Contributions à l'évaluation des risques en assurance tempête et automobile

Alexandre Mornet

THÈSE

Pour obtenir le grade de

Docteur

Délivré par l'**Université Claude Bernard Lyon I**

I.S.F.A

Laboratoire de Sciences Actuarielle et Financière

Contributions à l'évaluation des risques en assurance tempête et automobile

École Doctorale Sciences Économique et de Gestion (ED 486)

présentée et soutenue publiquement le 30 Septembre 2015

par

Alexandre Mornet

Composition du Jury

M. Stéphane Loisel (Directeur de thèse)
M. Hansjoerg Albrecher (Rapporteur)
M. Olivier Lopez (Rapporteur)
Mme. Pauline Barrieu (Examinatrice)
M. Michel Luzi (Examinateur)
M. Christian Robert (Examinateur)
M. Jean-Claude Augros (Co-directeur de thèse)

Doctorat de l'Université Claude Bernard Lyon I
I.S.F.A Arrêté du 17 Août 2006

Résumé :
Dans cette thèse nous étudions la garantie tempête consacrée aux dommages causés par le vent et un développement de l'assurance comportementale à travers le risque automobile. Nous associons des informations extérieures comme la vitesse du vent aux données de l'assurance. Nous proposons la construction d'un indice tempête pour compléter et renforcer l'évaluation des dégâts causés par les tempêtes majeures. Nous définissons ensuite un partage du territoire français en 6 zones tempêtes, dépendant des corrélations extrêmes de vent, pour tester plusieurs scénarios. Ces différents tests et considérations nous permettent d'améliorer notre indice tempête. Nous nous appuyons sur les modèles de la théorie des valeurs extrêmes pour montrer l'impact de la variabilité sur le calcul des périodes de retour et besoins en fonds propres. Nous soulignons ainsi les difficultés rencontrées pour dégager des résultats robustes en lien avec les événements extrêmes. Pour ce qui est de l'assurance automobile, nous testons différentes méthodes pour répondre aux évolutions techniques et réglementaires. Nous caractérisons la partition homme/femme en utilisant la procédure logistique, l'analyse des correspondances multiples ou les arbres de classification. Nous montrons qu'il est possible de compenser l'absence de la variable sexe par d'autres informations spécifiques à l'assuré ou à son véhicule et en particulier l'utilisation de relevés kilométriques. Enfin, nous nous intéressons à l'expérience acquise par les conducteurs novices. Nous étudions le comportement sur la route de l'assuré pour créer de nouvelles classes de risques.

Abstract :
In this Ph.D. Dissertation we study the storm guarantee dedicated to the damage caused by the wind and a development of the behavioral insurance through the automobile risk. We associate external information like the wind speed to insurance data. We propose the construction of a storm index to complete and strengthen the evaluation of the damages caused by the major storms. Then we define a partition of the French territory in 6 zones storms, depending on extreme wind correlations to test several scenarios. These various tests and considerations allow us to improve our storm index. We lean on extreme value theory models to show the impact of the variability on the calculation of return periods and capital requirements. We underline the difficulties to obtain strong results in connection with the extreme events. Concerning car insurance, we test various methods to answer the technical and legal evolutions. We characterize the man/woman partition by using the logistic procedure, the multiple correspondence analysis or the classification trees. We show that it is possible to compensate for the absence of the sex variable with other information specific to the insurants or to their vehicle and in particular the use of kilometric data. Finally, we are interested in the acquired experience by young drivers. We study the behavior on the road of the insurants to create new classes of risks.

Keywords : Storm Index, Wind Speed, Insurance, Extreme Value Theory, Extreme Dependence, Return period, Solvency II, Pay-as-you-drive.

Adresse : Laboratoire de Sciences Actuarielle et Financière
50, Avenue Tony Garnier
69366 LYON CEDEX 07

Remerciements

Je tiens d'abord à remercier Pauline Barrieu, Hansjoerg Albrecher, Olivier Lopez et Christian Robert d'avoir accepté de faire partie du jury lors de la soutenance de cette thèse.
Ce travail n'aurait jamais été possible sans les équipes enseignante et administrative de l'ISFA qui m'ont soutenu durant la poursuite de mes études suite à mon accident. Je remercie successivement Daniel Serant, Jean-Claude Augros et Nicolas Leboisne qui ont oeuvré pour que ma reprise se fasse dans de bonnes conditions.
Je suis très reconnaissant envers Luc Hartemann, mon professeur de droit, qui n'a pas hésité à se déplacer à Montpellier durant ma convalescence et avec qui j'ai pu valider mon premier partiel. Il n'a pas cessé de me soutenir et nous sommes désormais amis.
Je dois beaucoup à Stéphane Loisel, d'abord pour avoir organisé la logistique me permettant d'accéder aux cours et participé aux ascensions épiques des escaliers menant à l'ISFA lorsque les ascenseurs du bâtiment étaient en panne. Ensuite pour m'avoir extrêmement bien conseillé et su me motiver durant les quelques années nécessaires à l'aboutissement de cette thèse. J'en profite pour remercier chaleureusement Anne Eyraud-Loisel qui m'a aussi soutenu à de nombreuses reprises.
Je remercie également Pauline Barrieu pour sa participation et ses conseils au début de mes investigations.
Cette thèse est aussi le fruit d'un partenariat qui n'aurait pas vu le jour sans l'implication de Jean Michel Eyraud, avec deux équipes d'Allianz. Les nombreux échanges que j'ai eus avec Michel Luzi ont été essentiels pour moi. Ils ont ponctués l'avancée de mes recherches sur les tempêtes et même au-delà de cette problématique ils m'ont appris énormément sur le fonctionnement des assurances et sur la façon de structurer mes résultats. J'ai aussi beaucoup apprécié les réunions toujours conviviales avec Patrick Leveillard et son équipe, en particulier Laurence Serant qui a pris le temps de m'éclairer sur le fonctionnement des bases de données et de la tarification automobile. Je remercie enfin Bernard Bailleul qui a pris le relais de M Luzi et m'a tout de suite mis en confiance dans la poursuite de mon travail.

J'ai bénéficié au cours de mes travaux d'un crédit de recherches de la part de Météo France et je remercie ici Annick Auffray pour son soutien.

J'ai une pensée pour mes professeurs et amis Michèle Devallet et Michel Valadier qui ont donné sans compter de leur temps pour m'aider dans mes traductions d'anglais et ma compréhension des probabilités depuis mon retour à Montpellier.

C'est aussi à Montpellier que j'ai rencontré Thomas Opitz alors doctorant à l'Institut de Mathématiques et de Modélisation de Montpellier et désormais membre de l'Unité de Recherche INRA de Biostatistique et Processus Spatiaux. Je le remercie pour ses judicieux conseils dans mes recherches notamment en programmation informatique et en théorie des valeurs extrêmes.

Je me souviens également de mon professeur de Mathématiques en classe préparatoire, Jean-Marie Exbrayat, qui m'a judicieusement orienté vers le concours d'entrée de l'ISFA.
Plusieurs rencontres et discussions m'ont poussé à entreprendre une thèse combinant les problématiques climatiques avec celles de l'actuariat. Je voudrais remercier ici Pierre Arnal, Jean-Jacques Orgeval, Adolphe Nicolas qui m'a reçu dans son laboratoire de Tectonophysique, Philippe Naveau dont les présentations m'ont souvent inspiré, sans oublier Mathieu Ribatet et Pierre Ribereau qui ont toujours répondu à mes questions au début de mes recherches.
Je n'oublie pas non plus Esterina Masiello qui m'a apporté ses corrections dans le domaine des modèles linéaires généralisés.
J'ai beaucoup apprécié l'écoute et les suggestions des professeurs Sidney Resnick et Gennady Samorondnitsky qui m'ont agréablement accueilli lors de mon séjour à Cornell organisé grâce à Stéphane et la participation de l'équipe d'Allianz.

Si j'ai pu mener à bien mes différents travaux et la rédaction de cette thèse, c'est aussi avec l'aide de mes auxiliaires de vie qui ont participé durant de nombreuses après-midi à l'écriture de ces pages et ont aussi rendu possible mes déplacements. Parmi les plus valeureux, je compte Joan aussi actif en cuisine que devant les tableaux Excel et Bruno toujours partant pour m'accompagner vers de nouvelles destinations.

Je tiens à terminer ces remerciements en exprimant ma gratitude, d'abord envers mes grand-parents qui ont été très présents durant toute ma scolarité et après mon accident. Ensuite mes parents, Danièle et Dominique qui n'ont jamais cessé de me soutenir et de me conseiller. Enfin, ma soeur Pauline qui tout en poursuivant ses études d'économie, trouve toujours le temps de m'apporter son aide précieuse.

Table des matières

Introduction générale

Même chez les Bororo, on ne vainc la nature qu'en reconnaissant son empire et en faisant à ses fatalités leur part [52].

Claude Lévi-Strauss

0.1 Introduction

L'idée générale à l'origine de ce travail de recherche est de développer le lien entre le secteur de l'assurance et la part environnementale au sens large de certains domaines de risques. Nous aborderons dans cette thèse deux aspects fondamentalement différents de la tarification. Il est en effet possible de distinguer d'une part les sinistres déterminés par une cause globale et en un sens externes aux biens ou aux individus qui les subissent. Et d'autre part, les sinistres davantage liés aux caractéristiques propres et aux comportements des assurés. Pour évaluer ces dommages de natures différentes, pour pouvoir les tarifer, les approches employées sont nécessairement différentes. Des méthodes spécifiques peuvent être développées pour que le calcul du risque reflète au mieux l'origine de l'événement qu'il décrit. Autrement dit, les sinistres se dissocient selon que leur aléa se génère, se modélise, à partir de variables majoritairement internes ou externes au champ des données dont l'assureur dispose. Ce champ de l'assurance recouvre à la fois la diversité des variables qui peuvent expliquer le risque mais aussi l'historique à notre disposition pour évaluer ce risque de façon robuste. Si la période de retour de l'aléa auquel on s'intéresse dépasse le cadre temporel de nos données, cet aléa sort aussi inéluctablement du champ de l'assurance.

Pour illustrer ces deux points de vue sur la tarification, nous proposons d'étudier dans un premier temps la garantie tempête consacrée aux dommages causés par le vent et par la suite un développement de l'assurance comportementale à travers le risque automobile. Nous avons ainsi d'un côté un aléa associé à un phénomène naturel, météorologique et de l'autre des accidents dans lesquels sont impliqués des conducteurs et leur véhicule. De façon grossière, d'un côté un sinistre global, subi simultanément par un ensemble d'assurés (se trouvant sur la trajectoire de la tempête) et de l'autre des sinistres individuels dont les victimes peuvent être responsables ou non-responsables. En réalité, les situations sont évidemment plus complexes, les dégâts d'une rafale de vent peuvent être amplifiés par la vétusté d'une toiture ou même une fenêtre laissée ouverte. Les facteurs à l'origine d'un accident de la route sont multiples et parfois indépendants du conducteur comme l'état de la chaussée. Il n'en demeure pas moins que la tarification de ces risques demande des approches adaptées en fonction des variables disponibles et de leur significativité par rapport à l'aléa que l'on souhaite mesurer. Les deux premiers chapitres de cette thèse traitent le sujet des tempêtes en France, leur évaluation à l'aide d'un indice météorologique et la sensibilité des résultats aux critères retenus. Le troisième et dernier chapitre s'intéresse aux évolutions techniques et réglementaires de la tarification en assurance automobile.

C'est dans ce contexte que nous nous intéresserons d'abord à la garantie tempête, grêle, neige (TGN) et plus particulièrement aux dépressions météorologiques majeures qui ont un impact très important sur l'équilibre d'un portefeuille d'assurance. Nous introduisons la notion d'événement tempête, nous donnons une estimation de l'ampleur des dommages pour ensuite détailler les besoins d'un assureur en terme de tarification. Il faut ici comprendre à quel point il peut être difficile d'intégrer certains sinistres à la

fois très rares et très coûteux. Pour s'en faire une idée, il suffit de constater que sur les vingt dernières années, seules quatre tempêtes peuvent être considérées en France comme exceptionnelles. Elles écrasent à elles seules tous les autres événements de la même garantie. Il s'agit des cyclones extra-tropicaux Lothar et Martin, qui se sont enchaînés en décembre 1999 coûtant aux assureurs quasiment autant que la totalité de la garantie DaB (Dommage aux Biens). Puis, dix ans plus tard, Klaus en janvier 2009 et Xynthia en février 2010.

Les travaux du premier chapitre proviennent d'un article publié en mai 2015 dans la revue américaine *Risk Analysis* et ayant pour titre *Index for Predicting Insurance Claims from Wind Storms with an Application in France* [57] [1]. L'idée principale de cette étude est d'associer des informations extérieures comme la vitesse du vent aux données de l'assurance. Sur une période de plusieurs dizaines d'années, les variables météorologiques montrent moins de non-stationnarités que les historiques des portefeuilles d'assurance. Nous proposons la construction d'un indice tempête pour compléter et renforcer l'évaluation des dégâts causés par les tempêtes majeures. La gestion du risque lors des événements extrêmes [38] est un sujet qui touche au delà de l'assurance toute la société. Les méthodes de modélisation des extrêmes à partir de données seulement financières comme [28], [42] montrent la forte sensibilité des résultats aux hypothèses retenues. L'utilisation de données météorologiques en complément des données d'assurance est une solution possible pour améliorer la précision et la compréhension des tempêtes. Les études des précipitations extrêmes au Royaume-Uni [30] ou en Californie avec un calcul des périodes de retour selon les paramètres de la loi extrême [41] sont des exemples de modélisation d'un événement extrême à partir de variables climatiques mais sans lien direct avec le coût des dommages. Quelques articles proposent une comparaison entre la sinistralité et des données météorologiques, avec par exemple un modèle utilisant des covariables météo brutes pour prédire le nombre de déclaration par jour en Norvège [76] ou du nombre de victimes des Ouragans et des Typhons aux États-Unis [82]. Pour une évaluation plus précise des dégâts engendrés par l'événement, plusieurs approches sont envisagées comme le calcul d'une fonction d'endommagement pour les tempêtes et tremblement de terre [13] ou la mise en place d'un indicateur d'événements climatiques extrêmes à grande échelle [36] et [19]. D'autres études ont cherché un lien empirique entre les vitesses de vent et la vulnérabilité aux tempêtes pour calculer ensuite une fonction d'endommagement [37]. La méthode qui se rapproche le plus de la nôtre est celle de la construction d'un indice tempête permettant de repérer les événements majeurs en terme de charges d'assurances en se basant sur les vitesses de vent et des critères géographiques et démographiques. Plusieurs articles proposent leur propre indice. En Allemagne Klawa et Ulbrich [45] obtiennent de bons résultats mais sur une base de sinistralité annuelle et avec des données exploitables provenant de seulement 23 stations météo. Dans cet article comme dans ceux de Pinto et al. [64] et de Donat et al. [24], les dommages sont proportionnels à un rapport de vitesse de vent à la puissance 3. Ce choix est motivé par des considérations physiques liant le cube des vitesses de vent avec son énergie. Dorland et al. [25] de même que Prettenthaler

1. Indice pour évaluer les sinistres d'assurance dus aux tempêtes avec une application en France.

et al. [68] en Autriche proposent quant à eux de lier les dommages à une exponentielle des vitesses de vent. En plus des spécificités de sensibilité au vent de chaque région, Prettenhaler et al. prennent aussi en compte les prédispositions d'un bâtiment à être endommagé. Nous proposons avec notre indice de combiner un exposant libre et l'utilisation d'une exponentielle pour obtenir la meilleure correspondance possible entre les valeurs de l'indice et les dommages engendrés par les tempêtes majeures de ces dernières décennies.

La deuxième chapitre est aussi un article faisant suite à notre première publication. Le papier s'intitule *Gestion du risque tempête : sensibilité du calcul de la période de retour et répartition sur le territoire*. Nous utilisons la dépendance des extrêmes des vitesses de vent pour partager la France métropolitaine en six zones tempêtes. Ces zones permettent de mieux appréhender la répartition des sinistres sur le territoire et de calculer notre indice tempête de façon plus homogène. Nous montrons ensuite la sensibilité des calculs aux choix de modélisation avec notamment l'estimation de la Value at Risk (VaR) préconisée par la directive Solvency II. La forte variabilité des résultats renforce, s'il est nécessaire, notre perception des problèmes rencontrés pour dégager des résultats robustes en lien avec les événements extrêmes. Plusieurs difficultés rencontrées vont être décrites dans ce chapitre pour expliquer la variabilité des résultats qui en découlent. Pour commencer, il faut un accès aux relevés des stations météorologiques. En France métropolitaine, les tempêtes peuvent a priori balayer un territoire d'environ 552 000 km^2. La précision du maillage va dépendre du nombre de stations utilisées. Les dégâts observés concernent des bâtiments assurés dont la localisation n'est pas toujours très précise. On note de plus une grande variabilité des vitesses enregistrées selon la position géographique de la station [74]. L'étude des phénomènes de ruptures [78] nous a permis de repérer des décalages, peut-être dus à des changements de matériel, mais qui méritent d'être corrigés. D'autres phénomènes comme les changements de rugosité des sols ou le changement climatique peuvent aussi perturber les données. Étant donné ces difficultés, une solution possible est une classification des départements français selon les vitesses extrêmes de vent. Ce regroupement en 5 à 7 zones de risques permet d'agir à une échelle intermédiaire entre le département (dimension 95) et la France entière. Ce découpage se justifie aussi par le fait qu'il y ait des départements avec très peu de dommages forts (car peu de contrats) où des approches statistiques par département seraient problématiques. Nous utilisons pour cela l'algorithme de classification des `k-médoides`[17] associé à une mesure de dépendance tenant compte de la dépendance extrémale [16]. Le travail sur ces zones nous permet de tester différentes méthodes d'agglomération des vitesses de vent. Il nous permet aussi de considérer différents scénarios de trajectoires de tempêtes [39], [49]. Ces trajectoires pourraient évoluer suite au changement climatique ou balayer plusieurs zones consécutivement. La communication sur l'incertitude qui accompagne la prévision des événements extrêmes est importante et s'avère souvent négligée. Il existe pourtant des moyens techniques pour déterminer cette incertitude mais une réticence sociologique vis à vis de l'aléa limite sa diffusion. Ce fut le cas lors de la tempête Juno de 2015 qui fut en grande partie surestimée comme l'explique Winkler [81].

Le contenu du dernier chapitre porte sur l'étude de garanties dont la tarification devra prendre en compte de façon plus directe le comportement de l'assuré. Dans un premier temps, la coopération avec l'équipe assurance automobile Allianz a permis de répertorier l'ensemble des données disponibles et des critères retenus pour la tarification. La façon dont ces données sont collectées par l'assureur, le possible élargissement à d'autres types de critères et d'autres types de transmission des informations [1] sont à la base du projet. Cette transition est permise par l'afflux d'informations accessibles et regroupées dans ce qu'on appelle la datamasse ou Big Data [31]. Les variables traditionnelles se voient alors remplacées ou enrichies par des données contextuelles et par une circulation accrue des informations entre l'assureur et l'assuré. Les solutions actuelles font intervenir l'utilisation du véhicule. Ces dernières années, l'assurance au kilomètre ou *Pay As You Drive : PAYD* ([2], [4]) s'est largement diffusée dans le système de l'assurance automobile. Le partenariat entre Allianz et l'entreprise SOFCA (Société Française des Compteurs Automobiles) a rendu possible nos modélisations et a débouché sur une réflexion quant à l'évolution du produit pour davantage d'impact. L'article [56] présenté a été accepté et publié en juin 2015 au *Bulletin Français d'Actuariat* (BFA). La connaissance du kilométrage annuellement parcouru constitue une nouvelle variable très significative [48]. Bien que la relation entre la fréquence des sinistres et le kilométrage annuel puisse sembler évidente, peu d'études ont jusqu'à présent insisté sur l'importance de travailler avec des données fiables sur le kilométrage parcouru par les assurés ([3], [50]). Pourtant, la précision de cette information s'avère essentielle à la construction de catégories de risques lors de la tarification. Le facteur principal dans le choix d'une assurance automobile par les consommateurs demeure le prix. Par conséquent, une tarification plus flexible devrait constituer une offre plus intéressante et aurait en plus l'avantage écologique [11] d'inciter à un usage moins systématique des véhicules individuels lorsque cela est possible [9]. Légalement, il faut cependant noter que certaines limites doivent être observées dans le respect des libertés individuelles. Le projet PriPAYD [80] propose par exemple de mettre en place un système de calcul de la prime à l'intérieur de chaque véhicule pour ne transmettre à la compagnie d'assurance que des données agrégées sans fuite d'informations de localisation. Dans ce dernier chapitre, nous observons la sinistralité des différentes catégories de risques selon la partition homme/femme. L'approche graphique permet de mettre en évidence des différences statistiques. Pour faire évoluer leurs tarifications les assureurs peuvent envisager deux approches : mettre en place un proxy ou créer de nouveaux modèles. Partant de ce constat, nous essayons dans une deuxième partie de caractériser le sexe du conducteur en fonction d'autres variables spécifiques à l'assuré, à son environnement ou à son véhicule. Nous utilisons ensuite les modèles additifs généralisés (GAM) et les modèles linéaires généralisés (GLM) pour comparer l'efficacité des modèles avec et sans la variable sexe [54]. Nous démontrons qu'il est possible d'améliorer les résultats des prédictions en faisant abstraction du critère homme/femme. La dernière partie est consacrée aux novices et à l'acquisition d'expérience selon le kilométrage. Nous explorons à travers différentes tranches kilométriques le comportement et l'évolution des jeunes conducteurs durant leurs trois années de noviciat.

Les recherches présentées dans cette thèse s'inscrivent dans la perception et l'estimation du risque par les assureurs. Au delà d'une optimisation des pratiques courantes, les changements attendus dans le secteur de l'assurance demandent un véritable travail sur la conception et la tarification des produits en fonction de la nature de l'aléa, des nouvelles technologies disponibles et de l'évolution vers un marché de services globalisé. La quantité d'informations désormais accessibles et traitées entraîne de nouvelles réglementations, une redéfinition des variables discriminantes et des limites légales à leurs utilisations. Si certaines avancées sont techniques, d'autres se fondent sur la maîtrise et la qualité des données à la base de toutes interprétations.

0.2 La notion d'événement tempête

Une tempête majeure est un phénomène qui se produit à une échelle synoptique[2]. La plupart des données relatives aux sinistres sont disponibles avec une précision journalière, cependant le passage d'une tempête sur un territoire de plus de 550 000 km^2 comme la France métropolitaine peut durer plusieurs jours. La déclaration de sinistre peut aussi être une source d'imprécisions lors de la modélisation des événements les plus importants. En effet, différentes circonstances, comme la possession d'une résidence secondaire ou un assuré absent de son domicile pour des vacances, peuvent différer la date de cette déclaration. Lorsqu'on considère les résultats d'un portefeuille d'assurance, on observe généralement un étalement de charges très élevées sur les quelques journées qui précèdent et qui suivent directement la survenance d'une tempête. Ce décalage peut fausser les résultats d'une fonction de distribution et peut aussi poser problème si on veut utiliser des covariables telles que la vitesse du vent ou la température qui sont mesurées quotidiennement. Même en interne, les clauses de réassurance tempête s'étendent sur une période comprise entre 48 et 72 heures. L'échelle "jour" n'est donc pas la mieux adaptée à l'étude des événements les plus importants. Il est alors préférable pour la suite de notre étude de regrouper les charges journalières dépassant un certain seuil en "blocs" d'une durée déterminée (deux jours, une semaine, un mois) pour travailler à une échelle adaptée aux tempêtes. Nous verrons dans les chapitres 1 et 2 à quel point ces choix peuvent influencer les résultats.

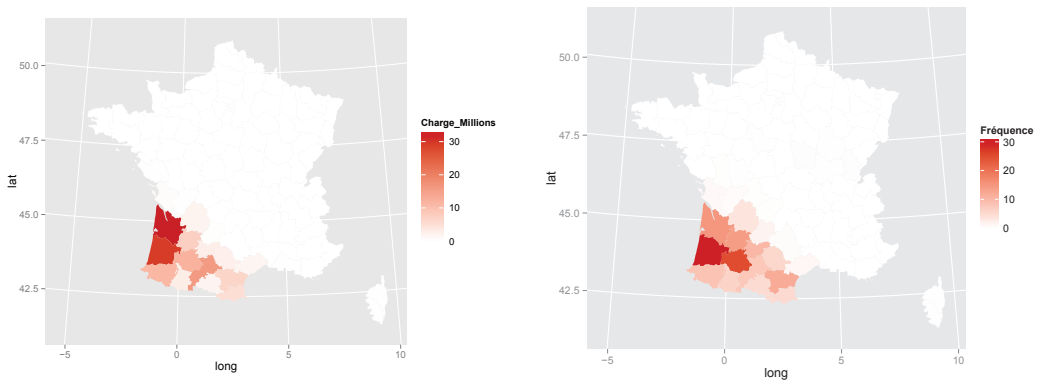


FIGURE 1 – Charges (en millions d'euros) et Fréquence (en %) durant la tempête Klaus

Il est possible d'observer à l'échelle nationale les dégâts causés par une tempête en associant pour chaque département les dommages enregistrés sur une période de deux ou trois jours autour de la date exacte de son passage. Sur la Figure 1 nous avons

2. ce terme est utilisé en météorologie pour des systèmes atmosphériques qui s'étalent à la fois dans l'espace et dans le temps [62].

représenté à gauche la somme des charges entre le 23 et 25 janvier 2009, dates du passage de la tempête Klaus dans le Sud-Ouest de la France. Cette approche graphique permet de délimiter relativement clairement la zone touchée par la tempête, cependant les charges dépendent du nombre de contrats ce qui peut fausser les amplitudes entre des départements inégalement peuplés ou inégalement représentés dans le portefeuille Allianz. La Gironde par exemple est la plus touchée sur cette carte mais elle compte avec Bordeaux une forte densité d'habitations et donc de risques potentiels. Nous verrons dans la suite de cette étude qu'un modèle tenant compte de la localisation des sinistres et de leurs fréquences peut apporter des solutions intéressantes.

La partie droite de la figure présente aussi les dégâts liés à la tempête Klaus mais cette fois-ci en termes de fréquences. L'avantage de cette approche est qu'elle ne dépend plus du nombre de contrats dans le portefeuille Allianz. Elle permet donc de mieux observer l'intensité relative du passage de la tempête dans chacun des départements touchés. Contrairement à la carte précédente basée sur les charges, la Gironde n'est plus la zone la plus fortement sinistrée avec 15% des contrats touchés soit presque moitié moins que les Landes où 31% des assurés ont déclaré un sinistre entre le 23 et 25 janvier. L'approche fréquence correspond davantage au passage de la tempête avec une diminution plus progressive des intensités. Il sera donc utile d'utiliser à la fois les charges, les nombres de sinistres et la répartition géographique du portefeuille dans la recherche d'une corrélation avec des variables météorologiques comme la vitesse du vent ou la température.

0.3 Les dommages causés par le vent

Pour les assureurs, les dommages causés par le vent sont pris en charge par la garantie TGN qui recouvre à la fois les tempêtes, la grêle et la neige. Une étude publiée par l'AFA (Association Française de l'Assurance) [26] en février 2012 (Tab I) permet d'observer à l'échelle nationale et pour des historiques remontant à 1982 la volatilité des coûts annuels (entre 70 millions et 7 milliards). On dispose depuis 1984 de la distinction entre tempête et autres natures d'événements. Sur une base en euros courants on constate la forte sensibilité des coûts aux tempêtes qui comptent pour 87.5 % de la garantie. L'année 1999 représente à elle seule plus du tiers de la charge totale.

Année	Tempête	Grêle/Neige	TGN		Année	Tempête	Grêle/Neige	TGN
1982			488		1998	220	30	250
1983			458		1999	6860	-	6860
1984	200	95	295		2000	240	60	300
1985	35	35	70		2001	135	75	210
1986	130	80	210		2002	160	40	200
1987	530	55	585		2003	420	150	570
1988	160	20	180		2004	400	90	490
1989	180	40	220		2005	300	120	420
1990	1315	180	1495		2006	360	215	575
1991	65	10	75		2007	380	135	515
1992	190	120	310		2008	360	90	450
1993	180	60	240		2009	1980	260	2240
1994	305	95	400		2010	855	255	1110
1995	170	10	180					
1996	200	10	210		Total	16450	2355	18805
1997	120	25	145		Part	87.5%	12.5%	100.0%

TABLE I – Charges non actualisées (en million d'euros)

L'actualisation des charges pour les rendre comparables à celles de 2012 constitue une part importante dans la solidité des résultats de notre étude. Bien que certaines approximations restent inévitables, nous essayons d'être le plus exhaustif possible en considérant :

- l'évolution des indices FFB pour les particuliers et RI pour les entreprises [3] ;
- le poids relatif des segments particulier/entreprise dans le portefeuille ;
- le taux de diffusion de la garantie tempête qui n'est obligatoire que depuis juillet 1990 ;
- la progression du parc immobilier en France.

Les charges actualisées (Tab II) permettent de relativiser l'importance des événements de 2009 et 2010. L'année 1999 montre à la fois la volatilité et l'intensité des dispersions avec une valeur actualisée dépassant les 12 milliards soit l'équivalent d'une année de

3. FFB (Fédération Française du Bâtiment, 3.93% / an) et RI (Risques Industriels 3.38% / an) sont deux indices du coût de la construction en France

Année	Tempête	TGN		Année	Tempête	TGN
1982		3745		1997	228	276
1983		2719		1998	408	464
1984	988	1457		1999	12394	12394
1985	148	296		2000	423	529
1986	493	797		2001	225	350
1987	1876	2070		2002	257	321
1988	499	561		2003	648	879
1989	494	603		2004	590	723
1990	3227	3668		2005	416	583
1991	155	179		2006	481	768
1992	434	707		2007	471	639
1993	400	534		2008	425	531
1994	651	854		2009	2218	2509
1995	353	374		2010	950	1234
1996	389	408				

TABLE II – Charges actualisées/2012 (en million d'euros)

prime DAB (Dommages Aux Biens). Les charges supportées en 1990 et 1999 sont le fruit de plusieurs événements majeurs dont notamment en décembre 1999 la succession des tempêtes Lothar et Martin. On perçoit le caractère exceptionnel de cette année qui représente à elle seule 30% de la charge totale. La période de retour d'événements d'une ampleur similaire voire supérieure aura une influence considérable sur les résultats. Par exemple selon que l'on étale l'impact de Lothar et Martin sur 50 ou 100 ans la charge moyenne annuelle TGN varie entre 1.3 et 1.1 milliards d'euros si on considère la période 1982-2010. Si on se contente de commencer en 1984 où la distinction tempête apparaît, la même charge moyenne annuelle baisse d'environ 15% pour se situer entre 1.1 et 0.9 milliards. Le niveau des primes relatives aux TGN (Tab III) était quant à lui de 1.3 milliards d'euros en 2010 soit 8.9% de l'ensemble des cotisations des assurances de DAB. Cependant il s'agit de primes commerciales qui intègrent des chargements comme les frais de gestions et les commissions évalués à environ 35%. On voit ici que ces primes sont sous évaluées au niveau du marché et qu'elles peuvent être largement dépassées lors d'événements majeurs.

Année	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
En millions d'euros	950	1000	1060	1100	1100	1115	1200	1210	1250	1310

TABLE III – Historique des cotisations du marché national depuis 10 ans (estimation)

La modélisation d'un phénomène d'une telle volatilité nécessiterait des informations précises et détaillées sur la période d'observation la plus longue possible, sur un portefeuille dont l'évolution est connue et pour une couverture géographique du territoire homogène. Nous ne disposons pas d'une telle base de données qu'il serait d'ailleurs difficile d'obtenir compte tenu des purges informatiques et des actualisations de garanties que connaissent les différentes sociétés d'assurance. Néanmoins des données ont été mises à notre disposition par Allianz et nous les avons utilisées pour expliquer

l'évolution des dommages liés au vent au cours de ces dernières années.

0.4 Les données et leurs utilisations

Pour une étude approfondie de l'évolution et l'impact du coût des tempêtes, disposer de 2000 ans de relevés de sinistres actualisés constituerait une base de données solides. Cela permettrait de disposer de toutes les informations utiles pour maîtriser les résultats relatifs aux problèmes de tarification et de distribution des résultats dans le temps à une échelle nationale ou locale. Force est de constater que nous sommes très loin de disposer d'une telle information et que cela rend les analyses plus délicates. Nous sommes parvenus à récupérer des données détaillées sur une quinzaine d'années. Sur les évènements les plus importants, il est possible de construire des historiques sur une quarantaine d'années [51]. Dans les deux cas, nous sommes loin d'atteindre le millier d'années. De plus, même si nous disposons de techniques pour actualiser les informations, ces techniques ne sont pas parfaites.

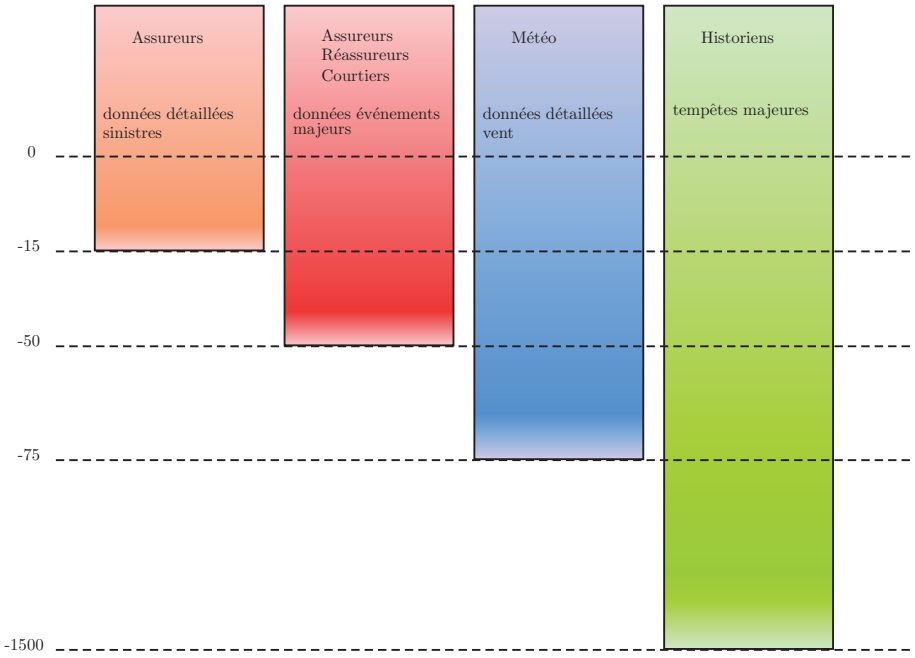


FIGURE 2 – Les données classées par strates historiques

Dans ce contexte, tout ce qui permet de compléter l'information insuffisante est utile. Cependant il faut toujours prendre garde à ne pas faire un transfert d'objectif. Construire et utiliser un indice tempête constitue une opération délicate. Tous les développements effectués dans cette thèse montrent ces difficultés. Ajoutons à cela le fait que l'on détermine une fonction d'ajustement basée sur des critères météorologiques eux-mêmes restreints. Nous ne disposons au mieux que de 200 stations pour l'ensemble du territoire métropolitain et les enregistrements varient également dans le temps (cf

chapitre 2.3). Si ces données permettent de disposer d'informations locales (1 à 3 stations par département) sur une quarantaine d'années, elles sont toujours loin de répondre complètement aux questions concernant la tarification et a fortiori la distribution des sinistres. Ce défaut se ressent particulièrement sur la queue relative aux évènements les plus extrêmes. Pour remonter aussi loin dans le temps, on ne peut imaginer qu'utiliser des données relevées par des historiens qui seront de nature très différentes et ne porteront que sur des évènements majeurs. La Figure 2 confronte les historiques disponibles selon la nature et l'origine des données tempête. En complément à ces types de données disponibles ou pas, on peut s'interroger sur les besoins rencontrés par les assureurs ou réassureurs en termes de tarification. Il sera utile de savoir en quoi ces diverses informations peuvent être utiles pour résoudre les problèmes rencontrés par les compagnies face aux événements naturels.

0.5 Les besoins d'un assureur en termes de tarification

Du point de vue de l'assurance, une fonction de distribution des montants des sinistres ne suffit pas pour répondre aux besoins en terme de tarification. Il faut également intégrer la dimension fréquence. Dans ces conditions, en quoi une information météorologique peut-elle être plus pertinente que celle des observations des assureurs ? Si l'on utilise des variables externes, comment valider les choix sur les hypothèses retenues ? Pour cela, on peut distinguer plusieurs niveaux d'évènements. Pour simplifier et sans donner de limites précises, nous avons :

- un très grands nombre de journées sans phénomène significatif ;
- un nombre encore significatif de journées supportant des évènements remarquables, mais d'un coût encore très modeste ;
- un petit nombre d'évènements présentant des charges significatives ;
- un nombre exceptionnel d'évènements présentant les charges les plus volumineuses.

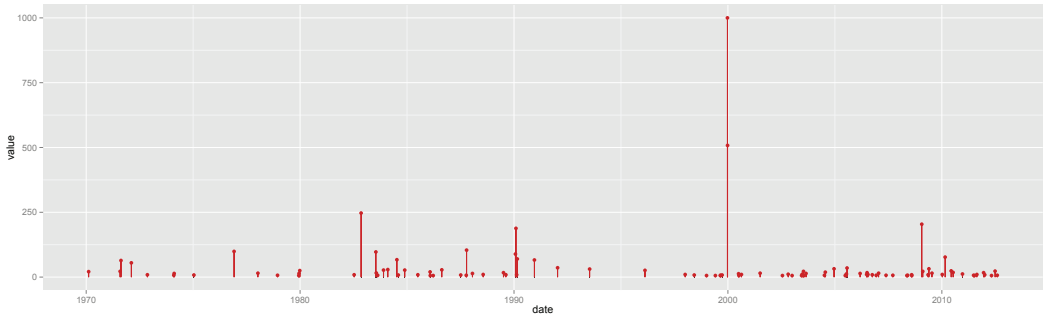


FIGURE 3 – Charges actualisées des tempêtes en France depuis 1970

Sur la Figure 3 on observe les charges actualisées des principales tempêtes depuis 1970 d'après les données du portefeuilles Allianz fournies par Michel Luzi [51].

FIGURE 4 – Charges actualisées des tempêtes en France depuis 1998

À une autre échelle, depuis 1998, on peut voir les mêmes résultats mais sur l'ensemble des charges journalières sur la Figure 4. Sur le graphique du haut on constate que les 4 principales tempêtes que sont Lothar et Martin en 1999 suivies de Klaus et Xynthia en 2009 et 2010 écrasent tous les autres événements. Sur celui du bas, hors évènements extrêmes, on constate que seulement quelques évènements restent un peu marquants à l'échelle nationale. En dessous d'un certain seuil, nous avons toutefois suffisamment d'observations pour considérer que la période observée ne biaise pas trop la recherche de lois. Plus clairement, pour les 2 premières classes (non représentées sur la Figure 3), les assureurs n'ont pas besoin d'utiliser un artifice à la place de leurs données. Leurs historiques, de 15 à 20 ans, sont largement suffisants pour répondre à leurs questions. Pour ce qui est des évènements extrêmes, il en va tout autrement. En fait pour cette partie, ils transfèrent le problème aux réassureurs, qui déterminent l'exposition et la traduisent en besoin de couverture et en prime de réassurance. Ensuite, la compagnie d'assurance répercute la prime dans ses tarifs. Reste le 3ème niveau qui peut poser problème, car non représentatif au niveau du portefeuille et pas complètement réassuré. Remarquons également, que le fait de se réassurer ne devrait pas dédouaner la compagnie d'assurance d'analyser son exposition sur les niveaux extrêmes. Sinon comment apprécier le coût du service rendu par le réassureur ? Pour revenir à la

question initiale, on peut constater que s'il peut être utile d'utiliser des informations complémentaires pour les évènements majeurs, il n'y a aucun intérêt d'utiliser des artifices sur la grande majorité d'évènements sans grande importance. C'est pourtant ce que font les modèles commerciaux proposés sur le marché. Puisqu'ils présentent un volumineux catalogue d'évènements (plusieurs milliers ou dizaine de milliers) qui donne l'impression de sérieux, mais porte essentiellement que sur des évènements de très faible importance, donc sans intérêt pour le problème posé. Il est donc possible de mesurer la charge (coût x fréquence) selon divers critères, géographiques (département, zone tempête), nature des expositions (particulier [individuel, collectif], professionnel, agricole, entreprise, etc), nature des activités. Ici encore, en quoi un indicateur artificiel peut-il être plus efficace que les données observées ? Comme on le constate, la combinatoire de nombreux paramètres (souvent inconnus) peut poser de sérieux problèmes de rapprochement. Il faut donc en priorité limiter le périmètre des prétentions aux sujets pouvant trouver une solution efficace. La vraie difficulté porte sur la meilleure maîtrise des fonctions (coûts, fréquences) des évènements extrêmes. C'est sur la partie des évènements extrêmes que les données externes peuvent apporter certains éléments de réponse. Nous devons toutefois remarquer que si le marché (assureurs, réassureurs, courtiers de réassurance) voulait coopérer, il serait possible de reconstituer les historiques des principaux évènements depuis le début des années 70. On arrive légèrement plus loin avec les données météorologiques (et encore, avec un questionnement sur les appareils de mesures et le nombre de stations disponibles). En définitive, il faudrait extrapoler avec des périodes de 100, 200, 500 ans, voire 2000 ans pour fiabiliser un seuil de 99,5%.

0.6 Points de vues selon l'intensité des tempêtes

Lorsqu'on étudie la répartition des charges selon l'intensité des tempêtes, on dispose de différents points de vue pour mieux comprendre leur impact. On cherche à savoir si cette répartition est homogène ou si au contraire la charge moyenne augmente lors des événements les plus importants en termes de charges globales et de fréquences pour l'assureur. On travaille ici à partir du portefeuille d'Allianz pour les particuliers en France entre 1998 et 2012. Ceci représente plus de 300 000 déclarations de sinistres. On commence par représenter la distribution des charges pour :

- les 2 principales tempêtes (Lothar, Martin)(T2) ;
- les 2 suivantes (Klaus, Xynthia) (T4) ;
- les 10 suivantes (T10) ;
- les 100 suivantes (T100).

Dans le tableau V on a regroupé les quantiles, la moyenne et le maximum des charges individuelles enregistrées par Allianz lors des tempêtes classées par intensité. À première vue les charges des tempêtes ayant causé le plus de dégâts à l'échelle globale sont plus élevées que les autres. La charge individuelle moyenne sur les deux principaux événements de la période 1998-2012 (Lothar et Martin) est de 3 560 euros alors qu'elle descend à 2 350 pour les deux suivants (Klaus et Xynthia). Elle descend à nouveau à 1 900 euros pour les dix tempêtes suivantes puis remonte à 2 400 euros pour les cent

Quantiles	Charges individuelles actualisées			
	T2	T4	T10	T100
Min.	0	0	0	0
Qu 0.25	828	533	470	475
Qu 0.5	1 977	1 240	1 030	1 135
Moy.	3 563	2 351	1 894	2 433
Qu 0.75	4 141	2 237	1 940	2 408
Qu 0.9	6 766	4 866	3 818	5 127
Qu 0.99	27 205	20 366	15 225	21 217
Qu 0.999	81 414	61 576	52 809	69 556
Max.	1 440 000	530 600	140 900	467 000

TABLE IV – Répartition empirique des charges actualisées individuelles

dernières. On obtient le même classement pour les différents quantiles et la plus grosse différence est obtenue pour le maximum avec une charge dépassant 1.4 millions d'euros après le passage de Lothar. Il semble donc que la charge individuelle augmente lors des tempêtes les plus intenses et que cette augmentation touche toutes les tranches de sinistres. À partir des données dont nous disposons il est impossible de fournir une explication directe à ces variations. On peut cependant apporter des hypothèses explicatives comme :

1. À une échelle globale,
   - Le prix des réparations peut augmenter lors d'une forte demande. Si les couvreurs montent les prix après le passage d'une tempête majeure la charge de l'assureur augmente alors que les dégâts subis par le particulier sont peut-être équivalents.
   - Les réparations faites suite à une tempête majeure vont a priori renforcer le bâti et diminuer les dommages de la tempête suivante.
   - Les choix d'actualisations étant donné que les deux principaux événements ont eu lieu en début de période (99).
   - La localisation et la topographie de la zone sinistrée [23].
   - L'ampleur de la zone sinistrée.
2. À une échelle individuelle,
   - Une chute d'arbre risque d'entraîner de lourds dégâts si une habitation est touchée.
   - La qualité du bâti (neuf, ancien, bien ou mal entretenu).
   - La vulnérabilité du bâtiment dont les portes ou les fenêtres peuvent rester ouvertes et amplifier les dégâts.

Il existe donc de nombreux facteurs, dépendants à la fois du marché, des individus, du climat ou de la géodynamique, susceptibles de faire varier la charge. Il faut cependant bien distinguer les facteurs globaux entraînant des variations propres à un événement et les facteurs individuels qui peuvent toucher indifféremment l'ensemble des événements. Graphiquement on peut voir en comparant les deux premiers histogrammes (Figures 5 et 6) que l'allure de la fonction de répartition des tempêtes les plus intenses n'est

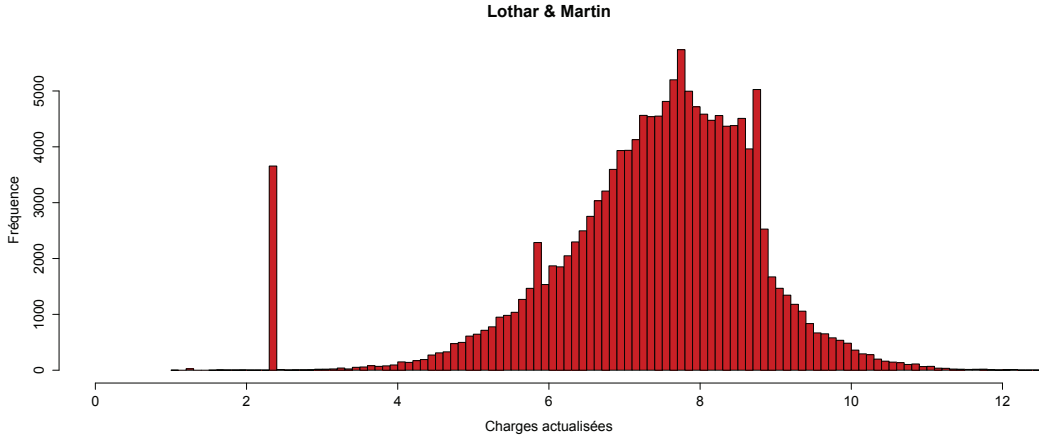


FIGURE 5 – Répartition empirique des charges actualisées individuelles pour Lothar et Martin (échelle logarithmique)

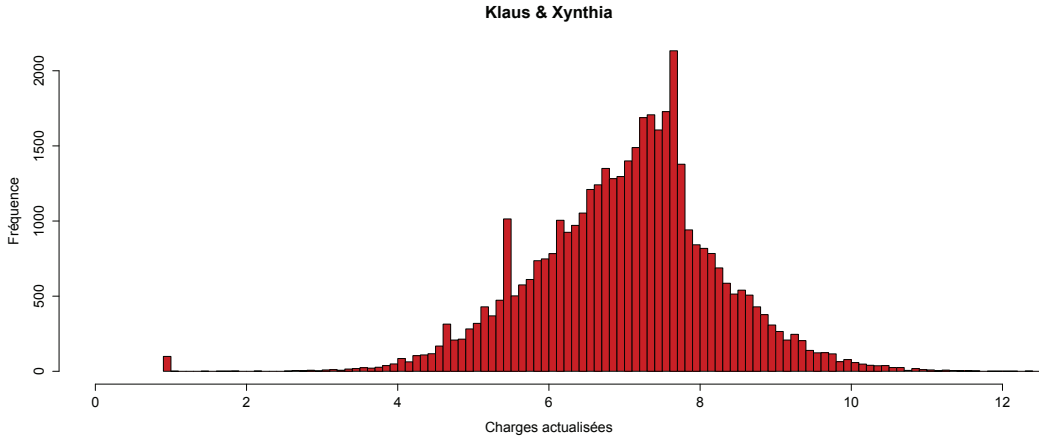


FIGURE 6 – Répartition empirique des charges actualisées individuelles pour Klaus et Xynthia (échelle logarithmique)

pas symétrique, contrairement aux deux suivants (Figures 9 et 8) qui prennent plus la forme d'une loi normale. On constate aussi un pic de fréquence entre les valeurs 7.5 et 8 (soit environ entre 2 000 et 3 000 euros).

Sur une échelle annuelle, les charges moyennes sont quasiment toutes comprises entre 2 000 et 2 500 euros, mis à part en 1999 où elle culmine à plus de 3 500 euros. On peut ainsi nuancer l'hypothèse d'une différence due à l'actualisation en comparant la valeur de 99 avec celles de 98 et 2000 qui sont inférieures à 2 500 euros. En terme de dispersion l'écart-type tourne autour de 5 000-6 000 euros et augmente en 1999, 2001 et 2008 pour dépasser les 8 000 euros. Cependant en 2001 et 2008 le nombre total de sinistres enregistrés est relativement faible (inférieur à 5 500) et par conséquent il suffit

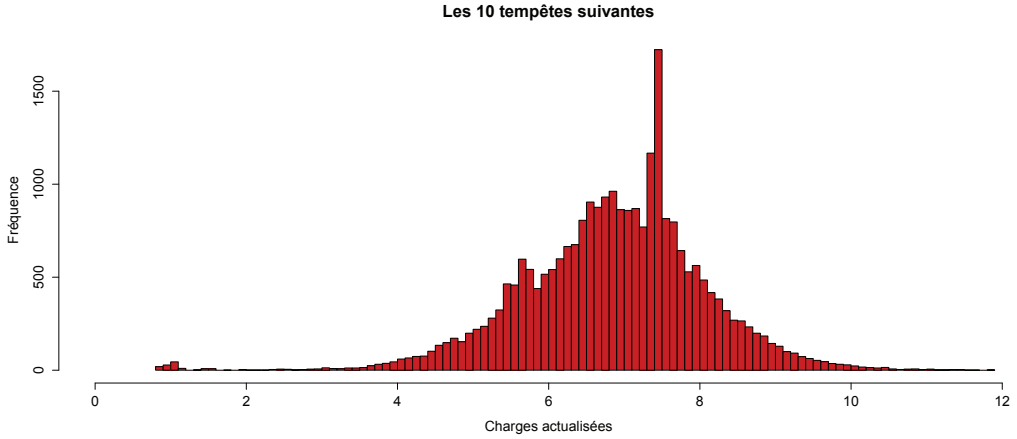


FIGURE 7 – Répartition empirique des charges actualisées individuelles pour les 10 tempêtes suivantes (échelle logarithmique)

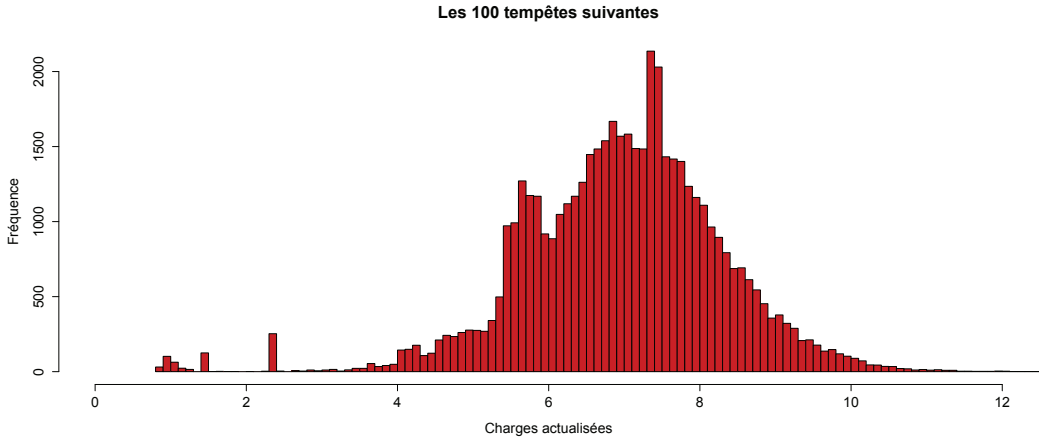


FIGURE 8 – Répartition empirique des charges actualisées individuelles pour les 100 tempêtes suivantes (échelle logarithmique)

d'un coût élevé pour augmenter la variance.

En observant les quantiles on retrouve à tous les niveaux le surcoût de 1999. La médiane notamment nous indique que la moitié des charges ont été supérieures à 2 000 euros cette année là, soit près du double des autres.

Année	Moy.	Médiane	Qu.75	Max.	SD	nb. (milliers)
1998	2 469	1 115	2 450	166 700	5 917	2.2
1999	3 541	1 949	4 101	1 440 000	9 211	150.0
2000	2 427	1 207	2 533	200 400	5 196	4.4
2001	2 485	1 158	2 533	295 600	8 041	2.6
2002	1 928	988	2 073	85 690	3 768	4.2
2003	2 599	1 215	2 729	128 800	4 780	7.7
2004	2 217	1 053	2 254	140 900	4 435	8.1
2005	2 548	1 097	2 574	181 800	6 343	2.3
2006	2 333	1 048	2 369	133 700	4 824	6.1
2007	1 938	850	1 910	319 400	6 176	6.5
2008	2 477	995	2 374	467 000	8 623	5.5
2009	2 377	1 263	2 241	530 600	5 606	36.5
2010	2 062	1 002	2 009	239 400	5 338	15.6
2011	1 925	1 559	1 708	325 500	5 184	6.6

TABLE V – Répartition et écart type des charges actualisées individuelles

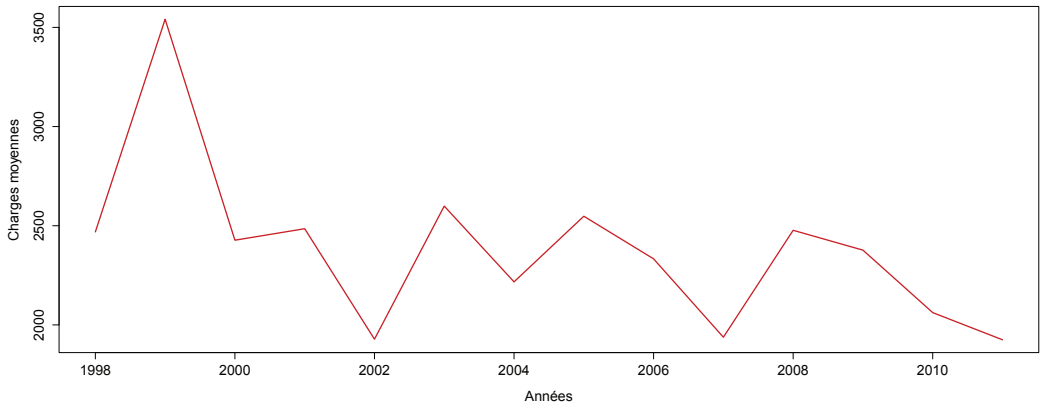


FIGURE 9 – Moyennes des charges actualisées individuelles entre 1998 et 2011

0.7 Autour de la vitesse du vent

0.7.1 Définition

Le vent est un mouvement au sein d'une atmosphère. Mécaniquement, il est décrit par la mécanique des fluides. Les molécules qui composent l'air ne sont pas solidaires les unes des autres, ce qui complique la prévision de ces trajectoires mais va permettre au flux de s'adapter aux configurations topographiques qui ne manqueront pas d'influencer son écoulement [12].

En tout point fixe de l'atmosphère où il doit être mesuré, le vent varie généralement au cours du temps de façon brusque, rapide et parfois importante. Sa mesure isolée à un instant donné, ou vent instantané, ne serait guère représentative de la valeur d'ensemble de ce mouvement horizontal au site où on l'observe, ni des écarts à cette valeur.

De plus, ces particularités du comportement dynamique de l'air se remarquent aussi bien pour la mesure de la vitesse du vent que pour celle de sa direction.
La valeur de référence de la vitesse et de la direction en un point et à un instant de mesure donnés sont donc la vitesse moyenne et la direction moyenne du vent en ce point et à cet instant, c'est-à-dire les moyennes des vitesses et directions instantanées sur l'intervalle assez long (généralement de 10 minutes) qui précède l'instant de mesure. Le vent ayant ces vitesses et directions moyennes est le vent moyen. Le vent évalué en météorologie est, sauf mention explicite du contraire, ce même vent moyen. Les bulletins météorologiques y font toujours référence, tout en mentionnant éventuellement la mesure observée ou l'intensité prévue des rafales (dont les vitesses peuvent parfois dépasser de plus de moitié celles du vent moyen) (Source : Météo France). Il est aussi possible de mesurer la composante verticale du vent avec un anémomètre tridimensionnel. On peut ainsi étudier les mouvements subsidents et ascendants de l'air.

Le vent trouve son origine dans la différence de pression atmosphérique. On identifie deux types de vents. Le premier est appelé vent synoptique, il est engendré par des phénomènes d'échelle continentale ou planétaire. Il est durable, étendu et parfois violent. Le deuxième se situe à l'échelle locale. Il est dû aux inégalités de températures, ce sont les brises thermiques. Pour comprendre les dégâts occasionnés par le vent, il est important de distinguer ces deux catégories. En effet, le comportement et l'évolution du flux de vent seront différents selon son origine. Le vent synoptique sera généralement freiné par un obstacle (naturel ou artificiel) alors que les brises thermiques se verront renforcées par le relief. Ces deux cas sont illustrés respectivement à gauche et à droite de la Figure 2.18.

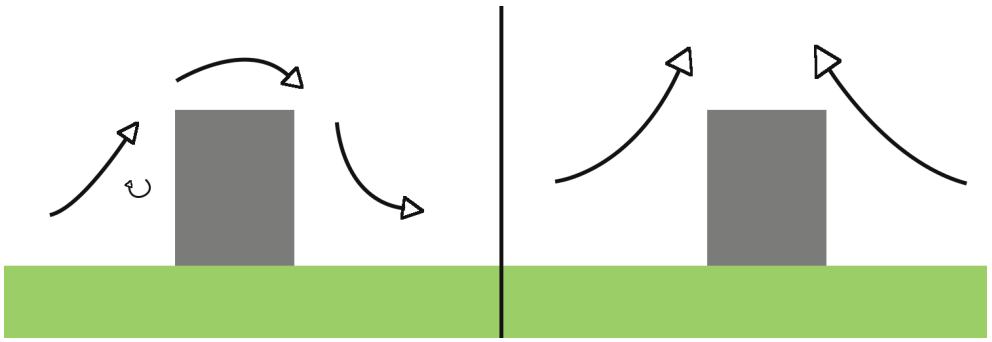


FIGURE 10 – Deux types de comportements du vent (synoptique à gauche, thermique à droite)

Notons ici que nos sources d'information ne sont pas aussi exhaustives. Nous avons donc essayer de faire pour le mieux avec les données à notre disposition.

0.7.2 Les données Météo France

Météo France propose 7 types de relevés associés au vent :
– Vitesse vent moyen maximale ;

– Moyenne des vitesses du vent (à 10m) ;
– Direction du vent instantané maximal (à 10m) ;
– Vitesse du vent instantané maximal ;
– Vitesse du vent instantané sur 3 secondes maximale ;
– Vitesse du vent instantané maximal à 2 mètres ;
– Moyenne des vitesses du vent à 2 mètres ;

Par exemple, la vitesse vent moyen maximale est définie de la façon suivante. Il s'agit de la vitesse maximale du vent moyenné sur 10 minutes, relevée entre 00 UTC le jour J et 00 UTC le lendemain (J+1). Hauteur de la mesure : 10 mètres. Unité : mètres par seconde et dixièmes (1 ms-1 = 3,6 km/h = 1,945 noeuds).

Les stations météorologiques professionnelles du réseau de Météo-France, appelé réseau Radome, sont au nombre de 554 en métropole (une tous les 30 km) et 67 en Outre-Mer. Ces stations mesurent de façon automatique les paramètres de base que sont la température et l'humidité sous abri, les précipitations et le vent (vitesse et direction) à une hauteur de 10 mètres. Certaines stations mesurent des paramètres complémentaires comme la pression, l'humidité dans le sol, le rayonnement, la visibilité, l'état du sol, etc. En effet, la densité nécessaire des mesures diffère selon les paramètres. Ainsi, la mesure de la pression nécessite un maillage plus lâche que celle de la température mais elle est systématiquement mesurée sur les aéroports où sa connaissance est indispensable. D'autres paramètres répondent à des besoins plus spécifiques et sont concentrés sur des zones sensibles : aérodromes, zones de risque d'inondations, d'incendies de forêt, d'avalanche. Outre les mesures traditionnelles effectuées par les baromètres (pression), hygromètres (humidité de l'air), anémomètres (vent) et pluviomètres (pluie), les progrès technologiques permettent aujourd'hui de mesurer de façon automatique un nombre de plus en plus grand de paramètres comme l'humidité du sol, l'état du sol, le rayonnement, le temps présent, la visibilité, la hauteur des nuages, etc.

Ces données sont très complètes mais présentent le désavantage d'être payantes. La référence tarifaire pour les données climatiques de base est de 40 centimes d'euros pour 10 relevés. A titre d'exemple, notre indice calculé entre 1973 et 2013 sur 170 stations nécessite plus de 2.5 millions de relevés soit un coût de 100 000 euros environ. Pour un accès illimité et annuel aux données climatiques de base, la redevance est de 200 000 euros. Nous avons obtenu en 2014 un crédit de recherches de la part de Météo France qui nous a permis de mener à bien les travaux présentés dans le chapitre 2.

0.7.3 Les données de la NCDC

Le site américain de la National Climatic Data Center (NCDC) propose seulement trois types d'informations sur le vent mais il a l'avantage d'être en libre accès.
– *Mean wind speed* ;
– *Maximum sustained wind speed* ;
– *Maximum wind gust* ;

Sur ce site, nous avons pu collecter les relevés de 173 stations météorologiques en France entre 1973 et 2013.

0.7.4 Problème de rupture des données

Parmi les autres problématiques abordées, il a aussi été question de la fiabilité des données météorologiques. En effet, lorsqu'on s'intéresse à une longue période dans le temps, les appareils de mesures et les méthodes de restitutions des données peuvent changer et introduire un biais dans les résultats [78]. Ces impacts peuvent être considérables, spécifiquement sur l'écart-type de la pente qui est inversement proportionnel à $T^{3/2}$ où T est la longueur de la période d'observation. Il est donc intéressant de savoir détecter les phénomènes de rupture et de pouvoir les corriger. A travers l'observation des moyennes/médianes/variances annuelles, pour chaque station, on essaie de détecter un changement dans ces séries.

0.7.5 Détection de ruptures dans nos relevés

On utilise le package `changepoint` du logiciel `R` pour détecter les éventuelles ruptures parmi les données de vitesses de vent que nous utilisons. Les différentes méthodes d'optimisation sont décrites dans l'article de Killick, R. et al. [44]. Formellement, pour une suite de données $(y_1, ..., y_n)$, on dit qu'il existe une rupture $\tau \in \{1, ..., n-1\}$ si les propriétés statistiques des sous-ensembles $\{y_1, ..., y_\tau\}$ et $\{y_{\tau+1}, ..., y_n\}$ sont différentes selon un certain critère. La détection multiple de ruptures se fait ensuite dans la littérature en minimisant :

$$\sum_{i=1}^{m+1} \left[\mathcal{C}\left(y(\tau_{i-1}+1) : \tau_i\right)\right] + \beta f(m) \tag{1}$$

avec $\mathcal{C}$ une fonction de coût pour chaque segment (par exemple la log-vraisemblance négative) et $\beta f(m)$ une pénalité pour éviter la sur-segmentation des données. Dans cette thèse, nous utilisons l'algorithme PELT (Pruned Exact Linear Time). PELT effectue d'abord un test de rupture sur l'ensemble de la période. Si une rupture est détectée alors les données sont divisées en deux à la position de la rupture. La détection est ensuite effectuée à nouveau sur les deux jeux de données ainsi créés. Cette procédure continue jusqu'à ce qu'aucune rupture ne soit plus trouvée. Pour éviter les problèmes de saisonalité, on se concentre ici sur les moyennes annuelles de chaque station. De plus, comme notre indice est construit à partir des vitesses les plus élevées, la détection se fait ici sur les dépassements au-delà d'un seuil fixé au quantile q à 80%.

Nous allons comparer les vitesses de vent maximales journalières obtenues selon deux sources différentes. Avant toute chose, il faut définir le plus précisément possible la façon dont ces mesures sont enregistrées. Rappelons que Météo-France utilise la vitesse du vent instantané maximal exprimée en mètres par seconde et dixièmes. Sur le site américain de la National Climatic Data Center (NCDC), on a moins de précisions, le paramètre est *maximum sustained wind speed* et l'unité de mesure est le noeud, une

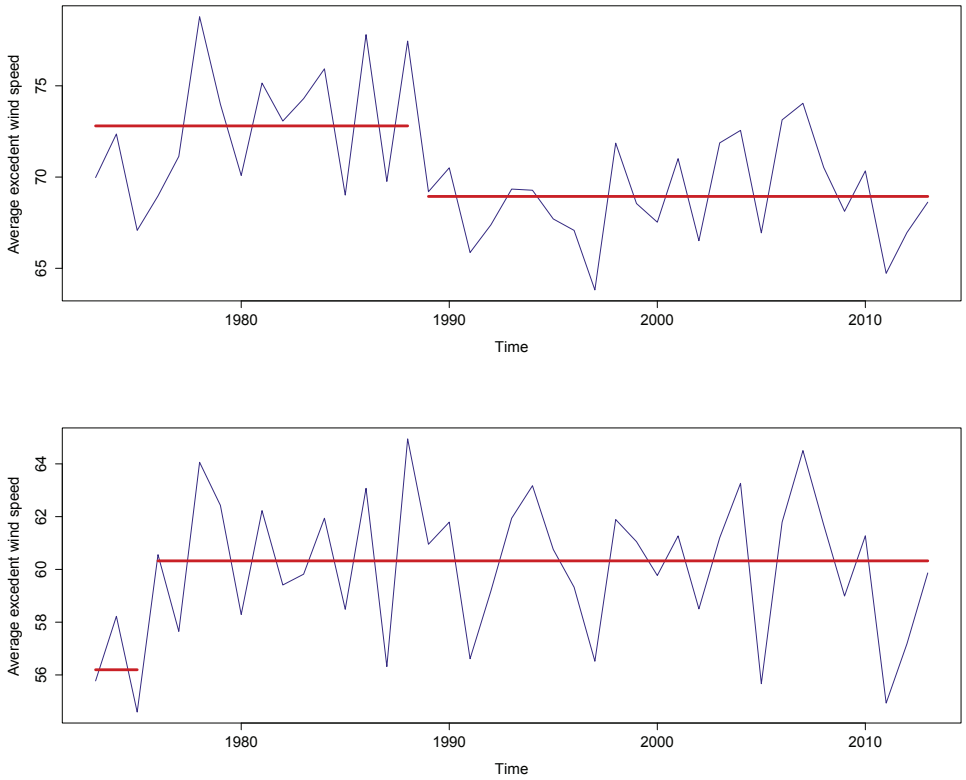


FIGURE 11 – Détection de ruptures sur les dépassements moyens de vitesses de vent à Bordeaux (Haut : NCDC - Bas : Météo France)

unité de vitesse utilisée en navigation maritime et aérienne [4]. Cependant, les faibles vitesses de vent obtenues après conversion en km/h nous ont conduit à penser que l'unité de mesure serait plutôt en m/s. Sur la Figure 11 on peut observer en haut les résultats de la détection de ruptures sur une moyenne annuelle des dépassements de vitesse de vent de la station Bordeaux-Mérignac. La période est découpée en deux séquences : de 1973 à 1988 puis de 1988 à 2013. Pour cette station, les vitesses enregistrées durant la première séquence sont surévaluées par rapport à celles de la deuxième séquence. Il serait donc nécessaire d'homogénéiser les données pour Bordeaux. Jusqu'en 1988, la moyenne des dépassements est de 72.8 km/h et à partir de 1989, elle est de 68.9 km/h. Si on compare avec les relevés de Météo France (bas de la Figure 11) la coupure est décalée en 1975 et ensuite on a une longue séquence plus stable jusqu'en 2013. Les vitesses sont aussi différentes, mais cela peut tenir à une définition différente de la vitesse maximale du vent instantané (sur 1 ou 3 secondes).

Sur la Figure 12 on compare les moyennes annuelles des dépassements sur la station d'Ajaccio. Avec la NCDC (en haut), la période n'est pas homogène et se divise en trois séquences. La première assez courte montre des vitesses peu élevées jusqu'en 1977.

4. 1 nœud correspond à 1 mille marin par heure, soit exactement 1,852 km/h.

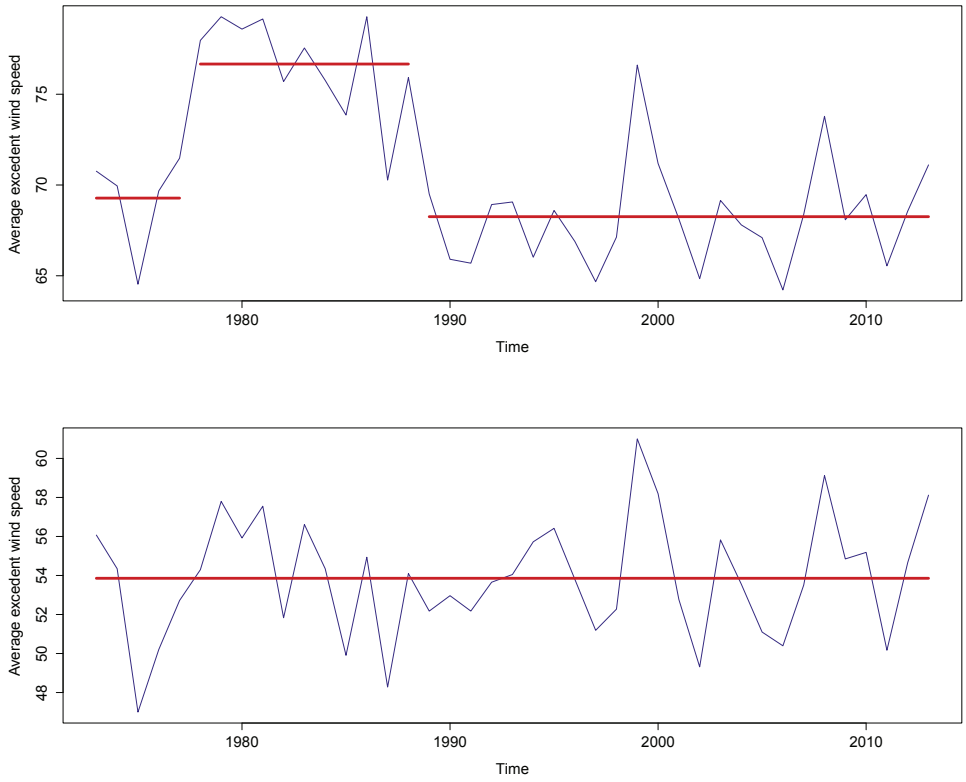


FIGURE 12 – Détection de ruptures sur les dépassements moyens de vitesses de vent à Ajaccio (Haut : NCDC - Bas : Météo France)

Ensuite de 1978 à 1988 on a une augmentation pour finalement revenir à un niveau globalement bas jusqu'en 2013. Pour Météo-France (en bas) les relevés sont beaucoup plus homogènes. Même si on retrouve la même allure générale, on ne détecte pas de rupture.

Sur la Figure 13 on compare les moyennes annuelles des dépassements sur la station de Montélimar. Les relevés de la NCDC se divisent clairement en deux séquences. La première s'étale jusqu'en 1988. Comme pour les stations bordelaises et ajacciennes on retrouve ce point commun de rupture. Les vitesses sont alors nettement plus élevées (avec une moyenne annuelle dépassant 100km/h en 1973) que sur la période plus récente (1989-2013). Chez Météo-France, la différence est moins marquée. On retrouve le pic de 1973 mais bien moins élevé (environ 82km/h). La rupture a cette fois-ci lieu en 1993.

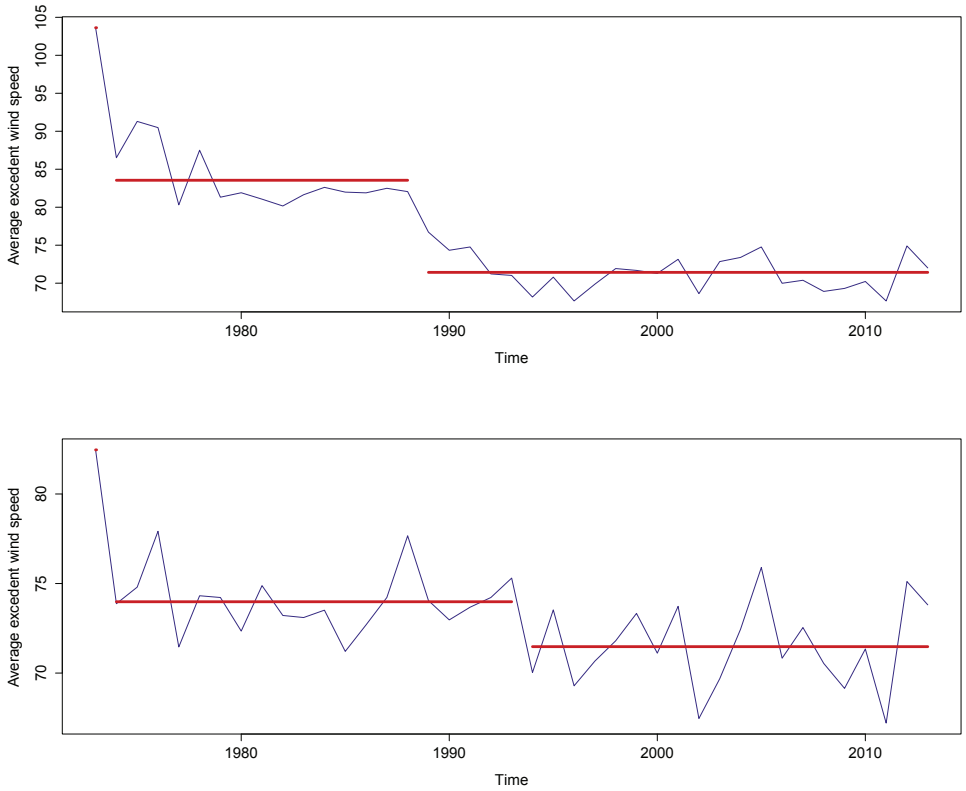


FIGURE 13 – Détection de ruptures sur les dépassements moyens de vitesses de vent à Montélimar (Haut : NCDC - Bas : Météo France)

Sur la Figure 14 on compare les moyennes annuelles des dépassements sur la station de Nantes-Bouguenais. Les deux sources montrent une division de la période en trois séquences. Si on retrouve la même allure générale dans les deux courbes, les échelles et les écarts diffèrent. Pour la NCDC, les ruptures ce situent en 1990 et 2002. Les vitesses enregistrées vont en diminuant avec le temps. Pour Météo-France, les ruptures ont lieu en 1985 et 2002. Cette fois c'est la période intermédiaire qui présente les vitesses les plus élevées.

Dans la table VI, nous avons regroupé les années de ruptures deux par deux pour pouvoir compter le nombre de stations ayant subi simultanément un changement dans la moyenne de leurs vitesses de vent mesurées. On peut voir que la période 1987-1990 compte la majorité des ruptures avec en tout 22 occurrences. Inversement les périodes de 1991 à 1998 puis post 2008 sont très peu représentées et présentent donc une plus grande homogénéité dans les relevés. Les observations les plus récentes des vitesses sont aussi les plus fiables grâce au progrès technologiques faits sur les anémomètres, d'où l'alignement des périodes trouvées avec la détection des ruptures.

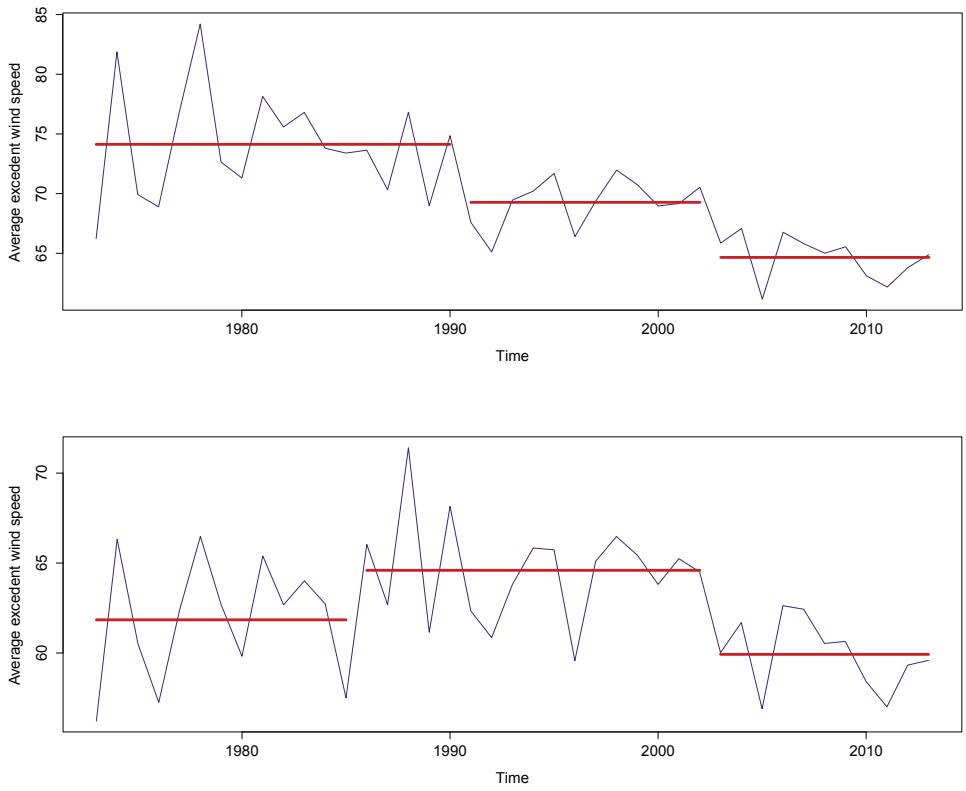


FIGURE 14 – Détection de ruptures sur les dépassements moyens de vitesses de vent à Nantes (Haut : NCDC - Bas : Météo France)

1974-1975	4		1989-1990	13
1976-1977	6		1991-1993	6
1979-1980	2		1995-1998	5
1981-1982	6		1999-2001	3
1983-1984	4		2002-2003	5
1985-1986	3		2004-2005	2
1987-1988	9		2007-2008	6

TABLE VI – Nombre de stations pour chaque rupture détectée

Plusieurs explications peuvent être fournies pour tenter d'expliquer l'origine de ces ruptures. Une modification de la topologie autour de la station suite, par exemple à la construction d'un bâtiment à proximité ou le changement du type d'anémomètre utilisé sont des sources de perturbation des mesures. On apprend dans l'article de Sacré, C. et al [74] que l'utilisation dans les années 70 des anémomètres Papillon a posé problème notamment pour estimer les fortes vitesses de vent. La figure 2.2 représente le transmetteur de l'appareil en question [61]. Ces anémomètres de type mécanique sont sensibles au gel et à l'humidité, ce qui amène à sous estimer les vitesses de vent

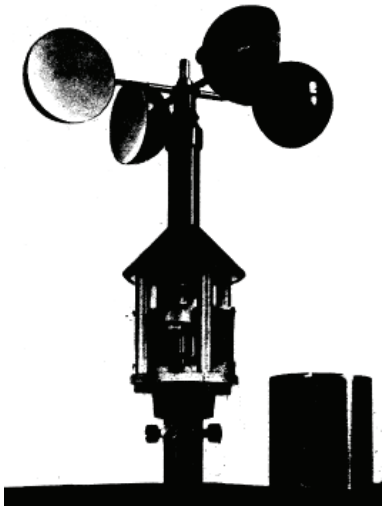


FIGURE 15 – Transmetteur de l'anémomètre Papillon

hivernales en montagne. L'appareil qui s'impose actuellement utilise des émetteurs-récepteurs d'ultrasons directionnels. Pour autant, même sophistiqué, les mesures ne sont pas à l'abri de problèmes techniques. La station de l'aéroport de Nice par exemple, présente des erreurs récurrentes malgré son profileur de vent qui s'avère incapable de détecter la brise de terre et la brise de montagne de la vallée du Var.

0.7.6 Recentrage des vitesses enregistrées

Il est parfois nécessaire d'homogénéiser les données lorsqu'une station présente des ruptures importantes. Pour Bordeaux par exemple, jusqu'en 1988, la moyenne des dépassements est de 72.8 km/h et à partir de 1989, elle est de 68.9 km/h. On harmonise la série en recentrant les dépassements de la première séquence sur ceux de la deuxième. On obtient la Figure 16.

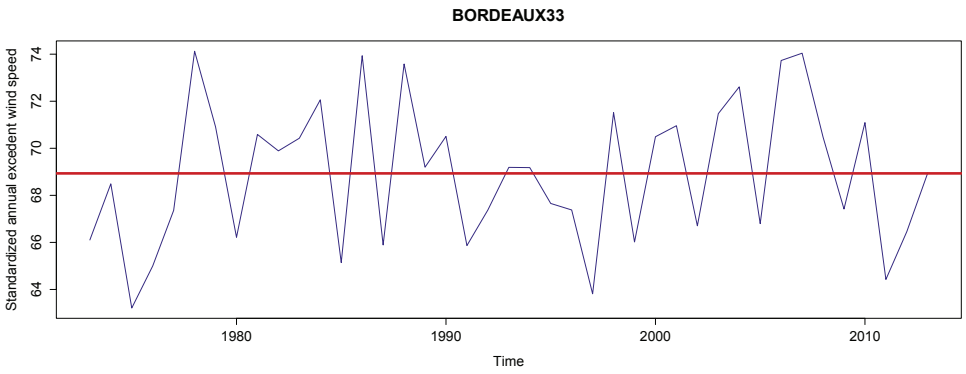


FIGURE 16 – Dépassements moyens de vitesses de vent à Bordeaux après recentrage

0.8 Informations complémentaires sur les modèles climatiques

0.8.1 Modèle multi-site pour le vent

Dans son étude sur le vent, Bessac et al. ([5] et [6]) construit des modèles stochastiques permettant de reproduire les propriétés statistiques de données spatio-temporelles de vent. L'étude se localise dans le Nord-Est de l'Atlantique (18 sites au large de la Bretagne). Ces modèles sont utilisé car les séries observées sont généralement trop courtes et présentent trop de valeurs manquantes pour estimer des probabilités d'évènements extrêmes. Les données utilisées proviennent de l'ECMWF (European Center of Medium-range Weather Forecast). Elles sont disponibles gratuitement pour les recherches à but scientifique : data-ECMWF. Les caractéristiques principales des modèles sont les suivantes :

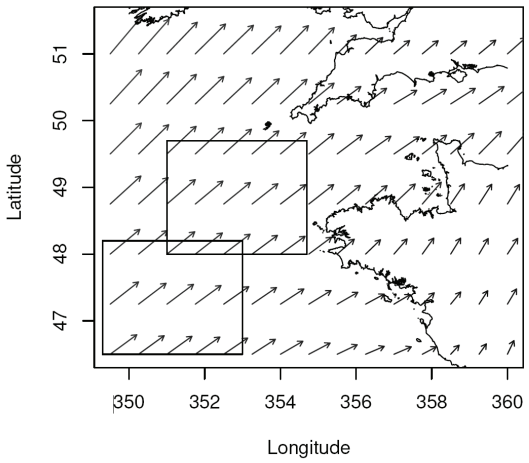


FIGURE 17 – Carte des zones étudiées - les flèches correspondent à la moyenne des composantes cartésiennes du vent

- La forte corrélation entre les stations météo suggère l'utilisation d'un signal commun à chaque site contenant une majeure partie de l'information.
- Vitesse et direction du vent sont étroitement liées.
- La direction du vent est un bon descripteur des conditions météorologiques synoptiques (i.e : système atmosphérique de large échelle)
- La modélisation est faite à partir de modèles autorégressifs à changement de régimes cachés markoviens (MS-AR)

Une critique de ces modèles est qu'ils sont indépendants dans le temps. Par exemple le changement de conditions de vents cycloniques au temps t à des conditions anticycloniques au temps $t+1$ ne dépend pas du temps t. Pourtant ce changement a généralement lieu quand le vent souffle du Nord et s'avère très rare quand il souffle du Sud. Les auteurs proposent donc un modèle non homogène qui dépend des directions

passées du vent.

0.8.2 Vélocité d'une tempête

Rychlik et Mao [73] décrivent variabilité des vitesses de vent à travers une étude de relevés maritimes lors de traversées de l'océan Atlantique. Les paramètres du modèle ont une interprétation physique naturelle et ils sont statistiquement adaptés pour représenter les données observées dans la base de vitesse de vent de l'ERA (aussi disponibles en libre accès).
Une tempête qui a lieu au temps t et au lieu p est une région où la vitesse du vent W vérifie $W(p,t) \geq u$ avec par exemple $u = 15$m/s. La limite d'une tempête est le contour de niveau u : {p : W(p,t)=u }. Cette frontière évolue selon que la tempête se déplace, grandit ou diminue.

La direction θ est appelée l'azimut principal de la tempête. Par convention on considère que la direction Sud-Nord a pour azimut $\theta = 0$ alors que la direction d'Ouest en Est a pour azimut $\alpha = 90$. La **vélocité** d'une tempête dans une direction fixée θ est alors définie par :

$$\vec{V_\theta} = -\frac{W_t}{W_\theta}, \tag{2}$$

avec W_t est la dérivée temporelle de la vitesse du vent et W_θ la dérivée directionnelle ayant pour azimut θ.

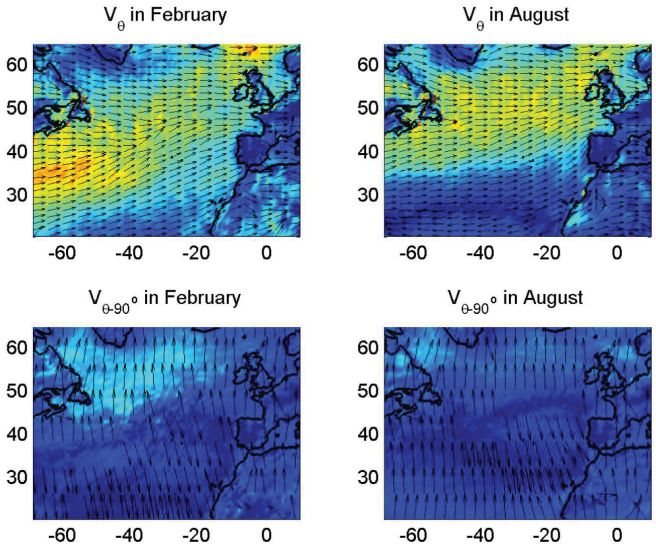


FIGURE 18 – Estimation des vélocités moyenne (km/h)

Sur la Figure 18 on peut observer pour les mois de Février et d'Août les vélocités moyennes. Les champs de vent se déplacent dans la direction θ. La couleur correspond

à la vitesse (orange pour les plus élevées et bleue pour les plus faibles). On constate que les vitesses sont plus faibles dans la direction $\theta - 90$.

0.8.3 Modélisation des rafales de vent extrêmes

La dépendance spatiale entre les rafales maximales de vent observées et prévues est ici modélisée par un processus bivarié de Brown-Resnick. L'étude se base sur les données de 110 stations dans le nord de l'Allemagne. La prévision des pics de vitesse de vent est difficile car les rafales de vent extrêmes sont rares et spatialement volatiles. On considère que la vitesse observée maximale du vent V_{max}^{obs} à un endroit fixé dépend de N observations réparties de façon régulière sur une journée. Oesting et al. [59] proposent de modéliser le lien entre les rafales de vent observées $V_{max}^{obs}(l,d)$ et prévues $V_{max}^{pred}(l,d)$ en utilisant un processus bivarié de Brown-Resnick. Les paramètres $l \in \mathbb{R}^2$ et $d \in \mathbb{N}$ sont respectivement la localisation et la date associées à la vitesse de vent. Pour les prévisions, les auteurs utilisent la loi des valeurs extrêmes généralisées (GEV) :

$$V_{max}^{pred} \sim G\left(\xi^{pred}, \mu^{pred}(l), \sigma^{pred}(l)\right). \tag{3}$$

En observant les 3 paramètres de la loi extrême, on peut voir que ξ est considéré comme indépendant dans le temps et l'espace alors que μ et σ dépendent de leur localisation. Avec une transformation adéquate, V suit une distribution standard de Gumble pour chaque localisation et chaque jour. Ainsi le champ spatial bivarié aléatoire $\{(X_{max}^{obs}(l,d), X_{max}^{obs}(l,d)), l \in \mathbb{R}^2\}$ peut alors être modélisé selon le processus de Brown-Resnick.

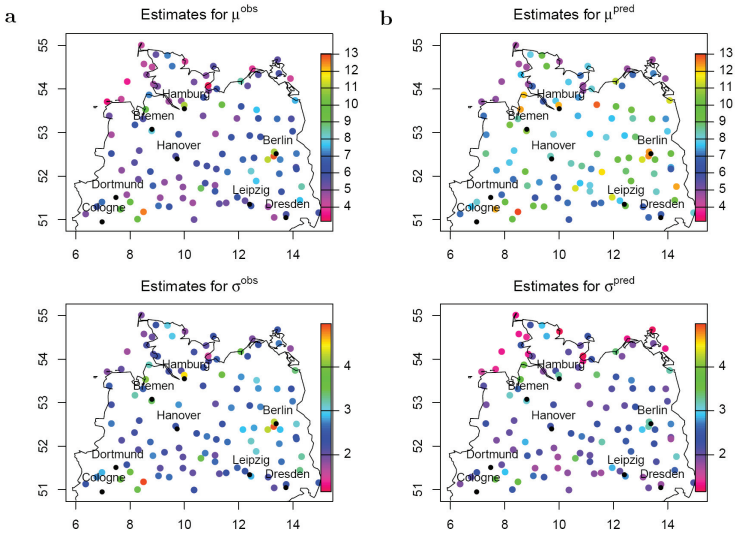


FIGURE 19 – Estimation des paramètres μ et σ pour les observations (a) et pour les prédictions (b)

Sur la Figure 19, on compare les estimations des paramètres de position et d'échelle

de la loi GEV. Les prédictions semblent assez différentes des observations. Pourtant, les corrélations entre leurs vecteurs respectifs sont fortement corrélés : 0.86 pour μ et 0.84 pour σ.

0.8.4 Lien avec le réchauffement climatique

Dans un cadre moins spécifique que celui des valeurs extrêmes, l'Onerc [58] a récemment publié son dernier rapport sur les effets du réchauffement climatique sur différents secteurs avec en particulier un groupe "risques naturels et assurance". Le groupe a étudié en particulier le risque d'inondation, le risque côtier et le risque lié au retrait-gonflement des sols argileux. En revanche, il n'a pas retenu le risque tempête, estimant que malgré les tempêtes de fin 1999 et janvier 2009, le quatrième rapport du GIEC préconise une augmentation faible du risque dans la partie Nord de la France et nulle dans le Sud.
L'actualisation des sommes en jeu se fait via "l'indicateur du pouvoir d'achat". Le modèle choisi reste assez basique car binomial, supposant que le coût moyen des dommages annuels est soit de type "bruit de fond" 200 millions d'euros, soit exceptionnel (comme par exemple en 2003) 1350 millions.
Le nombre de maisons individuelles croît dans la simulation de 0.925% par an. Le rapport va vers une véritable influence du réchauffement climatique sur la sinistralité en considérant les scénarios A2 et B2 du GIEC. le premier étant plutôt pessimiste et le second optimiste. Le changement de climat sur la France est tiré des simulations réalisées par le CNRM/Météo-France, avec le modèle Arpège-Climat, qui se situe dans la moyenne des modèles climatiques du GIEC en terme de réchauffement. Par exemple en Languedoc Roussillon, le coût des dommages liés aux aléas "submersion permanente" et érosion sur 100 ans est évalué entre 15 et 35 milliards d'euros 2008.

Le modèle de l'Onerc reste trop réducteur pour servir à notre étude sur les extrêmes mais pourra servir d'inspiration dans son utilisation des hypothèses d'évolution de climat, d'actualisation et de développement des infrastructures.

0.9 Méthodes de calcul

On présente ici différentes méthodes de calculs dont la nôtre, pour modéliser un indice de tempête à partir de données météorologiques. Ces indices ont pour but de lier les coûts des dommages enregistrés par les assureurs à des informations sur la vitesse du vent.

0.9.1 Méthode 1

Pays : France
Période : 1963-2013
Données assurance : sinistres (1998-2013)/ sinistres majeurs (1973-1998) - Allianz
Données météo : vitesses du vent journalières sur 173 stations

La méthode que nous avons développée dans l'article *Index for Predicting Insurance Claims from Wind Storms with an Application in France* [57] regroupe deux étapes :

- Construction d'un indice vent I_w à l'échelle locale.
- Construction à partir de cet indice vent d'un indice tempête I_S à l'échelle globale.

Notre indice de vent se calcule comme une différence entre $w^d(s)$ la vitesse maximale du vent enregistrée par la station s au jour d et $w_q(s)$ le quantile à $q\%$ sur l'ensemble des vitesses de la période dans cette même station. On ne retient que les dépassements avec la partie positive. L'exposant α sert à contrôler l'influence des dépassements selon leur intensité.

$$I_w^d(s) = \left([w^d(s) - w_q(s)]_+\right)^\alpha \tag{4}$$

Notre indice tempête se construit ensuite comme une agglomération pondérée des indices de vent :

$$I_S = \exp\left(\beta \sum_{d\in E} \sum_k \frac{I_w^d(s) \times G(s)}{N_a^d}\right), \tag{5}$$

avec $I_w^d(s)$ l'indice de vent enregistrée dans chaque station s à la date d. L'exposition au risque est considérée en pondérant cette vitesse par le portefeuille global d'Allianz (professionnels + particuliers) $G(s)$. Nous tenons aussi compte de la taille de la zone sinistrée à travers une aggregation géographique des dégâts, nous sommons sur la durée du passage de la tempête E selon les événements. Enfin le nombre de stations actives chaque jour est exprimé par N_a^d. Les formules ont été établies empiriquement. L'optimisation s'est ensuite faite sur les deux paramètres α et β de façon à minimiser l'écart entre les coûts globaux et l'indice de tempête lors des événements majeurs de la période 1973-2013.

Cette méthode a ensuite été affinée dans un deuxième article. Pour les vitesses de vent, nous avons tenu compte des problèmes de rupture que nous avons corrigés. Pour

la construction de l'indice, nous avons délimité 6 zones tempêtes. Dans chacune des six zones $A(k)$ précédemment délimitées, nous construisons un indice tempête spécifique associant les indices de vent à l'exposition de chaque département en terme de risque portefeuilles. Pour $k \in \{1, ..., 6\}$ nous définissons $I_S(k)$ comme

$$I_S(k) = \sum_{s \in A(k)} R(s) \times \max_{d \in E} \left(\frac{I_w^d(s)}{N^d} \right), \tag{6}$$

avec $R(k)$ les risques portefeuille. Notons ici un changement de méthode sur le choix du maximum et non plus de la somme des indices sur le cluster temporel de chaque tempête. Contrairement aux charges d'assurance qui sont additionnées sur la durée de l'événement, nous avons constaté que les pics de vent reflètent mieux l'intensité relative des phénomènes climatiques. On obtient alors un indice tempête global sur l'ensemble de la France en agglomérant les six indices tempête pondérés à la fois par le portefeuille et par un paramètre $B = \{\beta_1, ..., \beta_6\}$. Ce paramètre supplémentaire permet de sur ou sous pondérer chacune des zones selon l'impact en terme de dommage aux biens qu'aurait le passage d'une tempête sur son territoire. Pour notre étude, les β_i sont uniformément répartis entre 0.1 et 2 et indépendants. La formule devient

$$I_S = \sum_{k=1}^{6} R(k) \times I_S(k) \times \beta_k, \tag{7}$$

avec $R(k)$ le poids relatif de chaque zone dans le portefeuille Allianz. Nous avons optimisé cet indice pour qu'il reflète au mieux la répartition des dommages les plus extrêmes en terme de rang. Notre classement de référence pour les tempêtes les plus importantes depuis les années 70 est issu des travaux de M. Luzi [51]. Nous avons déduit de ce classement un dernier ajustement au niveau des écarts entre les valeurs de l'indice pour qu'elles soient proportionnelles aux charges actualisées des dommages. Pour distendre la distribution, nous employons la fonction exponentielle et un dernier paramètre : γ. Le lien entre notre indice tempête et le coût C pour l'assureur est le suivant :

$$C = \exp(\gamma \times I_S). \tag{8}$$

0.9.2 Méthode 2

Pays : Suède
Période : 1982-1993
Données assurance : sinistres individuels liés aux tempêtes - Länsförsäkringar
Données météo : pressions atmosphériques et vitesses de vent uniquement relevées pour 58 tempêtes sur 6 stations qui sont ensuite étalées sur un maillage de 50 kilomètres

En 2001, Rootzén et Tajvidi [72] proposent deux modèles :

– un modèle log-linéaire 9

$$\log(\text{loss}_i) = \alpha_0 + \sum_{j=1}^{18} \alpha_j p_j + \epsilon_i \tag{9}$$

avec loss_i la somme des dommages dans la région i, p_j les pressions du vent aux 18 points du maillage et les ϵ_i indépendants.

– un modèle suivant la distribution généralisée de Pareto 10 (GPD)

$$H(x) = 1 - \left(1 + \gamma \frac{x}{\sigma}\right)_+^{\frac{-1}{\gamma}} \tag{10}$$

avec σ une somme de fonctions exponentielles de la pression du vent aux différents points du maillage.

0.9.3 Méthode 3

Pays : Suède
Période : 1982-2005
Données assurance : sinistres individuels et professionnels tempêtes - Länsförsäkringar
Données météo : non utilisées

L'étude de Brodin et Rootzén [10] publiée en 2008 se focalise sur le coût des tempêtes majeures enregistrées en Suède (104 événements dépassent 1.5 millions de couronnes suédoises soit 200 000 euros). Deux modèles sont proposés, d'abord univarié puis bivarié.

Modèle 1

Dans ce modèle les petits sinistres sont modélisés non paramétriquement à partir de leurs moyennes et de leurs variances et les dépassements au-delà du seuil u suivent une loi de Pareto généralisée.

$$F(x) = 1 - \left(1 + \gamma \frac{x}{\sigma}\right)_+^{-1/\gamma} \tag{11}$$

dont ils déduisent le quantile supérieur à $p\%$ dans l'intervalle $[0, T]$

$$x_{T,p} = u + \frac{\sigma}{\gamma}\left(\frac{(\lambda T)^\gamma}{(-log(1-p))^\gamma} - 1\right) \tag{12}$$

Modèle 2

Ici les sinistres concernant l'industrie forestière (F) sont traités séparément de ceux concernant les bâtiments (B). Le modèle est donc un modèle bivarié symétrique logistique généralisé de Pareto : BGPD. On représente les excès $(X, Y) = 5B - u_B, F - u_F)$. Ensuite, la fréquence des événements est gérée par un processus de Poisson. Les charges sont supposées iid et indépendantes du processus de Poisson.

On modélise le coût total (M_T) lorsque l'exposition du portefeuille forestier augmente d'un facteur a soit un coût total pour la tempête égal à $B + aF$. Avec le seuil $v = u - u_B - au_F > 0$, on obtient la formule

$$P[M_T \leq u] = \exp(-\lambda T P[X + aY > v]), \tag{13}$$

avec λ l'intensité du processus de Poisson.

0.9.4 Méthode 4

Pays : Allemagne
Période : 1980-1997
Données assurance : résultats annuels de l'institut des assureurs allemands (GDV)
Données météo : vitesses de vent journalières sur 24 stations

D'autres formules ont déjà été proposées pour calculer l'indice de vent. M. Klawa et U. Ulbrich [45] utilisent un indice cubique appliqué à un ratio entre la vitesse maximum du vent sur une journée et le quantile à 98% des vitesses de vent enregistrées sur une station. CI_w :

$$CI_w(k) = \left(\frac{w(k)}{w_{98}(k)} - 1\right)^3 \tag{14}$$

Pour chaque évènement tempête, les auteurs évaluent le coût pour les assureurs à partir de l'équation suivante :

$$\text{loss} = c \times \sum_k \text{pop}(k) \left(\frac{w_{\max}(k)}{w_{98}(k)} - 1\right)^3 \tag{15}$$

où $\text{pop}(k)$ représente la population dans chaque district.

0.9.5 Méthode 5

Pays : Allemagne
Période : 1997-2007
Données assurance : sinistres particuliers - Verbundene Wohngebäude Versicherung
Données météo : vitesses de vent journalières issues de modèles atmosphériques sur un maillage de 79 km (0.7degrés) - ERA-Interim et NCEP

La formule de l'indice de vent est ici légèrement complexifiée avec la formule de Donat et al.[24] :

$$\text{loss ratio}(k) = A(k) \left(\left[\frac{w(k)}{w_{98}(k)} - 1\right]_+\right)^3 + B(k) \tag{16}$$

avec $A(k)$ la pente de la régression spécifique à chaque région et $B(k)$ l'ordonnée à l'origine. Cette approche tient compte des spécificités locales de chaque région. Par exemple pour une même vitesse de vent, une zone fortement urbanisée aura généralement plus

de dégâts qu'une zone rurale.

Pour calculer la sinistralité au niveau national, on a la formule suivante qui pondère pour chaque événement e la somme des pertes par la somme des valeurs assurées sur une année y :

$$\text{cumulated loss ratio}(e) = \frac{\sum_k \text{value}(k) \times \text{loss ratio}(k, e)}{\sum_k \text{value}(k, y)} \tag{17}$$

Cette approche est reprise par Pinto et al. [65], dans le contexte du réchauffement climatique. A partir des scénarios des modèles de circulation générale de l'ECHAM5, ils évaluent l'évolution des pertes et des périodes de retour durant le 21ème siècle. Après 2060, les charges maximales pourraient augmenter d'environ 65% selon les scénarios A1B et A2. La plupart des pays d'Europe centrale et de l'ouest sont touchés. Des changements significatifs sont attendus dès 2027.

0.9.6 Méthode 6

Pays : États-Unis (Côte Est et Golfe)
Période : 1900-2005
Données assurance : Coût économique historique - Monthly Weather Review / Storm summary data - National Hurricane Center (NHC)
Données météo : non utilisées

Pielke et al. [63] proposent une étude historique du coût des ouragans aux États-Unis. L'actualisation sur la période qui dépasse les 100 ans se fait avec les deux formules suivantes :

$$D_{2005} = D_y \times I_y \times RWPC_y \times P_{2005/y} \tag{18}$$

avec D_{2005} le coût actualisé des dommages en 2005, D_y le coût historique, I_y le taux d'inflation, $RWPC_y$ le taux d'évolution des richesses par personne (real wealth per capita) et $P_{2005/y}$ l'évolution de la population côtière.

$$D_{2005} = D_y \times I_y \times RWPHU_y \times HU_{2005/y} \tag{19}$$

avec D_{2005} le coût actualisé des dommages en 2005, D_y le coût historique, I_y le taux d'inflation, $RWPHU_y$ le taux d'évolution des richesses par habitation (real wealth per housing unit) et $HU_{2005/y}$ l'évolution du parc immobilier côtier. Par exemple l'ouragan Frederic de 1979 est évalué entre 10.3 et 11.5 milliards de dollars selon les deux formules. L'évènement majeur de la période a eu lieu en 1926, il s'agit du grand ouragan de Miami. Après actualisation, il est évalué à 72 milliards de dollars de 1995, à 157 milliards en dollars de 2005 et si on extrapole, il atteindrait les 500 milliards en valeur équivalente en 2020.

0.9.7 Méthode 7

Pays : Chine
Période : 1948 - 1999
Données assurance : non utilisées
Données météo : Températures et activité cyclonique - National Centers for Environmental Prediction (NCEP)/National Center for Atmospheric Research (NCAR)

Qian et al. [69] proposent un indice décrivant la fréquence des tempêtes de poussière en Chine. Il s'avère que ces événements étaient deux fois plus fréquents entre les années 50 et 70 que depuis le début des années 80. Ce changement pourrait s'expliquer par un réchauffement en Mongolie associé à un refroidissement dans le nord de la Chine qui auraient pour effet une diminution du gradient méridional de température. L'indice DI proposé est le suivant :

$$DI = r_C \frac{\Delta C}{\delta_C} - r_T \frac{\Delta T}{\delta_T} \tag{20}$$

avec T les températures de l'hiver précédent, C les cyclones ayant eu lieu durant le printemps et r_C, r_T des paramètres de pondération. La corrélation obtenue est supérieure à 0.5 avec un intervalle de confiance à 99%.

0.9.8 Méthode 8

Pays : Mongolie intérieure (Chine)
Période : 1961 - 2000
Données assurance : non utilisées
Données météo : vitesses de vent, vortex polaire et indices de circulation sur 37 stations

Zhao et al. [83] proposent un autre indice pour la fréquence des tempêtes de poussière se basant cette fois sur des facteurs climatiques. Les données regroupent le nombre de jours avec des rafales (NDWG], des indices d'intensité et de zone du vortex polaire de l'hémisphère nord (IINHPV et AINHPV) et du vortex polaire asiatique (IIAPV et AIAPV) et d'autres indices de circulation. Après régression, trois facteurs sont retenus pour exprimer la fréquence de tempête Y :

$$Y = -85.27 + 0.72\text{NDWG} + 0.48\text{IIAPV} + 0.82\text{AINHPV} \tag{21}$$

Ces facteurs permettent d'expliquer la diminution de la fréquence des tempêtes depuis les années 60 à travers des facteurs climatiques corrélés aux circulations d'air froid sur une large échelle.

0.10 Effets du clustering sur le calcul des périodes de retour

Nous avons exploré plusieurs approches pour mesurer la sensibilité de nos résultats. Le choix de recoupement des données journalières par événements et par clusters est ici étudié. Remarquons avant tout qu'il faut toujours partir et revenir à la base de ce que l'on cherche à traiter en pratique. Dans notre cas, comme il s'agit de mesurer les charges tempêtes annuelles supportées par un assureur, voire les charges nettes de réassurance, il faut toujours pouvoir retraduire les solutions en ces termes. Pour les assureurs et les réassureurs, les événements sont définis par un laps de temps et une étendue géographique. Bien sûr, il faut pouvoir tenir compte des contraintes administratives pour respecter la réalité du terrain (délais de déclaration).

En ce qui concerne les clusters, il faut être capable de relativiser leur importance dans ce type de travail. Si certaines conditions atmosphériques peuvent favoriser le déclenchement de tempêtes ou au contraire les empêcher, les historiques d'assurance disponibles sur quelques décennies, intègrent déjà ces conditions. Donc implicitement, les historiques contiennent des périodes à cluster (printemps 1990, décembre 1999) et des périodes plus calmes. Peut-on réellement approfondir cette problématique avec nos données ? Quel gain cela peut-il apporter dans la précision de nos résultats ? Les trois aspects suivants résument la position actuelle des assureurs :

- l'effet cluster doit être pris en compte (les perturbations atmosphériques extrêmes peuvent avoir des conséquences pendant des périodes de quelques semaines)
- l'effet cluster n'impacte pas la prime pure
- l'effet cluster augmente la volatilité des prévisions de charge annuelle

Sous ces hypothèses, l'intégration d'un effet cluster ne peut ainsi avoir qu'un impact de second ordre sur la prime pure lié à la rémunération de capitaux sous risques plus importants.

0.10.1 Par événements : clusters = 2-3 jours

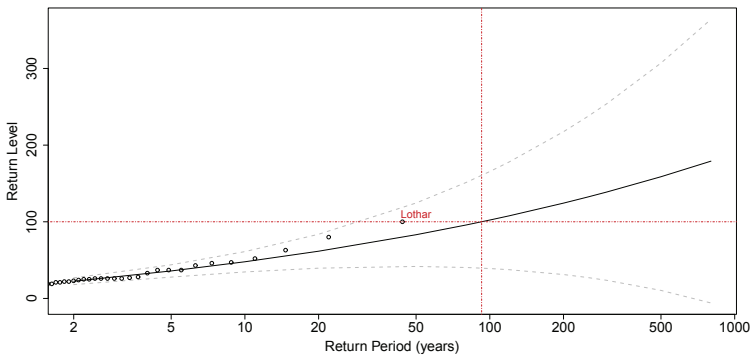


FIGURE 20 – Période de retour pour l'indice tempête.

Ici, chaque cluster correspond à un événement tempête selon les critères de l'assureur

(période durant laquelle sont reçues les déclarations de sinistres). Sur la Figure 20, on a une tempête de l'importance de Lothar tous les 93 ans. Ce niveau de retour apparaît en pointillés rouge.

0.10.2 Par semaines : clusters = 7 jours

Nous distinguons deux types d'approches avec d'un côté la somme des valeurs de l'indice tempête enregistrées sur la durée du cluster et de l'autre côté le maximum de ces valeurs.

Somme des valeurs

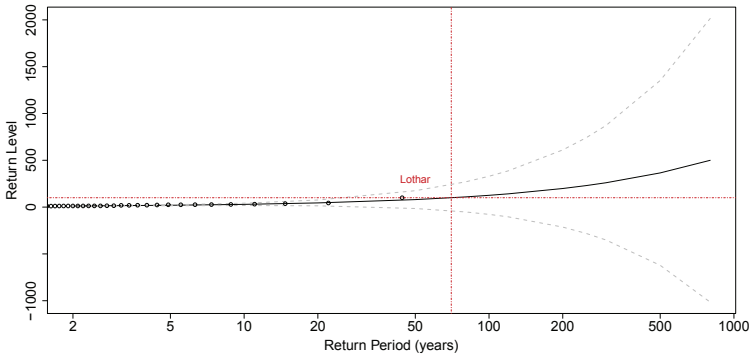


FIGURE 21 – Période de retour pour la somme des indices tempête.

Sur la Figure 21, on obtient une semaine du niveau de Lothar et Martin combinés tous les 70 ans.

Maximum des valeurs

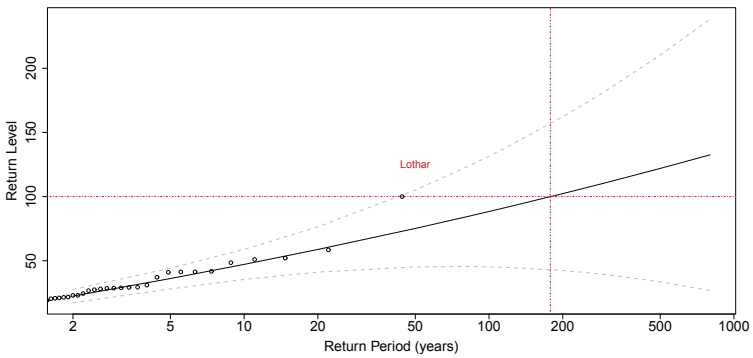


FIGURE 22 – Période de retour pour le maximum des indices tempête.

Sur la Figure 22, on a une semaine du niveau de Lothar et Martin combinés tous les 178 ans. Après ces deux tests, nous constatons que la somme des valeurs sur un cluster d'une semaine a tendance à diminuer la période de retour alors que l'approche par maximum augmente fortement cette même période.

0.10.3 Par mois : clusters = 28-31 jours

Somme des valeurs

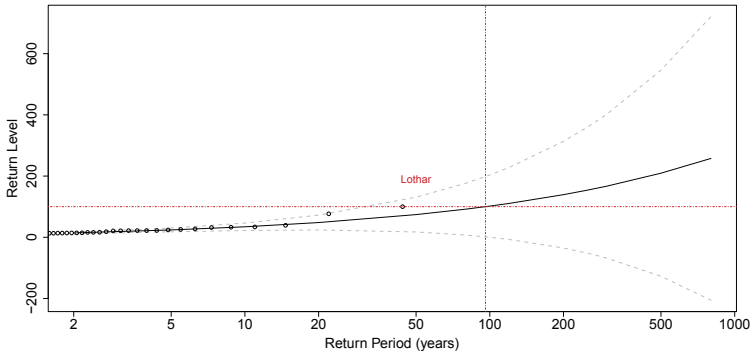


FIGURE 23 – Période de retour pour la somme des indices tempête.

Sur la Figure 23, on obtient un mois du niveau de Lothar et Martin combinés à d'autres tempêtes de moindre importance environ tous les 96 ans.

Maximum des valeurs

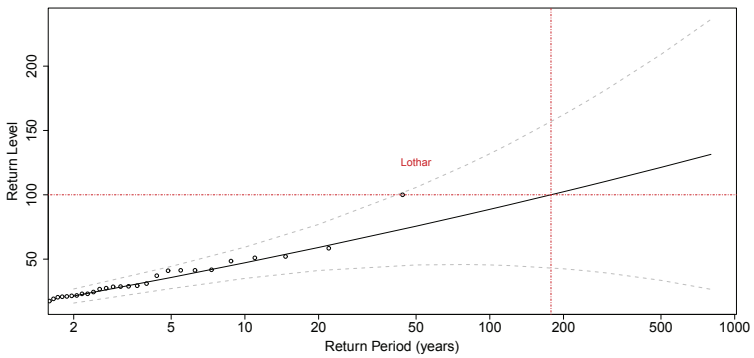


FIGURE 24 – Période de retour pour le maximum des indices tempête.

Sur la Figure 24, on obtient un mois du niveau de Lothar et Martin combinés à d'autres tempêtes de moindre importance environ tous les 178 ans.

0.10.4 Récapitulatif

Dans la Table VII nous proposons un récapitulatif des variations de la période de retour du cluster le plus marquant de la période (contenant dans tous les cas Lothar mais associé à d'autres tempêtes ayant eu lieu dans le même cluster) selon le type de découpage et la méthode d'agglomération choisie.

Méthode	somme			maximum	
Type de cluster	event	week	month	week	month
Lothar Cluster Return Period	93	70	96	178	178

TABLE VII – Calculs de périodes de retour selon différents scénarios

En conclusion, nous ne constatons aucun effet notable de l'agrégation par la somme sur une durée plus ou moins longue sur les périodes de retour. Cette stabilité est plutôt rassurante. En revanche, les différences des périodes de retour selon l'utilisation de la somme ou du maximum sont considérables.

0.11 Assurance automobile et comportementale

Cette partie regroupe différents résultats incluant le kilométrage au calcul du risque. Plusieurs approches sont présentées. Quelques résultats proviennent du secteur de l'assurance, mais d'autres ont été publiés par des instituts de recherche sur les transports. Certaines observations sont confrontées aux prévisions obtenues à partir de modèles linéaires généralisés. En complément du kilométrage, d'autres pistes sont explorées dans le cadre d'une assurance comportementale. Les conclusions sur l'impact de variables explicatives comme le sexe ou la période de la journée diffèrent selon l'utilisation de taux ajustés ou non. Le taux de sinistres est fortement corrélé au kilométrage, mais cette relation n'est ni monotone ni linéaire. Nous allons voir à travers ces différents exemples que les méthodes et les mesures employées influencent fortement les résultats. La création d'un modèle pertinent doit prendre en compte l'ensemble de ces interprétations pour une évaluation du risque la plus cohérente possible. Nos résultats sur ce sujet sont présentés dans le chapitre 3 de cette thèse.

0.11.1 Historique

Le sociologue Roger Roots [71] donne une vision nuancée des dangers représentés par l'automobile à travers la perspective historique des transports par chemin de fer et à cheval. Un calcul du risque basé sur les distances parcourues plutôt que sur la population amène à considérer l'automobile comme beaucoup plus sûre que ses prédécesseurs. À l'heure actuelle, les accidents liés aux véhicules motorisés représentent néanmoins la première cause de mortalité non intentionnelle et la plus grande source de mortalité chez les enfants et les jeunes adultes.

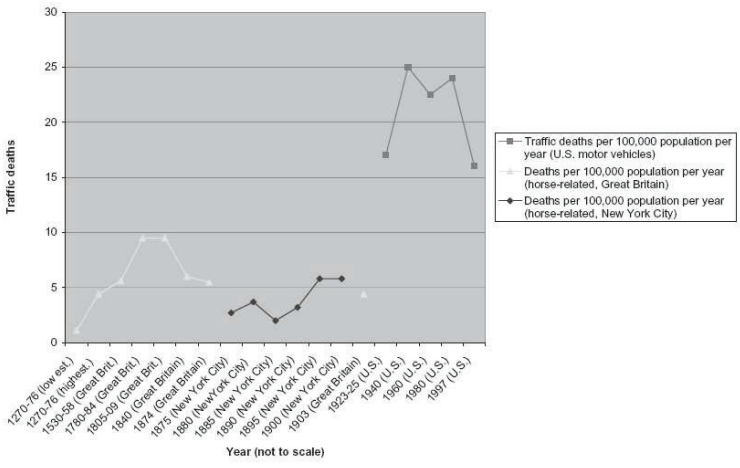


FIGURE 25 – Taux d'accidents mortels par personne – Chevaux comparés aux Automobiles – Sources : McShane (1994) ; Hair (1971).

Après avoir constaté un taux de mortalité pour 100 000 personnes 3 fois plus élevé pour les trajets en voiture de nos jours que pour ceux à cheval au début du 19ème siècle,

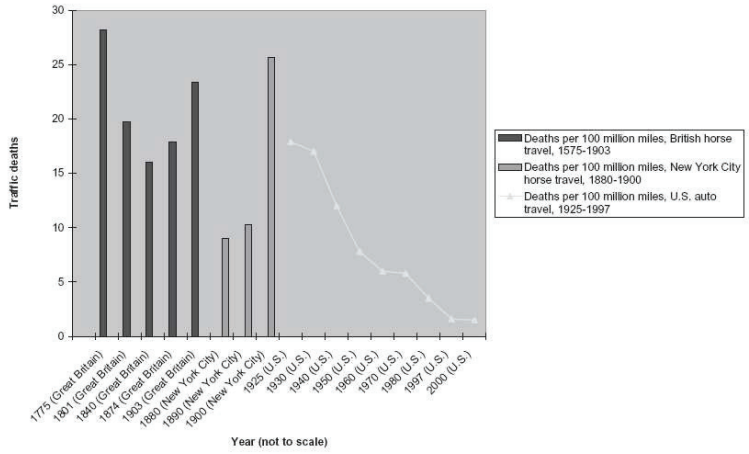


FIGURE 26 – Taux d'accidents mortels par miles – Chevaux comparés aux Automobiles – Sources : McShane (1994) ; Hair (1971).

Roots introduit le critère du kilométrage pour une nouvelle appréciation. Les américains sont aujourd'hui beaucoup plus mobiles qu'ils ne l'étaient en 1900, en grande partie de part leur dépendance aux véhicules personnels motorisés. Parcourir régulièrement de grandes distances est entré dans la culture et le mode de vie américain. Le taux de mortalité dans des accidents auto est inférieur à 2 pour 100 millions de miles parcourus annuellement. En comparaison, ce taux était 15 fois supérieur avec les transports à cheval en 1775. Cette évolution est en partie due aux améliorations des infrastructures et de la prise en charge médicale mais doit aussi correspondre à une amélioration de la sécurité apportée par l'automobile.

0.11.2 Tarification

La tarification du risque automobile connaît actuellement une évolution à la fois technique et dans son rapport avec les assurés. L'objectif principal reste d'obtenir la meilleure adéquation possible entre le risque et le tarif. Cependant, de nouveaux critères comme le kilométrage on fait leur apparition et cette tendance vers une assurance comportementale est en plein essor.

Litman du Victoria Transport Policy Institute [50] a collecté les kilométrages annuels obtenus lors de contrôles techniques qu'il a fait correspondre aux indemnités d'assurance pour plus de 700 000 véhicules en Colombie britannique. Sur la Figure 27 on peut comparer la sinistralité selon trois catégories de dommages : non responsable, responsable et corporel. Il a ainsi montré que pour une classe d'assurés donnée, la fréquence annuelle des accidents augmente avec le kilométrage. La croissance la plus marquée concerne les sinistres responsables.

Litman justifie aussi une tarification des primes d'assurances basée sur le kilométrage

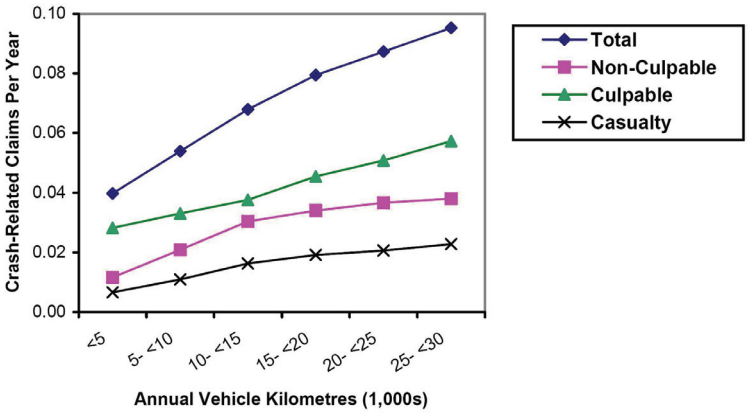


FIGURE 27 – Taux d'accidents selon le kilométrage annuel

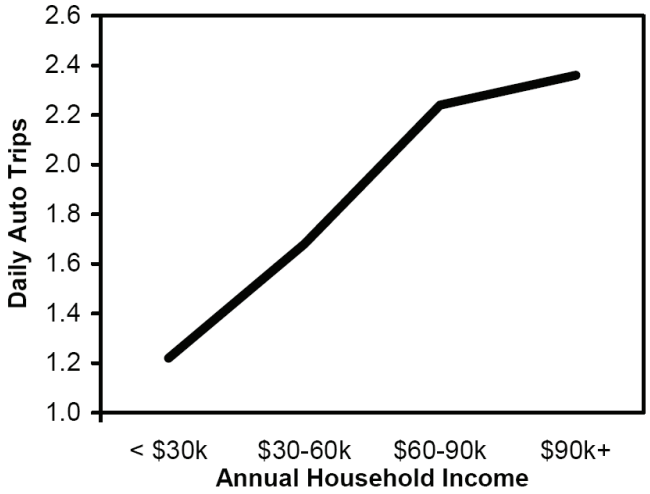


FIGURE 28 – Nombre de trajets journaliers selon le revenu

en faisant observer un lien entre revenus et kilomètres parcourus. Sur la Figure 28 nous constatons en effet que le nombre de trajets journaliers augmente avec les revenus annuels par foyers.

L'institut de recherche sur les transports de l'université du Michigan [54] fournit une étude sur les effets de 4 variables explicatives – (l'âge, le sexe, la période de la journée, et le kilométrage moyen) – sur le taux d'accidents automobile. Il confronte observations et projections. D'après les résultats à leur disposition, le taux de décès pour 100 millions de miles parcourus en automobile (MPA) était supérieur de 55% pour les hommes (3,5 contre 2,2). Au contraire, les femmes avaient un taux 26% supérieur à celui des hommes pour les accidents non mortels (2,3 contre 1,8). De plus, en considérant les Dommages Matériels Uniquement (DMU), les femmes ont aussi un taux 12% supérieur (4,2 contre 3,7). Cette tendance moins connue a aussi été relevée lors d'études australiennes (Drummond et Yeo [27], 1992), britanniques et canadiennes (Broughton 1990,

Downs 1988, Grime 1987, Mercer 1989) puis norvégiennes avec 63% de risque en plus pour les femmes (Bjørnskau [8], 1994). Une possible explication serait que les conducteurs acquièrent de l'expérience avec les kilomètres.

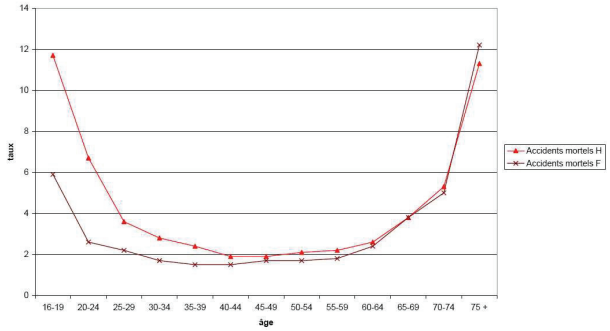


FIGURE 29 – Taux d'accidents mortels / 10^8 miles

Sur la Figure 29, on compare les taux d'accidents mortels pour un kilométrage global donné (100 millions de miles parcourus par des automobiles). Ces taux sont observés selon le sexe du conducteur et pour différentes tranches d'âge. Le premier constat est une augmentation générale du nombre de sinistres pour les jeunes conducteurs (jusqu'à 25 ans) et pour les seniors (à partir de 70 ans). Les hommes ont une sinistralité plus forte pour quasiment toutes les tranches d'âge mise à part la tranche 75 ans et plus.

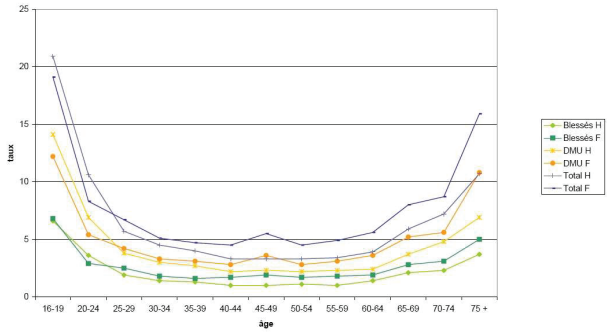


FIGURE 30 – Taux d'accidents par catégorie / 10^6 miles

En différenciant selon les catégories de dommages, les résultats ne sont plus les mêmes. Sur la Figure 30 on peut voir par exemple que dans le cas des sinistres corporels, les femmes ont plus d'accidents que les hommes pour toutes les tranches d'âge sauf les plus jeunes (16 à 24 ans). En ce qui concerne les dommages matériels, le taux de sinistres des hommes n'est supérieur à celui des femmes que pour la tranche 16-29 ans, ensuite il reste inférieur pour toutes les tranches plus âgées.

Les auteurs proposent ensuite leur modèle. La méthode statistique considère 4 variables explicatives dont 2 sont catégorielles. Le sexe prend la valeur 0 pour les hommes

et 1 pour les femmes, la période prend la valeur 0 le jour (de 6 à 21h) et 1 la nuit. L'âge du conducteur et le kilométrage annuel moyen sont traités comme des variables continues, elles sont centrées et réduites. Les raisons de cette sélection sont les suivantes : l'âge et le sexe sont clairement reliés au taux d'accident comme l'ont montré de nombreuses études antérieures. Le kilométrage est ici testé. Le risque d'implication dans un accident varie aussi sur la période de la journée. Le taux d'accident est généralement plus élevé la nuit que le jour. De plus, une interaction période/sexe pourrait décrire un taux d'accidents non mortels supérieur pour les femmes durant la journée mais équivalent durant la nuit. Le modèle utilisé est un modèle de Poisson avec lien logarithmique. Des modèles séparés ont été développés par décrire les accidents mortels, avec blessés et DMU. Un lien quadratique est décelé entre l'âge et le logarithme du taux d'accident, la variable « âge^2 » est donc incluse au modèle. Au final, les variables explicatives retenues en plus du sexe et de l'âge par une méthode descendante sont :

– kilométrage, période
– sexe × période, sexe × kilométrage
– période × âge, période × âge^2

Pour les trois types de modèles, le coefficient associé au kilométrage est négatif, ce qui correspond à une diminution du taux d'accident avec le kilométrage. L'interaction sexe/kilométrage renforce cet effet pour les femmes, le taux d'accident diminue plus fortement lorsque le kilométrage annuel augmente chez les femmes que chez les hommes.

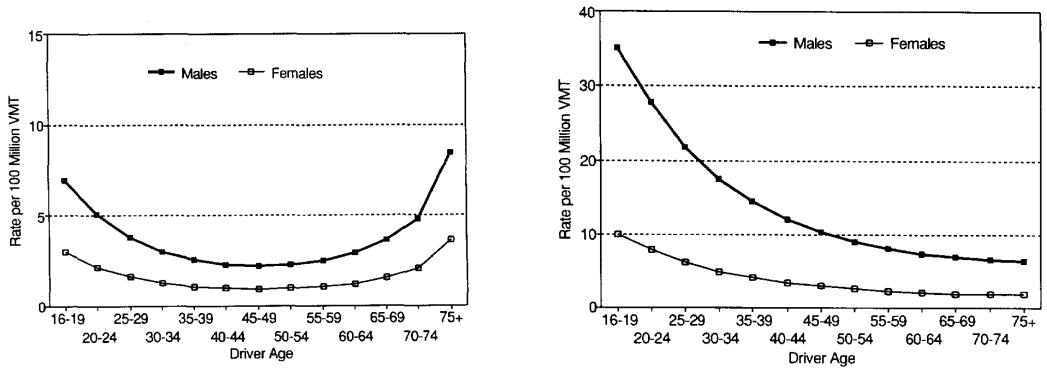


FIGURE 31 – Taux d'implication dans un accident mortel de jour (à gauche) et de nuit (à droite)

Les Figures 31 et 32 sont obtenues à partir du modèle proposé et d'un taux ajusté au kilométrage moyen par tranche d'âge. Les résultats sont divisés en 2 groupes selon la période de la journée. Les deux premiers graphiques sont consacrés aux accidents mortels. Les courbes sont différentes le jour et la nuit. En journée, les juniors et les seniors sont les plus touchés avec un maximum chez les hommes de 75 ans et plus. Pour toutes les tranches d'âges le taux de sinistres mortels est supérieur chez les hommes.

Les deux graphiques suivants concernent les dommages corporels non mortels. En

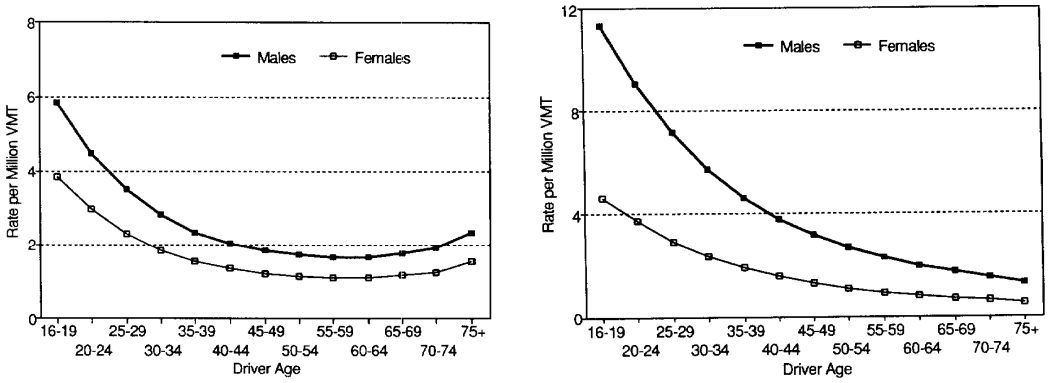


FIGURE 32 – Taux d'implication dans un accident avec blessé(s) de jour (à gauche) et de nuit (à droite)

journée, ce sont les jeunes qui ont le maximum d'accidents. Chez les seniors on note une légère augmentation mais bien moins marquée que pour les accidents mortels. On retrouve avec les modèles linéaires les mêmes tendances de taux d'accident supérieurs chez les hommes pour les accidents mortels, mais inférieurs pour les autres types d'accident. Ces tendances sont amplifiées par les taux ajustés.

Parmi les assureurs proposant d'ores et déjà une tarification de leurs primes auto directement liée au kilométrage, la compagnie belge Corona Direct [22] est un des seuls à diffuser directement ses prix sur son site internet. On peut dès lors avoir une première idée du taux de prime par kilomètres parcourus et de la prime fixe :

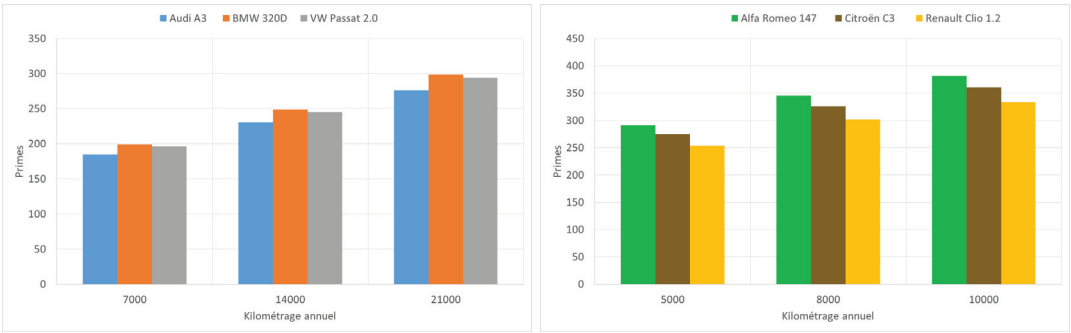


FIGURE 33 – Tarification pour un homme de 46 ans (à gauche) et une femme de 28 ans (à droite)

Dans l'exemple présenté sur le site (Figure 33), trois types de véhicules différents sont proposés selon le sexe du conducteur. Les deux individus sélectionnés sont un homme de 46 ans habitant à Bastogne, une petite ville de la région wallonne et une femme de 28 ans habitant à Malines, une ville moyenne de la région flamande. Pour les hommes, le site nous propose une Audi A3, une BMW 320D et une Volkswagen Passat 2.0 et une sélection de trois tranches de kilométrages annuels : 7 000, 14 000 et 21 000. Pour les femmes une Alfa Romeo 147, une Citroen C3 et une Renault Clio et les trois

tranches sont 5 000, 8 000 et 10 000 kilomètres parcourus dans l'année.

Le premier constat est que l'évolution des prix est linéaire avec le kilométrage. Les pentes observées sont sensiblement les mêmes quels que soient les véhicules sélectionnés avec un léger facteur d'augmentation en fonction du prix de la prime de base. En moyenne, on obtient respectivement pour l'homme de 46 ans et la femme de 28 ans une prime fixe de 145.51 et de 187.73 euros à laquelle s'ajoute une prime de 0.68 et de 1.71 cents par kilomètres parcourus.

Comme nous le verrons dans le chapitre 3, les critères retenus pour la tarification de l'assurance automobile sont soumis à réglementation. À ce sujet, le projet `PriPAYD`, une version de la tarification PAYD respectueuse de la vie privée des assurés, est développée par Troncoso et al. [80]. Dans la procédure classique d'assurance PAYD, les informations utilisées pour calculer le tarif sont la plupart du temps collectées par une boîte noire à l'intérieur du véhicule puis transférées à la compagnie d'assurance. Cette dernière a donc la capacité de suivre les allées et venues de n'importe lequel de ses usagers, facilement et avec précision. Dans le modèle `PriPAYD`, le calcul de la prime est effectué dans la boîte noire de la voiture et uniquement les informations minimales nécessaires à la facturation sont envoyées à l'assureur. Ces deux aspects sont schématisés sur la Figure 34.

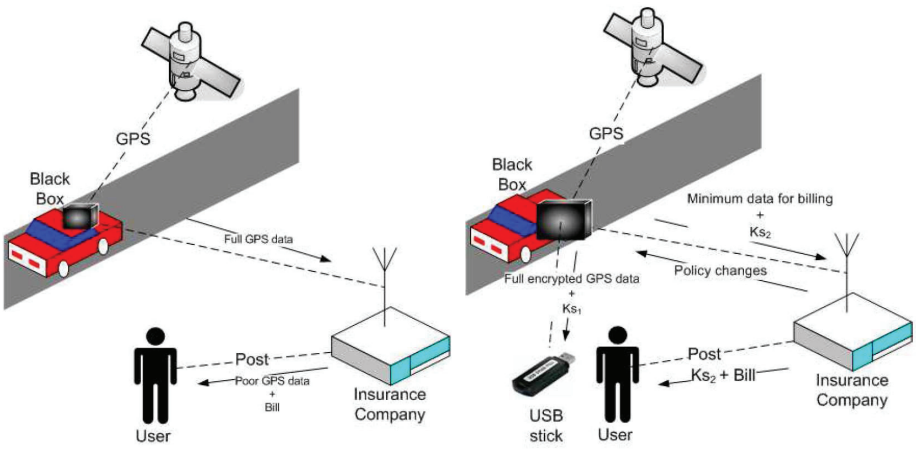


FIGURE 34 – Modèle classique et Nouveau modèle d'assurance PriPAYD

La solution proposée consiste à crypter les données qui sont transmises de façon distincte à l'assuré et à sa compagnie d'assurance. De cette façon l'assuré est le seul à avoir un accès complet à l'ensemble des informations et il peut vérifier que le calcul de sa prime correspond bien à sa conduite. Nous avons ici un point important de la procédure PriPAYD, même si la boite noire appartient à l'assureur, son contenu est la propriété exclusive de l'assuré. La protection de la vie privée des clients en matière d'assurance automobile comportementale est un sujet important. En novembre 2005, la CNIL a déjà refusé la mise en oeuvre par la MAAF d'un traitement automatisé de données à caractère personnel basé sur la géolocalisation des véhicules. La raison de ce

refus était que l'assureur aurait disposé d'informations sur les dépassements de vitesse, ce qui est en contradiction avec l'Article 9 de la directive sur la protection des données selon laquelle une entité privée n'est pas autorisée à utiliser des informations relatives à une infraction.

0.11.3 Projection

Une autre approche américaine est celle proposée par l'institut national d'Oak Ridge[40] avec l'appui de General Motors. Cette étude se focalise sur les accidents mortels impliquant les personnes de plus de 65 ans. Le taux d'accident est exprimé en fonction du kilométrage pour une meilleure appréciation du risque. On commence par observer les taux d'accident de 1983 à 1995 par tranche d'âge sur la Figure 35 et par zone géographique sur la Figure 36.

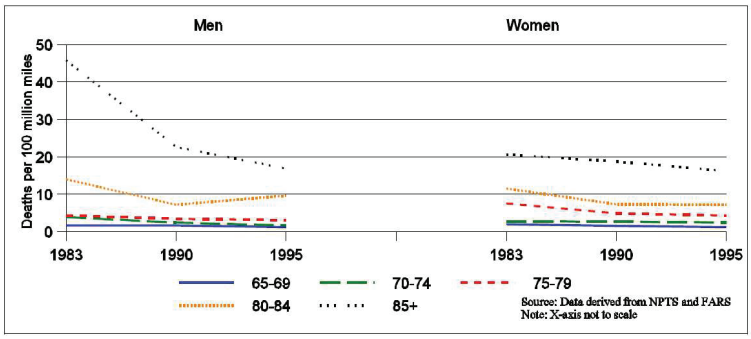


FIGURE 35 – Taux historique d'accidents mortels par âge

Les hommes de plus de 85 ans affichent la plus forte mortalité sur la route dans les années 80, mais aussi la plus forte diminution avec un taux qui passe de plus de 40 morts par 100 millions de kilomètres parcourus en 1983 à moins de 20 morts en 1995. Les femmes ont une sinistralité plus faible en général, mais pas pour la tranche 75-79 ans.

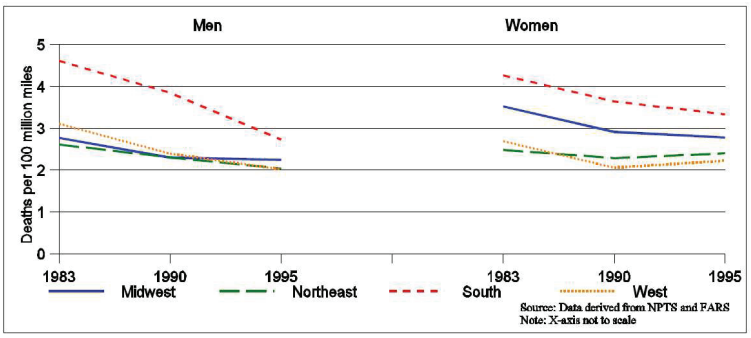


FIGURE 36 – Taux historique d'accidents mortels par régions

La zone géographique a aussi une forte influence sur la sinistralité. Le Sud est largement plus touchés pour les hommes comme pour les femmes. Le Midwest est la seule région où les femmes ont plus d'accidents mortels que les hommes. D'une manière générale, le taux de sinistre a diminué avec le temps. On peut noter que cette décroissance s'accélère chez les hommes après 1990 alors qu'elle ralentit chez les femmes. L'Ouest et le Nord Est sont nettement moins touché par la mortalité sur la route.

L'étude propose dans un second temps une projection à l'horizon 2025 selon trois composantes importantes. D'abord le vieillissement de la population avec une augmentation du nombre de personnes âgées estimée entre 50 % et 150 % d'ici 2025. Ensuite, la prise en compte des déterminants historiques dans la décision de conduire, illustrée par une projection du pourcentage de conducteurs actifs dans la population. Enfin, l'évolution du kilométrage moyen par personne et par an en perpétuelle augmentation pour les plus de 65 ans et ce en particulier pour les femmes.

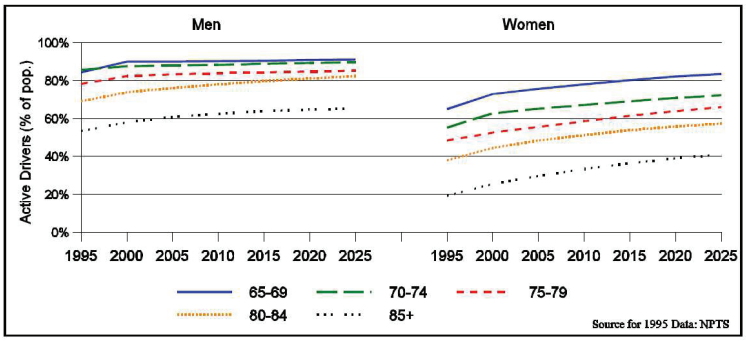


FIGURE 37 – Proportion de conducteurs dans la population

Sur la Figure 37 nous observons l'évolution de la proportion des conducteurs seniors dans la population entre 1995 et 2025. Pour les hommes, après une légère augmentation jusqu'en 2005, le pourcentage de conducteurs actifs stagne jusqu'en 2025, toutes tranches d'âge confondues. En revanche chez les femmes, nous avons de plus en plus de conductrices avec le temps avec une forte augmentation d'environ 20% même chez les plus âgées. Sur la Figure 38 nous pouvons comparer le kilométrage annuel moyen des hommes et des femmes. Les hommes parcourent nettement plus de distance. Et cette tendance va en augmentant dans le futur. Les seniors les plus jeunes (65-69 ans) sont ceux dont le kilométrage croît le plus vite tous sexes confondus. Plus on se rapproche de la tranche 85 ans et plus, plus la croissance est faible.

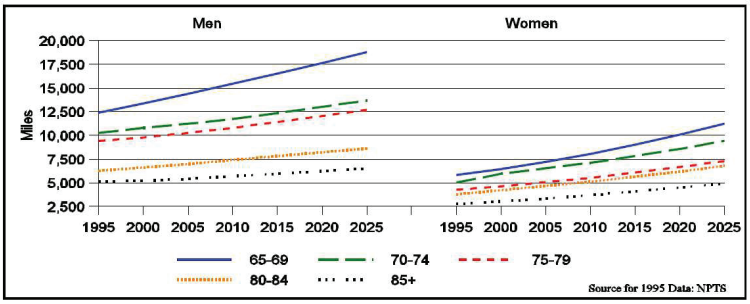


FIGURE 38 – Kilomètres parcourus en moyenne par personnes

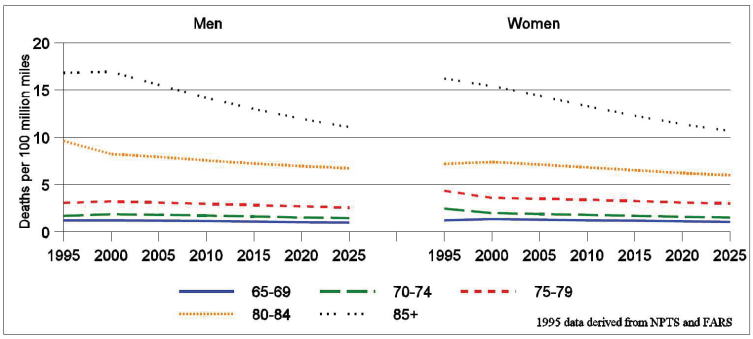


FIGURE 39 – Taux projeté d'accidents mortels par âges

L'étude se termine avec les résultats du modèle sur la Figure 39. Les taux d'accidents mortels par tranches d'âges et pour cent millions de miles parcourues sont projetés jusqu'en 2025. Les courbes sont quasiment constantes pour les moins âgés (de 65 à 79 ans). Ce sont les plus vieux (85+) qui améliorent nettement leur sinistralité avec une diminution d'environ 6 points pour les hommes comme pour les femmes.

Ces différentes approches montrent la complexité des facteurs qui peuvent être à l'origine d'un accident de la route. Elles permettront de mieux cerner notre contribution à la tarification du risque automobile, présentée dans le dernier chapitre de cette thèse.

The wind puffed, and Mr. Pickwick puffed, and the hat rolled over and over as merrily as a lively porpoise in a strong tide : and on it might have rolled, far beyond Mr. Pickwick's reach, had not its course been providentially stopped, just as that gentleman was on the point of resigning it to its fate. [21].

Charles Dickens

Chapitre 1

Index for Predicting Insurance Claims from Wind Storms with an Application in France

Sommaire

Alexandre Mornet Université de Lyon, Université Claude Bernard Lyon 1, Institut de Science Financière et d'Assurances, 50 Avenue Tony Garnier, Lyon F-69007, France and Allianz, Coeur Défense, 82 Esplanade du Général de Gaulle, Courbevoie F-92400, France, alexandre.mornet@allianz.fr
Thomas Opitz Biostatistics and Spatial Processes Unit, National Institute of Agronomic Research, Avignon, France, topitz@paca.inra.fr
Michel Luzi Non-life actuarial affairs former Director, Research ans Development Director at Allianz France, qualified Member of Institute of Actuaries, 132, rue du Président Wilson, Levallois-Perret F-92300, France
Stéphane Loisel Université de Lyon, Université Claude Bernard Lyon 1, Institut de Science Financière et d'Assurances, 50 Avenue Tony Garnier, Lyon F-69007, France, stephane.loisel@univ-lyon1.fr

1.1 Abstract

For insurance companies, wind storms represent a main source of volatility, leading to potentially huge aggregated claim amounts. In this article, we compare different constructions of a storm index allowing us to assess the economic impact of storms on an insurance portfolio by exploiting information from historical wind speed data. Contrary to historical insurance portfolio data, meteorological variables show fewer non-stationarities between years and are easily available with long observation records; hence, they represent a valuable source of additional information for insurers if the relation between observations of claims and wind speeds can be revealed. Since standard correlation measures between raw wind speeds and insurance claims are weak, a storm index focusing on high wind speeds can afford better information. A storm index approach has been applied to yearly aggregated claim amounts in Germany by Klawa and Ulbrich [45] with promising results. Using historical meteorological and insurance data, we assess the consistency of the proposed index constructions with respect to various parameters and weights. Moreover, we are able to place the major insurance events since 1998 on a broader horizon beyond 40 years. Our approach provides a meteorological justification for calculating the return periods of extreme storm-related insurance events whose magnitude has rarely been reached.

Keywords. Storm Index, Wind Speed, Insurance, Extreme Value Theory, Extreme Dependence

1.2 Introduction

For property damage insurance, natural events are the main source of volatility. Historical data covering a 30-year period show that, among these events, storms cause more than half of the costs [26]. This represents a challenging problem for insurers that have to decide on pricing, reinsurance and capital requirements. Classical techniques for the prediction of future claim amounts are based solely on companies' information (portfolios and damage), which leads to some typical problems. The insurers keep detailed historical data only for a short span of around 15 to 20 years. This lack of information makes it more difficult to get reliable results, in particular for the most extreme events that are also the least frequent ones. As we work on historical data, it is necessary to find appropriate normalisations correcting for various heterogeneities like property values, growth of real estate or the spreading rate of storm guarantee.

Managing extreme risks [38] is an issue involving the whole society and not only insurers. When we model extreme events using only damage data ([28], [42]), results show a strong dependence on model hypotheses which are difficult to verify in practice. In our context, adding meteorological data to insurance data is one possible solution to improve the quality of predictions and the fundamental understanding of how storms generate damage. Several authors suggest comparing damage numbers with meteorological data. [76] apply a model using raw meteorological variables to predict the number of claims per day in Norway. [82] seek to predict the number of casualties of storms and typhoons in the USA. For a more accurate assessment of damages, [13] consider the calculation of a damage function for storms and earthquakes, whereas [36] and [19] create an indicator of extreme climate events on a large scale. Other studies have tried to find an empirical link between the wind speeds and the vulnerability to storms in order to calculate a damage function [37]. Approaches similar to ours have been proposed, leading to the construction of storm indices that indicate the major events in terms of insurance costs and that are based on wind speeds as well as geographical and demographic variables. For the German territory, Klawa and Ulbrich [45] obtain good results for yearly aggregated claims with data from the relatively small number of 23 meteorological stations. In [45], [64] and [24], damage is assumed to be proportional to wind speed to a power of 3. This choice is justified by physical reasoning, with the cube of wind speed being proportional to wind energy. Other authors [25, 68] propose to link damage to an exponential of wind speeds and include information about wind direction. Prettenthalher and al. [68] also take into account the vulnerability of a building. For the construction of our index, we operate on a fine spatial and temporal resolution, using daily data from 130 weather stations covering the territory of France. We study constructions based either on powers or on exponential expressions of wind speed. Parameter like the exponent are considered as free, so we can adjust them to maximize the correspondence between the storm index value and damage for the major storms in the last decades.

For several decades, the insurance industry has covered storm damages. The sector has access to historical data structured according to various aspects (chronological,

geographical, costs, frequencies, ...). However, as pointed out above, one must be aware of the limitations and problems of using such data, especially when we deal with the most extreme events.

Our work aims at creating a simple but useful index to assess the economic impact of storms. As we can rely on meteorological data that are linked to storms, we try to find solutions to make use of that kind of information. Meteorological variables such as wind speed present the advantage of being easily available with long observation records over a dense network of weather stations. A major part of observed wind speeds are small or moderately large and cause claim amounts that are negligible compared to damages caused by the most extreme wind speeds. We focus on extreme wind speeds to build a storm index that adds useful information to our research. To ensure the validity of the index we must verify the solidity of the conditions. We test the coherence of the index by comparing it to the meteorological and insurance history. We bring out its sensitivity with respect to its parameters and weighting schemes. Following multivariate Extreme Value Theory (EVT) approach [28], [14], we show that our storm index presents strong tail dependence with actual claim frequency, which is highly desirable for risk management purposes. Moreover, the index approach enables us to analyze the major events on the broadest possible observation period.

Our paper is organized as follows : first, we recall different definitions for a storm. Then, we present the economic issues and we describe the available insurance and meteorological data to try to tackle them. In the following section we explain step by step the construction of the wind index and the storm index. The last part is dedicated to comparisons between claims and different index. From these comparisons, we define the best adapted Formula and we study the index evolution during the most extreme and expensive events for the insurer.

1.3 What is a storm ?

1.3.1 Meteorological point of view

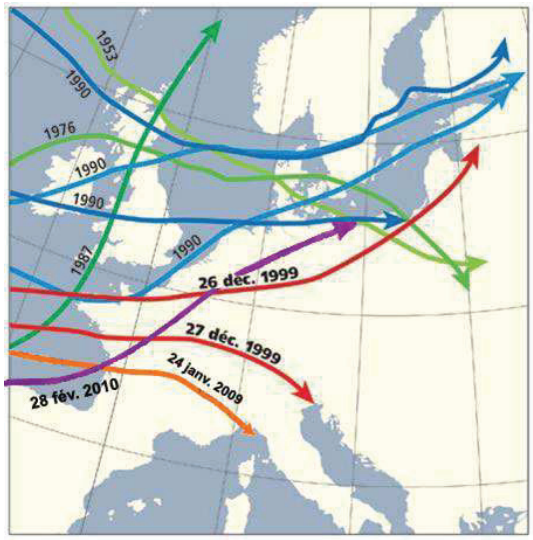
Can we give a single simple definition of a storm event ? A precise description of the phenomenon is hard to formulate. From a meteorological point of view, storms are considered as atmospheric disturbances that involve pressure and wind variations that can bring about risks of property damage. At sea, it means an atmospheric depression that brings about an average 90 kph wind speed, i.e. force 10 or more on the Beaufort scale (from 0 to 12). On land, a storm is a depression that brings about gusts of wind that cause damage and heavy precipitations. Atlantic storms that reach France before most European countries are due to two concomitant causes : the frequent lasting jet stream at high altitudes born above Newfoundland that blows at impressive speeds (200 kph) above Northern Europe and the midday decrease of earth temperature around the globe from the Equator to the North.

For Meteo France a storm is declared when more than 5% of the weather stations record wind speed above 100 kph during 3 consecutive days.

1.3.2 Insurance point of view

From an insurance standpoint, the claims registered in storm guarantee range from damages due to the slightest wind breath to those caused by large-scale phenomena (strong intensity, large affected territorial area, period which could exceed 3 days). For example, in 1999, Lothar hit the Northern half of France, with wind speed above 140 kph, between 25 and 27 December.

Note that the insurance of storms in France follows a regulatory framework . The 25th June 1990 law rules that the storm guarantee is compulsory for all risks that are covered for fire. It entails that all the insured risks are necessarily covered without adverse selection. According to the rules of the market, hail and snow guarantees are associated with storm guarantee.



Storm	Date
Lothar	Dec. 25-27/1999
Martin	Dec. 26-27/1999
Klaus	Jan. 23-25/2009
Xynthia	Feb. 26-01/2010

FIGURE 1.1 – Storm trajectory over France

Generally storm data are available on a daily basis. However a storm over a 550,000 km^2 territory like France can last several days as illustrated on Figure 1.1. The main storms since the 50s are represented. Four of them crossed France : Lothar and Martin in 1999, Klaus in 2009 and Xynthia in 2010. All of them lasted more than one day.

If the major storms take time to cross the country, the disaster declaration can also be more or less inaccurate, especially by absent people at the time of the facts. For these reasons, the treaties of reinsurance plan generally to group in the same event the claims declared for a period from 2 to 3 consecutive days. Lothar and Martin constitute a particular case because these two storms took place in the same period by one day but on different areas. We could then think of grouping them in one cluster, but for the

insurers it consists in two different events. This definition of the disaster is important in particular for the reinsurance which also works by event. In our article these two storms are treated separately.

Combining the meteorological and insurance approaches, we can conclude that a storm event is characterized by the wind intensity, the geographic spread of the damaged area and its duration that may exceed one day.

1.4 Insurance issues

In personal and commercial insurance, wind property damages are covered through TGN (Tempête, Grêle et Neige : Storm, Hail and Snow) guarantee. A study published by AFA (Association Française de l'Assurance : French Insurance Association) [26] in February 2012 shows the volatility of yearly costs related to wind on a national scale. The current costs vary between 70 million and 7 billion euros. Since 1984, storms represent 87.5% of the claims of the TGN guarantee. With a total of about 12 billion, the 2012 cost of damage caused by Lothar and Martin storms (year 1999) is at the same level as the average annual cost for the whole Property coverage. Regarding premiums, the sum of all property damages amounts to 16 billion among which 1.3 billion are dedicated to the TGN guarantee, about 8 %.

In order to compare costs during a long period, normalization is key to the validity of results. We choose here to convert all the costs into 2012 euro value. This stage of data treatments is generally underdeveloped. We nevertheless underline that depending on the used hypotheses, the results vary appreciably. For example, according to the choice of the construction index (FFB or RI[1]), the values 2012 of year 1982 vary by 18%. If we add the progress of the exposures (growth of the housing stock), the values 2012 of year 1999 increase by 21%.

Though some approximations cannot be avoided, we try to be as exhaustive as possible by taking into account the evolution of FFB rate for private individuals and RI rate for business concerns, the relative weight of segments private individuals / business in the portfolio, as well as the spreading rate of storm guarantee that only became compulsory in July 1990 and the growth of real estate in France.

Figure 1.2 refers to the 2012 annualized costs related to storms since 1984 (the figures come from FFSA's report : French Federation of Insurance Companies) . It shows both the volatility and the severity of the costs. We notice that the great majority of the years have values lower than the TGN annual average premium (dotted orange line : 1.3 billion). Only 4 years exceed this level, among which year 1999 which is very appreciably higher. We point out that the burden of the year 1999 represents more than 43% of the whole period costs. The return period of events of the same or even

1. FFB (Fédération Française du Bâtiment : French Building Federation, 3.93% / year) and RI (Risques Industriels : Industrial Risks, 3.38% / year) are two French building discount rates

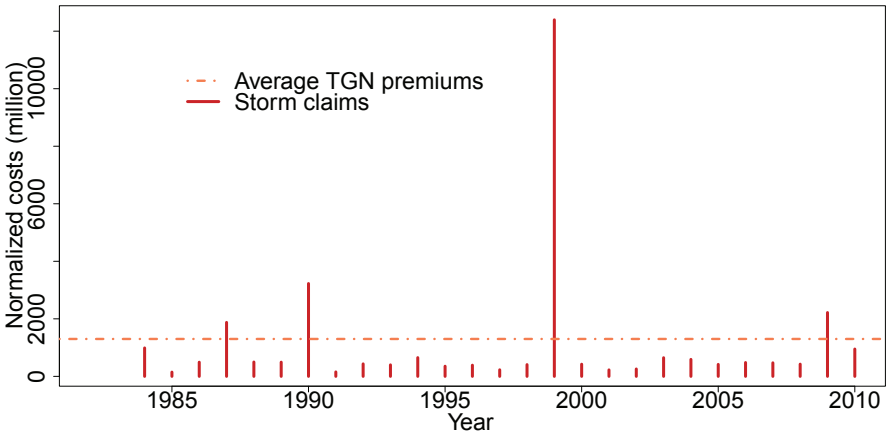


FIGURE 1.2 – Normalized annual costs for storm guarantee between 1984 and 2010

greater importance has considerable influence on results. For instance, if one spreads the impact on Lothar and Martin over 50 or 100 years, the annual TGN cost tends to rise substantially.

We have to differentiate here the notions of annual costs and costs by event. The quality and accuracy of these data are different. Except for the major storms, the statistics are not followed by event at Market level. Consequently, the use of data from an insurance company brings additional details.

1.5 Insurance data

1.5.1 1998-2012

Database description

We use the Allianz France data base. This portfolio contains all the individual claims between 1998 and 2012. It is a base of 520,000 claims, among which 310,000 for personal lines and the rest for commercial lines. For every claim, we know the amount, the date of occurrence and the place of the claim (department [2], even municipality).

Comparaison with the market

If the use of the data of a single company can skew the results, we notice that the annual loads are strongly correlated to those of the market. Also, if we study the geographical distribution of the portfolio, we notice that the distribution is close to that of the French population. Even if the results are not totally identical everywhere (parts between 4% and 14% according to departments, for an average value close to 9%), they

2. In metropolitan France, the territory is subdivided into 95 departments which correspond to administrative areas.

are not very different from those of the market.

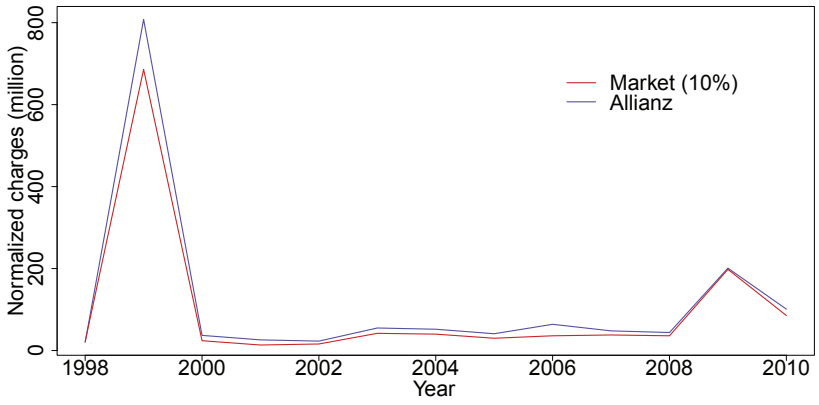


FIGURE 1.3 – Comparison between Allianz annual costs and those of the market

Figure 1.3 compares Allianz' wind property damage costs (in blue) with 10% of the market (in red). It reports that the Allianz' loads are very highly related (0.99 %) to those of the market.

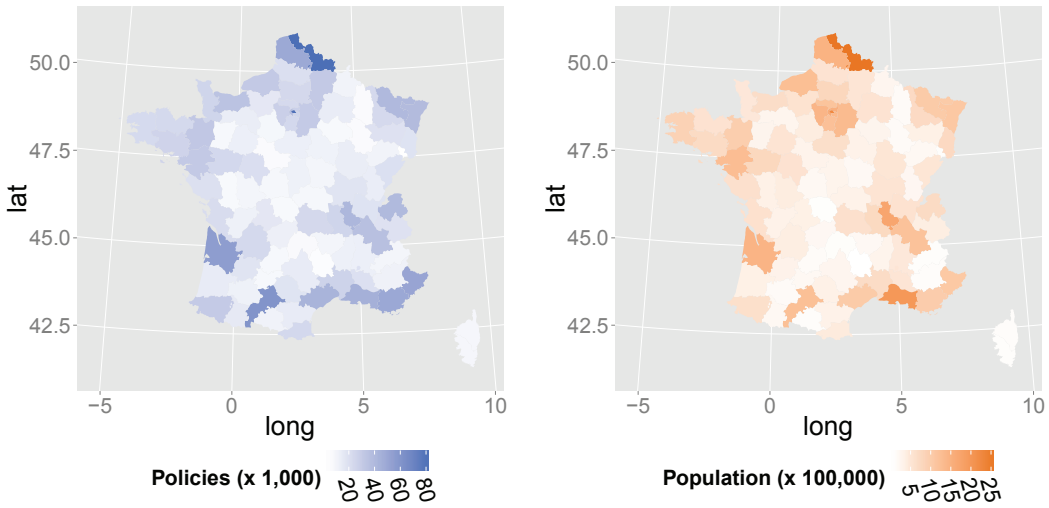


FIGURE 1.4 – Distribution of Allianz portfolio and French population in 2009

Figure 1.4 shows the distribution of the number of contracts in each department in Allianz portfolio and the French population. On the left map we notice great geographical differences in the representation of policyholders. For instance, there are less than 5,000 contracts in some departments, whereas there are more than 50,000 contracts in some other ones. However, the comparison with the second map makes it possible to

relativize differences which are generally in proportion with population. In fact, the most populated areas are also those where we find the greatest number of contracts, and vice-versa.

Thus, it is possible to rely on Allianz data for a study of damage on a national level.

Database analysis

The distribution of individual claims brings an essential information about the scale and the nature of the damages caused by storms. Table I shows the various quantiles as well as the average and the maximum of the 430,000 positive individual claims [3] registered by Allianz during the period from 1998 to 2012. The values are normalized in euros 2012. This distribution is very uneven with a large number of small claims and only around thirty exceptional disasters exceeding one million euros. The average individual claim amount approximately 4,600 euros [4]. Half of the claims are lower than 1,700 euros and 99% of them are lower than 45,000 euros. Yet, the maximum claim takes a much higher value by reaching more than 5 million euros.

In this portfolio, the personal lines represent 260,000 positive claims out a total of 430,000. By looking closer at the values, we notice that the personal lines and the commercial lines have not the same claim distribution. The damages sustained by professionals are superior in every point than those of private individuals. The average claim is 2.3 times higher for the professionals and the difference increases with a 5 times higher maximum claim.

Quantiles	2012 values	commercial lines	personal lines
Min.	1	1	1
Qu 0.25	677	749	640
Qu 0.5	1 667	2 005	1 542
Moy.	4 581	7 032	2 993
Qu 0.75	3 879	5 062	3 279
Qu 0.9	7 865	11 983	6 263
Qu 0.99	45 184	82 682	24 050
Qu 0.999	244 621	455 144	73 982
Max.	> 5 M	> 5 M	> 1 M

TABLE I – Distribution of individual claims (1998-2012)

The period from 1998 to 2012 spans 5479 days. Only 958 days (17.5%) are without claim. But less than half of the days record more than 5 claims. Less than 1% of the days record more than 500 claims. On the most damaged day the number of claims

3. We choose to retain only the positive claims so that our results are not influenced by the non positive claims (more than 16%) which can be bound to cases without continuation, amounts lower than the franchise or other unknown reasons.

4. The average cost by retaining all claims (positive and non positive) would rather be 3,500 euros.

reached 200,000. The distribution in numbers and costs is quite uneven.

The results presented below focus on all the days when Allianz recorded at least one claim with a strictly positive value.

Quantiles	2012 values	Number
Min.	3	1
Qu 0.25	1 921	2
Qu 0.5	6 557	5
Moy.	434 800	116
Qu 0.75	25 630	12
Qu 0.9	103 160	39
Qu 0.99	1 708 116	494
Qu 0.999	35 007 641	14 056
Max.	915 000 000	196 000

TABLE II – Distribution of daily damages (1998-2012)

Table II shows the distribution of daily damage over the 1998 - 2012 periods. Costs are normalized in euros 2012. In fact, generally, most of the stricken days bring relatively light costs. In one quarter of the days, costs do not exceed 2000 euros and in one half, they are under 7000 euros. Only 10% of the days are over 100,000 euros. Yet, the average claim reaches over 430,000 euros. Daily costs can literally rocket during exceptional events such as Lothar and Martin in 1999 or Xynthia in 2010. We then get 78 days over 1 million euros, 10 days over 10 million euros, and a maximum 915 million euros (Dec. 26th, 1999). If we consider the number of events, the result is nearly identical : 90% of impacted days register fewer than 40 claims and half of the days register fewer than 5 claims. The parts under 10,000 euros represent 2/3 of the whole claims. The top part (over 10 million) seems to be very unfrequent.

Thus we generally deal with damage due to the wind but which are too localized to be called storms. According to the first results it brings to study the data from the standpoint of events. Once the data will be grouped in events, we obtain approximately 130 significant events during the period. Table III shows the distribution of the main storms that crossed France since 1998. The 4 most important ones are named and the others are grouped in the last column of the Table. Claims recorded during storms Lothar and Martin in December, 1999 are relatively higher at any point than those of more recent events like Klaus or Xynthia. The average claim varies between 6,000 euros for Lothar and 2,700 euros for Xynthia. At first sight the individual claims of storms having caused globally most damages are higher than the others.

From these data, it is impossible to provide a detailed explanation for these variations. However we can bring background hypotheses. The cost of the building repairs can increase during a strong demand. The normalization choices, given that two main

Quantiles	Lothar	Martin	Klaus	Xynthia	Others
Min.	1	1	1	1	1
Qu 0.25	820	1 000	646	476	677
Qu 0.5	2 141	2 318	1 485	1 075	1 667
Moy.	5 968	5 190	3 388	2 707	4 549
Qu 0.75	4 850	4 940	2 717	2 212	3 878
Qu 0.9	9 884	9 124	6 245	5 070	7 865
Qu 0.99	62 748	46 672	31 830	25 746	45 164
Qu 0.999	355 300	230 379	143 271	111 763	243 642
Max.	> 5 M	> 2 M	> 2 M	> 1 M	> 2 M

TABLE III – Distribution of main event damages (1998-2012)

events took place in the beginning of period, the size and the location of the damaged area are not known and could nevertheless strongly influence the results.

1.5.2 1970-2012

If it is not possible to get back older historical data with the same level of detail, it is however possible to have information relative to the major events. This information is generally available in the departments responsible for reinsurance.

These data present certain advantages. It is possible to identify the dates when significant events occurred. We also have an information about the level of the damages supported for every event, what allows us to have a certain hierarchy. With these data which could go back up on several decades, it is possible to specify better the return periods of rather rare events.

On the other hand, it is not possible to work on a detailed level, on an individual claim level or on a department level. Besides, considering the age of information held or the length of time for which it may be kept, the problems of updating are even more acute. Very often, the studies use only an updating on the amounts of the claims, based on a building index. In fact, the subject is more complex and the gaps from methods can lead to very different results.

Table IV presents the normalized cost of storms since the seventies (according to Luzi's work [51]). Therefore, we propose two kinds of approaches : **Act IC** corresponds to an inflation only based on costs, **Act IG** corresponds to an inflation that takes into account costs, extensions of parks and guarantee spreading.

Thanks to this data recovery, we have about twenty major events over the period from 1970 to 2012, while we limited ourselves to 7 events over the period from 1998 to 2012.

The normalization choice modifies appreciably the results, especially those from the

Date	Act IC	Act IG	Events
12/25/1999	85	100	Lothar
12/27/1999	43	51	Martin
11/07/1982	10	25	Nov 82
01/23/2009	19	20	Klaus
02/03/1990	14	19	Herta
10/15/1987	6	10	87J
11/30/1976	4	10	Nov 76
07/18/1983	4	10	Jul 83
01/25/1990	6	9	Daria
02/27/2010	7	8	Xynthia
02/26/1990	5	7	Vivian
07/11/1984	3	7	Jul 84
12/15/1990	5	7	Dec 90
08/18/1971	2	6	
02/11/1972	2	6	
01/15/1992	3	4	Jan 92
07/27/2005	3	3	Jul 05
12/17/2004	3	3	Dec 04
05/25/2009	3	3	
07/15/1993	2	3	

TABLE IV – List of major storms since the 70s

oldest exercises. Even the ranking of the events can be impacted. Throughout this paper, we use the **Act IG** index which offers the advantage to take into account the inflation of costs, the extensions of the building parks and the guarantee spreading.

In spite of the care taken to value the claims, these data will of course never reach the level of precision of a purely physical measure.

1.6 Meteorological data

1.6.1 Stations

The use of meteorological data is quite natural in the context of a storm study in order to complete insufficient insurance information. Those data are available on a individual day basis and with a relatively long history (since 1973). We use the National Climatic Data Center (NCDC) which gets information from Météo France for France. It is an American website which offers a free and unlimited access. From time to time, we noticed some differences with data produced by Météo France.

From 1973, it is possible to have recordings on numerous stations. Before this date, the information is poorer. From 1998, weather station network allows us to obtain at least one source of information for each department.

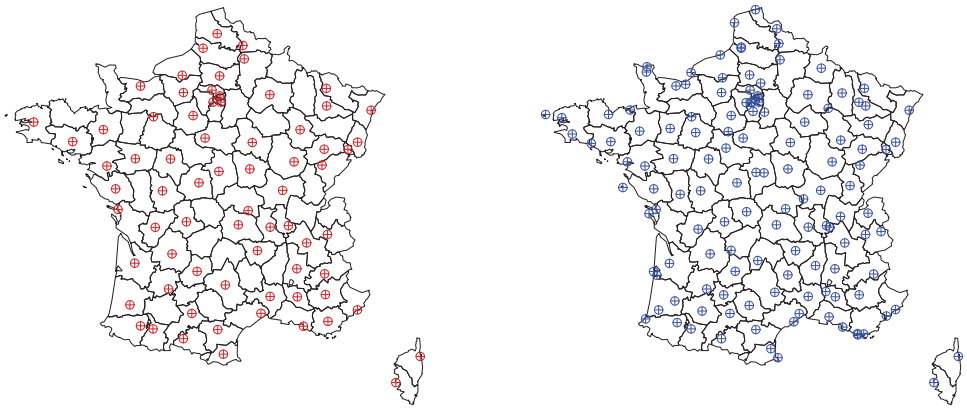


FIGURE 1.5 – Location of active stations : 1973 - 1998 (red) and 1998 - 2012 (blue)

We work with on average 90 stations between 1973 and 1998, then up to 130 stations between 1998 and 2012 (Figure 1.5). If every station is very precisely located, the claims can occur everywhere in the department. The distance between the station and certain buildings can be furthermore of 100 km.

1.6.2 Wind speed

On the National Climatic Data Center (NCDC) we have collected different kind of information about the wind speed : average wind speed, maximum sustained wind speed and maximum wind gust. The air flows generally irregularly pulling a strong variability of the wind in direction and in force. This is why the meteorologists measure the instantaneous wind which varies ceaselessly and the average wind calculated over a period of 10 minutes. When the instantaneous wind speed exceeds that of the average wind of more than 10 knots (18 kph) the meteorologists speak about gust.

We convert all data into kph. For every station, we establish a distribution of max wind speed per day. It leads us to highlight very different situations from one station to the other. In order to account for these differences, we show in Table V and Figure 1.6 how speeds are distributed in the towns of Montpellier (from department 34) and Charleville-Mézières (from department 08).

Sometimes they are over 200 kph in some stations, while others hardly reach 100 kph. In Montpellier the highest recorded speed is 183.2 kph, 10% of the recordings are over 82 kph and 1% over 111.2 kph. Though it is relatively protected by the nearly hills, the station faces Mistral and Tramontane winds. On the contrary, wind speeds in the Charleville-Mézières stations are far lower. The peak speed is only 104 kph and only 1% of the recordings are over 68.4 kph. Consequently an unusual wind speed will

Station	Min.	Qu 0.25	Qu 0.5	Moy.	Qu 0.75	Qu 0.9	Qu 0.99	Max.	Var.
(34)	6.80	35.60	50.40	53.33	64.40	82.4	111.2	183.20	395.04
(08)	6.80	28.80	35.60	34.76	43.20	50.4	68.4	104.00	169.29

TABLE V – Daily maximum wind speed in Montpellier and Charleville Mézières between 1973 and 2012

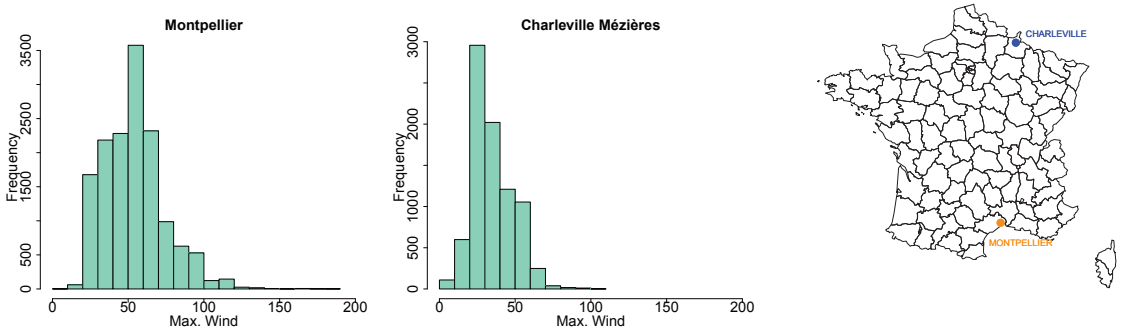


FIGURE 1.6 – Daily maximum wind speed in Montpellier and Charleville Mézières between 1973 and 2012

not always be a sufficient criterion. It will also be necessary to take into account the scale of the phenomenon, and not to overestimate the individual overtakings (as we shall see it in the Subsection 1.8.4, page 98).

More generally, these differences between the various zones can be observed all over the French territory. Figure 1.7 shows the average wind speeds and the 99% speed quantile in each of the 95 departments. Generally speaking, coasts record wind average higher than in the center of the country. Thanks to the 99% quantile, the highest speed in each department can be observed. Then the discrepancies are very much increased with a 60 kph minimum and a 160 kph one. This 100 kph difference shows the lack of homogeneity in the distribution of wind speeds in France. So it seems that some departments have developed a behavior of adaptation to the wind. It may have an influence on the local ways of building and of the reactions of the insured about storm risks. For instance, house roofs will be stronger if they are often exposed to gusts over 100 kph. Our study will take these disparities into account by observing wind-speeds according their local distribution.

In our paper, we consider meteorological data as more stable and stationary over time than insurance data. All studies that we are aware of (reports of FFSA and of Météo France) show that no significant trend could be detected since 1950, which corresponds to the period considered in our study. Nonetheless, we heed the consequences of a changing climate that would alter this stationarity. We cannot claim a reliable prediction for horizons of 30, 50 or 100 years. Fortunately, our prediction horizon in

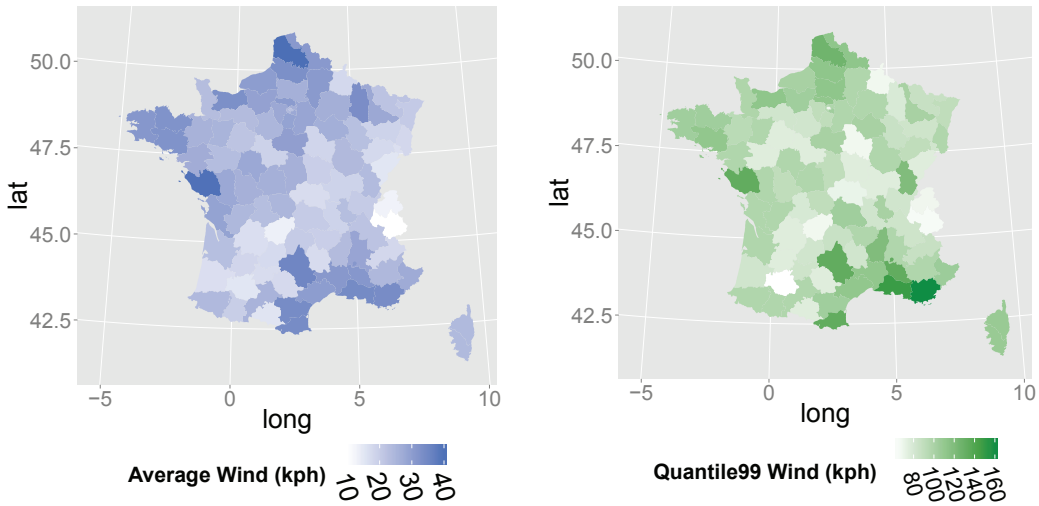


FIGURE 1.7 – Average wind speed and 99%-quantile in France between 1973 and 2012

the insurance context is usually only 1 to 2 years.

1.7 Index construction

We aim at constructing an index to capture the intensity of storms and to obtain a clear relation with the number of claims or with the aggregated claims. A wind index is based on wind speed data from a network of weather stations and can be defined on a local level for each of the 95 departments (*local wind index*). By aggregation of all departments, we obtain a *storm index* on the national level. For our study, we define storm events as being characterized by an observation of strong wind speeds or high claims in France, spanning a period of 1 to 3 consecutive days. A large number of days without significant effects are taken out of consideration by this approach.

1.7.1 Local wind index

Several formula have been proposed in the literature, some of them including parameters that need to be calibrated. Based on a series of daily maximum wind speed measurements $w^d(s)$ for a sequence of days $d \in D$ at station s, we here propose the following definition :

$$I_w^d(s) = \left([w^d(s) - w_q(s)]_+\right)^\alpha, \tag{1.1}$$

where w_q is the q%-quantile of $w_d(s)$ for $d \in D$. In the following section, different quantile values are tested. By adding the exponent α to this index, we increase or decrease the influence of the most extreme wind speeds. Since we will compare the wind index I_w^d to claims observed at the department level, we have to define $w^d(s)$ accordingly. If several weather stations exist for a departement, we define $w^d(s)$ as the maximum value over these stations. We remark that threshold exceedances $w^d(s) - w_q(s)$ as in (2.5) are

used in extreme value theory in the context of the peaks-over-threshold approach (see Chapter 4 in [15]), such that our approach is natural for capturing extreme behavior.

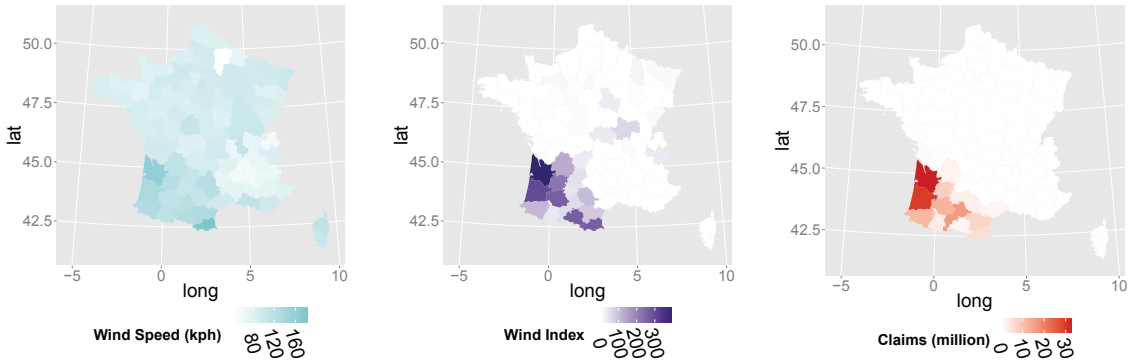


FIGURE 1.8 – 3 illustrations of the KLAUS Storm

The first test is a comparison, during a storm, between the maximum wind speeds and the wind index in each department and over 2 or 3 days, to try and find some similarity with the damages registered by Allianz. Figure 1.8 shows the maximum wind speeds, the wind index and the normalized costs of damages (in million euros) during the Klaus storm between Jan 23rd and 25th 2009. Klaus is a severe windstorm that struck the South West of France. From the wind speeds, the most impacted zone can be delimited but with perfectible accuracy. When you study the 2nd map with the wind index in every department, you can see that it is better connected with claims, so that zones that are little impacted by the storm can be excluded. With this graphic approach, the correlation between winds and their frequency is more obvious.

Other formula have been used in the literature. [45] use a cubic index based on a ratio between the maximum wind speed of the day and a quantile :

$$CI_w^d(s) = \left(\frac{w^d(s)}{w_{98}(s)} - 1 \right)_+^3, \tag{1.2}$$

which essentially is a special case of (2.5) with $q = 98$, $\alpha = 3$ and a multiplicative rescaling such that $CI_w^d(s) = w_{98}(s)^{-3} I_w^d(s)$.

1.7.2 Storm index

We have already mentioned that strong wind speeds are an essential cause of damages, but they are not enough to define a storm. From a geographical point of view, every station corresponds to a different exposure area. Several stations cover very populated departments with numerous insured risks. Others, on the contrary, cover more desert zones. It is thus necessary to balance the results of every station by a variable corresponding at best to the local exposure. Theoretically, it would be necessary to know the data that is specific to each event. Failing that, we can, at least, use the

geographical distribution of the population [5].

It is also necessary to notice that in certain cases, several storms affect the country on the same day. It was particularly the case on December 26th, 1999 with Lothar and Martin. It is then necessary to distinguish departments touched by the different events. It is not always easy. So, the first step of the construction of our storm index consist in multiplying the wind index on date d, $I_w^d(s)$ and the number of risks balanced by the number of contracts $C(s)$ in Allianz portofolio.

$$I_w^d(s) \times C(s)$$

Then the size of the damaged area must be considered. So we apply a geographic aggregation of all station weighed wind speeds. We obtain a first Formula of the storm index on date d, I_S^d.

$$I_S^d = \sum_k I_w^d(s) \times C(s)$$

The definition of this index has been established empirically and could be improved. However, it is based on established standards of damage-assessment. Here again, the choice of variables and stations will be adapted according to insurance data.

From a temporal point of view, the lasting days consideration allows to imagine several solutions. At this stage of the construction of the index we have to choose between two methods of aggregation of the values, the sum or the maximum. We decide to keep the sum by considering that the trajectory of the storm can bring about strong values of wind on several days in different places. This choice was then supported by the results obtained in term of correlation. In that way, we take into account both the course of the storm that can stretch over a few days and the frequent delays in damage claims. We define I_S as

$$I_S = \sum_{d \in E} \sum_s \frac{I_w^d(s) \times C(s)}{N_a^d}, \tag{1.3}$$

where in each station $I_w^d(s)$ is the wind index on date d and the number of risks is balanced by the number of contracts $C(s)$. We also take into account the size of the damaged area (geographic aggregation), the duration of the storm event E [6] and the number of active stations on date d, N_a^d.

For reasons of confidentiality, the results are established by assigning the value 100 to the most important event. Thus we point out the most striking events in the period according to our index.

One advantage of the storm index is its relatively long period of observation. In fact, we can use reports of wind speeds since 1973 in one station at least in each department.

5. It would be interesting to know also the number of all the buildings classified by categories (individual houses, buildings, etc.) but we did not obtain this information.

6. For each event $E \subset D$ we operate a daily aggregation, for example, during the storm Klaus : $E = \{01/23/09, 01/24/09, 01/25/09\}$.

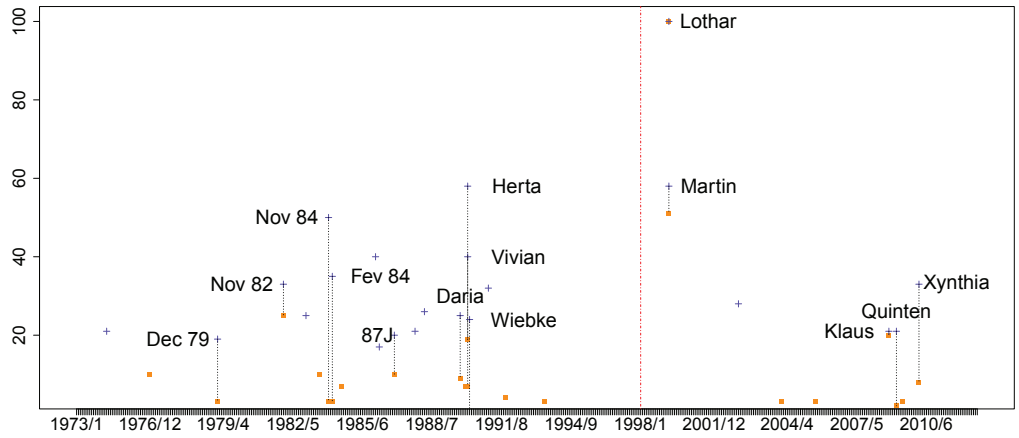


FIGURE 1.9 – Storm index in connection with population between 1970 and 2012

This first result is encouraging (Figure 1.9). The blue cross represent the storm index values and the orange squares represent insurer classification. All the data picked out by the index are those of known storms (Meteo France history). Lothar storm in December 1999 is remarkable with an index that is almost twice higher than Martin's. As we expected, this broadening of the observation period makes it possible to remark other storms in the seventies, eighties, and nineties. Yet the ordering of events in decreasing values does not express exactly the impact they may have had in terms of damage and insurers' costs. For instance, Xynthia storm is clearly above Klaus, though the latter cost twice as much. Herta storm gets more or less the same index as Martin though it was not so important. At this step, the link between our index and insurance results is not obvious. We find unmatched points in both categories of information. We will see (Figure 1.22 and 1.25) to what extent optimization of the used factors and the variable choices improve correlations.

1.8 Comparisons

1.8.1 Objectives

We compare results stemming from meteorological data with observed results from insurance. Our purpose is to study the correlations between the diverse parameters, to measure the relevance of the correlations, to select the most appropriate parameters. If these comparisons are relatively classic, it is necessary to underline that none of these results is perfect.

1.8.2 Some issues

Concerning claims, the issues come from the updating and from the geographical representativeness of a limited portfolio. We work here on the Allianz'database combining the personal and the global lines at the department level. We use the claims values normalized and balanced by sum of premiums (loss ratio). On the other hand, the numbers of claims are not normalized but balanced by the number of policies to obtain claim frequencies. After the treatments of updating and by referring to portfolios insured every year, we can hold more normalized values allowing to make comparable the results between departments. So the 3 types of insurance information about claims are : the normalized loss ratio, the average normalized claim cost and the frequency of claims.

Concerning the wind, the issues come from the low granularity of stations and relevance of the information (daily average speed, maximum speed, gust). We can have several stations on the same department and some departments without station. For departments with several stations, the recordings can be very different from a place to the other on the same day. Generally, the average speed of a station tends to smooth observations and on the contrary the only maximum gust tends to exaggerate the phenomena [74]. We tried to determine an intermediate solution by coupling the data of several stations for every department. We thus retain at least 3 stations for every department. Every station is balanced according to its closeness with the most urbanized zones. In case the number of stations is insufficient, we also use the data of the nearby department stations. We test here 4 types of wind speed measures : the weighted department stations average of daily average wind speed, the maximum on department stations of daily average wind speed, the weighted department stations average of daily maximum wind speed and the maximum on department stations of daily maximum wind speed.

EVT intuition suggests to use maximum of maxima as averaging out extremes is usually a bad idea. However, for our problem, we shall see that the best solution is the third one, as the storm causes many claims if it is strong and widespread. Our comparison concerns events. Yet the scale of day is not always relevant to realize the comparisons. Even, at the departmental level, the disasters relative to the same event

can be declared on several days. For the wind, a storm can take several days to cross France, or stagnate more or less time. We cannot focus on comparisons only. We thus grouped certain days between them to correspond better to the events.

1.8.3 Wind Speed and claims at department level

This step is essential to determine the most suited information at the station level. We have detailed information over 15 years (5,480 days) and on 95 departments (more than 500,000 references). If the volume is very important, we also notice that a very big part of these days are without wind and without claim. We thus have to limit our works in the significant days for our study. Several solutions were tested and finally, we choose to select the days which correspond to **the upper 3% of wind speed or the upper 3% of claims**. By not knowing the best criterion for every element, the selection is the widest for every parameter. When we combine the selected days it represents **around 35,000 data**.

Optimization of the claim criteria

Table VI shows Pearson's correlations between the wind speed and the values and the frequency of claims. We are aware that linear correlation is not always the best tool to analyse correlation. We use it at this stage to get a rough idea of the quality of an index. Later on, we carry out a more in depth analysis, focusing in particular on correlations of extremes. For raw wind speed, we observe a positive trend with total cost as well as with frequency. Nevertheless, the correlation does not exceed 17%. In the case of average costs the correlation is negative and weak (about 5%). By comparing the results of the global and personal lines, we can consider that the claim criterion does not bring fundamental difference. We thus focus for the rest of the study on the global portfolio.

Lines	loss ratio (2012 euros)	average cost (2012 euros)	frequency
Global	0.137	-0.062	0.164
Personal	0.139	-0.047	0.161

TABLE VI – Correlation with average of maximum wind speed

On Figure 1.10, the strongest claims are located in the upper right corner concerning loss ratio and frequency. We notice a certain progress of claim rates according to wind speed. Nevertheless, we observe many points along the x-axis, which is indicative of low loads or a low frequency of claims though the wind speed was high. We notice that the claim loss ratio and frequency remain very low under a certain wind threshold.
For average costs, the main claims seem to take place during days with a low wind speed. No clear relation appears with wind speed. The average costs are not thus a good information source for our work.
As the results of frequencies and loss ratio are rather close, we shall favor the frequencies as the claim criterion to optimize the criteria of the wind.

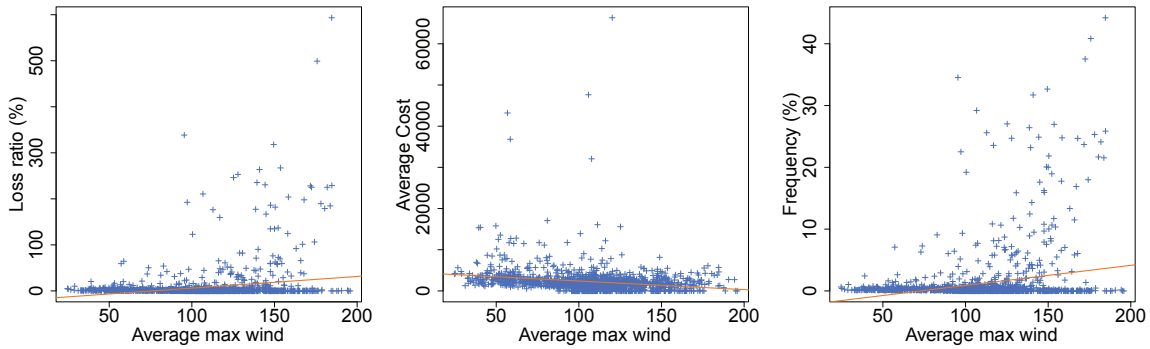


FIGURE 1.10 – Correlation between claims and average of maximum wind speed

Optimization of the wind criteria

In a general way, the results obtained on the basis of average winds are unsatisfactory. We obtain better results from the maximum wind speed.

Average speed	Average of maximum speed	Maximum of maximum speed
0.090	0.164	0.146

TABLE VII – Correlation with 3 types of estimation of the wind speed

On Table VII we compare the correlation obtained with 3 different ways to estimate the wind speed in a department. Among the various options to express the wind speed in every department, the best solution is to consider a weighted average of the maximal speeds (best correlation with more than 16%). On the left-hand graph (1.11) : average of winds, the highest claim frequencies are gathered in the central part and not to the right with the fastest winds. The right-hand graph corresponding at maximum speed on all the weather stations of the department improves the relation wind/claim but remain unsatisfactory. The middle graph corresponding to the average of the maximal speeds balanced on the department offers the best distribution of points.

Several graphs show a no loss ratio for high speeds of wind. Would there be an effect concerning the habituation in the wind ? To test this phenomenon, we looked for Formulae allowing to take into account this effect. Several solutions were envisaged. First, we try a difference between wind speed and a quantile : $(w - w_q)^\alpha$, then, we compare the obtained results with a ratio between wind speed and a quantile : $\left(\frac{w}{w_q}\right)^\alpha$. As we use quantiles as thresholds the local reference is established from a rather rarely reached speed in the department. We can then test if the results are sensitive to the different threshold (5%, 3%, 2% ...).

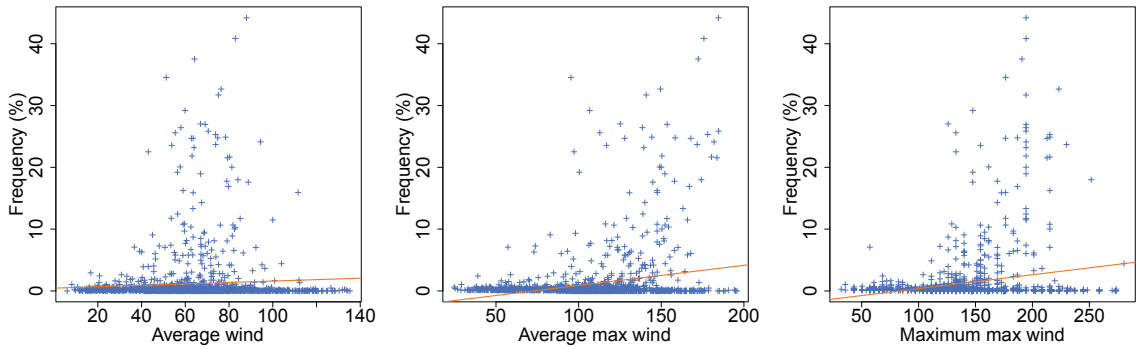


FIGURE 1.11 – Correlation between claims and 3 types of estimation of the wind speed

Furthermore, we are also going to show that working under the shape of moving averages allows to present results of a lesser volatility.

Moving average principle

A moving average is commonly used with time series data to smooth out short-term fluctuations and highlight longer-term trends or cycles. The threshold between short-term and long-term depends on the application, and the parameters of the moving average will be set accordingly. From the ratio index, we can then observe the moving averages to underline the improvements brought by the weighted average of the maximal wind speed. Here we chose to calculate a moving average centered on 12 events.

Denote by $(x_{(s)}, y_{(s)})$ the points where the first coordinate is ranked from the smallest to the largest one and $y_{(s)}$ is associated to $x_{(s)}$. The red points are obtained thanks to a moving averaging principle with the Formula :

$$\left(x_{(s)}, \frac{1}{25} \sum_{i=k-12}^{k+12} y_{(s)} \right). \tag{1.4}$$

On Figure 1.12 we can compare the scattered plot from the weighted average of maximum wind speed (blue dots), the ratio wind index (green dots) and the corresponding moving average (red circles) calculated from the relation with claim frequencies. This presentation shows clearly the sensitive improvement brought by introduction of the notion of habituation in the wind. On the right-hand graph which represents only the averages, we can distinguish three kinds of data. The first part groups a very large number of the observations, with moderated wind speeds and low levels of claims. The intermediate part (about 1400 points) contains some stronger wind speeds and higher claim rates. The top part, concerning not much case (about 65 points), presents a stronger accentuation of the loss ratio with regard to the increase of the wind speed.

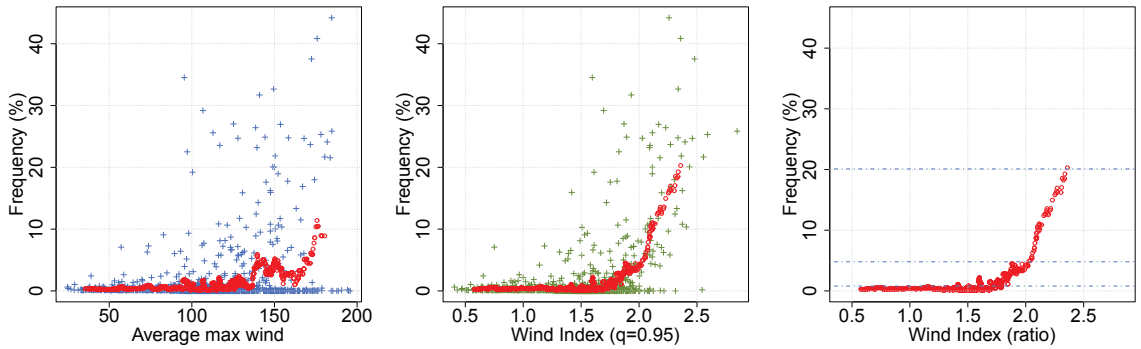


FIGURE 1.12 – Moving average between claim frequency and wind ratio

From these results, we choose to concentrate on **the most striking 1500 events** of the period which constitutes the great majority of important claims. For the wind speeds we shall use the weighted average of the maximums. For the claims we shall select the frequencies and loss ratio, and we shall abandon the average costs which give poor results. It is now necessary to study several subjects. Which function presents the best adjustment between the loss ratio and the wind ? Does the use of an exponent reduce the spread ? What extrapolation can we imagine (at the most, frequency cannot exceed 100%) ? Do the extreme values (strong loss ratio with low wind speed or worthless loss ratio with high wind speed) present common causes (type of department, habituation in the wind...) ?

Wind index Formulae

Now, we test various Formulae of the wind index at the level of the department and during the 1,500 most important events of the period 1998-2012 from insurance and meteorological points of view. We begin with a difference between the maximum wind speed w and a quantile w_q (like in our wind index 2.5) : $(w - w_q)^\alpha$. We test quantiles between 95% and 99.5%.

On Figure 1.13, we present the scattered plot (purple dots) and the moving average (red circles) calculated between the claim frequency and the difference index. From left to right, the quantiles raise from 95% to 99% and 99.5%. For all parameter choices the use of an index improves the results with regard to raw wind speeds. The distribution of big claims agrees better with strong values of the wind index. However visually the three graphs are rather similar. This shows that the choice of the quantile has no strong influence on the quality of the results.

Tables VIII shows the different correlations between claims and the difference index. We try various upper quantiles (from 95% to 99.5%) as threshold of wind speed. We also test values between 1 and 3 for the exponent α. The best link is obtained with q_{99} and $\alpha = 3$ with a correlation over 63%. It is necessary to note that the Pearson

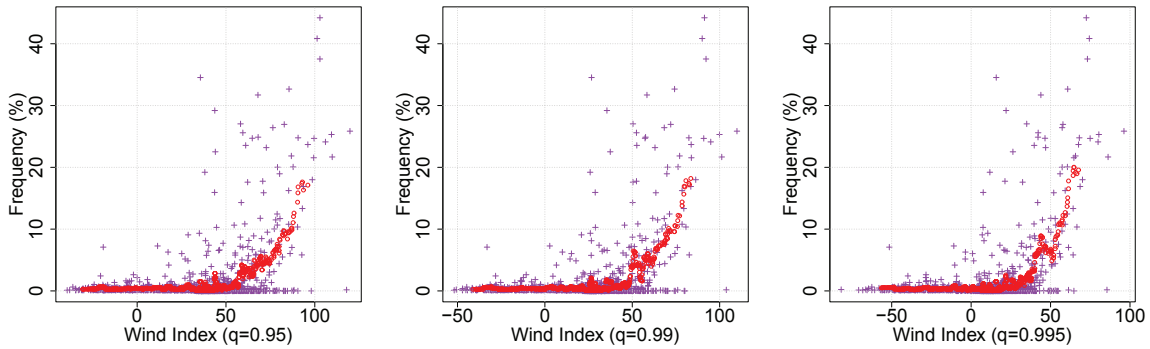


FIGURE 1.13 – Correlation between claim frequency and wind index (difference)

α	q_{95}	q_{97}	q_{98}	q_{99}	$q_{99.5}$
1	0.349	0.358	0.362	0.362	0.375
2	0.545	0.581	0.600	0.600	0.526
3	0.597	0.622	0.633	0.633	0.569

TABLE VIII – Correlation between wind index (difference) and claim frequency

correlation values are linked to the data scale. Consequently, the increase of the correlation when $\alpha \geq 2$ could be partially due to the strong values taken by the index. We thus have only weak differences depending on quantile choices. We can just notice that the correlation increases when the exponent increases and that the most appropriate exponent seems to be $\alpha = 3$. For other variations no trend is clearly recognizable. Nevertheless, if we look all the values of the Table, changes between quantiles are not yet significant. These differences are not sufficient to determine clearly the best Formula, in particular considering the other sources of uncertainties. Furthermore, even if the correlations are higher with a cubic index than with the raw wind speeds, they remain around 50%.

We can conclude from these various tests that the only notable improvement comes from the consideration of the adaptation to the specific wind speeds in every department via the use of a quantile. The correlation seems also to increase with the exponent. Other adjustments and parameterizations remain marginal considering all the inaccuracies which accompany at the same time the meteorological measures and the insurance results.

Spearman's correlation and wind

The results presented so far are those of Pearson's classical correlation. We use here the same approach as previously but by comparing this time not anymore the values but the ranks of the various variables. The obtained correlation is called Spearman's correlation and graphs correspond to the empirical copulas. The empirical copula was

defined in Deheuvels [18] as

$$C_n(u) = F_n\left(F_{n1}^{-1}(u_1), F_{n2}^{-1}(u_2)\right)$$

, where F_n is the empirical joint cumulative distribution function of the n-sample and F_{nj}^{-1} is the marginal quantile function of the jth coordinate sample. The empirical copula is invariant under monotone increasing transformations of the marginals, so it depends on the data only through the ranks.

As previously, we begin with raw wind speeds. In each department we consider the average of weighted maximum wind speed (the most appropriate measure from previous Subsections). We compare the speed ranks with that of the claims expressed in loss ratio, average cost and frequency. The correlations presented in Table IX, all the values are strongly negative (between -28 and -14%). So, if we only consider these Figures, the wind factor seems to influence the severity of storms for the insurers in a negative way, what seems illogical. We have to observe the empirical copulas to understand these negative correlations.

Events nb	loss ratio (2012 euros)	average cost (2012 euros)	frequency
1500	-0.196	-0.277	-0.149
150	-0.221	-0.134	-0.228

TABLE IX – Spearman correlation with wind speed

In Figures 1.14, 1.15 we show rank plots of a sub-dataset composed of couples whose at least one coordinate is in the 3% largest values. This explains that we have four quadrants in most of those figures : the bottom left-hand corner is usually less populated as it corresponds to small values for both coordinates. The two rectangles (bottom right and top left) contain points for which one coordinate is extreme but the other is not : they enable one somehow to quantify cases where we have complete mismatch between average max wind speed and the y coordinate. The top right-hand quadrant corresponds to points contained in the sub-dataset for which both coordinates are part of their the 0.3% largest values. This quadrant usually features some survival Clayton shape for loss ratio and claim frequency. With this approach, we both have a view on the probability to miss large losses with a low index (top left-hand rectangle) or to strongly overestimate the consequences of large wind speeds (bottom right-hand rectangle), as well as on the desired correlation for extreme values. Therefore, the correlations computed on this sub-dataset and presented in Tables IX, X correspond to a summary of negative correlation coming from the two rectangles and reinforced by this somewhat unfavorable sub-dataset choice and of positive correlation of extremes exhibited in the top right-hand quadrant.

On Figure 1.14, we observe the empirical copulas between the weighed average of maximum wind speed and loss ratio, average cost and frequency of claims. For average costs (middle graph) the point concentrations in the upper left and lower right parts of the square corresponds to the negative correlation that we have already met previously.

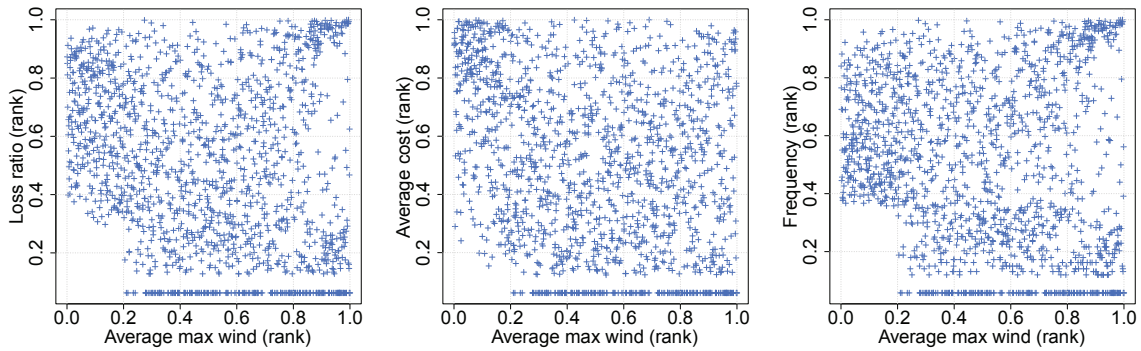


FIGURE 1.14 – Copula between claims and wind speed

On the rest of the square the distribution of points is rather uniform and it remains difficult to establish a link between average cost for the insurers and the wind speed. For loss ratio and frequency (left and right graph respectively) the distribution of points is rather similar. The most marked zone in term of point density is located at the top right corner, what corresponds well to the most extreme events for both variables. However the rest of the square is also filled relatively uniformly. It is thus difficult at this step to identify a copula corresponding to this distribution.

Then, we compare claim ranks to those of a difference between the maximum wind speed w and several wind quantiles w_q. To obtain the values of the index, we use the formula $\text{rank}\,(w - w_q)^{\alpha}$.

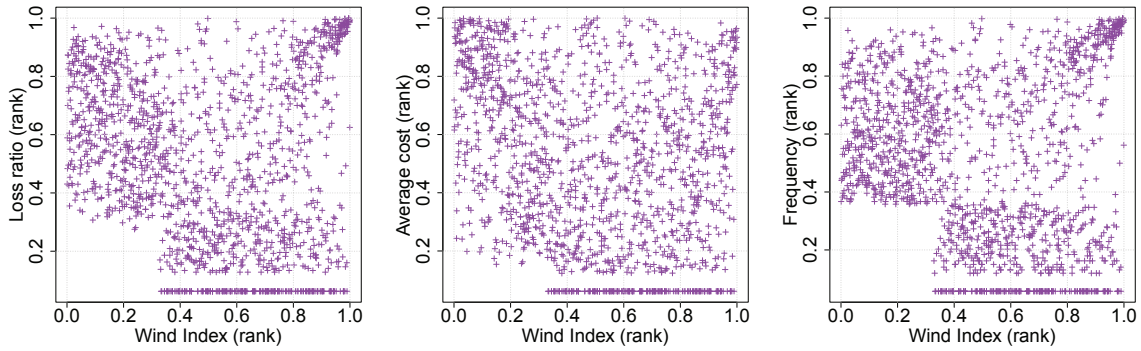


FIGURE 1.15 – Copula between claims and wind index (difference)

Figure 1.15 presents the empirical copulas obtained by comparing the ranks of the index with those of the claims. We denote a better point concentration along the diagonal for the first one and the last graph corresponding to loss ratio and to frequency. The values from the top right corner (corresponding to the major claims) are very concentrated. Graphically, the shape of these scatter plot gets closer to functions of

the survival Clayton copula or Gumbel copula. They are asymptotically dependent copulas, so this shape goes along with the notion of extreme dependence. Nevertheless, zones along x-axis and y-axis still present a significant density of points. This explains the negative values of the obtained correlations.
For the average cost the point distribution remains uniform on the main part of the square with a concentration on the top left corner. No relation is thus visible. We present the value of the corresponding Spearman's correlation in the next Table (X).

Events nb	loss ratio (2012 euros)	average cost (2012 euros)	frequency
1500	-0.155	-0.264	-0.097
150	0.106	0.146	0.114

TABLE X – Spearman correlation between wind index (difference) and claims (rank)

The values are still negative (between -9% and -26%). It would be necessary to concentrate only on the most extreme events to obtain again a positive correlation. This will be detailed in the next Subsection on tail dependance index.

We see here the limits of the improvements brought by a Formula. For a better understanding of all the important claims from a meteorological index, other criteria are doubtless necessary (as the nature of the buildings and houses, the duration of wind gusts). The precision of the data can also be improved if we had direct sources like Meteo France.

We obtain nevertheless rather satisfactory results and the gaps observed at a local level can cure with a global observation of the phenomenon. We shall see these improvements in the next section focused on main events thanks to the tail dependance index and from the section 7.4 dedicated to the storm index.

Tail dependance index

The standard way of measuring the correlation for the risk management is to zoom in on extremes by looking at the tail dependence index to the right. It is the limit when epsilon tends towards 0 of the conditional probability that the second coordinate is in the ϵ % bigger values (of the second coordinates), given that first coordinate is in the ϵ % bigger values (of the first coordinates). Mathematically, for two random variables X and Y with distribution function F_X and F_Y, the right tail dependence index is defined by :

$$\lambda_\epsilon = \lim_{\epsilon \nearrow 1} \left(\mathbf{P}(Y > F_Y^{-1}(\epsilon) | X > F_X^{-1}(\epsilon)) \right) \tag{1.5}$$

One advantage of this measure is that it is symmetric between the two variables.

We look at the same time at the cases where we do not predict a storm with a strong wind index, and the cases where we have a storm while the wind index was low.
The theory says that there is some strong tail dependence on the right if the limit is strictly positive. Variables are asymptotically independent if the limit is zero. We thus estimate empirically the indices corresponding to various thresholds. We choose to observe the conditional probability that the loss ratio is in the ϵ bigger values given that the index is also in the ϵ bigger values.

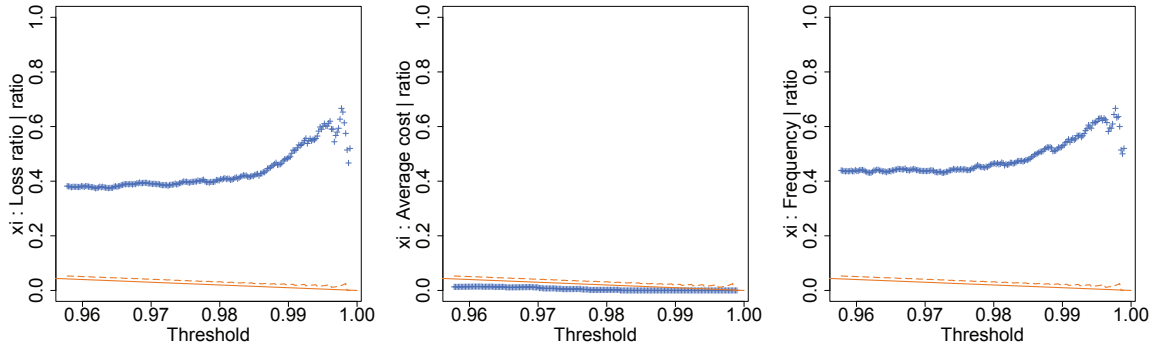


FIGURE 1.16 – Empirical tail dependence index for loss ratio, average cost, frequency, and wind index ratio

Figure 1.16 shows the evolution of the dependence index for different numbers of peak exceedances going from 1000 to 1. The higher the threshold, the closer we get to extreme events. From left to right we can see the dependence between the frequency, the average cost, the total cost and the ratio wind index. The full line represents the asymptotic independence with in dotted lines a confidence interval. Points situated over the dotted lines denote the existence of extreme dependence between claims and wind index. These graphs (we look at the limit of the blue curve when we aim towards 1, because the threshold depends on $1 - \epsilon$, what is standard) confirm the quality of the results for the frequency and the total cost (tail dependence index between 0,35 and 0,4). It also confirms that the average cost is not correlated in extremes to the indication wind index (it was a priori only few or not globally correlated thus it is logical) because the limit is 0 for the curve.

It is interesting to look also at the conditional probability that the second coordinate is in the $k.\epsilon$ % bigger values (of the second coordinates), given that the first coordinate is in the ϵ % bigger values (of the first coordinates), and conversely. It allows us to extend the perspective and better understand the variation of the index. Here we observe the evolution of the conditional probability that the frequency is in the $k.\epsilon$ % bigger frequencies, given that the index is among the ϵ % bigger index. We choose $k \in \{2, 3, 5\}$ and ϵ in a way that ϵ % corresponds to 1000 points.

The analysis carried out in 7.3.5 could give the impression that correlation is quite

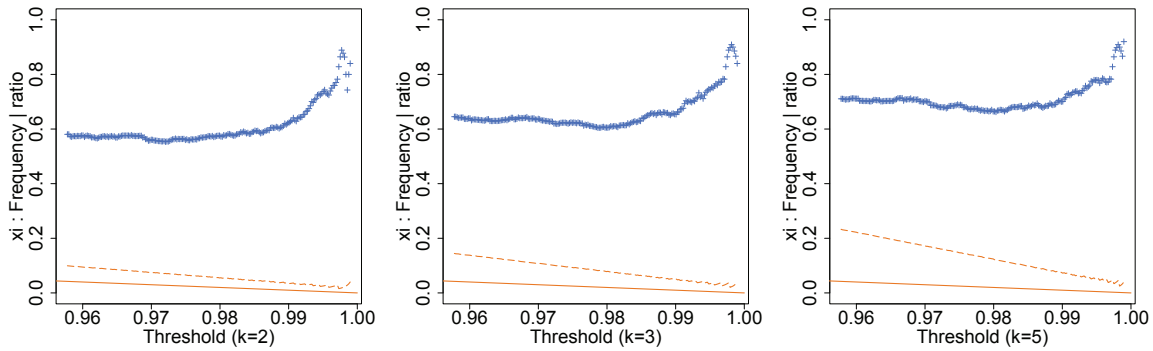


FIGURE 1.17 – Empirical tail dependence index for claims frequency and wind index ratio for $k \in \{2, 3, 5\}$ (from left to right)

low for the largest values. Figure 1.16 shows that strong correlation of extremes is really present for loss ratio and claim frequency, and completely absent for average cost. We see that the conditional probability to exceed $(1-\varepsilon)$-VaR for one variable given that the other one exceeds its $(1-\varepsilon)$-VaR increases in $(1-\epsilon)$ when $1-\epsilon \geq 98\%$. The right tail dependence index is the limit and is approximately 60%, which is quite high. Figures 1.14, 1.15 exhibited many points in the bottom right-hand and top left-hand rectangles, which could give the impression that when one coordinate is extreme, the other one might be extreme or anything. Figure 1.17 shows that the conditional probability to exceed 97.5%-VaR for one variable given that the other one exceeds its 99.5%-VaR is approximately 80%.

Overall, even if one cannot avoid false alarms or failure to represent a storm with the index, the desirable property of strong correlation of extremes between the index and frequency (or loss ratio) is satisfied, with high values of right strong tail dependence index. This shows that our index seems suitable for risk management purposes, where extremes matter.

1.8.4 Relation between wind and claims at event scale

On the basis of the events from 1998 to 2012

Here the comparison concerns only the significant events by taking into account all the stations affected by the same storm. We work with the sums of the local claim values and numbers which are used to match the storm index I_S defined in the previous section (2.7).

Concerning the damages, we work with the normalized **total cost** and with the **number** of claims. We want to focus on the main storms. We choose here to retain only the 200 main events. So, we lay the emphasis on the days over 150,000 euros for costs and 60 for daily claim numbers. The major events therefore have no risk of being

eliminated with such a threshold.

Concerning the wind, we use the **weighted average of maximum speeds** in each department. The results which follow are obtained by building this index from the upper 99% quantile and our Formula of the wind index (2.5). In order to select the main events, we have chosen a threshold of 20 for the index (between 0 and 100).

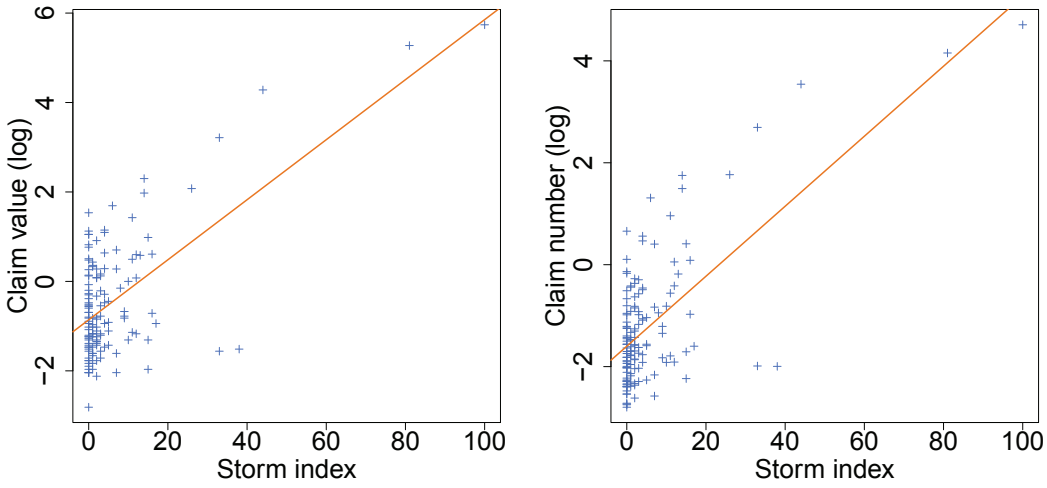


FIGURE 1.18 – Correlation between claims and storm index during the 200 main events

We do not consider anymore the days taken one by one but groups of days corresponding to the storm events. The selection of the events was made by **associating the days when the index or the claims exceed their respective thresholds**. The grouping of the days is done on a case-by-case basis by spotting the highest values over a period from 1 to 4 days around an event. On Figure 1.18 we can observe the repartition of the 200 main events. We calculate a storm index agglomerated on each event. Two approaches have been considered. The first one consists of keeping only the maximum value of the index for each event and the second one uses the sum of the index values for each event. The sum gives better correlations (more than 71%). We obtain the highest values by compiling the storm indices of each episode. This approach by episodes corresponds more to the economic reality than the daily approach. It also allows us to reflect more clearly a linear trend between the major storms and the highest indices. For the less important events we can see graphically that the relationship with the index is not as good. But our study focuses on the major events which are at the origin of the most important damages.

For example, we can estimate fairly accurately the numbers of claims (N) bred by Lothar, Martin, Klaus and Xynthia on the Allianz portfolios via the relationship :

$$N = 1.2 \times I_S - 28.6$$

On Figure 1.19, we have shown in an orange dotted line the linear regression on the whole set of values and in red on the 4 main events. We shall see later on that this relationship can still be improved.

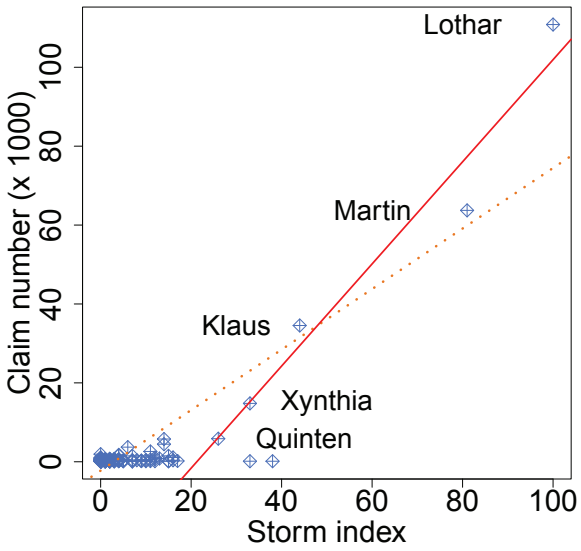


FIGURE 1.19 – Linear regression between the number of damages and the Storm index

The relation between claims and storms cannot be controlled from a single formula. There are several slices of events among which the causes and the consequences are very different. For our part only the high slice constituted by about twenty major events has a real importance on the variability of the insurer results. It will thus be important to define a threshold from which the number of claims must follow a law depending on the storm index. From a statistical perspective, the threshold is loosely defined such that the population tail can be well approximated by an extreme value model, obtaining a balance between the bias due to the asymptotic tail approximation and parameter estimation uncertainty [75]. The decisions concerning the choice of the threshold and the number of events which we hold can be made according to various methods. Graphical diagnostics like the mean residual life plot, the Hill estimator, the threshold stability plot or usual distribution fit diagnostics, are the most used. But other solutions are also available like the rules of thumb (square root rule or empirical driven rule), the computational approaches (resampling method for estimation of the optimal tale fraction (Hall) or bootstrap procedure for tail index estimation) or the mixture models (parametric, semiparametric and nonparametric bulk models).

Spearman's Correlation and storm

With Pearson's correlation the strong values have more weight than the other ones. For the insurer the major events are also the main source of damage. This bias is

therefore not necessarily detrimental to our model but it can be avoided by using a correlation based on rankings and independent of the spread of values. We present here the approach by episodes with the Spearman's correlation as an element of comparison. The weakest correlations obtained with this method are mainly due to the events of small size (both in term of costs and index) which are the most common. We nevertheless obtain a Spearman correlation around 40% which implies a link between the rankings of storms according to the index and the insurers. Figure 1.20 shows empirical copula. On the 2 graphs, only the values from the top right corner (corresponding to the major events) are concentrated. As in the case of wind index, the shape of these scatter plot gets closer to asymptotically dependent copula (like survival Clayton copula or Gumbel copula).

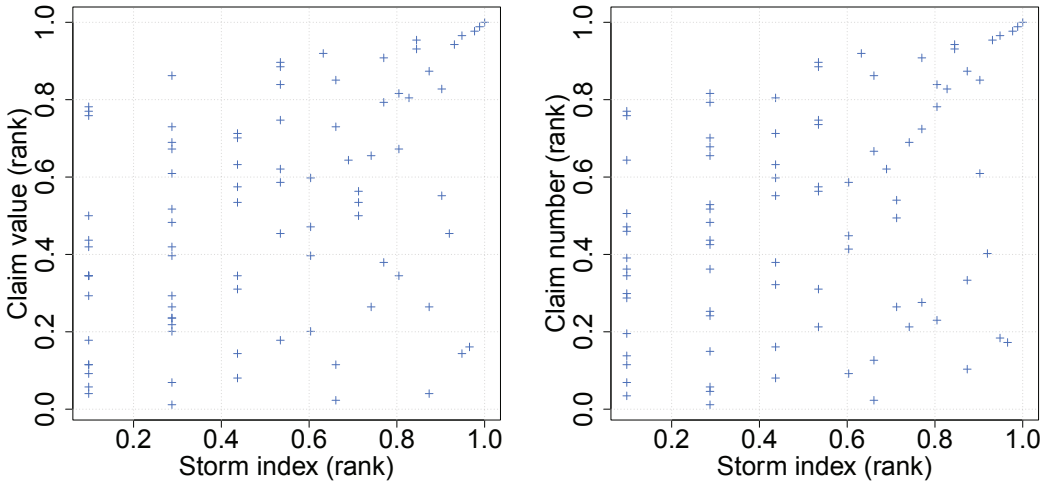


FIGURE 1.20 – Spearman's rank Correlation between normalized claims and storm index in France by event

On the basis of the events from 1970 to 2012

We take into consideration the list of major storms prepared by Luzi [51] that we now confront with different formulae of storm index. As we have seen previously the results are very sensitive to the methods used. We have therefore opted for a comparison based on both the values and on the rankings of major storms in the form of summary tables associated with a graph of empirical copula. We tested successively different weights, different quantiles and different approaches for the wind index to obtain the strongest relationship between storms and index.

For the weighting of the wind index, we first try demographic information like the population of each department or the density. Then we use Allianz' portfolio information like the number of contracts in the individual and in the global lines. For the

quantiles, we notice that small values give unsatisfactory results and we decide to test different values between 95% and 99.5%. The wind index Formulae presented in Subsection 6.1 are again optimized according to the events on a national scale. The presented results are shown here by a method of optimization of the absolute differences between the index and the normalized costs. The 1998-2012 period is used to calibrate the parameters values. Then the Formula is tested on the whole period (1973-2012).

In the following table, we have only kept the most important 20 storms according to the index I_S in the descending order. For each of these storms, the classification of insurers appears in the third column. If the storm is not part of the 20 most important ones in term of cost of insurance, we denote >20 in the place of the classification. The graph associated with each table allows the comparison of the values of the storm index with the normalized values of their cost.

Model 1 : wind index for 90 stations

As a first step, a representative meteo station is chosen in every department (for example the station of the most important city in terms of population). Then, the model 2 with 130 stations will show us if the volatility of the wind speeds on a local scale is important for our analysis (and according to the results, it seems to be)

Storm	date	rank(IG)	I_S	IG
Lothar	12/25/1999	1	100	100
Martin	12/27/1999	2	83	51
Herta	02/03/1990	5	60	19
	11/22/1984	>20	60	1
Nov 82	11/07/1982	3	48	25
	12/17/1986	>20	48	1
Vivian	02/26/1990	11	44	7
	12/08/1991	>20	40	1
	02/07/1984	>20	39	1
Xynthia	02/27/2010	10	37	8
Klaus	01/23/2009	4	33	20
	03/24/1988	>20	29	1
Daria	01/25/1990	9	29	9
87J	10/15/1987	6	28	10
	11/26/1983	>20	28	1
	02/03/2002	>20	27	1
Wiebke	03/28/1990	>20	24	1
Quinten	02/09/2009	>20	23	1
	01/22/1988	>20	23	1
	02/10/1974	>20	22	1

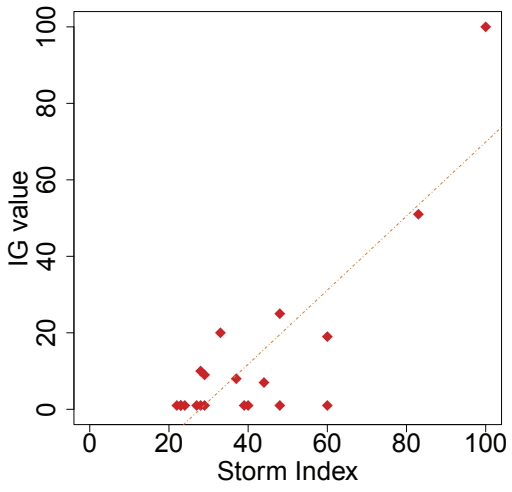


FIGURE 1.21 – Storm classification with 90 stations

With these new weights, storm Martin stands out of the rest of the values to be closer to Lothar. In the classification Klaus goes from the 17th to the 11th position before Daria and Wiebke, but stays nevertheless behind Xynthia. By focusing on the

values that constitute the index during major events, we realize that wind speeds above 100 kph entail significant overruns and strongly influence the value of the index, which depends on their squares. The choice of stations used to represent each department therefore plays an important role. Indeed, it is possible to overestimate or under-estimate the index of a storm if the measurement of the selected station for a department shows a significant gap in relation to the other stations in that same department. We would then favour a non representative station. The sub index assessment of Klaus could be due to the position of stations in relation to the trajectory of the storm. For Lothar for example this speed of 100 kph is exceeded 72 times on 83 active stations for Martin : 57/82, for Xynthia : 60/87 and for Klaus only 41/89. To verify this hypothesis, a denser meshing of stations during major events will provide more details.

Model 2 : wind index for 130 stations

On the NCDC's website the readings of 130 stations located throughout France have been collected. So the accuracy of wind speeds is improved. If we consider Klaus again, focusing on the Gironde, we now have not only Bordeaux but two other readings at Cap Ferret and Cazaux. On January 23, 2009, during the passing of the storm on the department the bordelaise station showed a maximum speed of 172.8 kph. With the other two stations we learn that the maximum speed in Cazaux was of 140.4 kph but especially that an exceptionally high reading has been recorded in the station of Cap Ferret with 251.6 kph. This important difference confirms the usefulness of a database giving the widest possible scope in particular during extreme events. We then calculate a new index I_{S_2} based on the 130 stations.

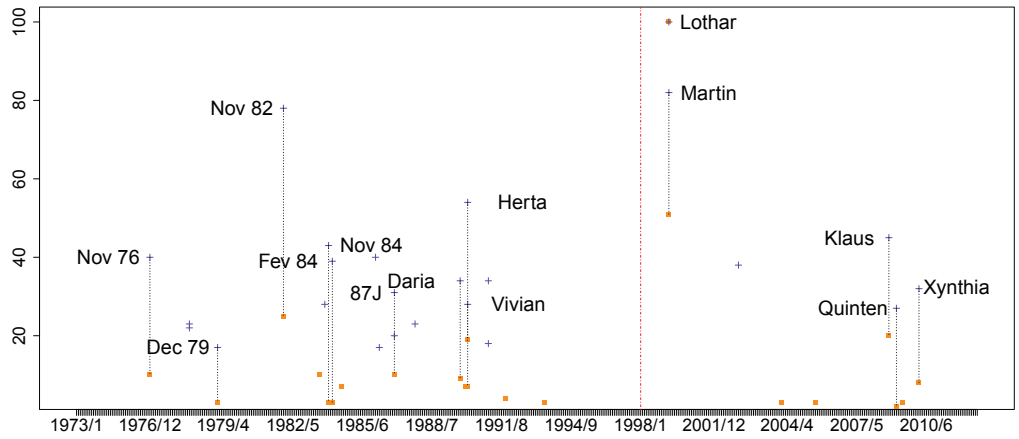


FIGURE 1.22 – Storm index with 130 stations between 1973 and 2012

With this new information, Lothar remains the main storm, still followed by Martin. The storm of early November 1982 comes inserted before Herta and thus appears to the ranking of major events. This storm has swept across western Europe causing

Storm	date	rank(IG)	I_{S_2}	IG
Lothar	12/25/1999	1	100	100
Martin	12/27/1999	2	82	51
Nov 82	11/07/1982	3	78	25
Herta	02/03/1990	5	54	19
Klaus	01/23/2009	4	45	20
Jul 83	07/13/1983	8	44	10
Nov 76	11/30/1976	7	40	10
	12/17/1986	>20	40	1
	02/07/1984	>20	39	1
	02/03/2002	>20	38	1
Daria	01/25/1990	9	34	9
Xynthia	02/27/2010	10	32	8
87J	10/15/1987	6	31	10
Vivian	02/26/1990	11	28	7
	12/01/1984	>20	28	1
Quinten	02/09/2009	>20	27	1
	01/24/1978	>20	23	1
	01/22/1988	>20	23	1
	11/01/1978	>20	22	1
	12/08/1991	>20	20	1

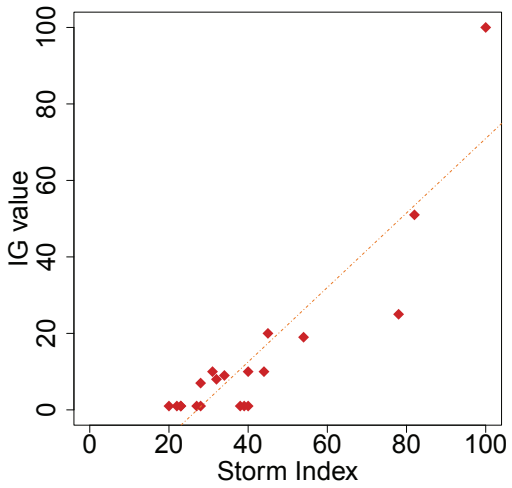


FIGURE 1.23 – Storm classification with 130 stations

catastrophic destruction in some thirty departments and many victims according to Météo France that ranked it among the most violent of these past thirty years. Furthermore the storm Klaus comes now in 5th position and shows especially an index much higher than that of Xynthia according to the insurance results . The storms of February 1990 will also be recorded as well known events of our period of observation. Daria overpasses Xynthia and Vivian.

This comparison shows that it is useful to use more stations. With I_S certain events such as storms of November 82 or Klaus are under-estimated and some among the most costly were not even as that of November 76. For I_{S_2} the most complete updating (**Act IG**) offers the most fairly estimated points. The 5 main events of the period are the same with an inversion between Klaus and Herta but relative because the gap is very low in the two approaches. The classification is then more difficult to compare but the order of arrival of Daria, Xynthia and Vivian remains the same.

Model 3 : Ratio Cubic Index for 130 stations

We perform here the same comparison but with the cubic index CI_S (1.2) proposed by Klawa et al. This index showed itself slightly more effective than ours during the local approach (department by department) but from now on, it is not the case anymore. In Table, storms are less well classified than previously. In particular, two minor events for the insurers, the storms of December 86 and January 78 are inserted between Lothar and Martin. During the 1986 storm we notice two stations with wind

speeds over 175 kms/h but few damaged departments in Allianz' historical data. This storm was too local to cause important damages. We thus have poorer results with this index at a global level.

Storm	date	rank(IG)	CI_S	IG
	12/17/1986	>20	100	1
Lothar	12/25/1999	1	90	100
	01/24/1978	>20	68	1
Martin	12/27/1999	2	53	51
	12/01/1984	>20	47	1
Nov 82	11/07/1982	3	47	25
Herta	03/02/1990	5	44	19
	02/03/2002	>20	35	1
Klaus	01/23/2009	4	30	20
Jul 83	07/13/1983	8	24	10
Vivian	02/26/1990	11	24	7
Xynthia	02/27/2010	10	21	8
	02/02/1986	>20	18	1
	08/02/1984	>20	17	1
Daria	01/25/1990	9	17	9
	02/13/1976	>20	15	1
87J	10/15/1987	6	14	10
Nov 76	11/30/1976	7	13	10
	08/20/1976	>20	12	1
	10/10/1987	>20	12	1

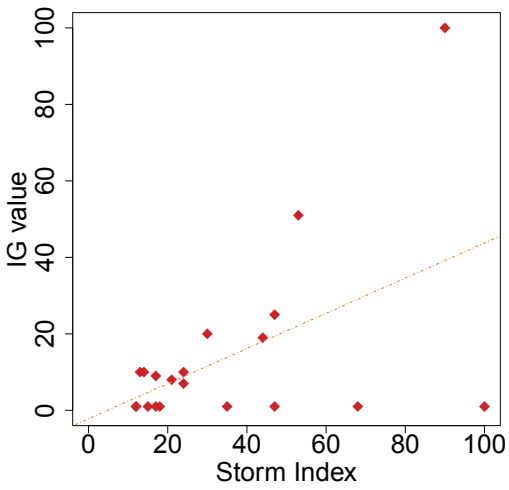


FIGURE 1.24 – Storm classification with cubic index and portofolio weights

When the wind index exponent α is greater or equal to 3, the order of the storm index values is very perturbed. A possible explanation is that localized storms with high wind speed are overestimated.

From the meteorological point of view it is surely possible to optimize the parameters of the storm index to obtain a better representativeness of observed events. Not all parameters are known, the duration of gusts of wind for example could refine some values. However, it must be borne in mind that, given the complexity of the phenomena at stake, a totally indisputable model will be difficult to obtain.

From the economic point of view it is possible to improve both the accuracy and the length of this type of historical background based on more comprehensive data from the market. The results of the reinsurance could for example throw another light on this subject.

The obtained hierarchy between the major storms of the period is satisfactory. The gaps and the orders of magnitude can be improved. We propose to modify the index so as to amplify the differences. The idea here is to take the exponential of the sum of the indices of wind recorded during a storm episode in order to highlight the extra cost

which seems to go with the most destructive events. In addition, it would be interesting to be able to change the weight given to the number of stations because an event can be overestimated if it takes place in an area well represented by the stations while their total number remains relatively low.

Model 4 : New parameter

At this stage, we have added a new parameter in order to check the scale of storms. This adaptation of the formula may seem specific to our database but in our opinion it responds to a more general problem. We can indeed observe a gap between the most important costs and the rest of the values. In our case, the gaps between the three biggest storms double. This peculiarity is not an isolated case and finds itself for example in the classification of hurricanes proposed by Pielke et al. [63]. In their study, the worst hurricane recorded between 1900 and 2005 is the one of Miami in 1926 and its value is twice as large as the second hurricane of the ranking. Our new formula associates the exponential function with a new parameter β. The new index formula becomes

$$I_{S_3} = \exp\left(\beta \sum_{d\in E} \sum_{s} \frac{I_w^d(s) \times C(s)}{N_a^d}\right), \tag{1.6}$$

where in each station $I_w^d(s)$ is the wind index on date d and the number of risks is balanced by the size of Allianz global portofolio $C(s)$. We also take into account the size of the damaged area (geographic aggregation), the duration of the storm event E (daily aggregation) and the number of active stations on date j, N_a^d.

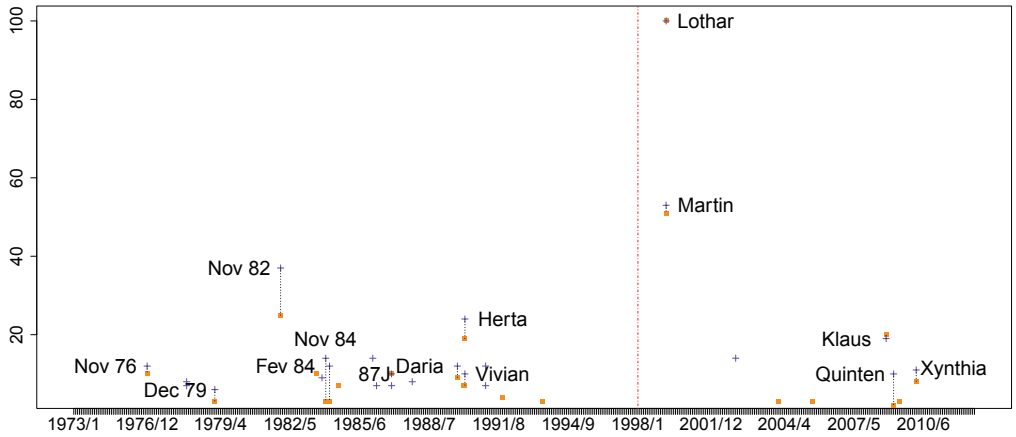


FIGURE 1.25 – Storm index with 130 stations between 1973 and 2012

With this latter result, we find a much more important parallel with the insurance

data from the insurance. Storm Lothar now clearly appears with an index almost twice as large as that of Martin. The following differences also correspond to the discrepancies of the normalized costs. Graphically, we obtain a much better alignment than previously for the costs, the frequencies and number of claims. The points corresponding to the major storms are closer to the linear regression.

Storm	date	rank(IG)	I_{S_3}	IG
Lothar	12/25/1999	1	100	100
Martin	12/27/1999	2	53	51
Nov 82	11/07/1982	3	37	25
Herta	02/03/1990	5	24	19
Klaus	01/23/2009	4	19	20
Jul 83	07/13/1983	8	16	10
	02/03/2002	>20	14	1
	11/22/1984	>20	14	1
Jul 84	02/07/1984	12	12	7
Daria	01/25/1990	9	12	9
Nov 76	11/30/1976	7	12	10
Xynthia	02/27/2010	10	11	8
87J	10/15/1987	6	10	10
Vivian	02/26/1990	11	10	7
Quinten	02/09/2009	>20	10	1
	12/01/1984	>20	9	1
	01/22/1988	>20	8	1
	01/24/1978	>20	8	1
	05/10/1987	>20	7	1
	12/08/1991	>20	5	1

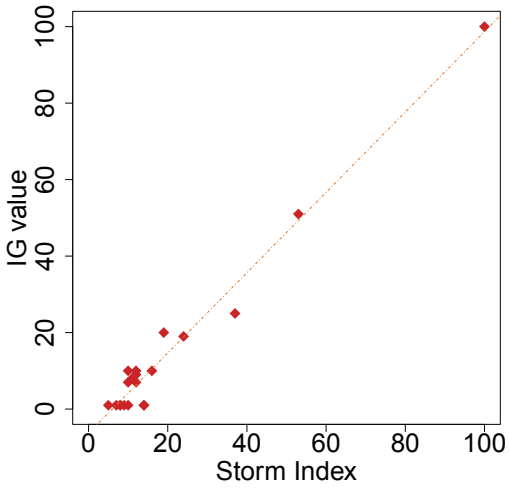


FIGURE 1.26 – Storm classification with portofolio weights and β parameter

The relationship between this new index and the loss experience of the 4 major events of the 1998-2012 period becomes even more direct with a slope equal to 1 :

$$N = I_{S_3} + 9.1.$$

On Figure 1.27, we have shown in an orange dotted line the linear regression between the set of values and in red between the 4 main events. The very strong correlations and this linear relationship allow us to support the relevance of our storm index.

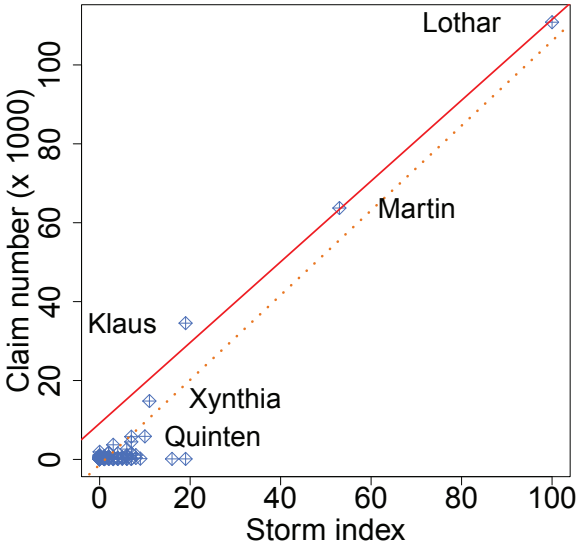


FIGURE 1.27 – Linear regression between the number of damages and the Storm index

1.8.5 Wrap up

Thanks to these comparisons, we have shown how parameters and variables can influence the quality of a wind index and a storm index. Concerning the claim criteria, it can be stated that the approach according to the average costs is doomed to failure (nevertheless, it seems to be the basis of all the storm software). The frequency of claims offers the strongest relation with wind speed. The use of other claim criteria than frequency (loss ratio ; individual / global lines) does not lead to fundamentally different results.

In our search for functions of damage at the level of the department, we have highlighted that it is better to use a wind criterion based on a weighted average over each department of the maximum daily speeds. This stage requires a real work of data processing to select stations with regard to departments. We use the concept of moving average to clarify the plots and focus on the main events. Then we show how the modulation according to the local wind habituation brings a sensitive improvement to the correlations between wind and claims. We complete the first comparisons based on Pearson's correlation by using the Spearman's correlation based on variable ranks. We notice that it is very difficult to obtain a robust relation between the wind index and the claims, for the small and medium events. We have to focus on the main events to observe a clear dependance. So we decide to calculate the tail dependance index and we obtain good results which prove that there exist a strong extreme dependance.

Finally, we present the optimization of the storm index at a global level. The successive models are based on the period 1998-2012 and we try to track down the main

events registered since the 70s. The Formula (2.7) overestimates small events (very localized but with strong wind gust) when it is associated with the cubic index. So, at a national scale, our wind index allow us to improve the relation between claims and storm index. By using a new parameter and the exponential function, we finally obtain the last storm index Formula (1.6). This Formula reflects both the order and the scale of the main storms that made landfall in France since the 70s.

Our next target is to study in a more precise way the sensibility of costs and return periods of a major event on a scale comparable or superior to that of Lothar in France. This subject will be dealt with in an article to come with a modeling directly based on the values of the storm index according to extreme value theory ([10] and [72]). The variability of the results to the assumptions adopted will be tested within the framework of the European statutory reform of the insurance, Solvency II.

Acknowledgements

We thank the referees for useful remarks and suggestions. We are also grateful to S. Resnick and G. Samorodnistsky for relevant comments. This work has been supported by Allianz France, by the research chair *Actuariat Durable* sponsored by Milliman Paris, by the GICC Miraccle and ANR McSim projects.

Chapitre 2

Gestion du risque tempête : sensibilité du calcul de la période de retour et répartition sur le territoire

Sommaire

Alexandre Mornet Université de Lyon, Université Claude Bernard Lyon 1, Institut de Science Financière et d'Assurances, 50 Avenue Tony Garnier, Lyon F-69007, France and Allianz, Coeur Défense, 82 Esplanade du Général de Gaulle, Courbevoie F-92400, France, alexandre.mornet@allianz.fr
Thomas Opitz Biostatistics and Spatial Processes Unit, National Institute of Agronomic Research, Avignon, France, topitz@paca.inra.fr
Michel Luzi Non-life actuarial affairs former Director, Research ans Development Director at Allianz France, qualified Member of Institute of Actuaries, 132, rue du Président Wilson, Levallois-Perret F-92300, France
Stéphane Loisel Université de Lyon, Université Claude Bernard Lyon 1, Institut de Science Financière et d'Assurances, 50 Avenue Tony Garnier, Lyon F-69007, France, stephane.loisel@univ-lyon1.fr
Bernard Bailleul Direction Métiers ABR, Actuaire non vie, ALLIANZ, Coeur Défense, 82 Esplanade du Général de Gaulle, Courbevoie F-92400, France.

2.1 Abstract

Les modèles et les projections des dommages engendrés par les tempêtes sont un sujet majeur pour les compagnies d'assurance. Dans cet article, nous voulons insister sur la sensibilité du calcul de la période de retour des événements les plus extrêmes. De nombreux éléments entrent en jeu, comme la qualité des données (localisation précise des bâtiments assurés, homogénéité des relevés météorologiques), le manque de mises à jour (historique de portefeuilles pour l'assurance, changement de rugosité des sols, changement climatique), l'évolution du modèle suite à un événement sans précédent comme Lothar en Europe et le choix du pas de temps (clusters de 2 ou 3 jours relatifs aux événements ou clusters fixes d'une semaine). Un autre aspect important concerne la trajectoire des tempêtes qui pourrait changer suite au réchauffement climatique ou balayer des zones plus importantes. Nous définissons ici un partage du territoire français en 6 zones tempêtes, dépendant des corrélations extrêmes de vent, pour tester plusieurs scénarios. Nous nous appuyons sur notre indice tempête − défini dans un précédent article [57]− pour montrer les difficultés rencontrées pour dégager des résultats robustes en lien avec les événements extrêmes.

Keywords. Storm Index, Volatility, Insurance, Extreme Value Theory, Extreme Dependence, Return period, Solvency II

2.2 Introduction

L'étude des tempêtes relève d'une part d'un phénomène naturel lié à des facteurs climatiques et géographiques, d'autre part de ses conséquences en termes de sinistres liés à des facteurs économiques, historiques et humains. La gestion du risque lors des événements extrêmes [38] est un sujet qui touche au delà de l'assurance toute la société. Avant toutes choses, il faut, pour comprendre notre démarche, considérer les problématiques que représentent la gestion et l'évaluation des tempêtes. Les informations économiques et structurelles disponibles au sein d'une société d'assurance pour la garantie tempête rendent compte de la volatilité et l'ampleur des dommages. Cependant elles ne permettent qu'un travail basé sur des données historiques limitées et difficiles à actualiser [42]. C'est pourquoi nous avons choisi de compléter nos données par des relevés météorologiques. Nous avons exploré les différentes variables susceptibles d'être liées aux tempêtes pour finalement nous focaliser sur les vitesses de vent (comme [45], [64] and [24]).

L'évaluation de l'impact d'une tempête n'est pas pour autant aussi simple et directe qu'on pourrait le souhaiter. Plusieurs difficultés rencontrées vont ici être décrites pour expliquer la variabilité des résultats qui en découlent. S'il est possible de disposer de données sur les sinistres et sur les portefeuilles à des niveaux relativement fins (la commune, voire l'adresse postale), on ne dispose généralement plus que de relevés au niveau départemental dès que l'on utilise des données de marché ou des historiques anciens. De plus, si l'on se réfère aux données météorologiques, on ne dispose que d'un nombre limité de stations, en gros une à deux par département, ce qui limite les possibilités de travailler avec précision. Nous notons aussi une grande variabilité des vitesses enregistrées selon la position géographique de la station [74]. Les stations en altitude ou sur le littoral présentent souvent des vitesses supérieures à celles du reste du département qui doivent être pondérées selon leur distance avec les principales agglomérations. Ensuite, pour une même station, l'homogénéité des enregistrements sur une longue période n'est pas forcément vérifiée. L'étude des phénomènes de ruptures [78] nous a permis de repérer des décalages, peut-être dus à des changements de matériel, mais qui méritent d'être corrigés. D'autres phénomènes comme les changements de rugosité des sols ou le changement climatique peuvent aussi perturber les données.

Étant donné ces difficultés, une solution possible est un regroupement des départements français selon les vitesses extrêmes de vent. Paradoxalement, on serait tenté de chercher à travailler avec une maille plus fine, au moins pour éviter de comparer des vitesses du vent à plusieurs dizaines de kilomètres des risques. Or, ici, on retient une solution inverse. Notre découpage se justifie par le fait qu'il y ait des départements avec très peu de dommages forts (car peu de contrats). À cela s'ajoute le nombre limité des événements significatifs et l'hétérogénéité des distances entre risque et station. Une approche statistique par département manquerait donc de robustesse. Nous avons opté pour une répartition en 6 zones de risques, ce qui permet d'agir à une échelle intermédiaire entre le département (dimension 95) et la France entière. Nous utilisons pour cela l'algorithme

de classification des `k-médoides`[17]. Le travail sur ces zones nous permet de tester différentes méthodes d'agglomération des départements (dépendance extrémale [16], vitesse maximale, distance géographique). Il nous permet aussi de considérer différents scénarios d'amplitudes de tempêtes comme dans [39], [49]. Ces amplitudes pourraient évoluer suite au changement climatique (apparition de tempêtes de type Medicane [1] par exemple) ou balayer plusieurs zones consécutivement. Ces projections représentent un moyen cohérent de palier le faible nombre de tempêtes majeures dans notre base d'observation.

Sur un plan plus technique, la construction d'un indice tempête implique le choix d'un pas de temps. Il faut ici dissocier les regroupements de journées pour coller avec la notion de l'évènement retenu comme définition par les assureurs et la notion de cluster qui correspond à une période favorable au déclenchement d'événements successifs. Nous avons réalisé que contrairement à l'évaluation des charges qui nécessite l'addition des sinistres sur la durée de l'événement, le maximum journalier de notre indice est la meilleure information sur l'intensité relative des tempêtes. Ces décisions auront évidemment un impact sur la projection des tempêtes. Nous avons comparé les résultats des approches statistiques selon l'emploi de données quotidiennes ou hebdomadaires. Dans nos modèles nous avons aussi voulu montrer la sensibilité du calcul de la période de retour à l'inclusion ou non des événements les plus importants comme Lothar et/ou Martin en 1999.

Pour un assureur, la gestion des dégâts liés au vent dépend très fortement de la période de retour que l'on attribue aux tempêtes les plus extrêmes. Ces événements sont par définition très rares et les historiques de sinistralité s'avèrent insuffisants pour apporter des réponses solides aux questions relatives à la tarification, à la volatilité des résultats, donc au besoin en fonds propres. La communication sur l'incertitude qui accompagne la prévision des évènements extrêmes est importante et s'avère souvent négligée. Il existe pourtant des moyens techniques pour déterminer cette incertitude mais une réticence sociologique vis à vis de l'aléa limite sa diffusion. Ce fut le cas lors de la tempête Juno de 2015 qui fut en grande partie surestimée comme l'explique Winkler [81]. L'objectif de cet article est d'apporter un nouvel éclairage sur la modélisation et le calcul de l'impact de ces événements extrêmes en s'appuyant sur l'indice tempête défini dans notre précédent article [57]. On veut ici insister sur la grande sensibilité des résultats aux hypothèses retenues, tout en proposant des résultats concrets appuyés sur des données météorologiques disponibles sur une période d'une quarantaine d'années et moins perturbées par les problèmes d'actualisation que les données d'assurance. Cette sensibilité sera illustrée à la fois par le calcul des périodes de retour et par l'estimation des besoins en fonds propres requis par la directive Solvency II.

1. Medicane est la contraction de Méditerranée et hurricane (ouragan en anglais). Ce phénomène climatique existe depuis une vingtaine d'années. Il est du à la conjonction de l'air froid en altitude et d'un réchauffement inhabituel de la mer Méditerranée.

2.3 Autour de la vitesse du vent

2.3.1 Définition

Le vent est un mouvement au sein d'une atmosphère. Mécaniquement, il est décrit par la mécanique des fluides. Les molécules qui composent l'air ne sont pas solidaires les unes des autres, ce qui complique la prévision de ces trajectoires mais va permettre au flux de s'adapter aux configurations topographiques qui ne manqueront pas d'influencer son écoulement [12].

En tout point fixe de l'atmosphère où il doit être mesuré, le vent varie généralement au cours du temps de façon brusque, rapide et parfois importante. Sa mesure isolée à un instant donné, ou vent instantané, ne serait guère représentative de la valeur d'ensemble de ce mouvement horizontal au site où on l'observe, ni des écarts à cette valeur. De plus, ces particularités du comportement dynamique de l'air se remarquent aussi bien pour la mesure de la vitesse du vent que pour celle de sa direction.
La valeur de référence de la vitesse et de la direction en un point et à un instant de mesure donnés sont les moyennes des vitesse et direction instantanées sur un intervalle assez long (généralement de 10 minutes) qui précède l'instant de mesure. Les bulletins météorologiques font référence au vent moyen, tout en mentionnant éventuellement la mesure observée ou l'intensité prévue des rafales (dont les vitesses peuvent dépasser de plus de moitié celles du vent moyen). Il est aussi possible de mesurer la composante verticale du vent avec un anémomètre tridimensionnel. On peut ainsi étudier les mouvements subsidents et ascendants de l'air. Cependant nos sources d'information ne sont pas aussi exhaustives. Nous avons donc essayé de faire pour le mieux avec les données à notre disposition.

2.3.2 Les données Météo France

Météo France propose 7 types de relevés associés au vent : vitesse vent moyen maximal, moyenne des vitesses du vent (à 10 mètres), direction du vent instantané maximal (à 10 mètres), vitesse du vent instantané maximal, vitesse du vent instantané sur 3 secondes maximal, vitesse du vent instantané maximal à 2 mètres, moyenne des vitesses du vent à 2 mètres. Par exemple, la vitesse vent moyen maximal est définie de la façon suivante. Il s'agit de la vitesse maximale du vent moyenné sur 10 minutes, relevée entre 00 UTC le jour J et 00 UTC le lendemain (J+1). La hauteur conventionnelle de la mesure est de 10 mètres, l'unité est le mètre par seconde et dixième (1 ms-1 = 3,6 km/h = 1,945 noeuds).

2.3.3 Problème de rupture des données

Parmi les problématiques abordées, il est aussi question de la fiabilité des données météo. En effet, lorsqu'on s'intéresse à une longue période dans le temps, les appareils de mesures et les méthodes de restitutions des données peuvent changer et introduire

un biais dans les résultats [78]. Ces impacts peuvent être considérables, spécifiquement sur l'écart-type des estimations de la tendance qui est inversement proportionnel à $T^{3/2}$ où T est la longueur de la période d'observation. Il est donc intéressant de savoir détecter les phénomènes de rupture et de pouvoir les corriger. A travers l'observation des moyennes/médianes/variances annuelles, pour chaque station, on essaie de détecter un changement dans ces séries.

2.3.4 Détection de ruptures dans nos relevés

On utilise le package `changepoint` du logiciel `R` [70] pour détecter les éventuelles ruptures parmi les données de vitesses de vent que nous utilisons. Les différentes méthodes d'optimisation sont décrites dans l'article de Killick et al. [44]. Formellement, pour une suite de données $(y_1, ..., y_n)$, on dit qu'il existe une rupture $\tau \in \{1, ..., n-1\}$ si les propriétés statistiques des sous-ensembles $\{y_1, ..., y_\tau\}$ et $\{y_{\tau+1}, ..., y_n\}$ sont différentes selon un certain critère. La détection multiple de ruptures se fait ensuite dans la littérature en minimisant :

$$\sum_{i=1}^{m+1} \left[\mathcal{C}\left(y(\tau_{i-1}+1) : \tau_i\right)\right] + \beta f(m) \tag{2.1}$$

avec $\mathcal{C}$ une fonction de coût pour chaque segment et $\beta f(m)$ une pénalité pour éviter la sur-segmentation des données. Dans ce document, nous utilisons l'algorithme PELT (Pruned Exact Linear Time). PELT effectue d'abord un test de rupture sur l'ensemble de la période. Si une rupture est détectée alors les données sont divisées en deux à la position de la rupture. La détection est ensuite effectuée à nouveau sur les deux jeux de données ainsi créés. Cette procédure continue jusqu'à ce qu'aucune rupture ne soit plus trouvée. Pour éviter les problèmes de saisonalité, on se concentre ici sur les moyennes annuelles de chaque station. De plus, comme notre indice est construit à partir des vitesses les plus élevées, la détection se fait ici sur les dépassements au delà d'un seuil fixé au quantile q à 80%.

Sur le graphique du haut de la Figure 2.1 on peut observer les résultats de la détection de ruptures sur une moyenne annuelle des dépassements de vitesse de vent. La période est découpée en deux séquences : de 1970 à 1983 puis de 1984 à 2013. Pour cette station, les vitesses enregistrées durant la première séquence sont surévaluées par rapport à celles de la deuxième séquence. Il est donc nécessaire d'homogénéiser les données pour Nîmes. Jusqu'en 1983, la moyenne des dépassements est de 39.87 m/s. et à partir de 1984, elle est de 35.07 m/s. Nous harmonisons la série (bas de la Figure 2.1) en recentrant les dépassements de la première séquence sur ceux de la deuxième plus récente dont nous supposons que les relevés sont plus fiables. Notons néanmoins que cette approche reste simpliste et qu'il faut donc rester prudent en interprétant les résultats qui impliquent des données avant les années 80. Les données anciennes sont beaucoup moins fiables et plus bruitées, même après avoir fait des prétraitements.

Plusieurs explications peuvent être fournies pour tenter d'expliquer l'origine de ces ruptures. Une modification de la topologie autour de la station suite, par exemple à

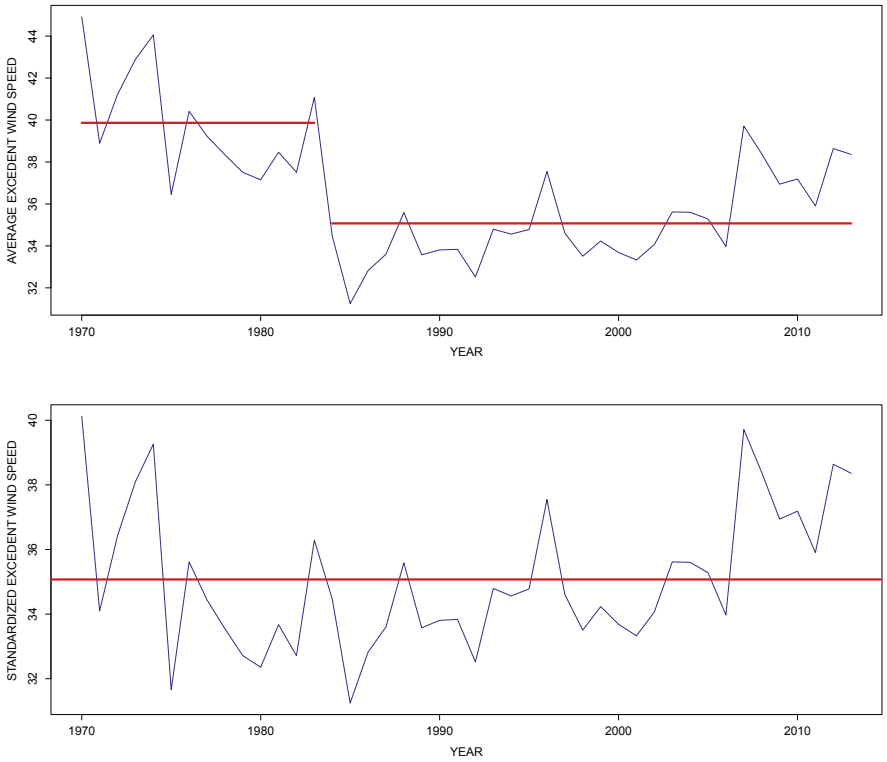


FIGURE 2.1 – Dépassements moyens de vitesses de vent à Nîmes avant recentrage (en haut) et après recentrage (en bas)

la construction d'un bâtiment à proximité ou le changement du type d'anémomètre utilisé sont des sources de perturbation des mesures. On apprend dans l'article de Sacré et al [74] que l'utilisation dans les années 70 des anémomètres Papillon a posé problème notamment pour estimer les fortes vitesses de vent. La figure 2.2 représente le transmetteur de l'appareil en question [61]. Ces anémomètres de type mécanique sont sensibles au gel et à l'humidité, ce qui amène à sous estimer les vitesses de vent hivernales en montagne. L'appareil qui s'impose actuellement utilise des émetteurs-récepteurs d'ultrasons directionnels. Pour autant, même sophistiqué, les mesures ne sont pas à l'abris de problèmes techniques. La station de l'aéroport de Nice par exemple, présente des erreurs récurrentes malgré son profileur de vent qui s'avère incapable de détecter la brise de terre et le brise de montagne de la vallée du Var, le département de Nice.

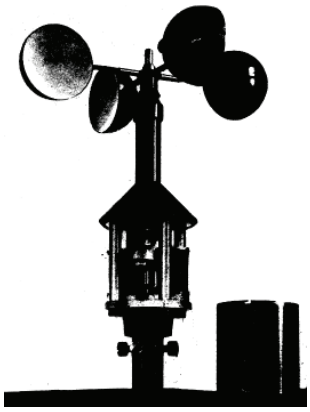


FIGURE 2.2 – Transmetteur de l'anémomètre Papillon.

2.4 Répartition des risques et des stations sur le territoire

Les stations météorologiques professionnelles du réseau de Météo-France, appelé réseau Radome, sont au nombre de 554 en métropole (une tous les 30 km) et 67 en Outre-Mer. Ces stations mesurent actuellement de façon automatique les paramètres de base que sont la température et l'humidité sous abri, les précipitations et le vent (vitesse et direction). Certains paramètres répondent à des besoins plus spécifiques et sont concentrés sur des zones sensibles : aérodromes, zones de risque d'inondations, d'incendies de forêt, d'avalanche. Outre les mesures traditionnelles effectuées par les baromètres (pression), hygromètres (humidité de l'air), anémomètres (vent) et pluviomètres (pluie), les progrès technologiques permettent aujourd'hui de mesurer de façon automatique un nombre de plus en plus grand de paramètres comme l'humidité du sol, l'état du sol, le rayonnement, la visibilité, la hauteur des nuages.

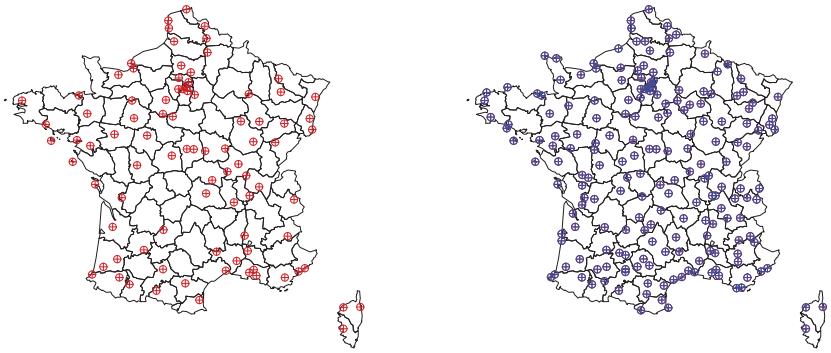


FIGURE 2.3 – Cartes des stations actives à partir de 1963 (à gauche) et de 1998 (à droite).

Cependant, ce nombre de station est relativement récent. Plus on remonte dans le

temps moins le maillage sera précis. Comme nous souhaitons travailler sur un historique de 50 ans, nous devons considérer ici l'évolution de la situation entre 1963 et 2013. Sur la Figure 2.3 sont représentées l'ensemble des stations actives en 1963 (début de notre période d'observation), puis en 1998 (début des historiques des sinistres avec une précision journalière chez Allianz). On peut voir sur ces cartes une nette évolution du maillage des stations sur le territoire. En 1963, on dispose des relevés de 89 stations. Certains départements ne sont pas représentés et on remarque même un important manque d'informations autour du Limousin et de la Champagne-Ardennes. Trente cinq ans plus tard, on a plus que doublé ce nombre avec plus de 190 stations actives et quasiment aucun département manquant à part les Hauts-de-Seine et la Seine Saint-Denis en région parisienne.

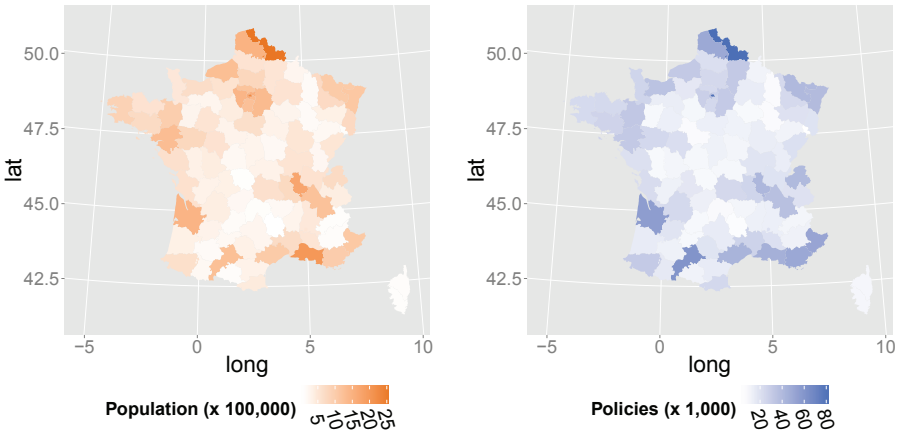


FIGURE 2.4 – Cartes de la population (à gauche) et du nombre de polices Allianz (à droite).

Cette répartition des stations est à mettre en parallèle avec les densités de population et en particulier le nombre de contrats du portefeuilles Allianz dans chacun des départements (Figure 2.4). Ceci va avoir un impact important sur notre recherche d'un lien entre les informations météorologiques et les dégâts enregistrés par les assureurs. Aux vues de ces disparités, on perçoit que l'échelle du département n'est pas forcément la plus adaptée à la construction d'un indice tempête. Nous allons voir dans la section suivante qu'il est possible de travailler sur des zones plus larges et plus homogènes en associant les départements selon leur exposition aux fortes vitesses de vent.

2.4.1 Classification des départements selon des zones de risque tempête homogènes

L'objectif est ici l'agrégation des données dans des territoires plus larges que le département pour faciliter la recherche du lien entre les vitesses de vent et les charges. En effet, travailler sur toute la France ne tiendrait pas du tout compte des hétérogénéités du territoire, alors qu'une approche par département pose un problème par rapport à la disponibilité, fiabilité et hétérogénéité des données. D'où l'intérêt de cette approche

intermédiaire qui permet de regarder les résultats pour les zones une par une (ce qui aurait été difficile pour 95 départements). La France offre un territoire au relief et au climat très diversifié, il semble ainsi normal de chercher à délimiter des zones susceptibles d'être touchées simultanément. On espère ainsi mettre en évidence des trajectoires privilégiées par les tempêtes ou au contraire des régions relativement épargnées par les dégâts du vent. De plus, les incertitudes rattachées à des résultats calculés par département auraient été très fortes, alors que le regroupement permet de les diminuer. Pour l'assurance, cette approche pourrait permettre de voir les régions où il y a des risques plus forts ou plus faibles par rapport à la moyenne nationale.

Pour k fixé, l'algorithme des k-médoïdes cherche à trouver k zones en minimisant les distances entre un département central et des départements associés. Le médoïde est le représentant central d'une zone, c'est à dire le département central dans notre cas. Le choix de la distance est donc important, nous avons utilisé comme base l'indice de dépendance des extrêmes λ qui mesure la dépendance des queues de distribution. Cet indice représente la probabilité de réalisations extrêmes simultanées et prend donc ses valeurs entre 0 (indépendance asymptotique) et 1 (dépendance totale). Pour deux variables aléatoires X et Y ayant pour fonction de répartition F_X et F_Y l'indice est défini par

$$\lambda = \lim_{v \to 1} \left(\mathbf{P}(Y > F_Y^{-1}(v) | X > F_X^{-1}(v)) \right)$$

,

pourvu que cette limite existe. Nous avons calculé cet indice à partir des vitesses de vent journalières en se fixant pour seuil les quantiles à 80, 98 et 99%. Ceci revient à retenir successivement pour chaque département les 3700, 370 et 185 vitesses les plus élevées étant donné que notre période d'observation s'étale sur 50 ans. Pour une représentation spatiale des dépendances, il faut pouvoir passer de cet indice à une distance compatible avec l'algorithme des k-médoïdes. Nous proposons

$$d_1(x, y) = 1 - \lambda. \tag{2.2}$$

Au passage, notons que l'article de Naveau et al. [17] utilise une distance très similaire définie pour les maxima annuels afin de regrouper des sites de mesures de précipitations. Cette méthode permet d'obtenir des groupes assez homogènes géographiquement. Cependant certains départements peuvent être très éloignés de leur médoïde, ce qui empêche d'avoir des zones clairement délimitées. Pour cette raison nous avons entrepris de modifier légèrement la distance utilisée en lui rajoutant une composante dépendant de la distance euclidienne géométrique (d_{geo}) entre chaque point considéré. De plus, cette composante permet aussi de tenir compte d'éventuelles différences dans la gestion des collectivités territoriales par rapport aux tempêtes. Pour simplifier les calculs nous utilisons ici la latitude et la longitude de chaque centre de département comme des coordonnées cartésiennes, ce qui revient à négliger la courbure du globe terrestre. Cette approximation ne devrait pas avoir d'influence pour notre étude. les séries de 95 mesures des distances utilisées ont chacune été centrée-réduite afin de les rendre comparables et de les pondérer ensuite. La distance devient :

$$d_2(x, y) = p \times d(x, y) + (1 - p) \times d_{geo}(x, y), \tag{2.3}$$

où p pris entre 0 et 1 permet de pondérer l'influence de cette distance géométrique sur notre distance basée sur les extrêmes. L'inclusion d'un critère de distance géographique permet de compenser (au moins un peu) des problèmes d'anémomètre imprécis. Cette distance d_2 est intégrée à l'algorithme des `k-médoides` [43] pour diviser la France métropolitaine en 6 zones. Les analyses préliminaires que nous avons effectués nous ont conduit à fixer $p = 0.7$. Les résultats de l'algorithme des `k-médoides` pour les trois seuils sélectionnés (0.8, 0.98 et 0.99) sont présentés dans les Figures 2.5, 2.6 et 2.7. Sur les cartes, on peut voir que l'appartenance d'un département à un groupe est plus ou moins marquée en fonction de la taille de son point central et du trait qui le relie au département central (le médoïde). On détermine la solidité de chaque groupe à partir de la *silhouette* qui rassemble les informations sous forme d'un histogramme dont la taille des barres est proportionnelle à l'appartenance au groupe. Pour un département dep donné, sa silhouette $s(\text{dep}) \in [-1, 1]$ est la différence entre $a(\text{dep})$, défini comme la moyenne des distances avec les autres départements de sa zone d'appartenance, et $b(\text{dep})$, défini comme le minimum de cette distance moyenne calculée par rapport à toutes les autres zones. Cette différence est normalisée par le maximum de $a(\text{dep})$ et $b(\text{dep})$; ainsi, $s(\text{dep}) = (a(\text{dep}) - b(\text{dep})/max(a(\text{dep}), b(\text{dep}))$. L'algorithme n'optimise pas directement la silhouette (ce qui serait trop lourd en calcul), rendant possible l'existence de valeurs négatives. La silhouette est donc un bon moyen pour voir si, globalement, l'algorithme réussit à bien délimiter des zones (il ne devrait pas y avoir trop de valeurs négatives, idéalement il n'y en a pas). En face de chaque histogramme on peut lire le nombre de départements qui constitue le groupe.

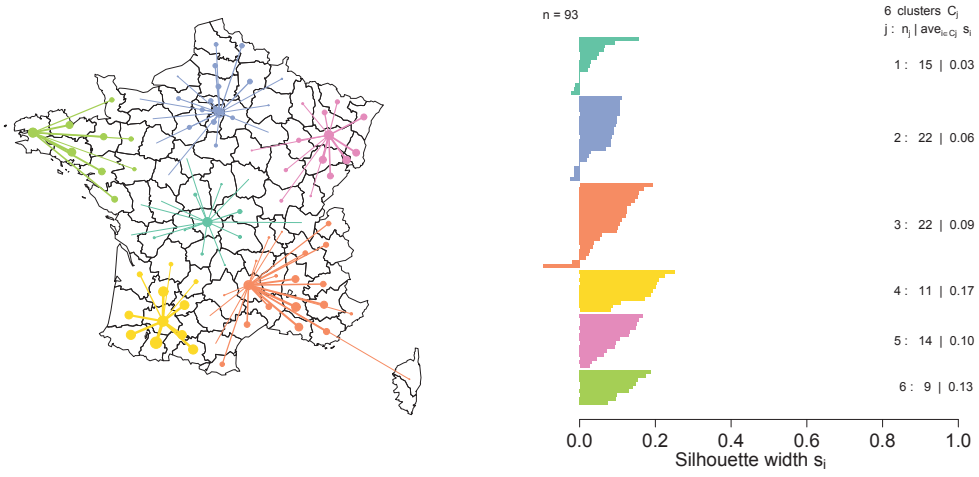


FIGURE 2.5 – Partition et silhouette de la France selon le quantile à 80% des vitesses de vent.

Les premiers découpages apparaissent dans la Figure 2.5. Les groupes 1, 2 et 3 qui comptent respectivement 15 et 22 départements sont les moins solides. En observant

leurs silhouettes dans le détail, on constate des valeurs négatives qui correspondent à des départements non fortement rattachés à leur médoïde. En revanche, les groupes 4, 5 et 6 présentent des silhouettes entièrement positives avec le meilleur résultat pour le groupe 4 (0.17). Sur la carte cette solidité s'exprime à travers la taille des points situés au centre des départements.

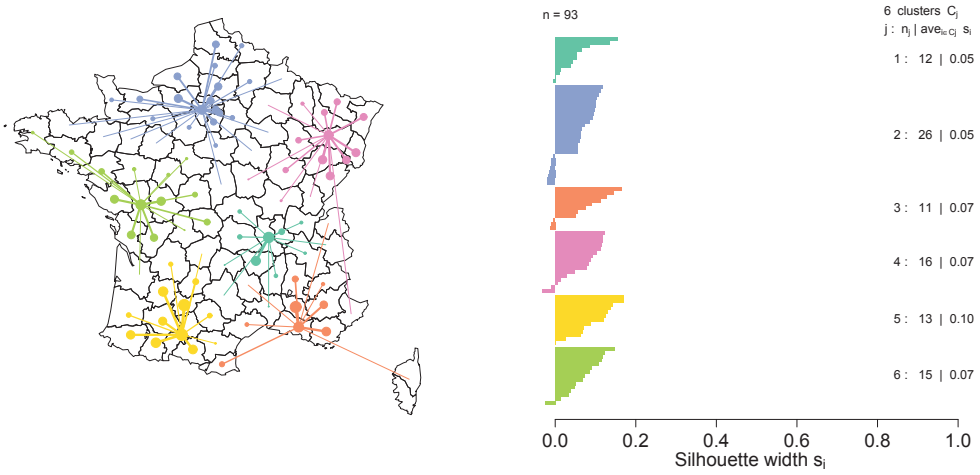


FIGURE 2.6 – Partition et silhouette de la France selon le quantile à 98% des vitesses de vent.

La Figure 2.6 permet de voir l'évolution géographique des 6 zones lorsque le seuil passe de 0.80 à 0.98. On s'intéresse donc plus particulièrement aux vitesses de vent extrêmes et inhabituellement élevées pour chaque département. D'après la silhouette, la solidité moyenne baisse légèrement (de 0.09 à 0.07) et il n'y a plus que le groupe 5 dont l'histogramme soit entièrement positif. Le groupe 2, dans le nord de la France devient le plus important en réunissant 26 départements. La zone 6 s'étale désormais de la Bretagne jusqu'au centre. Dans le sud le découpage est moins homogène avec des départements isolés de leur médoïde.

La dernière carte (Figure 2.7) est celle obtenue en fixant le quantile à dépasser à 99%. La solidité moyenne de la silhouette diminue encore et passe à 0.05. Les groupes ainsi formés sont assez différents des approches précédentes. Les départements le long de la frontière Est de la France se retrouvent réunis avec la Corse. La seule zone entièrement positive est la deuxième dans le Nord-Est avec une silhouette à 0.08. Si le découpage peut sembler moins homogène que précédemment, on peut néanmoins remarquer que les zones obtenues correspondent plutôt bien aux grands bassins versants français. Ces bassins sont délimités par les lignes de partage des eaux et leur connaissance est fondamentale dans l'étude des risques naturels [79]. La vulnérabilité à une tempête dépend à la fois de facteurs structurels, géographiques et conjoncturels. Notre travail sur les zones de risque a pour but de mieux retranscrire ces vulnérabilités.

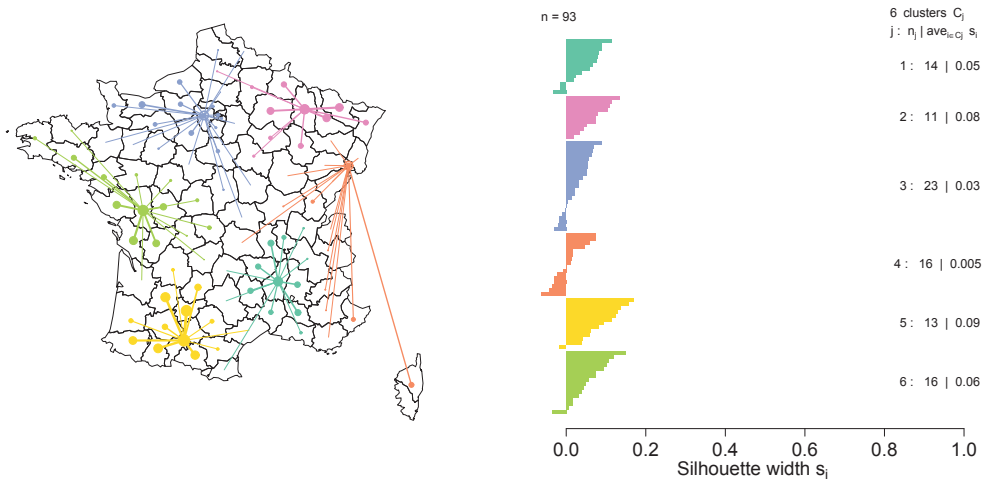


FIGURE 2.7 – Partition et silhouette de la France selon le quantile à 99% des vitesses de vent.

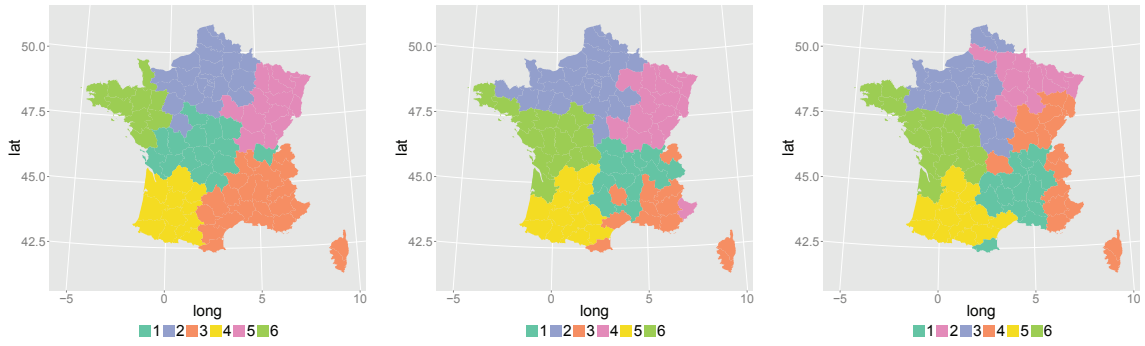


FIGURE 2.8 – Découpage des 6 zones de vent avec d_2 et selon les seuils 0.8, 0.98 et 0.99 (de gauche à droite).

La Figure 2.8 récapitule, les trois découpages différents selon la valeur du quantile de vent choisi. La première carte (à gauche) présente les zones les plus homogènes dans l'ensemble et exprime une sensibilité territoriale à des tempêtes de faible envergure nationale. Lorsqu'on se focalise sur les dépassements les plus extrêmes, le découpage devient moins précis géographiquement mais il est davantage susceptible d'exprimer des possibles trajectoires de tempêtes exceptionnelles. Outre la dépendance des extrêmes et la distance géographique, nous avons alors rajouté un troisième critère : la ressemblance des séries entières de vitesses du vent par département. Nous avons alors construit d_{tot} qui est la norme euclidienne entre les vitesses de vent brutes de deux stations différentes. Notre objectif est d'améliorer la solidité des groupes et ainsi de délimiter au mieux, en combinant ces trois approches, des zones homogènes où la circulation du vent est très sensiblement la même. Les valeurs de ces différentes distances ont été centrées et réduites pour éviter tout problème d'échelle. Nous obtenons finalement la définition suivante pour notre distance :

$$d_3(x,y) = p_1 \times d(x,y) + p_2 \times d_{geo}(x,y) + (1 - p_1 - p_2) \times d_{tot}(x,y), \qquad (2.4)$$

avec $p_1 = p_2 = 0.33$ choisies dans une analyse préliminaire. Cette nouvelle formule est utilisée comme précédemment avec les trois quantiles servant de seuil dans la distance des extrêmes. Nous constatons à chaque fois une amélioration de la solidité des zones de façon globale et aussi dans le détail. Pour q_{80} la solidité moyenne de l'ensemble des groupes atteint 0.16. Le groupes 1 qui compte 21 départements est le moins solide (0.05). En observant sa silhouette dans le détail, on constate des valeurs négatives qui correspondent à des départements non fortement rattachés à leur médoïde. En revanche, les autres groupes présentent des silhouettes majoritairement positives avec le meilleur résultat pour le groupe 2 (0.28) dans le Nord jusqu'en région parisienne. Pour q_{98} le découpage est sensiblement le même. La solidité moyenne baisse de 3 points (de 0.16 à 0.13) et les groupes 2 et 5 restent les plus homogènes. Le groupe 1, dans le Sud de la France demeure le plus important en réunissant 19 départements, mais aussi le moins solide. La zone 3 gagne un département vers le Nord. La zone 4 s'installe plus nettement sur l'Est de la France du Nord au Sud et la zone 6 s'élargit avec deux départements supplémentaires. Enfin, pour q_{99} la solidité moyenne de la silhouette diminue seulement d'un point et passe à 0.12. Les groupes ainsi formés sont assez différents des approches précédentes (et leur numérotation est aussi changée). Vingt-trois départements le long de la frontière Est de la France se retrouvent réunis de l'Alsace à la Corse. C'est la zone autour de la région parisienne qui compte désormais le plus de départements (24). La zone la plus homogène est la quatrième dans le Sud-Ouest avec une silhouette à 0.21, elle prend la même place que la zone 5 dans les précédentes répartitions. La nouvelle zone 5 se retrouve centrée sur la Bretagne et devient la plus petite en terme de superficie. Les trois cartes obtenues apparaissent dans la Figure 2.9.

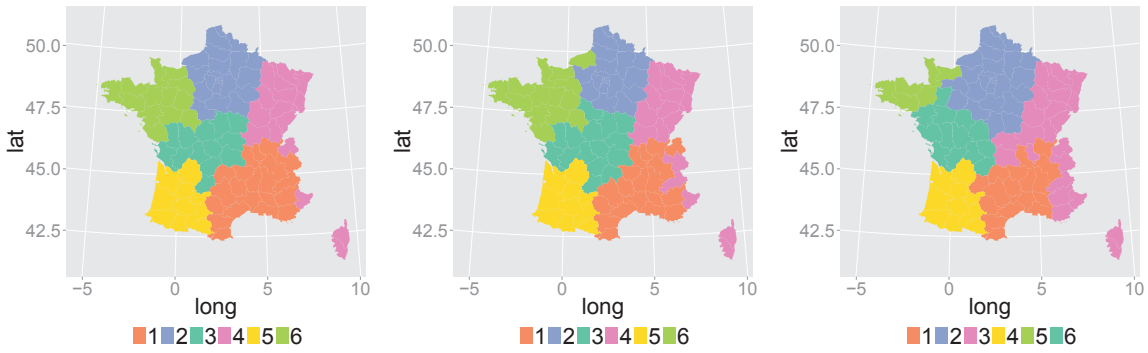


FIGURE 2.9 – Découpage des 6 zones de vent avec d_3 et selon les seuils 0.8, 0.98 et 0.99 (de gauche à droite).

Ces différentes approches nous amènent au concept de la matérialité des périls. Cette matérialité s'exprime à travers les dérives potentielles occasionnées par les dégâts d'une tempête sur une échelle géographique et temporelle. À ce stade, une question intéressante est de savoir à quel point la morphologie des zones est dépendante de la

réalisation de certains événements et en particulier des plus extrêmes. Nous reprenons donc la simulation de l'algorithme k-médoides mais sans les vitesses de vent enregistrées entre le 24 et le 28 décembre 1999 qui sont associées aux tempêtes Lothar et Martin. Nous utilisons la dernière distance définie 2.4 pour délimiter les nouvelles zones.

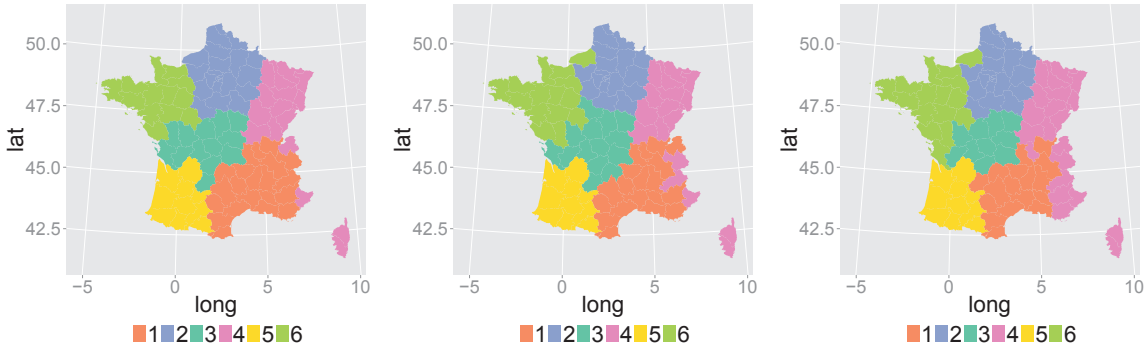


FIGURE 2.10 – Découpage des 6 zones de vent en omettant Lothar selon les seuils 0.8, 0.98 et 0.99 (de gauche à droite).

Sur la Figure 2.10 nous avons représenté les six zones de risque tempête obtenues en omettant les données qui concernent Lothar et Martin. La solidité moyenne reste inchangée (0.16 pour le quantile à 80%, 0.13 pour le quantile à 98% et 0.11 pour le quantile à 99%). C'est la répartition des départements dans les groupes et donc la forme géographique des zones qui est susceptible de changer. Pour les deux premiers seuils (q_{80} et q_{98}) on ne remarque quasiment pas de modifications. Il faut attendre le seuil le plus élevé (q_{99}) pour pouvoir observer des différences. La plus flagrante concerne la zone 6 autour de la Bretagne qui s'élargit nettement au détriment des zones 2 et 3. La zone 4 dans l'Est de la France s'étend toujours du Nord au Sud mais ne s'étale plus jusqu'au centre. Enfin, dans le Sud, les zones 1 et 5 restent stables, ce qui peut sembler normal étant donné qu'elles ont été les moins fortement touchées par le passage des tempêtes Lothar et Martin.

La gestion de l'aléa est ici compliquée par le fait que des éléments extérieurs au champ des variables dont dispose l'assureur, influent sur la sinistralité. Un phénomène naturel se traduit à la fois par un champ d'action (espace), une intensité (impact) et une récurrence (fréquence). Il est possible d'employer la statistique spatiale [34] pour bâtir des simulations de dommages adaptées à chacune des zones de vent précédemment définies. Pour la suite de l'étude, c'est le découpage obtenu avec le quantile à 99% et l'ensemble des vitesses de vent (Figure 2.11) qui sera utilisé.

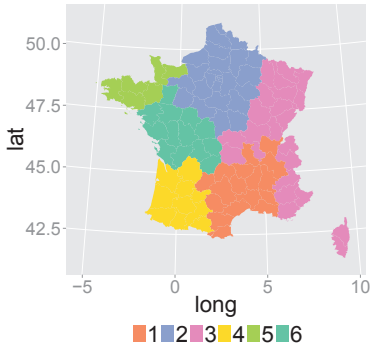


FIGURE 2.11 – Découpage retenu pour les 6 zones de vent.

2.4.2 Répartition du portefeuille sur les zones

Pour comprendre le poids relatif de chaque zone en terme de portefeuille d'assurance, nous présentons leur répartition sous trois angles essentiels. Le nombre de contrats, la somme des charges et la fréquence des sinistres. On s'intéresse à la répartition des dommages et plus particulièrement à la moyenne annuelle de la somme des charges de chaque zone. Pour estimer cette moyenne, nous utilisons la technique du bootstrap en créant 2500 échantillons différents sur la période d'observation. De cette façon, nous pouvons rendre compte de la variabilité de cette moyenne en fonction de la présence ou non de tempêtes majeures comme celles de 1999 au sein des observations. Nous devons cependant noter que dans notre situation avec des lois à queues très lourdes cette méthode non-paramétrique a des défauts. Un petit nombre d'événements à une très forte influence sur les résultats, ce qui empêche la consistance de la méthode en termes statistiques [35]. Cette dernière est néanmoins intéressante car elle repose sur un minimum d'hypothèse de modèles. Nous l'appliquons pour une première étude empirique avant de passer à des modèles plus robustes basés sur la théorie des valeurs extrêmes.

Dans la Table I, nous fournissons d'abord une estimation du nombre de contrats dans chacune des six zones sur une base de 1 million. On retrouve quasiment les mêmes proportions que celles du nombre de départements constituant les groupes. C'est la zone 2 qui comptabilise à la fois le plus de départements et la plus grande part du portefeuille. Nous avons donc une répartition homogène des contrats d'assurances entre les zones, comme c'était déjà le cas sur l'ensemble du territoire. Les lignes suivantes contiennent les quantiles à 2.5% et à 97.5% des charges puis du nombre de sinistres pour chaque zone calculés à partir de l'échantillon bootstrap. On constate que toutes les distributions ont une variance élevée. Les charges estimées sont en général proportionnelles à la taille des groupes, mis à part pour la zone 6 dont le quantile à 97.5% avoisine ceux des zones 2 et 3 (autour de 55 millions d'euros). Les écarts sont encore une fois très importants avec les nombres de sinistres, on passe par exemple d'environ 2800 sinistres par an dans la zone 2 en échelle basse à presque 13000 en échelle haute. Contrairement à ce que nous avions constaté pour les charges, la zone 6 (avec entre 800 et 7100

sinistres) n'apparaît plus comme surexposée, quand on la compare aux autres zones. La fréquence de sinistres semble plus proportionnelle à l'exposition en terme de risque portefeuille. D'une manière globale, on comprend ici que la période d'observation aura toujours une très forte influence sur les résultats d'un modèle incluant des événements extrêmes.

Zones	1	2	3	4	5	6
Nb. Contrats	155475	265222	248957	147899	52033	130413
Ratio	16%	27%	25%	15%	5%	13%
Nb. Dep.	17	25	24	10	6	13
Agrégation par jour						
Charges moyennes (en millions d'euros)						
$q_{2.5}$	4.8	9.2	8.3	2.4	1.8	2.6
$q_{97.5}$	36.1	55.7	52.7	33.9	20.1	48.4
Nombres moyens						
$q_{2.5}$	1539	2806	2689	899	517	821
$q_{97.5}$	6498	12868	11480	7580	3353	7117
Agrégation par semaine						
$q_{2.5}$	3.6	6.3	6.7	1.9	1.4	2.2
$q_{97.5}$	42.3	66.2	57.4	35.3	21.2	51.4

TABLE I – Répartition du nombre de contrats et de la distribution des moyennes par zone. Les quantiles ont été obtenus à l'aide d'un bootstrap des charges agrégées par jour ou par semaine.

Dans la Figure 2.12, on peut comparer les statistiques de ce bootstrap sur les zones 2 (en haut) et 6 (en bas). La zone 2 s'étend de la région parisienne au nord de la France. L'histogramme de gauche représente les estimations de la charge moyenne annuelle exprimée en millions d'euros. Cette répartition prend la forme d'une loi asymétrique avec une queue de distribution plus étalée à droite. Elle varie beaucoup selon la sélection et s'étale entre 10 et 90 millions d'euros avec un mode à 30 millions. Le graphique de droite est un QQplot associant les quantiles empiriques à ceux d'une loi normale centrée réduite. L'allure fortement non linéaire indique des queues plus lourdes dans les données, ce qui correspond bien à l'allure de l'histogramme. Dans les Figures du bas, nous représentons à nouveau les statistiques du bootstrap mais cette fois dans la zone 6. Cette zone est moins large que la zone 2, elle ne compte en effet que 13 départements contre 25 dans cette dernière. L'allure de la distribution est sensiblement différente. On observe plusieurs pics, comme si la distribution des charges ne répondait plus à une loi unique mais à un mélange de lois à supports disjoints. Cette zone illustre en fait les défauts de la méthode employée ici. La périodicité dans cette répartition est certainement due à un seul très grand événement, et les pics apparaissent pour les cas où l'événement est tiré 0, 1, 2, 3 fois ou plus lors du tirage. Par exemple, la théorie du bootstrap nous dit qu'un événement particulier ne se trouve pas dans un échantillon bootstrap avec une probabilité d'à peu près $\exp(-1) = .37$ lorsque n (le nombre d'événements au total) est grand. On perçoit ici l'importance de l'échelle géographique à laquelle on travaille en plus de l'échelle temporelle dans la stabilité des résultats.

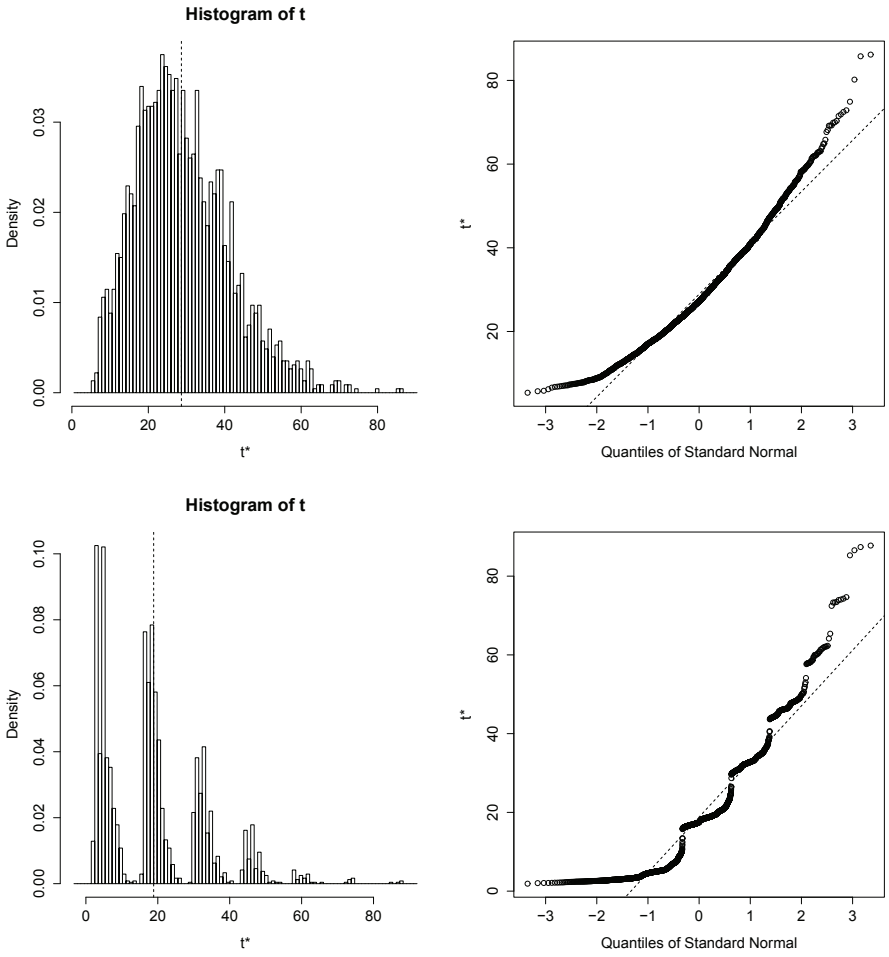


FIGURE 2.12 – Statistique du bootstrap des charges sur les zone 2 (en haut) et 6 (en bas).

Cette approche par rééchantillonnage est basée sur les données journalières. Elle ne retranscrit pas la notion d'événement et elle néglige les dépendances entre jours. Nous proposons donc de confronter nos résultats avec un bootstrap des charges agrégées sur une semaine pour mieux capter les événements. Dans la dernière partie de la Table I, nous présentons les quantiles 2.5% et à 97.5% des charges en moyenne annuelle et en millions d'euros. Nous observons une plus forte variabilité de la distribution de la moyenne due à un regroupement des évènements en clusters de plusieurs jours. Les estimations pour le quantile de début de distribution (2.5%) sont plus faibles que celles de l'approche journalière. Par exemple, dans la zone 2 on passe de 9.2 à 6.3 millions soit une baisse de plus de 30%. En revanche, pour le quantile de fin de distribution, les estimations basées sur les données hebdomadaires sont supérieures à celles des données quotidiennes. Toujours dans la zone 2, on passe de 55.7 à 66.2 millions d'euros, soit une hausse de presque 19%.

Dans la Figure 2.13, on observe à nouveau la répartition des charges annuelles pour les zones 2 et 6 mais cette fois à partir d'une agrégation par semaine. Les défauts remarqués précédemment apparaissent amplifiés. La sinistralité varie selon la période d'observation. Nous pouvons finalement conclure de cette approche qu'il nous faut des méthodes plus raffinées, comme un modèle paramétrique (GPD), tenant compte des clusters.

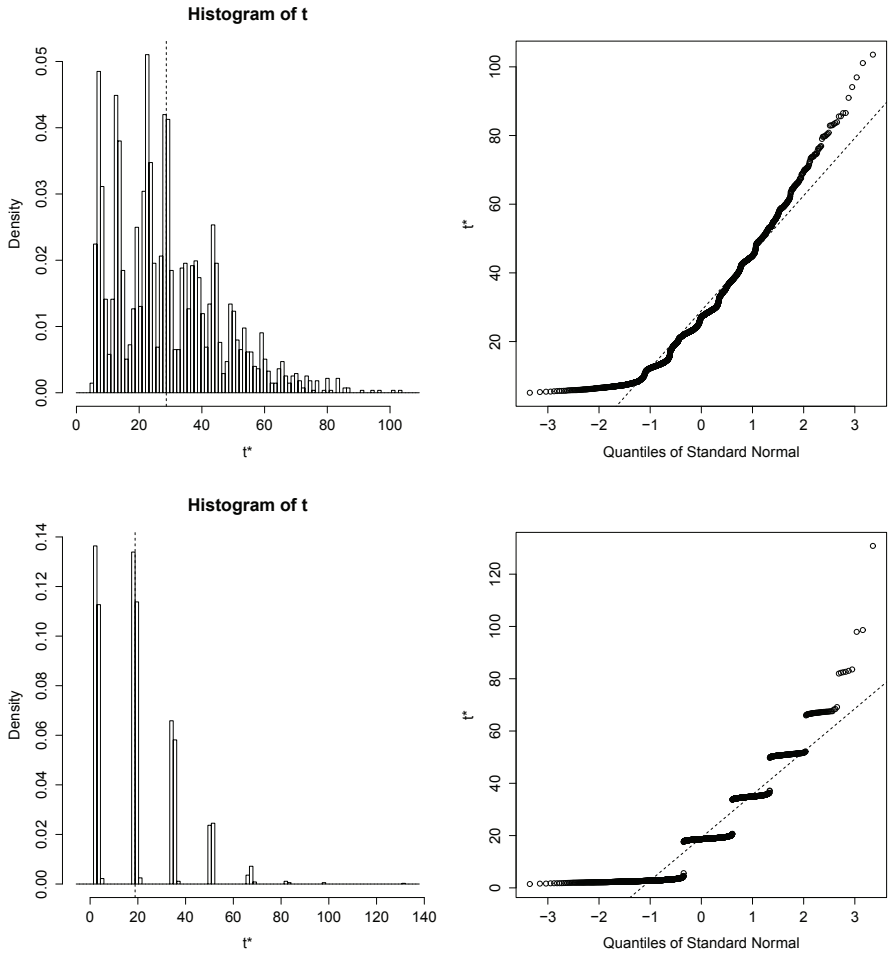


FIGURE 2.13 – Statistique du bootstrap des charges avec clusters semaine sur les zone 2 et 6.

2.4.3 Représentation spatiale des principales tempêtes

Notre première approche consiste à représenter sur la carte de France la répartition des sinistres lors d'événements majeurs en terme de charges et de les comparer avec les variables climatiques à notre disposition. Les différents essais que nous avons pu faire concernaient les températures, les vitesses de vent, les précipitations et les oscillations de l'Atlantique nord (NAO). À partir de ces données nous avons constaté que seules les vitesses de vent permettent de matérialiser géographiquement la trajectoire d'une tempête. Les indices de vent apportent davantage d'informations car ils permettent de repérer les régions où une force de vent inhabituellement élevée s'est produite. Les dégâts occasionnés sont en effet à la fois liés à la force de la tempête et à l'état des bâtiments qui, selon les régions, bénéficient d'une architecture ou d'une réalisation plus ou moins résistante aux intempéries. La formule de l'indice de vent utilisée ici est celle définie dans l'article [57]. Elle se base sur les maximums de vitesses de vent journaliers $w^d(s)$ pour une séquence de jours $d \in D$ à la station s. Nous proposons la définition suivante :

$$I_w^d(s) = \left([w^d(s) - w_q(s)]_+\right)^\alpha, \tag{2.5}$$

où w_q est le q%-quantile de $w_d(s)$ pour $d \in D$. L'utilisation des dépassements d'un quantile dans la construction d'un indice de risque permet de se focaliser sur les valeurs les plus extrêmes. Cette méthode est assez largement répandue [53]. On choisit donc de représenter par départements et sur une période de 3 jours autour de l'événement observé, la somme des charges, les vitesse maximales de vent et l'indice de vent.

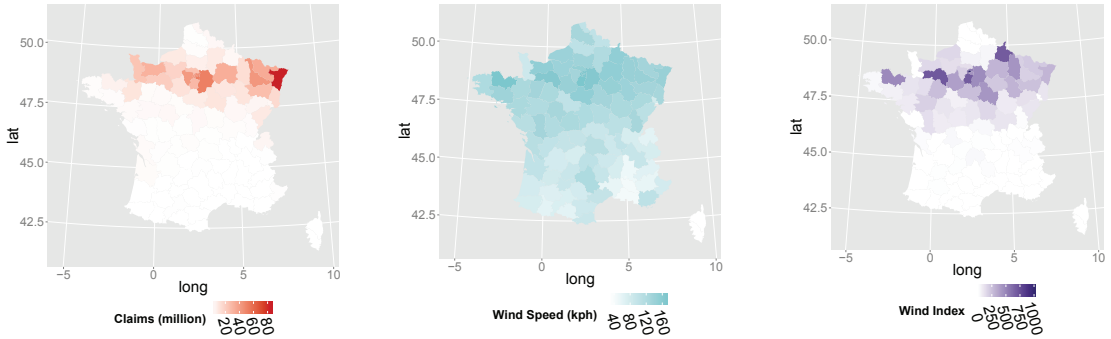


FIGURE 2.14 – Tempête Lothar du 26/12/1999 en France en terme de charges, de vitesse du vent puis d'indice de vent.

Sur le graphique de gauche de la Figure 2.14 (représentant les charges actualisées en millions d'euros) on retrouve la trajectoire de Lothar qui a balayé la France d'Ouest en Est sur un axe Bretagne (vers 4h) - Lorraine et Alsace (11h) avec un front mesurant 150 km de large. Cette tempête n'était pas un ouragan (cyclone tropical), bien que ce nom lui soit donné par certains, mais une dépression des latitudes moyennes exceptionnellement intense pour l'Europe. Le graphique du milieu montre les vitesses de vent maximales enregistrées entre le 24 et le 26 décembre. On observe dans les zones les plus touchées d'un point de vue assurance des vitesses de vent parmi les plus fortes mais cette information ne suffit pas à localiser clairement les zones sinistrées. Le gra-

phique de droite montre, lui, les variations de l'indice de vent. On retrouve alors assez nettement la trajectoire de la tempête dans le Nord de la France. Les départements les plus touchés ne sont pas exactement les mêmes, mais cela s'explique par l'absence de la notion d'exposition au risque (le nombre de contrat) dans la formule de notre indice. Ce paramètre apparaît plus tard dans la construction de l'indice tempête.

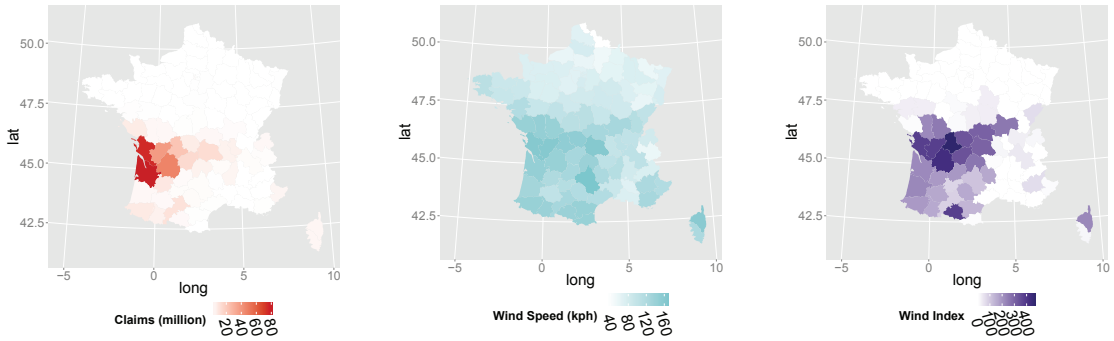


FIGURE 2.15 – Tempête Martin du 27/12/1999 en France en terme de charges, de vitesse du vent puis d'indice de vent.

La trajectoire de Martin (Figure 2.15) a principalement touché une large moitié sud de la France. Ici aussi les zones sinistrées correspondent à de fortes vitesses de vent. L'indice de vent permet d'éliminer des zones peu touchées par la tempête en termes de dégâts d'assurance comme le sud-est et la Bretagne. En revanche un fort indice n'entraîne pas toujours de forts dégâts pour Allianz comme on peut le voir dans les départements du centre de la France mais ceci est dû à une plus faible densité d'habitations et donc de risques dans cette zone.

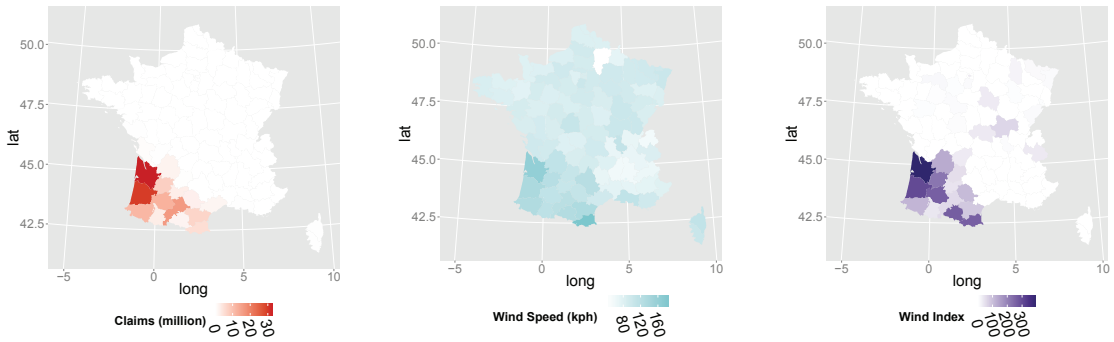


FIGURE 2.16 – Tempête Klaus du 24/01/2009 en France en terme de charges, de vitesse du vent puis d'indice de vent.

La trajectoire de Klaus (Figure 2.16) a principalement touché le sud-ouest de la France (les régions Aquitaine, Midi-Pyrénées et en partie le Languedoc-Roussillon et le Poitou-Charentes). Les vitesses de vent sont assez élevées dans le sud-ouest mais l'ensemble de la France n'a pas été épargné par le vent. L'indice de vent permet à nouveau de recentrer les zones les plus touchées sur le coeur de la tempête.

La tempête Xynthia (Figure 2.17) a principalement touché l'Espagne, le Portugal, la France (Aquitaine, Poitou-Charentes, Pays de la Loire, Bretagne et Normandie), la Belgique, le Luxembourg, l'Allemagne et dans une moindre mesure, le Royaume-Uni, la Scandinavie et les pays bordant la mer Baltique. Si les vitesses de vent permettent de retrouver les zones les plus sinistrées, l'indice de vent délimite plus clairement le couloir emprunté par la tempête, avec une intensité un peu trop faible dans les Pyrénées Atlantique.

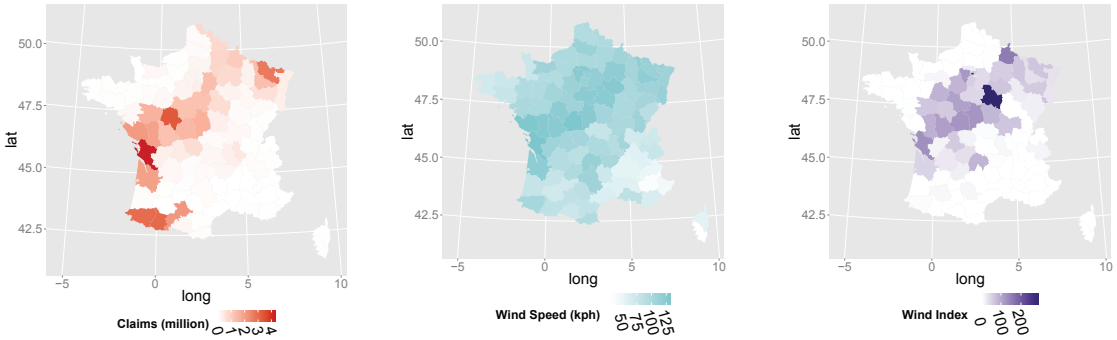


FIGURE 2.17 – Tempête Xynthia du 28/02/2010 en France en terme de charges, de vitesse du vent puis d'indice de vent.

Comme on peut le constater à travers les représentations de ces quatre tempêtes, la vitesse du vent permet d'expliquer une bonne part des dégâts, mais elle ne suffit pas à décrire dans sa totalité un phénomène aussi complexe qu'une tempête. Cette complexité justifie qu'il faille à chaque fois que possible accompagner les estimations d'une discussion sur les incertitudes, en particulier pour les évènements extrêmes [81]. Le vent trouve son origine dans la différence de pression atmosphérique. On identifie deux types de vent. Le premier est appelé vent synoptique, il est engendré par des phénomènes d'échelle continentale ou planétaire. Il est durable, étendu et parfois violent. Le deuxième se situe à l'échelle locale. Il est dû aux inégalités de températures, ce sont les brises thermiques. Pour comprendre les dégâts occasionnés par le vent, il est important de distinguer ces deux types. En effet, le comportement et l'évolution du flux de vent seront différents selon son origine. Le vent synoptique sera généralement freiné par un obstacle (naturel ou artificiel) alors que les brises thermiques se verront renforcées par le relief. Ces deux cas sont illustrés respectivement à gauche et à droite de la Figure 2.18.

La direction du vent a aussi son importance, combinée à la vitesse, elles forment la vélocité. La connaissance de ce paramètre comme de la durée des rafales pourrait être des compléments permettant une meilleure compréhension des tempêtes.

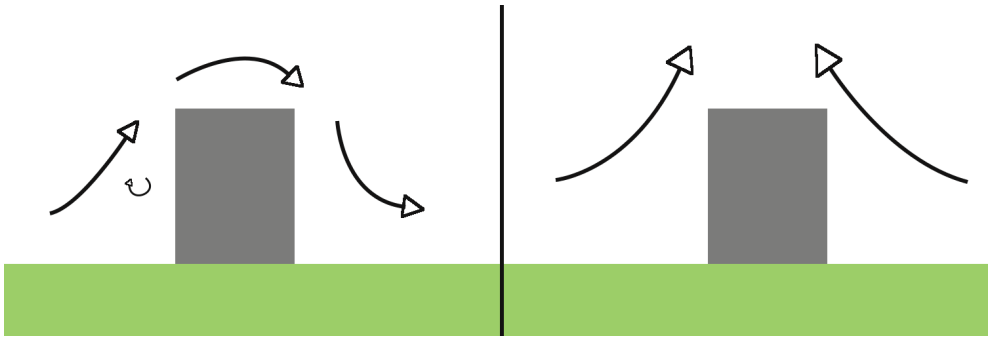


FIGURE 2.18 – Deux types de comportements du vent (synoptique à gauche, thermique à droite).

2.5 Modélisations

2.5.1 Définition d'un indice tempête

Construire un modèle à partir de données concernant des événements extrêmes nous a amené à nous poser plusieurs questions préliminaires sur le choix des variables de départ et sur leurs mises en forme. Il est en effet important de comprendre ici que les mêmes données brutes peuvent conduire à des résultats très différents selon leur interprétation. Dans notre cas, plusieurs approches peuvent être envisagées pour décrire l'impact d'une tempête dans l'espace et dans le temps. Principalement, nous allons nous intéresser ici à trois critères de recherche et à leurs influences sur le modèle. Le premier est la longueur de l'historique des observations. Le deuxième dépend du clustering : en quoi le fait de travailler à une échelle journalière ou au contraire de regrouper les informations en clusters de quelques jours ou d'une semaine change t-il notre perception de la fréquence des tempêtes ? Y a-t-il une durée plus appropriée que les autres à la compréhension de ce phénomène météorologique ? Enfin, le troisième critère est lié à l'existence et à la connaissance de l'événement considéré comme le plus extrême de la période. Nous voulons tester les variations de la distribution modélisée et plus particulièrement de la période de retour en fonction de la présence ou non de cyclones tels que Lothar et Martin dans nos données.

Dans cette section, on propose de projeter une version actualisée de l'indice tempête que nous avions défini dans un précédent article [57]. Cet indice peut être étudié sur une période plus longue que les données des dégâts et il présente moins de non-stationnarités. Il permet d'exploiter une plus fine résolution spatiale. De plus, cette approche est utile pour faire une extrapolation des valeurs de l'indice au delà des valeurs observées. Les données complémentaires à notre disposition nous ont permis d'affiner notre approche et de la rendre plus flexible. Dans chacune des six zones $A(k)$ précédemment délimitées, nous construisons un indice tempête spécifique associant les indices de vent à l'exposition de chaque département en terme de risque portefeuilles.

Pour $k \in \{1, ..., 6\}$ nous définissons $I_S(k)$ comme

$$I_S(k) = \sum_{s \in A(k)} R(s) \times \max_{d \in E} \left(\frac{I_w^d(s)}{N^d} \right), \tag{2.6}$$

où dans chaque station $I_w^d(s)$ est l'indice de vent à la date d et le nombre de risques $R(s)$ est pondéré par le nombre de contrats. Nous tenons également compte de la taille de la zone touchée, de la durée de l'évènement tempête E[2] et le nombre de stations actives à la date d, N^d. Notons ici un changement de méthode sur le choix du maximum et non plus de la somme des indices sur le cluster temporel de chaque tempête. Contrairement aux charges d'assurance qui sont additionnées sur la durée de l'événement, nous avons constaté que les pics de vent reflètent mieux l'intensité relative des phénomènes climatiques. On obtient alors un indice tempête global sur l'ensemble de la France en agglomérant les six indices tempête pondérés à la fois par le portefeuille et par un paramètre $B = \{\beta_1, ..., \beta_6\}$. Ce paramètre supplémentaire permet de sur ou sous-pondérer chacune des zones selon l'impact en terme de dommage aux biens qu'aurait le passage d'une tempête sur son territoire. Pour notre étude, les β_i peuvent varier entre 0.1 et 2 et indépendants. La formule devient

$$I_S = \sum_{k=1}^{6} R(k) \times I_S(k) \times \beta_k, \tag{2.7}$$

avec $R(k)$ le poids relatif de chaque zone dans le portefeuille Allianz. Nous avons optimisé cet indice pour qu'il reflète au mieux la répartition des dommages les plus extrêmes en terme de rang. Notre classement de référence pour les tempêtes les plus importantes depuis les années 70 est issu des travaux de M. Luzi [51]. Nous avons déduit de ce classement un dernier ajustement au niveau des écarts entre les valeurs de l'indice pour qu'elles soient proportionnelles aux charges actualisées des dommages. Pour distendre la distribution, nous employons la fonction exponentielle et un dernier paramètre : γ estimé avec une méthode des moindres carrés. Le lien entre notre indice tempête et le coût C pour l'assureur est le suivant :

$$C = \exp(\gamma \times I_S). \tag{2.8}$$

2.5.2 Modélisation statistique de l'indice tempête

La théorie des valeurs extrêmes [28] s'est développée autour de deux approches complémentaires, le maximum par bloc et le dépassement de seuil. On utilise ici la méthode du dépassement de seuil via une modélisation par la loi de Pareto généralisée (GPD). Avec ce modèle et pour un seuil adapté, u, la fonction de distribution de $Z = (X - u)$, sachant que $X > u$, est définie par :

$$H(z) = 1 - \exp\left(-\frac{z}{\sigma}\right), \xi = 0,$$

2. Pour chaque évènement $E \subset D$ nous sélectionnons le maximum journalier, par exemple, pour Klaus : $E = \{01/23/09, 01/24/09, 01/25/09\}$.

ou

$$H(z) = 1 - \left(1 + \frac{\xi z}{\sigma}\right)^{\frac{-1}{\xi}}, \xi \neq 0,$$

selon la valeur du paramètre de forme ξ qui contrôle la queue de distribution. Nous avons utilisé le package `extRemes` [32] du logiciel `R` pour calculer les différentes distributions des extrêmes de cette section. Nous avons opté pour deux périodes d'observations différentes. La première étalée sur vingt ans (1993/2013) et la seconde sur plus de quarante ans (1970/2013). Ce choix a été fait selon les données les plus homogènes à notre disposition. Avec les historiques provenant de l'assurance, il s'avère difficile de remonter sur des périodes aussi longues. Nous travaillons avec l'indice tempête I_S. Nous commençons par déterminer le seuil à dépasser. `extRemes` permet la sélection d'un seuil selon deux méthodes. La première consiste à simuler plusieurs fois la distribution avec des seuils différents et à estimer la solidité des paramètres de forme et d'échelle. La seconde représente les dépassements moyens au delà du seuil avec leurs intervalles de confiance.

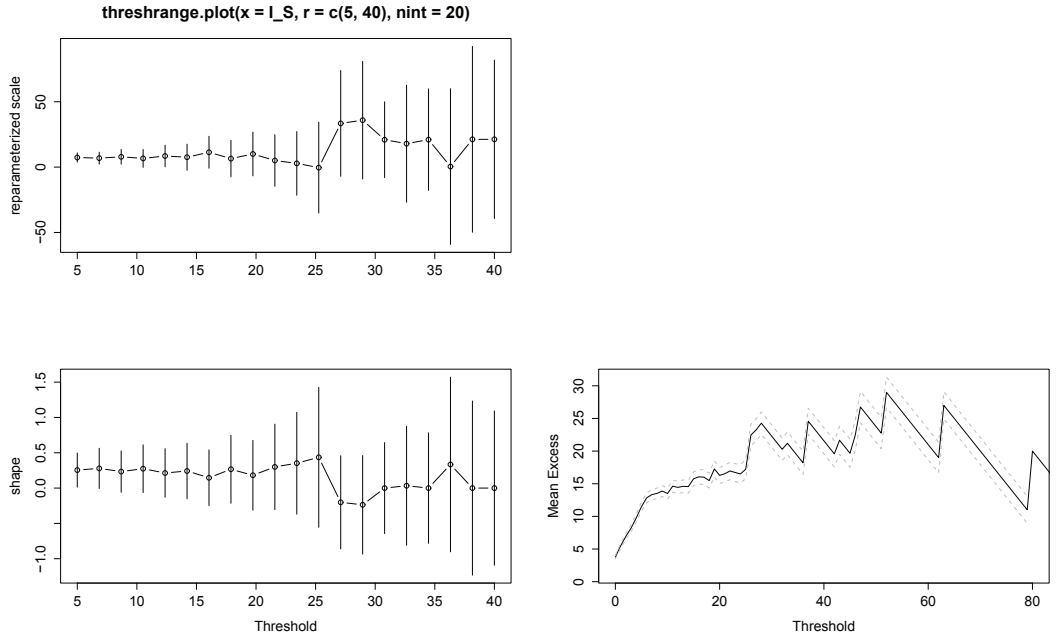


FIGURE 2.19 – Graphique de sélection de seuils de notre indice tempête.

La Figure 2.19 regroupe ces deux approches de la recherche de seuils. Les deux graphiques de gauche représentent les variations de σ et ξ avec leur intervalle de confiance à 95% selon les valeurs de seuils testées. Elles sont comprises entre $u = 5$ et $u = 40$, sachant que l'indice est calibré pour que son maximum observé (Lothar en décembre 1999) soit 100. En observant les intervalles de confiance qui s'élargissent, on peut se faire une idée de la solidité du modèle et surtout du seuil maximum à ne pas dépasser. Dans notre cas, la valeur $u = 20$ semble être la limite avant que la variabilité des estimations des paramètres ne deviennent trop importante. Le graphique de droite nous

montre les dépassements moyens avec en pointillés l'intervalle de confiance à 95% pour une plage de seuils allant de 0 à 80. Ici aussi la courbe devient erratique autour de $u = 20$ ce qui confirme notre première impression. Par conséquent nous avons retenu les seuils $u = 10$ et $u = 20$ comme base de nos modèles.

Ensuite, on propose une analyse de sensibilité par rapport à Lothar et Martin, les deux événements les plus importants de notre période d'observation. Notre méthode consiste ici à observer comment changent les temps de retour si l'on enlève ou inclut Lothar et Martin dans les données. Nous voulons aussi comprendre les fragilités d'un modèle mesurant des phénomènes de périodes de retour de plus de 70 ans, voire 200 ans avec un historique de 50 ans maximum.

2.5.3 u = 10 et période = 1970-2013

Nous commençons par estimer les paramètres de la loi de Pareto généralisée avec un seuil $u = 10$ et sur la période 1970-2013. Ce seuil est dépassé à 62 reprises soit à environ 1.4 fois par an sur la période d'observation. Les paramètres sont estimés selon la méthode maximum de vraisemblance (MLE). La Figure 2.20 regroupe le QQ plot des quantiles des données par rapport aux quantiles du modèle, une comparaison des quantiles d'une simulation du modèle par rapport aux données, la densité empirique associée à la densité du modèle et un graphique de la période de retour.

	MLE	Std. Err.
Scale (σ):	10.78	2.13
Shape (ξ):	0.20	0.15
Negative log-likelihood : 222.08		

TABLE II – Estimations du modèle : I_S.

Le paramètre ξ est positif (0.20 dans la Table II) ce qui correspond à une queue lourde et une fonction de répartition appartenant au domaine d'attraction de Fréchet. Les deux graphiques du haut sont des QQ plot. Sur celui de gauche, on note un léger écart dans les fortes valeurs de quantiles mais un bon alignement global et sur celui de droite, l'alignement des points est très bon pour la comparaison des probabilités du modèle avec les probabilités empiriques. La fonction de répartition du modèle est confrontée à une estimation empirique à noyau sur le graphique en bas à gauche. On note un écart au départ mais une assez bonne correspondance sur la queue de distribution, ce qui importe le plus étant donné que notre étude se focalise sur les extrêmes. En termes de niveaux de retour, l'ensemble des points appartiennent bien à l'intervalle de confiance à 95% qui est matérialisé par les pointillés gris. Nous représentons l'axe des abscisses à l'échelle logarithmique pour pouvoir discerner le type de distribution des extrêmes à partir de sa forme (les distributions à queues lourdes sont concaves, les queues légères sont en ligne droite et les distributions admettant une borne supérieure sont convexes avec une asymptote le long de cette borne supérieure). Dans notre cas,

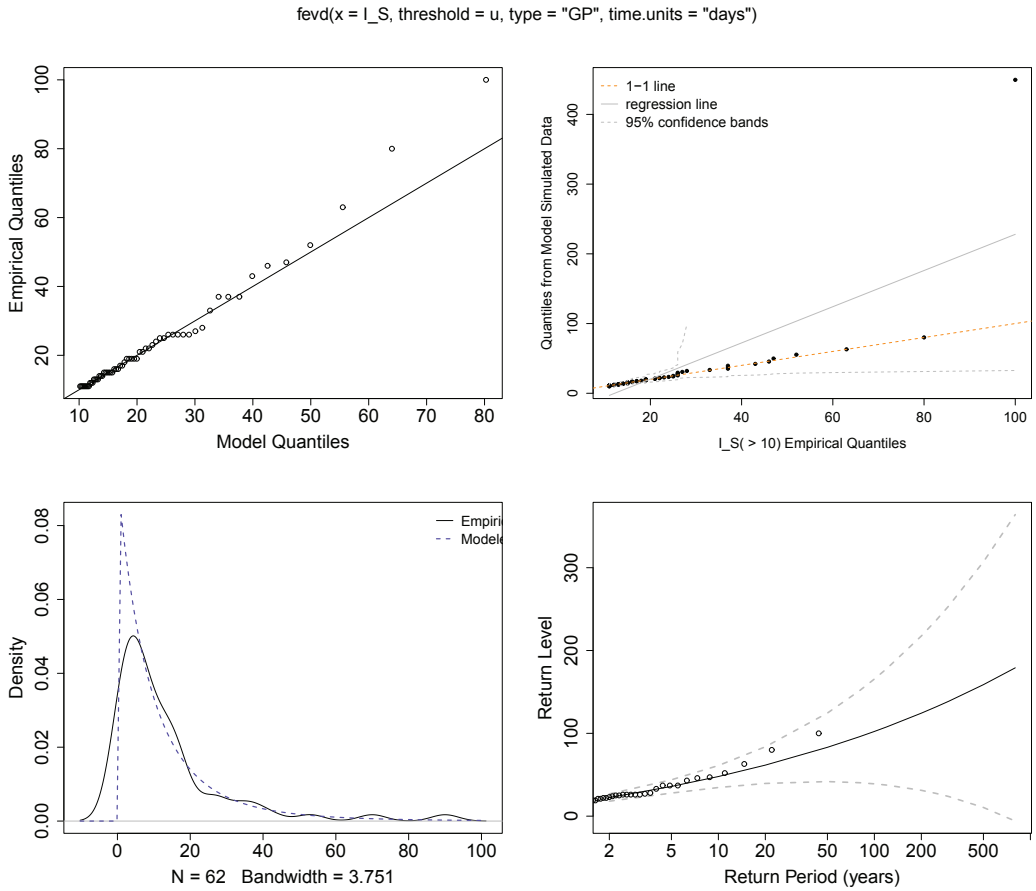


FIGURE 2.20 – Résultats de l'analyse GPD pour l'indice tempête.

la distribution principale ainsi que sa borne de confiance inférieure indiquent le domaine d'attraction borné de Weibull, alors que la borne supérieure tendrait vers une répartition à queue lourde. Si on regarde plus en détails ce dernier graphique (Figure 2.21), on obtient une tempête de l'importance de Lothar environ tous les 93 ans. Ce niveau de retour apparaît en pointillés rouge. Cependant au delà de notre période d'observation d'une quarantaine d'années, l'incertitude devient importante et la limite basse de l'intervalle de confiance commence à décroître ce qui pose un problème d'interprétation.

2.5.4 u = 10 et période = 1993-2013

Toujours avec le même seuil, on veut maintenant comparer les résultats précédents avec ceux obtenus sur une période deux fois plus courte, vingt ans entre 1993 et 2013. Sur cette période, le seuil n'est dépassé que 20 fois, soit une fois par an en moyenne. L'estimation des paramètres apparaît dans la Table III. Les erreurs standards sont plus importantes.

Le paramètre d'échelle σ a diminué (9.87 pour 10.78 précédemment). Le paramètre ξ a augmenté et atteint 0.50, nous avons donc toujours à faire à une répartition à

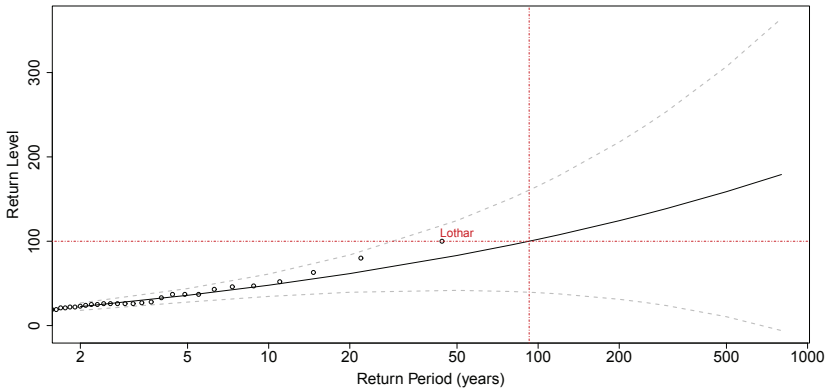


FIGURE 2.21 – Période de retour pour l'indice tempête.

	MLE	Std. Err.
Scale (σ):	9.87	4.18
Shape (ξ):	0.50	0.38
Negative log-likelihood : 75.76		

TABLE III – Estimations du modèle : I_S.

queue lourde. La période de retour d'un événement de type Lothar est fortement revue à la baisse comme on peut le constater sur la Figure 2.22. On estime désormais qu'une tempête de niveau 100 sur notre indice pourrait se produire tous les 31 ans. On constate ici que la durée de la période d'observation à une très forte influence sur le calcul de la période de retour. Avec un historique plus réduit, on a tendance à surestimer la probabilité qu'un tel événement se réitère. Nous nous concentrerons ainsi pour la suite de cette étude sur la période la plus longue à notre disposition : 1970-2013.

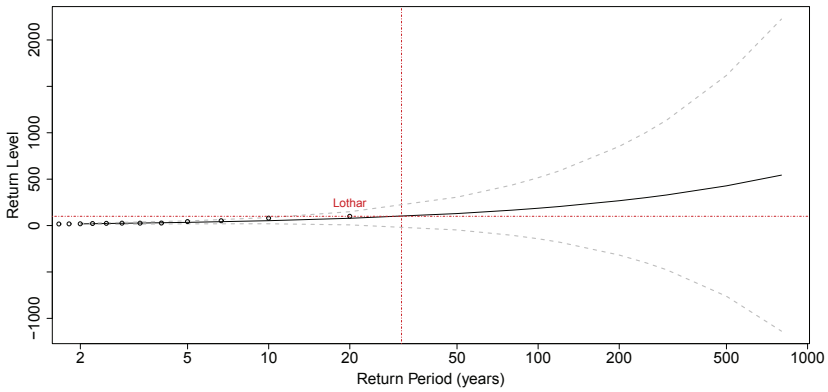


FIGURE 2.22 – Période de retour pour l'indice tempête.

2.5.5 u = 20 et période = 1970-2013

Le test suivant se fait sur la période 1970-2013 mais avec un nouveau seuil plus élevé : $u = 20$. Ce seuil est dépassé par notre indice tempête à 26 reprises, soit 0.6 fois en moyenne par an. Les estimations des paramètres du modèle sont retranscrites dans la Table IV.

	MLE	Std. Err.
Scale (σ):	12.96	4.20
Shape (ξ):	0.21	0.26
Negative log-likelihood : 98.09		

TABLE IV – Estimations du modèle : I_S.

Le paramètre d'échelle σ a augmenté (12.96 pour 10.78 au départ). Le paramètre de forme ξ vaut 0.21 soit une valeur très proche de celle obtenue avec le seuil $u = 10$ sur la même période. En revanche, l'erreur standard est plus importante et recouvre même la valeur 0 synonyme d'une queue légère de distribution. En termes de niveaux de retour, (Figure 2.23), on obtient cette fois ci une tempête de l'importance de Lothar tous les 88 ans. Cette prévision est légèrement plus faible que celle du premier modèle mais reste dans le même ordre de grandeur.

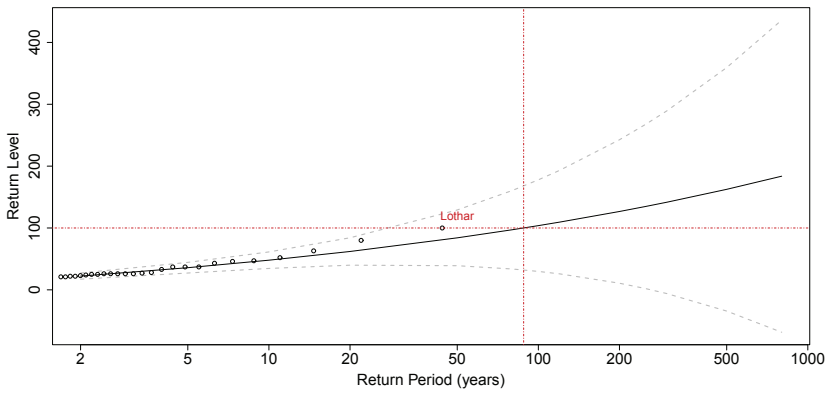


FIGURE 2.23 – Période de retour pour l'indice tempête.

En modifiant le seuil u du modèle entre 10 et 20, la période de retour de l'événement majeur de notre période d'observation (Lothar) baisse seulement de 1.06% en passant de 93 à 88 ans. Pour tester davantage la sensibilité des résultats, on propose dans les sections suivantes de reproduire les mêmes estimations en supprimant Lothar de la base de données.

2.5.6 u = 10 sans Lothar et période = 1970-2013

Nous reprenons le modèle de base ($u = 10$) et sur la période la plus longue (1970-2013) mais nous voulons simuler une situation dans laquelle la tempête Lothar n'aurait

pas eu lieu. Autrement dit, le sinistre le plus important depuis les années 70 n'aurait pas dépassé le niveau de la tempête Martin (deux fois inférieure à Lothar en terme de dommages enregistrés par les compagnies d'assurance en France). Les estimations des paramètres du modèle sont retranscrites dans la Table V.

	MLE	Std. Err.
Scale (σ):	11.02	2.15
Shape (ξ):	0.10	0.15
Negative log-likelihood : 213.54		

TABLE V – Estimations du modèle : I_S

Le paramètre d'échelle σ a augmenté (11.02 pour 10.78 au départ). Le paramètre ξ est encore positif (0.10) mais diminue par rapport au modèle avec Lothar (0.20). Cette diminution s'explique assez naturellement car ayant supprimé la plus forte valeur la queue de distribution devient moins lourde. L'erreur standard est plus importante proportionnellement à la valeur de ξ. En ce qui concerne les périodes de retour du modèle (Figure 2.24), on obtient une tempête de l'importance de Lothar environ tous les 280 ans. Sans avoir eu connaissance de Lothar on multiplie par 3 la période de retour d'un événement d'une ampleur comparable. Le changement est ici considérable. Nous ne pouvons que reconnaître avec cet exemple la fragilité de ce type de modélisation dont les prédictions changent radicalement lorsqu'un nouveau record est enregistré.

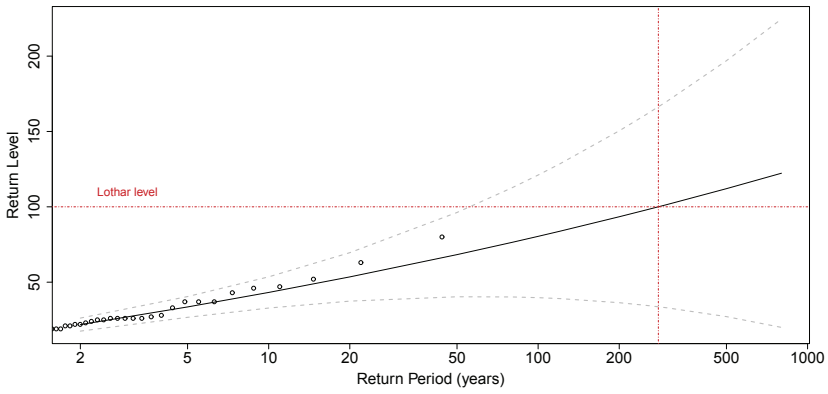


FIGURE 2.24 – Période de retour pour l'indice tempête sans Lothar.

2.5.7 u = 10 sans Martin et période = 1970-2013

Toujours sur le même principe que précédemment, nous proposons à cette étape de reconsidérer la même situation mais en l'absence de la deuxième plus importante tempête à notre connaissance, c'est à dire Martin. Les estimations des paramètres du modèle sont retranscrites dans la Table VI.

Le paramètre d'échelle σ diminue légèrement (10.61 pour 11.02 précédemment). Le paramètre ξ est positif (0.16) et diminue à peine par rapport au modèle complet

	MLE	Std. Err.
Scale (σ):	10.61	2.04
Shape (ξ):	0.16	0.14
Negative log-likelihood : 214.60		

TABLE VI – Estimations du modèle : I_S.

(0.20). En termes de niveaux de retour, (Figure 2.25), on obtient une tempête de l'importance de Lothar tous les 160 ans. Sans avoir eu connaissance de Martin on obtient une période de retour 1.7 fois plus importante pour Lothar. Dans cette autre configuration des événements, le modèle est aussi fortement impacté par l'omission d'une tempête majeure. Le changement d'échelle n'est pas aussi important qu'avec l'omission de Lothar mais il demeure conséquent.

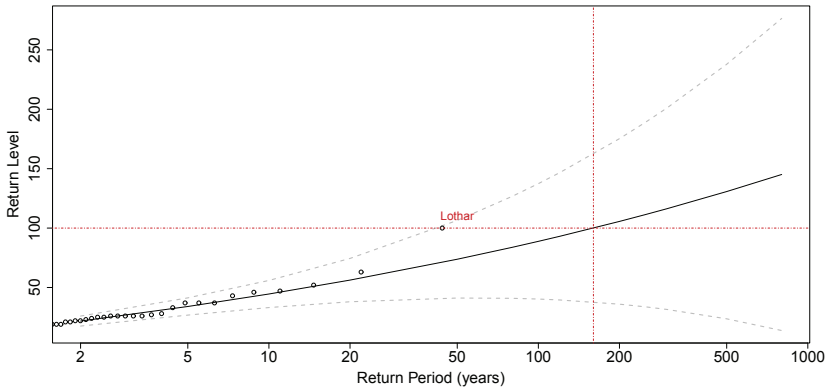


FIGURE 2.25 – Période de retour pour l'indice tempête sans Martin.

2.5.8 Récapitulatif et complément

Le modèle GPD dépend fortement de la valeur la plus extrême pour l'estimation du paramètre ξ qui contrôle la queue de distribution. L'estimation de ξ diminue en l'absence de l'indice de Lothar et le modèle à queue lourde se rapproche d'un modèle à queue légère (ξ =0) si on considère l'erreur standard. En revanche le paramètre d'échelle σ reste stable. Par conséquent les périodes de retour d'événements extrêmes augmentent fortement à leur tour. La période d'observation joue elle aussi un rôle déterminant, nous avons pu constater qu'une durée trop faible pour entraîner une forte sous-estimation des périodes de retour. Pour une tempête de l'importance de Lothar, cette période varie ainsi entre un minimum de 31 ans et un maximum de 280 ans. La suppression d'un événement intermédiaire important comme Martin perturbe aussi les résultats du modèle dans le sens d'une augmentation de la période de retour. Dans la Table VII nous proposons un récapitulatif des variations de la période de retour d'un événement du niveau de Lothar selon les méthodes et les scénarios envisagés. Pour compléter cette

analyse, nous avons ajouté les estimations obtenues pour un événement qui dépasserait Lothar en terme de dommages avec une proportion 1.5 fois plus importante : $L(1.5)$. Pour déterminer la valeur de l'indice I_S associée à ce niveau de sinistre, nous avons repris la formule 2.8.

$$L(1.5) = \frac{\ln(1.5)}{\gamma} + 100 \tag{2.9}$$

Période	1970-2013				1993-2013
Seuil choisi : u	10	20	10wL	10wM	10
Lothar Return Period	93	88	280	160	31
L(1.5) Return Period	136	128	490	252	84

TABLE VII – Calculs de périodes de retour selon différents scénarios.

Ce tableau nous permet de comparer les variations de période de retour selon l'amplitude de la tempête. Pour les deux premières estimations, entre 1970 et 2013 et avec l'ensemble des données, nous avons une augmentation quasiment linéaire de la fréquence de l'événement en fonction des dommages induits. Cette augmentation est un peu plus que linéaire et atteint un maximum de 490 ans dans les scénarios avec omission d'une tempête majeure. La variation la plus forte est celle observée sur la période 1993-2013 avec une multiplication par 2.7 pour passer de 31 à 84 années. Ces durées peuvent sembler extérieures à notre cadre de réflexion, mais ce n'est pas le cas. De telles différences auront des conséquences directes sur l'estimation tarifaire et les coûts de la réassurance. De plus, la perspective recommandée par la directive Solvency II aux compagnies d'assurance impose de travailler sur des quantiles à 99.5%, soit sur une échelle temporelle des événements ayant une période de retour de 200 ans.

2.6 Les besoins en fonds propres

Les tempêtes sont une cause d'intensité et de corrélation majeures sur un portefeuille IARD [3]. Lothar et Martin en 1999 ont représenté une charge de sinistres équivalente à une année normale. Toutes choses égales par ailleurs, en une seule année les compagnies d'assurances ont dû supporter l'équivalent de 2 années de sinistres. Ce qui n'est pas sans risque sur leurs fonds propres. Dans le cadre du projet européen Solvency II, les autorités de contrôle cherchent à déterminer des règles permettant de limiter le risque de ruine des compagnies d'assurance. Le premier pilier de cette réglementation définit les normes de calcul des fonds propres réglementaires. Le SCR (Solvency Capital Requirement) correspond au besoin en capital nécessaire à l'assureur pour tenir ses engagements sur un horizon d'une année. Un intervalle de confiance à 99.5% a été retenu pour les charges annuelles. Notons ici qu'il existe des mécanismes de couverture (la réassurance) qui permettent de lisser les situations extrêmes en constituant une forme de mutualisation, dans l'espace et dans le temps. Ils sont à la fois techniques et suivent des lois de marché. Cependant, la mesure des besoins en fonds propres se fait généralement au net de réassurance.

Compte tenu de la complexité de ces systèmes, propres à chaque compagnie, nous nous focalisons sur la sensibilité des résultats bruts dans le cadre de Solvency II. Nous étudions le capital cible nécessaire pour absorber le choc provoqué par un risque résultant d'événements extrêmes et irréguliers. L'autorité européenne de supervision des institutions d'assurance et de retraite (EIOPA) a défini une formules standard du SCR. Pour une charge annuelle variable C, la formule est la suivante :

$$\mathrm{SCR} = \mathrm{VaR}_{99.5}(C) - \mathbf{E}(C)$$

avec $\mathbf{E}(C)$ la charge moyenne annuelle calculée sur la période d'observation et $\mathrm{VaR}_{99.5}(C)$ la value at risk à 99.5 % de la distribution de C.

2.6.1 Calcul de la charge moyenne annuelle

Pour comprendre à quel point une charge exceptionnelle comme celle de l'année 1999 peut influencer la charge moyenne annuelle, nous proposons ici de l'inclure dans nos calculs selon les différentes périodes de retour obtenues dans la section précédente. D'après les résultats historiques [26] de la FFSA, 1999 représente à elle-seule 30% de la charge totale sur la période 1982-2012. Dans l'exemple qui suit nous allons considérer que la charge totale actualisée sur 30 ans d'une compagnie d'assurance s'élève à 5 milliards d'euros. On en déduit que la charge de l'année 1999 aurait atteint 1.5 milliards d'euros. La charge moyenne annuelle hors 1999 est alors égale à 116.7 millions.

Sur le tableau VIII on peut voir que la somme des charges actualisées de Lothar représente plus de deux tiers de la totalité des charges de la période. Selon la période de

3. Incendie Accidents Risques Divers.

1999 Return Period	31	88	93	160	280
Claims : 1999	48.4	17.0	16.1	9.4	5.4
Average annual claims	165.1	133.7	132.8	126.1	122.1

TABLE VIII – Evaluation de la charge moyenne annuelle (en millions d'euros).

retour choisie, la charge moyenne annuelle varie entre 165.1 millions (31 ans) et 122.1 millions d'euros (280 ans). L'année 1999 qui regroupe les deux principaux événements de la période est évaluée à 48.4 millions ou à 5.4 millions selon les deux scénarios les plus éloignés.

Ces écarts montrent qu'il est difficile de chercher un résultat qui est alimenté de façon arbitraire en amont. Le fait d'utiliser des outils de lissage et d'extrapolation ne crée qu'une illusion de technicité qui ne joue que de façon marginale. Le principal résultat ne dépend pas des modèles, mais du soin que l'on met à évaluer la période de retour de ces phénomènes.

2.6.2 Mesures de la dispersion

Sur quoi devons nous mesurer la dispersion ?

Pour aborder le problème de la dispersion, il faut rester dans les conditions vécues par l'assurance. Les compagnies travaillent généralement sur une base de charge annuelle nette de réassurance. La recherche de ce type de résultats nets pose le problème de la réassurance appliquée qui intègre à la fois des conditions techniques, stratégiques, commerciales et variables dans le temps. Il est donc plus simple, dans une étude théorique de rester au niveau du brut de réassurance. Les éléments fournis doivent cependant permettre de simuler la réassurance, ce qui impose de disposer des charges par évènement. Donc pour aborder cette question, nous devons travailler sur ce type de mesures, ou alors nous ne répondons pas à la question.

De quelles observations disposons-nous pour effectuer ces mesures ?

Sur les évènements les plus importants, il est possible de construire des historiques sur une quarantaine d'années [51]. Cette base *marché*, n'est pas une référence absolue mais plutôt un point de repère des événements marquants des quatre dernières décennies. L'actualisation sur une période relativement longue à l'échelle de l'assurance tempête pose quelques difficultés. Il faut, entre autres, considérer l'évolution des indices de construction (FFB ou RI), le poids relatif des segments particulier/entreprise dans le portefeuille, le taux de diffusion de la garantie tempête qui n'est obligatoire que depuis juillet 1990 ou encore la progression du parc immobilier en France. Sur la Figure 2.26 on observe les charges actualisées des principales tempêtes depuis 1970 d'après les données du marché. Le caractère exceptionnel de l'année 99, sans antécédent connu en France, illustre bien la très forte variabilité de la garantie tempête.

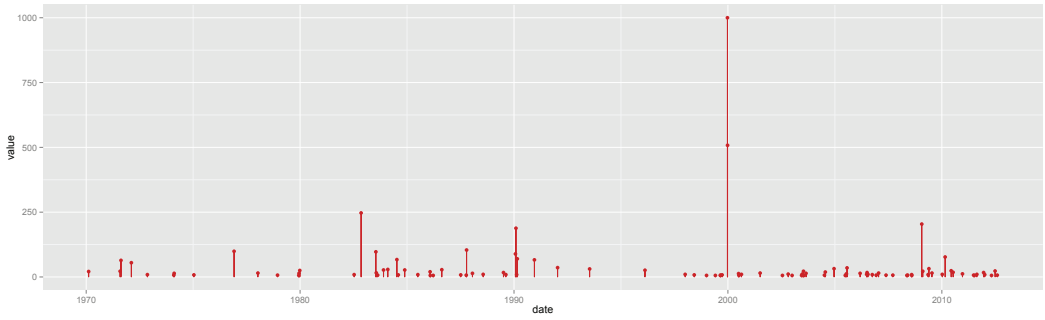


FIGURE 2.26 – Charges actualisées des tempêtes en France depuis 1970.

Il est aussi possible de travailler avec des bases *compagnie*, sur des périodes plus courtes et sur l'ensemble des sinistres. Ces bases fournissent des informations plus complètes sur les événements majeurs. Nous disposons, avec Allianz, d'une base de données répertoriant depuis 1998 l'ensemble des coûts des sinistres individuels causés par le vent avec une précision géographique départementale et temporelle journalière.

Ceci représente un peu plus de 520 000 sinistres en majorité pour des particuliers (310 000). Sur la base des primes comme sur celle des sinistres la part de marché d'Allianz pour la garantie tempête en France métropolitaine tend vers les 10% sur l'ensemble de la période. De plus les charges supportées par cette compagnie sont très fortement corrélées (0.99) avec celles du marché. Nous devons cependant noter que les difficultés d'actualisation demeurent présentes.

Quels constats pouvons-nous faire ?

Comme nous l'avons montré dans la section 5.1, les historiques sont trop courts pour se prononcer sur la charge moyenne. Ils le sont d'autant plus pour prétendre déterminer une distribution complète de la loi des charges annuelles. A ce stade, nous devons distinguer trois tranches de sinistres. Une tranche basse qui regroupe la plupart des événements, une tranche intermédiaire qui compte une dizaine d'événements par année et enfin la tranche supérieure qui concerne seulement les tempêtes les plus extrêmes de la période d'observation. En ce qui concerne la terminologie des événements, nous parlons d'événements fréquents, intermédiaires et extrêmes/rares. Dans la pratique, ces trois tranches ne sont pas modélisées selon une loi unique mais selon des lois différentes et adaptées. L'incertitude n'est pas identique sur tous les segments de la distribution. Il s'avère que les données disponibles sont suffisantes pour maîtriser les parties basse et intermédiaire de la sinistralité annuelle. En revanche, pour la partie des charges les plus extrêmes, il en va tout autrement.

Solutions existantes et envisagées

Différentes solutions sont déjà employées par les acteurs du marché. Elles se distinguent à la fois par les données de base utilisées pour leurs constructions et par les méthodes de calcul. La fonction de distribution est parfois estimée de façon empirique à partir des historiques dont dispose l'assureur. Dans d'autres cas elle se base sur des modèles permettant de multiplier les tirages pour simuler un très grand nombre d'années. Nous distinguons ici trois catégories d'approches : des modèles classiques (fonctions de fréquences et d'intensités annuelles, par niveaux de charges) ; des modèles de marché (utilisant des catalogues d'évènements, des fonctions d'endommagement, la localisation du portefeuille) et notre modèle construit à partir des relevés journaliers de charges et d'indice tempête.

Ces solutions apportent-elles des réponses indiscutables ?

Il n'existe pas à notre connaissance de solutions exemptes de difficultés. Qu'elles soient d'ordre théoriques, techniques ou matérielles, ces difficultés vont perturber la robustesse des résultats. Il est important de les souligner et de les associer aux modèles présentés pour mieux comprendre ses qualités et ses défauts et pour pouvoir utiliser les résultats en toute connaissance de cause.

En ce qui concerne les modèles classiques, le paramétrage des fonctions s'appuie sur un petit nombre d'observations. La fragilité croît avec le niveau des charges. Les paramétrages des lois extrêmes ne portent que sur un nombre très limité d'observations. Il faut parfois se contenter d'une unique observation annuelle, voire aucune observation. C'est ensuite par extrapolation d'une loi mal connue que l'on obtient la distribution.

Les modèles de marché essaient de s'exonérer des données observées, par l'utilisation de données externes (catalogue d'évènements synthétiques, fonctions d'endommagement). Mais comment valider ces paramètres si l'on ne peut pas les comparer à des observations ? Comment valider un catalogue de 18.000 évènements pour la France avec seulement 600 observations ? Comment valider les fonctions d'endommagement alors que nous n'avons aucune observation sur les critères retenus. Aucun dossier sinistre en France n'a enregistré la vitesse du vent, aucun dossier n'a enregistré le taux d'endommagement. Ces modèles ne nous renseignent pas sur les données à la base des fonctions d'endommagement. Les travaux menés en 2012 et 2013 par Luzi [51] ont montrés à quel point les résultats restitués par l'un de ces modèles étaient déconnectés de la réalité (en particulier en moyenne sur des charges non extrêmes et sur les restitutions géographiques).

Le choix d'utiliser des données journalières est motivé par la conservation d'un maximum de richesse d'informations dans un contexte d'historique limité dans le temps. Pour autant, cette démarche impose de réaliser ensuite un passage de la loi des charges quotidiennes vers une loi de charges annuelles. Autrement dit, dans le cas (avéré) du rejet de l'hypothèse d'indépendance entre les jours, nous devons étudier cette dépendance et l'estimer de façon à pouvoir l'intégrer ensuite dans notre modèle.

Notre modèle

Nous avons choisi de modéliser les charges journalières actualisées selon une loi GPD. Nous considérons les charges d'abord sans covariable puis avec l'indice tempête en covariable. Dans la Table IX nous présentons une estimation des paramètres de la loi de distribution, d'abord sur l'ensemble de la France puis selon les six zones tempêtes définies précédemment.

Zone	Global	1	2	3	4	5	6
Scale (σ):	1.51e+06	4.28e+05	5.88e+05	3.26e+05	2.10e+05	0.96e+05	3.29e+05
Shape (ξ):	1.12	1.03	1.17	1.12	1.26	1.14	1.09
Global Negative log-likelihood					1228.63		
Global AIC					2461.26		
Global BIC					2465.92		

TABLE IX – Estimations du modèle : (charges).

Dans tous ces modèles, le paramètre de forme ξ est supérieur à 1 pour les charges ce qui correspond à une distribution des extrêmes à queue très lourde. L'espérance

mathématique d'une telle loi est infinie ce qui montre la forte influence des extrêmes sur l'ensemble des tranches de sinistres. La différence entre le ξ des charges et le ξ de l'indice tempête, s'explique avec la formule 2.8. Il faut passer par une fonction exponentielle pour que l'indice reflète l'explosion de coûts qui apparaît pour les tempêtes les plus extrêmes. En distinguant les paramètres selon les zones géographiques de tempête, nous observons des différences avec un paramètre de forme variant entre un minimum de 1.03 dans la zone 1 et un maximum de 1.26 dans la zone 4. En ce qui concerne les paramètres d'échelle, les variations semblent être plus proportionnelles à l'exposition aux risques selon le portefeuille. Ces modèles peuvent être ajustés en ajoutant la covariable indice tempête à un de ces paramètres. Nous avons le choix entre le paramètre de forme ξ et le paramètre d'échelle σ. Notre décision s'appuie sur la Figure 2.27 qui confronte le logarithme des dépassements de charge avec l'indice tempête pour les vingt cinq dépassements les plus importants.

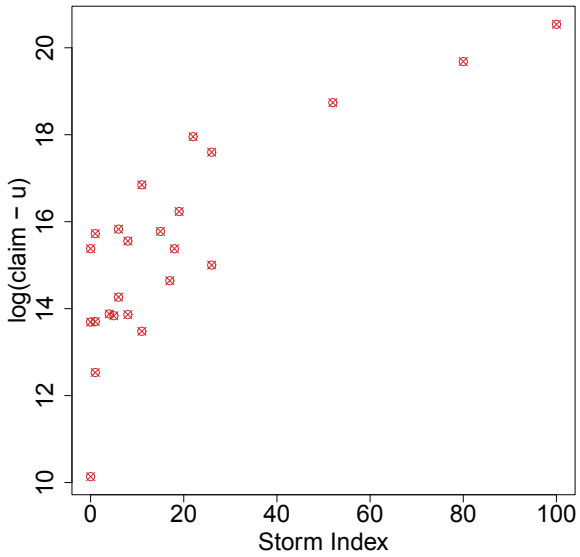


FIGURE 2.27 – Logarithme des charges en fonction de l'indice tempête.

La répartition quasi-linéaire des points, en particulier les quatre derniers, nous indique une bonne approximation par une loi de Pareto dont le paramètre de forme est lié de façon linéaire à l'indice tempête. En effet avec cette approximation, le log des charges suit une loi exponentielle de paramètre ξ. Dans la Table X nous rajoutons la covariable indice tempête à l'évaluation du paramètre de forme. La covariable améliore les résultats avec une vraisemblance plus forte d'environ 18 points comparée au modèle sans covariable. Cette différence est suffisante selon les critères AIC et BIC. On peut donc parler d'une amélioration.

Actuellement, nous avons un modèle correct pour les dépassements de seuil journaliers (GPD), donc pour la tranche la plus extrême de la distribution des sinistres. Nous souhaitons construire un modèle pour l'ensemble des charges annuelles. Étant dans une

Zone	Global
Scale (σ):	1.51e+06
Shape (ξ_0):	1.50e-03
Shape (ξ_1):	9.25e-02
Global Negative log-likelihood	1210.78
Global AIC	2427.55
Global BIC	2434.55

TABLE X – Estimations du modèle : (charges) avec covariable (sur le paramètre de forme).

situation où l'indice de forme de la queue de distribution est très élevé (ξ autour de 1, même supérieur à 1), on peut dire que les parties basses et intermédiaire des lois ont peu d'influence sur la loi annuelle car les quelques valeurs les plus fortes vont fortement dominer dans la somme des valeurs journalières. Pour correctement transformer cette loi journalière vers les lois annuelles, il nous faut donc la loi du nombre de dépassements sur une année. Sous l'hypothèse d'indépendance des jours, on aurait une loi binomiale ($n = 365$, p) avec p la probabilité de dépassement. Puisque n est grand et p est petit, nous pouvons l'approximer par la loi des évènements rares, la loi de Poisson ($\lambda = n \times p$). Pour cette loi, on a un rapport variance/espérance= $\lambda/\lambda = 1$, aussi appelé indice de dispersion.

Cependant, il peut y avoir une dépendance entre les jours, ce qui amène à l'occurrence de clusters et à la surdispersion, c'est-à-dire à un indice de dispersion variance/espérance > 1. Cette variabilité plus forte dans le nombre de dépassements se traduit par une variabilité plus forte des charges annuelles. Une loi adaptée à la surdispersion est la loi binomiale négative, une extension de la loi de Poisson avec un paramètre supplémentaire pour tenir compte du degré de surdispersion.

Comme les sources principales de clustering dans les dépassements de seuil des charges dues à des tempêtes agrégées par jour, nous pouvons considérer (en ordre décroissant de portée temporelle) le comportement saisonnier, les perturbations atmosphériques s'étendant sur plusieurs jours ou semaines et les tempêtes extrêmes s'étendant sur au moins deux jours. Bien sûr, ces trois sources sont étroitement liées.

Le comportement saisonnier dans les vitesses de vent et les sinistres associés ont pour résultat la mise en place de clusters naturels à une relativement grande échelle temporelle, voir Figure 2.28. La plupart des événements extrêmes, étant donnés ici comme des dépassements au-delà d'un seuil élevé, arrivent en été (juin à août) et en hiver (décembre à février). Les évaluations de l'index extrémal basé sur l'estimateur de Ferro-Segers ([29], certainement biaisées par le clustering saisonnier) sont montrées dans la Figure 2.29 (graphique de gauche), indiquant un clustering plus faible lors des évènements les plus extrêmes. L'indice de dispersion $\mathbb{V}(N)/\mathbb{E}(N)$ pour le nombre N de dépassement annuel au dessus d'un seuil élevé est donné dans la Figure 2.29 (graphique de droite), ce qui nous laisse entendre qu'il existe sûrement une surdispersion étant donné que les valeurs estimées sont comprises entre 2 et 3. Elles devraient être proches de 1, la valeur de référence pour une distribution de Poisson en l'absence

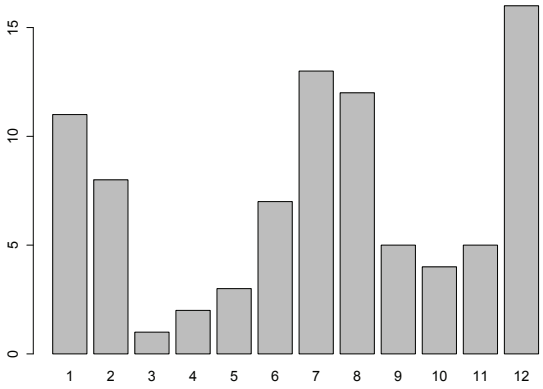


FIGURE 2.28 – Nombre de dépassements du vent selon les mois.

de clustering. L'estimation de l'indice est basée sur le paramètre de surdispersion θ d'une distribution binomiale négative paramétrée par $\mathrm{NB}(\mu, \theta)$ avec la moyenne μ et la variance $\mu + \mu^2/\theta$. Les données utilisées pour ces estimations se composent de 17 observations de comptage (pour 17 années d'observation).

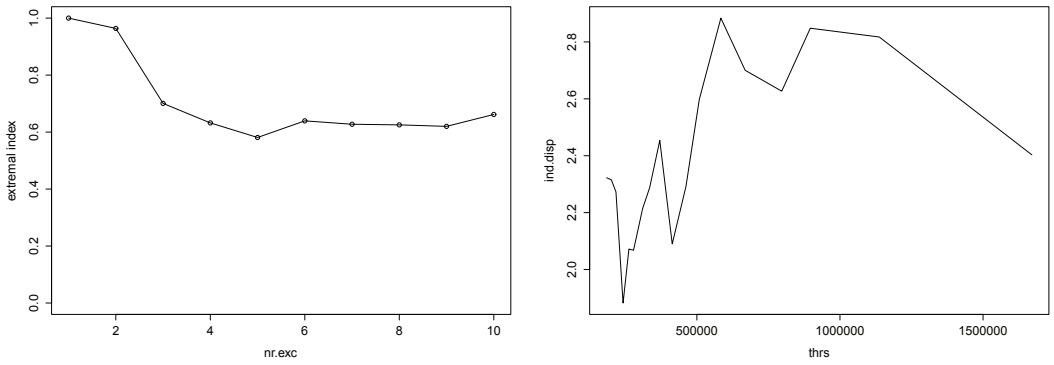


FIGURE 2.29 – Estimation de l'indice extrémal (gauche) et indice de dispersion (droite).

Un de nos objectifs principaux dans un contexte d'assurance est de prévoir le montant des charges annuelles agrégées et sa variabilité. Pour des raisons réglementaires, une bonne prédiction des quantiles élevés de ces charges annuelles ("Value at Risk") est cruciale. En raison d'évaluation d'indices de queue supérieures à 1 (voir le graphique de Hill dans la Figure 2.30 pour les charges quotidiennes agrégées), les queues sont très lourdes et, au moins théoriquement, peuvent correspondre à des distributions où ni l'espérance ni la variance ne sont finis. En pratique, les limites de couverture d'assurance imposent une borne supérieure finie à cette distribution. Cependant, dans notre contexte les charges annuelles agrégées vont être fortement dominées par les quelques événements journaliers les plus extrêmes qui se sont produits durant l'année et souvent

même l'événement journalier le plus extrême peut dominer à lui seul, c'est-à-dire qu'il aura une taille comparable à la somme de tous les autres événements.

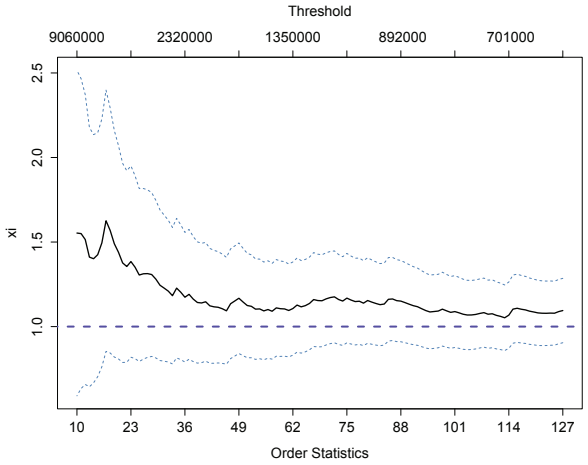


FIGURE 2.30 – Hill plot pour les charges journalières.

Notre approche se base sur des simulations de Monte-Carlo des charges annuelles agrégées. Nous nous focalisons ici sur un modèle approprié pour le nombre de dépassements annuels au dessus d'un seuil élevé, ensuite nous utilisons un modèle de Pareto généralisé pour la distribution des dépassements journaliers. Nous ne modélisons pas explicitement la distribution des charges journalières en dessous du seuil, mais à la place, nous utilisons une approche de bootstrap par bloc pour retranscrire correctement leurs dépendances temporelles et leurs contributions aux valeurs annuelles simulées. Pour simuler le montant des charges annuelles, nous fixons un seuil élevé u et utilisons les trois éléments suivants : 1) un modèle pour le nombre de dépassements N, ici la distribution sur-dispersée d'une binomiale négative, 2) un modèle GP des dépassements $Y \overset{d}{=} (X - u)|X > u$ pour simuler les évènements extrêmes $u + Y$, 3) un bootstrap par bloc pour les évènements journaliers en dessous du seuil, par exemple basés sur des blocs mensuels (en omettant les évènements extrêmes). L'avantage de cette approche est sa robustesse car nous n'avons pas à modéliser explicitement les clusters des évènements extrêmes journaliers. Par exemple, supprimer l'influence des comportements saisonniers ou tenir compte des perturbations atmosphériques s'étalant sur plusieurs semaines serait assez complexe.

La Figure 2.31 montre la distribution des charges annuelles agrégées (transformées en $\log_{10}$), basée sur 100000 simulations dans notre modèle. Nous proposons quatre choix de seuils correspondant à $3, 6, 9, 12$ dépassements annuels en moyenne. Les VaR pour des probabilités de 0.95 et 0.995 sont indiquées. Les choix de seuils n'ont que peu d'influence sur les quantiles (entre $10^{9.97}$ et $10^{10.10}$ pour $p = 0.95$, et entre $10^{10.69}$ et $10^{10.54}$ pour $p = 0.995$). Nous observons ensuite que l'impact du ré-échantillonnage en dessous du seuil est négligeable pour des quantiles aussi élevés. En effet, la somme

des valeurs ré-échantillonnées en dessous du seuil dépasse rarement 10^8 sur une année. Avec seulement 3 dépassements en moyenne, nous discernons un second mode dans la distribution qui disparaît avec un nombre élevé de dépassements. Ce second mode est dû au ré-échantillonnage en dessous du seuil, ce qui engendre une distribution moins lisse lorsque le seuil des dépassements est fixé très haut.

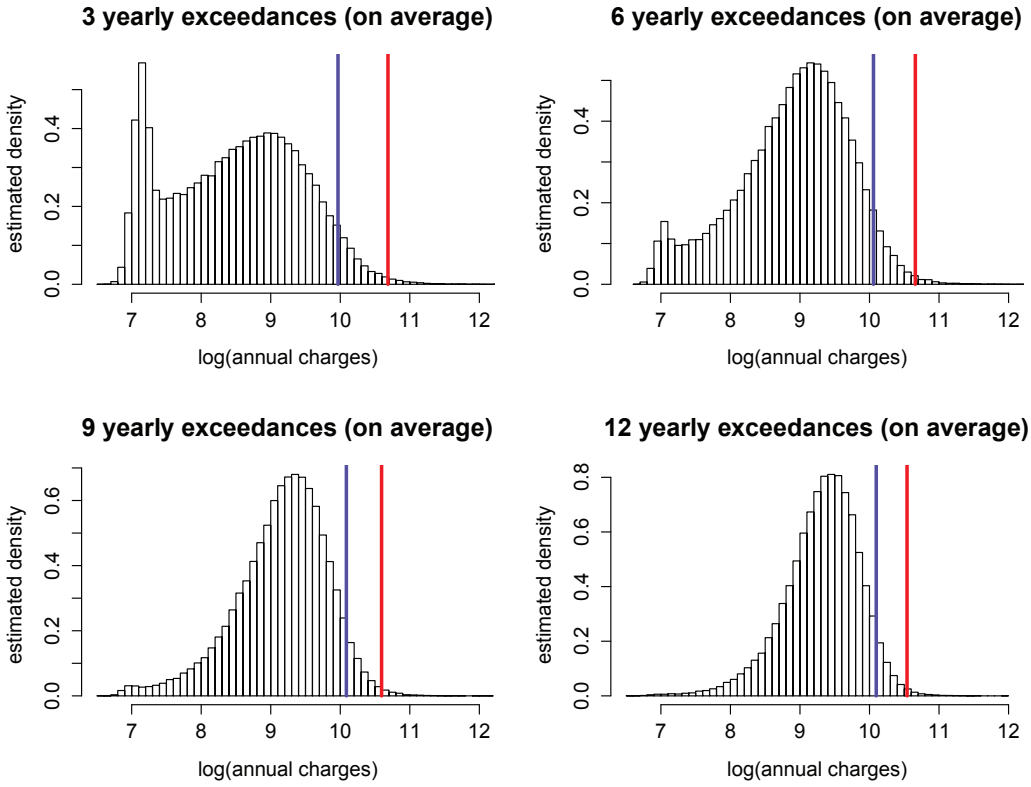


FIGURE 2.31 – Simulation de la distribution des charges annuelles.

2.7 Remerciements

Les données Météo France sont très complètes mais présentent le désavantage d'être payantes. Pour cette étude, nous avons bénéficié d'un crédit de recherche qui nous a permis de travailler sur les relevés de 192 stations météorologiques sur une période allant jusqu'à 50 ans. Ce travail a été soutenu par Allianz France.

2.8 Conclusion

Nous avons développé un indice tempête permettant d'estimer et de comparer l'ampleur des dommages causés par le vent lors des évènements les plus extrêmes observés

sur plusieurs dizaines d'années. L'intérêt de cette approche réside dans le fait que nous n'avons pas les mêmes difficultés et les mêmes besoins pour mesurer tous les phénomènes. En dessous d'un certain seuil, nous avons suffisamment d'observations pour considérer que la période observée ne biaise pas trop la recherche de lois. Soyons clairs, pour les petits et moyens évènements, les assureurs n'ont pas besoin d'utiliser un artifice à la place de leurs données. Leurs historiques, de 15 à 20 ans, sont largement suffisants pour répondre à leurs questions. Pour ce qui est des évènements extrêmes, il en va tout autrement. En fait pour cette partie, ils transfèrent le problème aux réassureurs, qui déterminent l'exposition et la traduisent en besoin de couverture et en prime de réassurance. Néanmoins certains problèmes demeurent, car certains évènements non représentatifs au niveau du portefeuille ne sont pas complètement réassurés. Remarquons également, que le fait de se réassurer ne devrait pas dédouaner la compagnie d'assurance d'analyser son exposition sur les niveaux extrêmes. Sinon comment apprécier le coût du service rendu par le réassureur ? En fin de compte, il peut être utile d'utiliser des informations complémentaires mais seulement pour les évènements majeurs.

Après une discussion sur la nature des vitesses de vent et les problèmes de ruptures inhérents à ce type de données, nous avons défini six zones géographiques de risques. Retenons que du fait du nombre limité des évènements significatifs, du nombre limité des risques touchés et de l'hétérogénéité des distances entre risques et stations, les résultats obtenus à l'échelle départementale manquent de robustesse. Nous avons réalisé des regroupements permettant de dégager des territoires plus homogènes. Ces zones ont été délimitées avec l'algorithme k-médoides à partir de la dépendance extrémale combinée aux maximums de vitesse de vent et à la distance euclidienne entre les centres de départements. Ces améliorations nous ont permis de construire un indice à la fois flexible et en adéquation avec les résultats des assureurs. Nous avons comparé les approches par jour et par semaine avec la méthode du bootstrap. Nous avons ensuite testé différents scénarios et différentes méthodes pour modéliser le risque tempête.

La distribution généralisée de Pareto (GPD) est à la base de notre modèle. Les paramètres estimés à chaque simulation nous renseignent sur la nature de la queue de distribution, mais ce sont les périodes de retour associées aux événements majeurs qui nous apportent le plus d'informations. Elles sont en effet très utiles pour calculer la charge moyenne annuelle d'un assureur. Il reste cependant difficile de retrouver des périodes de retour qui dépendent d'événements extrêmes très rares. Pour répondre aux exigences de la directive Solvency II, nos historiques laissent une grande place à l'incertitude. Les paramètres et les hypothèses de départ ont une forte influence d'abord sur la valeur de l'indice et ensuite sur sa modélisation. Les résultats différents d'un modèle à l'autre montrent cette sensibilité en particulier lors de l'évaluation des périodes de retour des tempêtes les plus extrêmes.

L'utilisation de données complémentaires comme les vitesses de vent associées à l'expertise des assureurs pour établir les zones de risque reste une bonne solution pour

la gestion du risque tempête. Cependant, malgré toutes les améliorations que nous pourrons apporter, la vraie volatilité des résultats dépend du choix des intensités et fréquences des sinistres extrêmes, données que ni nos historiques trop courts, ni nos modèles ne peuvent établir avec certitude. En fait, il faudrait d'autres moyens pour approfondir sensiblement les historiques sur les évènements extrêmes. Les simulations obtenues par des modèles du risque tempête, même en très grand nombre, ne peuvent pas remplacer des historiques incomplets et laissent donc une grande place à l'incertitude. Elles créent souvent une illusion de précision qui dépend très fortement d'hypothèses invérifiables.

How are the results of an electronic computer checked? By feeding data to a second computer of identical design. But two computers are not sufficient. If each computer arrived at a different answer, it is impossible to tell a priori which is correct. The solution, based on a careful study of statistical method, is to utilize a third computer to check the results of the first two. In this manner, a so-called majority report is obtained. [20].

Philip K. Dick

Chapitre 3

Comment répondre aux évolutions techniques et réglementaires de la tarification en assurance automobile ?

Sommaire

Alexandre Mornet Université de Lyon, Université Claude Bernard Lyon 1, Institut de Science Financière et d'Assurances, 50 Avenue Tony Garnier, Lyon F-69007, France and Allianz, Coeur Défense, 82 Esplanade du Général de Gaulle, Courbevoie F-92400, France, alexandre.mornet@allianz.fr
Patrick Leveillard ALLIANZ, Coeur Défense, 82 Esplanade du Général de Gaulle, Courbevoie F-92400, France.
Michel Luzi Non-life actuarial affairs former Director, Research ans Development Director at Allianz France, qualified Member of Institute of Actuaries, 132, rue du Président Wilson, Levallois-Perret F-92300, France
Stéphane Loisel Université de Lyon, Université Claude Bernard Lyon 1, Institut de Science Financière et d'Assurances, 50 Avenue Tony Garnier, Lyon F-69007, France, stephane.loisel@univ-lyon1.fr

3.1 Abstract

Dans le secteur de l'assurance automobile, la tarification évolue en fonction des avancées techniques et de la réglementation. Le contexte actuel offre de nouvelles perspectives avec la mise en place de dispositifs comme la géolocalisation et impose de nouvelles contraintes avec une limitation des informations à caractère personnel que l'assureur sera en droit d'utiliser. Nous en avons un exemple concret fin 2012 avec l'entrée en vigueur de la décision de la Cour de Justice de l'Union européenne selon laquelle il n'est plus possible de pratiquer des tarifs dépendant du sexe. Lorsqu'une variable couramment utilisée comme le sexe de l'assuré devient interdite, une première idée peut être de la remplacer par un proxy. Cependant la construction de modèles innovants apporte davantage de solutions. Dans cet article, nous testons différentes méthodes pour répondre à ces problématiques. Nous proposons de caractériser la partition homme/femme en utilisant la procédure logistique, l'analyse des correspondances multiples (ACM) ou les arbres de classification (CART). Nous montrons ensuite qu'il est possible de compenser l'absence de la variable sexe par d'autres informations spécifiques à l'assuré ou à son véhicule et en particulier l'utilisation de relevés kilométriques [48]. Dans une dernière partie, nous nous intéressons à l'expérience acquise par les conducteurs novices. Le critère sexe est en effet particulièrement pertinent sur cette classe de conducteurs. Nous proposons d'étudier le comportement sur la route de l'assuré durant ces trois années de noviciat pour créer de nouvelles classes de risques.

3.2 Introduction

La tarification du risque automobile connaît actuellement une évolution à la fois technique et dans son rapport avec les assurés. L'objectif principal est d'obtenir la meilleure adéquation possible entre le risque et le tarif. Tout en tenant compte des différentes réglementations à l'échelle nationale, européenne ou internationale, les compagnies d'assurance veulent proposer des produits novateurs répondant aux attentes des clients à la recherche d'offres de plus en plus personnalisées. Elles doivent cependant maintenir une bonne mutualisation des risques en évitant toute anti-sélection. Les différentes solutions développées jusqu'à présent peuvent nous permettre de mieux comprendre cette problématique. Classiquement, la tarification se base sur un nombre important de critères dont l'importance dépend des pondérations à l'intérieur du modèle. On distingue d'abord les variables relatives à l'identification des personnes parties, intéressées ou intervenantes au contrat comme l'état civil ou les coordonnées, celles relatives à la situation familiale et économique et à la situation professionnelle. D'autres données sont aussi nécessaires à l'appréciation du risque comme la situation géographique et des renseignements sur le bien assuré : le type et les caractéristiques du ou des véhicules assurés, le permis de conduire et l'usage à titre professionnel ou non. Viennent ensuite les données relatives à la gestion des sinistres et des prestations, puis à l'évaluation des préjudices. À toutes ces données, il faut désormais rajouter des informations qui peuvent s'avérer plus sensibles comme celles relatives à la localisation des personnes ou des biens et celles qui concernent la vie personnelle et les habitudes de vie.

L'assurance comportementale et l'assurance à l'usage comptent parmi les solutions envisagées par les assureurs. Ces nouveaux modèles d'offres visent à améliorer la connaissance du client et à se tourner davantage vers le serviciel. Cette rupture est permise par l'afflux d'informations accessibles et regroupées dans ce qu'on appelle la datamasse ou Big Data [31]. Les variables traditionnelles se voient alors remplacées ou enrichies par des données contextuelles et par une circulation accrue des informations entre l'assureur et l'assuré. Les solutions actuelles font intervenir l'utilisation du véhicule. Ces dernières années, l'assurance au kilomètre ou *Pay As You Drive : PAYD* ([2], [4]) s'est largement diffusée dans le système de l'assurance automobile. Elle ne constitue pas une dépense supplémentaire pour les assurés mais plutôt un pas de plus vers une tarification plus proche de leurs comportements routiers. Grâce à des boîtiers embarqués, on peut désormais mieux connaître l'utilisation que le conducteur fait de son véhicule. En plus du kilométrage, il est possible de savoir la période d'utilisation du véhicule (jour/nuit), le type de routes empruntées (autoroute/ville/campagne) ou même le type de conduite (vitesse/accélération). Légalement, il faut cependant noter que certaines limites doivent être observées dans le respect des libertés individuelles. En particulier, les compagnies d'assurance et les constructeurs automobiles doivent tenir compte de la délibération 2010-096 du 8 avril 2010 de la Commission nationale de l'informatique et des libertés (CNIL) concernant les dispositifs de géolocalisation. Une des principales recommandations porte sur la gestion des données collectées dans le cadre de l'assurance *PAYD*. Contrairement à d'autres dispositifs comme la lutte

contre le vol (tracking) ou l'appel d'urgence, les informations ne sont pas transmises ponctuellement mais de façon systématique. La conservation de ces données en dehors du cadre nécessaire au calcul de la prime pourrait alors constituer une atteinte à la vie privée des personnes concernées et irait à l'encontre de la liberté d'aller et venir anonymement. Il serait aussi possible via ces relevés de constater des manquements au code de la route mais les assureurs ne sont pas habilités à constater de telles infractions. Le projet PriPAYD [80] propose par exemple de mettre en place un système de calcul de la prime à l'intérieur de chaque véhicule pour ne transmettre à la compagnie d'assurance que des données agrégées sans fuite d'informations de localisation.

De façon globale, l'assureur doit faire face à une nouvelle donne imposée par la réglementation européenne. Devant les éventuelles dérives que pourrait entraîner l'utilisation d'informations personnelles concernant les assurés, la tendance générale est à une disparition progressive de tous les critères discriminants qui entrent dans la méthode de tarification. La distinction selon le sexe est la première à disparaître, mais au delà on peut s'attendre à un élargissement de ces interdictions qui pourraient atteindre beaucoup de critères jusqu'alors couramment utilisés comme le bonus/malus ou l'âge de l'assuré. Pour ce qui est du sexe, la décision de la cour européenne de justice stipule que l'article 5(2) de la Directive 2004/113 n'est pas compatible avec l'article 6(2) du traité de l'Union Européenne. En clair, les assureurs ne sont plus autorisés depuis le 21 décembre 2012 à ajuster leurs tarifs en fonction du critère homme/femme car une telle distinction n'est pas compatible avec les principes d'égalité. La variable sexe a pourtant été largement employée par les assureurs avant cette directive car elle leur permettait de déterminer facilement deux profils de risque différents. En assurance automobile par exemple, il s'avère que les femmes ont statistiquement moins d'accidents que les hommes.

Pour répondre à ces nouvelles obligations, deux pistes différentes s'offrent aux assureurs. Ils peuvent soit essayer de construire des proxies, c'est à dire substituer de nouvelles variables aux variables interdites, soit explorer de nouveaux modèles faisant totalement abstraction des variables interdites et proposant dans le meilleur des cas des solutions à la fois innovantes et plus efficaces que les précédentes. Dans ce contexte, nous avons voulu décrire ce qui caractérise la partition homme/femme à partir des autres variables explicatives que l'assureur est en droit d'utiliser. Nous avons aussi montré qu'un modèle de prédiction des sinistres peut fonctionner sans la variable sexe et fournir d'aussi bons résultats. Parmi les informations complémentaires à la disposition de l'assureur, la connaissance du kilométrage annuellement parcouru constitue une nouvelle variable très significative [48]. Bien que la relation entre la fréquence des sinistres et le kilométrage annuel puisse sembler évidente, peu d'études ont jusqu'à présent insisté sur l'importance de travailler avec des données fiables sur le kilométrage parcouru par les assurés ([3], [50]). Pourtant, la précision de cette information s'avère essentielle à la construction de catégories de risques lors de la tarification. Le facteur principal dans le choix d'une assurance automobile par les consommateurs demeure le prix. Par conséquent, une tarification plus flexible devrait constituer une offre plus intéressante

et aurait en plus l'avantage écologique [11] d'inciter à un usage moins systématique des véhicules individuels lorsque cela est possible [9]. Dans nos recherches, nous avons considéré la distribution des sinistres directement liés à la circulation routière comme les dommages matériels responsables, mais aussi ceux qui peuvent en sembler déconnectés comme le vol ou l'incendie. Historiquement, comme le montre Roger Roots dans son rapport sur les dangers de l'automobile [71], depuis les premiers moyens de transport à cheval, la tendance était à une diminution linéaire de la fréquence des sinistres pour chaque kilomètre parcouru. Dans une perspective plus actuelle, la sinistralité se reflète dans l'expérience acquise au volant et dans le nombre d'années depuis l'obtention du permis. Les habitudes de conduite que l'on retrouve dans le kilométrage annuel peuvent nous aider à mieux comprendre le risque d'accident.

Dans une première partie, nous présentons les données d'assurance à notre disposition. Nous observons la sinistralité des différentes catégories de risques selon la partition homme/femme. L'approche graphique permet de mettre en évidence des différences statistiques. Pour faire évoluer leurs tarifications les assureurs peuvent envisager deux approches : mettre en place un proxy ou créer de nouveaux modèles. Partant de ce constat, nous essayons dans une deuxième partie de caractériser le sexe du conducteur en fonction d'autres variables spécifiques à l'assuré, à son environnement ou à son véhicule. Nous utilisons ensuite les modèles additifs généralisés (GAM) et les modèles linéaires généralisés (GLM) pour comparer l'efficacité des modèles avec et sans la variable sexe [54]. Nous démontrons qu'il est possible d'améliorer les résultats des prédictions en faisant abstraction du critère homme/femme. La troisième partie est consacrée aux novices et à l'acquisition d'expérience selon le kilométrage. Nous explorons à travers différentes tranches kilométriques le comportement et l'évolution des jeunes conducteurs durant leurs trois années de noviciat.

3.3 Description des données d'assurance

Pour cette étude, nous avons eu accès aux portefeuilles automobiles d'Allianz en France sur la période 2008-2010. Plusieurs catégories de données sont disponibles. Elles relèvent de différentes sources. La table **Polices** $\times$ **risques** contient les informations relatives à l'assuré et son contrat d'assurance auto. Elle compte plus de 2.7 millions d'observations réparties sur 192 variables. On y trouve par exemple la formule de garanties, l'âge du véhicule, la catégorie socio-professionnelle (CSP), l'usage qui est fait du véhicule, l'âge du conducteur, l'âge à l'obtention du permis, la présence d'un garage et les antécédents de sinistralité en nombre de sinistres. Les risques et les niveaux de franchises sont classifiés en trois zones : vol incendie, dommage et bris de glaces (VI, DOM et BDG). Les critères client comme le nombre de contrats auto ou l'existence de contrats multirisques habitation (MRH) sont aussi exploités. Cette table est associée à la table **Sinistres** qui rassemble la nature de l'événement, sa date, son coût détaillé. Pour chaque exercice elle contient autant de lignes par numéro de police que d'événements sur le contrat.

L'association **SRA** (Sécurité et Réparations Automobiles) fournit toutes les informations relatives aux véhicules en commercialisation. Elle nous renseigne sur des caractéristiques techniques comme la puissance réelle, le type de carrosserie, le segment ou la note de sécurité. Trois indicateurs permettant une classification sont proposés. Le groupe reflète la dangerosité intrinsèque des véhicules, il a pour finalité d'aider les assureurs à tarifer la garantie responsabilité civile. La classe de prix est liée au prix du véhicule neuf, elle est utilisable en cas de perte totale du véhicule. Enfin, la classe de réparation est un indicateur lié au coût des réparations. Le partenariat entre Allianz et l'entreprise **SOFCA** (Société Française des Compteurs Automobiles) nous permet d'obtenir les relevés kilométriques des assurés. L'idée d'utiliser ce type d'information n'est pas nouvelle, au début des années 60, avec la démocratisation de l'automobile, une prise en compte du kilométrage lors de la tarification des polices d'assurance fait déjà partie des réflexions sur la réforme du tarif français d'assurance automobile [77]. On note même un premier essai en 1936 d'assurance au kilomètre par la compagnie *La Préservatrice*. Les études statistiques étaient cependant difficilement réalisables à l'époque. Notre étude s'appuie sur l'ensemble des conducteurs ayant équipé leurs véhicules d'un compteur additionnel, soit 440 949 observations.

Quantile	Kilométrage
Max 100%	49632
99%	20038
95%	12691
90%	10060
Q3 75%	7191
Médiane 50%	4630
Q1 25%	2495
10%	1146
5%	626
1%	114
min 0%	1

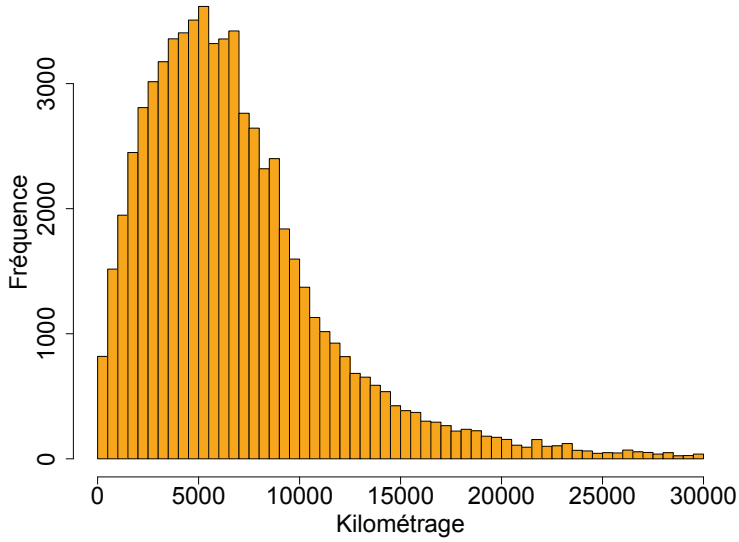


FIGURE 3.1 – Résultats Sofca : répartition des kilométrages annuels en 2008

La Figure 3.1 présente un histogramme de la répartition des kilométrages annuels sur l'année 2008. La distance annuelle moyenne parcourue est de 5378 km. On constate aussi que la majorité des assurés (90 %) font moins de 10 000 km. Nous verrons par la suite comment ces informations étalées sur 3 ans sont exploitées et comment nous avons construit différentes classes de kilométrage.

Nous utilisons des données de l'**INSEE** (recensement démographique) pour connaître

l'environnement démographique de l'assuré. Les données sélectionnées sont locales et assez diverses. Nous avons retenu la taille, l'évolution et l'âge de la population, le taux de naissance, le pourcentage de femmes/hommes entre 15 et 29 ans et la densité. En ce qui concerne les professions, deux groupes se distinguent à travers le pourcentage d'artisans, de commerçants, de chefs entreprise, de cadres et de professions intellectuelles supérieures et le pourcentage d'employés et d'ouvriers. À cela s'ajoute des catégories permettant de mieux comprendre les mouvements de population comme le nombre de personnes de 5 ans ou plus habitant depuis plus de 5 années le même logement, ceux habitant depuis plus de 5 années un autre logement de la même commune, ceux habitant depuis plus de 5 années dans une autre commune du même département, ceux habitant depuis plus de 5 années la même région et ceux habitant depuis plus de 5 années hors de France métropolitaine ou d'un département d'outre-mer. Enfin, dans notre étude, nous employons des critères spécifiques aux novices comme l'ancienneté de noviciat, conduite accompagnée, enfant d'assuré, variables relatives aux parents si enfant d'assurés.

3.3.1 Répartition de la fréquence des sinistres selon le sexe

D'une manière générale, dans toute étude statistique il faut envisager les différents biais inhérents à chaque catégorie de risques qui peuvent perturber les résultats. Dans notre cas, on relève certains comportements récurrents selon les âges qui seraient de nature à fausser l'identité réelle de la personne au volant. Par exemple, dans un couple, l'échange du véhicule entre le mari et sa femme est une pratique courante. On relève aussi une augmentation de la sinistralité dans la tranche 45/50 ans chez les femmes qui s'explique assez souvent par le prêt du véhicule à leurs enfants en âge de conduire. L'absence de la distinction homme/femme dans la tarification comme le stipule la directive européenne va surtout poser problème aux assureurs pour deux catégories de population, les conducteurs novices et les conducteurs seniors. En effet, on observe statistiquement une sinistralité plus forte chez les novices [1] hommes et chez les seniors [2] femmes, comme dans les résultats de Drummond, A. E. *et al.* [27] en Australie et de Massie, D. L. *et al.* de l'institut de recherche sur les transports du Michigan [54]. Ces différentes constatations nous ont menés à nous concentrer sur une seule catégorie de risques : les **novices**. Ils représentent 15% des entrées et 4% du portefeuille en agence [3].

Chez les conducteurs novices on remarque une plus forte sinistralité chez les hommes toutes catégories de risques confondues que chez les femmes. Sur la Figure 3.2, on peut comparer l'évolution du nombre moyen de sinistres selon le sexe sur 3 années consécutives. Le rayon des cercles correspond à la proportion respective d'hommes et de femmes assurés chez Allianz. Le portefeuille a légèrement diminué en taille durant la période 2008-2010, en revanche les proportions sont relativement les mêmes avec pour les hommes une part comprise entre 65.5 % et 64.2 % et pour les femmes entre 34.5 % et 35.8 %. L'écart de sinistralité évolue aussi à la baisse avec un écart relatif qui passe

1. Pour Allianz, ce sont les conducteurs avec moins de 3 ans d'assurance automobile.
2. Les assurés de plus de 60 ans
3. les taux Allianz sont très proches de ceux du marché

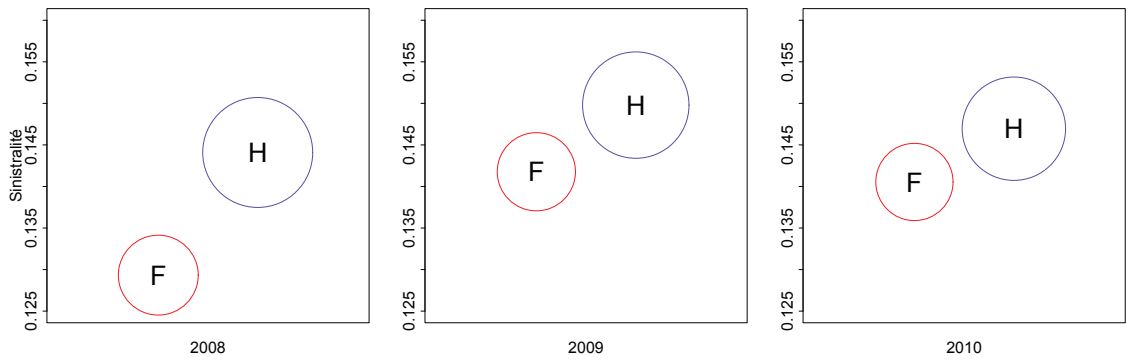


FIGURE 3.2 – Nombre moyen de sinistres pour les hommes et les femmes novices entre 2008 et 2010

de 10.2 % à 4.4 %. Pour plus de précisions on compare la moyenne de sinistres pour les différentes catégories de risques prises séparément.

Sur la Figure 3.3, on observe la fréquence des sinistres pour les hommes et les femmes selon la nature de la garantie touchée (cf. Table XXIII en annexe). Le premier constat concerne la répartition inégale des différents risques qui s'exprime à travers les différentes tailles de cercles. Les catégories les plus représentées en termes de nombre de sinistres sont les indemnisations directes (IDA), les bris de glaces (BDG) et l'assistance (ASSI). Les sinistres corporels non responsables et les vols/incendies sont les moins nombreux. Les hommes ont une sinistralité plus forte dans une majorité de catégories de risques, cependant ce n'est pas le cas pour les bris de glace, les vols et incendies auxquels on peut rajouter les sinistres corporels non responsables et les dommages. La sinistralité plus élevée chez les femmes concerne majoritairement des sinistres que l'on peut considérer comme non responsables. Ces catégories apparaissent en *italique* sur Table I. Nous entrevoyons avec ces premiers résultats que le relation entre la sinistralité et le sexe n'est pas à sens unique. Si de nombreuses études illustrent une plus forte implication des hommes dans les sinistres graves, il est plus rare de constater des taux plus élevés chez les femmes. C'est le cas lorsque l'on s'intéresse à certaines catégories de risque en particulier, comme les sinistres corporels non mortels qui sont selon une étude norvégienne [8] bien plus fréquents chez les femmes. Dans cet article, notre propos sera de montrer que ces différences sont liées au comportement des assurés qui s'avère plus complexe qu'un simple découpage entre homme et femme.

3.3.2 Répartition du coût des sinistres selon le sexe

Pour la modélisation des coûts, on travaille généralement sur les sinistres fermés, on évite ainsi d'intégrer aux résultats les sinistres forfaitaires. A priori, les écarts les plus marqués entre hommes et femmes apparaissent surtout au niveau des fréquences des sinistres. Sur la Figure 3.4 on a représenté la charge moyenne selon le sexe pour l'ensemble des garanties ainsi que le nombre de sinistres (proportionnel au rayon des cercles). Il y a bien une différence entre ces charges moyennes calculées sur trois années (de 2008

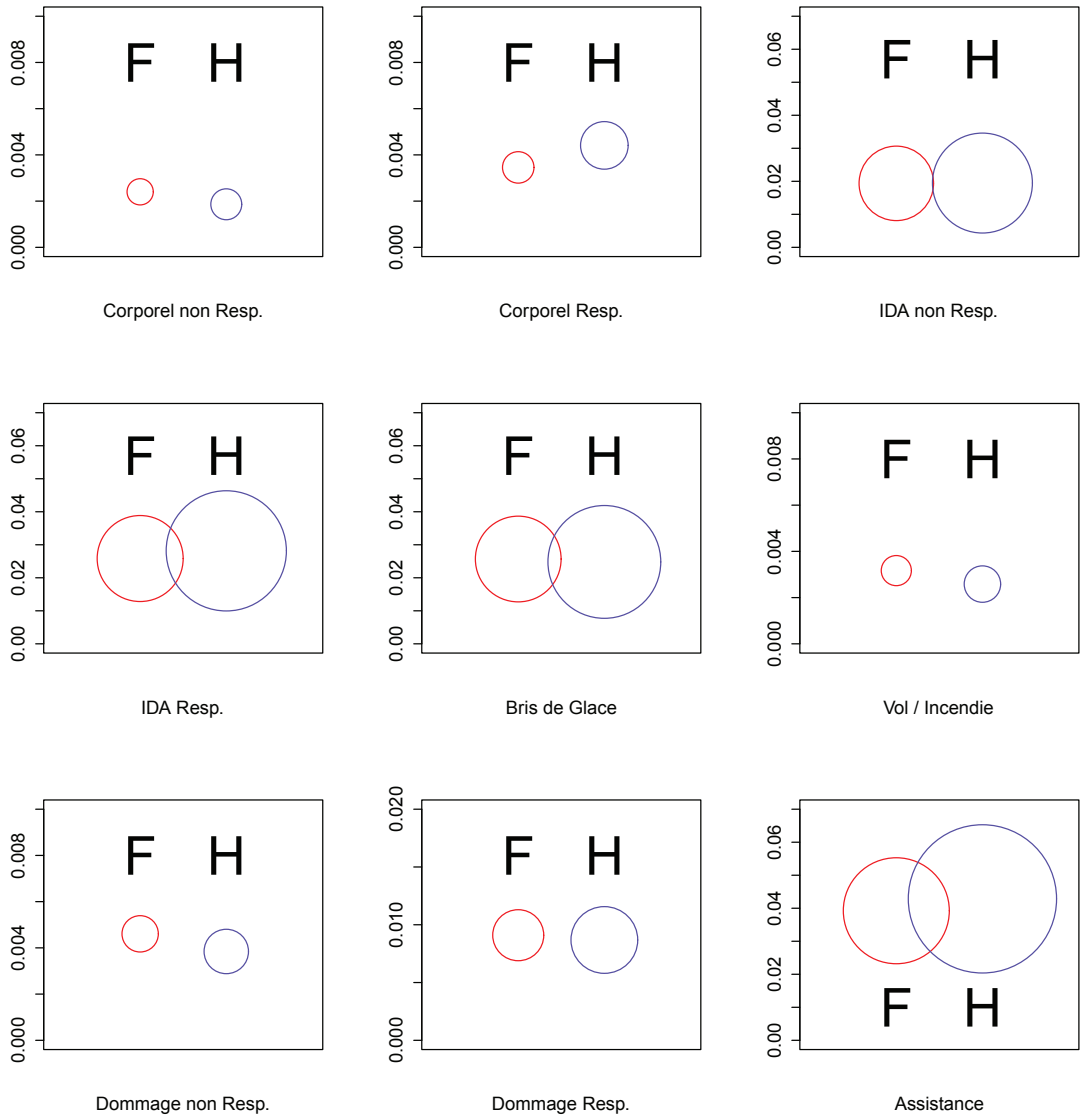


FIGURE 3.3 – Fréquence de sinistres hommes et femmes novices en 2010

à 2010), cependant elle est relativement faible : 3 % de plus pour les hommes. Pour le nombre de sinistres, on retrouve un ordre de grandeur correspondant aux constatations de la section précédente avec 68,3 % des sinistres du côté des hommes et 31,7 % du côté des femmes.

La Table II montre les coûts moyens des sinistres chez les hommes et les femmes novices selon les types de garanties. Lorsque l'on observe cette différence garantie par garantie, on voit qu'elle reste assez faiblement supérieure chez les hommes avec la plus forte différence observée pour les bris de glace avec un pic à 7 %.

	Non Responsable			Responsable			Autres		
2010	*Corporel*	IDA	*Dommage*	Corporel	IDA	*Dommage*	*BDG*	*VI*	ASSI
F	0,24 %	1,94 %	0,46 %	0,35 %	2,58 %	0,91 %	2,57 %	0,32 %	3,92 %
H	0,19 %	1,95 %	0,38 %	0,44 %	2,82 %	0,87 %	2,48 %	0,26 %	4,28 %
2009	Corporel	*IDA*	*Dommage*	Corporel	IDA	Dommage	*BDG*	*VI*	ASSI
F	0,29 %	2,09 %	0,54 %	0,36 %	2,55 %	0,82 %	2,55 %	0,33 %	3,91 %
H	0,17 %	1,93 %	0,47 %	0,44 %	2,91 %	0,84 %	2,52 %	0,32 %	4,31 %

TABLE I – Fréquence de sinistres hommes et femmes novices par catégories

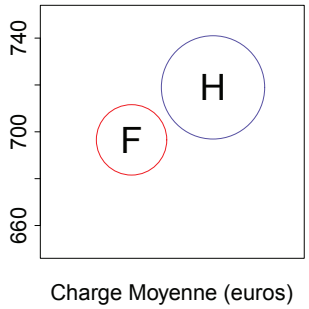


FIGURE 3.4 – Charge moyenne et proportion en nombre de sinistres chez les hommes et les femmes novices de 2008 à 2010

	Non Responsable			Responsable			Autres		
2010	Corporel	IDA	Dommage	Corporel	IDA	Dommage	BDG	VI	ASSI
F	791,8	697,8	887,8	1089,2	1041,7	1043,5	388,9	776,6	438,6
H	808,7	727,5	907,9	1104,2	1051,1	1052,7	418,1	809,6	448,4
Diff	2 %	4 %	2 %	1 %	1 %	1 %	7 %	4 %	2 %

TABLE II – Charge moyenne des sinistres hommes et femmes novices par catégories

En définitive, chaque critère, nombre, fréquence, coût, donne un éclairage potentiellement différent des autres. D'après le nombre moyen de sinistres, on retrouve l'idée assez communément admise, du moins chez les assureurs, selon laquelle les femmes ont moins d'accidents que les hommes. Ce constat est le même lorsque l'on compare les charges moyennes à la fois globales et selon les catégories de garanties. En revanche, les différentes fréquences vues dans le détail ne donnent pas un avantage aussi catégorique aux femmes. Dans certains cas, leur fréquence de sinistre est supérieure à celle des hommes. Pour comprendre ces différences, il faut analyser chaque sinistre plus en profondeur et dépasser la simple distinction selon le sexe qui ne suffit pas à tout expliquer. En matière de vol, par exemple, on constate que les petits modèles sont plus fréquemment touchés, et ce type de véhicules est plus souvent conduit par des femmes, mais cela reste une tendance. Nous voyons ici que pour construire un bon

modèle de tarification, il est évidemment préférable de se baser sur des critères fiables et directement liés au comportement du conducteur, plutôt que sur des associations qui ne s'avèrent pas toujours exactes.

3.4 Caractérisation de la partition Homme/Femme

Le choix de tarifer par genre offre de nombreux avantages : simplicité de recueil de l'information, existence de corrélation entre risque et variable choisie. Mais il ne faut jamais oublier que ces critères ne sont que des indicateurs de l'exposition de l'assureur aux risques [7]. Pour répondre à l'interdiction légale d'utiliser le sexe comme un critère discriminant dans leur tarification, une solution possible consiste à utiliser un proxy. En se servant de l'ensemble des variables toujours à leur disposition, les assureurs peuvent en effet tenter de retrouver le sexe de leurs clients et d'utiliser alors ce proxy dans un modèle classique. Cette solution n'est évidemment pas celle préconisée par la réglementation, mais il est naturel de l'envisager et c'est la raison pour laquelle nous abordons ce sujet ici. L'ensemble des résultats présentés sont obtenus à partir de la population novice du portefeuilles. Nous verrons par la suite qu'une autre approche est possible et peut même s'avérer meilleure. On se propose donc, dans les sections suivantes de caractériser la partition homme-femme à partir des variables catégorielles à notre disposition.

3.4.1 La procédure logistique

Plusieurs approches sont envisagées, la première est une modélisation couramment employée en assurance pour cibler une fraction de la clientèle : la régression logistique. Dans le cadre d'une régression binaire, la variable à prédire `SEXE` prend deux modalités possibles $\{1, 0\}$. Nous utilisons la procédure `logistic` du logiciel SAS et l'ensemble des variables prédictives sélectionnées par la suite pour les GLM. L'optimisation est effectuée sur un échantillon de 100 000 assurés. La méthode stepwise considère que 32 des 35 variables explicatives sont utiles. Les plus significatives sont le segment, la classe SRA, la CSP, l'âge du conducteur et la classe de véhicule. Les moins significatives proviennent de l'INSEE comme la densité de population ou le pourcentage d'employés et d'ouvriers de la commune. Les résultats de cette procédure selon les tests de Hosmer et Lemeshow sont dans la Table III. Lors de cette phase, les données sont rangées par ordre croissant des probabilités calculées à l'aide du modèle, puis partagées en 10 groupes. Le test est largement positif mais il est aussi malheureusement peu puissant. C'est le test du khi-deux qui est utilisé pour comparer les effectifs observés aux effectifs théoriques. Les prédictions globales sont assez précises.

La procédure est alors appliquée à un échantillon test de 125 000 assurés. Nous observons dans la Table IV la matrice de confusion. Cette matrice constituée de deux lignes et de deux colonnes permet de confronter directement le nombre de femmes et le nombre de hommes correctement identifiés par le modèle et le nombre d'erreurs. Les résultats sont concluants pour 71138 hommes et 15485 femmes. Nous lisons sur

		SEXE = H		SEXE = F	
Groupe	Total	Observé	Attendu	Observé	Attendu
1	10000	2743	2523	7257	7477
2	10000	3712	3807	6288	6193
3	10000	4640	4568	5360	5432
4	10000	5068	5195	4932	4805
5	10000	5592	5769	4408	4231
6	10000	6211	6335	3789	3665
7	10000	6897	6924	3103	3076
8	10000	7637	7529	2363	2471
9	10000	8289	8210	1711	1790
10	10000	9154	9083	846	917

TABLE III – Partition pour les tests de Hosmer et de Lemeshow

les données en apprentissage que le modèle de prédiction réalise $9954 + 28423 = 38377$ erreurs sur 125 000. Ceci représente un taux d'erreur inférieur à 31 %. Il faut remarquer ici que ce résultat relativement bon est en grande partie du à la forte proportion d'hommes qui sont correctement identifiés car pour les femmes le modèle se trompe plus souvent qu'il ne réussit. Nous allons comparer par la suite cette précision avec celle obtenue par les arbres de décision.

	F	H	Total
F	15485	28423	43908
	12,39	22,74	35,13
H	9954	71138	81092
	7,96	56,91	64,87
Total	25439	99561	125000
	20,35	79,65	100

TABLE IV – Matrice de confusion

3.4.2 Exploration de données

ACM

Dans un deuxième temps nous nous intéressons aux techniques d'analyse de données. Une solution adaptée à la prise en charge de variables catégorielles consiste à faire une analyse des correspondances multiples (ACM). L'ACM permet techniquement de projeter et donc représenter un nuage de points initialement situé dans un espace de très grande dimension (le nombre de modalités moins le nombre de variables) dans un espace de dimension plus réduite. La distance des points deux à deux y est maximale, donc l'espace est celui qui conserve le mieux la richesse de l'information de départ [47]. On peut alors observer une classification des variables selon leur importance dans le nuage et sur chaque axe. Pour notre projet de partition des individus selon le genre,

nous observons particulièrement le rang et la contribution de la variable SEXE dans le nuage de points et la constitution des axes. On pourra alors si par exemple la modalité homme a une contribution forte sur un des axes principaux associer à cette modalité celles des autres variables catégorielles qui influencent également l'axe auquel on s'intéresse. Nous utilisons la procédure CORRESP du logiciel SAS à laquelle est combinée la macro AIDEACM qui facilite l'interprétation des résultats de l'analyse. L'exécution du programme nécessite la construction d'un tableau disjonctif complet. Nous lançons la procédure avec des variables actives issues des critères classiques et spécifiques novices. Ces variables servent à la construction des axes.

NUAGE			AXE 1		AXE 2		AXE 3	
Variable	Modalité	Contrib	Contrib.	Rang	Contrib.	Rang	Contrib.	Rang
SEXE	Femme	1,9	1,4	21	0,8	20	0,4	28
	Homme	1,2	1	23	0,5	22	0,3	31
SEXE		3,0	2,4		1,3		0,7	
CSP	ETU	2,5	1,7	20	0,2	30	0,3	32
	FCT	2,9	0	43	0	38	0	41
	SAL	0,7	0,5	30	0,1	34	0,1	38
	SPR	2,9	0,1	37	0	43	0,2	35
CSP		9,0	2,3		0,4		0,6	
TOPVIT	TOPVIT1	2,7	12,1	1	5,7	9	0	42
	TOPVIT2	0,4	1,7	18	0,9	18	0	43
TOPVIT		3,1	13,8		6,6		0	
CLPRIX	ClPrix1	2,4	5,4	7	8,7	3	2,8	12
	ClPrix2	2,0	2,5	15	0,9	17	5,1	7
	ClPrix3	2,1	0,5	31	6	8	0,2	34
	ClPrix4	2,7	8,3	3	0,2	32	4,4	10
	ClPrix5	3,0	3,8	10	3,6	13	4,4	9
CLPRIX		12,2	20,6		19,5		16,8	
GPSRA	ClassA	2,2	6,4	5	6	7	0,8	23
	ClassB	3,0	0,1	38	0,5	25	0,8	24
	ClassC	1,9	0,6	29	8,1	4	6,4	5
	ClassD	2,1	7,2	4	0	39	8,7	4
	ClassE	3,0	4,2	8	10	1	10,4	3
GPSRA		12,1	18,4		24,7		27	

TABLE V – ACM : résultats sur les 3 axes principaux

Les résultats de l'ACM sur les 3 axes principaux apparaissent dans la Table V. Les valeurs indiquées sont les contributions et le classement de chacune des variables dans le nuage de points. La première constatation est le rôle non prépondérant variable SEXE dans cette analyse. Elle ne contribue que très faiblement à la constitution de chaque axe (jamais plus de 2,4 %) et seulement à 3 % sur l'ensemble du nuage. De même, si on observe le rang des modalités Femme/Homme on peut voir que leurs contributions ne se placent qu'en 21ème et 23ème position sur 43. Il semble donc difficile d'utiliser ces résultats pour constituer des classes dont le sexe serait l'élément discriminant. En revanche nous retenons de cette approche que les variables relatives aux prix des véhicules

(CLPRIX) et au groupe SRA (GPSRA) sont celles qui contribuent le plus à la constitution des axes. Finalement, cette analyse nous permet de relativiser le rôle discriminant que peut jouer la distinction homme/femme au sein de la population novice de notre portefeuilles. Cette méthode n'apporte pas de solution efficace pour la construction de proxies pour le sexe, mais elle montre que d'autres variables sont bien plus fédératrices.

Tableaux de contingence

Une autre solution envisagée est l'utilisation de la macro DESQUAL de l'INSEE : La macro édite les tableaux de contingence croisant une variable de classe (ici le sexe de l'assuré) avec chacune des variables qualitatives du portefeuilles automobile. Elle effectue des tests statistiques permettant de caractériser les classes de la partition par les modalités des variables explicatives. On obtient davantage de résultats avec cette approche. Pour chacune des modalités, la macro retient dans un premier temps une vingtaine de variables explicatives qu'elle classe par niveau de significativité.

Modalité	Variable	Effectif	Fréquence	Fréquence	proba	Val.Test
TOPVIT2	TOPVIT	8018	94,2	87,9	0,0000	23,6971
CLASSVEHA	gpsra	3069	36	27,9	0,0000	19,9448
ClPrix1	clprix	2138	25,4	18,5	0,0000	19,1433
ETU	CSP	2162	26,1	19,1	0,0000	19,1354
TOPSPORT2	TOPSPORT	8028	94,3	90,1	0,0000	16,9284
INJINDIRECTE	ALIM	5372	63,8	58,4	0,0000	12,3909
ESSENCE	ENERGIE	4748	55,8	50,5	0,0000	11,8445
SPR	CSP	478	5,8	3,8	0,0000	11,2274
ClPrix2	clprix	3376	40	35,3	0,0000	11,1154
CLASSVEHC	gpsra	3478	40,9	37	0,0000	8,956
...						

TABLE VI – Modalités sur-représentées chez les femmes

Modalité	Variable	Effectif	Fréquence	Fréquence	proba	Val.Test
SAL	CSP	13245	78,9	74,2	0,0000	23,989
CLASSVEHD	gpsra	6224	35,6	31	0,0000	23,942
TOPVIT1	TOPVIT	2661	15,2	12,1	0,0000	23,6971
CLASSVEHE	gpsra	812	4,6	3,5	0,0000	16,9379
TOPSPORT1	TOPSPORT	2091	12	9,9	0,0000	16,9284
ClPrix4	clprix	2814	16,4	14	0,0000	16,6754
ClPrix3	clprix	5568	32,5	29,9	0,0000	13,2938
INJDIRECTESUR	ALIM	5350	31	28,5	0,0000	12,9835
GASOIL	ENERGIE	9081	52	49,5	0,0000	11,8445
ClPrix5	clprix	503	2,9	2,3	0,0000	11,154
...						

TABLE VII – Modalités sur-représentées chez les hommes

Les Tables VI et VII contiennent les résultats d'abord pour les femmes puis pour les hommes. Dans la deuxième colonne on observe le classement des variables. La vitesse maximale du véhicule (TOPVIT), la catégorie socioprofessionnelle (CSP), le groupe SRA du véhicule (gpsra) et la classe de prix (clprix) sont les mieux placés, donc les plus discriminants à la fois pour les hommes et les femmes. L'ordre d'apparition est ensuite sensiblement le même pour les deux sexes. Pour voir apparaître les distinctions, il faut se référer à la première colonne de l'analyse qui caractérise à l'aide des modalités des variables les classes sur-représentées et sous-représentées chez les hommes et les femmes. Nous considérons ainsi le critère de vitesse réduite TOPVIT2, la classe de véhicules A (CLASSVEHA : inférieur à 7 600 euros) et le statut étudiant (ETU) comme plutôt féminin sur notre panel. Le statut de salariés (SAL), les classes D et E (CLASSVEHD, CLASSVEHE entre 10 500 et 13 000 euros) et les véhicules rapides TOPVIT1 caractérisent quant à eux davantage les hommes assurés chez Allianz.

Arbres de classification

L'idée inhérente à l'utilisation d'arbres est d'expliquer les modalités homme ou femme à partir de combinaisons de variables et non plus seulement par l'agencement de variables séparément significatives. Nous proposons donc ici de construire des arbres de décision selon la méthode CART, comme ceux proposés par A. Paglia et al. [60]. Ce type d'approche est aussi utilisé en assurance-vie pour expliquer le processus de décision d'un assuré voulant racheter son contrat [55]. Pour tester notre modèle sur le portefeuille Allianz, le logiciel gratuit Tanagra est utilisé. Plusieurs essais ont été réalisés sur différents types de variables et différentes tailles d'échantillons et même sur la totalité de la base de donnée puisque l'algorithme de Tanagra supporte une grande quantité de données. La procédure commence par créer l'arbre maximum en regroupant les modalités lorsque nécessaire, puis vient la phase d'élagage pour obtenir l'arbre le plus performant de la plus petite taille possible. Pour ce faire, on minimise le taux d'erreur de l'arbre dans sa phase de construction puis on se fixe un intervalle de confiance pour produire un arbre plus simple tout en conservant un bon niveau de performance. Les arbres qui suivent on été obtenus en sélectionnant les variables les plus significatives issues de la régression logistique et des macros de l'INSEE.

N	Nb. feuille(s)	Err (growing set)	Err (pruning set)	SE (pruning set)
67	1	0,3524	0,3559	0,0037
62	16	0,3164	0,3208	0,0036
59	22	0,3121	0,3187	0,0036
1	602	0,2685	0,3347	0,0037

Table VIII – Séquence de construction de l'arbre

La Table VIII présente les étapes successives de la construction de l'arbre. La performance maximum obtenue est d'environ 70 % de réussite (et ce à la fois sur la base qui sert de modèle et sur l'échantillon indépendant) mais il s'agit là d'une erreur individu

par individu et non d'une comparaison des erreurs globales comme c'est le cas pour la régression logistique. Ces résultats sont donc intéressants. L'arbre maximal obtenu compte 602 feuilles, la performance est alors de 27 % sur l'ensemble de construction et de 33 % sur l'ensemble d'élagage. L'arbre le plus performant selon les deux critères précédents compte lui 22 feuilles, mais on arrive à des performances équivalentes en élaguant jusqu'à ne conserver que 16 feuilles, le taux d'erreur est alors de 31,6 % sur l'ensemble de construction et de 32 % sur l'ensemble d'élagage. `Tanagra` nous indique que parmi les 50 000 observations dédiées à l'apprentissage, il a réservé 33 500 observations pour l'expansion de l'arbre (growing set), 16 500 pour le post élagage (pruning set). La partition a été effectuée de manière aléatoire.

L'arbre de classification associé à ce modèle est représenté dans la Figure 3.5. Sur les 21 variables descriptives retenues pour caractériser le sexe, le programme n'en retient que 11. Le premier noeud correspond à la variable `SEGMENT` qui serait donc la plus discriminante sur notre panel. La décision est prise directement si on n'appartient pas au `SEGMENT B`. Sinon, on regarde la `CSP` qui divise le panel en deux groupes de taille équivalentes, c'est ensuite la combinaison de choix parmi les classes de kilométrages, la marque, l'usage, la classe SRA, la vitesse, l'âge d'obtention du permis, la parenté avec un assuré, la puissance et le prix du véhicule qui répartissent les genres avec plus de complexité que l'approche des variables prises individuellement. Par exemple, une `FIAT` de classe de kilométrage `A` ou `F`, dont la vitesse n'est pas élevée et qui appartient à un salarié enfant d'assuré sera plutôt conduite par un homme, alors que les automobiles de la marque `FIAT` étaient plutôt féminines prises séparément. Les hommes qui sont plus nombreux dans le portefeuilles Allianz comptent 9 feuilles à leur actif pour seulement 7 pour les femmes.

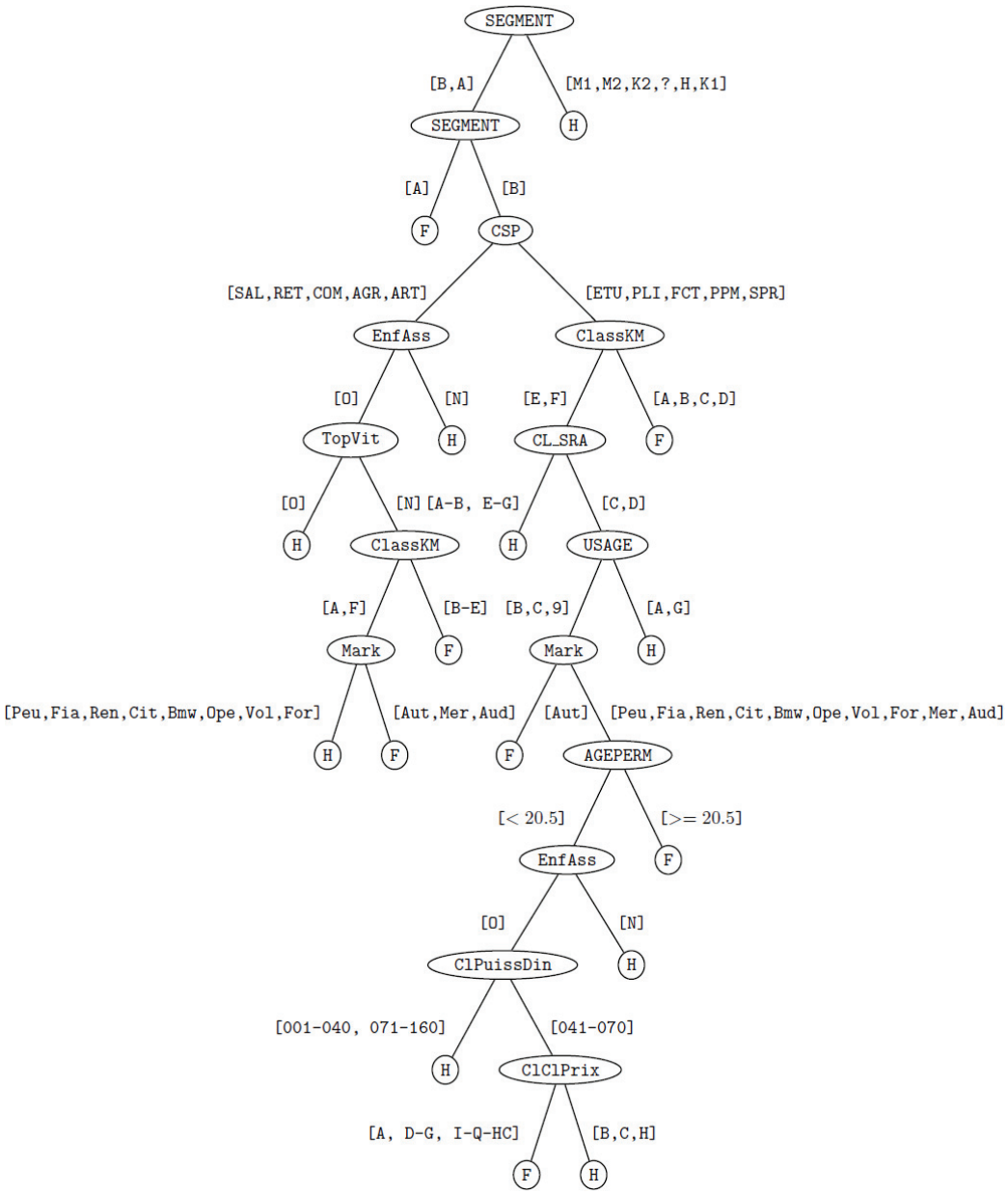


FIGURE 3.5 – Arbre de classification

On teste la solidité du modèle avec un échantillon test de 250 000 polices indépendantes des panels qui ont servis à la construction de l'arbre. La Table IX donne à la fois la précision et les valeurs de la matrice de confusion associée aux prédictions du modèle. Nous obtenons un taux d'erreur de 32%, ce qui est quasiment équivalent aux 31% d'erreurs de la procédure logistique présentée dans la section 3.1. En fin de compte, les différentes méthodes que nous avons envisagées pour tenter de retrouver le sexe des assurés à partir des autres variables en portefeuilles sont concluantes pour la majorité des individus mais demeurent imprécises. À travers cette exploration, nous avons décelé

d'autres variables qui ont aussi une place importante dans le portefeuilles. Il nous apparaît donc plus intéressant d'utiliser ces variables en tant que telles pour construire de nouveaux modèles plutôt que de les réduire à un substitut de la variable sexe désormais proscrite.

Taux d'erreur			0,3211			
Values prediction			Confusion matrix			
Value	Recall	1-Precision		H	F	Sum
H	0,8640	0,2929	H	140449	22108	162557
F	0,3347	0,4303	F	58176	29267	87443
			Sum	198625	51375	250000

Table IX – CART : résultats des tests

3.5 Création et validation d'un modèle sans sexe

Nous allons aborder ici la deuxième solution, et sans doute la meilleure, qui s'offre aux assureurs en réponse à l'interdiction d'utiliser le sexe dans leurs modèles de tarification. Notre approche utilise à la fois les GAM et les GLM. Les premiers servent à créer des classes à partir des variables numériques. Les deuxièmes nous permettent de tester le modèle ainsi créé et de valider les améliorations que nous apportons par rapport à un modèle classique.

3.5.1 Utilisation des GAM

Plusieurs paramètres comme l'âge, l'ancienneté du permis de conduire, le kilométrage parcouru ont une influence avérée sur l'expérience de l'assuré au volant de son véhicule. Ils peuvent avoir des répercussions sur sa sinistralité. Ces trois informations présentent l'avantage d'être quantifiables numériquement, ce qui les rends plus flexibles dans les modèles linéaires que nous utilisons. Nous poursuivons donc l'étude des risques à travers les variables explicatives numériques. Les modèles additifs généralisés (GAM) permettent une approche plus fine de l'influence de ces variables sur la sinistralité et aussi des éventuelles corrélations [46]. Ces méthodes sont assez largement utilisées dans le contexte de l'assurance automobile [67]. On propose ici des modèles univariés, l'objectif étant de mesurer la valeur explicative de la composante non linéaire (spline) de la variable. Lorsque le critère de 5 % de significativité est rempli, on modélise graphiquement la spline dont les variations nous permettront de délimiter des classes de risques plus homogènes que celles obtenues avec les modèles linéaires [66]. On commence par l'âge du conducteur modélisé selon les catégories de risques entre 2008 et 2010. Seules les catégories pour lesquelles la spline s'avère significative sont représentées.

La Table X présente les résultats du GAM pour le nombre total de sinistres en 2010. On a choisi de se focaliser sur la tranche 18 - 35 ans qui représente la majorité

Valeurs estimées des paramètres				
Paramètre	Valeur estimée des paramètres	Erreur type	Valeur du test t	$Pr > \|t\|$
Intercept	-1,08495	0,03504	-30,97	<,0001
Linear(AGECOND)	-0,03564	0,00157	-22,74	<,0001
Analyse du modèle de lissage				
Récapitulatif d'ajustement pour composantes du lissage				
Composante	Paramètre de lissage	DDL	GCV	Obs unique num
Spline(AGECOND)	0,323351	9,244633	0,001423	18
Smoothing Model Analysis				
Approximate Analysis of Deviance				
Source	DDL	Khi-2	$Pr >$ Khi-2	
Spline(AGECOND)	9,24463	197,4399	<,0001	

TABLE X – Analyse du modèle de régression

des novices (86 %), soit un panel d'environ 190 000 polices. Les tests permettent de conclure que pour cette garantie, l'âge du conducteur est à la fois significatif par sa composante linéaire et par sa composante spline qui est représentée ci-dessous.

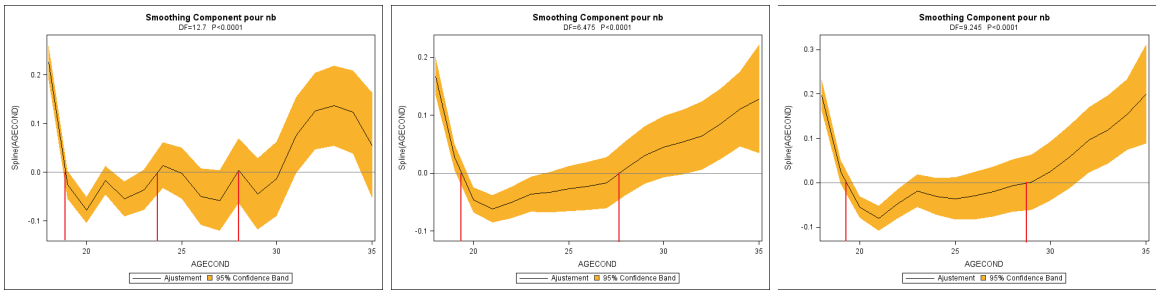


FIGURE 3.6 – Influence non linéaire de l'âge du conducteur sur la sinistralité totale entre 2008 et 2010

Les courbes que l'on observe sur les graphiques qui vont suivre correspondent à la partie non linéaire de chacune des variables explicatives. En effet, dans un GAM, l'influence de chaque variable s'exprime par la conjonction de deux composantes : linéaire et non linéaire. Ce sont les variations des courbes au dessus et en dessous de l'horizontale centrée en 0 qui nous informe sur l'impact positif ou négatif que va avoir la variable sur le résultat du modèle. Les lignes verticales délimitent ces variations et nous permettrons par la suite de définir plus efficacement les classes. Sur la Figure 3.6 on peut voir entre 2008 et 2010 l'évolution de la relation entre l'âge du conducteur et la sinistralité toutes catégories confondues. Selon l'étude graphique des splines de l'âge du conducteur, les variations les plus fortes correspondent aux tranches 18-19 ans et 19-23 ans. Mis à part en 2008, on a une certaine stabilité de 23 à 28 ans, l'augmentation est ensuite régulière après 28 ans. Pour les conducteurs les plus âgés dont la proportion augmente chaque année, on pourra se référer à l'étude américaine de l'Oak Ridge Institute [40].

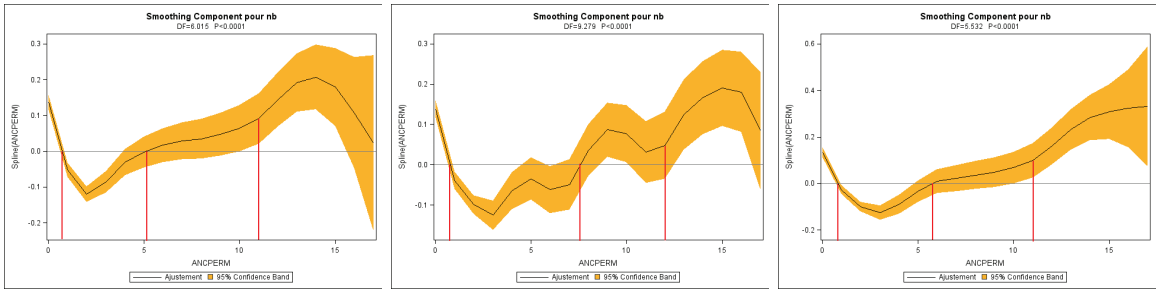


FIGURE 3.7 – Influence non linéaire de l'ancienneté du permis sur la sinistralité totale entre 2008 et 2010

Pour l'ancienneté du permis (Figure 3.7), les plus fortes variations sont entre 0 et 5 ans avec un minimum vers 3 ans, on a ensuite en 2008 et 2010 deux zones stables comprises entre 6 et 11, puis entre 11 et 15 ans de permis.

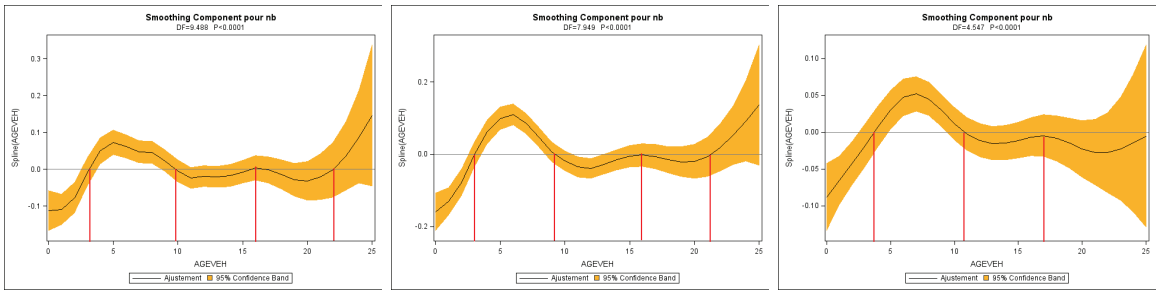


FIGURE 3.8 – Influence non linéaire de l'âge du véhicule sur la sinistralité totale entre 2008 et 2010

Pour l'âge du véhicule (Figure 3.8), on a choisi de se focaliser sur la tranche 0 - 25 ans qui représente la quasi totalité des novices (99 %). Les plus fortes variations sont entre 0 et 10 ans avec un maximum vers 6 ans, on a ensuite une zone stable comprise entre 10 et 17, au delà incertitude due à la faible représentativité l'emporte.

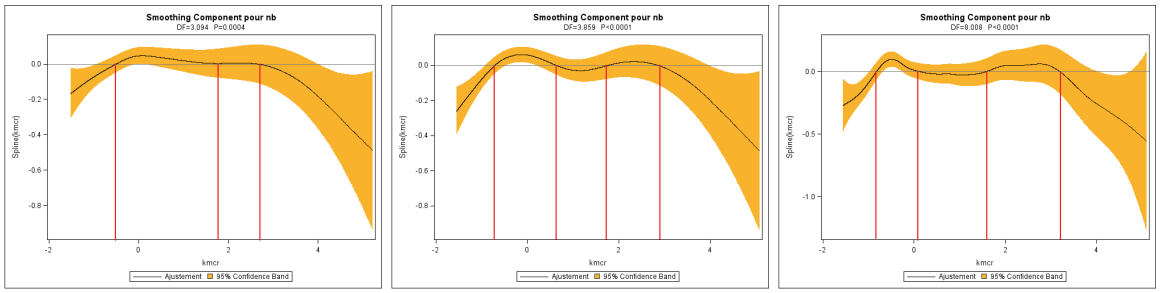


FIGURE 3.9 – Influence non linéaire du kilométrage annuel sur la sinistralité totale entre 2008 et 2010

En ce qui concerne le kilométrage (Figure 3.9), la quantité de données disponible est limitée par la présence d'un relevé SOFCA pour le véhicule. On récupère ainsi une moyenne de 60 000 assurés par ans. Le kilométrage annuel moyen se situe autour

de 5000 km et la grande majorité (99 %) des assurés parcourt moins de 30 000 km chaque année. Pour obtenir la convergence du GAM, nous avons dû utiliser des variables centrées réduites. La composante spline est alors significative et présente des variations notables entre -1 et 0,5 puis une certaine stabilité jusqu'à 1,5 et enfin une tendance linéaire après 3, ce qui correspond aux paliers réels 3000, 12 000, 18 000 et 27 000 km. Nous tenterons par la suite d'évaluer la part d'expérience acquise par les novices et donc la diminution de sinistralité induite selon leur tranche de kilométrage.

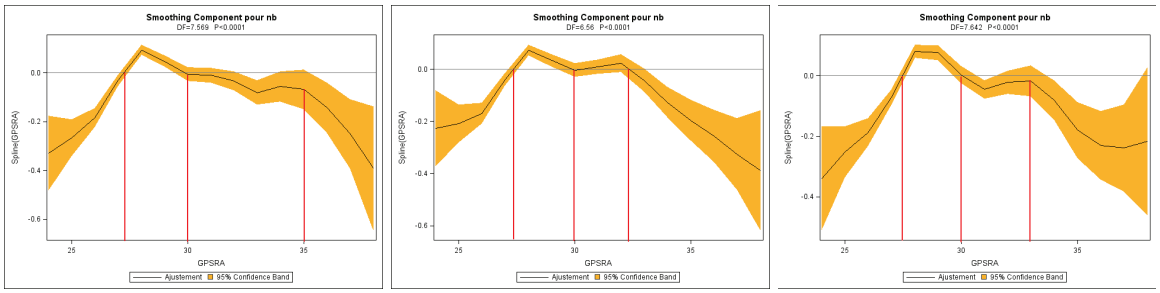


FIGURE 3.10 – Influence non linéaire du groupe SRA sur la sinistralité totale entre 2008 et 2010

Le groupe SRA (Figure 3.10) permet de classer les véhicules par type, on se focalise ici sur la tranche 25 - 35 qui représente la majorité des assurés. Sur les trois années, les courbes sont assez similaires avec une première tranche jusqu'à 28, puis à 30 et enfin après le groupe 33. Les classes que nous avons crées grace aux GAM sont différentes de celles que nous aurions choisies selon un découpage régulier ou proportionnel aux effectifs de chaque variable numérique. Nous espérons renforcer la significativité des variables catégorielles qui en découlent. Ces résultats sont exploités dans les modèles de la section suivante.

3.5.2 Utilisation des GLM

Différentes catégories de critères entrent en jeu lors de la création d'un modèle linéaire généralisé (GLM). Une bonne gestion de ces catégories doit permettre à l'assureur d'obtenir une modélisation à la fois flexible, précise et adaptée à sa tarification. On distingue d'abord les critères croisés comme la formule de garanties $\times$ l'âge du véhicule, la CSP $\times$ Usage, l'âge du conducteur $\times$ l'âge à l'obtention du permis, et les autres critères pris séparément comme les niveaux de franchises ou les antécédents de sinistralité et les critères SRA. On a ensuite des critères spécifiques aux novices et les informations issues de l'INSEE. Parmi ces variables, certaines sont numériques comme l'âge du conducteur ou le kilométrage d'autres, la plupart, sont catégorielles comme le sexe ou la CSP. Dans notre modèle, les variables numériques sont transformées en variables catégorielles dont les ajustements en différentes classes se font en accord avec des variations de la composante non linéaire des GAM réalisés en section 3.5.1. Avant toute chose, la multicolinéarité entre les différentes variables est testée par le facteur d'inflation de la variance (VIF) qui se base sur la méthodes des moindres carrés ordi-

naires. La Table XI présente la tolérance et la VIF pour les variables de base du modèle.

Variable	Tolérance	VIF
AGECOND	0,23759	4,20889
AGEVEH	0,87713	1,14008
ANCPERM	0,23957	4,17418
GPSRA	0,68133	1,46771
KMPar	0,97307	1,02768

TABLE XI – Facteur d'inflation de la variance : variables de base

Une tolérance inférieure à 0.2 correspondant à une VIF supérieure à 5 peut indiquer des problèmes de multicolinéarité. Dans la Table XI, on peut voir que lorsqu'on se limite aux variables de base il n'y a pas de multicolinéarité. En revanche lorsque l'on rajoute des variables spécifiques aux caractéristiques du véhicule (Table XII), la VIF dépasse le seuil de 5 et même 10, il faudra donc se limiter dans le choix des variables à utiliser dans nos modèles.

Variable	Tolérance	VIF
AGEVEH	0,72372	1,38176
GPSRA	0,09675	*10,33565*
KMPar	0,97453	1,02614
COUPLEMOTMAXI	0,87065	1,14856
VITMAXI	0,12134	*8,24098*
PTAC	0,4448	2,2482

TABLE XII – Facteur d'inflation de la variance : variables du véhicule

On choisit d'utiliser la procédure `GENMOD` du logiciel SAS pour modéliser l'ensemble des sinistres en RC responsable : RC0, non responsable : RC1 et bris de glace : BDG. La méthode retenue consiste à diviser notre panel d'environ 600000 polices en deux : une base de 400000 pour créer le modèle et un échantillon avec le reste des observations pour le tester. On se propose alors de comparer la solidité des modèles avec et sans la variable sexe, d'abord à partir des critères classiques comme la vraisemblance, et les critères d'information d'Akaike et bayésien, puis en appliquant les prédicteurs du modèle sur l'échantillon. Le portefeuille est aussi divisé selon que la police concerne un enfant d'assuré ou non, car les variables explicatives du modèle ne seront pas les mêmes.

On compare les modèles obtenus avec et sans la variable sexe du conducteur. Les critères d'évaluation de l'adéquation (Table XIII) indiquent d'abord une sous dispersion car la valeur/DDL de la déviance est faible (0.2 comparé à 1). On choisit ici `PSCALE` pour traiter les problèmes de dispersion. Une sous estimation des écarts types surestime les statistiques de test et augmente la significativité de nos variables explicatives. On introduire donc un terme de bruit qui correspond à la variance du nombre de sinistres non expliquée par les variables.

La qualité des deux modèles est proche avec un léger avantage pour le modèle sans sexe selon la vraisemblance (-24569,37 contre -24582,93), l'AIC (49408,75 contre 49437,87) et le BIC (50817,88 contre 50857,43).

		sans Sexe		avec Sexe	
Critere	DDL	Valeur	Valeur/DDL	Valeur	Valeur/DDL
Deviance	250000	54127,3863	0,2147	54078,5844	0,2145
Scaled Deviance	250000	37799,6975	0,15	37812,6824	0,15
Pearson Chi-Square	250000	360950,9351	1,432	360500,2271	1,4302
Scaled Pearson X2	250000	252069	1	252068	1
Log Likelihood		-24451,6207		-24465,0321	
Full Log Likelihood		-24569,3747		-24582,9329	
AIC (smaller is better)		49408,7494		49437,8658	
AICC (smaller is better)		49408,8951		49438,0136	
BIC (smaller is better)		50817,8786		50857,4329	

TABLE XIII – Critères d'évaluation de l'adéquation

La statistique LR pour Analyse de Type III (Table XIV et XV) donne les p-values de chaque variable indépendamment de leur ordre d'apparition. Il s'agit de voir d'une part si la variable SEXE est significative en terme de p-value (inf. à 5 %), d'autre part si son ajout au modèle diminue la significativité des autres variables.

Source	DDL	Khi-2	Pr>Khi-2	Khi-2	Pr> $Khi-2$
Cl_FormVeh	13	24,61	0,0259	24,1	0,0302
CSP	9	62,49	<,0001	53,45	<,0001
USAGE	4	68,86	<,0001	70,65	<,0001
Cl_age_perm	35	281,48	<,0001	270,38	<,0001
Cl_Energie	2	33,99	<,0002	33,58	<,0002
SEGMENT	7	22,32	0,0022	17,73	0,0132
Cl_PuissDin	15	37,45	0,0011	37,51	0,0011
Cl_Zone_RCDM	23	50,73	0,0007	49,41	0,0011
classkm	5	131,28	<,0001	126,9	<,0001
SEXE	*1*			*34,12*	<,0001
AncNov	2	35,07	<,0001	35,67	<,0001
FormPar	9	18,88	0,0263	18,8	0,027
DENSITE	5	44,6	<,0001	43,68	<,0001
Cl_NbAuto	5	145,1	<,0001	144,08	<,0001

TABLE XIV – RC1 : Statistique LR pour Analyse de Type III

Pour les sinistres responsables la variable SEXE est fortement significative, mais elle ne diminue que faiblement la p-value des autres variables. Seul le SEGMENT perd 1 % de significativité. La distinction homme/femme apporte donc une information utile dans ce cas mais non indispensable. Le modèle peut fonctionner sans cette variable tout en restant efficace.

Source	DDL	Khi-2	Pr>Khi-2	Khi-2	Pr> $Khi-2$
Cl_FormVeh	13	76,75	<,0001	76,3	<,0001
CSP	9	19,77	0,0194	19,66	0,0202
USAGE	4	56,12	<,0001	56,1	<,0001
Cl_age_perm	35	148,09	<,0001	147,39	<,0001
Cl_ClPrix	16	37,16	0,002	36,42	0,0025
Cl_Zone_RCDM	23	39,27	0,0185	39,27	0,0185
CRapPP	9	30,16	0,0004	30,16	0,0004
classkm	5	41,75	<,0001	41,72	<,0001
SEXE	*1*			*0*	*0,9681*
FormPar	9	22,13	0,0085	22,12	0,0085
TRPOP	5	28,41	<,0001	28,41	<,0001
CSPPLUS	4	13,14	0,0106	13,14	0,0106

TABLE XV – RC0 : Statistique LR pour Analyse de Type III

Pour les sinistres non responsables la variable SEXE n'est plus significative. On retrouve ici un résultat en accord avec l'approche graphique qui a été faite au début de cet article. Nous avions observés une sinistralité plus forte en terme de fréquence chez les femmes en cas de dommage non responsable tout en évoquant d'autres raisons à ce décalage. L'approche GLM nous permet de conclure que d'autres critères ont ici plus d'importance que le sexe pour expliquer la probabilité du sinistre.

	SINRC0	SINRC1	SINBDG
Modèles novices sans Sexe			
obs	3148	4124	3583
Pred	3180,36	4124,67	3512,6
Diff. Relative	1,028 %	0,016 %	1,965 %
Modèles novices avec Sexe			
obs	3148	4124	3583
Pred	3180,34	4126,66	3512,12
Diff. Relative	1,027 %	0,065 %	1,978 %

TABLE XVI – Résultats sur échantillon test

Après ces résultats relatifs à la construction du modèle, nous devons valider son efficacité. Nous utilisons alors un échantillon test indépendant des données ayant servi à calculer les prédicteurs pour vérifier la solidité du modèle. Le GLM pour les enfants d'assurés est obtenu sur une base de 250 000 polices. Il est ensuite testé sur un échantillon de 125 000 polices. Nous travaillons avec les sinistres en responsabilité civile non responsables (SINRC0) et les sinistres en responsabilité civile responsables (SINRC1) et les sinistres bris de glace (SINBDG). Les résultats apparaissent dans la Table XVI. Les prédictions du modèle sans la variable SEXE sont très proches de celles obtenues avec cette variable dans le cas non responsable (1,028 au lieu de 1,027% d'erreur) et sont même meilleures pour les garanties responsables (0,016 au lieu de 0,065% d'erreur)

et les bris de glace (1,965 au lieu de 1,978% d'erreur). Finalement, nous avons donc réussi à mettre en place un modèle dont les variables choisies de manières pertinentes permettent à la fois de respecter la réglementation européenne interdisant de distinguer les hommes et les femmes dans un tarif d'assurance et de prédire la sinistralité aussi efficacement, voire mieux qu'avec un modèle traditionnel.

3.6 Comment définir l'expérience des conducteurs novices ?

Les novices chez Allianz sont caractérisés par une durée d'assurance inférieure à 3ans. On se pose ici la question de l'expérience acquise par ces assurés au cours de leurs trois années de noviciat, et plus particulièrement l'évolution de leur sinistralité selon le nombre de kilomètres parcouru chaque année. D'autres rapports comme [4] ou [33] développent aussi la relation entre expérience et kilométrage dans le cadre du *Pay as You Drive*, mais nous nous focalisons ici sur les conducteurs novices. On peut en effet se poser la question de l'adaptation de cette période de 3 ans pour l'ensemble des assurés dont l'expérience au volant au cours de cette période est différente selon l'utilisation qu'ils font de leur véhicule [48]. D'après les résultats obtenus précédemment via les GAM, on choisit de créer différentes catégories de novices selon la durée mise par ces derniers pour passer le pallier des 12 000 km. Ce découpage particulier s'est avéré plus efficace qu'un découpage classique par tranche de kilomètres parcourus annuellement car il reflète davantage l'évolution de la conduite de l'assuré durant son noviciat. Les quatre catégories retenues sont les suivantes :

- `A` : Moins d'un an pour parcourir 12 000 km
- `B` : Moins de deux ans pour parcourir 12 000 km
- `C` : Moins de trois ans pour parcourir 12 000 km
- `D` : Plus de trois ans pour parcourir 12 000 km

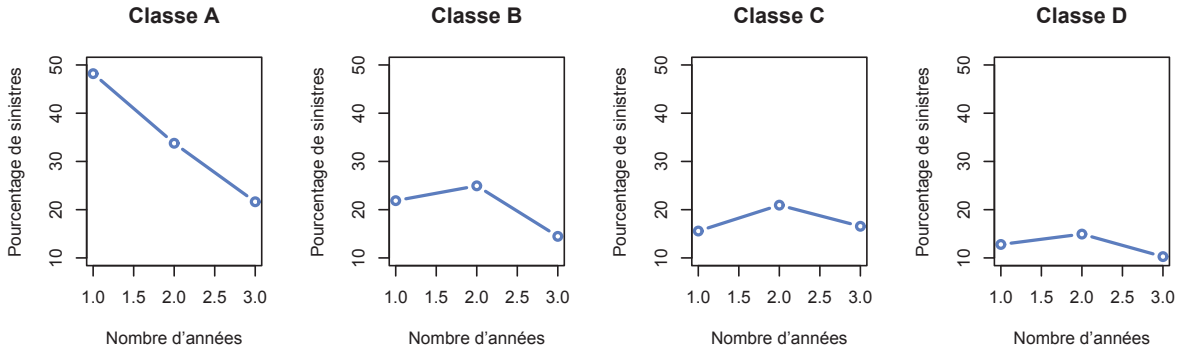


FIGURE 3.11 – Évolution de la sinistralité moyenne annuelle entre 2008 et 2010

Sur la Figure 3.11, on a représenté le nombre moyen de sinistres annuel. Un premier constat naturel est que les conducteurs parcourant un grand nombre de kilomètres

(classe A) sont plus exposés à la sinistralité et comptent donc un pourcentage de sinistres bien supérieur aux autres classes. Pour autant, ce sont les assurés de la classe A qui progressent le plus au cours de leurs trois années de noviciat avec une diminution de plus de 50 %. Pour les autres classes l'amélioration est beaucoup plus mesurée et passe même par une régression en cours de deuxième année. Cette légère augmentation de la sinistralité après un an de conduite pourrait s'expliquer par une prise de confiance prématurée de la part des novices qui roulent déjà depuis un certain temps mais n'ont pas parcourus assez de kilomètres pour minimiser leur exposition aux risques. On choisit donc d'observer à nouveau nos quatre classes mais cette fois-ci, en terme de nombre de sinistres par kilomètre parcouru pour faire apparaître l'exposition réelle du véhicule au risque sur une même distance.

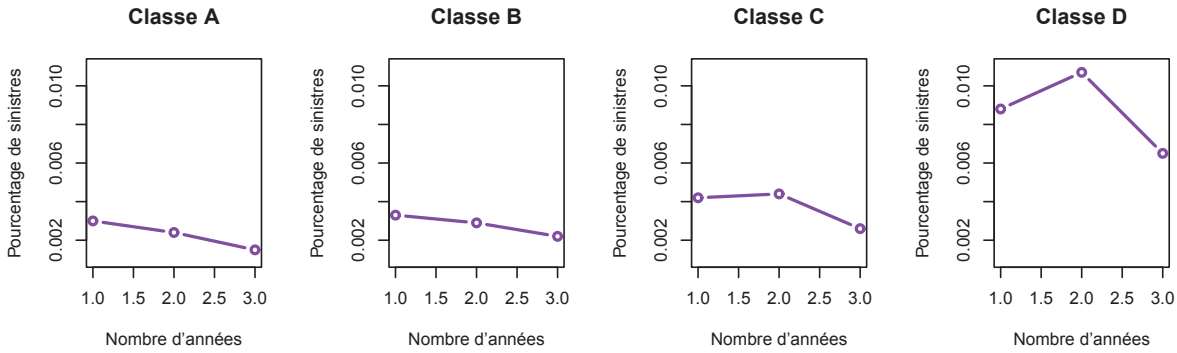


FIGURE 3.12 – Évolution de la sinistralité moyenne annuelle par kilomètre entre 2008 et 2010

Sur la Figure 3.12, la sinistralité moyenne est estimée en fonction du kilométrage. Une première constatation est l'inversion dans l'échelle de la sinistralité, la classe A devient la plus performante et la classe D présente un pourcentage de sinistres par kilomètre nettement supérieur aux trois autres classes. Concernant l'amélioration la classe A est aussi la meilleure avec une progression de 50 % alors que les classes B, C et D progressent respectivement de 33, 38 et 26 %. Les assurés qui roulent beaucoup deviennent donc a priori plus vite de bons conducteurs, ce qui pourrait justifier une sortie anticipée du noviciat, mais ils demeurent les individus les plus risqués pour l'assureur sur une période de contrat annuel. Autrement dit, S'il est clair, d'après les graphiques, que la classe A présente des fréquences plus faibles par km parcouru, la fréquence annuelle reste supérieure aux autres classes. Dans ce cas peut-on réellement parler de bons conducteurs ? Ou meilleurs que les autres ? Ce n'est le cas que si l'on fait une tarification au km.

En terme de garanties, on peut différencier les sinistres pour déterminer si cette diminution du risque s'avère plus importante lorsque l'assuré est responsable de son accident que lorsqu'il ne l'est pas. Nous représentons séparément les sinistres non-responsables et responsables.

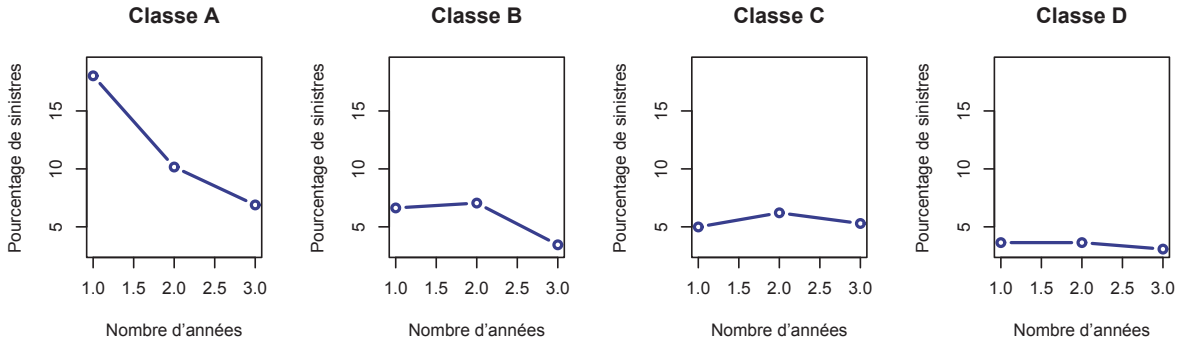


FIGURE 3.13 – Évolution du nombre moyen de sinistres responsables annuel entre 2008 et 2010

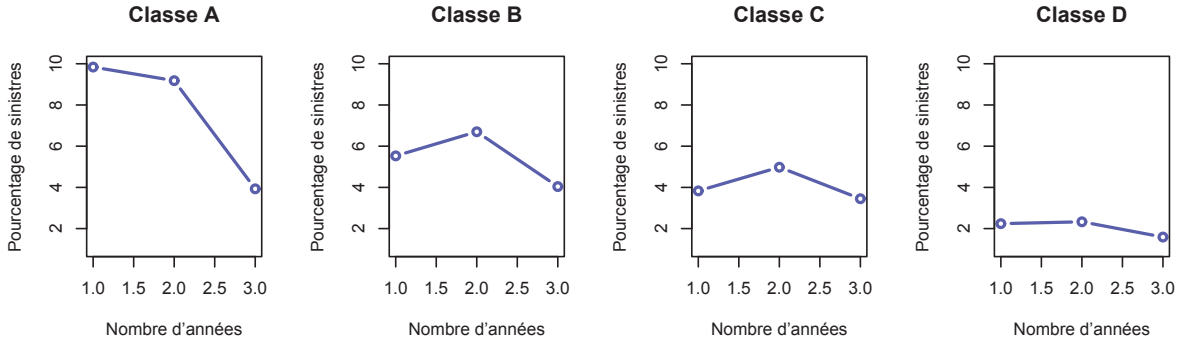


FIGURE 3.14 – Évolution du nombre moyen de sinistres non responsables annuel entre 2008 et 2010

En comparant les séries de graphiques 3.13 et 3.14, on constate une diminution plus rapide pour les sinistres responsables. La classe `A` en particulier obtient une courbe convexe dénotant une décroissance rapide dès la première année. Elle offre toujours la meilleure progression globale et se positionne en fin de période au même niveau de risque que les autres classes au début de leur noviciat. Les courbes sont toutes strictement décroissantes sauf celle de la classe `C` qui termine à peu près au même niveau de risque qu'elle a commencé. Les sinistres non responsables finissent par diminuer entre le début et la fin du noviciat, mais les courbes sont concaves pour l'ensemble des quatre classes. On note même entre la première et la deuxième année une légère augmentation de sinistralité pour les classes `B`, `C` et `D`. Ces résultats différents selon la nature du sinistre ne sont pas surprenants, une amélioration du comportement du conducteur sur la route lui évitera davantage d'être impliqué dans des accidents dont il est responsable, que non responsable. La maîtrise du véhicule permettrait donc avant tout de ne pas mettre en danger les autres usagers et dans une moindre mesure de diminuer son exposition aux risques.

On propose ensuite une nouvelle classification des conducteurs en lien direct avec leur kilométrage annuel. On commence par une approche statique, quelle que soit l'année

de noviciat, le conducteur se voit attribuer une classe allant de A à D correspondant aux tranches 0-3000, 3000-6000, 6000-12000 et plus de 12000 kilomètres. Il n'y a donc pas de suivi des assurés qui peuvent migrer d'une classe à l'autre au cours de leurs trois années d'observation. On compare le nombre moyen de sinistres selon les classes, année par année.

	Fréquence annuelle					Nombre d'assurés			
année	A	B	C	D		A	B	C	D
1	6,63 %	8,72 %	11,39 %	13,40 %		2424	5092	11219	7360
2	6,81 %	7,81 %	8,83 %	14,10 %		2645	4898	10545	7140
3	4,16 %	7,68 %	8,45 %	13,75 %		2676	4230	7429	3696

TABLE XVII – Sinistres responsables

	Fréquence annuelle					Nombre d'assurés			
année	A	B	C	D		A	B	C	D
1	5,83 %	5,65 %	8,34 %	9,68 %		2424	5092	11219	7360
2	5,26 %	5,89 %	8,36 %	12,79 %		2645	4898	10545	7140
3	2,77 %	4,37 %	6,76 %	9,56 %		2676	4230	7429	3696

TABLE XVIII – Sinistres non responsables

Dans les Tables XVII et XVIII on peut comparer les fréquences de sinistres et le nombre d'assurés de chaque catégorie de kilométrage selon la nature du sinistre (responsable ou non responsable). Les résultats de cette approche statique ne font que confirmer la plus faible sinistralité des classes de kilométrage les moins élevées. La troisième année de noviciat reste la meilleure en terme de diminution du risque, mais la progression la plus forte est observée chez les conducteurs de la classe A, ce qui, contrairement au résultat précédent, ne va pas dans le sens d'une meilleure maîtrise du véhicule lorsque l'on roule beaucoup. On peut noter que la diminution du risque est plus forte en cas de sinistre non responsable dans la classe A (-3,06 contre -2,47 %) et en cas de sinistre responsable dans la classe C (-2,94 contre -1,58 %). Cependant nous n'observons pas ici le kilométrage cumulé mais un état instantané qui ne reflète pas forcément l'expérience acquise par le conducteur au cours de ces trois ans. On opte alors pour une approche plus dynamique en considérant les tranches successives de kilométrage sur deux ans.

Les assurés sont ainsi classés en seize groupes de A à P, les tranches restant les mêmes, de 0 à 3000 km, de 3000 à 6000 km, de 6000 à 12000 km et plus de 12000 km, le kilométrage de la première année d'observation conditionnant l'appartenance au groupe de la deuxième année. Par exemple la classe C correspond à une distance parcourue comprise entre 0 et 3000 km la première année et entre 6000 et 12000 la deuxième. Les résultats suivants sont présentés dans quatre tableaux successifs pour les observations de sinistres responsables (Tables XIX et XX) et non responsables (Tables XXI et

1ère année	2ème année	Classe	Moy1	Moy2	nb
$0-3000$	$0-3000$	A	5,82 %	12,01 %	1132
	$3000-6000$	*B*	4,52 %	3,44 %	315
	$6000-12000$	*C*	10,19 %	7,33 %	195
	>12000	*D*	15,81 %	1,79 %	56
$3000-6000$	$0-3000$	*E*	3,64 %	2,29 %	313
	$3000-6000$	F	8,86 %	9,94 %	2150
	$6000-12000$	G	6,85 %	7,38 %	973
	>12000	*H*	12,95 %	9,28 %	245
$6000-12000$	$0-3000$	*I*	9,72 %	7,26 %	137
	$3000-6000$	*J*	6,81 %	6,75 %	594
	$6000-12000$	*K*	11,91 %	9,63 %	5865
	>12000	L	7,76 %	9,08 %	736
>12000	$0-3000$	*M*	15,02 %	0,00 %	20
	$3000-6000$	*N*	28,34 %	13,16 %	19
	$6000-12000$	O	35,38 %	36,64 %	55
	>12000	P	14,40 %	18,65 %	4070

TABLE XIX – Fréquence annuelle de sinistres responsables en première et deuxième année

XXII) des novices entre leur première et leur deuxième année puis entre leur deuxième et leur troisième année. On s'intéresse particulièrement aux catégories pour lesquelles on observe une diminution de la sinistralité d'une année sur l'autre. Ces groupes apparaîtront en italique.

2ème année	3ème année	Classe	Moy1	Moy2	nb
$0-3000$	$0-3000$	A	3,47 %	5,10 %	1176
	$3000-6000$	*B*	6,02 %	1,56 %	257
	$6000-12000$	C	3,22 %	10,86 %	141
	>12000	D	6,54 %	11,32 %	45
$3000-6000$	$0-3000$	E	0,77 %	2,90 %	261
	$3000-6000$	F	6,99 %	9,97 %	1891
	$6000-12000$	G	4,52 %	6,59 %	642
	>12000	*H*	8,52 %	8,37 %	123
$6000-12000$	$0-3000$	*I*	2,36 %	0,00 %	105
	$3000-6000$	*J*	3,34 %	2,60 %	444
	$6000-12000$	K	7,85 %	7,99 %	3932
	>12000	L	3,09 %	8,69 %	356
>12000	$0-3000$	M	0,00 %	0,00 %	8
	$3000-6000$	N	7,69 %	8,37 %	13
	$6000-12000$	O	4,17 %	6,65 %	32
	>12000	P	11,10 %	16,33 %	1811

TABLE XX – Fréquence annuelle de sinistres responsables en deuxième et troisième année

Une première constatation est que les effectifs stables (A, F, K, P) sont les plus représentés : les assurés ont donc tendance à rester dans la même tranches de ki-

lométrages d'une année sur l'autre. Les plus nombreux sur les deux premières années sont les groupes K et P qui correspondent aux tranches 6000-12000 km et plus de 12000 km avec combinés près de 10000 personnes. Entre la deuxième et la troisième année, les écarts sont moins marqués, le groupe K demeure le plus important. D'un autre côté les conducteurs ayant peu roulé une année et beaucoup la suivante (ou réciproquement) sont très peu nombreux. Les groupes D et M comptent seulement une cinquantaine et une vingtaine d'individus respectivement.

En ce qui concerne les sinistres responsables, on remarque que les groupes progressent plus au cours de la première que de la deuxième année. Parmi les groupes stables, seul le K présente entre la première et la deuxième année une diminution significative de la fréquence annuelle des sinistres (environ 2 %). Entre la deuxième et la troisième année de noviciat aucun grand groupe ne progressent et seuls les groupes B, H, I et J ont une diminution de la sinistralité. Contrairement à l'approche précédente, les plus gros rouleurs ne sont pas ceux qui progressent le plus, le groupe P notamment ne progresse ni en première ni en deuxième année.

1ère année	2ème année	Classe	Moy1	Moy2	nb
$0-3000$	$0-3000$	A	4,00 %	7,35 %	1132
	$3000-6000$	*B*	6,14 %	4,85 %	315
	$6000-12000$	*C*	5,23 %	2,34 %	195
	>12000	D	4,59 %	12,13 %	56
$3000-6000$	$0-3000$	*E*	5,17 %	3,39 %	313
	$3000-6000$	F	6,71 %	6,94 %	2150
	$6000-12000$	*G*	5,20 %	4,82 %	973
	>12000	H	12,15 %	12,98 %	245
$6000-12000$	$0-3000$	*I*	7,03 %	2,28 %	137
	$3000-6000$	J	4,35 %	4,84 %	594
	$6000-12000$	K	8,05 %	10,77 %	5865
	>12000	*L*	8,97 %	8,31 %	736
>12000	$0-3000$	M	0,00 %	0,00 %	20
	$3000-6000$	N	7,27 %	38,41 %	19
	$6000-12000$	O	10,12 %	42,18 %	55
	>12000	P	11,00 %	15,69 %	4070

TABLE XXI – Fréquence annuelle de sinistres non responsables en première et deuxième année

Pour les sinistres non responsables aucun des grands groupes ne progresse entre la première et la deuxième année, seuls quelques petits groupes (B, C, E, G, I, L) équitablement répartis dans les trois premières tranches de kilométrage montrent une légère diminution de la sinistralité. Entre la deuxième et la troisième année, on note davantage de progression, en particulier dans les grands groupes comme A, F et P. Avec cette catégorisation des assurés, on voit ressortir deux sortes d'évolutions selon la nature (responsable ou non) du sinistre. La deuxième année, les conducteurs des principales catégories progressent en moyenne moins bien en ce qui concerne leurs nombres de sinistres responsables, mais ils ont moins de sinistres non responsables. À travers

2ème année	3ème année	Classe	Moy1	Moy2	nb
$0-3000$	$0-3000$	*A*	3,22 %	2,12 %	1176
	$3000-6000$	*B*	4,11 %	4,08 %	257
	$6000-12000$	C	1,51 %	3,55 %	141
	>12000	*D*	14,37 %	4,45 %	45
$3000-6000$	$0-3000$	E	1,67 %	5,42 %	261
	$3000-6000$	*F*	5,87 %	4,47 %	1891
	$6000-12000$	G	4,36 %	5,29 %	642
	>12000	H	5,36 %	9,54 %	123
$6000-12000$	$0-3000$	I	4,72 %	4,76 %	105
	$3000-6000$	J	3,73 %	4,21 %	444
	$6000-12000$	K	6,91 %	7,28 %	3932
	>12000	*L*	8,09 %	6,61 %	356
>12000	$0-3000$	M	0,00 %	0,00 %	8
	$3000-6000$	N	0,00 %	7,69 %	13
	$6000-12000$	O	3,13 %	15,29 %	32
	>12000	*P*	11,26 %	10,66 %	1811

TABLE XXII – Fréquence annuelle de sinistres non responsables en deuxième et troisième année

ces différents résultats, on voit se dessiner plusieurs classes de risque d'une nouvelle nature faisant directement référence au comportement des assurés. Il serait intéressant de savoir dans quelle mesure ces résultats sont influencés par la nature des kilomètres parcourus (ruraux, urbains, autoroutes...). Y a-t-il des routes plus ou moins propices aux accidents ? Les informations de plus en plus nombreuses et précises à la disposition des assureurs devront permettre d'affiner ces recherches.

3.7 Conclusion

Pour répondre aux évolutions techniques et réglementaires du secteur automobile, les assureurs doivent proposer de nouvelles approches de tarification. Nous avons comparé différentes solutions qu'ils sont susceptibles d'envisager. L'utilisation d'un proxy pour pallier l'interdiction d'une variable est une méthode simple, mais nous avons montré les limites d'une telle démarche. Partant de ce constat, nous avons construit de nouveaux modèles faisant abstraction du sexe et nous les avons comparés aux modèles traditionnels. D'après nos résultats, la variable sexe, si elle permet de distinguer avec facilité des fréquences et des coûts de sinistres différents, ne s'avère pas indispensable à la construction d'un modèle de prédiction en assurance automobile. D'autres variables sur le type de véhicule, le cursus de l'assuré et son comportement en terme de kilométrage apportent une information aussi complète.

La dernière partie de cette étude se consacre à l'"expérience" de conduite des novices. Elle confirme l'utilité d'une information précise sur le kilométrage annuel de chaque assuré. Nous avons ainsi pu analyser l'évolution de la sinistralité des novices durant trois années consécutives. Les conducteurs ayant parcouru le plus de kilomètres progressent plus vite que les autres. Il faut donc différencier la fréquence annuelle de sinistres qui est naturellement plus élevée lorsque l'on conduit souvent, de la fréquence de sinistres par kilomètre qui reflète davantage la maîtrise du véhicule.

Ces constatations vont dans le sens des évolutions observées sur le marché autour de la conduite connectée et de la tarification au comportement, avec l'apparition de nouvelles options comme celle baptisée Allianz Conduite Connectée. Cette application utilise le GPS et les possibilités du téléphone mobile pour proposer de nouveaux services interactifs aux assurés. Le thème de la télématique dans le domaine de l'automobile est désormais d'actualité. La tarification ne se baserait plus seulement sur des caractéristiques du véhicule ou du conducteur mais aussi sur une mesure de la qualité et de la quantité des trajets effectués. Un exemple récent en est la compétition organisée sur le site de Kaggle par une compagnie d'assurance. À partir d'une base de données de plus de 50 000 trajets anonymes, les participants ont eu pour objectif d'identifier les conducteurs secondaires d'une voiture grâce à leur comportement routier. Cette expérience a permis le développement de signatures algorithmiques associées à des types de conduite. La personne au volant a t-elle tendance à favoriser les petits ou les longs trajets ? Fréquente-t-elle davantage les autoroutes ou les nationales ? Prend-elle les virages avec une vitesse élevée ? En combinant les réponses à ce genre de questions, il serait possible d'établir des profils quasiment uniques pour chaque conducteur.

L'utilisation de nouvelles sources d'information est possible à l'heure actuelle, mais elle ne va pas sans certaines difficultés de mise en place et sans certaines questions quant à la réglementation qui va accompagner leur diffusion. Pour récolter ces données, les assureurs pourraient s'associer directement aux constructeurs automobiles, aux fabriquants de GPS ou développer leurs propres capteurs télématiques. Dans cette

démarche, l'accueil du public à ces nouvelles pratiques sera déterminant. La confidentialité des données collectées est un point important. Une tarification à la fois compétitive et en adéquation avec les habitudes routières des assurés devrait ainsi voir le jour si les compagnies font bon usage de l'ensemble des informations potentiellement à leur disposition.

3.8 Lexique

Sinistre RC corporel non responsable	RCC0
Sinistre RC corporel responsable	RCC1
Sinistre IDA non responsable	IDA0
Sinistre IDA responsable	IDA1
Sinistre RC matériel non IDA non responsable	RCM0
Sinistre RC matériel non IDA responsable	RCM1
Sinistre Bris de glace	BDG
Sinistre Vol Incendie	VI
Sinistre Dommage non responsable	DOM0
Sinistre Dommage responsable	DOM1
Sinistre Assistance	ASSI

Table XXIII – Classification des sinistres

Variables dans le portefeuilles	
AGECOND	âge de la personne assurée
AGEVEH	ancienneté du véhicule
ANCPERM	nombre d'années depuis l'obtention du permis de conduire
ANCNOV	nombre d'années de noviciat
PUISSADM	puissance officielle du véhicule
NBPLACES	nombre de places
PUISS	puissance du véhicule
COUPLEMOTMAXI	taille du moteur du véhicule
TOPVIT	catégorie de vitesses
VITMAXI	vitesse maximale du véhicule
PTAC	poids total autorisé en charge
CDSEXCON	sexe du conducteur
CARROS	carrosserie
CSP	catégorie socioprofessionnelle
GPSRA	catégorie pour la sélection SRA
CDPRIME	catégorie de prime
CLPRIX	classe de prix du véhicule
CLASSKM	kilométrage annuel par catégorie
TXPRIME	taux de prime
ENERGIE	type d'énergie du véhicule
TRANSM	transmission
TYPE	type de véhicule
ALIM	alimentation
BOITEVIT	boîtier de vitesses
NBRAP	nombre de vitesses
SUSPENS	suspension
ASSISTFR	Assistance au freinage
ABS	anti-lock braking system
USAGE	type d'usage du véhicule
INSEE variables par communes	
TRPOP	Taille de la population
TRAGE	Age de la population
NAISSANCE	Taux de naissance
EVOL	Evolution de la population (2008/2010)
DENSITE	Nb habitants / superficie km^2
PCTCHGTLOGT	Pourcentage de changements de logement
PCTCHGTCOM	Pourcentage de changements de commune
PCTCHGTREG	Pourcentage de changements de région
PCTCHGTPAYS	Pourcentage de changements de pays

TABLE XXIV – Les variables explicatives utilisées

Conclusion

Nous avons construit et développé un indice tempête à partir de vitesses de vent que nous avons combinées avec l'exposition aux risques déterminée par les portefeuilles d'assurance. Nous avons d'abord optimisé notre approche en sélectionnant les données les plus appropriées, puis en testant les effets d'une paramétrisation. Nous avons aussi travaillé sur différentes échelles temporelles et géographiques pour mieux comprendre la relation entre les dommages enregistrés par l'assureur et l'ampleur de la tempête d'un point de vue météorologique. La sensibilité des résultats aux hypothèses de départ ainsi qu'aux méthodes employées pour évaluer les dégâts d'une tempête est un sujet important dans nos recherches. Nous voulons insister à travers les différents scénarios envisagés dans le deuxième chapitre sur la très forte variabilité des projections à partir du moment où elles concernent des événements extrêmes dont la période de retour dépasse la période d'observation dont nous disposons. Les simulations obtenues par des modèles du risque tempête, même en très grand nombre, ne peuvent pas remplacer des historiques incomplets et laissent donc une grande place à l'incertitude. L'utilisation de données complémentaires comme les vitesses de vent associées à l'expertise des assureurs pour établir les zones de risque reste néanmoins une bonne solution pour la gestion du risque tempête.

Dans le domaine de l'assurance comportementale, notre travail sur la garantie automobile a porté sur les évolutions techniques et réglementaires de la tarification. Dans un contexte de repositionnement des variables que l'assureur est en droit d'utiliser pour calculer le risque, nous avons montré que l'utilisation d'un proxy ne permet pas de pallier une interdiction. Nous avons proposé d'autres solutions en comparant de nouveaux modèles faisant abstraction de la variable sexe que nous avons comparés aux modèles traditionnels. Les résultats positifs que nous avons obtenus montrent que l'utilisation de variables complémentaires comme le kilométrage améliore les modèles de prédiction. Nous avons aussi étudié l'expérience de conduite des conducteurs novices. Plusieurs classes de risques ont étés déterminées selon l'évolution des distances parcourues et de la sinistralité durant leurs trois premières années en tant qu'assurés. Nous avons montré qu'en terme de sinistre par kilomètre les conducteurs roulant beaucoup dès la première année ont une meilleure marge de progression que les autres tout en restant à un niveau de sinistre annuel supérieur. La télématique offre au secteur de l'assurance automobile de nouvelles perspectives. La tarification peut désormais tenir compte de la qualité et de la quantité des trajets effectués mais doit aussi se fixer des limites quant aux informations personnelles qu'elle sera en mesure de demander aux assurés. Nous projetons sur ce thème un prochain article qui traiterait du *pay the way you drive*.

Annexe A

Quelques outils pour la Théorie des valeurs extrêmes

On expose ici plusieurs éléments classiques de la théorie des valeurs extrêmes [28]. Nous commençons par une présentation assez générale suivie de quelques résultats utiles sur la dépendance des extrêmes. Ces outils sont principalement employés dans la section 1.8.3 du chapitre 1 et dans les sections 3, 4 et 5 du chapitre 2.

A.1 Définition

On considère un échantillon aléatoire $\{X_i, 1 \leq i \leq n\}$ de fonction de distribution F. On s'interesse alors au maximum de cet échantillon : $X_{n,n}$. Le principal problème posé par la théorie des valeurs extrêmes est la recherche des distributions de X pour lesquelles il existe une suite de nombres $\{b_n, n \geq 1\}$ et une suite de nombres positifs $\{a_n, n \geq 1\}$ tels que :

$$P\left(\frac{X_{n,n} - b_n}{a_n} \leq x\right) \to G(x) \tag{A.1}$$

Comme dans le théorème central limit, on montre que toutes les distributions des valeurs extrêmes de type $G_\xi(x) = \exp(-(1+\xi x)^{-1/\xi})$ peuvent être une limite de (A.1). Une première condition nécessaire à l'application de la théorie est alors déterminée à partir de la fonction quantile $Q(y) = \inf\{x : F(x) \geq y\}$ ou de la fonction de la queue de quantile $U(y) = Q\left(1 - \frac{1}{y}\right)$. Il faut que

$$\lim_{x\to\infty} \{U(xu) - U(x)\}/a(x) = h_\xi(u) \tag{A.2}$$

pour une fonction positive a et $u > 0$, avec la fonction limite h non identiquement nulle. Enfin, la distribution F appartient au domaine d'attraction de G_ξ, $\mathcal{D}(G_\xi)$ si et seulement si pour une fonction auxiliaire b est $1 + \xi v > 0$

$$\frac{1 - F(y + b(y)v)}{1 - F(y)} \to (1 + \xi v)^{-1/\xi}. \tag{A.3}$$

Il existe différentes approches pour modéliser une distribution des valeurs extrêmes, nous présentons ici seulement les plus répandues. Dans le chapitre 2, nous avons opté pour la méthode du dépassement de seuil.

A.2 Approche par maximum de blocs

Une première approche de travail avec des données de valeurs extrêmes est de les grouper en blocs de longueur égale, et d'adapter la distribution aux maxima de chaque bloc. Le choix de la taille des blocs est importante. Il a été prouvé que les distributions des valeurs extrêmes sont les seules formes de limites pour un maximum normalisé d'un échantillon aléatoire, du moins lorsqu'une limite non dégénérée existe. Cette méthode est intimement liée à l'utilisation des familles de distributions des extrêmes généralisées (GEV). Cette généralisation dépend de 3 paramètres : μ pour la position, σ pour l'échelle et ξ pour la forme. Ce dernier indice s'avère être le paramètre le plus important en ce qui nous concerne car il détermine la forme de la queue de distribution. On définit successivement

$$G(z) = \exp\left(-\left[1 + \xi(\frac{z-\mu}{\sigma})\right]^{\frac{-1}{\xi}}\right) \qquad \text{si} \quad \xi \neq 0, \tag{A.4}$$

$$G(z) = \exp\left(-\exp\left[\frac{z-\mu}{\sigma}\right]\right) \qquad \text{si} \quad \xi = 0. \tag{A.5}$$

Cette approche est populaire dans les sciences environnementales, par exemple pour modéliser les températures maximales annuelles. On distingue trois types de lois selon le paramètre de forme ξ

1. si $\xi > 0$, F appartient au domaine d'attraction de Fréchet qui regroupe l'ensemble des lois à queue lourde, $\bar{F}(x)$ décroît comme une puissance de x à l'infini. Les distributions de Cauchy, Student, Pareto appartiennent à ce domaine,
2. si $\xi = 0$, F appartient au domaine d'attraction de Gumbel qui regroupe l'ensemble des lois à queue légère, $\bar{F}(x)$ décroît à vitesse exponentielle à l'infini. Par exemple les lois, Normale, Log-normale, Weibull, Gamma, Exponentielle,
3. si $\xi < 0$, F appartient au domaine d'attraction de Weibull qui regroupe l'ensemble des lois à queue finie, $\bar{F}(x) = 0$ pour $x > x_F$ appelé point terminal. Par exemple loi Uniforme et Beta.

A.3 Méthode du dépassement de seuil

La partie gauche dans A.3 peut être interprétée comme la fonction de survie conditionnelle des dépassements ($Y = X - t$) au delà d'un seuil t. À partir de la partie droite de la même équation apparaît une procédure statistique naturelle pour estimer la distribution $\bar{F}_t(y) \sim (1 + \xi y/b(t))^{-1/\xi}$. En interprétant $b(t)$ comme un paramètre d'échelle, on obtient la distribution généralisée de Pareto (GPD) :

$$1 - \left(1 + \frac{\xi y}{\sigma}\right)^{-1/\xi} \qquad \text{si} \quad \xi \neq 0, \tag{A.6}$$

$$1 - \exp\left(-\frac{y}{\sigma}\right) \qquad \text{si} \quad \xi = 0. \tag{A.7}$$

L'utilisation d'une GPD pour estimer les dépassements au delà d'un seuil peut aussi être motivée sur la base d'une caractérisation par processus ponctuel pour des dépassements élevés.

A.4 Processus ponctuels

Soit $\{X_i : i \in \mathcal{I}\}$ un échantillon aléatoire sur l'espace S. Un processus ponctuel N compte le nombre de points d'une région de S.

$$N(A) = \sum_{i \in \mathcal{I}} 1(X_i \in A)$$

Le nombre de points espérés est donné par la mesure d'intensité $\Lambda(A) = E[N(A)]$. Si Λ admet une densité λ et que l'espace S est euclidien alors $\Lambda(A) = \int_A \lambda(x)dx$, λ est appelé fonction d'intensité du processus. Le type le plus répandu de processus ponctuels correspond au processus de Poisson.

A.5 Approche par modèles multivariés

De nombreux problèmes modélisés par des valeurs extrêmes sont intrinsèquement multivariés. Qu'est-ce qui fait qu'une observation multivariée est extrême ? Suffit-il qu'une seule de ses coordonnées atteigne une valeur exceptionnelle ou doit-elle être extrême dans toutes ses dimensions ? Un problème fondamental qui apparaît dès que l'on s'intéresse à plus d'une variable est la dépendance. La théorie des extrêmes multivariés se divise donc en deux parties : les distributions marginales et la structure de dépendance.

Contrairement au cas univarié, les distributions extrêmes multivariées ne peuvent pas être représentées comme une famille paramétrique indexée par un vecteur de dimension finie. La raison est que la classe de dépendance est trop vaste.

A.6 Mesure caractéristique (exponent measure)

Pour étudier la structure de dépendance d'une distribution max-stable, il convient de standardiser les marginales. On commence par choisir des marginales de Fréchet standard. Soit $G_j^{\leftarrow}$ la fonction quantile de G_j, on rappelle que

$$G_j(x_j) = \exp\left(-\left[1+\xi_j(\frac{x_j-\mu_j}{\sigma_j})\right]_+^{\frac{-1}{\xi_j}}\right)$$

avec $x_j \in \mathbb{R}^d$, ξ, μ réels et $\sigma > 0$. On obtient

$$G_j^{\leftarrow}(e^{-1/z_j}) = \mu_j + \sigma_j \frac{z_j^{\xi_j}-1}{\xi_j}.$$

On peut alors écrire G_* la fonction de distribution de $(-1/\log G_1(Y_1), ..., -1/\log G_d(Y_d))$, telle que ses marginales soient de Fréchet standard

$$G(x) = G_*(-1/\log G_1(Y_1), ..., -1/\log G_d(Y_d)).$$

Soit μ_* la mesure caractéristique de G_*, on admet que μ_* est concentré sur $]0,\infty)$, tel que

$$\mu_*(]z,\infty)) = -\log G_*(z) = V_*(z).$$

Avec la fonction de dépendance de queue stable $l(v) = V_*(1/v_1, ..., 1/v_d)$, il est possible de reconstruire une fonction de distribution max-stable G à partir de ses marginales G_j en posant

$$-\log G(x) = l(-\log G_1(x_1), ..., -\log G_d(x_d)).$$

A.7 Mesure spectrale

Définissons la mesure S sur $\Xi = \mathbb{S}_2 \bigcap [0, \infty)$:

$$S(B) = \mu_*(z \in [0, \infty) : \|z\|_1 \geq 1, z/\|z\|_2 \in B)$$

La mesure S est appelée mesure spectrale, elle est déterminée uniquement par la mesure caractéristique μ_* et les deux normes choisies. La condition sur les marginales de G_* de Fréchet standard est équivalente à

$$\int_\Xi \frac{\omega_j}{\|\omega\|_1} S(d\omega) = 1, \qquad j = 1, ..., d.$$

Réciproquement, toute mesure S qui satisfait la condition ci-dessus est la mesure spectrale d'une distribution G_*. Et en terme de fonction de distribution G, on retrouve le lien avec les marginales via

$$\log G(x) = \int_\Xi \bigwedge_{j=1}^{d} \left\{ \frac{\omega_j}{\|\omega\|_1} \log G_j(x_j) \right\} S(d\omega), \qquad x \in \mathbb{R}^d.$$

A.8 Autres choix de marginales et copules

Même si le choix des distributions des marginales ne provoque pas de différences essentielles, certaines propriétés ou caractérisations sont plus naturellement visibles pour certains choix. On pourra par exemple considérer des marginales exponentielles, de Weibull, Gumbel ou Uniformes. À travers les marginales Uniformes, on retrouve la description de la structure d'une fdr multivariée via les copules. Il existe en effet, pour toutes fonctions F une fonction C_F à marginales uniformes sur $(0, 1)$ telle que

$$F(x) = C_F(F_1(x_1), ..., F_d(x_d)). \tag{A.8}$$

La copule d'une distribution des valeurs extrêmes multivariées G est donnée par

$$C_G(u) = \exp[-l(-\log(u_1), ..., -\log(u_d))]. \tag{A.9}$$

A.9 Aléa, période de retour et risque

On considère une suite $E_1, E_2, ...$ d'événements indépendants se produisant aux temps $t_1 < t_2 < ....$ Ces événements caractérisent une variable aléatoire $X \sim F$. On différencie $E_x^< = \{X \leq x\}$ et $E_x^> = \{X \geq x\}$ auxquels on associe $T_x^<$ et $T_x^>$ (les temps d'intervalle entre deux réalisations $E_x^<$ et $E_x^>$ successifs) et $N_x^<$ et $N_x^>$ (le nombre d'événements E_i entre deux réalisations). $N_x^<$ et $N_x^>$ suivent une loi géométrique de paramètres respectifs :

$$p_x^< = \mathbb{P}(E_x^<)$$
$$p_x^> = \mathbb{P}(E_x^>)$$

On définit les périodes de retour

$$\tau_x^< = \mathbb{E}(T_x^<) = \mathbb{E}(T_i)/p_x^<$$

$$\tau_x^> = \mathbb{E}(T_x^>) = \mathbb{E}(T_i)/p_x^>$$

Ce sont des nombres positifs correspondant au temps moyen écoulé entre deux réalisations successives d'un même événement. On note que $\tau_x^<$ et $\tau_x^>$ sont des fonctions décroissantes de $p_x^<$ et $p_x^>$. Lors d'application utilisant la méthode par bloc si l'on considère des maxima annuels, la taille du bloc sera 1 an et dans ce contexte $\mathbb{E}(T_i) = 1$. Le concept de risque combine à la fois l'occurrence d'un événement particulier et l'impact (ou les conséquences) de cet événement.

A.10 Bilan

La structure de dépendance d'une distribution max-stable peut être décrite de différentes façons, via une mesure caractéristique, une mesure spectrale, la fonction de queue de dépendance, la fonction de dépendance de Pickands, les copules etc. Ces quantités sont des objets de dimensions infinies et par conséquent pas toujours aisément utilisables. Une solution alternative est de résumer les principales propriétés de dépendance à travers des coefficients bien choisis.

Annexe B

Quelques outils pour l'analyse de données

Différentes méthodes sont employées dans le chapitre 3 pour décrire les relations entre les variables explicatives discrètes d'un modèle tarifaire en assurance automobile. Pour ce type de variables, nous avons exploré plusieurs procédures classiques de projections et de classifications. Ces techniques ne sont pas détaillées dans le chapitre mais nous les présentons brièvement dans cette annexe.

B.1 Analyse des correspondances multiples (ACM)

L'ACM permet techniquement de projeter et donc représenter un nuage de points initialement situé dans un espace de très grande dimension (le nombre de modalités moins le nombre de variables) dans un espace de dimension plus réduite. La distance des points deux à deux y est maximale, donc l'espace est celui qui conserve le mieux la richesse de l'information de départ [47]. On peut alors observer une classification des variables selon leur importance dans le nuage et sur chaque axe.

Considérons n individus décrits par p variables qualitatives ayant chacune $\{m_1, ..., m_p\}$ modalités. Dans la pratique, les calculs se font à partir d'une matrice appelée tableau disjonctif des indicatrices des variables. Ses valeurs propres sont notées μ_j et les coordonnées de chaque modalité sur un axe, a_j. La contribution d'une catégorie j d'effectif n_j est égale à

$$\frac{n_j \times a_j^2}{np \times \mu}.$$

On cherche les modalités dont la contribution est supérieure au poids, ce qui revient à observer sur la Figure B.1 les zones en dehors du rectangle bleuté où $|a_j| > \sqrt{\mu}$.

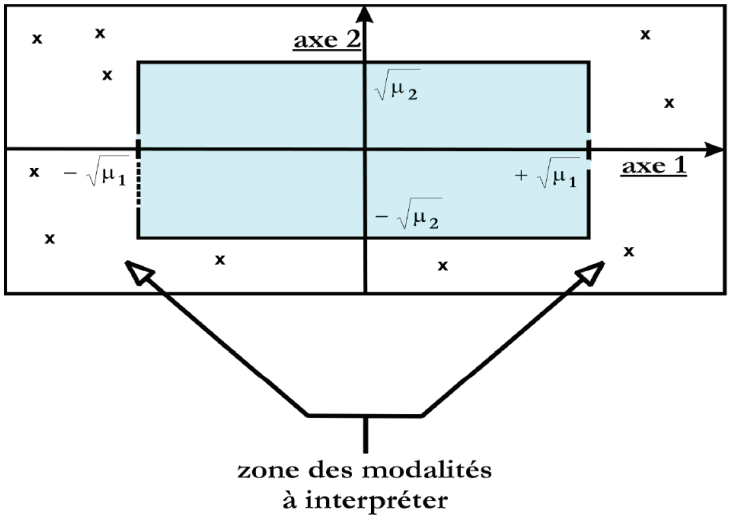


FIGURE B.1 – Représentation d'une ACM selon deux axes

Pour simplifier les calculs avec un grand nombre de variables, on distingue les variables actives qui déterminent les axes et les variables supplémentaires qui ne participent pas au calcul des valeurs propres et vecteurs propres. Dans notre article nous avons représenté ces catégories dans des tableaux séparés regroupant modalités, contribution et classement selon 3 axes.

B.2 Arbres de classification

Les arbres de classification nous permettent de créer des groupes homogènes en associant un certain nombre de modalités parmi les variables actives. Cette procédure a l'avantage de fournir des règles d'affectation utilisable à grande échelle. Les groupes sont interprétables directement. Enfin cette méthode est efficace pour traiter de grandes bases de données. Nous proposons dans notre article de construire des arbres de décision selon la méthode `CART` (Classification And Regression Tree), comme ceux proposés par A. Paglia et al. [60]. Par convention l'arbre est binaire pour éviter la fragmentation des données. Breiman et al. (1984) ont formulé les enjeux de la détermination de la taille **optimale** d'un arbre de décision. Avec la méthode `CART`, ils ont popularisé une approche, la construction en deux temps d'un arbre : **expansion - post-élagage**. La procédure commence par créer l'arbre maximum en regroupant les modalités lorsque nécessaire, puis vient la phase d'élagage pour obtenir l'arbre le plus performant de la plus petite taille possible. Comme pour l'ACM, on cherche à maximiser le gain d'inertie. Pour ce faire, on minimise le taux d'erreur de l'arbre dans sa phase de construction puis on se fixe un intervalle de confiance pour produire un arbre plus simple tout en conservant un bon niveau de performance. Dans la pratique, on sélectionne le plus petit arbre dont l'erreur n'excède pas 1 écart-type de l'erreur optimale.

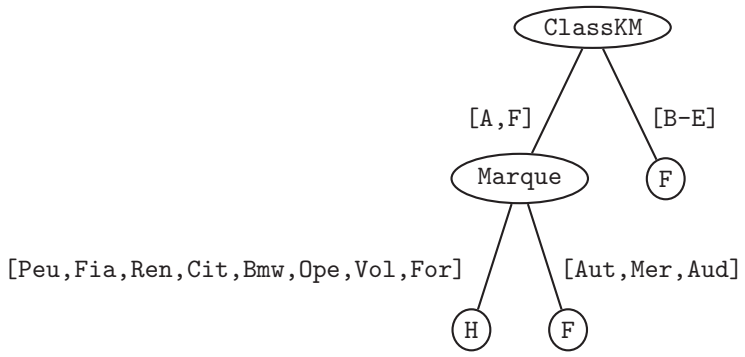


FIGURE B.2 – Extrait de l'arbre de classification

La Figure B.2 montre un extrait de l'arbre de classification que nous avons obtenu dans notre article [56]. Cet arbre compte deux noeuds (la classe de kilométrage et la maque du véhicule) et se termine par trois feuilles (F pour les femmes et H pour les hommes). D'après ce modèle on pourrait par exemple supposer qu'un assuré ayant un faible kilométrage annuel (classe A) et conduisant une Audi est une femme. Cependant, nous n'avons pas ici l'arbre complet et cette supposition mérite d'être complétée comme nous le verrons dans le chapitre 3.

B.3 Tableaux de contingence

Nous reprenons nos n individus et nous voulons tester la relation entre deux variables discrètes X et Y ayant respectivement k et p modalités. Le tableau de contingence I dénombre les modalités croisées de ces deux variables. Il compte donc k lignes et p colonnes. Les effectifs sont notés E.

	Y_1	.	.	Y_p	`Total`
X_1	E_{11}	.	.	E_{1p}	$E_{1.}$
.	.	.	.	.	.
.	.	.	.	.	.
X_k	E_{k1}	.	.	E_{kp}	$E_{k.}$
`Total`	$E_{.1}$	.	.	$E_{.p}$	$E_{..}$

TABLE I – Exemple de tableau de contingence

Comme ce tableau contient les effectifs bruts, il ne permet pas de comparer les proportions d'assurés selon la variable que l'on souhaite étudier. Par conséquent, nous avons présenté dans notre article des tableaux de profil indiquant les pourcentages et un test du χ^2 pour la significativité. Nous avons ainsi pu déterminer quelles caractéristiques étaient davantage sur-représentées ou sous-représentées chez les hommes et les femmes du portefeuille auto d'Allianz.

Bibliographie

[1] Insurance 2020 – innovating beyond old models. Technical report, IBM Global Business Services, 2006.

[2] Pay as you drive (payd) insurance pilot program phase 2 mid-course project report. Technical report, Progressive County Mutual Insurance Company, 2007.

[3] Use less, pay less - a simple concept that reduces the cost of car insurance now available to michigan and oregon drivers. Technical report, Progressive Direct, 2007.

[4] Pay-as-you-drive vehicle insurance - converting vehicle insurance premiums into use-based charges. Technical report, TDM Encyclopedia, Victoria Transport Policy Institute, 2014.

[5] P. Ailliot, J. Bessac, V. Monbet, and F. Pène. Non-homogeneous hidden markov-switching models for wind time series. pages 1–20, 2014.

[6] P. Ailliot, V. Monbet, and M. Prevosto. An autoregressive model with time-varying coefficients for wind fields. *Environmetrics*, 17(2) :107–117, 2006.

[7] Pierre Arnal and Romain Durand. Une vie sans sexes. comment le sexe devint genre, et comment le genre devint code : Le sort cruel d'une variable explicative. *Risques - Les cahiers de l'assurance*, 87 :1–7, 2011.

[8] T. Bjørnskau. A comparison of risks of different road user groups dependent on types of accident data. *Proceedings from 3rd International Conference on Safety and the Environment in the 21st Century*, pages 189–198, 1994.

[9] Jason E. Bordoff and Pascal J. Noel. Pay-as-you-drive auto insurance : A simple way to reduce driving-related harms and increase equity. *The Brookings Intitution*, pages 1–58, 2008.

[10] E. Brodin and H. Rootzén. Univariate and bivariate GPD methods for predicting extreme wind storm losses. *Insurance Math. Econom.*, 44(3) :345–356, 2009.

[11] Sandrine Carballes. Les véhicules particuliers en france. Technical report, ADEME, 2009.

[12] Pierre Carrega. Le vent : importance, mesures, modélisation et tribulations. *Bulletin de la Société géographique de Liège*, 51 :17–29, 2008.

[13] A.M. Chandler, E.J.W. Jones, and M.H. Patel. Property loss estimation for wind and earthquake perils. *Risk Analysis*, 21(2) :235–249, 2001.

[14] S. Coles. *An introduction to statistical modeling of extreme values*. Springer Series in Statistics. Springer-Verlag London Ltd., London, 2001.

[15] S. G. Coles. *An Introduction to Statistical Modeling of Extreme Values*. Springer, 2001.

[16] Stuart Coles, Janet Heffernan, and Jonathan Tawn. Dependence measures for extreme value analyses. *Extremes*, 2(4) :339–365, 1999.

[17] Dan Cooley, Philippe Naveau, and Paul Poncet. Variograms for spatial max-stable random fields. In *Dependence in probability and statistics*, volume 187 of *Lecture Notes in Statist.*, pages 373–390. Springer, New York, 2006.

[18] P. Deheuvels. La fonction de dépendance empirique et ses propriétés. Un test non paramétrique d'indépendance. *Acad. Roy. Belg. Bull. Cl. Sci. (5)*, 65(6) :274–292, 1979.

[19] P.M. Della-Marta, H. Mathis, C. Frei, M.A. Liniger, J. Kleinn, and C. Appenzeller. The return period of wind storms over europe. *Int. J. Climatol*, 29 :437–459, 2009.

[20] Philip K. Dick. *The Minority Report*. Fantastic Universe, 1956.

[21] Charles Dickens. *The Posthumous Papers of the Pickwick Club, Containing a Faithful Record of the Perambulations, Perils, Travels, Adventures and Sporting*. Chapman & Hall, 1837.

[22] Corona Direct. L'assurance au kilomètre. Technical report, Belgique, 2008.

[23] Matthias Dobbertin. Influence of stand structure and site factors on wind damage comparing the storms vivian and lothar. *For Snow Landsc Res*, 77(1/2) :187–205, 2002.

[24] 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*, 11(10) :2821–2833, 2011.

[25] C. Dorland, R.S.J. Tol, and J.P. Palutikof. Vulnerability of the netherlands and northwest europe to storm damage under climate change. *Climatic Change*, 43(3) :513–535, 1999.

[26] C. Douvillé and E. Burbaud. Tempêtes, grêle et neige : Résultats 2010. Technical report, Association française de l'assurance, FFSA and GEMA, 2012.

[27] A.E. Drummond and E.-Y. Yeo. The risk of driver crash involvment as a function of driver age. *Monash Universtity Accident Research Center*, 49 :1–32, 1992.

[28] P. Embrechts, C. Klüppelberg, and T. Mikosch. *Modelling extremal events*, volume 33 of *Applications of Mathematics (New York)*. Springer-Verlag, Berlin, 1997. For insurance and finance.

[29] Christopher AT Ferro and Johan Segers. Inference for clusters of extreme values. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 65(2) :545–556, 2003.

[30] Hayley J. Fowler, Daniel Cooley, Stephan R. Sain, and Milo Thurston. Detecting change in UK extreme precipitation using results from the climmate*prediction*.net BBC climate change experiment. *Extremes*, 13(2) :241–267, 2010.

[31] Philippe Gattet. Assurance et big data - opportunités et nouveaux écosystèmes. Technical report, Precepta, 2014.

[32] Eric Gilleland, Mathieu Ribatet, and AlecG. Stephenson. A software review for extreme value analysis. *Extremes*, 16(1) :103–119, 2013.

[33] Allen Greenberg. Designing pay-per-mile auto insurance regulatory incentives using the nhtsa light truck cafe rule as a model. Technical report, AIP, 2009.

[34] Xavier Guyon. Statistique spatiale. Technical report, Conférence S.A.D.A. 07 - Cotonou - Benin, 2007.

[35] Peter Hall. Asymptotic properties of the bootstrap for heavy-tailed distributions. *The Annals of Probability*, pages 1342–1360, 1990.

[36] M.J. Heaton, M. Katzfuss, S. Ramachandar, K. Pedings, E. Gilleland, E. Mannshardt-Shamseldin, and R.L. Smith. Spatio-temporal models for large-scale indicators of extreme weather. *Environmetrics*, 22(3) :294–303, 2011.

[37] P. Heneka, T. Hofherr, B. Ruck, and C. Kottmeier. Winter storm risk of residential structures model development and application to the german state of baden-wurttemberg. *Nat. Hazards Earth Syst. Sci.*, 6 :721–733, 2006.

[38] S. Hochrainer-Stigler and G. Pflug. Risk management against extremes in a changing environment. *Environmetrics*, 23(8) :663–672, 2012.

[39] Brian J Hoskins and Kevin I Hodges. New perspectives on the northern hemisphere winter storm tracks. *Journal of the Atmospheric Sciences*, 59(6) :1041–1061, 2002.

[40] Patricia S. Hu, Donald W. Jones, Timothy Reuscher, Richard S. Schmoyer Jr., and Lorena F. Truett. Projecting fatalities in crashes involving older drivers, 2000-2025. Technical report, Oak Ridge National Laboratory, 2000.

[41] R.W. Katz. Extreme value theory for precipitation : sensitivity analysis for climate change. *Advances in Water Resources*, 23 :133–139, 1999.

[42] R.W. Katz and B.G. Brown. Extreme events in a changing climate : Variability is more important than averages. *Climatic Change*, 21 :289–302, 1992.

[43] Leonard Kaufman and Peter J Rousseeuw. *Finding groups in data : an introduction to cluster analysis*, volume 344. John Wiley & Sons, 2009.

[44] R. Killick, P. Fearnhead, and I. A. Eckley. Optimal detection of changepoints with a linear computational cost. *J. Amer. Statist. Assoc.*, 107(500) :1590–1598, 2012.

[45] 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*, 3 :725–732, 2003.

[46] Nadja Klein, Michel Denuit, Stefan Lang, and Thomas Kneib. Nonlife ratemaking and risk management with Bayesian generalized additive models for location, scale, and shape. *Insurance Math. Econom.*, 55 :225–249, 2014.

[47] Romuald Le Lan. Analyse de données and classification sur données d'enquête - choix sur les variables, le nombre de classes and le nombre d'axes. Technical report, DREES, Bureau des professions de santé, 2003.

[48] Jean Lemaire, Sojung Park, and Kili Wang. The use of annual mileage as a rating variable. *STAT Discussion Paper*, pages 1–26, 2014.

[49] MLR Liberato, JG Pinto, RM Trigo, P Ludwig, P Ordóñez, D Yuen, and IF Trigo. Explosive development of winter storm xynthia over the subtropical north atlantic ocean. *Natural Hazards and Earth System Science*, 13(9) :2239–2251, 2013.

[50] Todd Litman. Pay-as-you-drive pricing in british columbia. *Victoria Transport Policy Institute*, pages 1–11, 2011.

[51] M. Luzi. Travaux sur les historiques tempêtes. *Directeur IARD - Allianz*, 2013.

[52] Claude Lévi-Strass. *Tristes Tropiques*. PLON, 1955.

[53] Cameron A MacKenzie. Summarizing risk using risk measures and risk indices. *Risk Analysis*, 34(12) :2143–2162, 2014.

[54] L.Dawn Massie, Paul E. Green, and Kenneth L. Campbell. Crash involvement rates by driver gender and the role of average annual mileage. *Accid. Anal. and Prev.*, 29 :675–685, 1997.

[55] Xavier Milhaud, Stéphane Loisel, and Véronique Maume-Deschamps. Facteurs explicatifs du rachat en Assurance-Vie : classification and prévisions du risque de rachat. In *42èmes Journées de Statistique*, Marseille, France, France, 2010.

[56] Alexandre Mornet, Patrick Leveillard, Michel Luzi, and Stéphane Loisel. Comment répondre aux évolutions techniques et réglementaires de la tarification en assurance automobile ? *Bulletin Français d'Actuariat*, 15(29) :75–112, 2015.

[57] Alexandre Mornet, Thomas Opitz, Michel Luzi, and Stéphane Loisel. Index for predicting insurance claims from wind storms with an application in france. *Risk Analysis*, 2015.

[58] ONERC : Observatoire national sur les effets du réchauffement climatique. Evaluation du coût des impacts du changement climatique et de l'adaptation en france. Technical report, Ministère de l'Écologie, de l'Énergie du Développement durable et de la Mer, 2009.

[59] M. Oesting, M. Schlather, and P. Friedrichs. Conditional modelling of extreme wind gusts by bivariate brown-resnick processes. *arXiv*, pages 1–22, 2013.

[60] Antoine Paglia and Martial V. Phelippe-Guinvarc'h. Tarification des risques en assurance non-vie, une approche par modèle d'apprentissage statistique. *Bulletin Français d'Actuariat*, 11(22) :49–81, 2011.

[61] J. Papillon. Les appareils de mesure de la vitesse du vent. anémomètres de pression et anémomêtres de vitesse. description des anémomètres actuellement utilisés par l'office national météorologique : Anémomètre électromagnétique papillon ; anémomètre magnétique à main richard. *Annales Francaises de Chronometrie*, 9 :289–303, 1939.

[62] Raymond Pearce, David Lloyd, and David McConnell. The post-christmas 'french'storms of 1999. *Weather*, 56(3) :81–91, 2001.

[63] R.A.Jr. Pielke, J. Gratz, C.W. Landsea, D. Collins, M.A. Saunders, and R. Musulin. Normalized hurricane damage in the united states : 1900-2005. *Natural Hazards Review*, 9(1) :29–42, 2008.

[64] J.G. Pinto, E.L. Fröhlich, G.C. Leckebusch, and U. Ulbrich. Changing european storm loss potentials under modified climate conditions according to ensemble simulations of the echam5/mpi-om1 gcm. *Natural Hazards and Earth System Science*, 7 :165–175, 2007.

[65] Joaquim G Pinto, Melanie K Karremann, Kai Born, Paul M Della-Marta, and Matthias Klawa. Loss potentials associated with european windstorms under future climate conditions. *Climate Research*, 54(1) :1–20, 2012.

[66] Sandra Pitrebois, Michel Denuit, and Jean-François Walhin. Personnalisation des primes-fréquence en assurance automobile par régression poissonienne en présence de données longitudinales. *STAT Discussion Paper*, pages 1–29, 2001.

[67] V. Pouna Siewe. Modèles additifs généralisés : Intérêts de ces modèles en assurance automobile. *Mémoire d'actuariat - ISFA*, pages 1–85, 2010.

[68] F. Prettenthaler, H. Albrecher, J. Köberl, and D. Kortschak. Risk and insurability of storm damages to residential buildings in austria. *The Geneva Papers*, 37 :340–364, 2012.

[69] Weihong Qian, Lingshen Quan, and Shaoyin Shi. Variations of the dust storm in china and its climatic control. *Journal of Climate*, 15(10) :1216–1229, 2002.

[70] R Core Team. *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015.

[71] Roger Roots. The dangers of automobile travel : A reconsideration. *American Journal of Economics and Sociology*, 66 :959–975, 2007.

[72] H. Rootzén and N. Tajvidi. Can losses caused by wind storms be predicted from meteorological observations ? *Scand. Actuar. J.*, (2) :162–175, 2001.

[73] I. Rychlik and W. Mao. Probabilistic model for wind speed variability encountered by a vessel. *Matematiska Vetenskaper*, pages 1–23, 2014.

[74] C. Sacré, J.M. Moisselin, M. Sabre, J.P. Flori, and B. Dubuisson. A new statistical approach to extreme wind speeds in france. *J. Wind Eng. Ind. Aerodyn.*, 95 :1415–1423, 2007.

[75] C. Scarrott and A. MacDonald. A review of extreme value threshold estimation and uncertainty quantification. *Revstats*, 10(1) :33–60, 2012.

[76] I. Scheel, E. Ferkingstad, A. Frigessi, O. Haug, M. Hinnerichsen, and E. Meze-Hausken. A Bayesian hierarchical model with spatial variable selection : the effect of weather on insurance claims. *J. R. Stat. Soc. Ser. C. Appl. Stat.*, 62(1) :85–100, 2013.

[77] André Thépaut. Quelques réflexions sur la réforme du tarif français d'assurance automobile. *Astin*, 2 :109–124, 1961.

[78] P. Thomson, B. Mullan, and S. Stuart. Estimating the slope and standard error of a long-term linear trend fitted to adjusted annual temperatures. *Statistics Research Associates Ltd*, 2014.

[79] Jean-Claude Thouret and Robert d'Ercole. Vulnérabilité aux risques naturels en milieu urbain : effets, facteurs et réponses sociales. *Cahiers des sciences humaines*, 32(2) :407–422, 1996.

[80] Carmela Troncoso, George Danezis, Eleni Kosta, Josep Balasch, and Bart Preneel. Pripayd : Privacy-friendly pay-as-you-drive insurance. *Dependable and Secure Computing, IEEE Transactions on*, 8(5) :742–755, 2011.

[81] Robert L. Winkler. The importance of communicating uncertainties in forecasts : Overestimating the risks from winter storm juno. *Risk Analysis*, 35(3) :349–353, 2015.

[82] S. Zahran, D. Tavani, and S. Weiler. Daily variation in natural disaster casualties : Information flows, safety, and opportunity costs in tornado versus hurricane strikes. *Risk Analysis*, 33(7) :1265–1280, 2013.

[83] Chunsheng Zhao, X Dabu, and Ying Li. Relationship between climatic factors and dust storm frequency in inner mongolia of china. *Geophysical Research Letters*, 31(1), 2004.