = \tilde{\sigma}^2 + (\hat{\alpha} + \hat{\beta} n_i)^2,$$

avec $(\hat{\alpha}, \hat{\beta}) = (-4.121e + 08, 464.4)$, comme estimé dans la section précédente.

Alors les intervalles de confiance à 95 % de notre estimation peuvent être approximés par

$$\widehat{L} \pm 1.96 \widehat{\sigma}^2.$$

Les résultats de cette estimation sont présentés dans le tableau 5.4. Selon France Assureurs, le coût de la sécheresse en France pour 2018 est de 900 millions d'euros, voir (*L'assurance des événements naturels en 2019* 2021) et elle pourrait atteindre 1 200 millions. En 2022 son

estimation est de 1 300 millions (*L'assurance des événements naturels en 2020* 2022). Le résultat du modèle d'agrégation montre des résultats assez similaires. Même si nous n'avons pas une très bonne précision au niveau de la commune, le coût général est cohérent avec les données observées.

Modèle	Valeur basse	Estimation	Valeur haute
GLMNET	461 125 885	579 350 811	697 575 737
RF	1 396 432 680	1 618 225 685	1 840 018 69
XGBOOST	839 262 189	977 086 655	1 114 911 122
AGGREGATE	796 820 728	965 750 651	1 134 680 547

Tableau 5.4 – Estimation et intervalle de confiance pour les prédictions du coût de l'année 2018 (en euros)

Parmi ces quatre classes de prédicteurs :
— le modèle linéaire généralisé pénalisé a tendance à sous-estimer considérablement le coût, étant nettement inférieur à l'évaluation de référence.
— Les trois autres méthodes fournissent des résultats plus plausibles, probablement parce qu'elles sont plus flexibles qu'une simple approche paramétrique, et ont donc plus de capacité à capturer des phénomènes complexes.
— Les forêts aléatoires produisent une estimation qui va même au-delà des évaluations les plus pessimistes du risque par le marché, ce qui semble plaider en faveur des estimations des deux autres approches.

La différence dans les coûts prédits est essentiellement due à une différence dans le nombre prédit de maisons. Qui, lui-même, est lié au nombre de communes prédites par chaque modèle, comme on peut le remarquer dans le tableau 5.5. Nous pouvons noter pour 2018 que RF prédit plus de communes avec un sinistre alors que GLMNET en prédit moins. Cela plaide une fois de plus pour l'utilisation de l'agrégation car elle moyenne les prédictions et nous permet de prendre le meilleur de chaque prédiction.

Modèle	Nombre de communes	Nombre de maisons en zone de susceptibilité
GLMNET	1 364	2 134 000
RF	5 525	4 371 000
XGBOOST	1 800	2 991 000
AGGREGATE	1 823	2 966 000

Tableau 5.5 – Prévision du nombre de maisons et de communes touchées par la sécheresse de 2018

5.5 Conclusion et discussion

Dans ce travail, nous avons développé une méthodologie permettant d'estimer le coût des conséquences de la sécheresse pour l'ensemble du marché français. Nous avons d'abord utilisé un modèle linéaire généralisé avec pénalisation Elastic-Net, des modèles Random Forest et Extreme Gradient Boosting avec différents seuils pour prédire les communes susceptibles d'être affectées par un sinistre. Sur la base de ces prédictions, nous avons calculé le nombre de maisons qui ont une propension au retrait et au gonflement de l'argile, puis nous avons calculé le coût total par le biais d'une régression linéaire.

Nous avons obtenu des résultats encourageants pour un phénomène aussi complexe, bien que plusieurs incertitudes subsistent. Malgré des résultats modérés pour la prédiction des communes impactées, nous avons obtenu des résultats cohérents pour la prédiction des coûts. La base de données que nous avons utilisée, le processus des catastrophes naturelles et la nature de ce risque de sécheresse rendent la modélisation très complexe et incertaine. En effet, notre base de données est fondée sur des sinistres passés, rapportés par différents assureurs et contient certaines imprécisions, qui peuvent avoir un impact sur les résultats de la prédiction de la probabilité pour une commune d'être affectée par un sinistre.

Une deuxième difficulté vient du processus des arrêtés de catastrophes naturelles. Pour pouvoir obtenir une indemnisation, et donc apparaître dans notre base de données, la commune doit avoir été reconnue par la commission. Il peut y avoir des demandes dans des communes qui n'ont pas été reconnues. Nos modèles peuvent prédire de telles demandes, mais nous ne sommes pas en mesure d'évaluer si la prédiction est correcte ou non. De plus, au cours des 20 dernières années, les critères pour qu'une commune soit reconnue en état de catastrophe naturelle ont changé six fois. Par conséquent, dans notre base de données de train, nous pouvons avoir différentes caractéristiques qui auront des effets différents en fonction du critère.

De plus, avec les variables météorologiques et géologiques dont nous disposions, nous n'avons abordé qu'une partie des facteurs à l'origine du risque. L'interaction entre la structure de la maison et la composition du sol joue un rôle important pour déterminer si la maison sera endommagée par un épisode de sécheresse. Nous avons pris en compte la nature du sol avec les indicateurs du BRGM mais il est très difficile de prendre en compte la nature exacte de cette interaction en raison du manque de données sur les différents types de fondations et de la présence de végétation. En pratique un expert, formé spécialement, vient faire une visite sur site, après la reconnaissance CatNat, pour déterminer si la sécheresse est bien à l'origine du sinistre. C'est un processus complexe qui nécessite une analyse à l'échelle très locale. C'est un des objectifs du chapitre précèdent, nous avons analysé les données des rapports d'expertise pour faire ressortir des informations sur l'endommagement à l'échelle fine du bâtiment. Nous avons essayé d'analyser les rapports d'expertise, permettant d'identifier les facteurs aggravants de la sécheresse. Ces informations pourraient à terme permettre d'améliorer les estimations du coût global en identifiant et cartographiant les facteurs de vulnérabilité, mais nous avons vu que nous manquions de données pour pouvoir faire cela.

Nous avons également rencontré des difficultés pour évaluer notre modèle. Comme nous l'avons mentionné plus haut, les résultats sont incertains en raison du processus de reconnaissance. Plus généralement, il est difficile de trouver le bon score pour juger un modèle, surtout avec des données déséquilibrées. De plus, les prédictions que nous faisons ne peuvent être vérifiées plusieurs années après si l'on veut avoir tous les sinistres clôturés. Les résultats sont encourageants mais doivent être consolidés par des prédictions plus précises.

La modélisation à l'échelle de la commune avec seulement des données météorologiques et d'expositions est donc nécessairement incomplète. Malgré ces difficultés nos modèles ont des résultats cohérents qui fournissent des indications pertinentes pour évaluer l'ampleur et qui nous ont permis d'améliorer la prédiction des coûts de la sécheresse. Les techniques que nous avons utilisées pourraient être améliorées avec une quantité supplémentaire de données, et avec des connaissances supplémentaires sur les phénomènes de dépendance spatiale entre les communes (à savoir comment deux communes proches peuvent coordonner ou non leurs réponses). Soulignons que le principal avantage de notre approche est de fournir une réponse rapide à la question du coût de tels événements naturels, dans un contexte où le temps de réaction est important pour optimiser la gestion des risques. Enfin, mentionnons que les méthodes que nous avons développées pourraient également être étendues pour approximer ou prédire l'indice utilisé par la commission CatNat, afin d'améliorer la prédiction. En effet, cet indice a récemment été rendu disponible.

Chapitre 6

Estimation du coût des inondations

Dans ce chapitre nous proposons deux méthodes permettant d'estimer le coût d'un événement inondation. La première méthode repose sur la théorie de la crédibilité. Pour déterminer un a priori nous utilisons des arbres de régression avec une loi de Pareto généralisée. La deuxième méthode se base sur une approche sévérité fréquence.

6.1 Introduction

6.1.1 Contexte

L'estimation du coût des inondations est enjeu majeur pour le secteur de l'assurance notamment pour évaluer leurs expositions. Néanmoins c'est un exercice difficile qui comporte de nombreuses incertitudes (Hall et Solomatine 2008 ; Eleutério 2012).

Dans ce chapitre nous proposons deux méthodes permettant d'estimer le coût des événements inondations. Nous nous concentrons sur l'estimation des conséquences des inondations et nous ne nous intéressons pas à la modélisation du phénomène. Nous nous inscrivons dans le cadre d'une mission d'appui à France assureurs pour dimensionner les réponses en cas de gestion de crise liée aux événements naturels. Nous essayons donc d'estimer le coût d'un événement inondation après son occurrence. Nous nous intéressons qu'aux événements « majeurs », qui relèvent de la combinaison des critères suivants :

— Étendue spatiale et temporelle de l'événement importante ;
— Retentissement médiatique de l'événement ;
— Nombre de décès provoqués par l'événement ;
— Spécificité ou rareté de l'événement.

La principale contrainte est d'être capable de fournir une estimation rapidement mais nous avons aussi attaché une attention particulière à faire une méthode intelligible, simple d'utilisation et avec des paramètres contrôlables. Cet exercice étant difficile et toute estimation amenant forcément des incertitudes nous avons essayé de laisser une part de contrôle aux gestionnaires de risques. Les méthodes sont construites pour reposer avant tout sur l'expertise métier et la comparaison aux données passées.

En France, du fait du régime CatNat, la Caisse Centrale de Réassurance est un acteur incontournable et propose de telles estimations pour les catastrophes naturelles. Elle développe des outils de modélisation pour estimer le coût d'un événement quelques jours après sa survenance mais aussi pour mesurer son exposition financière. On peut trouver des descriptions de leurs méthodes pour les inondations dans (Moncoulon et al. 2014 ; David Moncoulon 2014 ; David

Moncoulon et Quantin 2013) ou avec plus de détails dans la thèse de (Mao 2019). La CCR a développé un modèle déterministe pour estimer le coût et un modèle probabiliste pour estimer son exposition. Notre démarche est similaire au modèle déterministe développé, cependant nous utilisons ni les mêmes données ni les mêmes méthodes. Le modèle développé par CCR se base sur un modèle d'aléa, un modèle de vulnérabilité et un modèle de dommage.

- Le modèle d'aléa permet de simuler les écoulements d'eau en intégrant les différents processus hydrologiques conduisant à une crue. Il intègre la transformation de la pluie en débit (modèle pluie débit), l'écoulement, l'infiltration et l'hydrologie des cours d'eau.
- Le modèle de vulnérabilité est construit à partir des données de police d'assurance du marché et fournit pour chaque bien la localisation et les caractéristiques du risque.
- Le modèle de dommage combine ces deux informations pour estimer les dommages au niveau de chaque bien assuré. Il prend en compte une probabilité de sinistre, un taux de destruction, la probabilité de reconnaissance CatNat et la valeur du bien. Pour déterminer le taux de destruction et la fréquence de sinistres des distributions statistiques sont calibrées sur la base de l'intensité de l'aléa.

Notre approche est différente car nous ne modélisons pas l'aléa mais nous déterminons, via la cartographie décrite en 3.3.1, les zones qui sont les plus susceptibles d'être impactées. Nous mesurons la vulnérabilité en calculant le nombre de biens par zone et les dommages sont estimés à partir de ces données. Pour cela nous utilisons les méthodes décrites ci-dessous et notre démarche est similaire dans l'idée à celle de CCR, nous essayons d'ajuster des distributions statistiques, mais avec des données et des échelles d'analyses différentes.

6.1.2 Mode opératoire

Notre objectif est d'estimer le coût d'un événement rapidement après son occurrence. Pour estimer le coût des événements, France Assureurs fait une enquête statistique. Pour cela elle envoie un questionnaire directement aux sociétés d'assurance. Cette méthode donne des résultats très fiables car elle prend directement en compte le nombre de sinistres rapportés. Cependant elle prend du temps et les résultats ne sont pas connus avant une ou deux semaines. Nous nous inscrivons dans un temps plus court en estimant un ou deux jours après l'événement.

Cela pose un premier problème qui est de définir l'événement. En effet comme décrit en 3.2.1 notre définition des événements repose sur les reconnaissances en état de catastrophe naturelle. Or ces reconnaissances arrivent bien après, dans le cas d'une procédure accélérée on peut avoir les reconnaissances une ou deux semaines après et seulement pour les communes les plus sévèrement impactées. Dans le cas usuel on observe un délai d'un peu plus de deux mois pour les inondations. Nous ne pouvons donc pas utiliser les reconnaissances et nous nous tournons vers d'autres sources. Dans un premier temps on regarde les communes qui croisent un cours d'eau en vigilances orange et rouge sur le service d'information sur le risque de crues des principaux cours d'eau en France, Vigicrue. Cela nous donne un premier périmètre très susceptible d'être impactée par les débordements. Ensuite nous complétons ce périmètre avec les informations que nous retrouvons dans la presse locale. La définition d'un nouvel événement se base sur une procédure déterminée à l'avance mais peut varier en fonction des événements. Le gestionnaire de risque intervient pour corriger et amender ce périmètre en fonction des informations disponibles et de son expertise. Nous obtenons ainsi une liste de communes, qui constitue, avec la date, notre événement.

Cette définition d'un événement, différente de celle pour notre base historique, engendre une première source d'incertitude car selon les définitions on peut avoir un périmètre différent. Utiliser pour base d'apprentissage des événements définis selon la procédure appliquée pour les événements en temps réel pourrait permettre d'éviter cela. Cependant nous n'avons pas l'historique nécessaire et nous appliquons donc la base reposant sur les arrêtés CatNat.

Nous tirons profit des informations disponibles à la MRN sur les événements passés pour aider à caractériser les événements en cours. Pour cela nous reposons sur deux méthodes, une méthode d'arbre de régression en fonction du comportement de la queue de distribution et une méthode de type sévérité fréquence se basant sur une comparaison d'événements similaires.

6.2 Arbres de régression avec une loi de Pareto généralisée

6.2.1 Méthode générale

Notre première méthode consiste à estimer la distribution statistique du coût de nos événements. Pour cela nous appliquons la théorie des valeurs extrêmes couplée à des arbres de régression. Cette méthode est décrite dans l'article (Farkas, Heranval et al. 2021) et a premièrement été introduite par (Farkas, Lopez et Thomas 2021). Nous présentons ici le cas d'utilisation et nous présentons des résultats sur la consistence de cette méthode dans le chapitre suivant.

Elle se base sur le Peaks over Threshold (PoT) décrit en partie 2.1. Cela consiste à ajuster une distribution de Pareto Généralisée (GPD) aux valeurs en excès d'un certain seuil choisi au préalable. Nous nous plaçons ici dans le cadre d'une régression et nous essayons d'estimer les paramètres de cette GPD en fonction des valeurs de certaines covariables. Ces covariables influençant la distribution de la queue de la distribution. La particularité de notre méthode est d'utiliser, pour faire cette régression, les arbres de régression CART introduits en chapitre 2.3.3.

La régression des quantiles extrêmes est un sujet difficile et l'on peut trouver plusieurs travaux s'y rapportant. Les travaux de (Chernozhukov 2005 ; H. J. Wang, D. Li et He 2012), par exemple, estiment les quantiles extrêmes en assumant une forme de linéarités. D'autres approches modélisent les paramètres de la GPD comme une fonction des covariables, une fonction polynomiale, (Beirlant et Goegebeur 2004), ou des modèles additifs généralisés, (Chavez-Demoulin, Embrechts et Hofert 2016).

L'algorithme CART repose classiquement sur des critères d'erreurs moyennes pour faire une régression moyenne et d'autres fonctions de pertes ont été considérées dans (Chaudhuri et Loh 2002 ; Su, M. Wang et Fan 2004) qui utilise une perte log-vraisemblance (log-likelihood loss). Dans (Loh 2011 ; Loh 2014) on peut trouver une revue des fonctions de perte utilisées avec les arbres de régression. Dans notre méthode nous utilisons une fonction de perte GPD log-vraisemblance pour faire notre régression des valeurs extrêmes.

Pour rappel on cherche une fonction de distribution de la forme suivante pour nos valeurs en excès :

$$\overline{H}_{\sigma_{0u},\gamma_0}(y) = (1 + \gamma_0 \frac{y}{\sigma_{0u}})^{-1/\gamma_0}, y > 0.$$

Dans cette partie nous sommes dans le cadre d'une régression, nous cherchons à estimer l'impact de covariables $\mathbf{X}$ sur une variable réponse Y. En supposant que $\gamma_0(x) > 0, \forall x$ on peut réécrire que :

$$\lim_{t\to+\infty} \frac{\overline{F}(ty|x)}{\overline{F}(y|x)} = y^{-1/\gamma_0(x)}, \forall y > 0,$$

où $\overline{F}(y|x) = \mathbb{P}[Y \geq y|X = x]$ que l'on peut réécrire aussi

$$\lim_{u(x)\to+\infty} \sup_{y>0} |\overline{F}_{u(x)}(y|x) - \overline{H}_{\sigma_{0u(x)},\gamma_{0u(x)}}(y)| = 0,$$

avec $\overline{F}_{u(x)}(y|x) = \mathbb{P}[Y - u(x) > y|Y > u(x), X = x]$.

L'idée de cette méthode est d'appliquer la procédure du CART aux observations $(Y_i - u(X_i), X_i)$ avec $Y_i \geq u(X_i)$, en reposant, pour faire les séparations, sur la GPD log-vraisemblance, c'est à dire :

$$\phi(y, \theta) = -\log(\sigma) - \left(\frac{1}{\gamma} + 1\right) \log\left(1 + \frac{y\gamma}{\sigma}\right),$$

avec

$$\begin{aligned}\theta^*(x) &= arg\max_{\theta \in \Theta} \mathbb{E}[\phi(Y - u(X), \theta) | X = x, Y \geq u(x)] \\ &= arg\max_{\theta \in \Theta} \mathbb{E}[\phi(Y - u(X), \theta) \mathbf{1}_{Y \geq u(X)} | X = x]],\end{aligned}$$

où $\theta = (\sigma, \gamma)^T$. Par la convergence vers la GPD, $\theta^*(x)$ devait être proche de $\theta_0(x) = (\sigma_0(x), \gamma_0(x))^T$ pour u(x) assez grand. A la fin on obtient donc dans chaque feuille de l'arbre des classes d'événements qui ont un comportement de queue de distribution homogène, c'est à dire avec les mêmes $(\sigma_0(x), \gamma_0(x)$, commun dans chaque même feuille.

Des résultats théoriques sont décrits dans le chapitre 7 et les preuves sont reproduites en annexe A. On peut prouver que cette procédure est bien consistante. Dans un premier temps on obtient des résultats pour la consistance d'un arbre avec K feuilles, puis on peut montrer la consistance de la procédure d'élagage de l'arbre. Ces résultats reposent sur les inégalités de concentrations.

En comparaison avec d'autres méthodes de régressions de valeurs extrêmes notre procédure à l'avantage d'avoir la possibilité d'introduire des discontinuités dans la fonction de régression. Au contraire des approches paramétriques qui supposent la linéarité telle (Beirlant et Goegebeur 2003). Les méthodes les plus flexibles comme (Beirlant et Goegebeur 2004), reposent sur un lissage des données ce qui suppose une continuité dans les covariables. (Chavez-Demoulin, Embrechts et Hofert 2016) propose une approche semi-paramétrique pour séparer les variables continues des variables discrètes.

6.2.2 Application

Dans cette partie nous appliquons la méthode du CART GPD aux événements inondations pour mieux comprendre leurs comportements extrêmes et ainsi estimer leurs coûts. Nous observons dans les événements une certaine hétérogénéité en fonction de certaines caractéristiques telles que la région météorologique. Avec cette méthode nous allons faire des classes qui sont homogènes dans leurs comportements extrêmes ce qui d'un point de vue opérationnel, est très précieux.

On utilise la BD SILECC décrite en 3.2.3 enrichie de plusieurs covariables. En effet pour chaque événement on a rapporté, en plus du coût, la région météorologique (quand il y en a plusieurs pour un même événement on sépare l'événement par région), la saison, le nombre d'hydro-écorégions affectées, le nombre de logements individuels en zone de risque inondation et le nombre de professionnels en zone de risque inondation. La zone de risque inondation se base sur la carte de ruissellement faite, en 3.3.1. On calcule le nombre de particuliers et professionnels dans les zones de risques moyennes et fortes. Ces informations sont déterministes ou caractéristiques de l'événement et donc disponibles rapidement après son occurrence.

Nous avons 2 400 événements de 1999 à 2019. La variable d'intérêt, le coût des événements est très volatile. Elle va de 0 à 380 487 000 euros avec une variance empirique de $3.16e + 14$. La figure 6.1 montre la moyenne du coût des événements parmi les 10% les plus coûteux dans chaque région météorologique. Cela illustre bien l'hétérogénéité dans la sévérité des événements les plus importants. Nous pouvons aussi noter que les 10 événements les plus coûteux représentent 49%

du coût total de la base et les top 100 représente 87%.

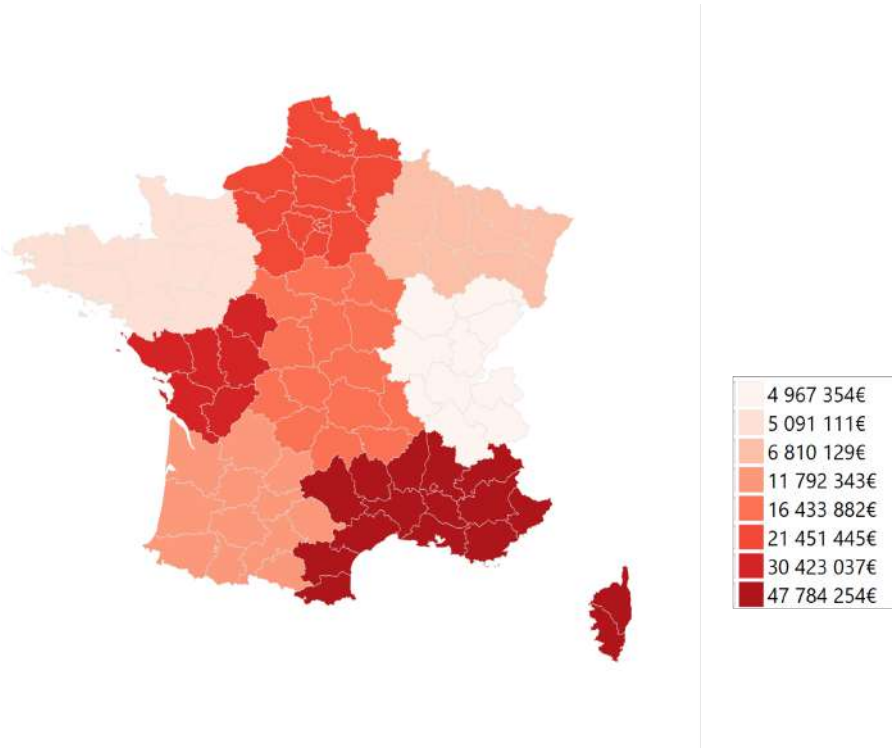


FIGURE 6.1 – Cartographie des coûts des événements inondations de 1999 à 2019. Pour chaque région météorologique, on montre la moyenne du coût des événements faisant partie des 10% les plus chers. Le rouge clair suggère un coût faible alors que le rouge foncé un coût plus important.

Nous cherchons à comprendre l'hétérogénéité du coût des événements inondations les plus sévères, les événements extrêmes. Comme décrit précédemment pour définir un événement extrême on choisit un seuil u qui correspond à un compromis biais variance. Nous choisissons ici un seuil $u = 100\ 000$ selon des considérations pratiques. Ce choix est validé par une analyse de la sensibilité, on a fait varier le seuil comme montré en figure 6.3. On a 820 événements au-dessus avec notre seuil $u = 100\ 000$.

La régression CART GPD est ensuite faite sur la sous-base des événements dont le coût est supérieur à 100 000 euros. Les variables de la base de données et leurs caractéristiques sont résumées en table 6.1 et 6.2. On peut encore remarquer la volatilité de la variable coût.

L'arbre obtenu avec cette méthode est disponible en figure 6.2. Les diagrammes quantile-quantile sont aussi disponibles en 6.4. Notre arbre possède 6 feuilles, avec des séparations selon 3 critères, le nombre de logements individuels en zone de risque inondations, le nombre de professionnels en zone de risque inondations et le nombre d'hydro-écorégions affectées. Cela semble cohérent car les deux premières covariables représentent l'exposition aux inondations mais aussi la densité de population de la zone touchée, la troisième covariable rend compte de l'étendue de l'événement. Le cas le plus extrême correspond à la feuille la plus à droite, avec un paramètre de forme de 0.92 et contenant 7% des événements. Elle correspond à une part importante de logement individuel touché et à une zone étendue. En table 6.3 on a pour chaque feuille calculée

les médianes et moyennes empiriques et théoriques. On peut rappeler que dans le cas d'une GPD de paramètre (γ, σ) la médiane théorique est égale à $\sigma(2^{\gamma} - 1)/\gamma$ et la moyenne théorique à $\sigma(1 - \gamma)$ pour $\gamma < 1$ et à ∞ pour $\gamma \geq 1$. Pour chaque feuille la médiane est bien inférieure à la moyenne suggérant que l'on est bien face à des événements extrêmes. Ensuite on observe un très bon ajustement avec des valeurs très proches dans toutes les feuilles pour les médianes théoriques et empiriques. Ensuite pour la moyenne les valeurs théoriques et empiriques sont aussi proches, sauf pour les feuilles 4 et 6 qui correspondent aux paramètres de formes les plus importants. Les paramètres semblent donc très bien ajustés à la distribution dans chaque feuille et la classification semble aussi pertinente.

Variable	Min	1erQ	Médiane	Moyenne	3èmeQ	Max
Coût	100 093	199 287	477 943	6 066 835	1 941 047	380 487 161
Nombre d'hydro-écorégions affectées	1	1	2	4	4	35
Nombre de logements individuels en zone de risque inondations	0	5 874	20 692	92 477	71 094	4 097 075
Nombre de professionnels en zone de risque inondations	0	2 230	8 163	44 830	26 321	2 050 165

TABLEAU 6.1 – Liste et résumé des statistiques descriptives des variables quantitatives utilisées.

Variable	Catégories	Nombre d'observations
Régions météorologiques	Centre	60
	Nord-Ouest	85
	Nord	135
	Nord-Est	87
	Est	96
	Sud	209
	Ouest	30
	Sud-Ouest	121
Saisons	Printemps	272
	Eté	279
	Automne	187
	Hiver	85

TABLEAU 6.2 – Liste et nombre d'observations des variables qualitatives utilisées.

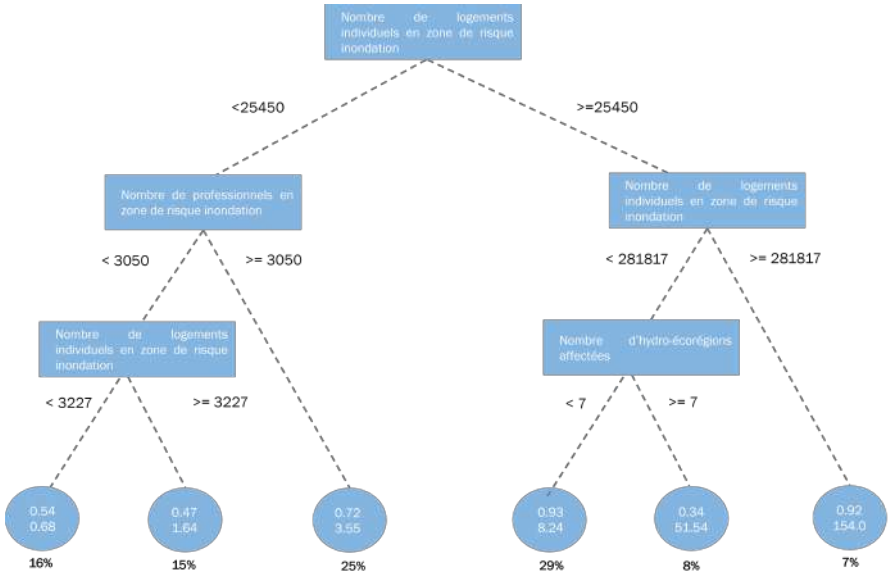
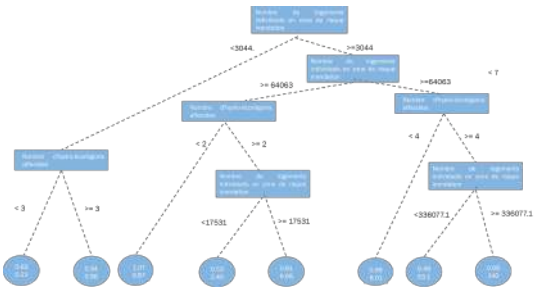
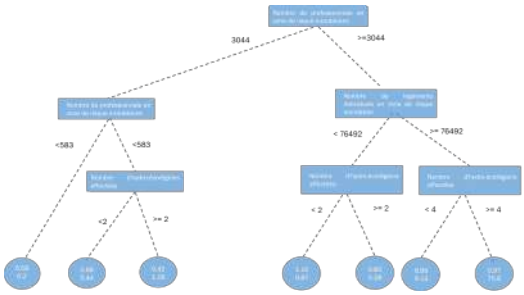


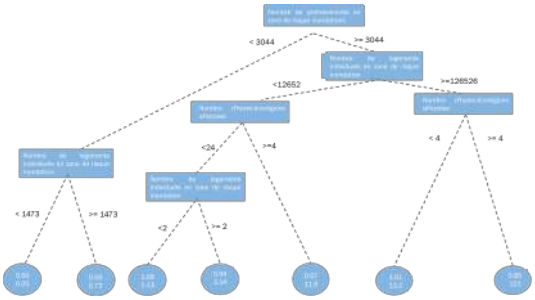
FIGURE 6.2 – Arbre de régression GPD obtenue pour les événements inondations. Pour chaque feuille on indique le paramètre de forme γ (première ligne), le paramètre d'échelle σ à 10^{-5} (deuxième ligne). Les pourcentages d'observations dans chaque feuille sont aussi présentés.



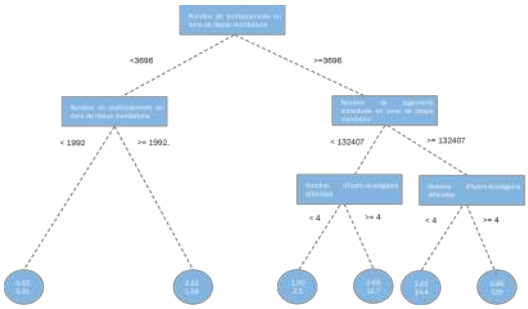
a) 2 000 plus grandes observations



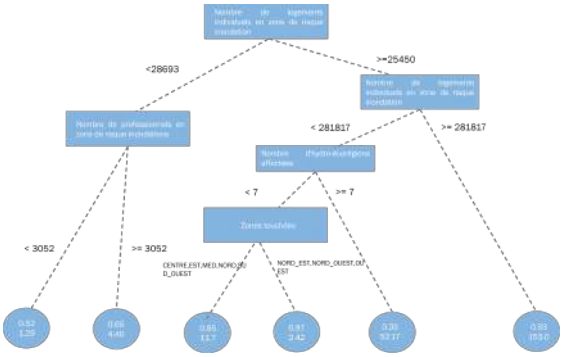
b) 1 750 plus grandes observations



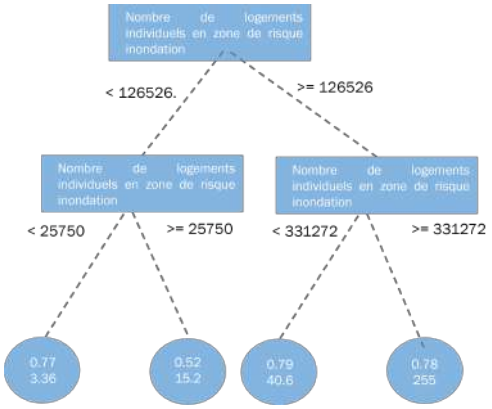
c) 1 500 plus grandes observations



d) 1 250 plus grandes observations



e) 750 plus grandes observations



f) 500 plus grandes observations

FIGURE 6.3 – Arbres obtenus pour le CART GPD en utilisant les plus grandes observations, on fait varier le nombre de 2 000 à 500 avec un pas 250 pour illustrer la sensibilité. Pour chaque feuille on donne une estimation du γ.

Feuille	Paramètre de forme	Médiane empirique	Médiane théorique	Moyenne empirique	Moyenne théorique
1	0.54	161 694	157 697	239 923	249 456
2	0.47	226 196	234 764	399 274	410 387
3	0.72	455 663	419 978	1 439 087	1 390 099
4	0.93	950 181	902 387	4 144 876	11 877 446
5	0.34	4 215 647	4 140 879	7 982 445	8 009 145
6	0.92	15 555 487	15 090 137	52 203 995	281 103 859

TABLEAU 6.3 – Médiane et moyenne empirique et théorique pour chaque feuille de l'arbre

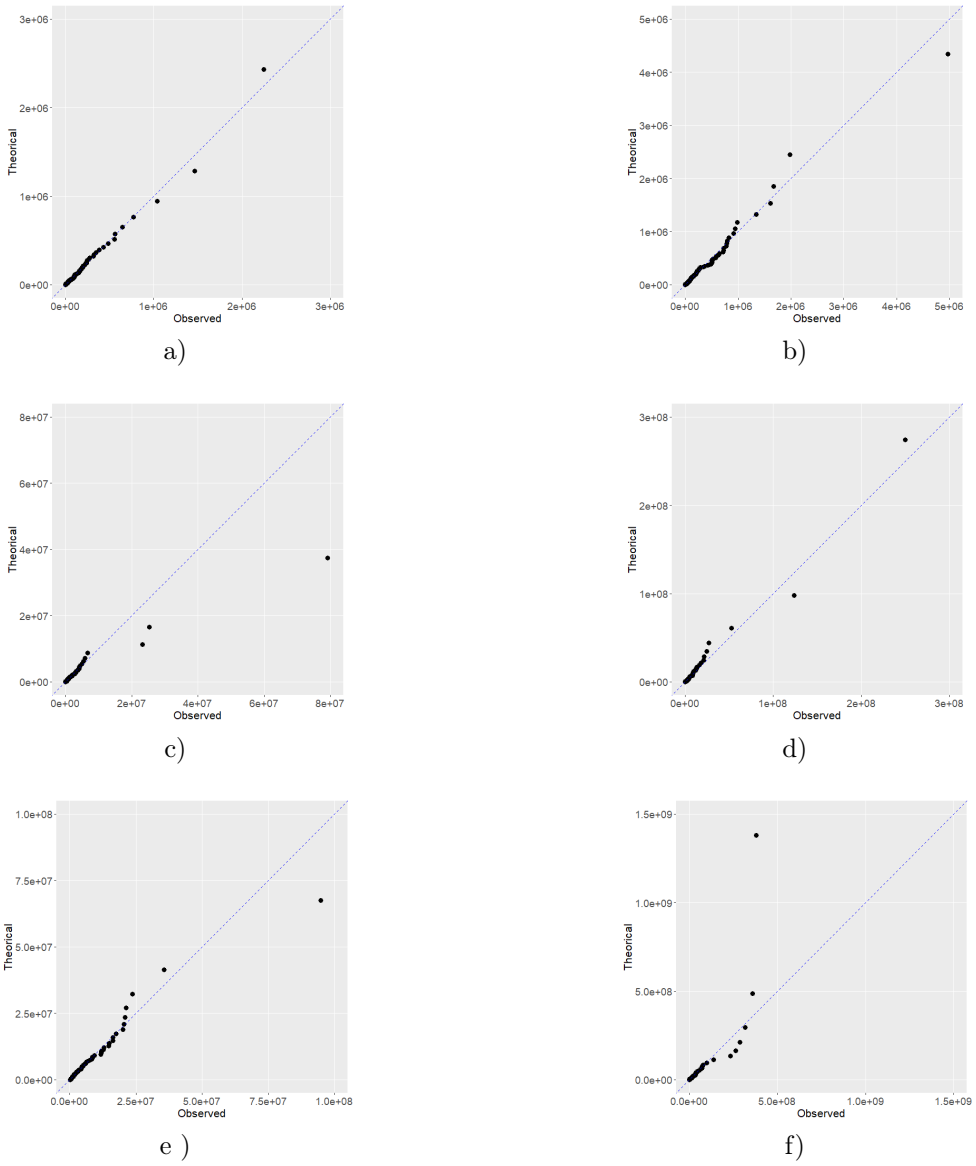


FIGURE 6.4 – Diagrammes quantile-quantile pour chaque feuille de l'arbre

6.3 Application de la théorie de la crédibilité

Nous obtenons donc avec cette méthode une classe avec une distribution pour chaque événement. Cependant cela ne nous permet pas de donner une estimation fixe pour chaque événement, ce qui est un des attendus pour notre méthode. Pour cela nous appliquons la théorie de la crédibilité bayésienne à l'échelle de la commune. Dans notre cas elle présente un intérêt particulier car elle permet de prendre en compte la distribution GPD obtenue par la classification CART. L'association de ces deux méthodes constitue le caractère novateur de notre approche. La classification CART GPD est un préalable à cette étude. Nous commençons par faire la classification CART GPD pour trouver les classes et distributions et après nous appliquons cette méthode.

Nous cherchons $Y_{i,j}$ le coût total pour la commune i sachant qu'elle fait partie d'un événement de type j, j étant la classe du CART.

Nous connaissons l'historique $(Y_{i,j,1}, ...Y_{i,j,k}, ...Y_{i,j,n})$ des événements passés pour une commune i d'un type j. Cet historique correspond à la base d'apprentissage de l'arbre, mais en se plaçant à l'échelle de la commune.

Or grâce à la classification CART GPD nous connaissons la distribution du coût des événements de type j. On suppose ensuite que :

$$Y_{i,j} \sim \mathbb{GPD}(\gamma_j, s_j),$$

avec $s_j = p_i\sigma_j$, où p_i est la proportion des primes de la commune par rapport au total des primes de l'événement et σ_j et γ_j les paramètres de la GPD de l'événement. On suppose donc que le coût est distribué uniformément dans un événement selon la répartition des primes. Les primes provenant aussi de la BD SILECC, cette hypothèse paraît acceptable, même si cela apporte une source d'erreur potentielle supplémentaire.

On se place ensuite dans un contexte de crédibilité bayésienne. On cherche $\mathbb{E}[Y_{i,j,n+1} \mid Y_{i,j,1}, ...Y_{i,j,n}]$ qui se base aussi sur le calcul de $\mathbb{E}[Y_{i,j} \mid \theta_{i,j}]$ avec $\theta_{i,j}$ le profil de risque de notre commune i pour un événement de type j que nous ne connaissons pas.

On suppose que :

$$Y_{i,j} \mid \theta_{i,j} \sim \mathbb{EXP}(\theta_{i,j}),$$

en supposant pour loi a priori une loi gamma $\theta_{i,j} \sim \Gamma(r_j, \lambda_j)$, la loi marginale de $Y_{i,j}$ est une loi de Pareto généralisée, plus précisément

$$Y_{i,j} \sim \mathbb{GPD}(\frac{1}{r_j}, \frac{\lambda_j}{r_j}).$$

Grâce à la $\mathbb{GPD}$ de la classification nous pouvons ainsi trouver les paramètres et renseigner le profil de risque.

Les paramètres et l'historique étant connus nous pouvons ensuite appliquer la crédibilité bayésienne pour estimer le coût sur une commune.

Pour cela on va approximer le coût de la commune i pour un événement k de type j à partir du montant de ses événements passées de type j on cherche : $\mathbb{E}[Y_{i,j,n+1} \mid Y_{i,j,1}, ...Y_{i,j,n}]$.

Empiriquement nous pouvons vérifier que les coûts des communes par classe j sont peu corrélés entre eux, utiliser les coûts passés par feuille semble être pertinent avec la classification, comme illustré en 6.4.

On peut ensuite déterminer la loi a posteriori de $\theta_{i,j}$.

Feuille	1	2	3	4	5	6
1	X	0.33	-0.07	-0.02	0.02	0.12
2	0.33	X	0.04	0.24	0.01	0.44
3	-0.07	0.04	X	0.03	0.26	0.03
4	-0.02	0.24	0.03	X	0.03	0.14
5	0.02	0.01	0.26	0.03	X	0.03
6	0.12	0.44	0.03	0.14	0.03	X

TABLEAU 6.4 – Coefficient de corrélation de Pearson pour le coût empirique des communes dans chaque feuille. On compare les moyennes des coûts dans les mêmes communes mais dans des feuilles différentes

Par les résultats de l'introduction on a :

$$\begin{aligned} f_{\theta_i|Y_{i,1}=y_1,...Y_{i,n_i}=y_{n_i}}(t) &= \frac{g_t(y_1)....g_t(y_{n_i}f_{\theta_i(t)})}{\int_s g_s(y_1)....g_s(y_{n_i}f_{\theta_i(s)})\mathrm{d}s} \\ &= C(Y_{i,j,1}...Y_{i,j,n})\prod_{k=1}^{n} te^{-ty_{i,k}}\frac{\lambda_j^{r_j}}{\Gamma(r_j)}t^{r_j-1}e^{-\lambda_j t} \\ &= C(Y_{i,j,1}...Y_{i,j,n})t^{n+r_j-1}e^{-(\sum_{k=1}^{n} y_{i,k}+\lambda_j)t}. \end{aligned}$$

On peut reconnaître que

$$\theta_{i,j} \mid Y_{i,j,1}...Y_{i,j,n} \sim \Gamma(n+r_j, \sum_{k=1}^{n} y_{i,k}+\lambda_j),$$

on cherche

$$\mathbb{E}[Y_{i,j,n+1} \mid Y_{i,j,1}, ...Y_{i,j,n}] = \mathbb{E}[\mathbb{E}[Y_{i,j} \mid \theta_{i,j}] \mid Y_{i,j,1}, ...Y_{i,j,n}].$$

Or

$$Y_{i,j} \mid \theta_{i,j} \sim \mathbb{EXP}(\theta_{i,j}),$$

donc

$$\mathbb{E}[Y_{i,j,n+1} \mid Y_{i,j,1}, ...Y_{i,j,n}] = \mathbb{E}[\frac{1}{\theta_{i,j}} \mid Y_{i,j,1}, ...Y_{,ij,n}],$$

on a $\theta_{i,j} \mid Y_{i,j,1}...Y_{i,j,n} \sim \Gamma(n+r_j, \sum_{k=1}^{n} y_{i,k}+\lambda_j)$, on cherche ici l'espérance d'une loi inverse-gamma.

On peut donc en déduire que :

$$\mathbb{E}[Y_{i,j,n+1} \mid Y_{i,j,1}, ...Y_{i,j,n}] = \frac{\sum_{k=1}^{n} y_{i,k}+\lambda_j}{n+r_j-1}.$$

En remplaçant avec les paramètres connus de départ on trouve :

$$\mathbb{E}[Y_{i,j,n+1} \mid Y_{i,j,1}, ...Y_{i,j,n}] = \frac{\sum_{k=1}^{n} y_{i,k}+(\frac{pi\sigma_j}{\gamma_j})}{n+\frac{1}{\gamma_j}-1}.$$

Sur la base de test de la section 6.5 nous observons que près de la moitié des communes que nous essayons d'estimer n'ont pas d'historique de sinistres ($n=0$) et pour seulement 20%

des communes nous avons plus d'un événement passé ($n > 1$). Cette méthode est donc particulièrement indiquée car l'expérience des événements passés n'est pas suffisante pour donner une estimation fiable et l'information ramenée par les classes du CART GPD permet d'enrichir l'estimation.

En synthèse le jour J nous avons un événement qui correspond à une liste de communes impactées. Nous calculons pour cet événement les variables d'entrées, qui seront utilisées pour la classification CART GPD. Ensuite pour chaque commune nous pourrons estimer un coût total dépendant de la classe obtenue. Le coût total correspondra à la somme des coûts des communes. Le coût total dépend donc de la liste de communes impactées, de l'historique du coût sur ces communes mais aussi du type d'événement. C'est très précieux pour notre étude car nous prenons en compte les spécificités locales des communes mais aussi l'événement dans sa globalité. De plus nous tirons pleinement profit de la méthode CART GPD en utilisant sa distribution de sortie comme a priori sur le profil de risque. A la fin nous estimons donc le coût d'un événement en cours par :

$$\hat{C} = \sum_{i=1}^{M} \mathbb{E}[Y_{i,j,n+1} \mid Y_{i,j,1}, ...Y_{i,j,n}].$$

6.4 Estimation fondée sur une approche type fréquence x sévérité

Nous avons aussi développé une autre méthode reposant sur des indicateurs à l'échelle de l'événement et un peu plus facile à interpréter. Cette méthode nous permet aussi d'avoir un élément de comparaison. C'est une approche de type fréquence sévérité, classique en actuariat. Ici on estime la fréquence et la sévérité d'un événement par comparaison avec des événements passés et l'on multiplie cette fréquence par l'exposition observée.

Pour mesurer l'exposition nous utilisons la carte de ruissellement précédemment décrite. Pour mesurer la fréquence passée nous calculons un taux de sinistralité en divisant le nombre de sinistres par le nombre de biens. La sévérité est mesurée par le coût moyen d'un sinistre à l'échelle communale. En multipliant ces trois quantités nous obtenons un coût. Pour introduire une variabilité et enlever une part de sinistralité résiduelle nous déclinons les mesures de sévérités et de fréquences sur des échantillons différents. Nous obtenons à la fin une table d'événements historiques de grandes ampleurs, possédant plusieurs mesures rendant compte de la sévérité et de la fréquence des sinistres à l'échelle des communes pour un événement.

Le jour J nous avons une liste de communes et nous pouvons calculer pour chacune une exposition. Ensuite nous allons regarder dans la table de référence quel événement passé est le plus semblable et quel échantillon permettait d'avoir les meilleurs résultats. Nous prendrons ensuite les mesures de sévérités et de fréquences correspondantes que nous appliquerons à l'exposition calculée. Pour augmenter la précision cette analyse est faite pour chaque zone de risques inondations de la cartographie 3.8. La comparaison avec les événements passés est faite directement par le gestionnaire de risque via des critères divers, comme par exemple la zone touchée, la saison, le type de crue, les communes importantes touchées. Cela laisse une part de contrôle qui est appréciée, le coût est d'une certaine façon construit « à la main ». Cependant cela rend cette méthode vulnérable à différents biais et il est aussi difficile de l'évaluer au préalable.

Détaillons cette méthode, dans un premier temps on cherche à mesurer l'exposition pour chaque zone de la cartographie inondations.

Exposition Le jour J nous avons un événement qui touche N communes. Nous allons calculer l'exposition pour chaque zone de risques aux inondations i et pour chaque commune j. Pour rappel nous avions 5 zones définissant un facteur de risque. Dans chaque zone nous pouvons compter le nombre de professionnels, que nous appellerons N_{pro_i}. Nous n'avons pas le détail des logements individuels par zone donc nous le calculons en utilisant la zone urbaine de l'INSEE. On peut croiser la zone urbaine et la cartographie pour déterminer la proportion de zone urbaine dans chaque zone de risques, appelons la P_{ZU_i}. Nous pouvons ainsi estimer le nombre de logements individuels par zone en multipliant le nombre de logements par cette proportion, en l'appelant N_{ind}, on a :

$$N_{ind_i} = P_{ZU_i} N_{ind}.$$

On peut ainsi calculer le nombre de biens exposé par zone de risques dans une commune j :

$$N_{tot_{i_j}} = N_{pro_{i_j}} + N_{ind_{i_j}} = N_{pro_{i_j}} + P_{ZU_{i_j}} N_{ind_j}.$$

Ensuite en sommant sur toutes les communes N on obtient une mesure du nombre de biens exposés à l'échelle de l'événement pour chaque zone i :

$$N_{tot_i} = \sum_{j=1}^{N} N_{tot_{i_j}}.$$

Sévérité Pour mesurer la sévérité nous regardons les événements passés de la BD SILECC. La liste de communes touchées n'est donc plus la même, supposons que nous avons N' communes pour l'événement étudié. Nous allons ici encore mesurer la sévérité pour chaque zone de risque i et dans chaque commune disons j'. Nous allons calculer la moyenne, arithmétique simple, du coût des sinistres dans chaque zone et chaque commune :

$$M_{i'_j} = \overline{C_{loc_{i_{j'}}}},$$

avec $C_{loc_{i_{j'}}}$ le coût d'un sinistre rapporté et localisé dans la BD SILECC pour la commune j' et le facteur de risque i. Pour connaître la zone de risque nous avons dû au préalable géo-localiser les sinistres pour pouvoir faire le croisement. Cela entraîne des pertes, en effet les adresses des sinistres ne sont pas systématiquement renseignées et la précision est parfois insuffisante pour localiser le sinistre. Nous utilisons donc ici un sous-échantillon des sinistres reçus que nous avons réussis à localiser, C_{loc}. Ensuite pour former une mesure de la sévérité à l'échelle de l'événement par zone de risque nous prenons la moyenne des coûts moyens.

$$S_i = \overline{M_{i_{j'}}} = \frac{1}{N'} \sum_{j'=1}^{N'} M_{i_{j'}}.$$

Fréquence Pour mesurer la fréquence nous restons sur le même périmètre que pour la sévérité, N' communes et pour chaque facteur de risque i. Nous cherchons à calculer pour chaque commune un taux de sinistralité. Le nombre de sinistres est donc déterminant et par conséquent nous prenons aussi en compte les sinistres que nous n'avons pas réussis à localiser. En effet ces sinistres, ne sont pas localisés assez précisément pour les rapprocher de la carte mais pour la plupart on a la commune de survenance. Nous pouvons donc comme pour la population les affecter à une zone de risque grâce à la zone urbaine, de la même façon, avec N_{nloc}, le nombre de sinistres non

localisés :

$$N_{nloc_i} = P_{ZU_i} N_{nloc}.$$

On peut ainsi calculer le nombre de sinistres, avec $N_{loc_{i_{j'}}}$ le nombre de sinistres localisés en zone de risque i pour une commune j' :

$$N_{sin_{i_{j'}}} = N_{loc_{i_{j'}}} + N_{nloc_{i_{j'}}}.$$

On peut ensuite calculer un taux de sinistralité T en divisant par le nombre de biens exposés, calculé comme décrit précédemment, dans chaque commune :

$$T_{i_{j'}} = \frac{N_{sin_{i_{j'}}}}{N_{tot_{i_{j'}}}}.$$

Pour en faire une mesure de la fréquence à l'échelle de l'événement on prend la moyenne géométrique :

$$F_i = \overline{T_{i_{j'}}} = \prod_{j'=1}^{N'} T_{i'_j}^{\frac{1}{N'}}.$$

Échantillons différents Nous obtenons donc deux mesures pour chaque zone de risque à l'échelle de l'événement, une de la sévérité, une de la fréquence. Pour faire varier les coûts on peut calculer S_i et F_i sur des échantillons différents. En effet pour enlever la sinistralité « résiduelle » et se concentrer sur les communes d'intérêts nous appliquons un filtre. Nous identifions les communes qui concentrent, 90%, 92.5%, 95%, 97.5%, 99% et 100% de la sinistralité de l'événement et nous calculons nos indicateurs sur ces sous-échantillons. Cela permet d'enlever un certain nombre de communes qui ont peu de sinistres et qui peuvent affaiblir la robustesse de nos indicateurs car ils sont très dépendants du nombre de sinistres.

On cherche donc

$$N'_{x\%} \text{ tel que } \sum_{j'=1}^{N'_{x\%}} CT_{j'} = 0.xCT,$$

avec $CT_{j'}$ représentant le coût total des sinistres pour la commune j' et CT le coût total de l'événement. On classe par ordre décroissant les communes en fonction du coût pour privilégier les communes qui contribuent le plus. On obtient ainsi plusieurs sous-échantillons de communes pour chaque événement et c'est sur ces sous-échantillons que l'on calcule nos indicateurs. On obtient donc 6 échantillons différents, $N'_{90\%}$, $N'_{925\%}$, $N'_{95\%}$, $N'_{975\%}$, $N'_{99\%}$, $N'_{100\%}$ qui vont nous donner des indicateurs différents qui vont nous permettre de choisir quelle méthode est la plus adaptée en fonction de la nature de l'événement.

Table de référence Nous calculons ces indicateurs pas pour tous les événements de la base SILECC mais seulement certains événements de grandes ampleurs. L'idée étant d'avoir un nombre restreint d'événement le jour J permettant de rapidement voir quel événement est similaire à celui en cours. On utilise les événements de grandes ampleurs rapportés par CCR dans son espace professionnels, cela nous permettra d'avoir une base de test comme décrit dans la section suivante. Nous faisons correspondre ces 56 événements à 56 événements de notre base SILECC. Les événements sont semblables mais avec un périmètre différent. En effet nous n'avons pas la même méthode que CCR et donc nous obtenons des listes de communes différentes pour des « mêmes » événements. Pour ces 56 événements en se basant sur les communes SILECC

nous construisons les indicateurs de sévérité et de fréquence selon les 6 sous-échantillons. Pour chacun de ces événements nous allons ensuite calculer le coût avec chaque sous-échantillon. Nous commençons par calculer l'exposition selon les communes de l'événement CCR, identique pour tous les sous-échantillons. Nous prenons la méthode décrite au début de cette section en utilisant les communes rapportées par CCR, on obtient un N_{tot_i}. On calcule ensuite en prenant les communes rapportées dans SILECC, $S_{i_{90\%}}$, ...$S_{i_{100\%}}$ et $F_{i_{90\%}}$, ...$F_{i_{100\%}}$. Finalement, le coût pour un sous-échantillon et dans une zone de risque est :

$$C_{i_{x\%}} = N_{tot_i} \times S_{i_{x\%}} \times F_{i_{x\%}}.$$

et l'on peut ensuite sommer sur toutes les zones de risques pour avoir le coût total par échantillon :

$$C_{x\%} = \sum_{i=1}^{5} C_{i_{x\%}}.$$

Le jour J on va avoir un événement, avec N communes. On va calculer N_{tot_i}, le nombre de biens exposés sur ces N communes. Ensuite on va regarder à quel événement de la base de références cet événement se rapporte. Une fois que l'on a trouvé l'événement le plus proche selon les caractéristiques choisies on regarde quel échantillon x permet d'avoir un coût $C_{x\%}$ le plus proche du coût réel de cet événement. Ces événements étant tous passés on peut avoir un coût réel rapporté pour tout le marché que nous avons préalablement actualisé selon l'indice de la FFB du coût de la construction (comme pour les coûts des sinistres). On va ensuite prendre les indicateurs $S_{i_{x\%}}$ et $F_{i_{x\%}}$ se rapportant à cet événement, c'est à dire les indicateurs qui ont permis d'obtenir $C_{x\%}$ dans la base de référence, et les appliquer à l'événement en cours. De telle sorte que, si l'on appelle S_{ref} et F_{ref} les indicateurs de cet événement de référence, le plus proche de l'événement et avec l'échantillon permettant le mieux possible de retrouver le coût total :

$$C = \sum_{i=1}^{5} N_{tot_i} \times S_{ref_i} \times F_{ref_i}.$$

6.5 Discussion des résultats

Pour évaluer et comparer nos méthodes nous les avons utilisées pour estimer des événements. On utilise la base des événements rapportés par CCR. Nous avons pour 56 événements les communes impactées et le coût pour tout le marché. Ces événements sont déjà présents dans notre base mais avec des communes différentes. En effet il est difficile de définir un événement et plusieurs méthodes peuvent aboutir à des événements différents. Ce sont des événements de grandes ampleurs donc avec des coûts conséquents, nous trouverons en table 6.5 un résumé de la variable d'intérêt. Nous n'avons retenu que les événements postérieurs à 2008 pour s'assurer que notre base soit la plus représentative possible. Ici contrairement à la sécheresse nous travaillons sur les coûts à l'échelle de la commune et sur des événements locaux, nous sommes d'autant plus sensibles à la répartition des sinistres et des portefeuilles.

Variable	Min	1erQ	Médiane	Moyenne	3èmeQ	Max
Coût	10 710 000	16 110 000	35 680 000	116 900 000	98 270 000	1 056 000 000

Tableau 6.5 – Résumé du coût des événements étudiés

Nous avons donc une liste de communes sur laquelle nous pouvons calculer nos paramètres d'entrées pour la classification et une liste d'événements de référence pour les indicateurs de la deuxième méthode.

Nous avons calculé pour chaque méthode l'erreur absolue moyenne (MAE).

MODELE	MAE
$C_{90\%}$	162 462 546
$C_{92.6\%}$	128 057 050
$C_{95\%}$	102 094 036
$C_{97.5\%}$	74 005 296
$C_{99\%}$	56 513 672
$C_{100\%}$	88 538 340
Méthode crédibilité GPD	80 996 208

Tableau 6.6 – Comparaison de la MAE des différentes méthodes

On peut déjà remarquer que les MAE sont assez importantes. Ceci s'explique car les méthodes ont du mal à estimer l'événement à 1 milliard et cela augmente fortement la moyenne d'erreur. Nous n'avons que 56 événements ce qui rend la MAE très sensible aux variations individuelles, surtout importantes. Ensuite on peut voir que la méthode se basant sur la crédibilité donne des résultats comparables aux autres méthodes. Les meilleures méthodes sont $C_{99\%}$ et $C_{97.5\%}$, ce qui conforte l'idée que les petits sinistres peuvent nuire à la qualité des indicateurs considérés. On peut aussi voir que les prédictions les plus éloignées le sont pour toutes les méthodes, cela pourrait donc être lié à un problème de représentativité globale de notre base. En effet notre base peut manquer certaines spécificités ce qui peut à l'échelle d'un événement impacter grandement la qualité des prédictions.

Ici nous comparons la MAE de chaque méthode mais dans l'utilisation prévue, nous utiliserons que la méthode qui correspond le mieux. C'est un avantage car cela laisse une part d'interprétation et permet de profiter de l'expertise métier de la MRN sur les événements en cours. Cependant cela limite aussi car si l'événement diffère trop de la base de référence alors nous ne pourrons pas l'estimer. Dans ce cas la méthode se basant sur le CART GPD est plus adaptée.

Nous obtenons des résultats encourageants avec les deux méthodes, qui seront utiles à la fédération dans son processus de gestion. C'est en effet dans une logique d'aide à la décision que ces méthodes ont été créées.

La principale incertitude de ces méthodes et une des difficultés reste de déterminer la liste des communes impactées le jour J. Ici toute l'expertise métier de la MRN est précieuse car tout repose sur cette liste.

Enfin ces méthodes peuvent sûrement être améliorées en utilisant des informations renseignant l'intensité de l'aléa, comme des variables météorologiques. En effet ici nous n'utilisons que des variables rapportant l'exposition et l'ampleur des événements. On ne prend pas en compte l'intensité de l'événement. C'est une information qui peut être difficile à utiliser à l'échelle de l'événement mais c'est la principale perspective pour améliorer ces estimations.

Chapitre 7

Arbres de régression avec une loi de Pareto généralisée

Dans ce chapitre nous étudions des résultats théoriques de la procédure d'arbre de régression avec une loi de Pareto généralisée (CART GPD) utilisée dans la partie précédente. Nous montrons que cette procédure est consistante. On commence par introduire les notations et hypothèses faites avant de montrer la consistance d'un arbre fixé à K feuilles, en séparant la partie stochastique de l'erreur et la partie de mauvaise spécification causée par l'approximation de la GPD. On étudie ensuite la consistance de l'élagage de l'arbre. Ce chapitre se base sur les résultats et preuves présents dans (Farkas, Heranval et al. 2021).

7.1 Notations

Dans l'approche PoT, on ne considère que les observations telles que $Y_i \geq u(X_i)$. Ici on se restreint au cas simple de $u(x) = u$. Nos résultats s'entendent cependant facilement au cas, $u(x) = \sum_{j=1}^{m} u_j \mathbf{1}_{x \in \mathbf{X}_j}$, avec $(\mathbf{X}_j)_{1 \leq j \leq m}$ qui sont des divisions de l'espace des covariables. Les résultats que nous fournissons tiennent uniformément pour $u \in [u_{min}\,;u_{max}]$, afin de permettre un choix adaptatif de ce paramètre. Les conditions suivantes sur u_{min} et u_{max} doivent être remplis :

Assumption 7.1.1. *Si on a n observations, appelons k_n une suite intermédiaire, telle que $k_n \to +\infty$ et $\frac{k_n}{n} \to 0$ quand $n \to \infty$. Alors $\frac{k_n}{n}$ correspond à la proportion moyenne de Y plus grand que u_{min}, soit $\mathbb{P}(Y \geq u_{min}) = k_n n^{-1}$. De plus on suppose que :*

$$\mathbb{P}(Y \geq u_{max}) = \frac{u_0 k_n}{n},$$

pour une certaine constante $u_0 \leq 1$.

Ici k_n, correspondra au nombre moyen d'observations utilisées pour ajuster le modèle. C'est donc lié à la vitesse de convergence de la procédure. L'hypothèse suivante introduit une contrainte supplémentaire sur la vitesse de k_n et sur l'espace des paramètres.

Assumption 7.1.2. *On suppose que $\Theta = \mathcal{S} \times \Gamma$ avec*
- *$\mathcal{S} = [\sigma_{min}\,;\sigma_n]$, avec $\sigma_n = O(n^{a_1})$, avec $a_1 > 0$,*
- *Γ est un ensemble compact $[\gamma_{min}\,;\gamma_{max}]$ avec $\gamma_{min} > 0$*

De plus on suppose que $k_n = O(n^{a_2})$*, avec* $a_2 > 0$ *et que le nombre de feuilles de l'arbre maximal* K_{max} *satisfait* $K_{max} \leq \kappa k_n$*, avec* $\kappa > 0$*.*

Ensuite introduisons des notations concernant l'arbre. On considère un arbre $T(u)$ avec K feuilles, dénotées, $T_l, l = 1, ..., K$. En introduisant la contribution (normalisée) de la log-vraisemblance de la lième feuille,

$$L_n^l(\theta, u) = \frac{1}{k_n} \sum_{i=1}^{n} \phi(Y_i - u, \theta) \mathbf{1}_{Y_i > u} \mathbf{1}_{X_i \in T_l},$$

avec

$$\hat{\theta}_l(u) = \arg\max_{\theta} L_n^l(\theta, u),$$

l'estimation de la valeur du paramètre dans la feuille T_l. Cette estimation est attendue proche de

$$\theta_l^*(u) = \arg\max_{\theta} L^l(\theta, u),$$

en introduisant $L^l(\theta, u) = k_n n^{-1} \mathbb{E}[L_n^l(\theta, u)]$. On désigne par $T^*(u|T)$ l'arbre avec les mêmes feuilles que T mais avec les paramètres $\theta_l^*(u)$. Ce n'est pas exactement ce que l'on recherche, on voudrait estimer :

$$\theta_{0,l}(u) = (\sigma_0(T_l, u), \gamma_0(T_l)),$$

telle que

$$\lim_{t \to \infty} \sup_{z>0} |\bar{F}_t(z|T_l) - \bar{H}_{\sigma_0(T_l,t), \gamma_0(T_l)(z)}| = 0, \tag{7.1.1}$$

où $\bar{F}_t(z|T_l) = \mathbb{P}(Y - t \geq z | X \in T_l, Y \geq t)$. On désigne $T_0(u|T)$ l'arbre avec les mêmes feuilles que T mais de paramètre $\theta_{0,l}(u)$. Si $\theta = (\theta_l)_{l=1,...,K}$, désigne l'ensemble des paramètres d'un arbre avec K feuilles $(T_l)_{l=1,...,K}$, on désigne $\theta(X)$ la fonction définie par :

$$\theta(X) = \sum_{l=1}^{K} \theta_l \mathbf{1}_{X \in T_l}.$$

On se concentre dans un premier temps, en partie 7.2 sur la différence entre $T(u)$ et $T^*(u|T)$ qui est la partie stochastique de l'erreur. Dans un deuxième temps, en partie 7.3, on étudie la différence entre $T^*(u|T)$ et $T_0(u|T)$ (et finalement la différence entre $\hat{\theta}(x)$ et $\theta_0(x)$). Cela peut être vu comme un terme de mauvaise spécification, dû au fait que les excès ne suivent pas exactement une GPD.

Pour $l = 1, ..., K$, appelons $\nabla_\theta L^l(\theta, u)$ le gradient de $L^l(\theta, u)$,

$$\nabla_\theta L^l(\theta, u) = \mathbb{E}\left[\begin{pmatrix} g_{\theta, l(Y_i - u)} \\ h_{\theta, l(Y_i - u)} \end{pmatrix} \mathbf{1}_{Y_i > u} \mathbf{1}_{X_i \in T_l} \right]$$

avec, pour $z > 0$,

$$g_\theta(z) = \partial_\sigma \phi(z, \theta) = \left(-\frac{1}{\sigma} + (1 + \frac{1}{\gamma}) \frac{\gamma z}{\sigma^2 (1 + \frac{\gamma z}{\sigma})} \right),$$

$$h_\theta(z) = \partial_\gamma \phi(z, \theta) = \left(-\frac{1}{\gamma^2} \log(1 + \frac{\gamma z}{\sigma}) + (1 + \frac{1}{\gamma}) \frac{z}{\sigma + \gamma z}) \right).$$

Pour la partie stochastique il est nécessaire d'ajouter quelques hypothèses. On a besoin d'une

condition de domination sur la classe des dérivées des fonctions $y \to \phi(y-u,\theta)$. Ces dérivées sont uniformément bornées par

$$\Phi(y) = C(1+\log(1+wy))$$

avec C une constante (ne dépendant pas de n) et $w = \frac{\gamma_{max}}{\sigma_{min}}$.

Assumption 7.1.3. *Pour un $\rho_0 > 0$, on suppose que*

$$m_{\rho_0} = \mathbb{E}[\exp(\rho_0 \Phi(Y))] < \infty.$$

Cette hypothèse est automatiquement satisfaite dès que l'hypothèse 7.1.2 tient, comme $\gamma(x) \geq \gamma_{min} \geq 0, \mathbb{E}[Y|1/(\gamma-\epsilon)] < \infty, \epsilon > 0$. Il nous faut aussi une hypothèse sur la régularité du critère L^l

Assumption 7.1.4.

$$M^l_{\theta_1,\theta_2,\theta_3,\theta_4}(u) = \mathbb{E}\left[\begin{pmatrix} \partial_\sigma g_{\theta_1}(Y-U) & \partial_\gamma g_{\theta_2}(Y-U) \\ \partial_\sigma h_{\theta_3}(Y-U) & \partial_\gamma h_{\theta_4}(Y-U) \end{pmatrix} \mathbf{1}_{Y_i>u} | X \in T_l\right]$$

On suppose qu'il existe une constance $\mathbb{C}_1 > 0$ tel quel

$$\inf_{a,b\in\mathbb{R}} \inf_{\theta_1,\theta_2,\theta_3,\theta_4\in\Theta} \inf_{u\in[u_{min};u_{max}]} \inf_{l=1,\dots,K} \left| M^l_{\theta_1,\theta_2,\theta_3,\theta_4}(u) \begin{pmatrix} a \\ b \end{pmatrix} \right| \geq \mathbb{C}_1 \max(|a|,|b|)$$

Cette hypothèse vient naturellement en utilisant un développement de Taylor. L'infimum par rapport à $\theta_1,\theta_2,\theta_3,\theta_4$ peut être restreint à θ_2,θ_3 appartenant à un proche voisinage de θ_1 et pas à l'ensemble Θ.

7.2 Limites de déviation pour notre estimateur

Dans cette section nous étudions la consistance d'un arbre ajusté, $T(u)$, un sous-arbre de l'arbre maximal $T_{max}(u)$ avec K feuilles $(T_l)_{l=1,\dots,K}$. On compare cette arbre ajusté à $T^*(u|T)$, qui est l'arbre basé sur les mêmes subdivisions, mais où dans chaque feuille l, le paramètre est $\theta^*_l(u)$, (au lieu de $\hat{\theta}_l(u)$ dans $T(u)$). La première étape est définir une distance entre les arbres. On définit $||(a,b)||_\infty = max(|a|,|b|)$ et pour deux arbres T et S

$$||T-S||_2 = \left(\int ||T(x)-S(x)||^2_\infty \mathrm{d}\mathbb{P}(x)\right)^{1/2}$$

Le résultat principal de cette section est une limite de déviation pour $||T(u)-T^*(u|T)||_2$, correspondant au théorème suivant :

Theorem 7.2.1. *Sous les hypothèses précédemment énoncées, en prenant $\beta > 0$ tel que $\beta a_2 \geq 10/\rho_0$ (ρ_0 définit en hypothèse 7.1.3) et pour $t \geq c_1 K(\log k_n) k_n^{-1}$ avec $c_1 > 0$*

$$\mathbb{P}\left(\sup_{u_{min}\leq u\leq u_{max}} ||T(u)-T^*(u|T)||^2_2 \geq t\right) \leq 2\left(\exp\left(-\frac{C_1 k_n t}{K\beta^2(\log k_n)^2}\right) + \exp\left(-\frac{C_2 k_n t^{1/2}}{K^{1/2}\beta \log k_n}\right)\right) + \frac{C_3 K}{k_n^{5/2} t^{3/2}},$$

avec C_1, C_2, C_3 des constantes positives.

La preuve de ce théorème peut être trouvée en annexe A. On peut noter que le terme exponentiel du côté droit vient des inégalités de concentrations prouvées par (Einmahl et Mason 2005). Le terme polynomialement décroissant est lié à la log-vraisemblance et au fait que ce soit une quantité non bornée quand on considère son espérance. En sous-produit on obtient le corollaire suivant (par intégration des limites du théorème)

Corollary 7.2.2.

$$\mathbb{E}\left[\sup_{u_{min}\leq u\leq u_{max}} ||T(u) - T^*(u|T)||_2^2\right] \leq C_4 \frac{K\beta^2 (log k_n)^2}{k_n}$$

De ce corollaire, on peut remarquer que la norme$-L^2$ de la partie stochastiques de l'erreur, $\mathbb{E}[\sup_{u_{min}\leq u\leq u_{max}} ||T(u) - T^*(u|T)||_2^2]^{1/2}$, est proportionnelle à $K^{1/2}$, et croît avec la complexité de l'arbre. D'un autre côté, l'erreur décroît presque à la vitesse de $k_n^{1/2}$, ce qui est la vitesse de convergence d'un estimateur standard, utilisée pour estimer les paramètres de la queue de distribution en l'absence de covariable. Une preuve peut être trouvée en annexe A.

7.3 Biais de mauvaise spécification

Pour $X = x$, le but final est d'estimer les paramètres de la queue de distribution $\theta_0(X) = (\sigma_{0u}(x), \gamma_0(x))$ par la maximisation de la GP-vraisemblance. La différence entre $\theta_0(x)$ et $\theta^*(x)$ peut être vue comme un terme de mauvaise spécification lié au fait que les observations au-dessus du seuil ne sont pas exactement distribuées selon une GP. Ce biais peut être contrôlé par des conditions de second ordre, ce qui est usuel en analyse des valeurs extrêmes.

En effet, avec des échantillons finis, on n'a pas exactement que les excès sont distribués selon une GP. Cela introduit un terme de biais. Pour contrôler ce terme, une condition de second ordre est nécessaire, une condition pour contrôler la vitesse de convergence. On peut exprimer cette condition de plusieurs façons possibles. Ici on considère la même condition que la C6 de (Beirlant et Goegebeur 2004). On commence par réécrire que,

$$\overline{F}(y|x) = y^{-1/\gamma_0(x)}\eta(y|x), \forall y > 0,$$

avec η une fonction variant peu telle que, $\eta(ty|x)/\eta(t|x) \to 1$ quand $t \to \infty, \forall y > 0$.

Assumption 7.3.1. *On suppose que pour tout x, il existe une constance c et une fonction ψ tel que*

$$\eta(ty|x)/\eta(t|x) = 1 + c\psi(t)\int_1^t v^{\rho-1}\mathrm{d}v + o(\psi(t))$$

quand $t \to \infty$ pour chaque $y > 0$ avec $\psi(t) > 0$ et $\psi(t) \to 0$ quand $t \to \infty$ et $\rho \leq 0$

Le résultat suivant nous assure que terme de biais tend vers 0 quand $u \to \infty$

Proposition 7.3.2. *Il existe une constante c et une fonction ψ tel que $\psi(u) > 0$ et $\psi(u) \to 0$ quand $u \to \infty$ et de sorte que pour $X = x$*

$$||\theta_0(x) - \theta^*(x)||_\infty \leq C_2(u)\frac{k_n}{n}(1 + c\gamma_{max}\psi(u) + o(\psi(u)),$$

avec $C_2(u)$ un constante dépendant de γ_{min} et γ_{max}

7.4 Consistance de l'étape d'élagage

Les résultats précédents permettent de couvrir le cas d'un arbre avec un nombre fixé de feuilles. En pratique, la question est de sélectionner le sous-arbre correct du l'arbre $T_{max}(u)$, l'arbre maximal obtenu une fois que l'étape précédente du CART s'est arrêté, avec un nombre optimal de feuille. C'est l'objectif de l'étape de l'élagage. Le corollaire nous dit que l'erreur stochastique au carré, croit proportionnellement à K. On choisit usuellement une pénalité proportionnelle à K, le théorème suivant corrobore ce choix.

Commençons par définir, pour une décomposition, $(T_l^k)_{l=1,\dots,K}$ de K feuilles, $T_K(u)$ l'arbre avec des paramètres $\hat{\theta}_l^K(u)$ estimés par la procédure CART, $T_K^*(u)$, l'arbre de paramètres :

$$\theta_l^{*k}(u) = \arg\max_{\theta\in\Theta} \mathbb{E}\left[\phi(Y-u,\theta)\mathbf{1}_{Y>u}\mathbf{1}_{X_i\in T_l^k}\right],$$

et $x \to \theta^{*k}(x) = \sum_{l=1}^K \theta_l^{*k}(u)\mathbf{1}_{x\in T_l^k}$, la fonction de régression correspondante. De plus,

$$K_0(u) = \arg\max_{K=1,\dots,K_{max}} \mathbb{E}[\phi(Y-u,\theta^{*k}(X))\mathbf{1}_{Y>U}]$$

Soit que, $T^*(u) = T^*_{K_0(u)}(u)$ est un sous-arbre de $T_{max}(u)$ qui est le plus proche de $x \to \theta^{*k}(x)$ dans la mesure où il maximise l'espérance de la (pseudo) log-vraisemblance.

Ensuite, on désigne le nombre de feuilles sélectionnées :

$$\hat{K}(u) = \arg\max_{K=1,\dots,K_{max}} \left\{ \frac{1}{k_n}\sum_{l=1}^K\sum_{i=1}^n \phi(Y_i-u,\hat{\theta}^K(X_i))\mathbf{1}_{Y_i>u}\mathbf{1}_{X_i\in T_l} - \alpha K \right\},$$

et $\hat{T}(u) = T_{\hat{K}(u)}(u)$ l'arbre correspondant sélectionné.

On définit la log-vraisemblance $L_n(T_k,u)$ associé à un arbre $T_k(u)$ avec K feuilles, $(T_l^k)_{l=1,\dots,K}$ de paramètres $\hat{\theta}^k(u) = \left(\hat{\theta}_l^k(u)\right)_{l=1,\dots K}$

$$L_n(T_k,u) = \sum_{l=1}^K L_n^l(\hat{\theta}_l^K,u).$$

Alors $L(T_k,u) = \mathbb{E}[L_n(T_k,u)]$. Enfin pour deux arbres T et S, $\Delta L_n(T,S) = L_n(T,u) - L_n(S,u)$ et de la même manière $\Delta L(T,S) = L(T,u) - L(S,u)$. Le théorème suivant, montre que le méthode d'élagage sélectionne un arbre $\hat{T}(u)$ qui atteint approximativement le même taux que $T_{K_0}(u)$, même si $K_0(u)$ est inconnu, à condition que la constante de pénalité λ appartienne à un intervalle raisonnable.

Theorem 7.4.1. *Prenons $D = \inf_u \inf_{K<K_0(u)} \Delta L(T^*(u), T_K^*(u))$ et supposons qu'il existe une constante $C_2 > 0$ telle que la constante de pénalisation λ satisfasse :*

$$c_2\{\log k_n\}^{1/2}k_n^{-1/2} \le \lambda \le (D - 2c_2\{\log k_n\}^{1/2}k_n^{-1/2})k_n^{-1},$$

alors pour tout $u \in [u_{min}\,;u_{max}]$,

$$\mathbb{E}\left[||\hat{T}(u) - T^*(u)||_2^2\right] \le \frac{C_5 K_0(u)(\log k_n)^2}{k_n},$$

avec C_5 une constante dépendant de $T^*(u)$, une preuve se trouve aussi dans l'annexe.

7.5 Conclusion

Dans ce chapitre, nous avons étudié la consistance des arbres de régression Pareto généralisée, appliqués à la régression des valeurs extrêmes. Les résultats que nous obtenons sont non-asymptotiques, et permettent de justifier la consistance de l'étape d'élagage utilisée pour sélectionner un sous-arbre approprié. Notons que les conditions sous lesquelles nos résultats tiennent sont relativement faibles, dans le sens où elles tiennent même si l'indice de queue γ est arbitrairement proche de zéro (le cas particulier $\gamma = 0$ est exclu) ou grand. De plus, aucune hypothèse de régularité sur les paramètres cibles n'est requise, en raison de la flexibilité de la procédure de l'arbre de régression.

Conclusion et perspectives

Le but de cette thèse était de mettre à profit des méthodes d'apprentissages statistique pour l'étude des risques naturels en France et en particulier pour l'évaluation de ces conséquences. Nous avons contribué à l'évaluation du coût des événements de catastrophes naturelles sécheresse et inondations, avec la contrainte d'être disponible rapidement après l'événement. Nous avons aussi apporté de la connaissance sur la nature et le coût de l'endommagement à l'échelle fine du bâti en analysant les données des rapports d'expertise. Ces travaux s'inscrivaient dans un objectif d'amélioration de la connaissance et de la prévention, pour les sociétés d'assurances mais aussi pour l'intérêt général. Pour cela nous utilisons largement les données fournies par les sociétés d'assurances et réseaux d'expertise, récoltées et traitées par la Mission Risques Naturels.

Nous avons dans un premier chapitre, étudié la sinsitralité à l'échelle fine du bâtiment. Nous avons pour cela analysé les données textuelles des rapports d'expertise. Nous avons appliqué des méthodes de classification textuelle pour faire correspondre ces données, des champs de texte libre, à des composantes du bâtiment prédéfinis. Cette classification est faite en utilisant des réseaux de neurones avec une couche de plongement lexical. Cela nous permettait d'avoir des coûts pour chaque catégorie de dommages en fonction des événements naturels. Nous arrivons à obtenir des résultats très intéressants, cependant nous n'avions accès qu'à des échantillons restreints et ces analyses gagneraient à être renforcées. Nous pouvons grâce aux développements en analyse de texte bien traiter les données provenant des rapports d'expertise avec une bonne précision. Nous arrivons aussi avec une bonne fiabilité à récupérer des informations textuelles décrivant le bien, dans les rapports en entier. C'est une très bonne source de données qui peut aider à comprendre la sinistralité à l'échelle du bâti et à identifier les zones vulnérables d'un bâtiment pour pouvoir à terme les renforcer. Une part importante du travail consiste cependant à nettoyer et classer des données qui pourraient être harmonisées en amont. Les rapports pourraient être « normés » et il suffirait alors de récolter et de regrouper les données des différents réseaux d'expertise. C'est ce qui se produit dans plusieurs réseaux d'expertise, ils récoltent de plus en plus de données exploitables. Cependant pour avoir un niveau de dommages fin et cohérent entre les réseaux il serait judicieux de se mettre d'accord sur une structure commune au préalable. En particulier pour les informations de chiffrages qui restent difficiles à exploiter automatiquement dans des rapports en entier du fait de leurs formatages et de leurs structures trop spécifiques.

Dans une deuxième application nous avons essayé d'estimer le coût d'un événement de sécheresse. Notre méthode reposait sur la comparaison de différents modèles statistiques tels que les modèles linéaires généralisés combinés avec des pénalités Lasso et Elastic-Net avec des algorithmes d'apprentissage automatique, tels que les forêts aléatoires ou l'Extreme Gradient Boosting. Nous avons aussi étudié une agrégation de ces trois modèles. La calibration de ces méthodes est effectuée sur une importante base de données couvrant environ 70 % du marché français de l'assurance dommage et sur des données météorologiques renseignant l'intensité de l'événement, et des données d'exposition permettant de connaître la susceptibilité au retrait gonflement des argiles. Nous avons obtenu des résultats encourageants pour un phénomène aussi complexe, notre

méthode donne une indication pertinente sur la gravité potentielle de l'événement sécheresse en cours. Nous avons cependant rencontré des difficultés et la prédiction donne des résultats discutables à l'échelle de la commune. Cela reflète la nature complexe de ce risque qui dépend de nombreux facteurs, il ne peut être réduit à des valeurs d'un indice météorologique et géotechnique. Pour pouvoir faire une prédiction plus fiable il faudrait avoir accès à des variables biens plus fines, à l'échelle du bâtiment, comme celle présentes dans les rapports d'expertise.

Nous avons enfin proposé deux méthodes d'estimations pour le coût des événements inondations. La première méthode consiste à estimer la distribution statistique du coût de nos événements en nous concentrant sur les événements extrêmes. Pour cela nous appliquons la théorie des valeurs extrêmes et en particulier le Peaks over Threshold (PoT), couplée à des arbres de régression. Cette méthode possède des résultats théoriques assurant la consistance de cette procédure. Pour compléter cette démarche et permettre l'estimation du coût des événements nous utilisons la théorie de la crédibilité. Avec cette approche, le coût total dépend donc de la liste de communes impactées, de l'historique du coût sur ces communes mais aussi du type d'événement. C'est très précieux pour notre étude car les spécificités locales des communes sont prises en compte mais aussi le profil de la distribution des extrêmes de l'événement. Ce profil renseigne le type d'événement et constitue le prior de l'approche bayésienne. Nous avons aussi proposé une deuxième méthode fondée sur une approche sévérité fréquence, très utilisée en assurance. Cette méthode repose sur des indicateurs à l'échelle de l'événement et peut-être légèrement plus facile à interpréter. Nous obtenons des résultats encourageants avec les deux méthodes, qui seront utiles à la fédération dans son processus de gestion. C'est en effet dans une logique d'aide à la décision que ces méthodes ont été créées. Ces résultats peuvent néanmoins sûrement être améliorés grâce à l'intégration de variable météorologique fines renseignant l'intensité de l'événement.

Comme décrit en introduction une augmentation du coût des événements naturels est à prévoir dans les prochaines années du fait de l'augmentation des biens et du changement climatique. Pour maintenir le haut niveau de couverture des dommages par l'assurance, la réduction du coût des catastrophes naturelles est un enjeu essentiel. Il ne peut être atteint que par une amélioration de la connaissance et de la prévention des risques. Nous avons essayé d'y contribuer en améliorant la connaissance des événements naturels et en particulier de leurs coûts. En effet nos approches d'estimations permettent aussi de mieux comprendre les risques naturels. L'approche pour la sécheresse permet de mieux établir les liens entre le coût et des covariables d'intérêt, notamment météorologiques et géologiques. L'approche pour les inondations nous donne une typologie d'événements avec un focus sur les extrêmes. Cette classification permet d'identifier des territoires potentiellement vulnérables. Enfin l'analyse des rapports d'expertise produit directement de la connaissance en identifiant les parties vulnérables dans le bâtiment. Ces informations peuvent contribuer à une meilleure prévention si elles sont prises en compte, notamment au moment de la construction.

Une des perspectives pour nos travaux est de participer à terme à l'amélioration de la résilience des bâtiments. Pour cela dans une nouvelle thèse CIFRE, la MRN va proposer un outil de Diagnostic de Performance Résilience (DPR) et un outil de cotation associé. Nos travaux pourraient servir à alimenter ce diagnostic en permettant par exemple de chiffrer les dommages évités. Cela permettrait de valoriser les conséquences positives des mesures de résilience au regard du coût de leur installation. Une utilisation possible de nos travaux serait alors de regarder les coûts observés pour une réparation dans les rapports d'expertise et de le comparer aux coûts d'installation des mesures de résilience permettant de réduire ou d'éviter cette réparation. Si ces coûts d'installation sont inférieurs à l'endommagement potentiel alors l'analyse sera une incitation pour encourager davantage les constructeurs à mettre en place des mesures de résilience performantes. Ces travaux ambitieux bénéficieraient néanmoins, comme vu précédemment, d'avoir accès à plus de données et dans une forme plus harmonisée. Le dernier rapport du GIEC (Pörtner et al.

2022) se concentre sur l'adaptation au effet du changement climatique et ces données et études s'inscrivent parfaitement dans cette optique.

Sur le plan académique, les travaux sur les arbres de régression ouvrent des perspectives intéressantes pour l'étude des événements extrêmes. En gardant une bonne interprétabilité cette méthode peut être appliquée à des situations diverses. Cependant les arbres de régression peuvent être instables et il pourrait être intéressant d'étudier comment d'autres méthodes comme les forêts aléatoires pourraient être appliquées, à l'étude de la régression des valeurs extrêmes. L'application de la théorie de la crédibilité à notre problème ouvre aussi un champ original en permettant de considérer le coût à la commune en fonction d'un profil extrême de risque de l'événement qui la touche. On pourrait étendre cette analyse à d'autres aléas et à d'autres données. L'association de ces deux méthodes pour l'étude de la régression pourrait donc être approfondie. Enfin il est toujours pertinent de confronter les modèles mathématiques et algorithmes d'apprentissage statistiques aux données réelles. Dans cette thèse nous avons appliqué ces méthodes aux données de l'assurance et évaluer les prédictions sur des problèmes concrets. En particulier pour les données textuelles, où les résultats obtenus peuvent ne pas correspondre aux standards de ce domaine mais reflètent bien de l'importance de la qualité des données et de la difficulté de travailler avec des données réelles.

Annexe A

Preuves

A.1 Proofs

In this Section, we present in details the proof of the results presented throughout this thesis. Concentration inequalities required to obtain the results are presented in Section A.1.1. These inequalities are used to obtain deviation bounds in Section A.1.2, which are the key ingredients of the proof of Theorem 7.2.1 (Section A.1.3), Corollary 7.2.2 (Section A.1.4), and Theorem 7.4.1 (Section A.1.6). Section A.2 shows some results on covering numbers that are required to control the complexity of some classes of functions considered in the proofs. Some technical lemmas are gathered in Section A.3.

A.1.1 Concentration inequalities

The proofs of the main results are mostly based on concentration inequalities. The following inequality was proved initially by (Talagrand 1994), see also (Einmahl et Mason 2005).

Proposition A.1.1. *Let $(\mathbf{V}_i)_{1\leqslant i\leqslant n}$ denote i.i.d. replications of a random vector $\mathbf{V}$, and let $(\varepsilon_i)_{1\leqslant i\leqslant n}$ denote a vector of i.i.d. Rademacher variables (that is, $\mathbb{P}(\varepsilon_i=-1)=\mathbb{P}(\varepsilon_i=1)=1/2$) independent from $(\mathbf{V}_i)_{1\leqslant i\leqslant n}$. Let $\mathfrak{F}$ be a pointwise measurable class of functions bounded by a finite constant M_0. Then, for all t,*

$$\mathbb{P}\left(\sup_{\varphi\in\mathfrak{F}}\left\|\sum_{i=1}^{n}\{\varphi(\mathbf{V}_i)-[\varphi(\mathbf{V})]\}\right\|_\infty > A_1\left\{E\left[\sup_{\varphi\in\mathfrak{F}}\left\|\sum_{i=1}^{n}\varphi(\mathbf{V}_i)\varepsilon_i\right\|_\infty\right]+t\right\}\right) \leqslant 2\left\{\exp\left(-\frac{A_2t^2}{nv_{\mathfrak{F}}}\right)+\exp\left(-\frac{A_2t}{M_0}\right)\right\},$$

with $v_{\mathfrak{F}}=\sup_{\varphi\in\mathfrak{F}}\mathrm{Var}(\|\varphi(\mathbf{V})\|_\infty)$, and where A_1 and A_2 are universal constants.

The difficulty in using Proposition A.1.1 comes from the need to control the symmetrized quantity $\mathbb{E}\left[\sup_{\varphi\in\mathfrak{F}}\|\sum_{i=1}^{n}\varphi(\mathbf{V}_i)\varepsilon_i\|\right]$. Proposition A.1.2 is due to (Einmahl et Mason 2005) and allows this control via some assumptions on the considered class of functions $\mathfrak{F}$.

We first need to introduce some notations regarding covering numbers of a class of functions. More details can be found for example in Chapter 2.6 of (Vaart 1998). Let us consider a class of functions $\mathfrak{F}$ with envelope Φ (which means that for (almost) all v, $f\in\mathfrak{F}$, $|f(v)|\leqslant\Phi(v)$). Then, for any probability measure $\mathbb{Q}$, introduce $N(\varepsilon,\mathfrak{F},\mathbb{Q})$ the minimum number of $L^2(\mathbb{Q})$ balls

of radius ε to cover the class $\mathfrak{F}$. Then, define

$$\mathcal{N}_\Phi(\varepsilon, \mathfrak{F}) = \sup_{\mathbb{Q}:\mathbb{Q}(\Phi^2)<\infty} N(\varepsilon(\mathbb{Q}(\Phi^2)^{1/2}), \mathfrak{F}, \mathbb{Q}).$$

Proposition A.1.2. *Let $\mathfrak{F}$ be a point-wise measurable class of functions bounded by M_0 with envelope Φ such that, for some constants $A_3, \alpha \geqslant 1$, and $0 \leqslant \sqrt{v} \leqslant M_0$, we have*

(i) $\mathcal{N}_\Phi(\varepsilon, \mathfrak{F}) \leqslant A_3 \varepsilon^{-\alpha}$, *for* $0 < \varepsilon < 1$,

(ii) $\sup_{\varphi\in\mathfrak{F}} \mathbb{E}\left[\varphi(\mathbf{V})^2\right] \leqslant v$,

(iii) $M_0 \leqslant \frac{1}{4\alpha^{1/2}} \sqrt{nv/\log(A_4 M_0/\sqrt{v})}$, *with* $A_4 = \max(e, A_3^{1/\alpha})$.

Then, for some absolute constant A_5,

$$\mathbb{E}\left[\sup_{\varphi\in\mathfrak{F}} \left\| \sum_{i=1}^n \varphi(\mathbf{V}_i)\varepsilon_i \right\| \right] \leqslant A_5 \sqrt{\alpha n v \log(A_4 M_0/\sqrt{v})}.$$

A.1.2 Deviation results

We first introduce some notations that will be used throughout Sections A.1.2 to A.2. In the following, f_θ is a function indexed by $\theta = (\sigma, \gamma)^\tau$ denoting either $\phi(\cdot, \theta)$ or $g_\theta = \partial_\sigma \phi(\cdot, \theta)$, or $h_\theta = \partial_\gamma \phi(\cdot, \theta)$. Let us note that the functions $y \mapsto g_\theta(y-u)$ and $y \mapsto h_\theta(y-u)$ are uniformly bounded (eventually up to some multiplication by a constant) by $\Phi(y) = \log(1+wy)$, where $w = \gamma_{\max}/\sigma_{\min}$ (see Assumption 7.1.3). On the other hand, $y \mapsto \phi(y-u, \theta)$ is bounded by $\log \sigma_n + \Phi(y) = O(\log(k_n)) + \Phi(y)$. We consider in the following a class of functions $\mathfrak{F}$ defined as

$$\mathfrak{F} = \{ y \mapsto f_\theta(y-u)\mathbf{1}_{y\geqslant u}\mathbf{1}_{\mathbf{x}\in\mathcal{T}_\ell},\ \theta \in \Theta,\ u \in [u_{\min}; u_{\max}], \ell = 1, ..., K \}. \tag{A.1.1}$$

Next, recall that for $\ell = 1, \ldots, K$

$$L_n^\ell(\theta, u) = \frac{1}{k_n} \sum_{i=1}^n \phi(Y_i - u, \theta)\mathbf{1}_{Y_i>u}\mathbf{1}_{\mathbf{X}_i\in\mathcal{T}_\ell},$$

is the (normalized) GP log-likelihood in the leaf ℓ of the tree $T(u) = (\mathcal{T}_\ell)_{\ell=1,\ldots,K}$. The key results behind Theorems 7.2.1 and 7.4.1 relies on studying the deviation of the processes

$$\begin{aligned} \mathcal{W}_0^\ell(\theta, u) &= L_n^\ell(\theta, u) - L^\ell(\theta, u), \\ \mathcal{W}_1^\ell(\theta, u) &= \nabla_\theta L_n^\ell(\theta, u) - \nabla_\theta L^\ell(\theta, u), \end{aligned}$$

indexed by θ, u and ℓ.

We study these deviations by decomposing $\mathcal{W}_i^\ell(\theta, u)$, for $i = 0, 1$, (which is a sum of i.i.d. observations) into two sums :

— the first one gathers observations smaller than some bound (more precisely, such that $\Phi(Y_i) \leqslant M_n$), which is considered in Theorem A.1.3. Since these observations are bounded (even if this bound in fact depends on n and can tend to infinity when n grows), we can apply a concentration inequality such as the one of Section A.1.1 ;

— in the second one, we consider the observations larger than this bound, and control them through the fact that the function Φ is assumed to have a finite exponential moment (see Assumption 7.1.3).

Corollary A.1.5, which provides deviation bounds for estimation errors in the leaves of the tree, is then a direct consequence.

Theorem A.1.3. *Let $M_n = \beta \log k_n$, with $\beta > 0$ and*

$$\underline{\mathcal{Z}}(M_n) = \sup_{f\in\mathfrak{F}} \left| \frac{1}{k_n} \sum_{i=1}^{n} \left(f(Y_i)\mathbf{1}_{\Phi(Y_i)\leqslant M_n} - \left[f(Y_i)\mathbf{1}_{\Phi(Y_i)\leqslant M_n} \right] \right) \right|$$

Then, under Assumptions 7.1.1, 7.1.2 and 7.1.4,

$$\mathbb{P}\left(\underline{\mathcal{Z}}(M_n) \geqslant t\right) \leqslant 2\left(\exp\left(-\frac{C_1 k_n t^2}{M_n^2} \right) + \exp\left(-\frac{C_2 k_n t}{M_n} \right) \right), \tag{A.1.2}$$

for $t \geqslant \mathfrak{c}_1 (\log k_n)^{1/2} k_n^{-1/2}$.

Démonstration. Let us stress that $\sup_{f\in\mathfrak{F}} \|f(y)\mathbf{1}_{\Phi(y)\leqslant M_n}\|_\infty \leqslant M_n$. From Proposition A.1.1,

$$\begin{aligned} & \mathbb{P}\left(\underline{\mathcal{Z}}(M_n) \geqslant A_1 \left\{ \mathbb{E}\left[\sup_{f\in\mathfrak{F}} \frac{1}{k_n} \left| \sum_{i=1}^{n} f(Y_i)\mathbf{1}_{\Phi(Y_i)\leqslant M_n}\varepsilon_i \right| \right] + t \right\} \right) \qquad (A.1.3) \\ \leqslant \; & 2\left(\exp\left(-\frac{A_2 k_n^2 t^2}{n v_{\mathfrak{F}}} \right) + \exp\left(-\frac{A_2 k_n t}{M_n} \right) \right). \end{aligned}$$

From Lemma A.3.1, $v_{\mathfrak{F}} \leqslant M_n^2 k_n n^{-1}$, which shows that the first exponential term on the right-hand side of (A.1.3) is smaller than

$$\exp\left(-\frac{A_2 k_n t^2}{M_n^2} \right). \tag{A.1.4}$$

We can now apply Proposition A.1.2 (combined with Lemma A.2.1) to this class of functions with $v = M_n^2 k_n n^{-1}$ and $M_0 = M_n$. Hence,

$$\mathbb{E}\left[\sup_{f\in\mathfrak{F}} \frac{1}{k_n} \left| \sum_{i=1}^{n} f(Y_i)\mathbf{1}_{\Phi(Y_i)\leqslant M_n}\varepsilon_i \right| \right] \leqslant \frac{A_6}{k_n} \sqrt{n v \mathfrak{s}_n} = A_6 \frac{\mathfrak{s}_n^{1/2}}{k_n^{1/2}},$$

where $A_6' > 0$ and $\mathfrak{s}_n = \log(\sigma_n^\alpha K^{4(d+1)(d+2)} n/k_n)$ ($\alpha > 0$ being defined in Lemma A.2.1). From Assumption 7.1.2, we see that $\mathfrak{s}_n = O(\log(k_n))$ (let us recall that K is necessarily less than n). Whence, if $\mathfrak{c}_1 = 2A_1A_6'$, for $t \geqslant \mathfrak{c}_1 \left\{\log\left(k_n\right)\right\}^{1/2} k_n^{-1/2}$,

$$\mathbb{P}\left(\underline{\mathcal{Z}}(M_n) \geqslant t\right) \leqslant \mathbb{P}\left(\underline{\mathcal{Z}}(M_n) \geqslant A_1 \left\{ \mathbb{E}\left[\sup_{f\in\mathfrak{F}} \frac{1}{k_n} \left| \sum_{i=1}^{n} f(Y_i)\mathbf{1}_{\Phi(Y_i)\leqslant M_n}\varepsilon_i \right| \right] + \frac{t}{2A_1} \right\} \right).$$

Equation (A.1.2) follows from (A.1.3) and (A.1.4) with $C_1 = A_2 A_1^{-2}/4$ and $C_2 = A_2 A_1^{-1}/2$. □

Theorem A.1.4. *Define*

$$\overline{\mathcal{Z}}(M_n) = \sup_{f\in\mathfrak{F}} \left| \frac{1}{k_n} \sum_{i=1}^{n} \left(f(Y_i)\mathbf{1}_{\Phi(Y_i)>M_n} \right) - \left[f(Y_i)\mathbf{1}_{\Phi(Y_i)>M_n} \right] \right|.$$

Then, under Assumptions 7.1.1, 7.1.2 and 7.1.3, for $M_n = \beta \log k_n = \beta a_2 \log n$ and $\beta a_2 \geqslant 10/\rho_0$, and $t \geqslant \mathfrak{c}_2 k_n^{-1/2}$,

$$\mathbb{P}\left(\overline{\mathcal{Z}}(M_n) \geqslant t\right) \leqslant \frac{C_3}{k_n^{5/2} t^3}. \tag{A.1.5}$$

Démonstration. Let $\beta' = \beta a_2$. $\overline{\mathcal{Z}}(M_n)$ is upper-bounded by

$$\frac{1}{k_n}\sum_{i=1}^{n}\left\{\Phi(Y_i)\mathbf{1}_{\Phi(Y_i)\geqslant M_n}\mathbf{1}_{Y_i\geqslant u_{\min}} + \mathbb{E}\left[\Phi(Y)\mathbf{1}_{\Phi(Y)\geqslant M_n}\mathbf{1}_{Y\geqslant u_{\min}}\right]\right\}.$$

A bound for $E_{1,n} = \mathbb{E}\left[\Phi(Y)\mathbf{1}_{\Phi(Y)\geqslant M_n}\mathbf{1}_{Y\geqslant u_{\min}}\right]$ is obtained from Lemma A.3.2, and $nE_{1,n}/k_n \leqslant \mathfrak{e}_1 k_n^{-1/2}$ if $\beta' \geqslant 2/\rho_0$.

Next, from Markov inequality,

$$\begin{aligned} t^3\mathbb{P}\left(\frac{1}{k_n}\sum_{i=1}^{n}\Phi(Y_i)\mathbf{1}_{\Phi(Y_i)\geqslant M_n}\mathbf{1}_{Y_i\geqslant u_{\min}} \geqslant t\right) &\leqslant \frac{nE_{3,n}}{k_n^3} + \frac{n(n-1)E_{2,n}E_{1,n}}{k_n^3} \\ &\quad + \frac{n(n-1)(n-2)E_{1,n}^3}{k_n^3}. \end{aligned}$$

From Lemma A.3.2, we get

$$\begin{aligned} \frac{nE_{3,n}}{k_n^3} &\leqslant \frac{\mathfrak{e}_3 n^{-(\rho_0\beta'/4-1/2)}}{k_n^{5/2}}, \\ \frac{n(n-1)E_{2,n}E_{1,n}}{k_n^3} &\leqslant \frac{\mathfrak{e}_2\mathfrak{e}_1 n^{-(\rho_0\beta'/2-3/2)}}{k_n^{5/2}}, \\ \frac{n(n-1)(n-2)E_{1,n}^3}{k_n^3} &\leqslant \frac{\mathfrak{e}_1^3 n^{-(\rho_0\beta'/4-5/2)}}{k_n^{5/2}}. \end{aligned}$$

Each of these terms is bounded by $\max(\mathfrak{e}_3, \mathfrak{e}_2\mathfrak{e}_1, \mathfrak{e}_1^3)k_n^{-5/2}$ for $\beta' \geqslant 10/\rho_0$. Thus, for $t \geqslant 2\mathfrak{e}_1 k_n^{-1/2}$ and $\beta' \geqslant 10/\rho_0$,

$$\begin{aligned} &\mathbb{P}\left(\overline{\mathcal{Z}}_n \geqslant t\right) \\ &\quad\leqslant \mathbb{P}\left(\frac{1}{k_n}\sum_{i=1}^{n}\Phi(Y_i)\mathbf{1}_{\Phi(Y_i)\geqslant M_n}\mathbf{1}_{Y_i\geqslant u_{\min}} \geqslant \frac{t}{2}\right) + \mathbb{P}\left(\mathbb{E}\left[\Phi(Y)\mathbf{1}_{\Phi(Y)\geqslant M_n}\mathbf{1}_{Y\geqslant u_{\min}}\right] \geqslant \frac{t}{2}\right) \\ &\quad\leqslant \frac{8\max(\mathfrak{e}_3, \mathfrak{e}_2\mathfrak{e}_1, \mathfrak{e}_1^3)}{t^3 k_n^{5/2}} \end{aligned}$$

□

We now apply these results to deduce deviation bounds on the estimators $\widehat{\theta}_\ell$ in the leaves of the tree.

Corollary A.1.5. *Under the assumptions of Theorem A.1.3 and A.1.4 and Assumption 7.1.4, for* $t \geqslant \mathfrak{c}_3(\log k_n)^{1/2}k_n^{-1/2}$,

$$\begin{aligned} \mathbb{P}\left(\sup_{\substack{\ell=1,\dots,K,\\ u_{\min}\leqslant u\leqslant u_{\max}}} \|\widehat{\theta}_\ell(u) - \theta_\ell^*(u)\|_\infty \geqslant t\right) &\leqslant 2\left(\exp\left(-\frac{C_4 k_n t^2}{\beta^2(\log k_n)^2}\right) + \exp\left(-\frac{C_5 k_n t}{\beta\log k_n}\right)\right) \\ &\quad + \frac{C_6}{k_n^{5/2}t^3}. \end{aligned}$$

Démonstration. For $1 \leqslant \ell \leqslant K$ and $u_{\min} \leqslant u \leqslant u_{\max}$, write $\theta = (s, \gamma)^\tau$ and $\theta_\ell^*(u) = (s_\ell^*(u), \gamma_\ell^*(u))^\tau$, and let $m_{u,\ell}(\theta) = \nabla_\theta L^\ell(\theta, u)$. From a Taylor expansion,

$$m_{u,\ell}(\theta) = \mathbb{E}\left[\begin{pmatrix} \partial_s g_{\tilde{s}_1,\gamma}(Y-u) & \partial_\gamma g_{s,\tilde{\gamma}_1}(Y-u) \\ \partial_s h_{\tilde{s}_2,\gamma}(Y-u) & \partial_\gamma h_{s,\tilde{\gamma}_2}(Y-u) \end{pmatrix} \mathbf{1}_{\mathbf{X}\in\mathcal{T}_\ell} \mathbf{1}_{Y \geqslant u}\right] (\theta - \theta_\ell^*(u))^\tau,$$

for some parameters $\tilde{\gamma}_j$ (resp. $\tilde{s}_j$) between γ and $\gamma_\ell^*(u)$ (resp. s and $s_\ell^*(u)$). From Assumption 7.1.4, we get, for all $\ell = 1, \ldots, K$,

$$\frac{n}{k_n} \|m_{u,\ell}(\theta)\|_\infty \geqslant \mathfrak{C}_1 \|\theta - \theta_\ell^*(u)\|_\infty.$$

Hence, for all $\ell = 1, \ldots, K$,

$$\mathbb{P}\left(\|\widehat{\theta}_\ell(u) - \theta_\ell^*(u)\|_\infty \geqslant t\right) \leqslant \mathbb{P}\left(\frac{n}{k_n} \|m_{u,\ell}(\widehat{\theta})\|_\infty \geqslant \mathfrak{C}_1 t\right).$$

Since for all $\ell = 1, \ldots, K$, $\nabla_\theta L_n^\ell(\widehat{\theta}) = 0$, $\mathcal{W}_1^\ell(\widehat{\theta}(u), u) = -\frac{n}{k_n} m_{u,\ell}(\widehat{\theta})$. Hence,

$$\mathbb{P}\left(\sup_{\substack{\ell=1,\ldots,K,\\ u_{\min}\leqslant u\leqslant u_{\max}}} \|\widehat{\theta}_\ell(u) - \theta_l^*(u)\|_\infty \geqslant t\right) \leqslant \mathbb{P}\left(\sup_{\substack{\ell=1,\ldots,K,\\ u_{\min}\leqslant u\leqslant u_{\max}}} \|\mathcal{W}_1^\ell(\widehat{\theta}(u), u)\|_\infty \geqslant \mathfrak{C}_1 t\right),$$

and the right-hand side is bounded by

$$\mathbb{P}\left(\overline{\mathcal{Z}}(M_n) \geqslant \frac{\mathfrak{C}_1 t}{2}\right) + \mathbb{P}\left(\underline{\mathcal{Z}}(M_n) \geqslant \frac{\mathfrak{C}_1 t}{2}\right).$$

The result follows from Theorem A.1.3 and A.1.4. □

A.1.3 Proof of Theorem 7.2.1

The proof of Theorem 7.2.1 then consists in gathering the results on the leaves obtained in Corollary A.1.5. Let $u_{\min} \leqslant u \leqslant u_{\max}$,

$$\|T(u) - T^*(u \mid T)\|_2^2 \leqslant \sum_{\ell=1}^K \|\widehat{\theta}_\ell(u) - \theta_\ell^*(u)\|_\infty^2 \leqslant K \sup_{\ell=1,\ldots,K} \|\widehat{\theta}_\ell(u) - \theta_\ell^*(u)\|_\infty^2.$$

Hence

$$\begin{aligned}&\mathbb{P}\left(\sup_{u_{\min}\leqslant u\leqslant u_{\max}} \|T(u) - T^*(u \mid T)\|_2^2 \geqslant t\right) \\ &\leqslant \mathbb{P}\left(\sup_{\substack{\ell=1,\ldots,K,\\ u_{\min}\leqslant u\leqslant u_{\max}}} \|\widehat{\theta}_\ell(u) - \theta_\ell^*(u)\|_\infty \geqslant t^{1/2} K^{-1/2}\right).\end{aligned}$$

The results follows from Corollary A.1.5, and from the assumption on $K \leqslant K_{\max} = O(k_n^3)$.

A.1.4 Proof of Corollary 7.2.2

Write

$$\mathbb{E}\left[\sup_{u_{\min}\leqslant u\leqslant u_{\max}}\|T(u)-T^*(u\mid T)\|_2^2\right]=\int_0^\infty \mathbb{P}(\sup_{u_{\min}\leqslant u\leqslant u_{\max}}\|T(u)-T^*(u\mid T)\|_2^2\geqslant t)dt.$$

Let $t_n = c_1 K(\log k_n)k_n^{-1}$, then

$$\begin{aligned}\int_0^\infty \mathbb{P}(\sup_{u_{\min}\leqslant u\leqslant u_{\max}}\|T(u)-T^*(u\mid T)\|_2^2\geqslant t)dt\\ \leqslant\quad t_n+\int_{t_n}^\infty \mathbb{P}(\sup_{u_{\min}\leqslant u\leqslant u_{\max}}\|T(u)-T^*(u\mid T)\|_2^2\geqslant t)dt.\end{aligned}$$

We now use Theorem 7.2.1 to bound the integral on the right-hand side. Since $\int_0^\infty \exp(-at)dt = \frac{1}{a}$, $\int_0^\infty \exp(-a^{1/2}t^{1/2})dt = \frac{2}{a}$, and $\int_1^\infty t^{-3/2}dt = 2$, we get

$$\begin{aligned}\mathbb{E}\left[\sup_{u_{\min}\leqslant u\leqslant u_{\max}}\|T(u)-T^*(u\mid T)\|_2^2\right] &\leqslant t_n+\frac{2K\beta^2(\log k_n)^2}{\mathcal{C}_1 k_n}+\frac{4K\beta^2(\log k_n)^2}{\mathcal{C}_2^2 k_n}+\frac{2\mathcal{C}_3 K}{k_n^{5/2}}\\ &\leqslant \frac{c_1 K\log k_n}{k_n}+\frac{2K\beta^2(\log k_n)^2}{\mathcal{C}_1 k_n}\\ &\quad+\frac{4K\beta^2(\log k_n)^2}{\mathcal{C}_2^2 k_n}+\frac{2\mathcal{C}_3 K}{k_n^{5/2}}\\ &\leqslant \frac{\mathcal{C}_4 K(\log k_n)^2}{k_n}.\end{aligned}$$

A.1.5 Proof of Proposition 7.3.2

Let $\mathbf{x}$ fixed, then,

$$\|\theta^*(\mathbf{x})-\theta_0(\mathbf{x})\|_\infty = \|\sum_{\ell=1}^{K_{\max}}(\theta_\ell^*-\theta_0(\mathbf{x}))\,\mathbf{1}_{\mathbf{x}\in\mathcal{T}_\ell}\|_\infty \leqslant \sum_{\ell=1}^{K_{\max}}\|\theta_\ell^*-\theta_0(\mathbf{x})\|_\infty \mathbf{1}_{\mathbf{x}\in\mathcal{T}_\ell}\,.$$

Now, from Taylor expansion, for $\ell=1,\dots,K$, conditionally on $\mathbf{X}\in\mathcal{T}_\ell$,

$$\begin{aligned}\nabla_\theta L^\ell(\theta_0(\mathbf{x}),u) &= \nabla_\theta L^\ell(\theta_\ell^*,u)+\nabla_\theta^2 L^\ell(\widetilde{\theta}_\ell)(\theta_0(\mathbf{x})-\theta_\ell^*)^\tau\\ &= \mathbb{E}\left[\begin{pmatrix}\partial_\sigma g_{\tilde{\sigma}_1,\gamma}(Y-u) & \partial_\gamma g_{\sigma,\tilde{\gamma}_1}(Y-u)\\ \partial_\sigma h_{\tilde{\sigma}_2,\gamma}(Y-u) & \partial_\gamma h_{\sigma,\tilde{\gamma}_2}(Y-u)\end{pmatrix}\mathbf{1}_{Y\geqslant u}\mid \mathbf{X}\in\mathcal{T}_\ell\right](\theta_0(\mathbf{x})-\theta_\ell^*)^\tau\end{aligned}$$

for some parameters $\tilde{\gamma}_j$ (resp. $\tilde{\sigma}_j$) between $\gamma_0(\mathbf{x})$ and γ_ℓ^* (resp. $\sigma_0(\mathbf{x})$ and σ_ℓ^*).

Thus, under Assumption 7.1.4,

$$\begin{aligned}\|\theta_0(\mathbf{x})-\theta_\ell^*\|_\infty &\leqslant \frac{1}{\mathfrak{C}_1}\|\nabla_\theta L^\ell(\theta_0(\mathbf{x}),u)\|_\infty\\ &\leqslant \frac{1}{\mathfrak{C}_1}\frac{k_n}{n}\max\left(\left|\left[g_{\theta_0(\mathbf{x})}(Z)\mid\mathbf{X}\in\mathcal{T}_\ell\right]\right|,\left[h_{\theta_0(\mathbf{x})}(Z)\mid\mathbf{X}\in\mathcal{T}_\ell\right]\right),\end{aligned}$$

where Z is a random variable distributed according to the distribution F_u defined in Section 6.2 with $\sigma_0(\mathbf{x}) = u\gamma_0(\mathbf{x})$ and with

$$\begin{aligned}
\left[g_{\theta_0(\mathbf{x})}(Z) \mid \mathbf{X} \in \mathcal{T}_\ell\right] &= -\frac{1}{u\gamma_0(\mathbf{x})} + \frac{1}{u^2\gamma_0(\mathbf{x})}\left(1+\frac{1}{\gamma_0(\mathbf{x})}\right)\left[\frac{Z}{1+Z/u} \mid \mathbf{X} \in \mathcal{T}_\ell\right] \\
\left[h_{\theta_0(\mathbf{x})}(Z) \mid \mathbf{X} \in \mathcal{T}_\ell\right] &= -\frac{1}{\gamma_0(\mathbf{x})^2}\left[\log(1+Z/u) \mid \mathbf{X} \in \mathcal{T}_\ell\right] \\
&\quad + \frac{1}{u\gamma_0(\mathbf{x})}\left(1+\frac{1}{\gamma_0(\mathbf{x})}\right)\left[\frac{Z}{1+Z/u} \mid \mathbf{X} \in \mathcal{T}_\ell\right].
\end{aligned}$$

Under Assumption 7.3.1, we have

$$\overline{F}_u(z) = \left(1+\frac{z}{u}\right)^{-1/\gamma_0(\mathbf{x})}\left\{1 + c\psi(u)\int_1^{1+z/u} v^{\rho-1}v + o(\psi(u))\right\}.$$

$$\begin{aligned}
\left[\frac{Z}{1+Z/u} \mid \mathbf{X} \in \mathcal{T}_\ell\right] &= \int_0^u \overline{F}_u\left(\frac{t}{1-t/u}\right)t \\
&= \frac{u}{1+1/\gamma_0(\mathbf{x})}\left(1+\frac{c\psi(u)}{1+1/\gamma_0(\mathbf{x})-\rho}+o(\psi(u))\right) \\
&\leqslant u\left(1+c\gamma_0(\mathbf{x})\psi(u)+o(\psi(u))\right)
\end{aligned}$$

and then

$$\begin{aligned}
\left[\log(1+Z/u) \mid \mathbf{X} \in \mathcal{T}_\ell\right] &= \int_0^u \mathbb{P}\left[Z \geqslant u(^t-1) \mid \mathbf{X} \in \mathcal{T}_\ell\right]t \\
&= \gamma_0(\mathbf{x})\left(1+\frac{c\psi(u)}{1/\gamma_0(\mathbf{x})-\rho}+o(\psi(u))\right) \\
&\leqslant \gamma_0(\mathbf{x})\left(1+c\gamma_0(\mathbf{x})\psi(\mathbf{x})(u)+o(\psi(u))\right).
\end{aligned}$$

Consequently,

$$\left|\left[g_{\theta_0(\mathbf{x})}(Z) \mid \mathbf{X} \in \mathcal{T}_\ell\right]\right| \leqslant \frac{1}{\gamma_{\min}}\left(1+\frac{1}{u}\left(1+\frac{1}{\gamma_{\min}}\right)\right)\left(1+c\gamma_0(\mathbf{x})\psi(u)+o(\psi(u))\right)$$

and

$$\left|\left[h_{\theta_0(\mathbf{x})}(Z) \mid \mathbf{X} = \mathbf{x}\right]\right| \leqslant \frac{1}{\gamma_{\min}}\left(1+\frac{1}{\gamma_{\min}}+\frac{\gamma_{\max}}{\gamma_{\min}}\right)\left(1+c\gamma_0(\mathbf{x})\psi(u)+o(\psi(u))\right).$$

Hence,

$$\|\theta_0(\mathbf{x}) - \theta_\ell^*\|_\infty \leqslant \mathfrak{C}_2(u)\frac{k_n}{n}\left(1+c\gamma_{\max}\psi(u)+o(\psi(u))\right),$$

where $\mathfrak{C}_2(u) = \frac{1}{\mathfrak{C}_1}\frac{1}{\gamma_{\min}}\max\left(1+\frac{1}{u}+\frac{1}{u\gamma_{\min}}, 1+\frac{1}{\gamma_{\min}}+\frac{\gamma_{\max}}{\gamma_{\min}}\right)$.

Finally,

$$\begin{aligned}\|\theta^*(\mathbf{x}) - \theta_0(\mathbf{x})\|_\infty &\leqslant \sum_{\ell=1}^{K_{\max}} \|\theta_\ell^* - \theta_0(\mathbf{x})\|_\infty \mathbf{1}_{\mathbf{x}\in\mathcal{T}_\ell} \\ &\leqslant \mathfrak{C}_2(u)\frac{k_n}{n}\left(1 + c\gamma_{\max}\psi(u) + o(\psi(u))\right)\sum_{\ell=1}^{K_{\max}} \mathbf{1}_{\mathbf{x}\in\mathcal{T}_\ell} \\ &\leqslant \mathfrak{C}_2(u)\frac{k_n}{n}\left(1 + c\gamma_{\max}\psi(u) + o(\psi(u))\right) .\end{aligned}$$

A.1.6 Proof of Theorem 7.4.1

The following lemma will be needed to prove Theorem 7.4.1.

Lemma A.1.6. *Let $\mathfrak{D} = \inf_u \inf_{K<K_0(u)} \Delta L(T^*(u), T_K^*(u))$ and $u \in [u_{\min}, u_{\max}]$ fixed. Suppose that there exists a constant $c_2 > 0$ such that the penalization constant λ satisfies*

$$c_2\{\log k_n\}^{1/2}k_n^{-1/2} \leqslant \lambda \leqslant (\mathfrak{D} - 2c_2\{\log(k_n)\}^{1/2}k_n^{-1/2})k_n^{-1},$$

then, for $K > K_0(u)$,

$$\begin{aligned}\mathbb{P}(\widehat{K}(u) = K) &\leqslant 2\left(\exp\left(-\frac{C_1k_n\lambda^2(K-K_0(u))^2}{\beta^2(\log k_n)^2}\right) + \exp\left(-\frac{C_2k_n\lambda(K-K_0(u)))}{\beta\log k_n}\right)\right) \\ &\quad + \frac{C_3}{k_n^{5/2}\lambda^3(K-K_0(u))^3},\end{aligned}$$

and, for $K < K_0(u)$,

$$\begin{aligned}\mathbb{P}(\widehat{K}(u) = K) &\leqslant 4\exp\left(-\frac{C_1k_n\{\mathfrak{D} - \lambda(K_0(u)-K)\}^2}{\beta^2(\log k_n)^2}\right) \\ &\quad + 4\exp\left(-\frac{C_2k_n\{\mathfrak{D} - \lambda(K_0(u)-K)\}}{\beta\log k_n}\right) \\ &\quad + \frac{2C_3}{k_n^{5/2}\{\mathfrak{D} - \lambda(K_0(u)-K)\}^3}.\end{aligned}$$

Démonstration. Let $u \in [u_{\min}, u_{\max}]$ fixed. If $\widehat{K}(u) = K$, this means that

$$\Delta L_n(T_K(u), T_{K_0(u)}(u)) := L_n(T_K, u) - L_n(T_{K_0(u)}, u) > \lambda(K - K_0(u)).$$

Decompose

$$\begin{aligned}\Delta L_n(T_K(u), T_{K_0}(u)) &= \{L_n(T_K, u) - L_n(T_K^*, u)\} + \{L_n(T_K^*, u) - L_n(T^*, u)\} \\ &\quad + \{L_n(T^*, u) - L_n(T_{K_0(u)}, u)\}.\end{aligned}$$

Since $L_n(T^*, u) - L_n(T_{K_0(u)}, u) < 0$,

$$\Delta L_n(T_K(u), T_{K_0(u)}(u)) \leqslant \{L_n(T_K, u) - L_n(T_K^*, u)\} + \{L_n(T_K^*, u) - L_n(T^*, u)\}.$$

For $K > K_0(u)$, $T_K^*(u) = T^*(u)$, hence,

$$\begin{aligned}\mathbb{P}(\widehat{K}(u) = K) &\leqslant \mathbb{P}\left(\Delta L_n(T_K(u), T_K^*(u)) > \lambda(K - K_0(u))\right) \\ &\leqslant \mathbb{P}\left(|\Delta L_n(T_K(u), T_K^*(u)) - \Delta L(T_K(u), T_K^*(u))| > \lambda(K - K_0(u))\right).\end{aligned}$$

For $K > K_0(u)$, a bound is then obtained from Theorems A.1.3 and A.1.4 if $\lambda(K - K_0(u)) \geqslant c_1\{\log(k_n)\}^{1/2}k_n^{-1/2}$, that is $\lambda \geqslant c_1\{\log k_n\}^{1/2}k_n^{-1/2}$. Now, for $K < K_0(u)$,

$$\begin{aligned}\Delta L_n(T_K^*(u), T^*(u)) &\leqslant |\Delta L_n(T_K^*(u), T^*(u)) - \Delta L(T_K^*(u), T^*(u))| + \Delta L(T_K^*(u), T^*(u)) \\ &\leqslant |\Delta L_n(T^*(u), T_K^*(u)) - \Delta L(T^*(u), T_K^*(u))| - \mathfrak{D}(K_0(u), K).\end{aligned}$$

where $\mathfrak{D} = \inf_{K<K_0(u), u\in[u_{\min}, u_{\max}]} \mathfrak{D}(K_0(u), K)$, Hence,

$$\begin{aligned}&\mathbb{P}(\widehat{K}(u) = K) \\ &\leqslant \mathbb{P}\left(\Delta L_n(T_K(u), T_K^*(u)) \geqslant \frac{\mathfrak{D} - \lambda(K_0(u) - K)}{2}\right) \\ &\quad + \mathbb{P}\left(|\Delta L_n(T^*(u), T_K^*(u)) - \Delta L(T^*(u), T_K^*(u))| \geqslant \frac{\mathfrak{D} - \lambda(K_0(u) - K)}{2}\right) \\ &\leqslant \mathbb{P}\left(|\Delta L_n(T_K(u), T_K^*(u)) - \Delta L(T_K(u), T_K^*(u))| \geqslant \frac{\mathfrak{D} - \lambda(K_0(u) - K)}{2}\right) \\ &\quad + \mathbb{P}\left(|\Delta L_n(T^*(u), T_K^*(u)) - \Delta L(T^*(u), T_K^*(u))| \geqslant \frac{\mathfrak{D} - \lambda(K_0(u) - K)}{2}\right).\end{aligned}$$

These two probabilities can be bounded using Theorems A.1.3 and A.1.4 provided that, for all $K < K_0(u)$,

$$\frac{\mathfrak{D} - \lambda(K_0(u) - K)}{2} \geqslant \mathfrak{c}_1\{\log(k_n)\}^{1/2}k_n^{-1/2},$$

that is,

$$\lambda \leqslant \mathfrak{D} - 2\mathfrak{c}_1\{\log(k_n)\}^{1/2}k_n^{-1/2}.$$

□

We are now ready to prove Theorem 7.4.1. Let $u \in [u_{\min}, u_{\max}]$ fixed.

$$\begin{aligned}
\left[\|\widehat{T}(u) - T^*(u)\|_2^2\right] &= \sum_{K=1}^{K_{\max}} \left[\|T_K(u) - T^*(u)\|_2^2 \mathbf{1}_{\widehat{K}(u)=K}\right] \\
&\leqslant \left[\|T_{K_0(u)}(u) - T^*(u)\|_2^2\right] + \sum_{K=1, K\neq K_0(u)}^{K_{\max}} K\mathbb{P}(\widehat{K}(u) = K) \\
&\quad + \sum_{K=1, K\neq K_0(u)}^{K_{\max}} \mathbb{E}\left[\|T_K(u) - T^*(u)\|_2^2 \mathbf{1}_{\|T_K(u)-T^*(u)\|_2^2 > K} \mathbf{1}_{\widehat{K}(u)=K}\right] \\
&\leqslant \left[\|T_{K_0(u)}(u) - T^*(u)\|_2^2\right] + \sum_{K=1}^{K_0(u)-1} K\mathbb{P}(\widehat{K}(u) = K) \\
&\quad + \sum_{K=K_0(u)+1}^{K_{\max}} K\mathbb{P}(\widehat{K}(u) = K) \\
&\quad + 2 \sum_{K=1, K\neq K_0(u)}^{K_{\max}} \mathbb{E}\left[\|T_K(u) - T_K^*(u)\|_2^2 \mathbf{1}_{\|T_K(u)-T^*(u)\|_2^2 > K}\right] \\
&\quad + 2 \sum_{K=1, K\neq K_0(u)}^{K_{\max}} \mathbb{P}(\widehat{K}(u) = K)\|T^*(u) - T_K^*(u)\|_2^2.
\end{aligned}$$

Firstly, from Theorem 7.2.1,

$$\begin{aligned}
&\mathbb{E}\left[\|T_K(u) - T_K^*(u)\|_2^2 \mathbf{1}_{\|T_K(u)-T^*(u)\|_2^2 > K}\right] \\
&= K\mathbb{P}\left(\|T_K(u) - T_K^*(u)\|_2^2 > K\right) + \int_K^\infty \mathbb{P}\left(\|T_K(u) - T_K^*(u)\|_2^2 > t\right) t \\
&\leqslant 2K\left(1 + \frac{\beta^2(\log k_n)^2}{\mathcal{C}_1 k_n}\right) \exp\left(-\frac{\mathcal{C}_1 k_n}{\beta^2(\log k_n)^2}\right) \\
&\quad + 2K\left(1 + \frac{2\beta(\log k_n)}{\mathcal{C}_2 k_n} + \frac{2\beta^2(\log k_n)^2}{\mathcal{C}_2^2 k_n^2}\right) \exp\left(-\frac{\mathcal{C}_2 k_n}{\beta(\log k_n)}\right) + \frac{2\mathcal{C}_3 K^{1/2}}{k_n^{5/2}}.
\end{aligned}$$

Secondly, recall that

$$\|T_K^*(u) - T^*(u)\|_2^2 = \int \|\theta^{K*}(\mathbf{x}) - \theta^*(\mathbf{x})\|_\infty^2 \mathbb{P}(\mathbf{x}) \leqslant K_{\max} \sum_{\ell=1}^{K_{\max}} \|\mu(\mathcal{T}_\ell)\theta_\ell^{K*} - \theta_\ell^*\|_\infty^2,$$

where $\mu(\mathcal{T}_\ell) = \mathbb{P}(\mathbf{X} \in \mathcal{T}_\ell)$. Following the same idea as in the proof of Proposition 7.3.2, from Taylor's expansion, under Assumptions 7.1.4 and 7.3.1,

$$\|\theta_\ell^{K*} - \theta_\ell^*\|_\infty^2 \leqslant \mathfrak{C}_2^2(u)\frac{k_n^2}{n^2}\left(1 + c\gamma_{\max}\psi(u) + o(\psi(u))\right)^2.$$

Hence,

$$\begin{aligned}\|T_K^*(u) - T^*(u)\|_2^2 &\leqslant \mathfrak{C}_2^2(u)\frac{k_n^2}{n^2}(1 + c\gamma_{\max}\psi(u) + o(\psi(u)))^2 \sum_{\ell=1}^{K_{\max}} \mathbf{1}_{\mathbf{x}\in\mathcal{T}_\ell} \\ &\leqslant \mathfrak{C}_3(u)\frac{k_n^2}{n^2}\,.\end{aligned}$$

Finally,

$$\left[\|\widehat{T}(u) - T^*(u)\|_2^2\right] \leqslant \frac{\mathcal{C}_5 K_0(u)(\log k_n)^2}{k_n},$$

for some constant $\mathcal{C}_5$. .

A.2 Covering numbers

Lemma A.2.1. *Following the notations of the proof of Theorem A.1.3, the class of functions $\mathfrak{F}$ satisfies*

$$\mathcal{N}_\Phi(\varepsilon, \mathfrak{F}) \leqslant \frac{\mathfrak{C}_4 K^{4(d+1)(d+2)} \|\Phi\|_2^{\alpha_1} \sigma_n^\alpha}{\varepsilon^\alpha},$$

for some constants $\mathfrak{C}_4 > 0$ and $\alpha > 0$ (not depending on n nor K).

Démonstration. Let

$$\begin{aligned} g_\theta(z) &= -\frac{1}{\sigma} + \left(\frac{1}{\gamma} + 1\right)\frac{\gamma z}{\sigma^2(1 + \frac{z\gamma}{\sigma})}, \\ h_\theta(z) &= -\frac{1}{\gamma^2}\log\left(1 + \frac{z\gamma}{\sigma}\right) + \frac{\left(\frac{1}{\gamma} + 1\right) z}{\sigma + z\gamma}, \end{aligned}$$

for $z > 0$. For θ and θ' in $\mathcal{S} \times \Gamma$, we have (from a straightforward Taylor expansion),

$$|g_\theta(y - u) - g_{\theta'}(y - u)| \leqslant C|\gamma - \gamma'| + C'|\sigma - \sigma'|,$$

for some constants C and C'. More precisely, one can take

$$\begin{aligned} C &= \frac{6}{\gamma_{\min}^2 \sigma_{\min}}, \\ C' &= \frac{1}{\sigma_{\min}^2}\left(1 + 3\left\{1 + \frac{1}{\gamma_{\min}}\right\}\right). \end{aligned}$$

Next, observe that

$$|g_{\theta'}(y - u) - g_{\theta'}(y - u')| \leqslant C''|u - u'|,$$

where $C'' = 4\gamma_{\max}^2/[\gamma_{\min}\sigma^3]$. Which leads to

$$|g_\theta(y - u) - g_{\theta'}(y - u')| \leqslant C_g \max(\|\theta - \theta'\|_\infty, |u - u'|),$$

for some constant $C_g > 0$. Similarly,

$$|h_\theta(y - u) - h_{\theta'}(y - u)| \leqslant C_1(4 + \log(1 + wy))|\gamma - \gamma'| + C_2|\sigma - \sigma'|,$$

Next,

$$|h_{\theta'}(y-u) - h_{\theta'}(y-u')| \leqslant C_7|u-u'|,$$

where $C_7 = 5/(\gamma_{\min}\sigma_{\min})$, leading to, for some $C_h > 0$,

$$|h_{\theta}(y-u) - h_{\theta'}(y-u')| \leqslant C_h \max(\|\theta-\theta'\|_\infty, |u-u'|).$$

On the other hand,

$$|\phi(y-u,\theta) - \phi(y-u,\theta')| \leqslant \frac{1}{\gamma_{\min}^2}(2+\log(1+wy))|\gamma-\gamma'| + \frac{3}{\gamma_{\min}\sigma_{\min}}|\sigma-\sigma'|,$$

and

$$|\phi(y-u,\theta') - \phi(y-u',\theta')| \leqslant \frac{1}{\sigma_{\min}}|u-u'|.$$

Define $\mathfrak{F}_1 = \{g_\theta(\cdot - u) : \theta \in \mathcal{S}\times\Gamma, u \in [u_{\min}, u_{\max}]\}$, $\mathfrak{F}_2 = \{h_\theta(\cdot - u) : \theta \in \mathcal{S}\times\Gamma, u \in [u_{\min}, u_{\max}]\}$, and $\mathfrak{F}_3 = \{\phi(\cdot - u, \theta) : \theta \in \mathcal{S}\times\Gamma, u \in [u_{\min}, u_{\max}]\}$. From Example 19.7 in (Vaart 1998), we get, for $i = 1, ..., 3$,

$$N(\varepsilon, \mathfrak{F}_i) \leqslant F_i \|\Phi\|_2^{\alpha_1}\sigma_n^{\alpha_1}\varepsilon^{-\alpha_1},$$

for some $\alpha > 0$ and constants F_i.

On the other hand, let

$$\mathfrak{F}_4 = \{\mathbf{x} \mapsto \mathbf{1}_{\mathbf{x}\in\mathcal{T}_\ell} : \ell = 1, \ldots, K\},$$

and

$$\mathfrak{F}_5 = \{y \mapsto \mathbf{1}_{y>u} : u \in \mathcal{U}\}.$$

From Lemma 4 in (Lopez, Milhaud et Thérond 2016), we have $N(\varepsilon, \mathfrak{F}_4) \leqslant m^k K^{\alpha_2}\varepsilon^{-\alpha_2}$, where $\alpha_2 = 4(d+1)(d+2)$, and where k is the number of discrete components taking at most m modalities. On the other hand, from Example 19.6 in (Vaart 1998), $N(\varepsilon, \mathfrak{F}_5) \leqslant 2\varepsilon^{-2}$.

From Lemma A.1 (Einmahl et Mason 2005), we get, for $i = 1, \ldots, 3$,

$$N(\varepsilon, \mathfrak{F}_i\mathfrak{F}_4\mathfrak{F}_5) \leqslant \frac{4m^k K^{\alpha_2}\max(C_g, C_h)\|\Phi\|_2^{\alpha_1}\sigma_n^{\alpha_1}}{\varepsilon^{\alpha_1+\alpha_2+\alpha_3}}.$$

Multiplying $\mathfrak{F}_i\mathfrak{F}_4\mathfrak{F}_5$ by a single indicator function $\mathbf{1}_{\Phi(Y_i)\leqslant M_n}$ does not change the covering number, and the result follows. □

A.3 Technical Lemmas

Lemma A.3.1. *With $v_{\mathfrak{F}}$ defined in Proposition A.1.1,*

$$v_{\mathfrak{F}} \leqslant \frac{M_n^2 k_n}{n}.$$

Démonstration. We have

$$\begin{aligned} v_{\mathfrak{F}} &\leqslant \mathbb{E}\left[\Phi(Y)^2 \mathbf{1}_{Y\geqslant u_{\min}}\mathbf{1}_{\Phi(Y)\leqslant M_n}\right] \\ &\leqslant M_n^2 \mathbb{P}(Y \geqslant u_{\min}) = \frac{M_n^2 k_n}{n}. \end{aligned}$$

□

Lemma A.3.2. *Define, for* $j = 1, 2, 3$,

$$E_{j,n} = \mathbb{E}\left[\Phi(Y)^j \mathbf{1}_{\Phi(Y)\geqslant M_n} \mathbf{1}_{Y\geqslant u_{\min}}\right].$$

Under the assumptions of Theorem A.1.4,

$$E_{j,n} \leqslant \frac{\mathfrak{e}_j k_n^{1/2}}{n^{1/2} n^{\rho_0 \beta a_2/4}}.$$

Démonstration. Applying twice Cauchy-Schwarz inequality leads to

$$E_{j,n} \leqslant \mathbb{P}(Y \geqslant u_{\min})^{1/2} [\Phi(Y)^{2j} \mathbf{1}_{\Phi(Y)\geqslant M_n}]^{1/2} \leqslant \frac{k_n^{1/2}}{n^{1/2}} [\Phi(Y)^{4j}]^{1/4} \mathbb{P}(\Phi(Y) \geqslant M_n)^{1/4}.$$

Next, from Chernoff inequality,

$$\mathbb{P}(\Phi(Y) \geqslant M_n) \leqslant \exp(-\rho_0 M_n)[\exp(\rho_0 \Phi(Y))] \leqslant \frac{m_{\rho_0}}{n^{\rho_0 \beta a_2}}.$$

□

Annexe B

Présentation du détail des composantes du bâtiment de notre étude

Composantes Secondaires	Composantes Principales
Sous œuvre	Fondations - Ouvrage enterré - Sous-œuvre
Longrine	Fondations - Ouvrage enterré - Sous-œuvre
Micropieux	Fondations - Ouvrage enterré - Sous-œuvre
Micropieux et longrine	Fondations - Ouvrage enterré - Sous-œuvre
Injection résine	Fondations - Ouvrage enterré - Sous-œuvre
Micropieux et injection	Fondations - Ouvrage enterré - Sous-œuvre
Puits	Fondations - Ouvrage enterré - Sous-œuvre
Géomembrane	Fondations - Ouvrage enterré - Sous-œuvre
Forages et sondages	Fondations - Ouvrage enterré - Sous-œuvre
Autres réparations des fondations	Fondations - Ouvrage enterré - Sous-œuvre
Structure	Structure
Plancher courant	Structure
Dallage sur terre-plein	Structure
Façade	Structure
Revêtement extérieur de façade	Structure
Escalier intérieur	Structure
Charpente	Charpente - Couverture - Toiture
Couverture	Charpente - Couverture - Toiture
Etanchéité	Charpente - Couverture - Toiture
Menuiseries extérieures	Menuiserie
Vitrerie	Menuiserie
Stores et fermetures	Menuiserie
Menuiseries intérieures	Menuiserie
Menuiserie non spécifiée	Menuiserie
Autres menuiseries	Menuiserie
Agencement	Agencement - Isolation - Cloison
Isolation	Agencement - Isolation - Cloison
Plafond	Agencement - Isolation - Cloison

Cloison	Agencement - Isolation - Cloison
Peinture	Embellissement
Papier peint	Embellissement
Autres revêtements mur	Embellissement
Parquet	Embellissement
Carrelage	Embellissement
Autres revêtements sol	Embellissement
Embellissement non spécifié	Embellissement
Gaz	Réseaux intérieurs - Autres équipements
Electricité	Réseaux intérieurs - Autres équipements
Plomberie et sanitaire	Réseaux intérieurs - Autres équipements
Autres réseaux intérieurs	Réseaux intérieurs - Autres équipements
Autres équipements	Réseaux intérieurs - Autres équipements
Climatisation	Equipement de génie climatique
Chauffage	Equipement de génie climatique
Ventilation	Equipement de génie climatique
Cheminée	Equipement de génie climatique
Autre équipement de génie climatique	Equipement de génie climatique
Voirie	Voirie et Réseaux Divers
Réseaux divers	Voirie et Réseaux Divers
Terrasse	Voirie et Réseaux Divers
Ouvrages de soutènements	Voirie et Réseaux Divers
Jardin	Autres ouvrages extérieurs
Piscine	Autres ouvrages extérieurs
Portail	Autres ouvrages extérieurs
Clôture	Autres ouvrages extérieurs
Véranda	Autres ouvrages extérieurs
Dépendances diverses	Autres ouvrages extérieurs
Démolition déblais	Mesures Conservatoires
Décontamination	Mesures Conservatoires
Nettoyage	Mesures Conservatoires
Assèchement	Mesures Conservatoires
Frais divers	Mesures Conservatoires
Intervention des secours	Mesures Conservatoires
Traitement désamiantage	Mesures Conservatoires
Autres mesures conservatoires	Mesures Conservatoires
Maitrise d'œuvre, ingénierie	Etude géotechniques, maitrise d'œuvre, ingénierie
Etude de sol	Etude géotechniques, maitrise d'œuvre, ingénierie
Mobilier divers	Mobilier
Electroménager / Blanc	Mobilier
Tv Hifi Vidéo / Brun	Mobilier
Matériel informatique / Gris	Mobilier
Linge / Vêtements	Mobilier
Objets de valeur	Mobilier
Alimentation	Mobilier
Autres mobiliers	Mobilier
Matériel et stock (pro)	Mobilier
Inclassable	Inclassable

Bibliographie

André, Camille (2013). "Analyse des dommages liés aux submersions marines et évaluation des coûts induits aux habitations à partir de données d'assurance : perspectives apportées par les tempêtes Johanna (2008) et Xynthia (2010)". Thèse de doct. Université de Bretagne occidentale-Brest.

Antonio, Katrien et Richard Plat (2014). "Micro-level stochastic loss reserving for general insurance". In : *Scandinavian Actuarial Journal* 2014.7, p. 649-669.

Arnold, Céline (2018). *Le parc de logements en France au 1er janvier 2018*. Rapp. tech. INSEE.

Assadollahi, Hossein (2019). "The impact of climatic events and drought on the shrinkage and swelling phenomenon of clayey soils interacting with constructions". Thèse de doct. Université de Strasbourg.

Avant de construire – Prendre en compte les risques du terrain (2014). Rapp. tech. Agence Qualité Construction.

Baillargeon, Jean-Thomas, Luc Lamontagne et Etienne Marceau (2021). "Mining actuarial risk predictors in accident descriptions using recurrent neural networks". In : *Risks* 9.1, p. 7.

Balkema, August A et Laurens De Haan (1974). "Residual life time at great age". In : *The Annals of probability* 2.5, p. 792-804.

Baudry, Maximilien (2020). "Quelques problèmes d'apprentissage statistique en présence de données incomplètes". Thèse de doct. Lyon.

Beirlant, Jan et Yuri Goegebeur (2003). "Regression with response distributions of Pareto-type". In : *Computational statistics & data analysis* 42.4, p. 595-619.

Beirlant, Jan et Yuri Goegebeur (2004). "Local polynomial maximum likelihood estimation for Pareto-type distributions". In : *Journal of Multivariate Analysis* 89.1, p. 97-118.

Berthet, Claude, Jean Dessens et José Luis Sánchez (2011). "Regional and yearly variations of hail frequency and intensity in France". In : *Atmospheric Research* 100.4, p. 391-400.

Bikel, Daniel M et al. (1998). "Nymble : a high-performance learning name-finder". In : *arXiv preprint*.

Bojanowski, Piotr et al. (2017). "Enriching word vectors with subword information". In : *Transactions of the association for computational linguistics* 5, p. 135-146.

Bourguignon, David (2014). "Événements et territoires-le coût des inondations en France : analyses spatio-temporelles des dommages assurés". Thèse de doct. Université Paul Valéry-Montpellier III.

Bousquet, Nicolas et Pietro Bernardara (2021). *Extreme Value Theory with Applications to Natural Hazards : From Statistical Theory to Industrial Practice*. Springer Nature.

Breiman, Leo (1996). "Bagging predictors". In : *Machine learning* 24.2, p. 123-140.

Breiman, Leo (2001). "Random forests". In : *Machine Learning* 45, p. 5-32.

Breiman, Leo et al. (1984). *Classification and regression trees*. CRC press.

Brodin, Erik et Holger Rootzén (2009). "Univariate and bivariate GPD methods for predicting extreme wind storm losses". In : *Insurance : Mathematics and Economics* 44.3, p. 345-356.

Brownlee, Jason (2020). *Imbalanced Classification with Python : Better Metrics, Balance Skewed Classes, Cost-Sensitive Learning*. Machine Learning Mastery.

Bühlmann, Hans et Alois Gisler (2005). *A course in credibility theory and its applications*. T. 317. Springer.

Cartographie de l'exposition des maisons individuelles au retrait-gonflement des argiles (2021). Rapp. tech. Commissariat général au développement durable (CGDD).

Charpentier, Arthur, Laurence Barry et Molly R. James (2021). "Insurance against natural catastrophes : balancing actuarial fairness and social solidarity". In : *The Geneva Papers on Risk and Insurance - Issues and Practice*.

Charpentier, Arthur, Molly Rose James et Hani Ali (2021). "Predicting Drought and Subsidence Risks in France". In : *Natural Hazards and Earth System Sciences Discussions*, p. 1-27.

Chaudhuri, Probal et Wei-Yin Loh (2002). "Non parametric estimation of conditional quantiles using quantile regression trees". In : *Bernoulli*, p. 561-576.

Chavez-Demoulin, Valérie, Paul Embrechts et Marius Hofert (2016). "An extreme value approach for modeling operational risk losses depending on covariates". In : *Journal of Risk and Insurance* 83.3, p. 735-776.

Chemitte, Jérôme (2008). "Adoption des technologies de l'information géographique et gestion des connaissances dans les organisations. Application à l'industrie de l'assurance pour la gestion des risques naturels". Thèse de doct. École Nationale Supérieure des Mines de Paris.

Chen, Tianqi et Carlos Guestrin (2016). "XGBoost : A Scalable Tree Boosting System". In : *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Francisco California USA : ACM, p. 785-794.

Chernozhukov, Victor (2005). "Extremal quantile regression". In : *The Annals of Statistics* 33.2, p. 806-839.

Chinchor, Nancy et Beth M Sundheim (1993). "MUC-5 evaluation metrics". In : *Fifth Message Understanding Conference (MUC-5) : Proceedings of a Conference Held in Baltimore, Maryland, August 25-27, 1993*.

Chneiweiss, Arnaud et José Bardaji (2020). *Les assureurs face au défi climatique*. Fondapol, Fondation pour l'innovation politique.

Chollet, Francois et al. (2015). *Keras*. url : `https://github.com/fchollet/keras`.

Circulaire numéro 91-50 du 12 février 1991 (1991). Rapp. tech. Ministère de la Transition écologique.

Coles, Stuart et al. (2001). *An introduction to statistical modeling of extreme values*. T. 208. Springer.

Collobert, Ronan et al. (2011). "Natural language processing (almost) from scratch". In : *Journal of machine learning research* 12, p. 2493-2537.

Contribution de Météo-France à l'analyse de la sécheresse géotechnique à l'attention de la Commission CatNat pour l'année 2019 (2020). Rapp. tech. Météo France, Direction de la Climatologie et des Services Climatiques.

Davison, Anthony C et Richard L Smith (1990). "Models for exceedances over high thresholds". In : *Journal of the Royal Statistical Society : Series B (Methodological)* 52.3, p. 393-425.

Denuit, Michel et Arthur Charpentier (2005). *Mathematiques de l'Assurance Non-Vie. Tome II : Tarification et Provisionnement*.

Dessens, J, C Berthet et JL Sanchez (2015). "Change in hailstone size distributions with an increase in the melting level height". In : *Atmospheric Research* 158, p. 245-253.

Devlin, Jacob et al. (2018). "Bert : Pre-training of deep bidirectional transformers for language understanding". In : *arXiv preprint arXiv :1810.04805*.

Ecoto, Geoffrey, Aurélien Bibaut et Antoine Chambaz (2021). "One-step ahead sequential Super Learning from short times series of many slightly dependent data, and anticipating the cost of natural disasters". In : *arXiv preprint arXiv :2107.13291*.

Einmahl, Uwe et David M Mason (2005). "Uniform in bandwidth consistency of kernel-type function estimators". In : *The Annals of Statistics* 33.3, p. 1380-1403.

Eleutério, Julian (2012). "Flood risk analysis : impact of uncertainty in hazard modelling and vulnerability assessments on damage estimations". Thèse de doct. Strasbourg.

Ellingsworth, Martin et Dan Sullivan (2003). "Text mining improves business intelligence and predictive modeling in insurance". In : *Information Management* 13.7, p. 42.

Elman, Jeffrey L (1990). "Finding structure in time". In : *Cognitive science* 14.2, p. 179-211.

Embrechts, Paul, Claudia Klüppelberg et Thomas Mikosch (2013). *Modelling extremal events : for insurance and finance*. T. 33. Springer Science & Business Media.

Etude : Changement climatique et assurance à l'horizon 2040 (2021). Rapp. tech. France assureurs.

Face aux risques – Le retrait gonflement des argiles (2007). Rapp. tech. Direction Générale de l'Aménagement du Logement et de la Nature.

Farkas, Sébastien, Antoine Heranval et al. (2021). "Generalized Pareto Regression Trees for extreme events analysis". In : *arXiv preprint arXiv :2112.10409*.

Farkas, Sébastien, Olivier Lopez et Maud Thomas (2021). "Cyber claim analysis using Generalized Pareto regression trees with applications to insurance". In : *Insurance : Mathematics and Economics* 98, p. 92-105.

Feldman, Ronen, James Sanger et al. (2007). *The text mining handbook : advanced approaches in analyzing unstructured data*. Cambridge University Press.

Fisher, Ronald Aylmer et Leonard Henry Caleb Tippett (1928). "Limiting forms of the frequency distribution of the largest or smallest member of a sample". In : *Mathematical proceedings of the Cambridge philosophical society*. T. 24. 2. Cambridge University Press, p. 180-190.

Francis, Louise A (2006). "Taming text : An introduction to text mining". In : *Casualty Actuarial Society Forum*. Citeseer, p. 51-88.

Fréchet, Maurice (1927). "Sur la loi de probabilité de l'écart maximum". In : *Ann. Soc. Math. Polon.* 6, p. 93-116.

Friedman, Jerome, Trevor Hastie et Rob Tibshirani (2010). "Regularization paths for generalized linear models via coordinate descent". In : *Journal of statistical software* 33.1, p. 1.

Gérin, Sarah (2011). "Une démarche évaluative des Plans de Prévention des Risques dans le contexte de l'assurance des catastrophes naturelles : Contribution au changement de l'action publique de prévention". Thèse de doct. Université Paris-Diderot-Paris VII.

Gey, Servane et Elodie Nedelec (2005). "Model selection for CART regression trees". In : *IEEE Transactions on Information Theory* 51.2, p. 658-670.

Gnedenko, Boris (1943). "Sur la distribution limite du terme maximum d'une série aléatoire". In : *Annals of mathematics*, p. 423-453.

Goldberg, Yoav (2017). "Neural network methods for natural language processing". In : *Synthesis lectures on human language technologies* 10.1, p. 1-309.

Goodfellow, Ian, Yoshua Bengio et Aaron Courville (2016). *Deep Learning*. MIT Press.

Goussebaile, Arnaud (2016). "Prevention and insurance of natural disasters". Thèse de doct. Université Paris-Saclay (ComUE).

Grislain-Letrémy, Céline et Cédric Peinturier (2010). "Le régime d'assurance des catastrophes naturelles en France". In.

Guillier, Flora (2017). "Evaluation de la vulnérabilité aux inondations : Méthode expérimentale appliquée aux Programmes d'Action de Prévention des Inondations". Thèse de doct. Paris Est.

Guillou, Armelle et Patrick Willems (2006). "Application de la théorie des valeurs extrêmes en hydrologie". In : *Revue de statistique appliquée* 54.2, p. 5-31.

Gumbel, Emil Julius (1958). *Statistics of extremes*. Columbia University Press.

Habets, Florence et al. (2008). "The SAFRAN-ISBA-MODCOU hydrometeorological model applied over France". In : *Journal of Geophysical Research : Atmospheres* 113.D6.

Hall, Jim et Dimitri Solomatine (2008). "A framework for uncertainty analysis in flood risk management decisions". In : *International Journal of River Basin Management* 6.2, p. 85-98.

Hastie, Trevor et al. (2009). *The elements of statistical learning : data mining, inference, and prediction*. T. 2. Springer.

Heranval, Antoine, Olivier Lopez et Maud Thomas (2021). "Application of machine learning methods for cost prediction of drought in France".

Hochreiter, Sepp et Jürgen Schmidhuber (1997). "Long short-term memory". In : *Neural computation* 9.8, p. 1735-1780.

Honnibal, Matthew (2017). *SPACY'S ENTITY RECOGNITION MODEL : incremental parsing with Bloom embeddings and residual CNNs*. Youtube.

Inondations, s'informer pour mieux se protéger (2019). Rapp. tech.

James, Gareth et al. (2021). "Statistical learning". In : *An introduction to statistical learning*. Springer, p. 15-57.

Joulin, Armand et al. (2016). "Fasttext. zip : Compressing text classification models". In : *arXiv preprint arXiv :1612.03651*.

Katz, Richard W, Marc B Parlange et Philippe Naveau (2002). "Statistics of extremes in hydrology". In : *Advances in water resources* 25.8-12, p. 1287-1304.

Kellermann, Patric et al. (2020). "The object-specific flood damage database HOWAS 21". In : *Natural Hazards and Earth System Sciences* 20.9, p. 2503-2519.

Klein, Robert W. et Shaun Wang (sept. 2009). "Catastrophe Risk Financing in the United States and the European Union : A Comparative Analysis of Alternative Regulatory Approaches". In : *Journal of Risk and Insurance* 76.3, p. 607-637.

Kolyshkina, Inna et Marcel van Rooyen (2006). "Text mining for insurance claim cost prediction". In : *Data Mining*. Springer, p. 192-202.

Krizhevsky, Alex, Ilya Sutskever et Geoffrey E Hinton (2012). "ImageNet classification with deep convolutional neural networks". In : *Advances in neural information processing systems* 25.

Kusner, Matt et al. (2015). "From word embeddings to document distances". In : *International conference on machine learning*. PMLR, p. 957-966.

L'assurance des événements naturels en 2019 (2021). Rapp. tech. France Assureurs.

L'assurance des événements naturels en 2020 (2022). Rapp. tech. France Assureurs.

La Fondation de prévention des établissements cantonaux d'assurance en Suisse, Cahiers spéciaux de la MRN (2019). Rapp. tech.

Latruffe, Laure et Pierre Picard (2005). "Assurance des catastrophes naturelles : faut-il choisir entre prévention et solidarité ?" In : *Annales d'économie et de statistique*, p. 33-56.

Lavrač, Nada, Vid Podpečan et Marko Robnik-Šikonja (2021). *Representation Learning : Propositionalization and Embeddings*. Springer.

Le, Hang et al. (2019). "Flaubert : Unsupervised language model pre-training for french". In : *arXiv preprint arXiv :1912.05372*.

LeCun, Yann, Yoshua Bengio et al. (1995). "Convolutional networks for images, speech, and time series". In : *The handbook of brain theory and neural networks* 3361.10, p. 1995.

Lettre d'information de la Mission Risques Naturels 30 (2019). Rapp. tech. Mission Risques Naturels.

Lettre d'information de la Mission Risques Naturels 34 (2020). Rapp. tech. Mission Risques Naturels.

Lettre d'information de la Mission Risques Naturels 36 (2021). Rapp. tech. Mission Risques Naturels.

Li, Jing et al. (2020). "A survey on deep learning for named entity recognition". In : *IEEE Transactions on Knowledge and Data Engineering* 34.1, p. 50-70.

Loh, Wei-Yin (2011). "Classification and regression trees". In : *Wiley interdisciplinary reviews : data mining and knowledge discovery* 1.1, p. 14-23.

Loh, Wei-Yin (2014). "Fifty years of classification and regression trees". In : *International Statistical Review* 82.3, p. 329-348.

LOI numéro 2021-1837 du 28 décembre 2021 relative à l'indemnisation des catastrophes naturelles (2021).

LOI numéro 82-600 du 13 juillet 1982 relative à l'indemnisation des victimes de catastrophes naturelles (1982).

Lopez, Olivier, Xavier Milhaud et Pierre-E Thérond (2016). "Tree-based censored regression with applications in insurance". In : *Electronic journal of statistics* 10.2, p. 2685-2716.

Ly, Antoine (2019). "Algorithmes de machine learning en assurance : solvabilité, textmining, anonymisation et transparence". Thèse de doct. Paris Est.

Ly, Antoine, Benno Uthayasooriyar et Tingting Wang (2020). "A survey on natural language processing (nlp) and applications in insurance". In : *arXiv preprint arXiv :2010.00462*.

Mao, Gwladys (2019). "Estimation des coûts économiques des inondations par des approches de type physique sur exposition". Thèse de doct. Université de Lyon.

Marquardt, Donald W et Ronald D Snee (1975). "Ridge regression in practice". In : *The American Statistician* 29.1, p. 3-20.

Martin, Louis et al. (2019). "CamemBERT : a tasty French language model". In : *arXiv preprint arXiv :1911.03894*.

McKee, Thomas B, Nolan J Doesken, John Kleist et al. (1993). "The relationship of drought frequency and duration to time scales". In : *Proceedings of the 8th Conference on Applied Climatology*. T. 17. 22. Boston, p. 179-183.

Météo-France dans le dispositif CATNAT sécheresse. (2020). Rapp. tech. Météo France.

Mikolov, Tomas et al. (2013). "Efficient estimation of word representations in vector space". In : *arXiv preprint arXiv :1301.3781*.

Minaee, Shervin et al. (2021). "Deep learning–based text classification : a comprehensive review". In : *ACM Computing Surveys (CSUR)* 54.3, p. 1-40.

Mission d'enquête sur le régime d'indemnisation des victimes des catastrophes naturelles, Rapport de synthèse (2005). Rapp. tech. Inspection générale des finances (IGF).

Moncoulon, D et al. (2014). "Analysis of the French insurance market exposure to floods : a stochastic model combining river overflow and surface runoff". In : *Natural Hazards and Earth System Sciences* 14.9, p. 2469-2485.

Moncoulon, David (2014). "Proposition d'une méthode d'estimation de l'exposition financière aux inondations pour le marché de l'assurance en France : modélisation hydrologique et économique probabiliste spatialisée". Thèse de doct. Toulouse 3.

Moncoulon, David et Antoine Quantin (2013). "Modélisation des événements extrêmes d'inondation en France métropolitaine". In : *La Houille Blanche* 1, p. 22-26.

Moore, Ian Donald, RB Grayson et AR Ladson (1991). "Digital terrain modelling : a review of hydrological, geomorphological, and biological applications". In : *Hydrological processes* 5.1, p. 3-30.

Mornet, Alexandre (2015). "Contributions à l'évaluation des risques en assurance tempête et automobile". Thèse de doct. Université Claude Bernard-Lyon I.

Mowbray, Albert H (1914). "How extensive a payroll exposure is necessary to give a dependable pure premium". In : *Proceedings of the Casualty Actuarial society.* T. 1. 1, p. 24-30.

Nasiboglu, Resmiye et Mustafa Gencer (2021). "Comparison of Spacy and Stanford libraries's pre trainer deep learning models for named entity recognition". In : *Journal of Modern Technology and Engineering* 6.2, p. 104-111.

Nelder, John Ashworth et Robert WM Wedderburn (1972). "Generalized linear models". In : *Journal of the Royal Statistical Society : Series A (General)* 135.3, p. 370-384.

Nicolet, Pierrick, Marc Choffet et al. (2015). "Assessing hail risk for a building portfolio by generating stochastic events". In : *EGU General Assembly Conference Abstracts*, p. 11602.

Nicolet, Pierrick, Jérémie Voumard et al. (2014). "Analysis and modeling of a hail event consequences on a building portfolio". In : *EGU general assembly conference abstracts*, p. 10447.

Norberg, Ragnar (1993). "Prediction of outstanding liabilities in non-life insurance". In : *ASTIN Bulletin : The Journal of the IAA* 23.1, p. 95-115.

Norberg, Ragnar (1999). "Prediction of outstanding liabilities II. Model variations and extensions". In : *ASTIN Bulletin : The Journal of the IAA* 29.1, p. 5-25.

Pennington, Jeffrey, Richard Socher et Christopher D Manning (2014). "Glove : Global vectors for word representation". In : *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, p. 1532-1543.

Pickands III, James (1975). "Statistical inference using extreme order statistics". In : *the Annals of Statistics*, p. 119-131.

Piette, Pierrick (2019). "Contributions de l'Apprentissage Statistique à l'Actuariat et la Gestion des Risques Financiers". Thèse de doct. Université de Lyon.

Pigeon, Mathieu, Katrien Antonio et Michel Denuit (2014). "Individual loss reserving using paid–incurred data". In : *Insurance : Mathematics and Economics* 58, p. 121-131.

Pörtner, Hans O. et al. (2022). "Climate change 2022 : impacts, adaptation and vulnerability". In.

Pritchard, O. G., S. H. Hallett et T. S. Farewell (déc. 2015). "Probabilistic soil moisture projections to assess Great Britain's future clay-related subsidence hazard". In : *Climatic Change* 133.4, p. 635-650.

Procédure de reconnaissance de l'état de catastrophe naturelle - Révision des critères permettant de caractériser l'intensité des épisodes de sécheresse-réhydratation des sols à l'origine de mouvements de terrain différentiels. (2019). Rapp. tech. Ministère de l'interieur.

Rabiner, Lawrence R (1989). "A tutorial on hidden Markov models and selected applications in speech recognition". In : *Proceedings of the IEEE* 77.2, p. 257-286.

Rau, Lisa F (1991). "Extracting company names from text". In : *Proceedings the Seventh IEEE Conference on Artificial Intelligence Application.* IEEE Computer Society, p. 29-30.

Référentiels de résilience du bâti aux aléas naturels (2022). Rapp. tech. Mission Risques Naturels.

Resnick, Sidney I (1997). "Discussion of the Danish data on large fire insurance losses". In : *ASTIN Bulletin : The Journal of the IAA* 27.1, p. 139-151.

Rijsbergen, CV (1979). "Information retrieval 2nd ed Buttersworth". In : *London.*

Rootzén, Holger et Nader Tajvidi (1997). "Extreme value statistics and wind storm losses : a case study". In : *Scandinavian Actuarial Journal* 1997.1, p. 70-94.

Rosenblatt, Frank (1961). *Principles of neurodynamics. perceptrons and the theory of brain mechanisms.* Rapp. tech. Cornell Aeronautical Lab Inc Buffalo NY.

Sabban, Isaac Cohen, Olivier Lopez et Yann Mercuzot (2020). "Automatic analysis of insurance reports through deep neural networks to identify severe claims". In : *Annals of Actuarial Science*, p. 1-26.

Saito, Takaya et Marc Rehmsmeier (2015). "The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets". In : *PLOS ONE* 10.3, e0118432.

Salagnac, JL (2015). *Adaptation du cadre bâti aux conditions climatiques actuelles et futures : le cas des canicules*. Rapp. tech.

Salagnac, JL et al. (2014). *Impacts des inondations sur le cadre bâti et ses usagers, rapport final*. Rapp. tech.

Schapire, Robert E (1990). "The strength of weak learnability". In : *Machine learning* 5.2, p. 197-227.

Schmitt, Xavier et al. (2019). "A replicable comparison study of NER software : StanfordNLP, NLTK, OpenNLP, SpaCy, Gate". In : *2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS)*. IEEE, p. 338-343.

Sécheresse Géotechnique, de la connaissance de l'aléa à l'analyse de l'endommagement du bâti (2018). Rapp. tech. Mission Risques Naturels.

Silge, Julia et David Robinson (2017). *Text mining with R : A tidy approach*. O'Reilly Media, Inc.

Smith, James A (1987). "Estimating the upper tail of flood frequency distributions". In : *Water Resources Research* 23.8, p. 1657-1666.

Sols argileux et catastrophes naturelles (2022). Rapp. tech. Cour des comptes.

Su, Xiaogang, Morgan Wang et Juanjuan Fan (2004). "Maximum likelihood regression trees". In : *Journal of Computational and Graphical Statistics* 13.3, p. 586-598.

Surminski, Swenja et Annegret H Thieken (2017). "Promoting flood risk reduction : The role of insurance in Germany and England". In : *Earth's Future* 5.10, p. 979-1001.

Talagrand, Michel (1994). "Sharper bounds for Gaussian and empirical processes". In : *The Annals of Probability*, p. 28-76.

Tibshirani, Robert (1996). "Regression shrinkage and selection via the lasso". In : *Journal of the Royal Statistical Society : Series B (Methodological)* 58.1, p. 267-288.

Vaart, A. W. van der (1998). *Asymptotic statistics*. T. 3. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, p. xvi+443.

Vaswani, Ashish et al. (2017). "Attention is all you need". In : *Advances in neural information processing systems* 30.

Vidal, J.-P. et al. (2012). "Evolution of spatio-temporal drought characteristics : validation, projections and effect of adaptation scenarios". In : *Hydrology and Earth System Sciences* 16.8, p. 2935-2955.

Vidal, Jean-Philippe et Jean-Marc Moisselin (2011). *Impact du changement climatique sur les sécheresses en France*. Rapp. tech.

Vincent, M, E Plat et S Le Roy (2007). "Cartographie de l'aléa Retrait-Gonflement et Plans de Prévention des Risques". In : *Revue française de géotechnique* 120-121, p. 189-200.

Vinet, Freddy (2002). "La question du risque climatique en agriculture : le cas de la grêle en France". In : *Annales de géographie*. JSTOR, p. 592-613.

Vylomova, Ekaterina et al. (2015). "Take and took, gaggle and goose, book and read : Evaluating the utility of vector differences for lexical relation learning". In : *arXiv preprint arXiv :1509.01692*.

Wang, Huixia Judy, Deyuan Li et Xuming He (2012). "Estimation of high conditional quantiles for heavy-tailed distributions". In : *Journal of the American Statistical Association* 107.500, p. 1453-1464.

Wright, Marvin N. et Andreas Ziegler (2017). "ranger : A Fast Implementation of Random Forests for High Dimensional Data in C++ and R". In : *Journal of Statistical Software* 77.1. arXiv : 1508.04409.

Xu, Shuzhe (2021). "Applications of Modern NLP Techniques for Predictive Modeling in Actuarial Science". Thèse de doct. Middle Tennessee State University.

Zappa, Diego et al. (2021). "Text mining in insurance : From unstructured data to meaning". In : *Variance Journal*, p. 26122.

Zou, Hui et Trevor Hastie (2005). "Regularization and variable selection via the elastic net". In : *Journal of the Royal Statistical Society : Series B (statistical methodology)* 67.2, p. 301-320.

Contributions des données de l'assurance à l'étude des risques naturels
Application de méthodes d'apprentissage statistique pour l'évaluation de la nature et du coût des dommages assurés liés aux événements naturels en France

Résumé

Dans un contexte d'augmentation du coût des dommages assurés lié aux événements climatiques, d'un niveau déjà élevé, les assureurs ont vocation à participer à l'amélioration de la connaissance et de la prévention. Dans cette thèse, nous présentons des applications de méthodes d'apprentissage statistique pour l'évaluation de la nature et du coût des dommages assurés dû aux risques naturels en France. Nous commençons par étudier la sinistralité à l'échelle fine du bâti. Pour cela, nous analysons les données textuelles des rapports d'expertise. Ensuite, nous présentons des travaux portant sur l'estimation du coût de la sécheresse en France. Enfin, nous proposons une méthode estimation du coût des événements inondations, rapidement après leurs occurrences. Nous introduisons une méthode combinant des arbres de régression et la théorie des valeurs extrêmes. Nous ajoutons une application de la théorie de la crédibilité pour compléter cette estimation.

Mots clés : assurance, risques naturels, apprentissage statistique, valeurs extrêmes, analyse de texte, théorie de la crédibilité

Abstract

In the context of increasing costs of insured damages due to climatic events, at a level already high, insurers have to contribute to the improvement of natural risk knowledge and reduction. In this thesis, we present applications of statistical learning methods for the evaluation of the cost of insured damages due to natural hazards in France. In the first place, we will begin with a study of the damage distribution at the scale of the building. For this purpose, we analyze the textual data of the expert's reports. Then, we will present work on the estimation of the cost of drought in France. Finally, we propose a method for estimating the cost of flood events, quickly after their occurrence. We will introduce a method combining regression trees and extreme value theory. We will add an application of credibility theory to this estimation.

Keywords: insurance, natural hazards, statistical learning, extreme values theory, text analysis, credibility theory

---

Laboratoire de Probabilités, Statistique et Modélisation
Sorbonne Université – Campus Pierre et Marie Curie – 4 place Jussieu – 75005 Paris – France