Mémoire présenté le : 04/05/2021

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

Par : Alix NGUYEN-KHOA-MAN

Titre : Estimation des coûts des tempêtes sur la réassurance en France à partir de méthodes de machine learning

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

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

Membres présents du jury de l'Institut des Actuaires	*Signature*
M. BAGARRY	
A. HASSLER	
A. COULOUMY	
Membres présents du jury de l'ISFA	
E. MASIELLO	

Entreprise :
Nom : OdysseyRe

Signature :

Directeur de mémoire en entreprise :
Nom : C. DELELIS-FANIEN

Signature :

Invité :
Nom :

Signature :

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

Signature du responsable entreprise

Signature du candidat

Résumé

Résumé

Mots-clés : *Tempêtes*, *Taux de Dommages*, *Réassurance*, *Modèles Linéaires Généralisés*, *Apprentissage Statistique*, *Arbre de Décision*, *Méthodes Ensemblistes*, *Forêt Aléatoire*, *Gradient Boosting Machine*, *Provisionnement*, *Tarification*.

Les tempêtes européennes présentent un risque majeur pour les assureurs français qui ont eux-mêmes besoin de s'assurer par le biais de la réassurance. En effet, si la majorité des événements ont une charge assez limitée et ne touche pas, ou très peu, la réassurance, des événements extrêmes se produisent occasionnellement et peuvent mettre à mal les assureurs comme les réassureurs. Cela a été le cas notamment en décembre 1999 lorsque les tempêtes Lothar et Martin se sont succédé. Des sociétés spécialisées proposent aujourd'hui des modélisations pour ce type d'événements, cependant, elles ont leurs limites : elles sont très onéreuses, et peuvent demander beaucoup de temps. Par ailleurs, les trois principaux modèles disponibles sur le marché ont souvent des estimations très différentes.

Ce mémoire vise à construire un modèle d'estimation des coûts d'une tempête sur la réassurance en partant de données d'exposition et de vent, et à l'aide de différentes techniques d'apprentissage statistique. Nous utilisons, plus spécifiquement, des modèles linéaires, des arbres de décision et enfin des méthodes ensemblistes de *bagging* et de *boosting* pour estimer, dans un premier temps, le coût sur le marché français. Nous en déduisons par la suite les montants pour chacune de nos cédantes et finalement, le montant à la charge d'OdysseyRe.

Enfin, nous verrons les apports de ce modèle pour OdysseyRe aussi bien dans le cadre du provisionnement que dans celui de la tarification. Pour le calcul de provisions, il fournira une estimation rapide à partir de peu d'informations sur la tempête, cette première estimation pouvant ensuite être ajustée petit à petit avec les nouvelles informations disponibles. Pour la tarification, il offre une nouvelle méthode de calcul d'*as-if* prenant mieux en compte les évolutions de portefeuilles.

Abstract

Abstract

Keywords : *Windstorms*, *Damage Ratio*, *Reinsurance*, *Generalized Linear Models*, *Machine Learning*, *Decision Tree*, *Ensemble Methods*, *Random Forest*, *Gradient Boosting Machine*, *Reserving*, *Pricing*.

European windstorms are a major risk for French insurers which, in turn, also need to be insured via reinsurance. Indeed, even though most events have a limited loss amount that does not affect the reinsurance, or at least not much, some extreme events can occasionally happen and be particularly severe for insurers as well as reinsurers. It was notably the case in late December 1999 when windstorms Lothar and Martin both hit France within a few days. Nowadays, several specialized companies offer modeling for this type of events, however they have their limits: they are quite expensive and time consuming. Moreover, the three main CAT modelers available on the market often give three very different estimations.

This thesis aims to build a model for estimating the reinsurance loss of a windstorm from exposure and wind data, and using statistical learning methods. More specifically, we use linear models, decision trees and finally ensemble methods of bagging and boosting to estimate the insurance market loss. From that, we then deduce the loss for each of our ceding companies and finally, the reinsurance loss for OdysseyRe.

Lastly, we will highlight the importance of such a tool for OdysseyRe in the contexts of reserving as well as pricing. In reserving, this model will provide a quick estimation from little data on the windstorm. This estimation can then be adjusted little by little with every new piece of information made available. For pricing, it offers a new method for deriving the as-if losses that better takes into account the evolution of the portfolio.

Remerciements

Je tiens à remercier Christophe Delelis-Fanien, mon tuteur en entreprise, ainsi que Stéphane Loisel, mon tuteur académique, pour leur encadrement lors de la rédaction de ce mémoire.

Je remercie par ailleurs toute l'équipe actuariat : Nil Atamer, Laetitia Boudon, Charles Keita et Manon Tessier pour leur accueil chaleureux au sein d'OdysseyRe, et pour avoir partagé avec moi leurs connaissances ainsi que leur expérience depuis le début de mon stage et tout au long de mon alternance. Et plus particulièrement Aurélien Bonnet et Denis Frebourg, ainsi que l'équipe souscription France d'OdysseyRe, qui m'ont accompagné dans ma recherche de données et guidé plus spécifiquement sur ce sujet.

Enfin, mes derniers remerciements vont au corps enseignant de l'ISFA qui m'a formé au cours de ces trois dernières années et m'a mené à l'obtention de mon diplôme.

Sommaire

Introduction

Le marché de l'assurance en France est un marché majeur à l'échelle mondiale et un de ses risques les plus importants est la tempête.

Contrairement aux assureurs qui reçoivent des informations directement et assez rapidement de la part des assurés déclarant leurs sinistres, lorsqu'une tempête touche la France, les réassureurs ne reçoivent les estimations de chacune de leurs cédantes qu'après un certain délai.

Selon la période lors de laquelle l'événement survient, le réassureur ne peut pas toujours attendre les estimations des cédantes avant de devoir fixer ses réserves. C'est le cas notamment pour les tempêtes survenant en décembre, soit en plein milieu de la saison des tempêtes européennes, pour lesquelles un montant de réserve doit être fixé avant la fin de l'année. La bonne évaluation du montant de l'événement est alors particulièrement importante puisque sa sous-estimation peut étaler les mauvais résultats engendrés sur une nouvelle année.

Nous avons alors créé un premier outil empirique qui nous permet de déterminer une estimation du coût d'une tempête à notre part à partir des premières données disponibles qui sont les vitesses de vent et premiers dégâts observés ainsi qu'un montant marché. Pour chaque département, nous fixons alors une intensité locale de l'événement sur une échelle de 0, s'il n'est pas touché, à 5 pour les départements les plus violemment touchés. Avec ces intensités, nous répartissons ensuite les sinistres sur les différents départements puis, avec les parts de marché locales de chaque assureur français, sur chacune de nos cédantes. Nous pouvons ainsi calculer une estimation du coût de l'événement pour chacune de nos cédantes et enfin, en appliquant les conditions des traités de réassurance correspondant, en déduire une estimation du coût de l'événement à notre part.

L'objectif de ce mémoire est d'approfondir ce premier outil à l'aide de techniques actuarielles. Nous voulons ainsi construire un modèle nous permettant d'estimer directement et plus précisément le coût d'un événement sur chacun des départements à partir des intensités locales de celui-ci et avec l'aide de méthodes de *machine learning*.

Le Chapitre 1 de ce mémoire présente plus précisément le contexte de l'étude en commençant par la tempête en France et en assurance, et jusqu'à la réassurance et les spécificités des contrats de réassurance pour la tempête.

Le Chapitre 2 présente la théorie derrière les outils statistiques utilisés dans le cadre de ce mémoire et notamment les différentes méthodes de *machine learning* envisagées dont l'arbre de décision et les méthodes ensemblistes que sont la forêt aléatoire et le GBM.

Dans le Chapitre 3, nous constituons notre base de données sur laquelle seront basés tous nos modèles. Nous rassemblons ainsi des données internes et des données de la base *Industry Exposure and Loss Database* de la société PERILS afin de sélectionner nos variables finales.

Dans le Chapitre 4, nous construisons enfin deux modèles à l'aide des outils présentés dans le deuxième chapitre. Un premier modèle estime directement les taux de dommages et un second prédit d'abord la présence ou non de sinistres puis estime les taux de dommages lorsqu'il y a des sinistres.

Le Chapitre 5 compare les résultats obtenus avec nos deux modèles et présente les applications possibles de ceux-ci. La principale application est au provisionnement en reprenant notre premier outil sur l'exemple de la tempête Alex du 2 octobre 2020. Nous présentons également les apports de ce modèle en tarification.

Enfin, dans le Chapitre 6 nous envisageons une autre approche du problème utilisant plutôt la théorie des valeurs extrêmes sur le modèle des travaux effectués par A. Mornet et al. [12].

Chapitre 1

Contexte

1.1 La tempête

1.1.1 Le phénomène météorologique

Le vent désigne les mouvements de l'air et est le résultat de diverses forces sur les particules aériennes. Parmi ces forces, la plus importante est la force du gradient de pression.

La force du gradient de pression

L'air à la surface de la Terre est chauffé inégalement par le soleil en raison de la forme sphérique de la planète et des différences à sa surface. Les continents s'échauffent plus que les océans et les déserts plus que les forêts par exemple.

L'air chaud se dilate et s'élève résultant en une zone de faible pression aussi appelée une dépression. Inversement, l'air froid descend formant une zone de haute pression appelée un anticyclone.

Autour d'une dépression (respectivement d'un anticyclone) la pression augmente (respectivement diminue) en s'éloignant du centre dessinant ainsi des lignes de pression égales, ou isobares. Le vent se forme le long de ces isobares et sa vitesse augmente proportionnellement au gradient de pression qui représente la variation de pression dans l'air en pascals par mètre.

Les autres forces

D'autres forces [2] jouent, à un plus petit degré, sur la vitesse et la direction du vent. La force de Coriolis qui est générée par la rotation de la Terre dévie la trajectoire des vents. La force centripète augmente légèrement les vitesses de vent proches des centres de haute pression. Enfin, selon les reliefs, la force de frottement peut ralentir et dévier les vents de façon non négligeables.

2. Ces autres forces sont présentées de façon plus détaillée en Annexe A.1.

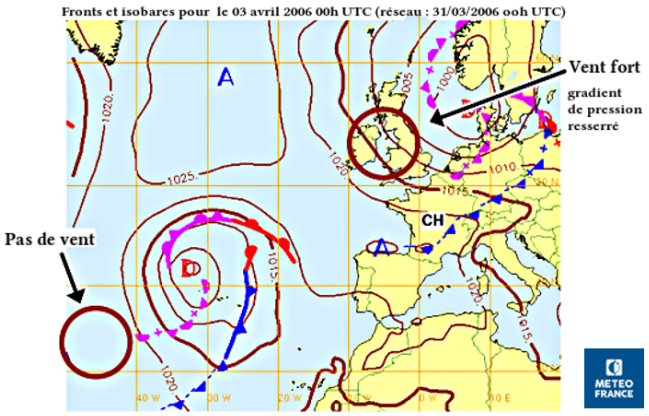


FIGURE 1.1 – Carte Météo France des isobares

Vents violents et tempêtes

C'est en météorologie marine que les termes de vents violents et tempête sont définis. Si le vent est décrit par deux mesures, sa direction et sa vitesse, ces appellations ne sont basées que sur les vitesses. La météorologie utilise par ailleurs deux mesures des vitesses de vent : le vent moyen, mesuré sur 10 minutes, et le vent instantané, mesuré sur environ 0,5 seconde, qui permet de connaître les vitesses des rafales.

Les seuils utilisés pour ces qualifications sont donnés par l'échelle de Beaufort [1]. Ainsi, les vents violents sont des vents de force 10 à 12 sur l'échelle de Beaufort, c'est-à-dire des vents dont la vitesse moyenne est supérieure à 89 km/h. Le terme tempête désigne alors la zone étendue de vents violents. Il est aussi utilisé pour désigner la dépression à l'origine de ces vents violents lorsque les vitesses de vent atteintes sont suffisamment élevées.

1.1.2 Les tempêtes historiquement en France

Les tempêtes observées en Europe proviennent principalement de l'Océan Atlantique et touche le Nord-Ouest du continent. Il arrive aussi que des vents violents se forment dans la Mer Méditerranée affectant ainsi plutôt le Sud-Ouest de l'Europe, à savoir l'Espagne, le Sud de la France et l'Italie.

Étant donnée sa localisation, la France est donc particulièrement exposée aux tempêtes européennes. Le plus souvent, elles touchent surtout la côte Ouest et le Nord de la France. Néanmoins, les tempêtes Lothar et Martin, qui ont marqué la mémoire collective des français en décembre 1999, montrent que tout le territoire peut être atteint. Lors de ces tempêtes, les vents avaient frôlé les 200 km/h, détruit près de 6% de la surface boisée française et fait 92 morts.

Cependant, ces deux événements sont très exceptionnels aussi bien en intensité qu'en étendue. Tous les ans, Météo France enregistre une quinzaine de tempêtes qui affectent la France métropolitaine, mais parmi celles-ci, seules quelques-unes peuvent être considérées comme fortes. Dans les années plus récentes, on retient surtout Klaus et Xynthia qui sont survenues en janvier 2009 et en février 2010 respectivement.

1. L'échelle de Beaufort est présentée en Annexe A.2.

Avant Lothar et Martin, les tempêtes de 1987 puis les tempêtes Daria, Herta et Vivian en 1990 révélaient déjà un besoin de surveiller les vents et la formation de tempêtes. L'établissement publique Météo France est ainsi créé en 1993 par le décret n° 93-861 du 18 juin avec pour mission, entre autres, de surveiller l'atmosphère, d'en prévoir les évolutions et de diffuser les informations correspondantes. Pour Météo France, le terme de tempête est utilisé lorsqu'au moins 5% des stations météorologiques du territoire enregistrent des vitesses de vent supérieures à 100 km/h pendant trois jours consécutifs.

La FIGURE 1.2 représente les vitesses de vent maximales enregistrées pour les 214 tempêtes les plus importantes survenues entre 1979 et 2015. On y retrouve les pics de vents en 1987, 1990 et 1999 qui correspondent aux tempêtes évoquées précédemment. On note aussi que ces vitesses de vent n'ont jamais été atteintes depuis, bien que les tempêtes Klaus et Xynthia s'en rapprochent en 2009 et 2010.

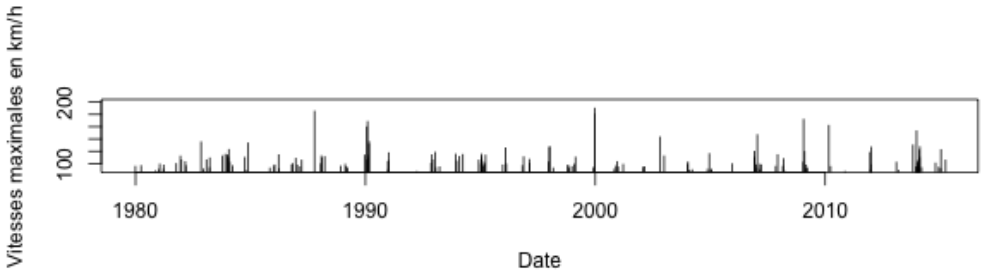


FIGURE 1.2 – Vitesses de vent maximales enregistrées pour les tempêtes de 1979 à 2015

1.1.3 Le risque tempête en assurance

Ces mêmes tempêtes de 1987 et 1990, ont aussi mis la lumière sur le besoin d'assurer le risque tempête. En effet, suite à la tempête d'octobre 1987, les Bretons notamment, ont été témoins du pouvoir destructif des tempêtes et de l'impact qu'elles peuvent avoir sur leur vie quotidienne. On pouvait ainsi lire dans la Une du Monde du 21 octobre 1987 : *"Dans les campagnes, on croirait être revenu un demi-siècle en arrière. Plus d'électricité, plus de télévision, plus de téléphone, plus d'essence et, souvent, plus d'eau : voilà dans quelles conditions vivent des centaines de milliers de ruraux"*. Si on pense d'abord à la destruction des logements et à l'impact sur les conditions de vie, l'activité des agriculteurs est elle-aussi très impactée. En effet, cette perte d'électricité rend inutilisables certaines machines et mène à des pertes de production et d'animaux qui ont besoin d'être conservés au frais, ou au contraire d'être réchauffés. Tout cela peut représenter des pertes économiques non négligeables pour un agriculteur.


Les conséquences d'un ouragan dévastateur

La Bretagne en état de choc

Plusieurs jours après l'ouragan qui a frappé les côtes ouest de l'Europe, dans la nuit du 15 au 16 octobre, les Bretons, qui étaient aux avant-postes de la tempête, sont encore en état de choc. Ils constatent que les dégâts subis constituent non seulement une perte économique considérable, mais un véritable désastre écologique.

Dans les campagnes on croirait être revenu un demi-siècle en arrière. Plus d'électricité, plus de télévision, plus de téléphone, plus d'essence et, souvent, plus d'eau : voilà dans quelles conditions vivent des centaines de milliers de ruraux.

Dans certaines communes où l'on transporte les groupes électrogènes d'un boulanger à l'autre, le pain est rationné : une seule boule par famille. Lundi, on comptait encore, dans l'ensemble des départements bretons, cent dix mille foyers sans courant. Au mieux, l'électricité reviendra jeudi, et dans certains hameaux samedi.

Les fermiers se trouvent soudain devant des problèmes invraisemblables, raconte Jean Le Naour, notre correspondant à Quimper. Comment traire trente, quarante, quatre-vingts vaches sans courant ? Alors, on s'y met, à la main, toute la nuit, à la lanterne, comme autrefois. Mais comment conserver le lait alors que les bacs de réfrigération ne fonctionnent plus ?

Chez les éleveurs, les porcelets meurent de froid à peine nés et les veaux accoutumés au lait tiède ne veulent plus s'alimenter. Sans parler des hangars abattus par centaines, des bâtiments sans toit, des animaux errant à présent à l'air libre, nuit et jour.

L'agriculture bretonne, parce qu'elle a voulu devenir l'une des plus performantes, est aussi extrêmement fragile. Dans toute l'Armorique, l'autarcie n'est plus qu'un souvenir.

Heureusement, la solidarité des campagnes a joué. Les congélateurs pleins de victuailles ont été transportés dans les bourgades disposant d'électricité. Des groupes électrogènes passent d'une ferme à l'autre et ronflent sans cesse de l'aube à minuit.

Aucun de ceux qui l'ont vécue n'oublieront cette nuit d'épouvante, raconte Jean Le Naour. Le vent soufflait à l'intérieur des terres, à plus de 200 kilomètres-heure.

(Lire la suite page 10.)

Le Monde

SCIENCES ET MÉDECINE

■ Prix Nobel : l'effondrement des dogmes génétiques ; Les travaux sur la supraconductivité ; La chimie « supramoléculaire ». ■ Les mirages du loch Ness. ■ Hausse sur le quotient intellectuel. ■ Aux Pays-Bas : mort douce pour des adolescents.

PAGES 17 à 19

LE MONDE DU VIN

PAGE 8

17 sch. ; Belgique, 30 fr. ; Canada, 1,75 $; Côte-d'Ivoire, 315 F CFA ; Danemark, 9 kr. ; Espagne, 145 pes. ; G.-B., 85 p. ; Portugal, 110 esc. ; Sénégal, 335 F CFA ; Suède, 11,50 cs. ; Suisse, 1,60 f. ; USA, 1,80 $; USA (West Coast), 1,75 $.

FIGURE 1.3 – Une du Monde, 21 octobre 1987

Ainsi, la garantie tempête, associée aux garanties grêle et neige dans ce qu'on appelle la garantie TGN (Tempête, Grêle, Neige), est rendue obligatoire par la loi n° 90-509 du 25 juin 1990. Tous les dommages causés par le vent doivent depuis être couverts pour tout contrat couvrant les dommages incendies.

Par ailleurs, la loi n° 82-600 du 13 juillet 1982, qui instaure une garantie catastrophe naturelle couverte par l'État pour les événements naturels d'intensité anormale, laissait un doute sur le cas des tempêtes. Suite à l'instauration de la garantie TGN, nous pouvions déjà supposer que les tempêtes n'entraient pas dans le cadre de la garantie catastrophe naturelle, cependant, en l'absence de précisions, des recours étaient encore possibles. C'est finalement la loi n° 92-665 du 16 juillet 1992 qui précise que cela concerne les dommages non assurables et donc pas les dommages des tempêtes qui sont déjà assurés par la garantie TGN. Les sinistres causés par les tempêtes sont donc à la charge des compagnies d'assurance et non de l'État. Cela implique aussi que tous les dommages causés par le vent sont couverts sans que l'État ait besoin de reconnaître le statut de catastrophe naturelle.

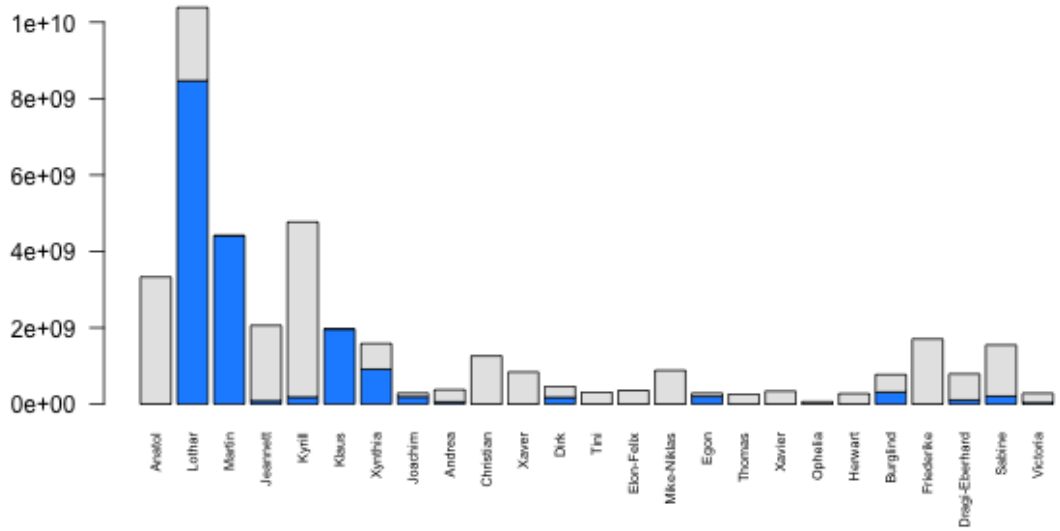


FIGURE 1.4 – Coûts indexés des tempêtes européennes depuis 1999

La FIGURE 1.4 présente les pertes assurantielles engendrées par les différentes tempêtes européennes depuis 1999 en Europe (en gris), et la part de la France (en bleu).

Ces montants sont indexés sur l'ICC [1] afin de prendre en compte l'évolution des coûts de la construction en France et représentent ainsi les montants en *as-if* [2] de l'inflation monétaire. Ils ne prennent cependant pas en compte les évolutions de l'exposition qui impacteraient aussi les coûts. Nous pouvons remarquer que les tempêtes de Noël 1999 sont inégalées aussi en dégâts causés. Elles représentent même près de 75% des coûts des tempêtes sur l'assurance en France sur les vingt dernières années.

1. l'Indice du Coût de la Construction (ICC), communément appelé le FFB, est un indice calculé trimestriellement par la Fédération Française du Bâtiment (FFB) prenant en compte les variations de coûts des différents éléments entrant dans la construction de bâtiments.
2. Les montants *as-if* sont les montants qui seraient engendrés si un événement identique se reproduisait dans le contexte actuel.

1.2 La réassurance

1.2.1 L'utilité de la réassurance

La garantie TGN, que les assureurs ont l'obligation de couvrir, augmente considérablement la volatilité de leurs portefeuilles. En effet, si plusieurs années consécutives peuvent être particulièrement peu touchées, comme cela a été le cas au début des années 2000, un événement comme Lothar et Martin peut causer en quelques jours plus de sinistres que ce qu'un assureur peut supporter. L'assureur a ainsi besoin à son tour de s'assurer en faisant appel à la réassurance.

Picard et Besson définissent l'opération de réassurance comme un *"contrat par lequel un réassureur (dit cessionnaire) vis-à-vis d'un assureur professionnel (dit cédant) qui répond seul et intégralement vis-à-vis des assurés des risques par lui assurés, prend en charge moyennant rémunération, tout ou partie de ce risque, s'engageant à lui rembourser dans les conditions déterminées, tout ou partie des sommes versées aux assurés à titre de sinistre"*. Ainsi, le réassureur n'a pas de responsabilité envers l'assuré directement mais seulement envers l'assureur avec qui il doit définir son engagement, les conditions de son intervention et le prix de cet engagement pour l'assureur.

La réassurance permet à l'assureur d'équilibrer son portefeuille et de lisser ses résultats en répartissant les charges annuelles exceptionnellement élevées sur plusieurs années. Par ailleurs, elle est prise en compte dans le calcul des marges de solvabilité et aide donc l'assureur à les respecter. Le réassureur, de son côté, peut équilibrer son portefeuille grâce à sa présence internationale et sur les multiples branches d'activité.

1.2.2 Les formes de réassurance

La réassurance peut prendre plusieurs formes. On distingue d'abord la réassurance proportionnelle de la réassurance non proportionnelle.

La réassurance proportionnelle

En réassurance proportionnelle, l'assureur verse au réassureur une part de sa prime en échange de laquelle le réassureur prend à sa charge cette même part des sinistres de l'assureur. Parmi les contrats de réassurance proportionnelle, on trouve les contrats de type Quote-Part, notés QP, pour lesquels un taux unique fixé est appliqué à tous les risques, et les contrats en excédent de plein pour lesquels le réassureur prend en charge la portion des risques dépassant un seuil appelé plein de rétention. Un taux est ainsi calculé pour chaque risque couvert par l'assureur. Après avoir déterminé ce taux, noté $x\%$, la prime P et les sinistres S du réassureur sont simplement calculés ainsi :

$$\begin{aligned} P_{\text{réassureur}} &= x\% \cdot P_{\text{assureur}} \\ S_{\text{réassureur}} &= x\% \cdot S_{\text{assureur}} \end{aligned}$$

La tarification des contrats proportionnels consiste alors à jouer sur des clauses telles que la commission de réassurance qui représente la participation du réassureur aux frais de gestion de l'assureur. Ces contrats peuvent aussi avoir une clause de participation aux bénéfices selon laquelle le réassureur doit rembourser à l'assureur une partie de ses bénéfices, et une clause de participation aux pertes selon laquelle l'assureur participe plus au paiement des sinistres dans le cas d'une sinistralité exceptionnellement élevée.

La réassurance non proportionnelle

En réassurance non proportionnelle, le montant à la charge du réassureur n'est pas proportionnel au montant du sinistre mais plutôt défini par une priorité et une portée.

Un contrat non proportionnel se note Portée XS Priorité.

La priorité fonctionne comme une franchise, c'est le montant du sinistre qui reste à la charge de l'assureur. Ainsi pour tous les sinistres inférieurs à la priorité, le réassureur ne paie rien. La portée, quant à elle, correspond au montant maximum que le réassureur s'engage à couvrir. Pour les sinistres dépassant la somme de la priorité et de la portée, aussi appelée le plafond, le réassureur ne prend en charge que la part du sinistre qui se trouve entre la priorité et le plafond. L'assureur doit donc prendre en charge la partie du sinistre dépassant le plafond en plus de la partie sous la priorité.

Le montant d'un sinistre à la charge du réassureur se calcule donc, à partir de la priorité et de la portée, avec la formule suivante [1] :

$$S_{\text{réassureur}} = \min\Big(\max(S_{\text{assureur}} - \text{Priorité}; 0); \text{Portée}\Big)$$

La prime de réassurance, quant à elle, est déterminée lors du processus de tarification et des négociations en fonction du risque couvert. Elle est généralement exprimée en pourcentage de la prime de l'assureur.

La FIGURE 1.5 représente la part des sinistres qui est à la charge de l'assureur et du réassureur dans les deux natures de réassurance. Le montant à la charge du réassureur évolue très différemment dans les deux cas en fonction du montant sinistre. Si, dans le cas non proportionnel, la charge peut être nulle, elle augmente très vite une fois que le sinistre dépasse la priorité. Dans le cas proportionnel, la charge augmente plus lentement, mais celle-ci n'a pas de limite et commence dès le premier euro sinistré.

1. Dans le cas de l'excédent de perte annuelle, présenté plus loin, c'est au Loss Ratio que cette formule est appliquée. Le montant sinistre à charge du réassureur se retrouve alors en multipliant par la prime.

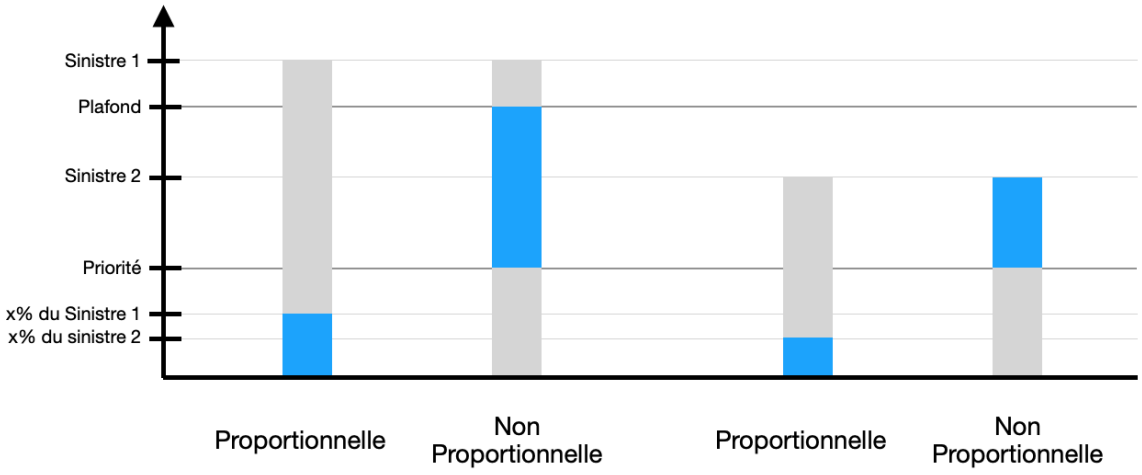


FIGURE 1.5 – Montants sinistre à charge de l'assureur (en gris) et du réassureur (en bleu)

Les traités de réassurance non proportionnelle prennent principalement trois formes : l'excédent de sinistre par risque, abrégé XS par risque, l'excédent de sinistre par événement, abrégé XS par événement, et l'excédent de perte annuelle.

Le XS par risque couvre l'assureur pour chaque risque individuel. Ainsi la priorité et la portée sont appliquées aux sinistres individuels. Les plus petits sinistres restent donc à la charge de l'assureur alors que le réassureur ne sera concerné que par les sinistres les plus importants.

Le XS par événement couvre l'assureur pour l'ensemble des polices sinistrées du fait d'une même cause. La somme de ces sinistres forme ainsi un événement dont la cause peut être, par exemple, une tempête, un incendie ou un attentat. Dans ce cas, la priorité et la portée sont appliquées à l'événement et non pas individuellement à chaque sinistre. Ces événements sont caractérisés par une très grande fréquence de sinistres de faibles intensités pour lesquels le XS par risque n'apporterait aucune protection. Les événements en eux-mêmes, au contraire, sont rares et leurs montants peuvent être très élevés.

Enfin, l'excédent de perte annuelle protège l'assureur contre les mauvais résultats en appliquant la priorité et la portée directement au ratio sinistres sur primes, noté S/P, annuel de la cédante. La priorité et la portée ne sont alors pas exprimées directement en montant sinistré, comme c'est le cas pour les XS par risque et par événement, mais plutôt sous la forme d'un S/P. Les montants dus par le réassureur sont alors obtenus en multipliant par la prime de la cédante.

Le risque tempête est principalement réassuré par des XS par événement. On trouve aussi des *excess aggregate* qui fonctionnent de manière similaire mais sans condition de même cause. L'événement considéré est alors la somme de sinistres similaires, correspondant généralement à une branche d'activité, sur l'année. L'assureur est alors protégé contre une grande fréquence d'événements semblables, de tempêtes par exemples, même si, individuellement, les événements ont une intensité relativement faible.

1.2.3 Les clauses pour un contrat non proportionnel

Les contrats non proportionnels peuvent être complétés par des clauses qui changent le calcul du montant à charge du réassureur.

AAD et AAL

Les clauses de franchise et de limite annuelles, ou AAD et AAL (pour *Annual Aggregate Deductible* et *Annual Aggregate Limit*), restreignent l'engagement du réassureur. L'AAD et l'AAL fonctionnent comme une franchise et une limite classique appliquées à la somme annuelle des montants dus par le réassureur, c'est-à-dire des montants sur lesquels ont déjà été appliquées la priorité et la portée prévues par le contrat.

Reconstitutions de garantie

Sans AAL, l'engagement du réassureur est illimité puisque le nombre de sinistres ou d'événements est illimité. L'AAL n'est cependant pas l'unique moyen de limiter cet engagement, les reconstitutions de garantie peuvent aussi être utilisées dans ce même but. Le principe des reconstitutions est de limiter l'engagement du réassureur à une fois la portée en donnant ensuite à l'assureur la possibilité de reconstituer sa garantie un certain nombre k de fois (en général entre 1 et 4 fois). Le réassureur s'engage donc à payer au maximum $(k+1)$ fois la portée.

La reconstitution de garantie peut être gratuite, l'assureur n'a alors pas de prime supplémentaire à payer. Dans ce cas son prix est en réalité compris dans le prix initial du contrat, c'est pourquoi on parle aussi de reconstitution prépayée. Si elle est payante, son prix est calculé à partir d'un taux fixé par le contrat, de la prime de réassurance et au prorata des capitaux absorbés, il est donc proportionnel au montant de la reconstitution.

$$\text{Prime de reconstitution} = \text{taux} \cdot \frac{\text{montant à reconstituer}}{\text{portée}} \cdot \text{prime de réassurance}$$

Les reconstitutions sont donc définies par leur nombre et un taux, pour chaque reconstitution, qui représente son prix et qui peut être nul (reconstitution gratuite), égal à 100%, inférieur ou supérieur à 100%.

Chapitre 2

Présentation des outils statistiques

2.1 Statistiques bivariées

La corrélation est une mesure qui met en évidence la relation entre deux variables continues X et Y. Elle est comprise entre -1 et 1. Une corrélation nulle indique que les variables sont indépendantes, alors qu'une corrélation s'éloignant de 0 révèle des variables de plus en plus liées positivement (si la corrélation est positive) ou négativement (si la corrélation est négative).

On retient généralement les trois mesures de corrélations suivantes :

- Le **coefficient de corrélation de Pearson** :

$$r = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

Avec $\mathrm{Cov}(X, Y)$ la covariance des variables X et Y et σ_X et σ_Y leurs écart-types respectifs.

- Le $\boldsymbol{\tau}$ **de Kendall** :

$$\tau = 2 \cdot \frac{N_{\mathrm{conc}} - N_{\mathrm{disc}}}{n(n-1)}$$

Avec N_{conc} et N_{disc} les nombres de paires concordantes et discordantes respectivement. Une paire d'observations (i, j) est dite concordante si l'ordre des deux observations selon la variable X et selon la variable Y coïncident, c'est-à-dire ($x_i \leq x_j$ et $y_i \leq y_j$) ou ($x_i \geq x_j$ et $y_i \geq y_j$), sinon elle est discordante.

- Le $\boldsymbol{\rho}$ **de Spearman** :

$$\begin{aligned} \rho &= \frac{\mathrm{Cov}(R_X, R_Y)}{\sigma_{R_X} \sigma_{R_Y}} \\ &= 1 - 6 \cdot \frac{\sum_{i=1}^{n} (R_{X_i} - R_{Y_i})}{n(n^2-1)} \qquad \text{si tous les rangs sont distincts pour chaque variable} \end{aligned}$$

Avec R_X et R_Y les variables rangs de X et Y respectivement. R_{X_i} désigne donc la position de X_i dans la liste ordonnée des $(X_i)_{1 \leq i \leq n}$.

Le coefficient de corrélation de Pearson ne met en avant que les relations linéaires entre deux variables, alors que le τ de Kendall et le ρ de Spearman, qui sont calculés avec les rangs, mettent aussi la lumière sur les relations non linéaires entre variables.

Le τ de Kendall est moins sensible aux erreurs et disparités dans les données que le ρ de Spearman, et est meilleur sur les plus petits échantillons. Néanmoins, les deux mesures mènent généralement aux mêmes conclusions, bien que le ρ de Spearman prenne habituellement des valeurs plus élevées que le τ de Kendall.

2.2 Méthodes d'apprentissage statistique

2.2.1 Apprentissage et test

Avant de construire un modèle d'apprentissage statistique, et lorsque la quantité de données le permet, il est préférable de partager les données aléatoirement en deux bases : une base d'apprentissage et une base de test. La base de données d'apprentissage est la sous-base sur laquelle est construit le modèle. La base de test est mise de côté lors de la phase d'apprentissage et n'est utilisée que pour tester le modèle une fois construit. Cela permet de contrôler si le modèle est bon sur des données jamais rencontrées. Dans ce mémoire, la base d'apprentissage représente $\frac{2}{3}$ de la base originale. La base de test correspond au $\frac{1}{3}$ restant.

L'un des enjeux de la construction d'un modèle d'apprentissage statistique est de choisir la bonne complexité pour le modèle. En effet, un modèle trop simple risque d'être incapable d'expliquer correctement les données d'apprentissage, cela s'appelle le sous-apprentissage. Au contraire, si le modèle est trop complexe, il correspond parfaitement aux données d'apprentissage mais réagit très mal face à de nouvelles données, c'est le sur-apprentissage. Nous choisirons nos paramètres en pensant à ces contraintes.

2.2.2 Modèles linéaires

Régression linéaire multiple

La régression linéaire multiple est la méthode la plus simple pour expliquer, c'est-à-dire prédire, une variable à partir de variables explicatives. Un tel modèle s'écrit de la façon suivante :

$$Y = \beta_0 + \sum_{j=1}^{p} \beta_j X_j + \epsilon$$

Dans cette équation, Y représente la variable à expliquer, les X_j sont les variables explicatives, les β_j sont les paramètres inconnus à estimer et ϵ est l'erreur du modèle.

Dans le cas où il n'y a qu'une seule variable explicative (régression linéaire simple), qui est plus facile à visualiser, β_0 et β_1 représentent respectivement l'ordonnée à l'origine et la pente d'une droite. La FIGURE 3.6 est un exemple de régression linéaire pour expliquer ln(`MDR`) à partir des vitesses de vent.

Ce modèle repose sur les hypothèses suivantes :

- Un même modèle est adapté à toutes les observations,
- La variable à expliquer est liée linéairement à chacune des variables explicatives,
- Espérance nulle des résidus, $\mathbb{E}(\epsilon_i) = 0$, hypothèse toujours vérifiée grâce à la présence de la constante β_0,
- Homoscédasticité des résidus, $\mathrm{Var}(\epsilon_i) = \sigma^2$ constant,
- Pour tout $i \neq j$, $\mathrm{Cov}(\epsilon_i, \epsilon_j) = 0$.

Les coefficients $(\beta_i)_{0 \leq i \leq k}$ sont estimés par la méthode des moindres carrés ordinaires qui consiste à minimiser la somme des carrés des erreurs :

$$S = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j X_{ij})^2$$

Le problème d'optimisation à résoudre s'écrit ainsi :

$$\min \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j X_{ij})^2$$
$$s.c. \begin{cases} \mathbb{E}(\epsilon_i) = 0 \\ \mathbb{E}(\epsilon_i X) = 0 \end{cases}$$

Dans le cas de la régression linéaire simple (n=1), la résolution de ce problème se fait en résolvant :

$$\frac{\partial S}{\partial \beta_0} = 0 \quad \text{et} \quad \frac{\partial S}{\partial \beta_1} = 0$$

Nous obtenons ainsi les coefficients :

$$\begin{cases} \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 &= \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \end{cases}$$

Où $\bar{x}$ et $\bar{y}$ sont respectivement la moyenne des réponses et la moyenne de la variable explicative.

Nous trouvons de la même manière les coefficients d'une régression linéaire multiple ($n > 1$).

Modèles linéaires généralisés

Les modèles de régression linéaire supposent une variance constante de la variable à expliquer, or ce n'est pas le cas pour tous les jeux de données.

Les modèles linéaires généralisés sont une généralisation des modèles linéaires applicable à des variables réponse dont la distribution appartient à la famille exponentielle.

Une distribution appartient à la famille exponentielle si sa densité peut s'écrire sous la forme :

$$f(y_i;\theta,\phi) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y;\theta)\right)$$

- θ et ϕ sont deux paramètres réels appelés, respectivement, paramètre de la moyenne (ou paramètre canonique) et paramètre de dispersion,
- a est une fonction non nulle définie sur $\mathbb{R}$,
- b est une fonction deux fois dérivable définie sur $\mathbb{R}$,
- c est une fonction définie sur $\mathbb{R}^2$.

Dans un modèle linéaire généralisé, ce n'est plus directement la variable réponse Y qui est expliquée, mais $g(Y)$, g étant appelée la fonction de lien. Cette fonction est choisie de sorte à lier la réponse Y au prédicteur linéaire η défini par :

$$\eta = \beta_0 + \sum_{j=1}^{p} \beta_j X_j$$

Le modèle devient alors :

$$g(Y) = \beta_0 + \sum_{j=1}^{p} \beta_j X_j + \epsilon$$

Bien que plusieurs fonctions puissent être choisies comme lien pour chacune des distributions de la famille exponentielle, la fonction de lien canonique est souvent préférée.

Par définition, la fonction de lien canonique est la fonction qui associe à la moyenne μ des réponses le paramètre canonique θ. La TABLE 2.1 synthétise les distributions usuelles de la famille exponentielle, leur lien canonique ainsi que les autres fonctions de lien acceptées par `R`.

Distributions Usuelles	Lien Canonique	Autres Liens Possibles
Normale	identité $g(\mu) = \mu$	log, inverse
Binomiale et quasi-Binomiale	logit $g(\mu) = \frac{\mu}{1-\mu}$	probit, cauchit, cloglog
Poisson et quasi-Poisson	log $g(\mu) = \ln(\mu)$	identité, racine carrée
Gamma	inverse $g(\mu) = \frac{1}{\mu}$	identité, log
Gaussienne Inverse	$g(\mu) = \frac{1}{\mu^2}$	identité, log, inverse

Table 2.1 – Distributions usuelles et leurs fonctions de lien

2.2.3 Arbres de décision

Les arbres de décision peuvent être utilisés aussi bien pour des problèmes de classification, pour lesquels la variable à expliquer est discrète, que de régression, pour lesquels la variable à expliquer est continue.

Leur principe est de segmenter l'espace formé par les variables explicatives en zones auxquelles sont associées une valeur pour la variable à expliquer. Cette valeur est la moyenne des valeurs observées dans la zone en question ou, dans le cas de la classification, la modalité la plus représentée dans la zone.

Un arbre de décision est construit en partant de la totalité de la base de données d'apprentissage, qui constitue la racine de l'arbre, en la divisant en deux sous-ensembles selon une règle de division choisie, puis en reproduisant l'opération de façon récursive jusqu'à atteindre une condition d'arrêt qui peut être sur la profondeur de l'arbre ou le nombre d'éléments par zone par exemple. Les zones obtenues à la fin sont appelées les feuilles de l'arbre, et toutes celles qui se trouvent entre la racine et les feuilles sont appelées nœuds.

À chaque étape, afin de choisir comment séparer les données, nous considérons toutes les paires de sous-ensembles :

$$R_1(j,s) = \{X | X_j < s\} \ et \ R_2(j,s) = \{X | X_j \geq s\}$$

Les X_j représentant les différentes variables explicatives.

À chacun de ces sous-ensembles est associée une valeur réponse $\hat{y}_{R_i(j,s)}$ qui est la moyenne des réponses des observations qui se trouvent dans le sous-ensemble $R_i(j,s)$, ou alors la modalité la plus représentée dans ce sous-ensemble.

Nous cherchons ensuite le meilleur couple (j,s), j représentant la variable explicative selon laquelle les données sont séparées et s la valeur qui délimite les deux sous-ensembles résultant.

Pour la régression, c'est le couple qui minimise le *Root Sum Squared* (RSS), ce qui revient à minimiser :

$$\sum_{i=1}^{n} \left((y_i - \hat{y}_{R_1(j,s)})^2 \mathbb{1}_{x_i \in R_1(j,s)} + (y_i - \hat{y}_{R_2(j,s)})^2 \mathbb{1}_{x_i \in R_2(j,s)} \right)$$

Pour la classification, c'est le couple qui minimise, au choix, l'erreur de classification, l'indice Gini ou l'Entropie qui sont donnés par les formules :

$$\begin{aligned} \text{Erreur de Classification} &= 1 - \max_k(\hat{p}_{mk}) \\ \text{Gini} &= \sum_{k=1}^{p} \hat{p}_{mk}(1 - \hat{p}_{mk}) \\ \text{Entropie} &= -\sum_{k=1}^{p} \hat{p}_{mk} \log(\hat{p}_{mk}) \end{aligned}$$

Où $\hat{p}_{mk}$ est la proportion d'observations du sous-ensemble m qui appartient à la classe k.

L'erreur de classification est une mesure d'erreur associée au nœud alors que l'indice Gini et l'Entropie représentent la pureté du nœud.

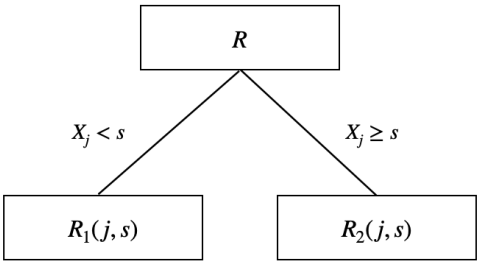


FIGURE 2.1 – Première étape de la construction d'un arbre

La FIGURE 2.1 représente le premier sous-arbre créé en partant de la racine R qui est l'ensemble des données d'apprentissage. Après avoir choisi le premier meilleur couple (j, s), les mêmes opérations sont réitérées sur les deux sous-ensembles obtenus et ce jusqu'à atteindre une condition d'arrêt.

Ce processus construit ainsi l'arbre maximal qui peut être de très grande taille. En raison de la complexité de cet arbre, le risque de sur-apprentissage est très élevé. Pour éviter cela, l'arbre maximal est ensuite élagué jusqu'à obtenir un sous-arbre de taille optimale.

2.2.4 Méthodes ensemblistes

Les méthodes ensemblistes de *machine learning* consistent à utiliser plusieurs fois un même algorithme d'apprentissage de base sur des échantillons différents et à combiner leurs résultats afin d'obtenir un meilleur pouvoir prédictif. Dans notre cas, l'algorithme d'apprentissage de base est l'arbre de décision. Plutôt que de construire un seul arbre très complexe avec un très grand risque de sur-apprentissage, nous construisons ainsi plusieurs arbres plus simples pour en déduire notre prédiction finale. Il existe deux types de méthodes ensemblistes : le *bagging* et le *boosting*.

Le *bagging*

Le *bagging*, pour *bootstrap aggregation*, consiste à construire simultanément plusieurs arbres à partir d'échantillons obtenus par *bootstrap*, puis à prendre comme réponse la moyenne des réponses données par chacun des arbres, ou la classe la plus représentée dans les réponses des différents arbres.

La forêt aléatoire est une méthode de *bagging* améliorée. Lors de la construction des arbres, le choix de la variable sur laquelle faire la division se fait, à chaque fois, sur un sous-ensemble aléatoire de m variables explicatives plutôt que sur leur ensemble complet. Cela permet d'éviter de construire des arbres trop similaires notamment dans le cas où une variable explicative se distingue des autres par un pouvoir prédictif nettement meilleur.

Le *boosting*

Le *boosting* est une méthode qui utilise aussi plusieurs arbres. En revanche, plutôt que de les construire simultanément, cette méthode consiste à construire les arbres de manière itérative de sorte que chaque nouvel arbre soit une amélioration du précédent. Par ailleurs, le *boosting* ne travaille pas directement sur la variable réponse y_i, mais sur les résidus r_i qui sont recalculés après chaque arbre. Le principe de l'algorithme est présenté ci-dessous :

- Initialisation : $\forall i \in [\![1, \mathrm{n}]\!],\ \hat{y}_i = 0$ et $r_i = y_i$
- $\forall k \in [\![1, \mathrm{N}]\!]$,
  - Construire un arbre de décision de profondeur d pour expliquer les résidus r_i
  - Ajouter à la réponse une version réduite des résidus prédits : $\hat{y}_i \leftarrow \hat{y}_i + \lambda \cdot \hat{r}_i$
  - Mettre à jour les résidus : $r_i \leftarrow r_i - \lambda \cdot \hat{r}_i$
- Sortie : $\forall i \in [\![1, \mathrm{n}]\!],\ \hat{y}_i = \sum_{k=1}^{n} \lambda \cdot \hat{r}_i^k$

$\hat{y}_i$ est la prédiction de y_i par l'algorithme de *boosting*, $\hat{r}_i$ est la prédiction des résidus par l'arbre de décision et $\hat{r}_i^k$ est plus précisément la prédiction du $\mathrm{k}^{\text{ème}}$ arbre.

N, d et λ sont des paramètres du modèle à choisir. Dans la fonction `gbm()` de `R`, ces paramètres sont appelés respectivement `n.trees`, `interaction.depth` et `shrinkage`. λ représente la vitesse d'apprentissage de l'algorithme, il sera fixé dans ce mémoire à 0,01.

2.3 Sélection de modèle

2.3.1 Les tests de significativité

Les tests de significativité permettent de mesurer, pour chaque coefficient d'un modèle de régression linéaire, s'il est statistiquement significatif et ainsi, l'importance de la variable concernée dans l'apport d'informations pour expliquer Y. Ces tests guident ainsi le choix des variables à utiliser, ou à retirer, pour ce type de modèles. Deux tests de significativité sont utilisés par `R` : le test de Student et le test de Wald.

Test de Student

Le test de Student, ou t-test, est un test statistique pour lequel la statistique suit une loi de Student lorsque l'hypothèse nulle est vérifiée.

Dans le cadre de la régression linéaire, l'hypothèse nulle est $H_0 : \beta_j = 0$, contre $H_1 : \beta_j \neq 0$. Nous testons la nullité d'un coefficient et donc si la variable associée à ce coefficient peut être retirée du modèle sans en affecter la qualité.

La statistique de ce test est :

$$T = \frac{\hat{\beta}_j}{\text{se}(\hat{\beta}_j)} \sim t_{n-p-1}$$

$\text{se}(\hat{\beta}_j) = \sqrt{\widehat{\text{Var}}(\hat{\beta}_j)}$ désigne l'erreur type (*standard error*), c'est-à-dire l'estimation de l'écart type de $\hat{\beta}_j$.

t_{n-p-1} est la loi de Student à $(n-p-1)$ degrés de liberté, n étant le nombre d'observations et p le nombre de variables explicatives.

Nous comparons alors la valeur prise par la statistique au quantile de la loi de Student de niveau α.

Si $|T| \geq t_{(n-p-1);(1-\frac{\alpha}{2})}$, l'hypothèse H_0 est rejetée. La variable considérée est donc significative.

Sous `R`, la fonction `lm()`, qui permet de construire des modèles linéaires, fournit les résultats de ce test pour chaque coefficient.

Pour chaque variable, cette fonction donne les valeurs suivantes :

- `Estimate` qui est la valeur estimée du coefficient β
- `Std. Error` qui est l'erreur type de $\hat{\beta}$
- `t value` qui est la valeur de la statistique, `t value` $= \frac{\texttt{Estimate}}{\texttt{Std. Error}}$
- `Pr(>|t|)` qui est la p-value du test, c'est-à-dire la probabilité que la loi de Student soit supérieure à la statistique, `Pr(>|t|)` $= P(t_{n-p-1} > |T|)$

La p-value est, de plus, marquée d'un symbole représentant un seuil. Tous les symboles, et les seuils correspondant, sont résumés dans la TABLE 2.3.1 ci-contre. Généralement, l'hypothèse H_0 peut être rejetée lorsque la p-value est inférieure à 5% et donc, lorsqu'elle est marquée d'au moins une étoile.

SYMBOLE	P-VALUE
***	< 0.001
**	< 0.01
*	< 0.05
.	< 0.1
	< 1

TABLE 2.2 – Notations R des niveaux de significativité

Test de Wald

Pour les modèles linéaires généralisés, dans le cas où le paramètre de dispersion ϕ de la loi de la famille exponentielle est connu, ce n'est pas un test de Student qui est effectué par R, mais un test de Wald. Dans ce cas ce n'est pas la `t value` qui est donnée mais la `z value`.

Le test de Wald repose sur le fait que le coefficient de régression $\hat{\beta}_j$ suit approximativement une loi normale lorsque n est suffisamment grand.

La statistique de ce test est alors :

$$Z = \frac{\hat{\beta}_j}{\text{se}(\hat{\beta}_j)} \sim N(0,1)$$

De la même manière que pour le test de Student, la fonction `glm()` donne les quatre valeurs suivantes :

- `Estimate`
- `Std. Error`
- `z value` $= \frac{\texttt{Estimate}}{\texttt{Std. Error}}$
- `Pr(>|z|)`

Les notations des différents niveaux de p-value sont aussi les mêmes que ceux présentés dans la TABLE 2.3.1.

Enfin, comme pour le test de Student, l'hypothèse H_0 est rejetée si $|Z| \geq N_{1-\frac{\alpha}{2}}$ ou si p-value $< 5\%$.

Si le paramètre ϕ n'est pas connu, en revanche, un estimateur de ϕ est utilisé et la loi de Student est souvent une meilleure estimation que la loi normale, notamment lorsque le nombre d'observations est limité. R effectue alors plutôt un test de Student comme dans le cas de la régression linéaire normale.

2.3.2 AIC et BIC

Le critère d'information d'Akaike ou AIC (*Akaike Information Criterion*) est un critère de sélection de modèle qui prend en compte à la fois la pertinence du modèle et sa parcimonie.

Pour un modèle linéaire classique, il est défini par :

$$\text{AIC} = n \ln\left(\frac{RSS}{n}\right) + 2(p+1)$$

Plus généralement, il est défini à partir de la log-vraisemblance l qui, pour une distribution de la famille exponentielle, s'écrit :

$$\begin{aligned} l &= \ln(L) \\ &= \ln\left(\prod_{i=1}^{n} f(y_i;\phi,\theta_i)\right) \\ &= \sum_{i=1}^{n} \ln(f(y_i;\phi,\theta_i) \\ &= \sum_{i=1}^{n} \left(\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i;\phi)\right) \end{aligned}$$

Pour un modèle linéaire généralisé, l'AIC s'écrit alors :

$$\text{AIC} = -2l + 2(p+1)$$

Le critère d'information Bayésienne ou BIC (*Bayesian Information Criterion*) est très proche de l'AIC, cependant, il tend à préférer des modèles plus parcimonieux que l'AIC en pénalisant plus les modèles ayant beaucoup de variables explicatives, avec un facteur $\ln(n)$ plutôt que 2.

De la même manière que pour l'AIC, pour un modèle linéaire normal, le BIC est défini par :

$$\text{BIC} = n \ln\left(\frac{RSS}{n}\right) + (p+1)\ln(n)$$

Et pour un modèle linéaire généralisé, il s'écrit :

$$\text{BIC} = -2l + (p+1)\ln(n)$$

L'AIC et le BIC ne représentent rien individuellement mais ils peuvent être comparés aux AIC et BIC d'autres modèles. Le meilleur modèle est alors celui qui minimise l'un de ces deux critères.

2.3.3 Matrice de confusion

Dans le cadre de la classification binaire, pour laquelle la réponse ne peut prendre que deux valeurs notées 0 et 1, la matrice de confusion permet de visualiser la performance du modèle. Elle représente sous forme de matrice les nombres de valeurs observées et prédites à 1 (*True Positives*), observées et prédites à 0 (*True Negatives*), observées à 0 et prédites à 1 (*False Positives*), observées à 1 et prédites à 0 (*False Negatives*). Pour une meilleure visualisation, elle est souvent colorée, une couleur plus sombre indiquant la présence d'un plus grand nombre de valeurs.

FIGURE 2.2 – Modèle de matrice de confusion

À partir de cette matrice, différentes mesures de la qualité de nos modèles peuvent être calculées.

- L'*Accuracy* représente la proportion de valeurs correctement prédites :

$$Accuracy = \frac{True\ Positives + True\ Negatives}{True\ Positives + True\ Negatives + False\ Positives + False\ Negatives}$$

- La *Precision* représente la précision du modèle lorsqu'il prédit un 1, c'est-à-dire parmi tous les 1 prédits quelle proportion sont réellement des 1. Une meilleure *Precision* permet de limiter le nombre de faux positifs qui sont très mauvais dans le cas de la détection de mails indésirables par exemple puisqu'ils pourraient causer la perte de mails importants.

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

- Le *Recall* représente la proportion de valeurs correctement prédites à 1 sur toutes les valeurs observées à 1. Un meilleur *Recall* permet de limiter le nombre de faux négatifs qui peuvent être dangereux dans le cas de la détection de fraude puisqu'une personne frauduleuse pourrait ne pas être détectée.

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

- Le *F1 Score* prend en considération ces deux problèmes. Il est aussi préféré à l'*Accuracy* notamment dans le cas où la distribution observée entre les deux classes est très déséquilibrée.

$$\begin{aligned} F1\ Score &= 2 * \frac{Precision * Recall}{Precision + Recall} \\ &= \frac{2 * True\ Positives}{2 * True\ Positives + False\ Positives + False\ Negatives} \end{aligned}$$

2.3.4 Courbe ROC et AUC

La fonction d'efficacité du récepteur, plus communément appelée la courbe ROC (pour *Receiver Operating Characteristic*), est une mesure de performance de modèles de classification binaire. C'est la courbe de la sensibilité par rapport à la spécificité, la sensibilité étant une autre appellation du *Recall* présenté précédemment, et la spécificité étant le taux de bonnes prédictions pour les valeurs observées à 0.

L'AUC (*Area Under the Curve*) est l'aire sous la courbe ROC.

Une courbe se rapprochant du coin supérieur droit, ou une AUC proche de 1, représente un meilleur modèle de classification.

2.3.5 RMSE et MAE

Pour les problèmes de régression, des mesures d'erreurs telles que le RMSE (*Root Mean Square Error)* et le MAE (*Mean Absolute Error)*, qui sont les plus communes, permettent de comparer les performances de différents modèles.

Le RMSE est la racine carrée de l'erreur quadratique moyenne :

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Le MAE est l'erreur absolue moyenne :

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Le RMSE est une mesure d'erreur quadratique, la présence du carré donne plus de poids aux grandes erreurs, et les pénalise donc plus, par rapport au MAE qui prend la valeur absolue des écarts.

Un meilleur modèle est un modèle qui minimise la mesure d'erreur choisie.

Chapitre 3

Constitution de la base de données

Avant de pouvoir construire un quelconque modèle, nous devons constituer une base de données. Les données que nous devons rassembler sont principalement les données d'expositions et de sinistres des différents assureurs présents sur le marché français, et plus spécifiquement celles qui concernent les tempêtes historiques. Nous compléterons ensuite ces données par des données non assurantielles.

3.1 Nos sources de données

3.1.1 Nos cédantes

En tant que réassureurs, et contrairement aux assureurs, nous n'avons pas accès directement aux données d'expositions et de sinistres puisque nous n'assurons pas directement les individus. Néanmoins, nos cédantes nous transmettent annuellement des informations sur leurs expositions ainsi que sur leurs sinistres dans le cadre de la campagne de renouvellement lors de laquelle nous avons besoin de ces données pour déterminer nos tarifs.

Par ailleurs, nous ne recevons pas des données détaillées par risque et par sinistre mais seulement des données agrégées, au mieux par zone CRESTA [2] et par type de risque, de la part des assureurs. Nous utiliserons donc nécessairement une approche de modélisation agrégée plutôt que détaillée.

Si les expositions nous sont transmises de manière assez régulière pour la France, nous permettant de reconstruire une base solide représentant près de la totalité du marché français et remontant jusqu'à 2008, il n'en est pas de même pour les sinistres. En effet, seules les plus grandes cédantes nous transmettent leurs sinistres avec un minimum de détails, et ce uniquement pour les événements les plus importants puisque les plus petits n'affectent pas la réassurance. Nous avons ainsi constitué une base rassemblant les montants sinistres par type de risque (agricoles, commerciaux, industriels et résidentiels) et par département des tempêtes Lothar, Martin, Klaus et Xynthia pour quatre cédantes. Nous avons reçu des sinistres d'autres événements et de la part d'autres cédantes cependant, nous n'avons qu'un montant global que nous pouvons difficilement intégrer à notre base de données.

2. En France, la norme CRESTA correspond aux départements.

3.1.2 Perils

Pour compléter nos données et obtenir une base plus solide, nous avons accès aux données payantes de la société Perils et plus précisément, à la base *Industry Exposure and Loss Database*. Perils récupère ses données directement auprès des assureurs, les anonymise et les rassemble dans une base représentant le marché. Pour la France, les assureurs transmettant leurs données à Perils représentent 60% du marché. Les chiffres sont ensuite rapportés à la totalité du marché en utilisant les parts de marché des assureurs basées sur leurs primes. La base d'expositions est ainsi mise à jour annuellement. Par ailleurs, pour chaque événement, Perils fournit une première estimation des pertes engendrées par pays lors de sa survenance, puis des estimations ajustées six semaines et trois mois après sa survenance. Enfin, si les pertes totales dépassent 200 millions€ en Europe, Perils transmet les sinistres détaillés par CRESTA et par type de risque ainsi qu'une mesure d'intensité de l'événement.

3.2 Les données

3.2.1 Les expositions

En ce qui concerne les expositions, nous disposons ainsi, d'une part, des nombres de risques de chacune de nos cédantes pour les années 2008 à 2019 et, d'autre part, de la base Perils des nombres de risques et sommes assurées marché qui couvre les années 2013 à 2020. Pour notre étude, Perils nous a également fourni les expositions correspondant aux événements entre 2009 et 2012. Ces deux sources nous permettent de vérifier la cohérence des données.

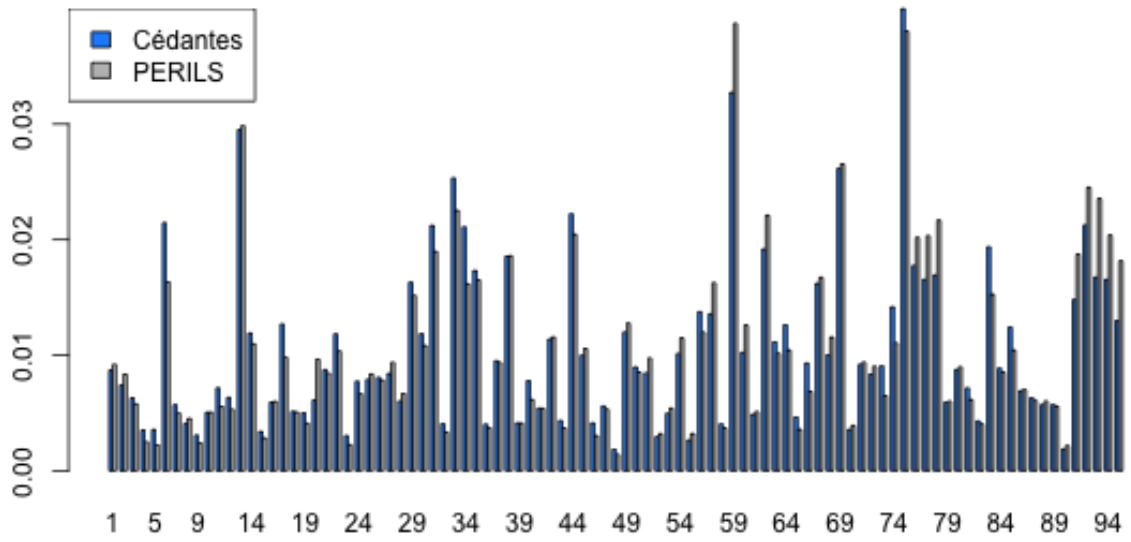


FIGURE 3.1 – Répartition des risques assurés par département en 2019

Les profils d'expositions par département, représentés sur la FIGURE 3.1, et par type de risque, représentés sur la FIGURE 3.2, sont similaires pour les données de nos cédantes et les données Perils, ce qui est rassurant. Néanmoins, on observe des petits décalages et la TABLE 3.1 révèle des écarts importants entre nos deux sources de données, même si les ordres de grandeurs restent les mêmes.

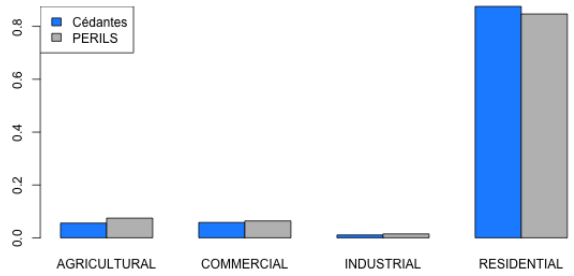


FIGURE 3.2 – Répartition des risques assurés par type de risque en 2019

ANNÉE	CÉDANTES	PERILS
2009	41 120 470	38 866 839
2010	41 355 733	38 866 839
2011	41 719 058	40 886 844
2012	42 538 191	40 886 844
2013	43 603 532	37 841 776
2014	44 025 068	38 646 576
2015	44 933 048	37 895 538
2016	45 426 087	39 713 060
2017	45 513 031	40 239 723
2018	46 511 817	42 074 644
2019	47 206 191	41 755 569

TABLE 3.1 – Nombres de risques

Ces décalages proviennent sûrement du décompte des risques qui peut beaucoup varier d'une personne à une autre. En effet, ce décompte peut être fait par risque ou par police alors qu'une même police peut couvrir plusieurs risques. Si, en général, les deux sont comptés par les assureurs, ce ne sont pas toujours les mêmes chiffres qui sont transmis.

Par ailleurs, en tant que réassureurs, nous demandons des chiffres annuels pour tarifer des contrats qui commencent presque toujours au 1er janvier (dans le cas de la France) alors que les assureurs, eux, souscrivent des contrats tout au long de l'année, et les nombres de risques assurés changent donc selon la date à laquelle ils sont transmis.

Enfin, les données que nous recevons de nos cédantes sont les chiffres réels et représentent près de la totalité du marché français alors que les chiffres de Perils ne représentent, initialement, que 60% du marché français et sont alors rapportés au marché complet en utilisant les parts de marché basées sur les primes des différents assureurs hors, la tarification des risques étant propre à chaque compagnie, ces parts ne correspondent généralement pas aux parts de marché basées sur les expositions.

Regardons maintenant l'évolution de ces chiffres dans le temps qui est représentée par la FIGURE 3.3. Alors que la croissance des expositions données par nos cédantes est linéaire, l'évolution des expositions de Perils est très irrégulière. Ces irrégularités peuvent être causées par l'évolution de l'échantillon source de Perils qui ne représentait au début, en 2010, que 42% du marché alors qu'il en représente aujourd'hui 60%.

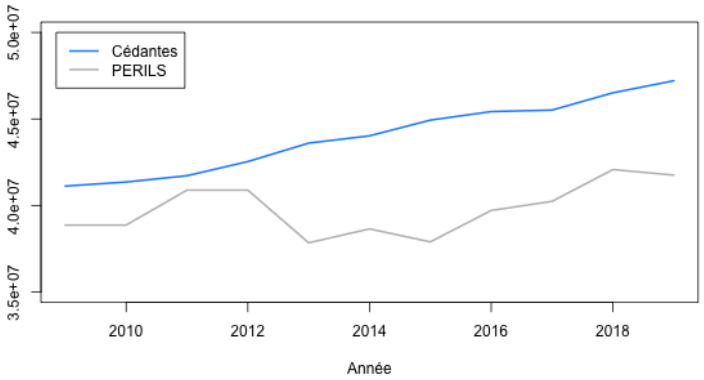


FIGURE 3.3 – Évolution des nombres de risques dans le temps

Malgré ces évolutions surprenantes, nous choisissons de travailler avec les expositions de Perils car elles correspondent aux sinistres de Perils et sont transformées de la même manière. Par ailleurs, dans les expositions Perils, nous avons en plus les sommes assurées totales ainsi que les valeurs assurées par type de garantie (bâtiment, contenu et perte d'exploitation). Nous indexons ces valeurs assurées sur l'ICC comme pour les sinistres afin de conserver les mêmes taux de dommages.

Nous notons qu'en France, l'évaluation des valeurs assurées est peu précise et varie d'un assureur à l'autre ce qui apportera un biais à notre modèle. En effet, la plupart des assureurs ne se basent que sur le nombre de pièces ou la surface habitable pour tarifer leurs contrats et n'évalue donc ni la valeur réelle du bien, ni le coût de sa reconstruction.

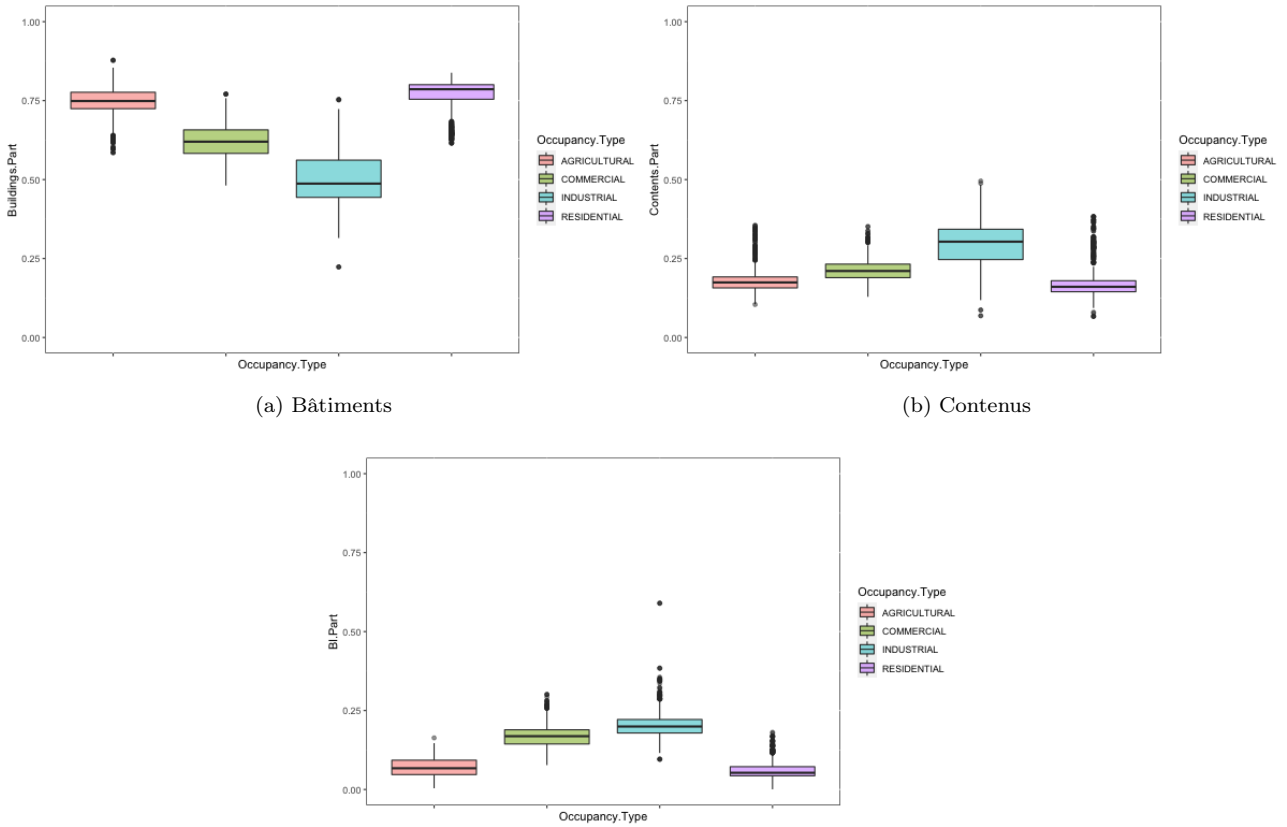
3.2.2 Les sinistres

La base sinistre est constituée des vingt-trois tempêtes européennes les plus importantes survenues entre 2009 et 2019. Seules neuf de ces tempêtes ont touché la France, celles-ci sont résumées dans la TABLE 3.2.

ÉVÉNEMENT	MONTANT	NOMBRE	SINISTRE MOYEN
Klaus	1 967 499 816	497 926	3 951
Xynthia	912 931 948	334 833	2 726
Joachim	191 779 021	95 178	2 015
Andrea	59 715 462	27 562	2 167
Dirk	177 657 708	71 309	2 491
Egon	200 566 859	78 624	2 551
Zeus	285 504 993	124 316	2 297
Burglind	314 313 033	126 832	2 478
DragiEberhard	107 621 176	42 352	2 541

TABLE 3.2 – Tempêtes historiques en France

Pour chacun de ces événements, nous avons les sinistres par département et par type de risque (agricoles, commerciaux, industriels et résidentiels). Par ailleurs, comme pour les sommes assurées, les sinistres sont détaillés par garantie, en bâtiment, contenu et perte d'exploitation. Cependant un nombre important de données manquantes nous oblige à mettre ces variables de côté.



(a) Bâtiments

(b) Contenus

(c) Perte d'exploitation

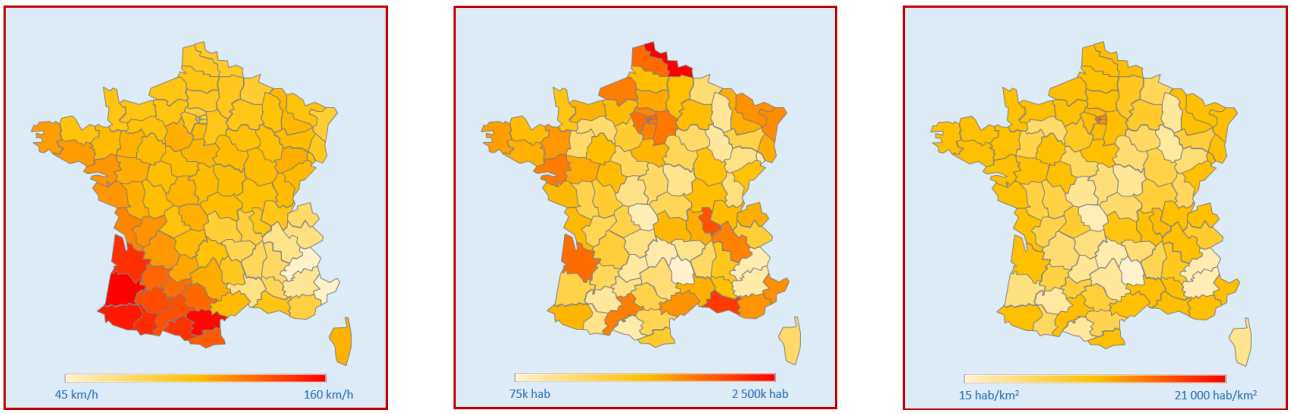
FIGURE 3.4 – Boîtes à moustaches des valeurs assurées par type de risque

Les boîtes à moustaches de la FIGURE 3.4 ci-dessus montrent que les sommes assurées concernent majoritairement des bâtiments. C'est très net pour les risques agricoles et résidentiels (en rouge et violet respectivement). Ça l'est moins pour les risques industriels (en bleu) mais le bâtiment représentent tout de même plus de 45% des sommes assurées dans 75% des cas et le contenu ou la perte d'exploitation ne dépassent que très exceptionnellement les 50% des sommes assurées. C'est pour cette raison que nous indexons les sinistres et les valeurs assurées sur l'ICC qui représente l'évolution du coût de la construction.

3.2.3 Les autres données

Nous complétons ces données assurantielles avec des données météorologiques ainsi que des données démographiques.

La FIGURE 3.5 représente ces données pour la tempête Klaus qui a touché la France le 24 janvier 2009.



(a) Vitesses de vent (b) Nombres d'habitants en 2009 (c) Densités de population en 2009

FIGURE 3.5 – Tempête Klaus de janvier 2009

Les données météorologiques nous sont fournies par Perils. Il s'agit, pour chaque département, d'une moyenne pondérée par la population des vitesses de vent maximales modélisées par la modèle Cosmo-7 de MeteoSwiss à partir de données atmosphérique mesurées au cours de la tempête. En traçant les ln des taux de dommages [1] par rapport à ces vitesses de vent dans la FIGURE 3.6, un lien important se dessine déjà entre les sinistres et les vitesses de vent qui auront donc sûrement un rôle important par la suite, lorsque nous construirons nos modèles.

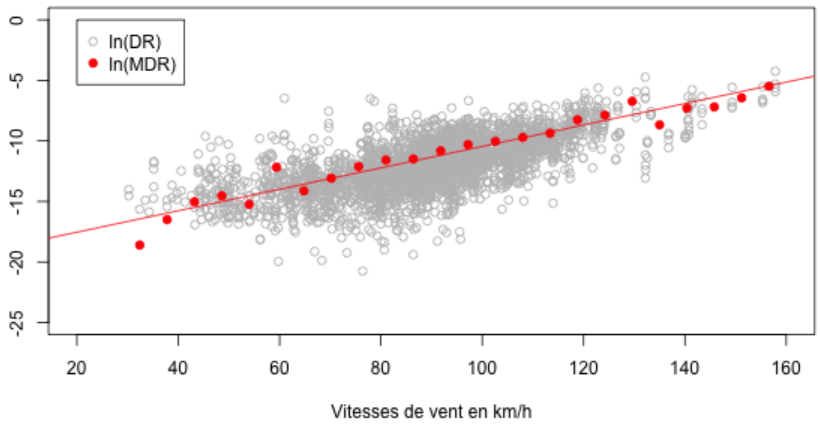


FIGURE 3.6 – Relation entre les taux de dommages et les vitesses de vent

Les données démographiques sont des données de l'INSEE qui sont en accès libre sur leur site, et plus spécifiquement les nombres d'habitants et densités de population. Ces données ont pour but d'apporter des informations sur les types de structures assurées que nous n'avons pas directement. En effet, nous pouvons imaginer que dans les zones les plus densément peuplées, comme Paris, ce sont plutôt des immeubles et des appartements qui sont assurés, alors que dans les zones plus rurales, ce sont plutôt des grandes maisons.

1. DR, pour *Damage Ratio*, désigne les taux de dommages individuels et MDR, pour *Mean Damage Ratio*, désigne les taux de dommages moyens par vitesse de vent

3.3 Sélection de variables

Dans cette partie, nous étudions les corrélations entre nos différentes variables, représentant les informations au niveau des CRESTA, afin de choisir les variables explicatives de nos modèles.

3.3.1 Variables d'exposition

Nous disposons de deux variables représentant l'exposition totale : les variables `TSI` (*Total Sum Insured*) et `Number.of.Policies` qui représentent respectivement les sommes assurées et les nombres de risques totaux.

Variable X	Variable Y	$r(X,Y)$
`TSI`	`Number.of.Policies`	0.9071035
`TSI`	`ASI`	-0.0424545

Table 3.3 – Coefficients de corrélation de Pearson

Sans surprise, avec un coefficient de corrélation de Pearson de $0,9$, ces deux variables sont fortement corrélées positivement. En effet, assurer un plus grand nombre de polices implique logiquement une augmentation des sommes assurées.

Nous préférons donc utiliser la variable `ASI`, qui est très peu corrélée à la variable `TSI`, plutôt que la variable `Number.of.Policies`. L'`ASI` (*Average Sum Insured*) est la valeur assurée moyenne par police. Elle est définie par :

$$\texttt{ASI} = \frac{\texttt{TSI}}{\texttt{Number.of.Policies}}$$

3.3.2 Variables de structure des portefeuilles

En plus des expositions totales, nous avons les sommes assurées détaillées par garantie en `Buildings.TSI`, `Contents.TSI` et `BI.TSI`, qui sont les valeurs assurées en bâtiment, en contenu ainsi qu'en perte d'exploitation respectivement.

$$\texttt{Buildings.Part} = \frac{\texttt{Buildings.TSI}}{\texttt{TSI}} \qquad \texttt{Contents.Part} = \frac{\texttt{Contents.TSI}}{\texttt{TSI}} \qquad \texttt{BI.Part} = \frac{\texttt{BI.TSI}}{\texttt{TSI}}$$

Nous créons les variables `Buildings.Part`, `Contents.Part` et `BI.Part` qui donnent plutôt les structures de portefeuilles. Ces variables représentent plus exactement les parts de `TSI` qui couvre chacune de ces trois garanties.

Ces trois variables ont été construites de sorte à avoir une somme égale à 1. Ainsi deux d'entre elles nous donnent déjà la totalité de l'information concernant la structure du portefeuille. Nous pouvons donc en retirer une que nous choisissons en calculant les corrélations présentées dans la Table 3.4.

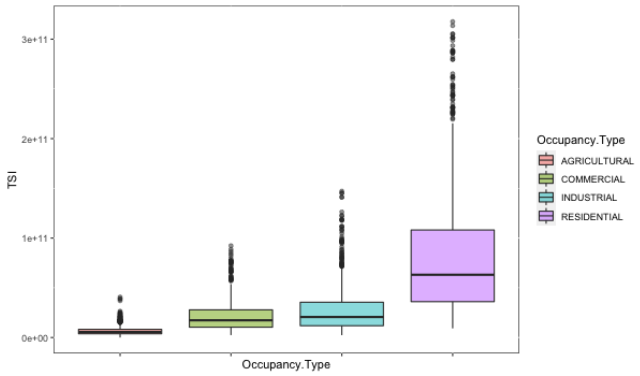
Variable X	Variable Y	$r(X, Y)$
`Buildings.Part`	`Contents.Part`	-0.872668
`Buildings.Part`	`BI.Part`	-0.878149
`Contents.Part`	`BI.Part`	0.532730

Table 3.4 – Coefficient de corrélation de Pearson

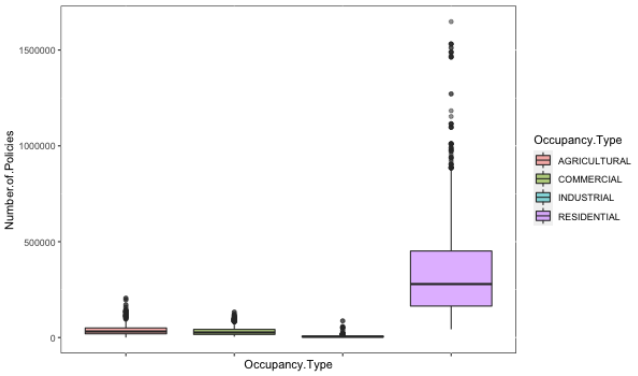
Nous choisissons de retirer la variable `Buildings.Part` qui est très fortement corrélée aux deux autres.

3.3.3 Variable `Occupancy.Type`

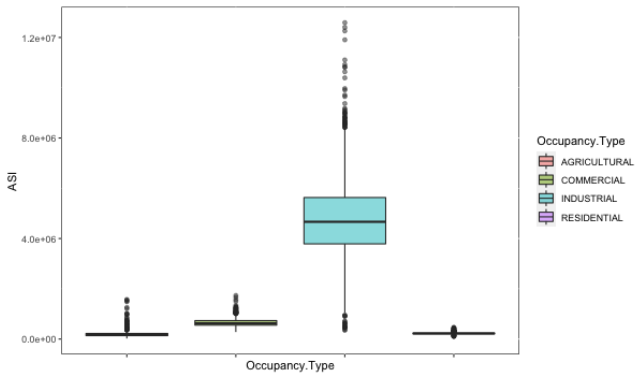
La variable `Occupancy.Type` représente le type de risque couvert. C'est une variable qualitative qui peut prendre quatre modalités : `AGRICULTURAL` pour les risques agricoles, `COMMERCIAL` pour les risques commerciaux, `INDUSTRIAL` pour les risques industriels et `RESIDENTIAL` pour les risques résidentiels. La Figure 3.7 complète la Figure 3.4 pour résumer les différentes variables quantitatives par types de risque sous la forme de boîtes à moustaches.



(a) Sommes assurées totales



(b) Nombres de risques



(c) Sommes assurées moyennes

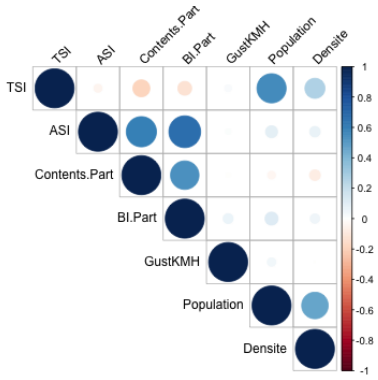
Figure 3.7 – Boîtes à moustaches des variables par type de risque

Nous retrouvons, dans les deux premières figures, le lien entre les variables `TSI` et `Number.of.Policies`. En effet, les deux variables ont des boîtes à moustaches similaires avec notamment des valeurs nettement plus élevées pour les risques résidentiels. Cela s'explique par l'obligation d'assurance contre les tempêtes. Tous les particuliers doivent s'assurer et donc, comme ils sont largement plus nombreux que les entreprises, un plus grand nombre de risques résidentiels sont assurés.

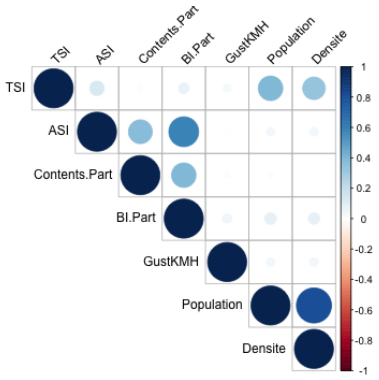
Les boîtes à moustaches des sommes assurées moyennes, en revanche, sont très différentes. Les `ASI` pour les risques résidentiels et agricoles sont très faibles par rapport aux risques industriels. Cela peut venir du fait que les risques résidentiels et agricoles concernent principalement du bâtiment (autour de 80% de la somme assurée totale) alors que les risques industriels sont plus partagés avec du contenu et de la perte d'exploitation qui représentent des sommes assurées plus élevées.

Les trois dernières variables `GustKMH`, `Population` et `Densite` ne sont pas représentées car elles ne varient pas pour les quatre types de risque. En effet, nous avons leurs valeurs par département et par événement mais la vitesse du vent, la population et la densité ne changent pas d'un type de risque à un autre. De plus, pour chaque événement et pour chaque département, nous avons une ligne pour chacun des types de risque.

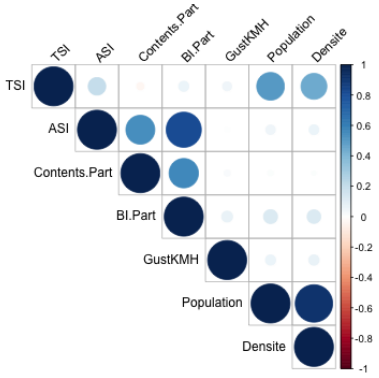
3.3.4 Variables sélectionnées



(a) Coefficient de corrélation de Pearson



(b) τ de Kendall



(c) ρ de Spearman

FIGURE 3.8 – Matrices de corrélations

Nous avons finalement sélectionné neuf variables explicatives dont les matrices de corrélations sont représentées dans la FIGURE 3.8 :

- Deux variables qualitatives : `CRESTA` et `Occupancy.Type`
- Sept variables quantitatives : `TSI`, `ASI`, `Contents.Part`, `BI.Part`, `GustKMH`, `Population` et `Densite`.

Les trois mesures de corrélation donnent des matrices de corrélations semblables. La plupart de nos variables sont peu corrélées. Nous pouvons tout de même noter deux corrélations remarquables : une première entre les variables `ASI` et `BI.Part` et une autre entre les variables `Population` et `Densite`.

La corrélation entre les variables `ASI` et `BI.Part` peut venir du fait que la somme assurée pour une police couvrant la perte d'exploitation est généralement plus élevée que pour une police couvrant le bâtiment. Par ailleurs, cette garantie coûte très cher et les plus petites entreprises n'ont souvent pas les moyens de se payer une telle couverture. Nous pouvons imaginer que, par un effet de régionalisation, les plus grosses entreprises se trouvent plutôt dans les zones urbaines alors que dans les départements plus ruraux, nous trouverons exclusivement des petites entreprises. Ainsi, une plus grande `BI.Part` représenterait la présence de grands groupes pour lesquels les sommes assurées par police sont plus élevées.

La corrélation entre les variables `Population` et `Densite`, quant à elle, semble assez naturelle étant donné la définition de la densité. Cependant le coefficient de corrélation de Pearson est assez faible indiquant une relation non linéaire entre les deux variables. En effet, les départements les plus peuplés étant aussi les plus petits en superficie, à savoir les départements d'Île-de-France, la densité augmente plutôt exponentiellement que linéairement avec la population.

Chapitre 4

Modélisation sous R

4.1 Premier modèle

Nous cherchons dans un premier temps à construire des modèles de régression qui expliquent les taux de dommages à partir des variables explicatives sélectionnées précédemment.

Les taux de dommages sont définis par : $\texttt{MDR} = \frac{\texttt{Loss}}{\texttt{TSI}}$.

Variable	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
MDR	0.00e0	0.00e0	0.00e0	2.82e-5	3.07e-6	1.43e-2
MDR|MDR > 0	1.00e-9	1.35e-6	6.39e-6	6.94e-5	3.05e-5	1.43e-2

Les taux de dommages varient entre 0% et $1,5\%$, mais plus de la moitié des valeurs sont nulles et seules $\frac{1}{4}$ des valeurs sont supérieures à $0,0003\%$. Si nous nous intéressons aux taux non nuls, plus de la moitié des valeurs se situent entre $0,0001\%$ et $0,0031\%$, les premier et troisième quartiles.

4.1.1 Régression linéaire multiple

Nous commençons par une régression linéaire multiple en utilisant la fonction `lm()` de R.

Plusieurs variables étant peu significatives, nous réalisons une régression dite *stepwise backward*, qui consiste à retirer une par une les variables les moins significatives, afin d'obtenir un modèle plus parcimonieux et un meilleur AIC. Nous retirons ainsi les variables `TSI`, `Contents.Part` et `BI.Part`.

Nous obtenons finalement les coefficients et les significativités de la Table 4.1.

Les modalités CRESTA1 et AGRICULTURAL des variables quantitatives `CRESTA` et `Occupancy.Type` n'apparaissent pas dans cette table car le département 1, c'est à dire l'Ain, et le type de risque agricoles sont choisies comme modalités de référence. Nous ne changeons pas ce choix de modalités de référence car tous les départements sont autant représentés dans notre base de données, et il en va de même pour les types de risque. Ces deux modalités conviennent donc comme référence.

Variable	Coef.	Sign.
(Intercept)	3.140e-04	*
CRESTA2	-7.028e-05	.
CRESTA3	-2.089e-04	**
CRESTA4	-3.222e-04	***
CRESTA5	-3.191e-04	**
CRESTA6	3.303e-04	***
CRESTA7	-2.082e-04	**
CRESTA8	-2.552e-04	***
CRESTA9	-2.959e-04	**
CRESTA10	-2.448e-04	***
...	...	...
COMMERCIAL	-3.252e-05	***
INDUSTRIAL	-1.098e-05	
RESIDENTIAL	-3.257e-05	***
ASI	-1.161e-11	**
GustKMH	2.530e-06	***
Population	-7.161e-10	***
Densite	1.340e-07	

Table 4.1 – Résultats du modèle linéaire (AIC)

Variable	Coef.	Sign.
(Intercept)	-1.019e-04	***
COMMERCIAL	-3.237e-05	***
INDUSTRIAL	1.850e-05	
RESIDENTIAL	-3.166e-05	***
ASI	-1.332e-11	***
GustKMH	2.238e-06	***

Table 4.2 – Résultats du modèle linéaire (BIC)

Les premiers coefficients indiquent ainsi que les taux de dommages seront plus élevés dans les Alpes-Maritimes (CRESTA6) que dans l'Ain, et plus faibles pour des risques commerciaux que pour des risques agricoles. Les derniers coefficients, pour les variables quantitatives `ASI`, `GustKMH`, `Population` et `Densite`, signifient que le taux de dommages croît avec la vitesse du vent et un peu moins vite avec la densité, et qu'il décroît à un degré encore plus faible avec les sommes assurées et le nombre d'habitants.

Nous avons également réalisé une régression *stepwise backward* en minimisant le critère BIC plutôt que l'AIC. Nous avons alors retiré, en plus, les variables `CRESTA`, `Population` et `Densite`. Le modèle obtenu est beaucoup plus parcimonieux puisqu'en enlevant la variable `CRESTA`, il retire 94 coefficients à déterminer. Les six coefficients restant sont donnés dans la Table 4.2 avec leurs significativités. Nous retenons plutôt le modèle linéaire minimisant l'AIC, car il minimise aussi largement le RMSE.

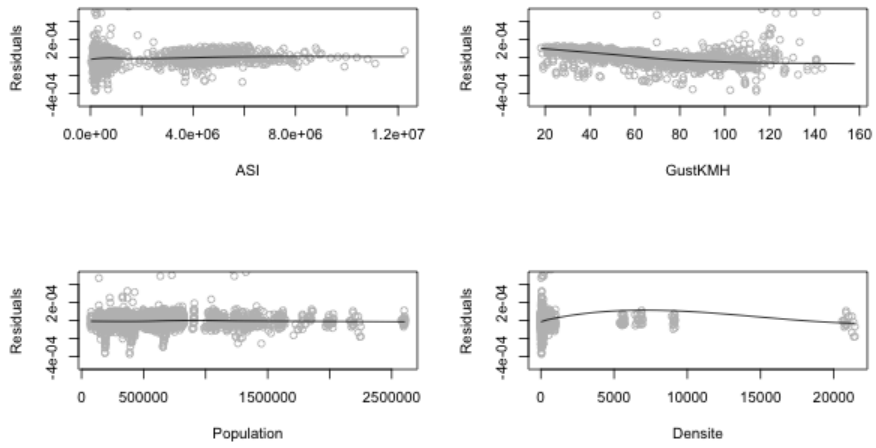


FIGURE 4.1 – Résidus par rapport aux variables explicatives

La FIGURE 4.1 représente les résidus par rapport aux différentes variables utilisées par le modèle. Pour chaque variable explicative, les résidus semblent avoir une variance à peu près constante autour de 0, ce qui confirme la linéarité entre les taux de dommages et ces variables.

Pour la variable `GustKMH`, les résidus ont tout de même tendance à s'éloigner beaucoup de 0 pour les plus grandes vitesses de vent. Nous avions d'ailleurs remarqué avec la FIGURE 3.6 qu'une relation linéaire lie plutôt ln(`MDR`) et les vitesses de vent que directement les variables `MDR` et `GustKMH`. Cependant, la présence de nombreux 0 dans les données pose problème dans l'utilisation d'un modèle linéaire généralisé avec un lien log, et nous ne voulons pas perdre les informations apportées par ceux-ci.

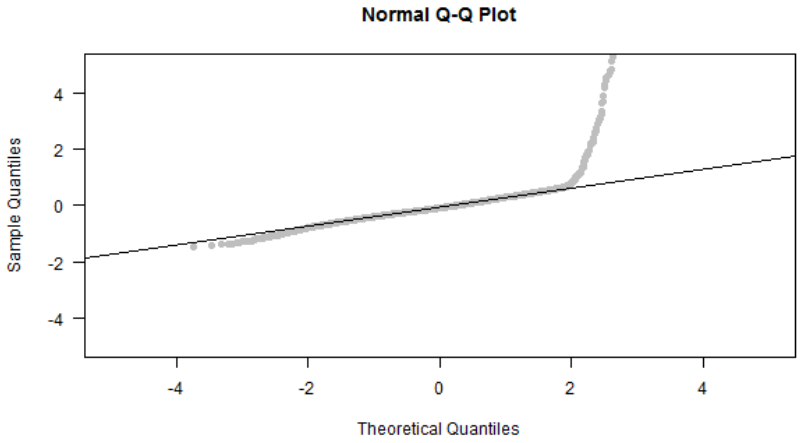


FIGURE 4.2 – Q-Q plot des résidus contre la loi normale

La FIGURE 4.2 est le diagramme Quantile-Quantile normal des résidus. Il représente les quantiles des résidus par rapport aux quantiles de la loi normale.

Ainsi, pour chaque point (x, y) de la courbe, $P(r < y) = P(N < x)$ [1]. Pour des résidus gaussiens, les points seraient tous sur la ligne. Ici, les résidus ont une queue de distribution plus lourde que celle de la loi normale, ce qui indique la présence de nombreux résidus élevés. On remarque sur la FIGURE 4.3 que les plus grands taux de dommages, qui sont plus rares, sont très mal prédits, ce qui explique ce grand nombre de résidus élevés. Globalement, les résultats de ce modèle ne semblent pas très bons.

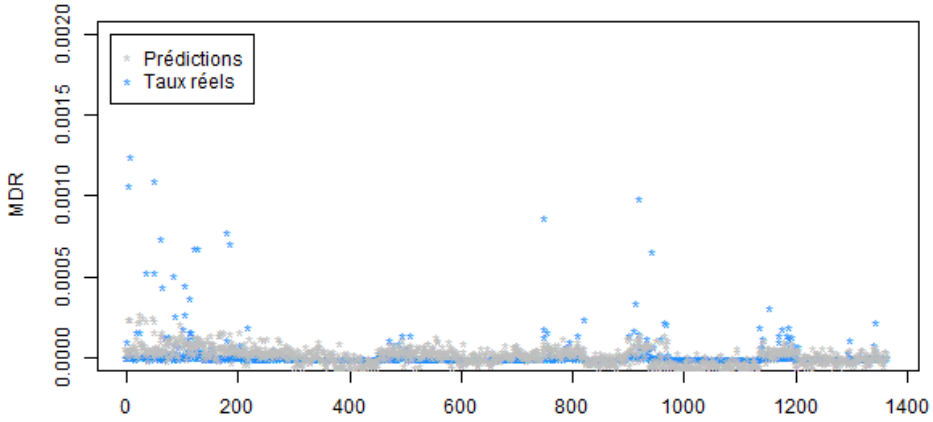


FIGURE 4.3 – Taux réels et taux prédits par le modèle linéaire

4.1.2 Arbre de décision

Nous espérons obtenir de meilleurs résultats avec un arbre de décision. Nous construisons alors, dans un premier temps, un arbre maximal, représenté dans la FIGURE 4.4, avec la fonction `rpart()` sans contrainte.

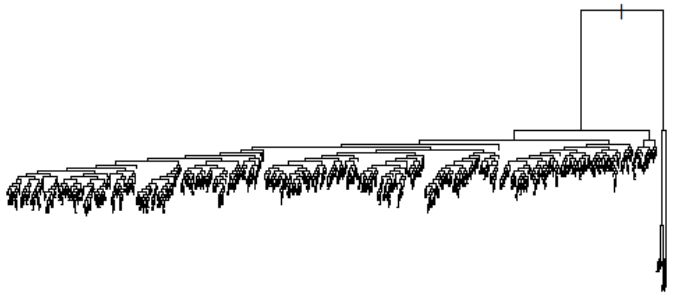


FIGURE 4.4 – Représentation de l'arbre maximal

1. Ici, r représente les résidus et N représente la loi normale.

Nous cherchons ensuite le paramètre de complexité `cp` qui minimise l'erreur pour procéder à l'élagage de notre arbre maximal et obtenir un arbre optimal. Ce paramètre représente l'amélioration minimale que doit apporter une nouvelle division pour être effectuée. Ainsi, toutes les "branches" de l'arbre qui n'améliorent pas suffisamment la qualité du modèle sont coupées.

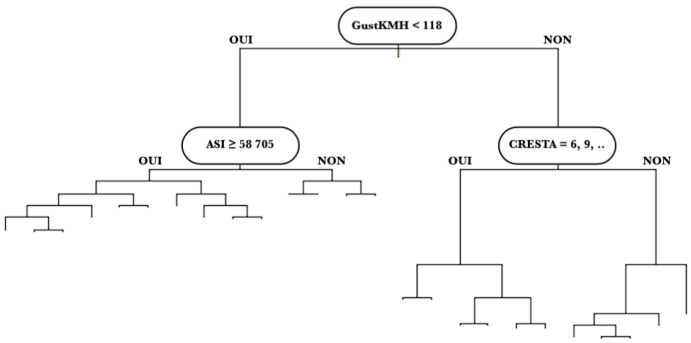


FIGURE 4.5 – Représentation de l'arbre optimal

Nous construisons alors ce second arbre, dit optimal, avec `cp` $= 2,16.10^{-3}$. Il en résulte l'arbre de la FIGURE 4.5 dont les divisions se font principalement, et en premier lieu, sur les vitesses de vent, puis sur les sommes assurées moyennes et les départements. Les variables `Occupancy.Type`, `Contents.Part` et `BI.Part` sont aussi utilisées pour les derniers découpages. En revanche, les variables `TSI`, `Population`, et `Densite` n'influent pas du tout sur les prédictions de notre arbre final. Si la variable `TSI` était aussi écartée par la régression *stepwise* et la variable `Densite` était conservée mais non significative, la variable `Population` par contre était très significative dans notre modèle linéaire.

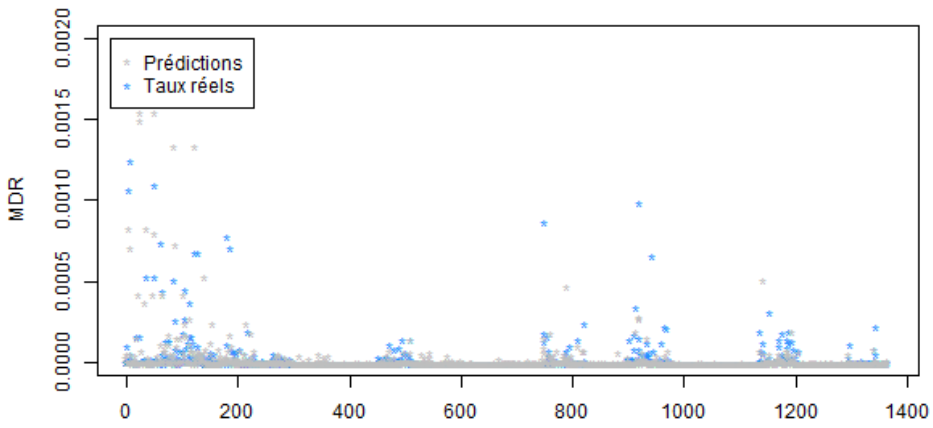


FIGURE 4.6 – Taux réels et taux prédits par l'arbre de décision

Les résultats de l'arbre de décision sur l'échantillon test, présentés dans la FIGURE 4.6, paraissent meilleurs que ceux du modèle linéaire, mais il y a encore de grands écarts entre les taux réels et les prédictions pour les plus grands taux. Dans le but d'améliorer ce modèle, nous nous tournons vers les méthodes ensemblistes de forêt aléatoire et de *gradient boosting machine* (GBM).

4.1.3 Forêt aléatoire

Sous `R`, la fonction `RandomForest()`, qui est utilisée pour construire des forêts aléatoires, ne gère pas les variables catégorielles de plus de 53 niveaux car cela ferait exploser la complexité de l'algorithme. Nous ne pouvons donc pas utiliser la variable `CRESTA` qui en a 95, pour les 95 départements français. Nous les regroupons donc par régions, selon l'ancien découpage, pour réduire le nombre de zones géographiques à 22.

Pour les forêts aléatoires, les paramètres les plus importants à ajuster sont `ntree`, qui correspond au nombre d'arbres à générer dans la forêt, et `mtry` qui est le nombre de variables, choisies aléatoirement, sur lesquelles nous pouvons choisir de diviser chacun des nœuds. On ne considère donc plus tous les couples de sous-ensembles $R_1(j, s)$ et $R_2(j, s)$ (tels qu'introduits dans la partie 2.2.3), mais seulement `mtry` couples choisis aléatoirement pour chaque nouveau nœud.

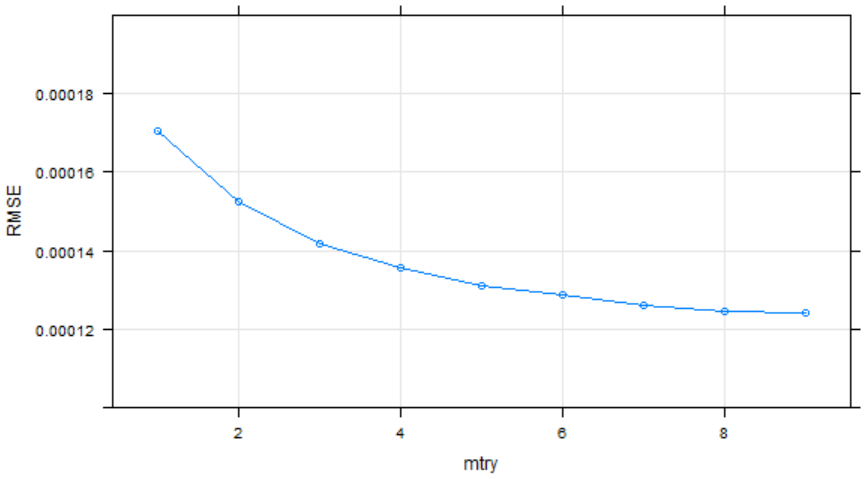


FIGURE 4.7 – Évolution de l'erreur en fonction de `mtry`

Nous cherchons la meilleure valeur pour `mtry` avec un algorithme *gridsearch* qui consiste à comparer les erreurs des modèles obtenus avec toutes les combinaisons de paramètres d'une grille. Ici, comme il y a neuf variables explicatives, le paramètre peut prendre pour valeur tous les entiers de 1 à 9. C'est donc sur ces valeurs que nous effectuons le *gridsearch*.

Les résultats de cette recherche sont donnés par la FIGURE 4.7. Selon le RMSE (les résultats sont similaires avec le MAE), le meilleur `mtry` est 9, cependant, l'amélioration du modèle est moindre après `mtry` = 5. Nous prenons donc plutôt `mtry` = 5.

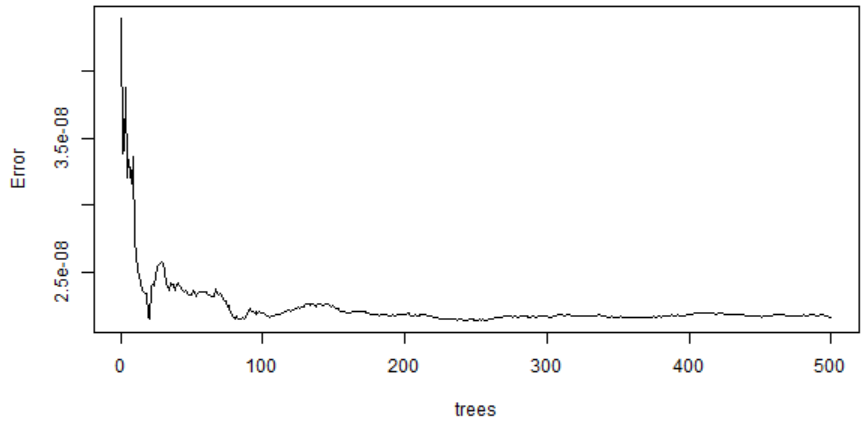


FIGURE 4.8 – Évolution de l'erreur en fonction du nombre d'arbres

Pour choisir le meilleur nombre d'arbres, nous regardons ensuite l'évolution de l'erreur du modèle par rapport au nombre d'arbres lorsque `mtry` est fixé à 5.

Nous obtenons ainsi la courbe de la FIGURE 4.8. Au delà de 100 arbres, l'erreur se stabilise, l'utilisation de plus d'arbres n'améliore pas significativement le modèle. Pour éviter le sur-apprentissage, nous retenons donc `ntree` = 100.

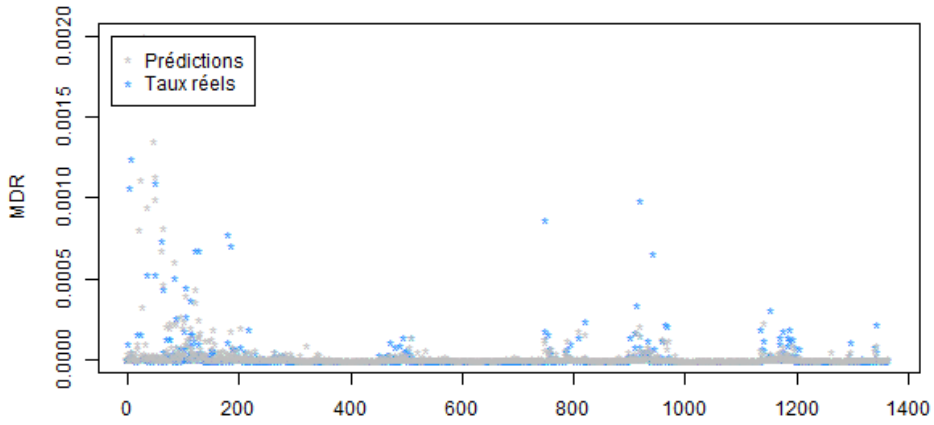


FIGURE 4.9 – Taux réels et taux prédits par la forêt aléatoire

Les résultats, visibles dans la FIGURE 4.9, sont meilleurs que ceux de l'arbre de décision, notre forêt aléatoire prédit notamment mieux les grands taux de destruction.

4.1.4 *Gradient Boosting Machine*

Les principaux paramètres à ajuster dans la fonction `gbm()` sont `n.trees`, qui est le nombre d'arbres à créer, et `interaction.depth` qui est la profondeur maximale de chaque arbre. En effectuant un *gridsearch*, nous obtenons la FIGURE 4.10. `# Boosting Iterations` correspond au paramètre `n.trees` et `Max Tree Depth` au paramètre `interaction.depth`.

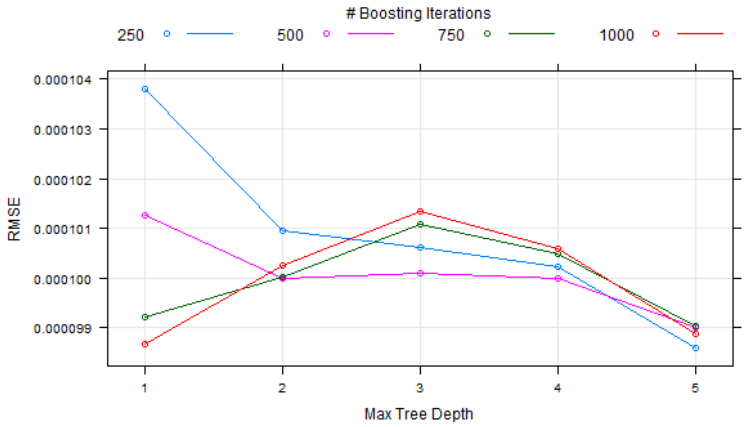


FIGURE 4.10 – Évolution de l'erreur en fonction du nombre d'arbres et de la profondeur maximale

Deux choix sont alors possibles. Nous pouvons choisir un grand nombre d'arbres simples, composés d'un seul nœud, ou peu d'arbres plus complexes. Nous choisissons finalement `n.trees` = 1000 et `interaction.depth` = 1.

VARIABLE	INFLUENCE
`GustKMH`	63.91379956
`CRESTA`	21.65311506
`TSI`	5.45728438
`ASI`	3.93832144
`Occupancy.Type`	2.44774486
`Contents.Part`	1.54500177
`BI.Part`	0.91134369
`Population`	0.10557990
`Densite`	0.02780934

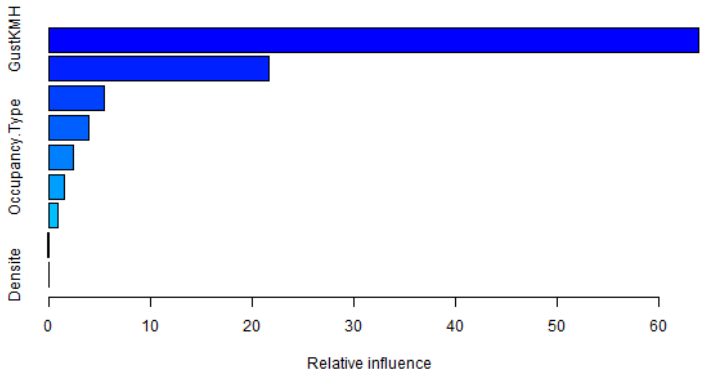


FIGURE 4.11 – Influence relative des variables

La FIGURE 4.11 représente l'influence de chacune des variables explicatives sur le modèle GBM. On retrouve, comme dans les modèles précédents, que les variables les plus imprtantes sont `GustKMH` et `CRESTA`. Par contre, les variables `Population` et `Densite` ont très peu d'influence sur ce modèle alors qu'elles faisaient partie des variables conservées par la régression *stepwise*. Ces deux variables étaient par contre aussi écartées par l'arbre optimal.

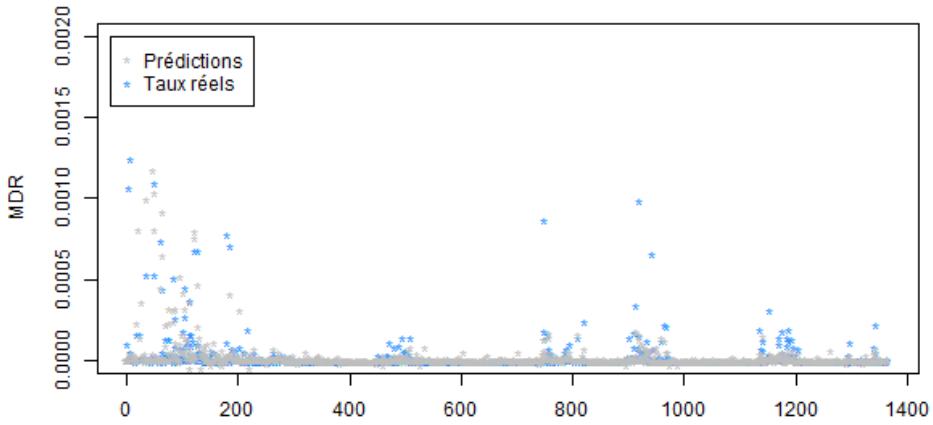


FIGURE 4.12 – Taux réels et taux prédits par le *gradient boosting*

Les résultats du GBM, représentés par la FIGURE 4.12, sont similaires à ceux de notre forêt aléatoire.

4.1.5 Sélection du meilleur modèle

Nous regardons maintenant les erreurs de nos différents modèles afin de pouvoir les comparer plus précisément et sélectionner le meilleur.

La TABLE 4.3 récapitule les RMSE et MAE [1] de nos modèles sur l'échantillon test.

MODÈLE	RMSE	MAE
Modèle linéaire	0.019848	0.0054494
Arbre de décision	0.016297	0.0027904
Forêt aléatoire	0.012458	0.0022125
GBM	0.012360	0.0024164

TABLE 4.3 – RMSE et MAE de nos différents modèles

L'arbre de décision améliore considérablement le modèle linéaire, divisant son MAE par deux.

L'amélioration des modèles ensemblistes par rapport à l'arbre est moins importantes, mais c'est tout de même le GBM qui minimise le RMSE et la forêt aléatoire qui minimise le MAE. Nous sélectionnons finalement la forêt aléatoire qui minimise le MAE et a un RMSE proche de celui du GBM.

1. Afin d'être plus lisibles, les valeurs ont été multipliées par 100 et représentent donc des pourcentages.

4.2 Deuxième modèle

Dans notre première série de modèles, le grand nombre de 0 observés biaise nos résultats et a tendance à nous faire sous-estimer les coûts lorsqu'il y a un sinistre et à les sur-estimer lorsqu'il n'y en a pas. Nous avons donc essayé de traiter séparément la présence d'un sinistre ou non, par classification puis, par régression sur les autres données, les montants pour les lignes sinistrées. Cela nous permet de diminuer le nombre de 0 dans la base sans pour autant perdre les informations qu'ils nous apportent. Dans ce but, nous créons une nouvelle variable à expliquer `hit` qui prend la valeur 0 si les pertes sont nulles et 1 si elles sont strictement positives.

4.2.1 Présence de sinistres

Modèle linéaire généralisé

Comme nous voulons expliquer une variable qui ne prend que les valeurs 0 et 1, nous faisons un modèle linéaire généralisé avec la famille `binomial` et sa fonction de lien canonique `logit`. Par régression `stepwise`, nous écartons les variables `Contents.Part`, `TSI` et `Densite`. Nous pouvons remarquer que les variables sélectionnées pour expliquer la présence ou non d'un sinistre diffèrent de celles pour expliquer les montants sinistrés. Ce modèle est un modèle de régression, il prédit des réels entre 0 et 1 et non pas les classes 0 ou 1. Afin d'avoir les résultats de la classification, nous attribuons donc la classe 1 aux valeurs supérieures à $\frac{1}{2}$, et la classe 0 aux valeurs strictement inférieures à $\frac{1}{2}$. Nous obtenons ainsi la matrice de confusion de la FIGURE 4.13.

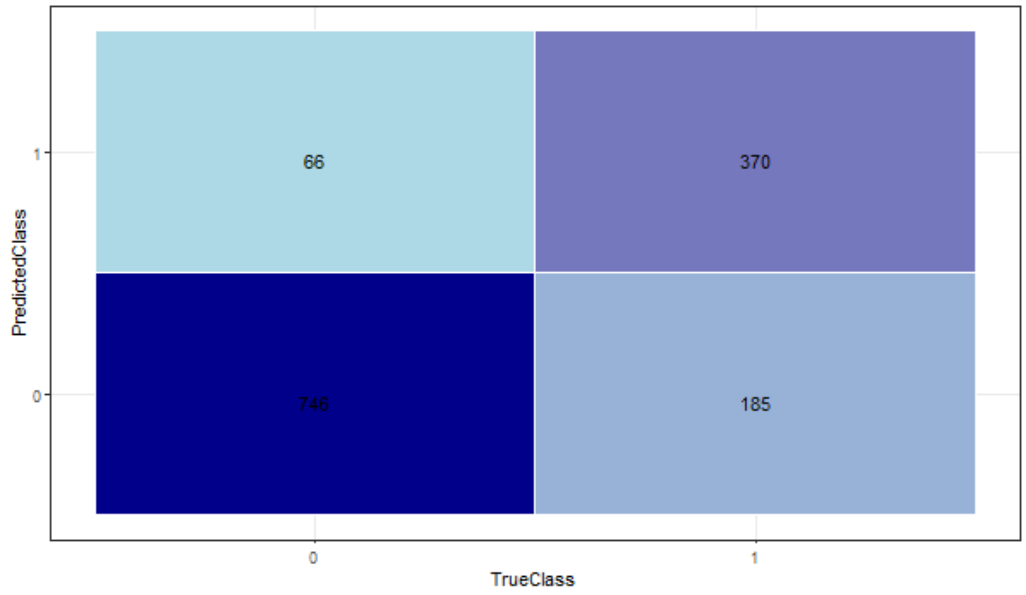


FIGURE 4.13 – Matrice de confusion du modèle linéaire généralisé

Ce modèle prédit très bien les valeurs qui sont observées à 0, mais a plus de difficultés à prédire les observations de 1 puisque près d'un tiers de ces observations sont des faux négatifs. Cela est très gênant car ce sont des éléments auxquels le modèle attribuera automatiquement une perte nulle alors qu'elle ne l'est pas. Les 66 faux positifs sont moins graves car ils pourront encore être prédits nuls ou du moins de très faible valeur par la partie régression pour déterminer les taux dans la suite.

Arbre de décision

(a) Arbre maximal (b) Arbre optimal

FIGURE 4.14 – Représentation des arbres obtenus

De la même manière que pour notre premier arbre de régression, nous construisons d'abord un arbre maximal, puis en déterminant le meilleur paramètre de complexité `cp`, nous l'élaguons pour obtenir l'arbre optimal de la FIGURE 4.14b. Ces deux arbres ont une structure très différentes de celles des arbres de régression construits précédemment. Néanmoins, ce sont toujours les variables `GustKMH` et `CRESTA` qui déterminent les premiers sous-ensembles. Pour les divisions qui suivent, la totalité des variables est finalement utilisée aussi bien par l'arbre maximal que par l'arbre optimal.

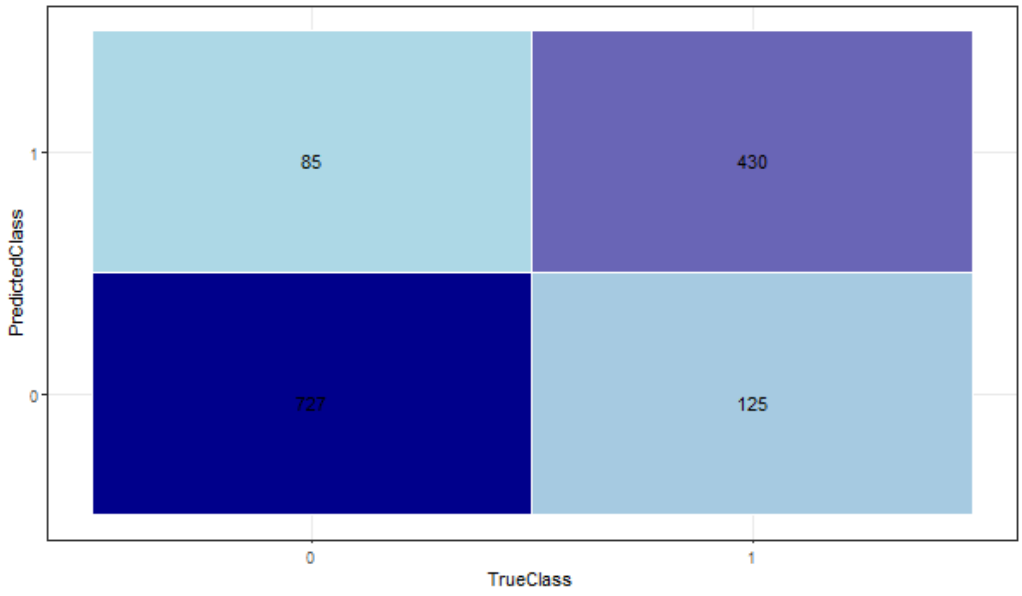
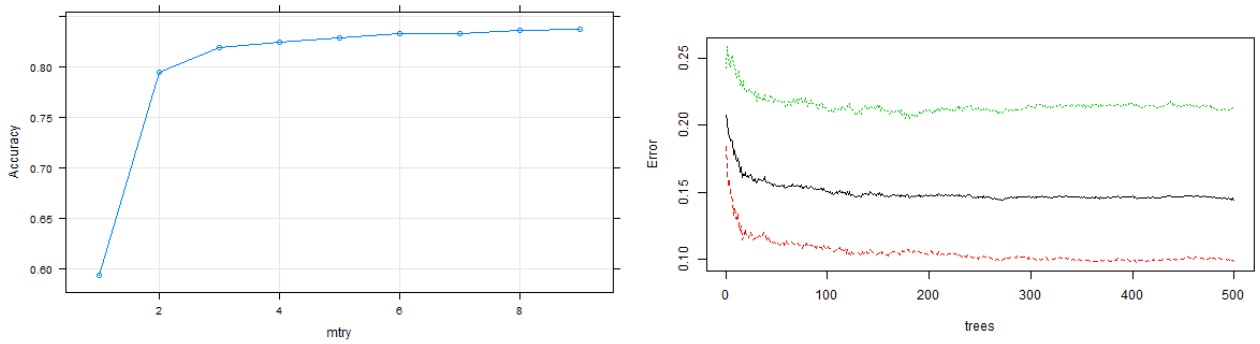


FIGURE 4.15 – Matrice de confusion de l'arbre de décision

La FIGURE 4.15 présente les résultats de notre arbre sur les données test. Ce modèle est meilleur pour prédire les 1 correctement, mais moins bon pour les 0 par rapport au modèle linéaire précédent. Nous préférons ce modèle au précédent car, dans notre cas et comme expliqué plus tôt, la présence de faux positifs est moins grave que celle de faux négatifs.

Forêt aléatoire



(a) Justesse en fonction de `mtry`

(b) Erreur en fonction du nombre d'arbres

FIGURE 4.16 – Ajustement des paramètres

Dans le cas de la classification, l'algorithme de *gridsearch* utilise l'*Accuracy* plutôt qu'une mesure d'erreur. Nous choisissons alors le paramètre qui maximise l'*Accuracy* sur la FIGURE 4.16a.

À partir des courbes de la FIGURE 4.16 et en suivant le même raisonnement que pour la forêt précédente, nous choisissons alors les paramètres `mtry` $= 4$ et `ntree` $= 100$ pour notre forêt.

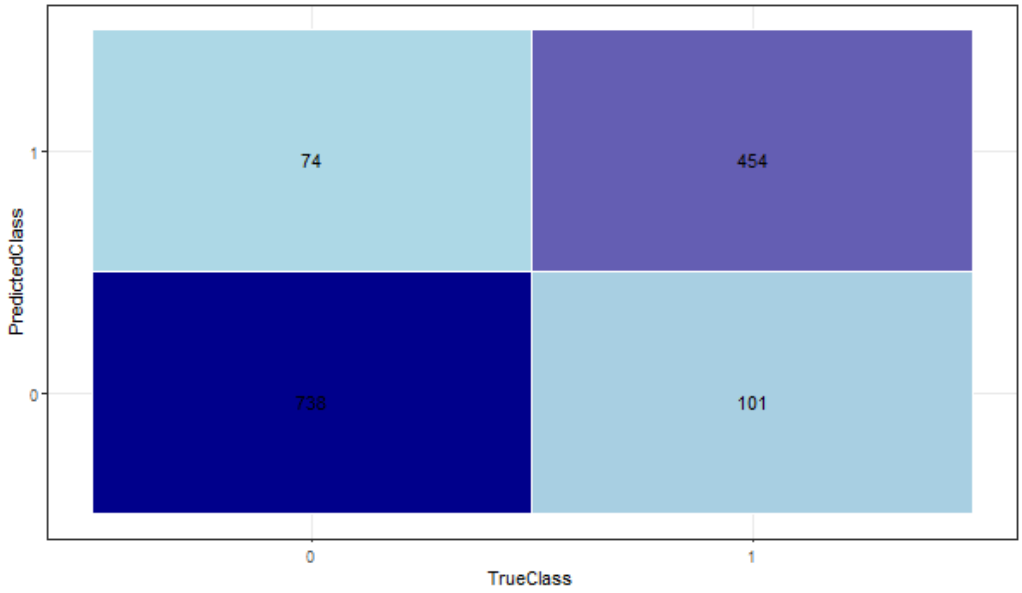


FIGURE 4.17 – Matrice de confusion de la forêt aléatoire

Cette forêt améliore les résultats de notre arbre. En effet, la FIGURE 4.17 montre que le nombre de faux positifs, comme celui de faux négatifs, est plus faible que pour l'arbre de décision précédent. Le nombre de faux positifs est, en revanche, toujours plus élevé que pour le modèle linéaire. Cependant, pour tous nos modèles de classification ce nombre de faux positifs est négligeable par rapport au nombre de vrais négatifs. Le nombre de faux négatifs varient de façon plus importante entre les modèles.

Gradient Boosting

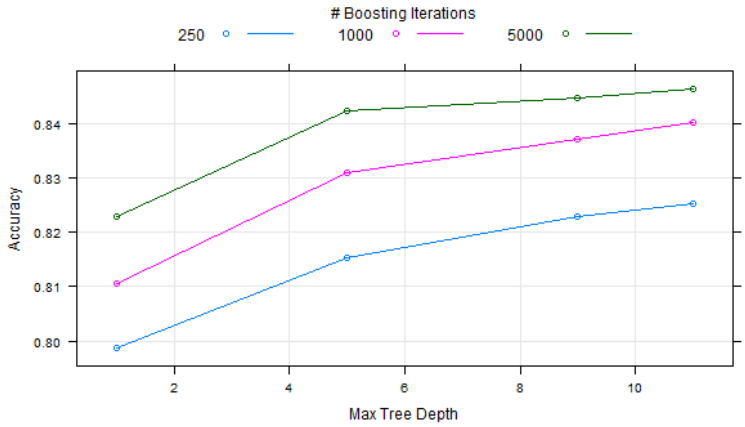


FIGURE 4.18 – Évolution de la justesse en fonction du nombre d'arbres et de la profondeur maximale

À l'aide d'un *gridsearch*, dont les résultats sont présentés dans la FIGURE 4.18, nous fixons les paramètres de notre nouveau GBM à `interaction.depth` = 5 et `n.trees` = 5000.

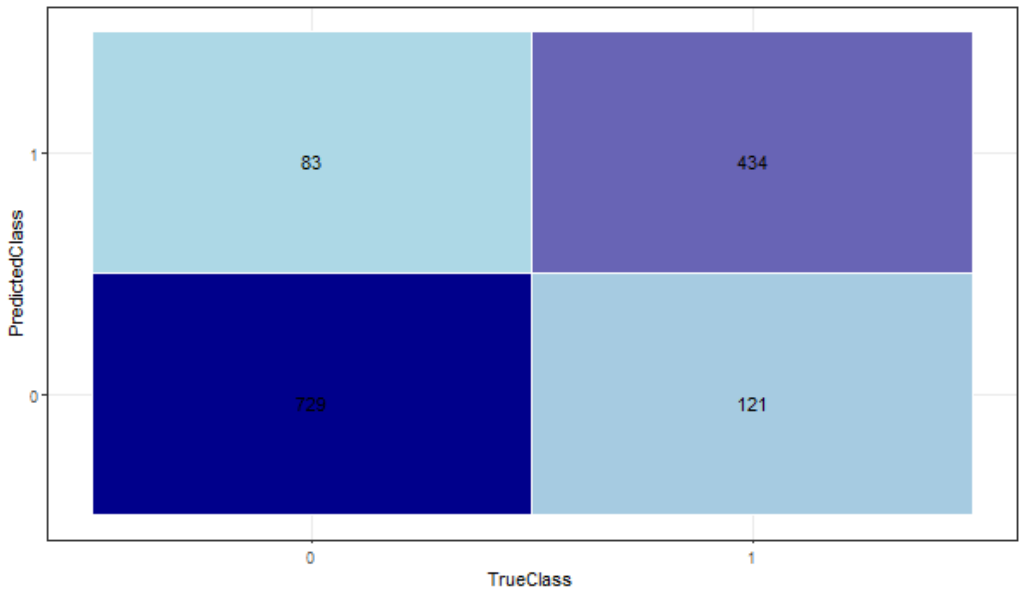


FIGURE 4.19 – Matrice de confusion du GBM

La FIGURE 4.19 présente une matrice de confusion moins bonne que celle de la forêt et qui se rapproche beaucoup de celle de l'arbre de décision. La forêt aléatoire semble alors se présenter comme le meilleur modèle de classification.

Sélection du modèle de classification

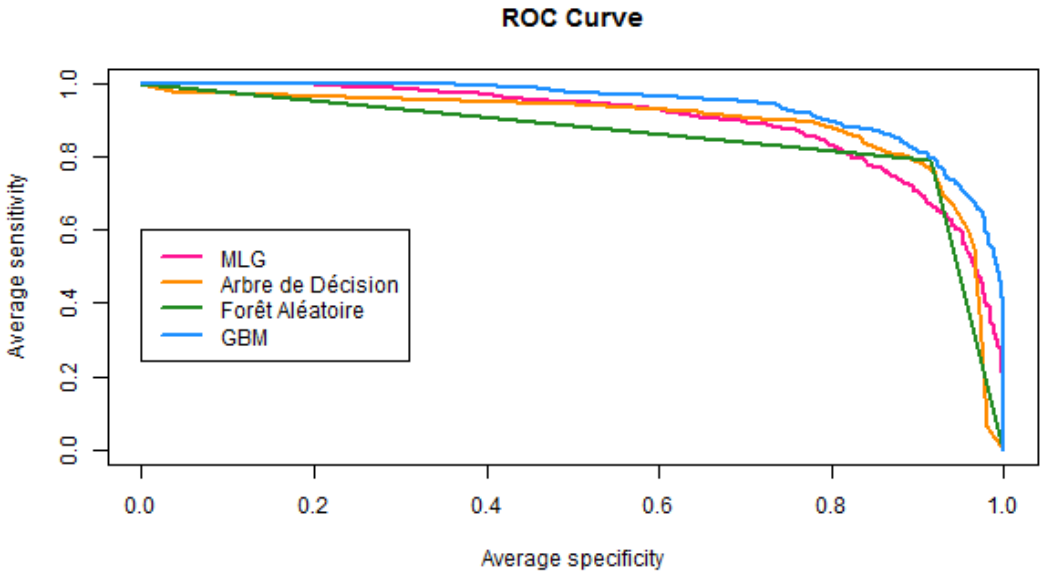


FIGURE 4.20 – Courbes ROC des différents modèles

Les tracés des courbes ROC de la FIGURE 4.20 nous permettent de comparer visuellement la qualité de nos différents modèles. Le GBM est le meilleur selon le critère de la courbe ROC puisqu'il se rapproche le plus du coin supérieur droit représentant un modèle parfait.

MODÈLE	ACCURACY	PRECISION	RECALL	F1 SCORE	AUC
Modèle linéaire	0.8164	0.8486	0.6667	0.7467	0.8944
Arbre de décision	0.8464	0.8350	0.7748	0.8037	0.8990
Forêt aléatoire	0.8720	0.8598	0.8180	0.8384	0.8634
GBM	0.8508	0.8395	0.7820	0.8097	0.9299

TABLE 4.4 – Scores des modèles

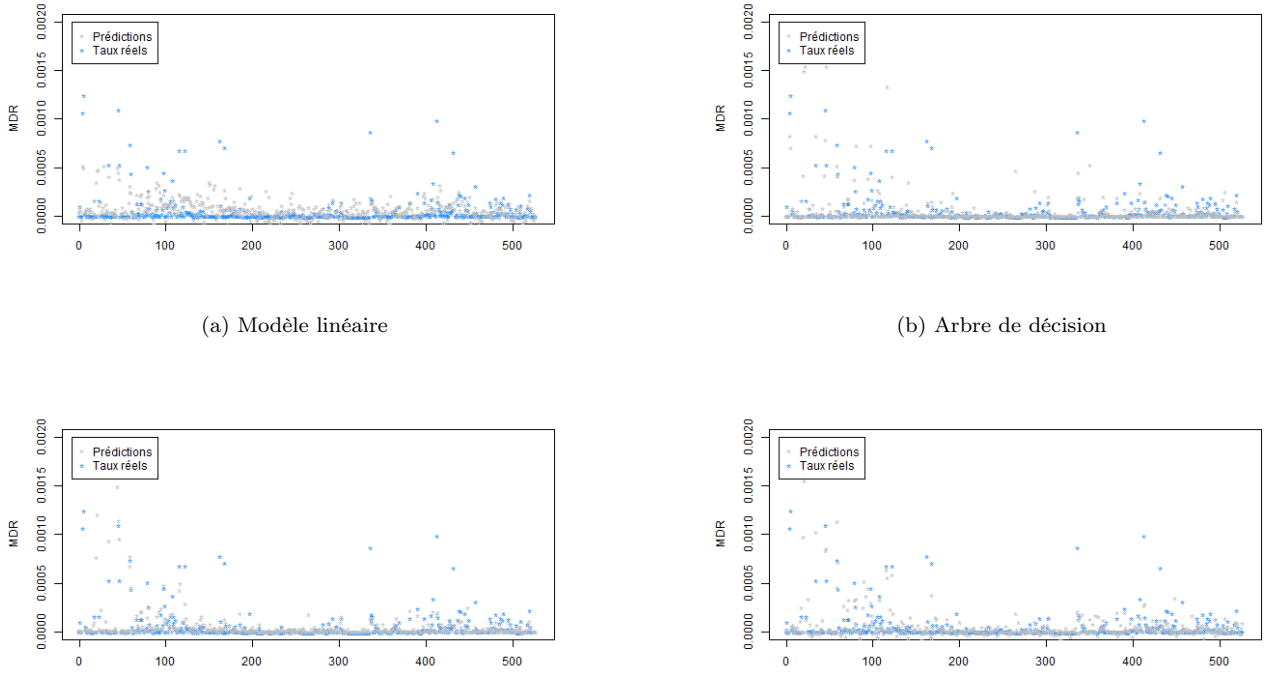
Les différents scores de la TABLE 4.4 nous permettent de comparer numériquement les quatre modèles.

Dans notre cas, comme nous souhaitons prioritairement limiter le nombre de faux négatifs, nous nous intéressons surtout au *Recall*. Selon ce critère, la forêt aléatoire est le meilleur modèle.

En regardant tout de même les autres scores, nous pouvons remarquer que la forêt est le meilleur modèle partout sauf en AUC. Nous sélectionnons donc la forêt aléatoire pour prédire la présence ou non de sinistres.

4.2.2 Montants sinistrés

Pour prédire ensuite les montants sinistrés, nous ne conservons, de notre base d'apprentissage, que les lignes pour lesquelles la forêt aléatoire de classification sélectionnée précédemment prédit la présence de sinistres, c'est-à-dire 1. Nous suivons ensuite le même processus que pour la première série de modèles de la partie 4.1. Après ajustement des paramètres, nous obtenons les résultats de la FIGURE 4.21.



(a) Modèle linéaire

(b) Arbre de décision

(c) Forêt aléatoire

(d) GBM

FIGURE 4.21 – Taux réels et taux prédits par nos différents modèles

Nous pouvons déjà remarquer que le modèle linéaire est meilleur sur les taux élevés que le premier modèle linéaire effectué sur la totalité des données, mais il a toujours tendance à beaucoup sur-estimer les plus petits taux. Les autres modèles semblent aussi plus précis que nos premiers modèles de régression. En considérant les mesures d'erreurs, données dans la TABLE 4.5, les méthodes ensemblistes sont toujours les meilleures avec des RMSE et MAE très proche. La forêt a un meilleur MAE, c'est donc le modèle que nous sélectionnons puisque les RMSE des deux dernières méthodes sont presque égaux.

MODÈLE	RMSE	MAE
Modèle linéaire	0.0294	0.00977
Arbre de décision	0.0265	0.00763
Forêt aléatoire	0.0194	0.00532
GBM	0.0191	0.00606

TABLE 4.5 – RMSE et MAE des modèles

4.3 Comparaison des deux modèles sélectionnés

Nous avons finalement construit deux modèles. Le premier consiste en une forêt aléatoire de régression directement sur les données. Le second est formé de deux modèles emboîtés, une première forêt aléatoire de classification détermine d'abord la présence ou non de sinistres, puis une seconde prédit ensuite les montants pour les lignes sinistrées.

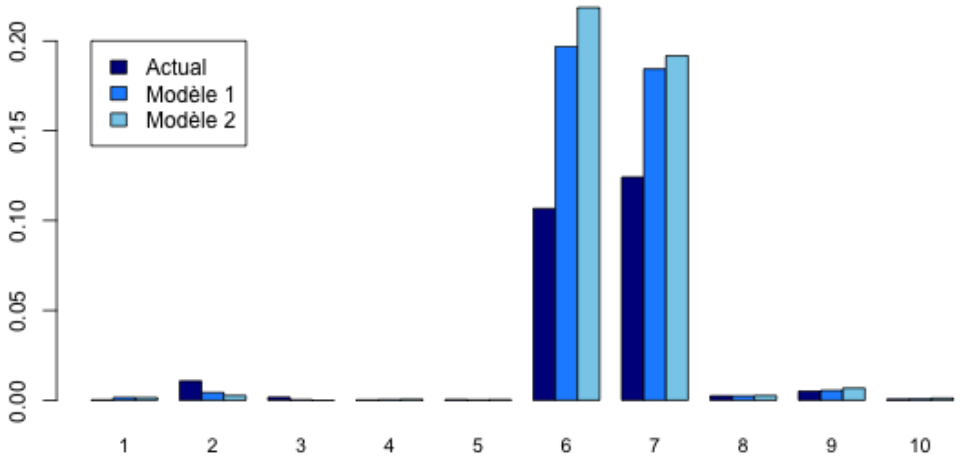


FIGURE 4.22 – Taux de dommages (en pourcentage)

En regardant quelques prédictions, dans la FIGURE 4.22, aucun des deux modèles ne se démarque clairement comme le meilleur. Les deux modèles se suivent, ils sous-estiment et sur-estiment de la même manière les taux réels et toujours avec deux prédictions très proches entre elles.

Le modèle 1 est ici le plus proche du taux réel à chaque fois, mais nous ne pouvons pas nous fier aux seuls résultats de dix lignes.

MODÈLE	RMSE	MAE
Modèle 1	0.012458	0.0022125
Modèle 2	0.012081	0.0020833

TABLE 4.6 – Erreurs des deux modèles sélectionnés

Les deux mesures d'erreur sont plus fiables pour déterminer le meilleur des deux modèles.

Contrairement à l'impression donnée par les dix taux regardés précédemment, c'est le deuxième modèle qui est le meilleur puisqu'il minimise à la fois le RMSE et le MAE dont les valeurs sont données dans la TABLE 4.6. Les deux modèles sont tout de même très proches, le deuxième n'améliore pas significativement le premier.

Regardons désormais à quoi ressemblent les prédictions des taux les plus élevés ainsi que celles des taux nuls pour les deux modèles. La FIGURE 4.23 représente les prédictions (en bleu) et les taux réels (traits noirs) des cinquante plus grands taux réels d'une part et de tous les taux réels nuls d'autre part.

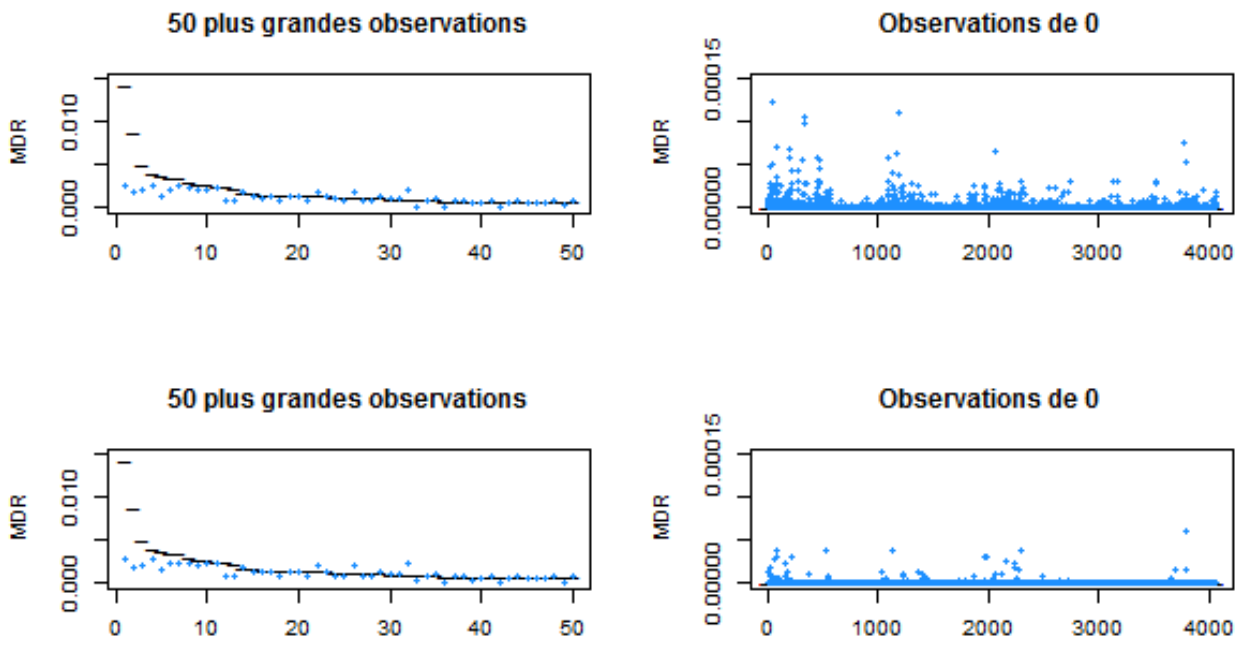


FIGURE 4.23 – Prédictions des modèles 1 (en haut) et 2 (en bas) pour les plus grands taux et les taux nuls

Les deux modèles semblent avoir des résultats très proches sur les plus grand taux de dommages observés. Par contre, les prédictions des taux qui sont réellement nuls sont très différentes. Un très grand nombre de taux réels nuls sont prédits non nuls par le premier modèle alors que le second modèle en prédit beaucoup moins et prédit, de plus, des taux plus limités.

Chapitre 5

Résultats et applications

Dans le chapitre précédent, nous avons construit deux modèles qui estiment les taux de dommages, et par conséquent les montants des sinistres, causés par une tempête par département français et par type de risque. Nous avons ensuite comparé les taux de destruction individuels prédits par les deux modèles par rapport aux taux de destruction réels.

Dans ce chapitre nous nous intéressons plutôt aux prédictions de nos modèles au niveau des événements. Nous en déduirons par la suite les coûts sur la réassurance.

Enfin, nous verrons quelles sont les différentes applications possibles de ces modèles à la tarification ainsi qu'au provisionnement.

5.1 Les prédictions par tempête

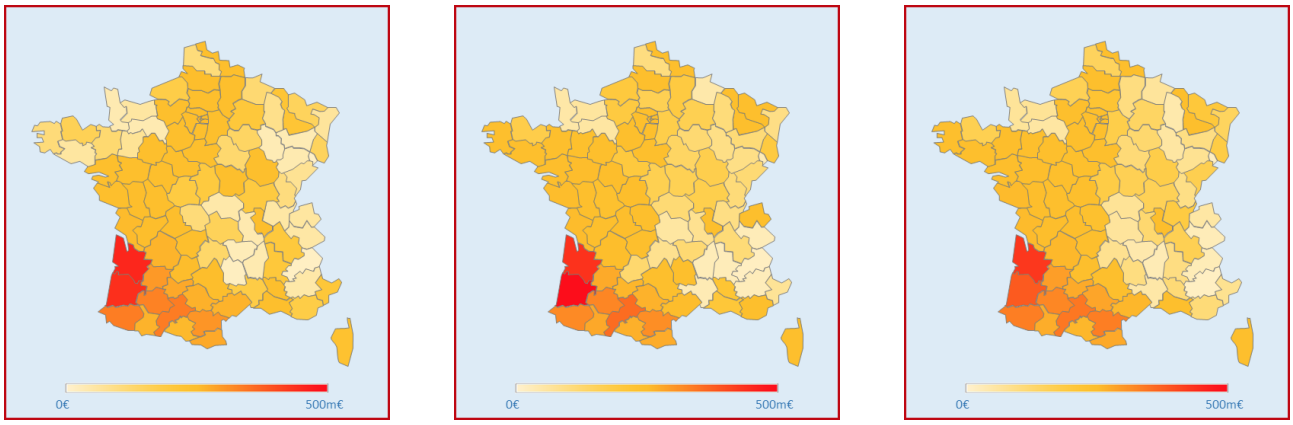
Nous commençons par regarder les événements Klaus et Xynthia qui sont les deux plus grosses tempêtes dont nous disposons dans notre base de données.

Nous regarderons ensuite les tempêtes Sabine et Victoria. Ces deux événements ont touché la majorité du Nord de l'Europe au début de l'année 2020, le 9 février et le 15 février respectivement. Si les pertes recensées pour ces deux événements approchent 2 milliards€ en Europe, celles-ci sont réparties sur de nombreux pays et les pertes en France dépassent à peine 350 millions€.

Comme ces événements sont plus récents, ils ne faisaient pas encore partie de notre base de données lors de la construction de nos modèles. Ils nous permettront donc de vérifier que les modèles s'adaptent bien à des événements complètement nouveaux puisqu'une partie des événements Klaus et Xynthia a été utilisée pour l'apprentissage de nos modèles. Par ailleurs, nous pourrons aussi voir comment les modèles réagissent face à des événements plus petits.

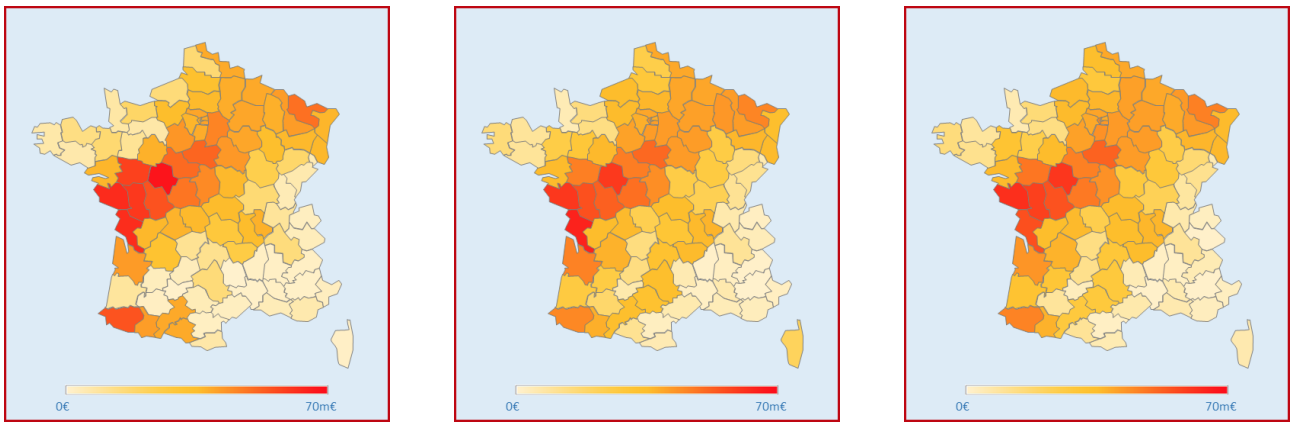
5.1.1 Klaus et Xynthia

Les FIGURES 5.1 et 5.2 donnent une première comparaison globale des prédictions de nos deux modèles aux pertes réelles des événements Klaus et Xynthia.



(a) Montants réels (b) Modèle 1 (c) Modèle 2

FIGURE 5.1 – Cartes des montants sinistres par département pour l'événement Klaus



(a) Montants réels (b) Modèle 1 (c) Modèle 2

FIGURE 5.2 – Cartes des montants sinistres par département pour l'événement Xynthia

Pour ces deux tempêtes, les cartes sont similaires. Les décalages observables sont principalement sur les départements les moins touchés. Les départements les plus touchés, qui sont les plus importants dans le calcul du coût global, semblent mieux prédits.

Nous nous concentrons donc maintenant sur les zones les plus sinistrées afin de voir plus précisément les écarts entre les prédictions de nos modèles et les montants réels.

Les histogrammes des FIGURES 5.3 et 5.4 présentent ces résultats sur les vingt départements les plus touchés.

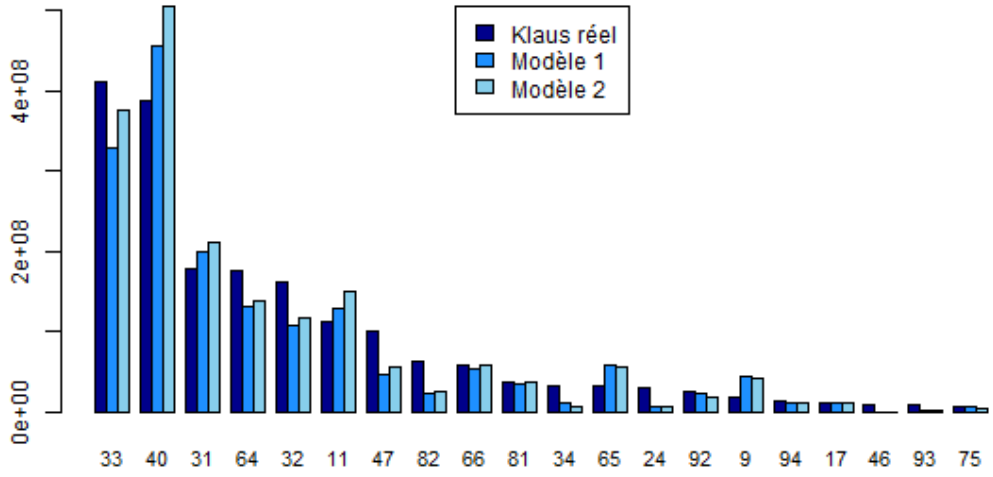


FIGURE 5.3 – Montants réels et prédits pour les vingt départements les plus touchés par la tempête Klaus

Pour Klaus, les sinistres se concentrent sur cinq ou six départements qui sont relativement bien prédits mis à part les Landes (département 40) qui est très sur-estimé par nos deux modèles. Pour Xynthia, les dommages sont répartis sur plus de départements dont plusieurs sont beaucoup sous-estimés. Au global, les montants prédits par nos modèles s'élèvent à 1,831 et 1,983 milliards€ contre 1,966 milliards€ réels pour Klaus et 858 et 860 millions€ contre 914 millions€ réels pour Xynthia.

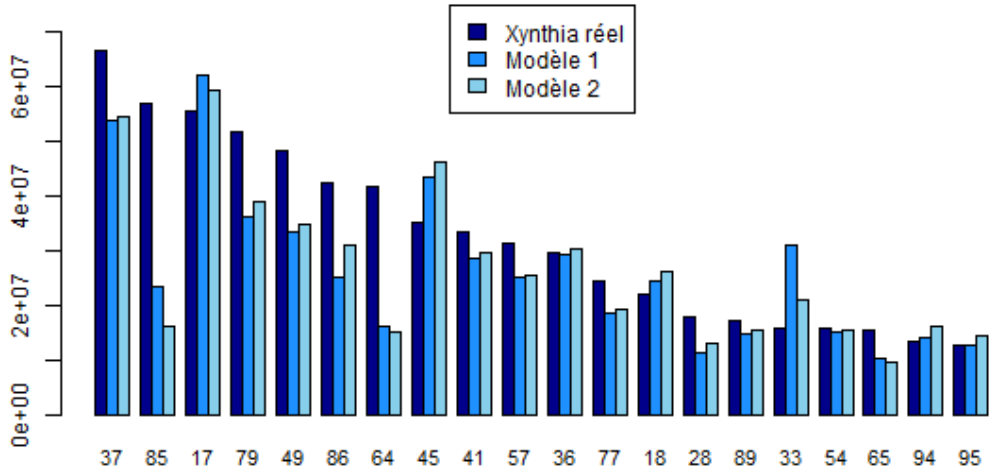
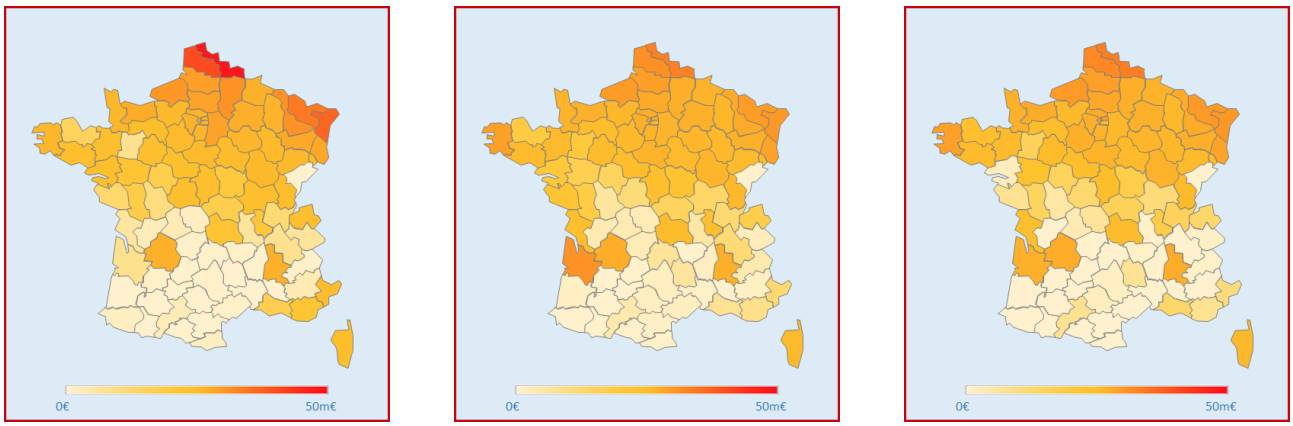


FIGURE 5.4 – Montants réels et prédits pour les vingt départements les plus touchés par la tempête Xynthia

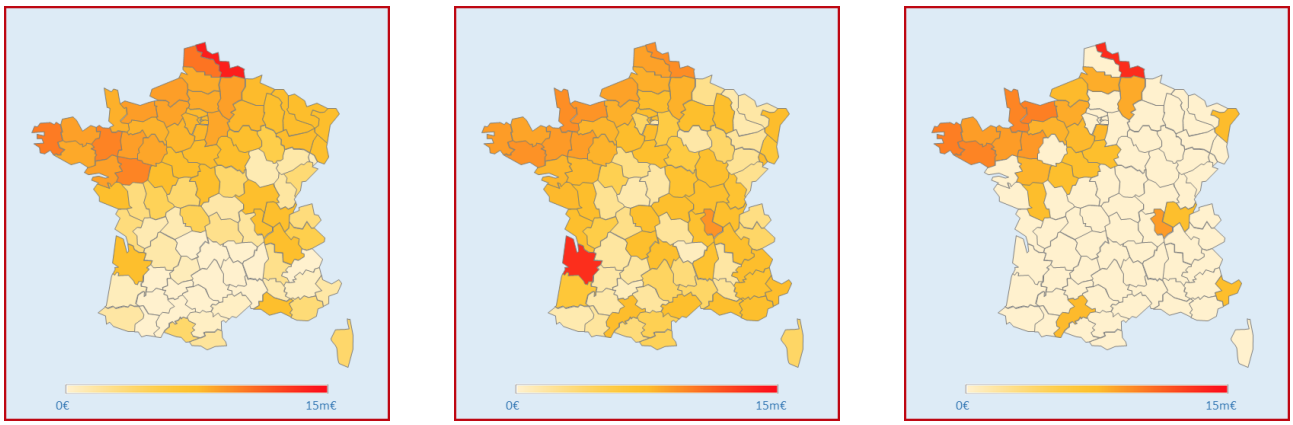
5.1.2 Sabine et Victoria

Les FIGURES 5.5 et 5.6 donnent une première comparaison globale des prédictions de nos deux modèles aux pertes réelles des événements Sabine et Victoria.



(a) Montants réels (b) Modèle 1 (c) Modèle 2

FIGURE 5.5 – Cartes des montants sinistres par département pour l'événement Sabine



(a) Montants réels (b) Modèle 1 (c) Modèle 2

FIGURE 5.6 – Cartes des montants sinistres par département pour l'événement Victoria

Pour Sabine et Victoria, les cartes de nos modèles changent plus de la carte des montants réels que cela avait été le cas pour Klaus et Xynthia. Nous observons notamment des écarts importants pour les départements les plus touchés.

Par ailleurs, si les cartes de Sabine restent assez proches, la carte du modèle 2 pour Victoria est très différentes de la carte des montants réels avec notamment de nombreux départements prédits à 0.

Nous notons que les écarts importants de couleurs sont aussi dus au changement d'échelle. En effet, les pertes par département pour Victoria ne dépassent pas 15 millions€ alors qu'elles atteignaient presque 500 millions€ pour Klaus.

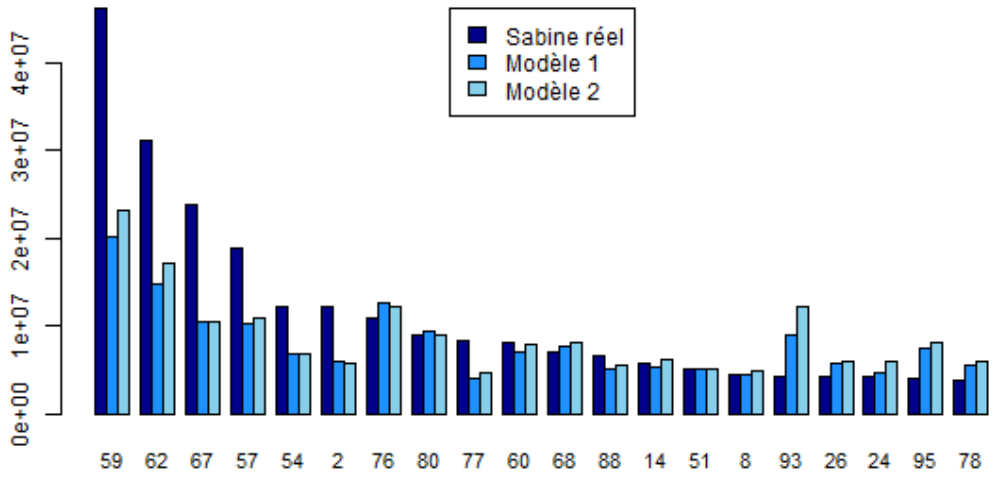


FIGURE 5.7 – Montants réels et prédits pour les vingt départements les plus touchés par la tempête Sabine

En nous concentrant sur les vingt départements les plus touchés par Sabine, nous retrouvons bien des sous-estimations importantes pour les cinq premiers départements. Les suivants sont mieux estimés. Pour Victoria nous retrouvons même plusieurs 0 prédits pour les départements pourtant les plus touchés en raison de la faible intensité de la tempête en France. Au global, les montants prédits par nos modèles s'élèvent à 257 et 292 millions€ contre 287 millions€ réels pour Sabine et 57 et 63 millions€ contre 73 millions€ réels pour Victoria.

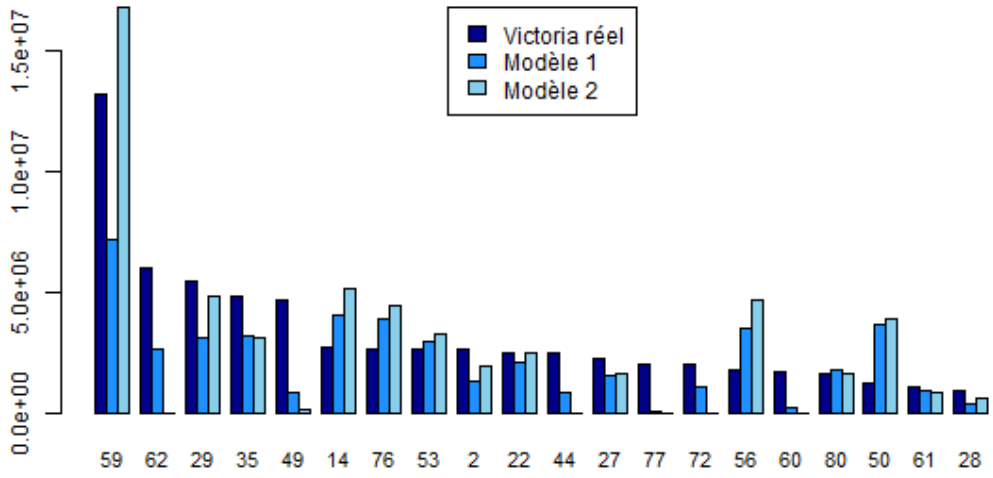


FIGURE 5.8 – Montants réels et prédits pour les vingt départements les plus touchés par la tempête Victoria

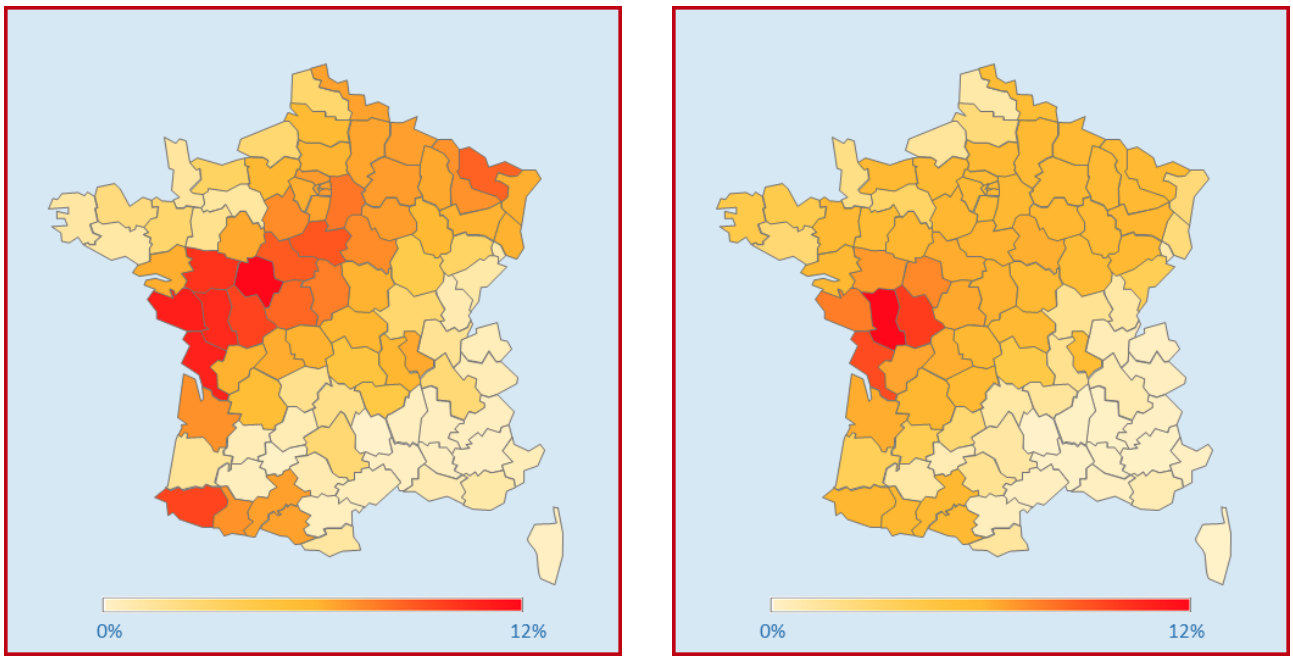
5.2 Les prédictions par cédante

Ces prédictions ne sont pas tout à fait celles qui nous intéressent puisque nous ne pouvons pas en déduire directement le montant à notre part.

En effet, nous réassurons vingt-sept assureurs contre la tempête sur le territoire français métropolitain, et nous avons avec chacune de ces vingt-sept cédantes un contrat de réassurance bien différent. Les priorités et portées changent déjà d'un contrat à l'autre, mais ce sont aussi les clauses qui varient. Il ne suffit donc pas d'appliquer une formule simple directement au montant total de la tempête ou même aux montants par département et par type de risque.

À partir des montants prédits, nous devons d'abord reconstituer les montants sinistrés pour chacune des cédantes avant de pouvoir finalement utiliser les différentes formules présentées dans la partie 1.2. Pour cela nous nous référons à notre base de nombres de polices de nos cédantes. Nous déterminons à partir de cette base, pour chaque département et pour chaque type de risque, les parts de marché des différents assureurs français.

Ces parts de marché nous permettent par la suite de répartir chaque ligne de sinistre, correspondant à un département et un type de risque, aux différents assureurs. La présence sur le territoire pouvant varier énormément d'une cédante à l'autre, la répartition des sinistres change aussi, comme le met en avant la FIGURE 5.9. Les parts de marché varient encore plus par type de risque puisque les risques agricoles par exemple, sont couverts, pour plus de 90%, par seulement deux cédantes, et la plupart des autres assureurs n'ont aucune police agricole.



(a) Pour tout le marché

(b) Pour une cédante

FIGURE 5.9 – Répartition des sinistres par département

Pour chaque cédante, nous obtenons ensuite le montant total de l'événement en sommant les sinistres de chaque département. Sur ces montants, nous pouvons enfin appliquer les priorités et les portées spécifiques à chaque assureur ainsi que les clauses d'*aggregate* (AAD et AAL) ou de reconstitutions de garantie s'il y en a.

5.2.1 Klaus et Xynthia

Nous arrivons finalement aux montants présentés dans les FIGURES 5.10 et 5.11. Ces deux histogrammes représentent les coûts à la part d'OdysseyRe obtenus en partant des coûts réels de Perils, et des coûts prédits par nos deux modèles pour les tempêtes Klaus et Xynthia respectivement.

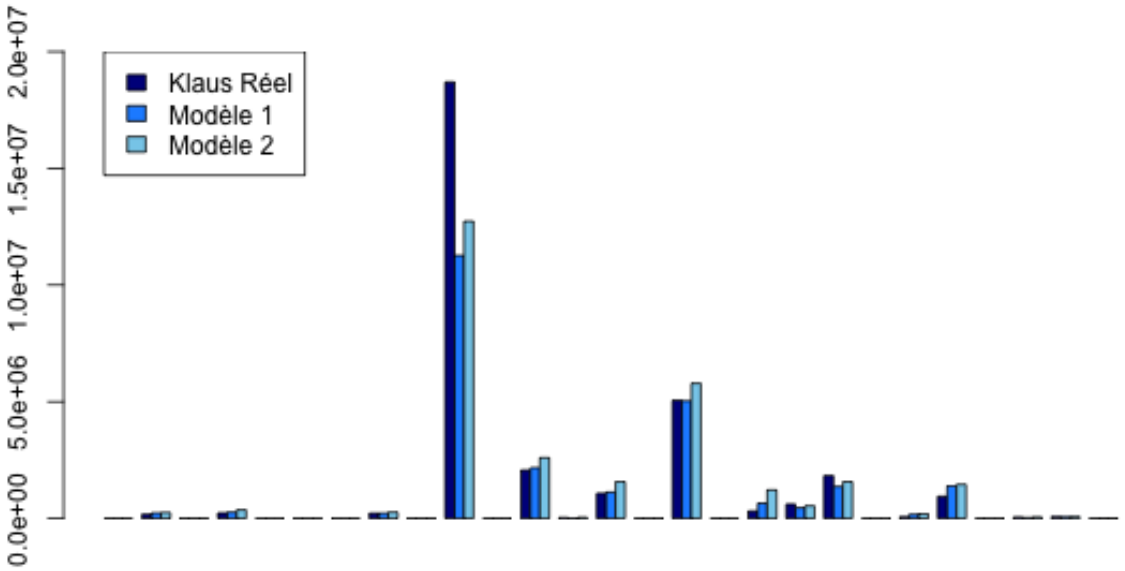


FIGURE 5.10 – Coûts pour OdysseyRe, par cédante, de la tempête Klaus

Pour Klaus, les coûts sont sous-estimés pour la cédante la plus touchée, sinon, les estimations sont très proches des montants réels. Nous remarquons pour ces deux événements que le modèle 2 donne des estimations plus élevées que le modèle 1. Au global, pour l'événement, les deux modèles estiment des coûts pour OdysseyRe s'élevant à 24 et 29 millions€ respectivement alors que les coûts obtenus à partir des montants réels est de 31 millions€.

Pour Xynthia, nous observons une sous-estimation des coûts pour cette même cédante. Les écarts relatifs entre les coûts déduits des prédictions et ceux déduits des montants réels sont plus importants. Nos deux modèles estiment des coûts à la charge d'OdysseyRe de 3,7 et 3,8 millions€ contre 4,4 millions€ en partant des montants réels.

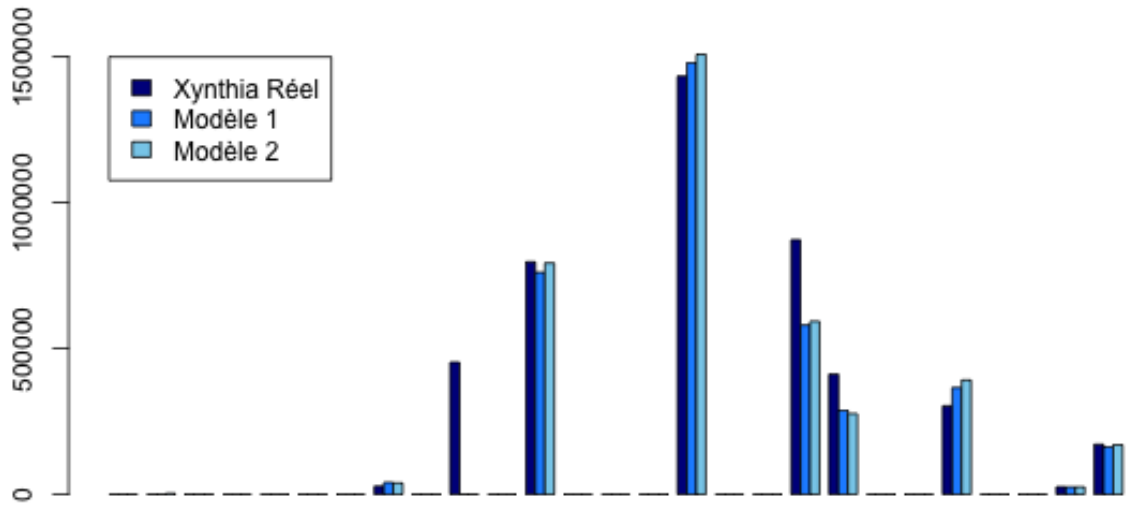


FIGURE 5.11 – Coûts pour OdysseyRe, par cédante, de la tempête Xynthia

Les résultats sont plutôt satisfaisant sur ces deux premiers événements. Nous ne pouvons toutefois pas nous arrêter sur ces résultats puisqu'une grande partie de ces événements a été utilisée pour construire les modèles. Ces derniers pourraient être beaucoup moins bons sur des données entièrement nouvelles.

C'est pour cette raison que nous regardons aussi les résultats sur les tempêtes Sabine et Victoria.

5.2.2 Sabine et Victoria

Ces deux événements sont très petits comparés à Klaus et Xynthia et ils ne touchent presque pas la réassurance. Nous ne pouvons donc pas les comparer de la même manière.

Nous regardons alors les coûts pour chacune de nos cédantes plutôt que les coûts à notre part qui sont nuls ou très faibles. Ces coûts sont ainsi présentés dans les FIGURES 5.12 et 5.13.

Les résultats sont aussi très bons pour ces deux nouvelles tempêtes. Nous pouvons observer que pour un petit événement tel que Victoria, et contrairement aux événements vus précédemment, le deuxième modèle a plutôt tendance à estimer des montants plus faible que le premier modèle. Cela est la conséquence de la prédiction de 0 pour un grand nombre de départements où l'intensité de l'événement est plus faible.

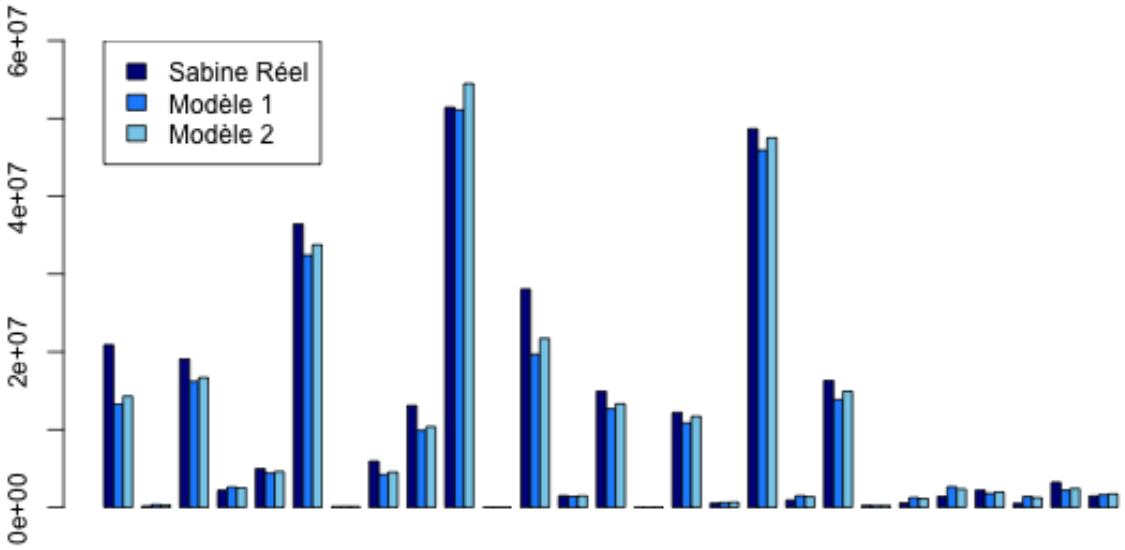


FIGURE 5.12 – Coûts par cédante de la tempête Sabine

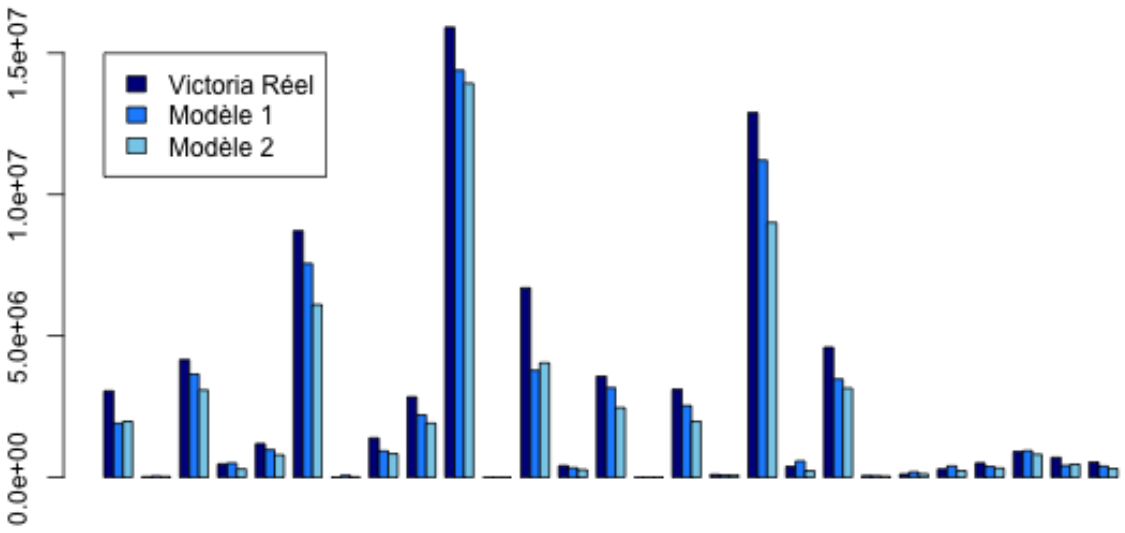


FIGURE 5.13 – Coûts par cédante de la tempête Victoria

5.3 Applications pour l'entreprise

Nous présentons enfin dans cette partie l'utilité de cet outil pour OdysseyRe au travers d'applications au provisionnement ainsi qu'à la tarification sur l'événement Klaus [1].

5.3.1 Application à la tarification

La tarification des traités de réassurance CAT peut se faire selon différentes méthodes. Les premières utilisées sont la méthode des *burning costs* qui repose sur l'expérience passée, et la méthode probabiliste, aussi appelée le modèle actuariel.

Pour les tempêtes européennes, suite aux événements de 1990, des outils de modélisation ont été développés par des sociétés spécialisées. L'utilisation de ces modélisations pour la tarification de ces traités est aujourd'hui très répandue chez les réassureurs.

La méthode des *Burning Costs*

La méthode des *burning costs*, ou méthode statistique, consiste à tarifer en se basant sur l'expérience passée de la cédante.

La cédante transmet au réassureur ses encaissements historiques d'une part et ses sinistres historiques d'autre part. Ces données sont mises en situation *as-if* par une revalorisation suivant deux indices : un pour les primes et un pour les sinistres. L'observation des évolutions historiques nous indiquant que la prime évolue moins vite que le sinistre moyen considéré, ces deux indices sont généralement choisis différents. Ceci s'explique notamment par la forte concurrence commerciale entre assureurs qui les pousse à limiter les augmentations tarifaires dans le but de préserver leurs parts de marché.

Les années historiques sont ainsi toutes ramenées à la situation de l'année du contrat. En appliquant les différentes conditions du contrat à tarifer, nous calculons alors les charges pour le réassureur correspondant à chaque année historique et les taux de charge annuelle en divisant par les encaissements de chaque année. Le taux pur peut ensuite être calculé de deux manières, soit comme la moyenne des taux obtenus pour chaque année, soit comme la somme des charges sur la somme des primes. Le tarificateur est responsable de choisir la méthode la plus adaptée ainsi que la période sur laquelle l'appliquer.

Enfin, l'application de ce taux pur à l'encaissement prévisionnel de l'année du contrat donne la prime pure de réassurance qui est ensuite complétée par des chargements représentant, par exemple, les coûts de gestion ou encore ceux du courtage si la cédante passe par un courtier.

Cette méthode suppose que le portefeuille assuré soit relativement stable d'une année à l'autre. Elle nécessite aussi une grande fréquence de sinistres, or les sinistres CAT sont rares. Pour les traités CAT, le calcul des *burning costs* sert principalement de référence pour calibrer les différents modèles.

1. Les montants présentés dans cette partie sont toujours indexés ce qui explique les éventuels écarts avec les montants réellement observés en 2009.

Le modèle actuariel

Le modèle actuariel est une méthode probabiliste qui consiste en un modèle fréquence/intensité (ou fréquence/coût). Cela signifie que la fréquence des sinistres est modélisée par une première loi de probabilité et les montants de ces sinistres par une seconde loi.

Le modèle le plus utilisé est le modèle Poisson/Pareto, la loi de Poisson modélisant le nombre de sinistres et la loi de Pareto leurs montants.

La loi de Poisson, aussi appelée la loi des événements rares, permet de modéliser des événements dont la fréquence est faible. Une telle loi est caractérisée par un seul paramètre λ qui est égal à sa moyenne et à sa variance. La fonction de masse de la loi de Poisson est donnée par :

$$\forall k \geq 0, \quad p_k = \frac{\lambda^k}{k!} e^{-\lambda}$$

La loi de Pareto permet de modéliser des distributions tronquées. Pour les sinistres CAT, les données sont tronquées par un montant, inférieur à la priorité du contrat, qui représente le seuil à partir duquel les sinistres sont partagés aux réassureurs. Une loi de Pareto est caractérisée par deux paramètres : x_0 représentant la troncature ou le plus petit sinistre disponible, et α qui représente la forme de la courbe. La fonction de répartition de la loi de Pareto est donnée par :

$$\forall x > x_0, \; F(x) = 1 - \left(\frac{x_0}{x}\right)^{\alpha}$$

Pour les tempêtes, α est communément choisi entre 0,7 et 1,5. À partir des n sinistres revalorisés x_i supérieurs à x_0, une estimation plus précise est donnée par la formule :

$$\alpha = \frac{n}{\sum_i \ln\left(\frac{x_i}{x_0}\right)}$$

Avec ces deux lois nous pouvons ensuite simuler un grand nombre d'années en simulant, pour chacune d'entre elles, le nombre d'événements attendu N par la loi de Poisson puis les N montants de ces événements par la loi de Pareto.

La suite de la méthode est la même que pour la méthode des *burning costs*. Nous appliquons les conditions du contrats aux sinistres pour obtenir les charges et les taux de charge annuelle pour le réassureur. Le taux pur est un taux moyen calculé à partir de ces taux de charge qui est complété ensuite par des chargements.

Les modèles vendus sur le marché

Dans le cadre de la tarification de traités de réassurance couvrant les tempêtes, la fiabilité des deux méthodes précédentes est très limitée par le manque de données. En raison de la faible fréquence des tempêtes en Europe et des priorités élevées, peu d'événements historiques sont disponibles. Il est par ailleurs courant que les assureurs transmettent aux réassureurs les informations sinistres sur une période limitée et il devient alors difficile de juger de la représentativité statistique de ladite période.

Afin de pallier ce manque de données, des modèles ont commencé à être développés à partir de 1990, suite à la survenance des tempêtes Daria, Herta et Vivian. Aujourd'hui, trois sociétés sont spécialisées dans ce type de modélisations : AIR, RMS et EQECAT.

Ces modèles sont composés de trois modules :

- Un **premier module** génère des événements fictifs à partir de modèles météorologiques complexes.
- Un **deuxième module** estime les dégâts causés à partir des expositions et des intensités locales des scénarios générés par le premier module.
- Un **troisième et dernier module** en déduit les pertes pour l'assurance selon les conditions des contrats.

Ils permettent ainsi d'étendre la période d'observation sur des centaines d'années simulées et d'augmenter le nombre d'événements à des milliers. Ces modèles, qui sont aujourd'hui largement préférés par les assureurs et réassureurs français pour la tempête, ont toutefois des limites.

Pour commencer, ils sont fortement dépendant de la qualité des données sous-jacentes. Même si, en France, les données météorologiques enregistrées sont complètes et disponibles sur un historique long, les données sur les biens assurés sont moins bonnes. En effet, la localisation de ces biens est parfois erronée notamment pour les résidences secondaires qui sont souvent rattachées à une adresse principale, similairement les risques industriels sont souvent rattachés à l'adresse du siège social. De plus, comme nous l'avons évoqué plus tôt, la valeur des biens est rarement évaluée précisément. Or ces données sont tout aussi importantes que les données météorologiques.

Par ailleurs, ces modèles renvoient directement des coûts pour l'assureur à partir de ses expositions et des conditions de ses contrats mais, en raison de la complexité des modèles sous-jacents, ces montants, et les ajustements qui ont souvent besoin d'être faits, sont difficiles à justifier et expliquer.

Enfin, ils ont un coût non négligeable et le temps nécessaire pour obtenir un résultat peut être très contraignant en particulier lors du dernier trimestre qui correspond, pour les réassureurs français, à la campagne de renouvellement au 1er janvier et lors duquel l'activité est largement accrue.

Les apports de notre modèle

Notre modèle ne peut pas complètement remplacer les modèles du marché car nous ne disposons pas des compétences nécessaires pour modéliser la survenance d'événements météorologiques. Nous ne sommes pas capables de générer des scénarios fictifs de tempêtes, néanmoins, notre modèle correspond aux deuxième et troisième modules. En partant de scénarios existant, caractérisés seulement par des vitesses de vent par département, et des expositions, nous pouvons estimer les pertes pour l'assurance.

Cela nous fournit une nouvelle méthode de calcul *as-if* des coûts des tempêtes historiques. Nous disposons de onze scénarios historiques qui ont touché la France et des expositions de l'année en cours ce qui nous permet d'estimer les coûts de ces scénarios s'ils avaient lieu aujourd'hui sur les nouveaux portefeuilles. Ce calcul d'*as-if* est plus correct que la simple indexation des montants des tempêtes historiques puisqu'il prend aussi en compte l'évolution du portefeuille d'une année à l'autre.

Il sera particulièrement utile dans le cas de grands changements de portefeuille. Par exemple, lorsque Klaus a frappé la France en 2009, de nombreuses forêts étaient assurées et sont en grande partie responsables du montant élevé de la tempête. Suite à cet événement, les portefeuilles ont beaucoup évolué en assurant notamment beaucoup moins de forêts.

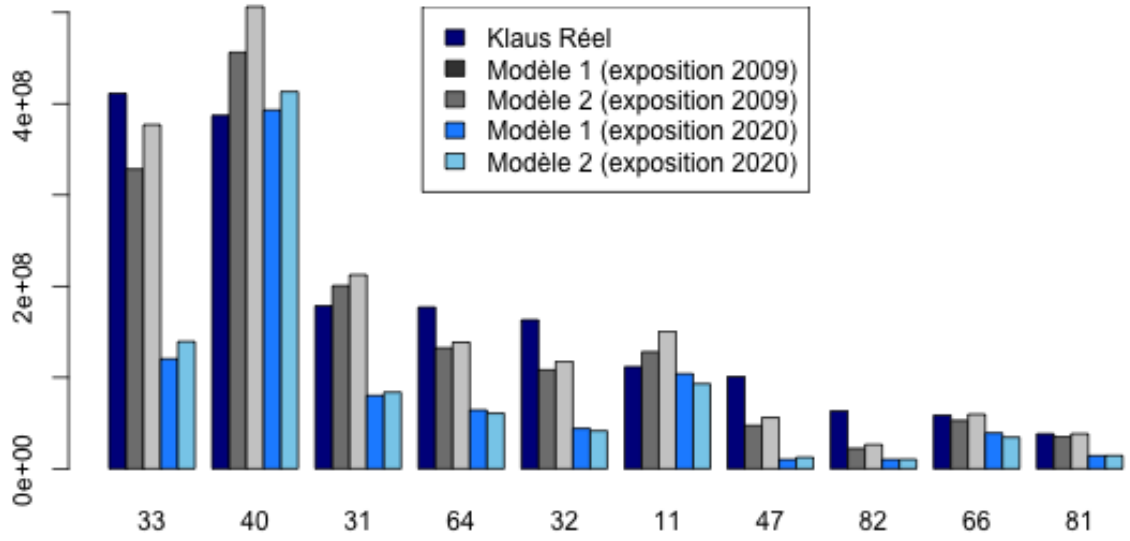


FIGURE 5.14 – Comparaison des estimations pour le scénario Klaus sur les expositions de 2009 et de 2020 pour les dix départements les plus touchés

En regardant les dix départements les plus touchés par Klaus, qui représentent à eux seuls plus de 85% des pertes, nous pouvons déjà remarquer les écarts majeurs des estimations selon le portefeuille considéré. Pour la plupart des départements, les estimations sur le portefeuille 2009 et les coûts réels revalorisés sont plus du double des estimations sur le portefeuille 2020.

Globalement, les pertes causées par Klaus qui, seulement revalorisées, atteignaient presque 2 milliards€, sont plus proches du milliard€ lorsque nous considérons en plus les changements d'expositions. Elles sont plus précisément estimées à 1,17 et 1,21 milliards€ par notre premier et deuxième modèle respectivement.

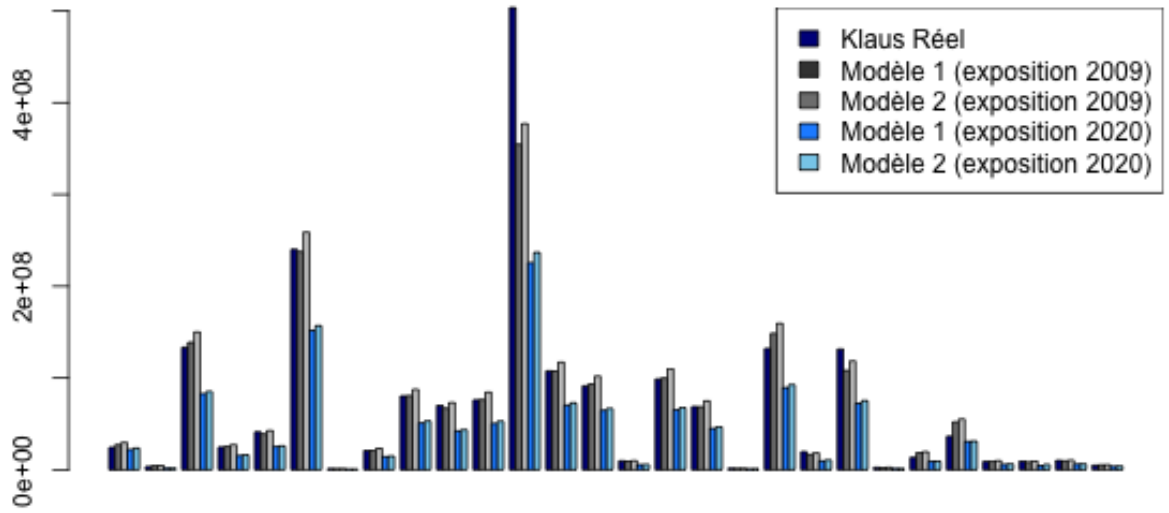


FIGURE 5.15 – Comparaison des estimations pour le scénario Klaus sur les expositions de 2009 et de 2020 par cédante

Après application des différent traités de réassurance, la charge à la part d'OdysseyRe qui dépasse, pour les montants réels revalorisés, 30 millions€, n'atteint même plus les 10 millions€ en considérant les évolutions du portefeuille. Les coûts pour OdysseyRe, en partant des nouvelles estimations de nos deux modèles sont de 8 et 9 millions€ respectivement.

5.3.2 Application au provisionnement

Cet outil apportera aussi une grande aide au calcul des provisions suite à la survenance d'une tempête. En effet, il permet de donner une première estimation rapide des coûts liés à une tempête à partir de peu d'informations sur celle-ci.

Lorsqu'une tempête frappe la France, les premières informations disponibles à son sujet sont les vitesses de vent remarquables. Comme nous disposons de toutes les expositions de l'année, ces vitesses nous permettent d'avoir, dans un premier temps, des estimations pour les départements les plus touchés. L'arrivée, plus tard, de nouvelles vitesses de vent et des premières estimations du montant marché nous permet dans la suite d'ajuster nos premières estimations.

Ces estimations seront particulièrement utiles dans le cas où une tempête touche la France quelques jours avant la clôture du trimestre comme cela avait été le cas avec Lothar et Martin. Dans cette situation, nous devons calculer des réserves avec très peu de temps et d'informations sur l'événement.

Nous étudions dans la suite le cas de la tempête Klaus.

Région	Département	Poste	Vent Météo France	GustKMH
Languedoc-Roussillon	66	Cap Béar	191	127
Languedoc-Roussillon	66	Perpignan	184	127
Languedoc-Roussillon	66	St Paul de Fenouillet	177	127
Aquitaine	33	Lège Cap Ferret	173	141
Aquitaine	40	Biscarrosse	173	158
Aquitaine	33	Bordeaux-Mérignac	160	141
Languedoc-Roussillon	11	Narbonne	159	155
Midi-Pyrénées	65	Vic en Bigorre	157	143
Languedoc-Roussillon	11	Leucate	155	155
Aquitaine	64	Pointe de Socoa	151	149
Midi-Pyrénées	31	St Felix Lauragais	150	130

Table 5.1 – Vitesses de vent diffusées par Météo France après la survenance de Klaus et vitesses Perils correspondantes

En 2009, suite à la survenance de Klaus, Météo France a donné les vitesses remarquables de la Table 5.1. Ces vitesses remarquables sont les plus grandes rafales observées durant la tempête. Cela explique qu'elles soient largement supérieures aux vitesses de Perils qui sont des vitesses moyennes. Ce type de données sont les premières informations disponibles sur les tempêtes suite à leur survenance.

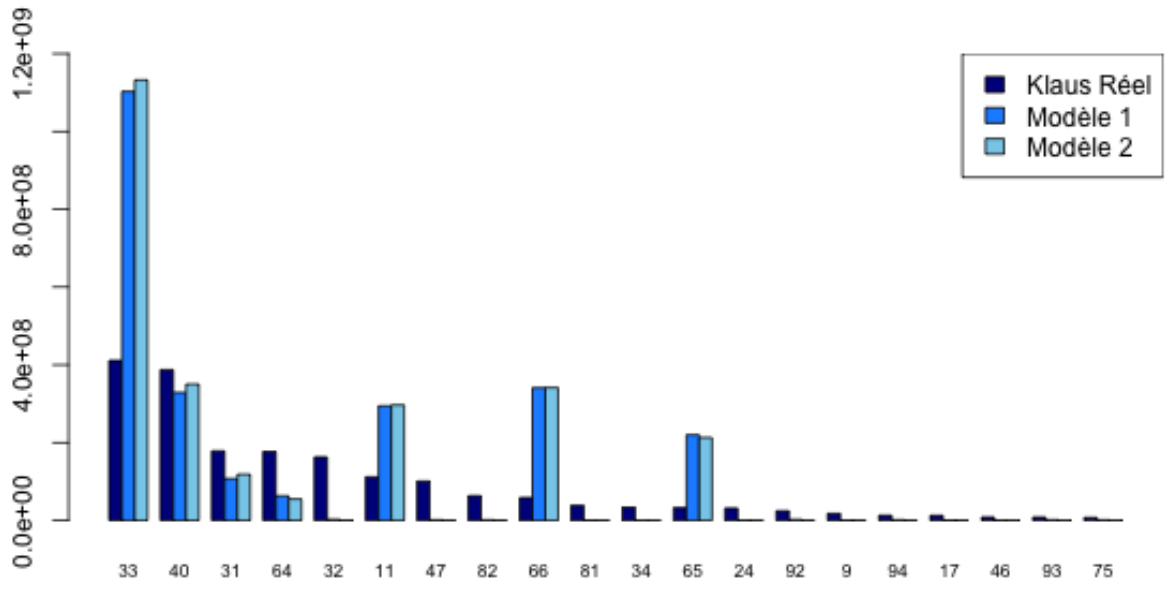


Figure 5.16 – Estimations obtenues avec les vitesses remarquables de sept départements

Pour Klaus nous avons ainsi des mesures de l'intensité de l'événement dans sept départements, lorsque nous avons plusieurs vitesses pour un même département, nous en retenons pour l'instant la moyenne. Avec ces vitesses, et en fixant les vitesses des autres départements à 0, nous pouvons alors faire tourner notre modèle sur les expositions de l'année 2009. Nous obtenons ainsi les estimations de la FIGURE 5.16.

Plusieurs départements sont beaucoup sur-estimés, ce qui est cohérent avec les écarts de vitesses remarqués précédemment. Tous les départements pour lesquels nous n'avons pas de vitesse sont estimés à 0 par le deuxième modèle ce qui est bien puisque sans vent il ne peut pas y avoir de dommages liés au vent. Le modèle 1, en revanche, prédit des pertes strictement positives mais celles-ci sont négligeables par rapport aux montants les plus élevés. Elles ne sont d'ailleurs pas visibles sur l'histogramme.

Sur la France entière, nos modèles estiment des pertes totales autour de 2,5 milliards€ alors que les pertes réelles sont de 2,0 milliards€.

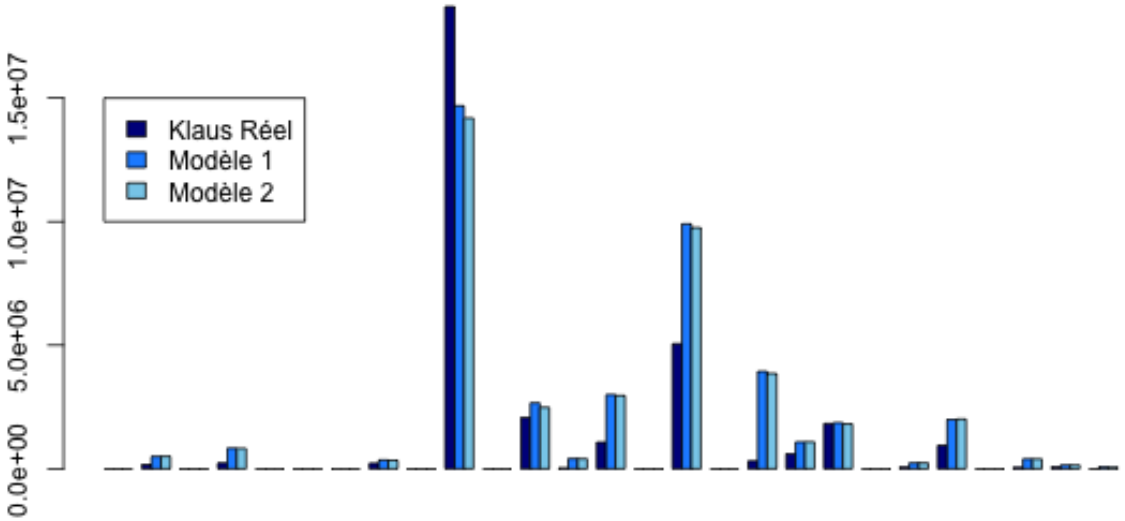


FIGURE 5.17 – Estimations des coûts pour OdysseyRe par cédante obtenus avec les vitesses remarquables de sept départements

Au niveau de la réassurance, les coûts pour OdysseyRe par cédante, représentés par le FIGURE 5.17, sont plus proches des montants réels mais ils sont toujours sur-estimés. Nous arrivons ainsi à un montant à la charge d'OdysseyRe de 42 millions€ alors que le coût réel est de 31 millions€.

Ces estimations nécessitent quelques ajustements. Ces ajustement peuvent concerner dans un premier temps les vitesses de vent disponibles. Comme ce sont les vitesses de rafales et non les vitesses de vent moyen, nous pouvons les réduire. Les FIGURES 5.18 et 5.19 correspondent à des vitesses diminuées de 10 km/h.

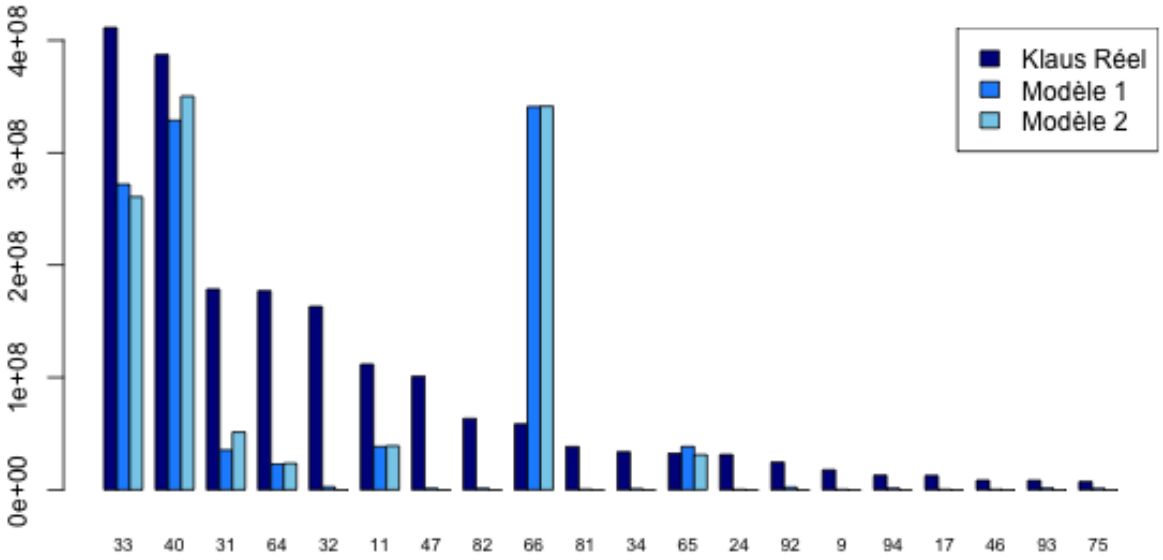


FIGURE 5.18 – Estimations obtenues avec les vitesses remarquables ajustées de sept départements

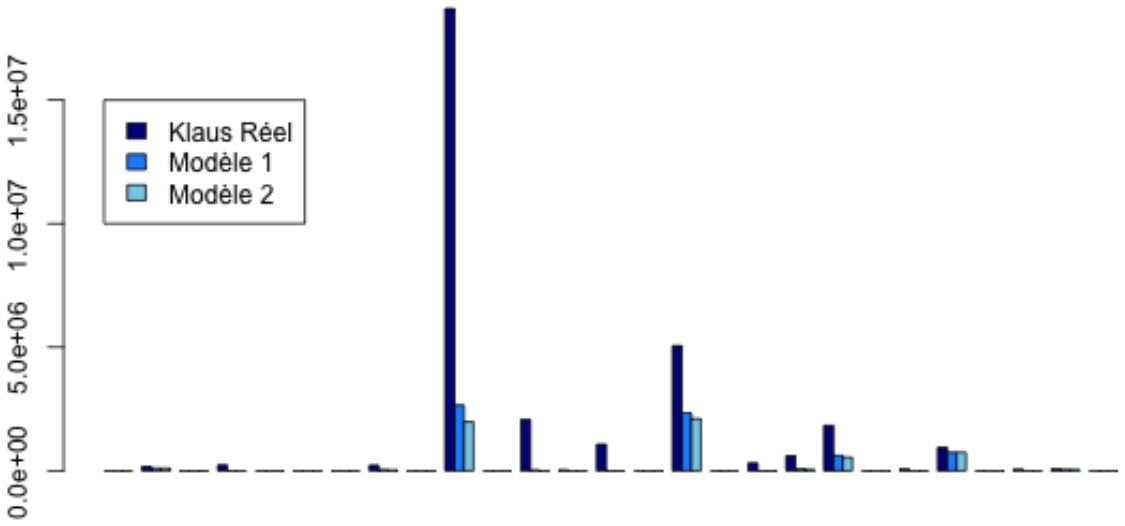


FIGURE 5.19 – Estimations des coûts pour OdysseyRe par cédante obtenues avec les vitesses remarquables ajustées de sept départements

Les estimations par département sont désormais largement sous-estimées sauf pour le département 66 pour lequel les rafales sont particulièrement élevées. Avec un tel ajustement, le montant total sur l'assurance en France descend à 1,2 milliards€. Le coût pour OdysseyRe est aussi maintenant largement sous-estimé et passe à 7 millions€.

Un ajustement département par département est peut-être plus judicieux. Par ailleurs, nous pouvons considérer, plutôt que la moyenne, le minimum des vitesses disponibles pour un même département par exemple. Selon les informations à disposition, nous pouvons aussi ajouter des vitesses non nulles aux autres départements que nous savons touchés par la tempête.

Enfin, nous pouvons réaliser des ajustements, plus tard, sur les montants à l'aide d'un montant marché global, ou encore avec des montants marché par type de risque.

5.3.3 La tempête Alex

Alex est la première tempête de la saison 2020/2021, elle a touché principalement la Bretagne, la Manche et la Loire-Atlantique dans la nuit du jeudi 1er octobre au vendredi 2 octobre 2020. Elle nous permet d'effectuer les premiers tests de nos modèles en situation réelle.

Cet événement est particulièrement précoce puisqu'en France les tempêtes sont le plus souvent observées en hiver, entre décembre et février. Historiquement, peu de tempêtes sont survenues en octobre, et presque toutes étaient de faible intensité. Nous retenons tout de même un événement historique d'intensité remarquable en octobre : l'ouragan de 1987, évoqué dans la partie 1.1.2, qui a touché la Bretagne et les côtes de la Manche entre le 15 et le 16 octobre.

Par ailleurs, Alex a la particularité d'être un événement multi-périls. En effet, les vents violents ont été accompagnés de fortes averses provoquant de nombreuses inondations notamment dans le sud du pays. Les inondations représentent une grande partie des sinistres causés par Alex, cependant, notre outil n'évalue que les dégâts causés par le vent.

Premières alertes et prévisions des vitesses de vent

Les premières alertes concernant Alex ont été partagées entre le 30 septembre et le 1er octobre 2020 par Perils, Météo France et d'autres sites de suivi des catastrophes naturelles tels que *CatNat.net*. Ces alertes sont accompagnées de premières estimations des vitesses de vent attendues lors de la tempête.

Météo France prévoit ainsi des rafales de vent entre 100 et 120 km/h dans le Nord-Ouest du pays et entre 130 et 140 km/h sur le littoral sud de la Bretagne.

CatNat.net prévoit quant à lui des vitesses de vent moyen entre 70 et 80 km/h et des rafales entre 100 et 130 km/h dans les terres, entre 120 et 150 km/h sur les côtes et des pointes de 150 à 170 km/h sur les îles.

Enfin, Perils donne des estimations des vitesses de vent moyen pour chaque département, nous retenons surtout les quatre vitesses supérieures à 100 km/h de la Table 5.2. Ces informations, que nous n'avions pas pour Klaus, nous permettront d'avoir des estimations pour tous les départements.

CRESTA	Département	Vitesse de vent
56	Morbihan	117 km/h
44	Loire-Atmantique	111 km/h
35	Ille-et-Vilaine	106 km/h
29	Finistère	100 km/h

Table 5.2 – Prévisions Perils de vitesses de vent supérieures à 100 km/h

Vitesses de vent observées

Dès le 2 octobre, juste après le passage de la tempête, Météo France et *CatNat.net* partagent les vitesses de vent remarquables enregistrées par les différentes stations météorologiques au cours de la tempête. Nous disposons ainsi des vitesses remarquables présentées dans la table 5.3 [1]

CRESTA	Département	Vitesse de vent
22	Côtes d'Armor	111 km/h
35	Ille-et-Vilaine	116 km/h
49	Maine-et-Loire	129 km/h
50	Manche	142 km/h
56	Morbihan	186 km/h
56	Morbihan	157 km/h
56	Morbihan	134 km/h
56	Morbihan	132 km/h

Table 5.3 – Vitesses remarquables observées lors de la tempête Alex

Ces vitesses se rapprochent toutes des vitesses records enregistrées par leurs stations respectives. Le problème de ces valeurs est qu'elles ont un biais dû à la localisation des différentes stations par lesquelles elles sont enregistrées. Par exemple, les deux plus grandes valeurs, les 157 et 186 km/h observées dans le Morbihan, ont été enregistrées sur des îles qui sont des zones où le vent souffle naturellement plus fort. De même, les stations se situant en altitude ou encore sur des pointes enregistrent naturellement des vitesses de vent plus élevées. Ces vitesses nous donnent néanmoins de bonnes informations sur les zones les plus touchées par l'événement et pourront ainsi guider nos choix d'ajustements.

1. Comme les vitesses provenant de nos deux sources coïncident, nous les avons rassemblées dans une seule et même colonne.

Nos estimations

Avec ces différentes vitesses, nous avons suffisamment d'informations pour donner une première estimation avec nos modèles. En nous basant sur les vitesses de vent prévisionnelles de Perils, nous estimons alors des pertes globales, pour la tempête, s'élevant à 85 millions€ et 70 millions€ avec les modèles 1 et 2 respectivement.

Ces estimations sont les résultats directs des modèles. Par expérience et à partir des premières informations disponibles, nous choisissons de faire un scénario ajusté à 100 millions€ en conservant la même répartition des sinistres par département.

Enfin, nous avons une cédante dont les polices sont très concentrées en Bretagne et qui est, par conséquent, susceptible d'être plus touchée que les autres et de se rapprocher de sa priorité. Pour cette raison, nous créons un dernier scénario "extrême" considérant que seule la Bretagne est touchée et que nous ajustons aussi à 100 millions€.

Les estimations de ces trois différents scénarios pour les 15 cédantes les plus touchées sont résumées dans la TABLE 5.4. Le scénario 1 correspond au modèle 1, les scénarios 2 et 3, au modèle 1 ajusté et au scénario extrême respectivement.

CÉDANTE	SCÉNARIO 1	SCÉNARIO 2	SCÉNARIO 3
Cédante 1	19,7	28,8	33,1
Cédante 2	7,4	10,7	9,8
Cédante 3	5,6	8,2	8,5
Cédante 4	4,1	5,9	4,1
Cédante 5	4,2	6,1	6,4
Cédante 6	3,7	5,4	5,4
Cédante 7	3,6	5,2	4,9
Cédante 8	3,1	4,6	4,1
Cédante 9	2,7	4,0	4,1
Cédante 10	2,0	2,9	0,8
Cédante 11	2,3	3,3	2,6
Cédante 12	1,9	2,8	2,8
Cédante 13	1,9	2,7	2,1
Cédante 14	1,8	2,6	5,3
Cédante 15	0,9	1,3	1,2

TABLE 5.4 – Estimations des coûts de la tempête Alex par cédante (en millions€)

En plein renouvellement, ces estimations nous permettent d'avoir en tête un ordre de grandeur des chiffres qui devraient nous être transmis et ainsi d'être mieux préparés pour les rencontres ou les échanges avec nos cédantes.

Les estimations des différents acteurs

Perils est le premier à donner une estimation des pertes dès le début de la tempête. Ces estimations sont basées sur ses propres estimations des vitesses de vent et non sur des observations de vitesses de vent ou des premiers dégâts constatés. Ainsi, le 30 septembre 2020, il estime un montant global s'élevant à 12 millions€. Ce n'est que plus tard qu'il donnera le montant réel obtenu à partir des données sinistres qu'il reçoit de la part des assureurs.

Après le passage de la tempête, les différents acteurs partagent à leurs tours leurs estimations respectives. Cependant, toutes ces estimations sont à prendre avec précautions car ces montants ne représentent pas toujours la même chose.

Ainsi, le 6 octobre, *CatNat.net* estime un montant s'élevant à 500 millions$ (soit 425 millions€). Ce montant s'éloigne beaucoup de nos estimations car c'est un montant économique qui prend en compte tous les dommages et non seulement les dommages assurés. En particulier, l'État étant son propre assureur, la plupart des routes et des ponts détruits ne sont pas assurés et ne sont donc pas pris en compte dans nos estimations. Par ailleurs, ce montant cumule les deux périls, la tempête et les inondations, alors que nous n'estimons que les dégâts causés par les tempêtes.

Au cours du mois d'octobre nous recevons ensuite, de la part de nos cédantes, leurs estimations des coûts d'Alex pour chacune d'entre elles. La Table 5.5 résume les dates auxquelles les estimations nous ont été données ainsi que les écarts entre les estimations des cédantes et les nôtres.

Cédante	Date	Écart	Périls
Cédante 3	07/10/2020	+7%	Tempête
Cédante 4	16/10/2020	+28%	Tempête
Cédante 6	07/10/2020	-36%	Tempête
Cédante 9	07/10/2020	-26%	Tempête
Cédante 14	15/10/2020	+71%	Tempête+Inondations

Table 5.5 – Estimations de nos cédantes

Ainsi, pour la cédante 3 par exemple, la cédante estime un montant supérieur de 7% à notre estimation alors que la cédante 6 estime un montant inférieur de 36% au nôtre.

Nous remarquons ici que l'écart entre l'estimation de la cédante 14 et la nôtre est particulièrement élevé par rapport aux autres cédantes ayant fourni une estimation. Cependant cet écart est probablement dû à l'inclusion des sinistres liés aux inondations en plus des sinistres tempêtes.

Bilan

Avec notre modèle et les informations disponibles, nous avons été capables de donner une estimation du coût de la tempête au moment même de sa survenance. Par ailleurs, ces estimations sont cohérentes avec les estimations de nos cédantes, pour les cédantes qui en ont fourni une.

Du point de vue de la réassurance, Alex est une tempête de faible intensité qui inquiète peu. Nos trois scénarios estiment qu'aucun de nos traités de réassurance ne sera touché. De même, nos cédantes estiment des montants bien inférieurs à leurs priorités respectives.

Tous ces résultats restent cependant des estimations. Nous ne recevrons les montants réels définitifs que bien plus tard, s'ils nous sont transmis.

Chapitre 6

Indice de vent

Pour finir, dans ce dernier chapitre nous envisageons l'utilisation d'un indice de vent, en nous inspirant des travaux de A. Mornet et al. sur le sujet [12], afin d'améliorer nos résultats.

Principe

L'idée derrière la création d'un indice de vent est de prendre en compte la distribution des vitesses de vent propre à chaque département. Cela permet d'identifier les vitesses exceptionnelles spécifiques à chaque département et ainsi de prendre en compte le fait que les départements les plus exposés aux vents sont aussi les mieux préparés pour faire face à des vents violents, notamment en terme de normes de construction.

Plusieurs formules d'indice se basant sur la vitesse de vent $w^j(k)$ de la station k à la date j et le quantile $w_q(k)$ d'ordre q de la distribution de la station k ont été utilisées dans le passé. A. Mornet et al. proposent un indice de la forme suivante :

$$I_w^j(k) = \left(\left[w^j(k) - w_q(k) \right]_+ \right)^\alpha$$

Avant eux, deux autres indices basés sur le rapport, plutôt que la différence, entre la vitesse et un quantile ont été proposés par M. Klawa et U. Ulbrich [7] et M.G. Donat et al. [2] respectivement :

$$CI(k) = \left(\frac{w(k)}{w_q(k)} - 1 \right)^\alpha$$

$$LR(k) = A(k) \left(\left[\frac{w(k)}{w_q(k)} - 1 \right]_+ \right)^\alpha + B(k)$$

Données supplémentaires

Les données de vent utilisées jusqu'ici pour construire nos modèles sont très limitées. En effet, avec Perils, nous n'avons accès qu'aux vitesses de vent des tempêtes ayant dépassé un coût total de 200 millions€ en Europe. Ces données sont largement insuffisantes pour construire un indice tel que présenté précédemment, nous avons donc besoin de nouvelles données.

La meilleure source pour obtenir les distributions de vent par département est Météo France, cependant ces données sont très chères [1]. Nous utiliserons donc plutôt une base mise à notre disposition par *CatNat.net* qui recense toutes les vitesses de vent moyen ayant dépassé 90km/h depuis 2016. Cette base nous apporte un peu plus d'informations que les données de Perils.

Dans cette nouvelle base, nous avons plus précisément toutes les vitesses de vent moyen dépassant 90km/h enregistrées toutes les heures sur une période de près de 2 000 jours. Pour chacun de ces jours et par département, nous retenons la vitesse maximale enregistrée dans la journée. Sur les 190 000 mesures que cela représente, nous n'avons que celles qui dépassent 90km/h, ce qui représente seulement 3 000 mesures, soit 1,6% du total. La TABLE 6.1 donne la proportion de jours pour lesquels nous avons des données pour chaque département.

CRESTA	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
% JOURS	1,1	0,4	1,0	2,7	3,3	3,6	2,7	0,3	3,3	0,5	7,9	4,0	3,3	1,4	2,9	1,1

CRESTA	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
% JOURS	1,9	1,0	0,6	6,9	0,9	2,7	1,3	0,6	0,6	2,7	0,7	0,4	3,8	4,7	3,4	0,3

CRESTA	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48
% JOURS	1,2	4,3	1,3	1,2	0,3	3,1	0,9	1,1	0,4	2,0	2,7	1,9	0,4	0,5	0,3	3,5

CRESTA	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64
% JOURS	0,9	2,9	0,9	0,9	0,4	0,6	0,7	0,4	2,9	0,7	0,4	1,6	0,8	1,0	3,0	3,1

CRESTA	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80
% JOURS	3,3	7,6	0,9	1,2	1,5	0,4	1,2	0,3	3,4	2,6	0,1	2,1	0,8	0,5	1,5	1,5

CRESTA	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95
% JOURS	4,1	0,5	3,3	2,7	2,0	1,3	0,7	0,9	0,4	0,4	0,4	0,1	0,2	0,1	0,4

TABLE 6.1 – Proportions de jours pendant lesquels une vitesse supérieure à 90km/h est observée

1. Le coût est de l'ordre de 40 centimes par lot de 10 mesures.

Les données de cette base sont encore très limitées puisque pour certains département tels que les Hauts-de-Seine (92), nous n'avons que les 0,1% supérieurs de la distribution. Pour près de 80% des départements, nous avons tout de même au moins 0,5% de la distribution. Cela risque d'être insuffisant puisque l'indice utilisé par Mornet et al. utilise le quantile d'ordre 99% et les deux autres indices utilisent le quantile d'ordre 98%.

Application

En raison des limites des données à notre disposition, nous avons peu de choix pour le quantile q, nous construisons donc un indice avec le quantile d'ordre 99,5%. Pour les départements pour lesquels il nous manque des données, nous fixons le quantile à 90km/h.

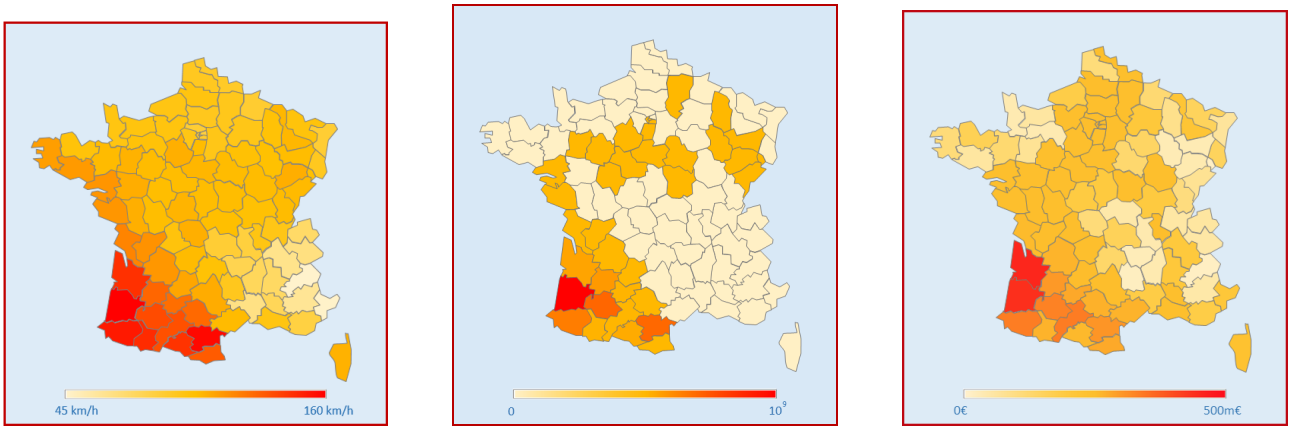
Cela nous laisse alors un seul paramètre à choisir : la puissance α. Nous regardons donc comment évolue la corrélation entre l'indice et la sinistralité observée selon le α choisi. Les corrélations obtenues sont présentées dans la TABLE 6.2.

VARIABLE	CORRÉLATION
Indice $\alpha = 1$	0,105
Indice $\alpha = 2$	0,130
Indice $\alpha = 3$	0,146
Indice $\alpha = 4$	0,154
Indice $\alpha = 5$	0,157
Indice $\alpha = 6$	0,158
Indice $\alpha = 7$	0,157
Vitesse de vent	0,225

TABLE 6.2 – Corrélations avec la sinistralité observée

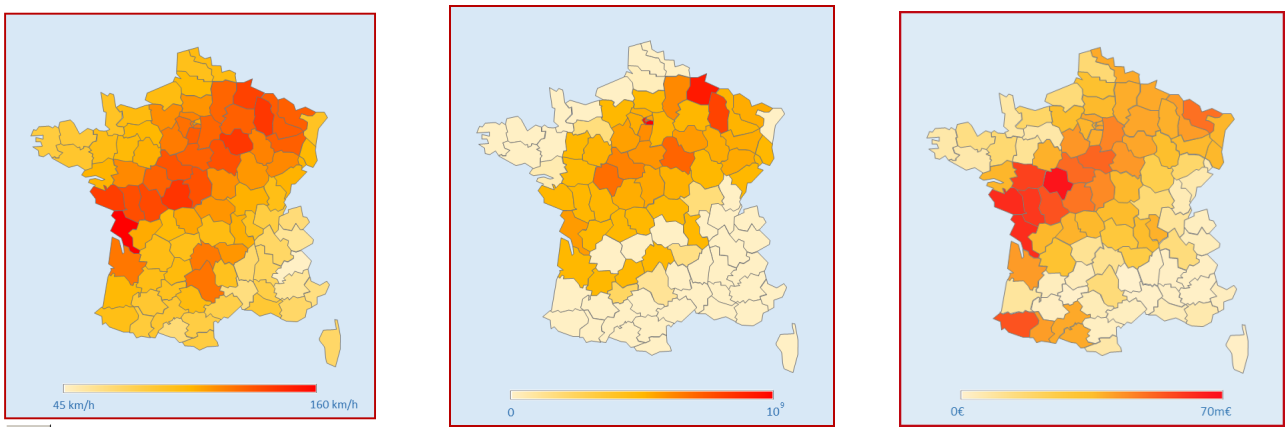
Les corrélations de chaque indice avec les sinistres observés sont faibles. Le meilleur indice est celui avec un α égal à 6 qui a une corrélation de 0,158 avec les sinistres observés. Nous remarquons cependant que cet indice est moins corrélé aux sinistres observés que les simples vitesses de vent. Cet indice est donc a priori moins bon que les vitesses de vent pour évaluer les sinistres.

Afin de mieux visualiser ces écarts de corrélation, Les FIGURES 6.1 et 6.2 représentent les vitesses de vent, les indices de vent et les sinistres sur des cartes pour les deux plus grosses tempêtes de notre base, Klaus et Xynthia. Nous observons ici aussi que l'indice de vent est mauvais pour prédire la sinistralité et rapprocher les vitesses de vent aux sinistres. Les résultats obtenus avec un autre α ou plutôt avec le rapport $\frac{w(k)}{w_q(k)}$ sont similaires et aussi mauvais qu'avec cet indice.



(a) Vitesses de vent (b) Indices de vent (c) Sinistres

FIGURE 6.1 – Klaus



(a) Vitesses de vent (b) Indices de vent (c) Sinistres

FIGURE 6.2 – Xynthia

Les mauvais résultats obtenus sont probablement dus aux multiples biais apportés par la nouvelle base de données utilisée pour construire cet indice. Pour commencer, cette base ne nous donne qu'une petite partie des distributions de vent de chaque département puisqu'elle ne retient que les vitesses supérieures à 90km/h. Par ailleurs, les sources fournissant les vitesses de nos deux bases ne sont pas les mêmes, les vitesses de *CatNat.net* provennant de Météo France alors que celles de Perils proviennent du modèle Cosmo-7 de MeteoSwiss. Les vitesses de *CatNat.net* sont aussi données par commune alors que Perils donne, pour chaque département, une vitesse moyenne pondérée par la population. Enfin, cette base couvre un historique court ce qui limite beaucoup le nombre de données pour certains départements et ne nous permet pas de travailler exclusivement avec ces vitesses.

Avec les données à notre disposition nous ne pouvons donc pas améliorer nos modèles en utilisant un indice de vent. Il nous faudrait pour cela une base de données météorologiques plus fines et sur un historique plus long.

Conclusion

Dans ce mémoire, nous avons construit deux modèles pour estimer les taux de dommages causés par le vent lorsqu'une tempête touche la France en partant des portefeuilles assurés et des vitesses de vent. Les deux modèles obtenus ont tendance à sous-estimer légèrement les dommages réels, le premier étant meilleur sur les plus petits événements et le deuxième sur les événements d'intensités plus élevées.

En tant que réassureurs, les plus petits événements nous intéressent peu, et nous préférerons donc, en général, plutôt notre deuxième modèle.

Celui-ci nous apportera des premières estimations rapides des coûts des tempêtes pour OdysseyRe à partir de très peu de données sur l'événement. Ces estimations seront particulièrement utiles lorsque les contraintes de temps seront fortes, en fin de trimestre par exemple, lorsque la clôture approche et qu'une estimation est nécessaire pour fixer des provisions. En fin d'année aussi, lorsque les enjeux seront encore plus grands puisque le mauvais provisionnement d'un événement peut étaler les mauvais résultats sur une deuxième année.

Ce modèle apporte aussi une nouvelle méthode de calcul des *as-if* pour les tempêtes. En effet, nous disposons d'une dizaine de scénarios historiques, caractérisés par les vitesses de vent par département, et des portefeuilles pour chaque année. Avec notre modèle, nous pouvons ainsi estimer les pertes indexées causées par chaque scénario selon le portefeuille assuré.

Les limites du modèle

Ce modèle compte malgré tout un certain nombre de biais.

Pour commencer, il repose en partie sur les sommes assurées or, en France, celles-ci sont le plus souvent mal évaluées. Elles ne sont évaluées qu'à partir des surfaces habitables et des nombres de pièces et les méthodes d'évaluation de la valeur des biens varient d'un assureur à l'autre.

Par ailleurs notre base de données n'est constituée que de deux tempêtes moyennes, Klaus et Xynthia et d'autres petites tempêtes. Elle manque de plus grosses tempêtes telles que celles qui ont eu lieu en 1987, 1990 et 1999. Cet absence d'événement d'intensité plus importante risque de limiter grandement les taux de dommages prédits. Par conséquent, si une grosse tempête venait à se produire, nos modèles risquent de beaucoup les sous-estimer.

Développements futurs possibles

Cet outil pourra être affiné régulièrement par des mises à jour. Annuellement, d'une part, il sera complété par les expositions de la nouvelle année et une nouvelle année sera aussi ajoutée à l'indice utilisé afin de revaloriser les données sur l'année en cours. D'autre part, chaque nouvelle tempête qui touche la France viendra compléter nos données et améliorer nos modèles.

L'outil pourrait aussi être amélioré en obtenant de nouvelles données. Des données plus détaillées sur les types de biens assurés ou des données de vent plus fines par exemple. La reconstitution des données pour les événements Lothar et Martin serait aussi intéressante pour palier notre manque de gros événements.

Enfin, cet outil pourra être étendu, dans un premier temps, aux autres pays affectés par les tempêtes européennes et, plus tard, aux autres périls pour lesquels Perils dispose de suffisamment données.

Bibliographie

[1] H. Chaumont. Vulnérabilité des risques assurantiels aux périls innondation et tempête. Master's thesis, ENSAE, 2015.

[2] M.G. Donat, T. Pardowitz, G.C. Leckebusch, U. Ulbrich, and O. Burghoff. High resolution refinement of a storm loss model and estimation of return periods of loss-intensive storms over Germany. *Natural Hazards and Earth System Sciences*, 2011.

[3] P.K. Dunn and G.K. Smith. *Generalized Linear Models With Examples in R.* Springer Texts in Statistics. 2018.

[4] Fédération Française du Bâtiment. Indice FFB du coût de la construction.

[5] T. Hastie, R. Tibshirani, and J Friedman. *The elements of statistical learning.* Springer Series in Statistics. 2009.

[6] G. James, D. Witten, T. Hastie, and R. Tibshirani. *An introduction to statistical learning.* Springer Texts in Statistics. 2013.

[7] M. Klawa and U. Ulbrich. A model for the estimation of storm losses and the identification of severe winter storms in Germany. *Natural Hazards and Earth System Sciences*, 2003.

[8] N. Langevin. Modélisation de la sinistralité tempête, apport de l'Open Data et du Machine Learning. Master's thesis, ENSAE, 2019.

[9] Légifrance.

[10] E. Mlynarczyk. *Techniques et pratiques de la réassurance.* 2014.

[11] G.G. Moisen. Classification and regression trees. 2008.

[12] A. Mornet, T. Opitz, M. Luzi, and S. Loisel. Index for predicting insurance claims from wind storms with an application in France. *Risk Analysis*, 2015.

[13] Météo France. Tempêtes en France métropolitaine.

Annexes

Annexe A

Compléments météorologiques

A.1 Les forces impactant les vents

A.1.1 La force de Coriolis

La force de Coriolis est une force d'inertie qui apparaît lorsqu'un référentiel est en rotation par rapport à un référentiel Galiléen au repos : la rotation d'un disque dans le référentiel terrestre par exemple.

En regardant le mouvement d'une bille sur ce disque tournant depuis le référentiel terrestre, nous observons un mouvement rectiligne alors que sur le disque, le tracé du chemin pris par la bille est un arc de cercle.

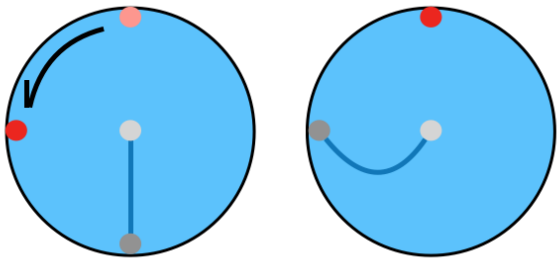


FIGURE A.1 – Mouvement d'une bille vu dans le référentiel terrestre (gauche) et dans le référentiel du disque (droite)

La Terre étant en rotation sur elle-même vers la droite, les objets en mouvement à sa surface, et notamment les vents, sont déviés vers la droite dans l'hémishphère Nord et vers la gauche dans l'hémisphère Sud, comme représenté dans la FIGURE A.2.

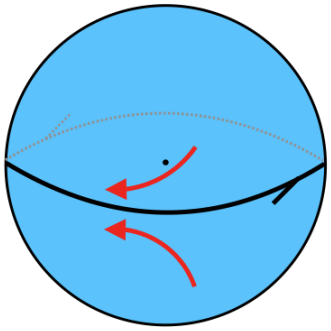


FIGURE A.2 – Force de Coriolis sur la Terre

A.1.2 La force centripète

La force centripète est une force perpendiculaire à un mouvement de trajectoire courbe qui tend à rapprocher du centre. Elle s'oppose à la force centrifuge qui tend à éloigner du centre.

Le mouvement d'une balle attachée à un fil sous tension et en rotation autour d'un axe, représenté dans la FIGURE A.3, donne un exemple de force centripète. Si la vitesse seule éloignerait la balle de l'axe de rotation, la tension est ici une force centripète qui maintient la balle à distance constante.

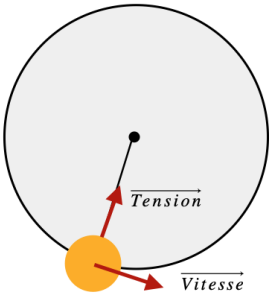


FIGURE A.3 – Exemple de force centripète

Dans le cas des vents, la force centripète dévie légèrement leur trajectoire et augmente leur vitesse à proximité des centres des dépressions.

A.1.3 La force de frottements

La surface de la Terre est très irrégulière et présente de nombreux obstacles aux vents. Si les océans reste relativement lisses, les frottements avec l'eau ralentissent tout de même les vents. Sur les continents, la présence de végétation, de villes et d'autres reliefs humains ou naturels ralentit et dévie les vents vers les zones de pression plus faibles.

A.2 L'échelle de Beaufort

L'échelle de Beaufort est une échelle de mesure de la vitesse moyenne des vents créée par l'amiral britannique Francis Beaufort en étudiant les effets du vent sur l'état de la mer et le mouvement des arbres. Il définit ainsi treize degrés de vent auxquels sont associés des effets sur la mer et sur la terre. Cela permettait alors aux marins d'avoir une idée de la vitesse du vent en l'absence d'anémomètre. Elle est adoptée internationalement comme unité de mesure de la force du vent en météorologie maritime en 1874 jusqu'à être remplacée par le nœud en 1946. Les treize degrés de Beaufort et les observations associées sont présentés dans la Table A.1

Degré	Vent	Observations maritimes	Observations terrestre
0	< 1 km/h	La mer est comme un miroir.	La fumée s'élève verticalement ; végétaux immobiles.
1	1 à 5 km/h	Il se forme des rides ressemblant à des écailles de poisson, mais sans aucune crête d'écume.	Dérive de la fumée à peine perceptible ; végétaux immobiles.
2	6 à 11 km/h	Vaguelettes, courtes encore mais plus accusées ; leurs crêtes ont une apparence vitreuse, mais elles ne déferlent pas.	Vent tout juste perçu au visage, fumée à 80° ; les feuilles frémissent.
3	12 à 19 km/h	Très petites vagues ; les crêtes commencent à déferler ; écume d'aspect vitreux ; parfois quelques moutons épars.	Fumée à 70° ; poussière soulevée ; brindilles agitées.
4	20 à 28 km/h	Petites vagues devenant plus longues ; moutons franchement nombreux.	Fumée à 50° ; cheveux dérangés et vêtements qui claquent ; petites branches agitées.
5	29 à 38 km/h	Vagues modérées prenant une forme plus nettement allongée ; naissance de nombreux moutons (éventuellement des embruns).	Fumée à 30° ; yeux gênés par les suspensions dans l'air ; sensation de picotement sur le visage si température négative ; le gréement commence à siffler.
6	39 à 49 km/h	Des lames commencent à se former ; les crêtes d'écume blanche sont partout plus étendues (habituellement quelques embruns).	Fumée à 15° ; manches gonflées par les côtés ; grandes branches agitées.
7	50 à 61 km/h	La mer grossit ; l'écume blanche qui provient des lames déferlantes commence à être soufflée en traînées qui s'orientent dans le lit du vent.	Fumée à 5 ou 10° ; picotement au visage par température inférieure à 3°C ; la marche devient difficile ; arbres entiers agités.
8	62 à 74 km/h	Lames de hauteur moyenne et plus allongées ; de bord supérieur de leur crêtes commencent à se détacher des tourbillons d'embruns ; l'écume est soufflée en très nettes traînées orientées dans le lit du vent.	Progression impossible en général ; les brindilles cassent.
9	75 à 88 km/h	Grosses lames ; épaisses traînées d'écume dans le lit du vent ; les crêtes des lames commencent à vaciller, s'écrouler et déferler en rouleaux ; les embruns peuvent réduire la visibilité.	Enfants renversés ; les branches cassent.
10	89 à 102 km/h	Très grosses lames à longues crêtes en panache ; l'écume produite s'agglomère en large bancs et est soufflée dans le lit du vent en épaisses traînées blanches ; dans son ensemble, la surface des eaux semble blanche ; le déferlement en rouleaux devient intense et brutal ; la visibilité est réduite.	Adultes renversés ; arbres déracinés. (Rarement observé à terre)
11	103 à 117 km/h	Lames exceptionnellement hautes (les navires de petit et moyen tonnage peuvent par instants être perdus de vue) ; la mer est complètement recouverte de bancs d'écume blanche élongés dans la direction du vent ; partout le bord des crêtes des lames est soufflée et donne de la mousse ; la visibilité est réduite.	Ravages étendus. (Très rarement observé à terre)
12	> 118 km/h	L'air est plein d'écume et d'embruns ; la mer est entièrement blanche du fait des bancs d'écumes dérivante ; la visibilité est très fortement réduite.	Ravages désastreux : violence et destruction. (En principe degré non utilisé)

Table A.1 – Échelle de Beaufort