INSTITUT DES ACTUAIRES

ISUP

Mémoire présenté le :
pour l'obtention du diplôme
de Statisticien Mention Actuariat
et l'admission à l'Institut des Actuaires

Par : Monsieur Sami BECK

Titre du mémoire : Modélisation du risque de grêle en France pour une compagnie d'assurance non vie

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

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

Membres présents du jury de l'Institut des Actuaires

signature

Membres présents du jury de la filière

Entreprise :

Nom : ***Pacifica***

Signature :

Directeur de mémoire en entreprise :

Nom : ***Valery CAUSSARIEU***

Signature :

Invité :

Nom :

Signature :

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

Signature du responsable entreprise

Signature du candidat

Résumé

Mots clés : grêle, SCR, modèle CAT, survenance, vulnérabilité, copules, régression logistique, *Random Forest*, garantie TGN.

La grêle est un risque atypique pouvant causer ponctuellement et occasionnellement des dégâts considérables sur tout type de biens. Les évènements majeurs de Juin 2013 et de la Pentecôte de 2014 ont représenté des charges considérables pour les compagnies d'assurances témoignant d'une augmentation de l'intensité du phénomène en raison du changement climatique. Cette évolution interroge les assureurs sur la possibilité de mettre en place de nouvelles méthodes tenant compte des évolutions climatiques afin de modéliser ce risque.

La réforme de Solvabilité 2 impose notamment le calcul du SCR (Solvency Capital Requirement) qui est le capital nécessaire afin que la probabilité de défaut à horizon 1 an de l'assureur soit inférieure à 0,5%. Le calcul du SCR par le biais d'un modèle interne offre la possibilité de refléter le profil de risque réel de l'assureur et de s'adapter à des changements de structure de portefeuille.

Une modélisation précise et rigoureuse du péril grêle pour une compagnie d'assurance est donc un enjeu clé pour la solvabilité des assureurs.

Dans ce mémoire, nous caractériserons les portefeuilles d'assurance automobiles et habitation de Pacifica. Ensuite, Nous mettrons en parallèle les données climatiques à la maille la plus fine disponible et la sinistralité du portefeuille Pacifica afin d'identifier les variables responsables de la survenance de grêle. Nous simulerons ensuite la sinistralité en tenant compte de la dépendance entre les différentes variables.
Enfin, nous étudierons l'impact de la grêle (fréquence de biens touchés, taux de destruction...) sur les différents biens caractérisés. L'objectif final sera de modéliser la charge annuelle et la charge par évènement afin de définir la meilleure couverture de réassurance possible pour Pacifica.

Abstract

Mots clés : Hail, CAT modeling, occurrence, vulnerability, copula, Generalized Linear Model, *Random Forest*.

Hail is an atypical risk which may cause occasionally considerable damage on a wide number of goods. Major Events of June 2013 and the 2014 Pentecôte represented important losses for insurance companies showing that the augmentation of hail events intensity in the past twenty years may be caused by global warming. This evolution lead insurers to think about new ways to model the risk considering global warming and its impact on other environmental parameters.

In this thesis we will consider Automotive and Property portfolios. Then, we will use the most accurate climate data available in perspective of the claim data from Pacifica in order to identify variables responsible of hail occurrence. We will then simulate next year future claims taking into account dependency among variables.

We will use Hail impact (frequency of insured goods damaged, destruction rates ...) in function of all goods characterized. The aim is to deduce the aggregate exceedance probability (AEP) and the occurrence exceedance probability (OEP) in order to define the best reinsurance structure for Pacifica.

Note de synthèse

Les épisodes de grêle se produisent en général lors d'orages violents formés par des unités dynamiques appelées cellules. Lors de la présence de couches humides dans la basse troposphère et d'un taux de changement de température assez élevé, les orages accumulent beaucoup d'énergie convective potentielle (CAPE) et peuvent devenir potentiellement « violents ». C'est dans ces courants ascendants assez forts que les grêlons se forment avant de tomber au sol sous forme d'averses de grêle.

En France, l'Anelfa (Association Nationale d'Etude et de Lutte contre les Fléaux Atmosphériques) a effectué des relevés au sol de chutes historiques de grêle sur une vingtaine de départements. Les résultats montrent que les chutes de grêles les plus destructrices se produisent dans leurs grandes majorités durant les mois d'été entre le début d'après-midi et le début de soirée. Dans sa thèse "Le risque-grêle en France étude géographique", Freddy Vinet corrobore ces observations en distinguant les petites grêles d'hiver, peu intenses et touchant majoritairement la côte Atlantique, des **grêles d'été** survenant dans les départements de plaines et de montagnes au cœur des terres.

La grêle en Assurance :

La grêle est un phénomène climatique qui peut toucher tous types de biens, il est pris en compte dans **la garantie TGN** (Tempête, Grêle, Neige) des contrats d'Assurances. C'est un risque atypique pouvant causer ponctuellement et occasionnellement des dégâts considérables sur tous types de biens.

Les chutes de grêle présentent la particularité d'être très localisées, ce qui complique les recensements. Il devient donc particulièrement difficile pour les assureurs d'identifier le risque et son évolution sur le territoire français. Les évènements majeurs de **Juin 2013** et de la **Pentecôte de 2014** ont représentés des charges considérables pour les compagnies d'Assurances, témoignant d'une augmentation de l'intensité du phénomène en raison du changement climatique. Cette évolution interroge les Assureurs sur la possibilité de mettre en place de nouvelles méthodes qui tiennent compte des évolutions climatiques afin de modéliser ce risque.

La réforme de **Solvabilité 2** impose notamment le calcul du SCR (Solvency Capital Requirement) qui est le capital nécessaire pour que la probabilité de défaut à horizon 1 an de l'assureur soit inférieure à 0,5%. Contrairement à la formule standard, le calcul du SCR par le biais d'un modèle interne offre la possibilité de refléter le profil de risque réel de l'assureur et de s'adapter à des changements de structure de portefeuille. Une modélisation précise et rigoureuse du péril grêle pour une compagnie d'assurance est donc un enjeu clé.

Un modèle CAT permet de modéliser l'impact d'un risque naturel sur le portefeuille d'une compagnie d'assurance. Il se compose traditionnellement en 3 modules : le module **Aléa** qui permet de modéliser les caractéristiques physiques inhérentes au risque, le module **Vulnérabilité** qui permet de lier les simulations du phénomène à l'exposition du portefeuille assuré et le module **Financier** qui va y associer les pertes de l'assureur. Nous conservons donc cette approche dans l'optique de modéliser la grêle.

Les données internes et climatiques :

D'une part, chez **Pacifica**, les évènements de grêle sont définis sur des plages de 1 jour. Sur une journée, un évènement de grêle peut toucher des zones très vastes comme très restreintes que l'on appelle des « empreintes ». A partir des bases de sinistre de l'ensemble des produits automobiles et habitations des évènements de grêle majeurs de Pacifica sur la période 1999 – 2018, nous réalisons une cartographie des évènements majeurs de grêle afin que chaque sinistre soit considéré à partir de son **code postal**, de

sa date et de sa charge.

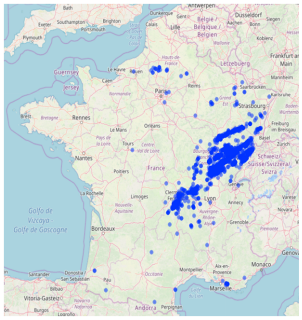
D'autre part, les données climatiques sont issues de 2 sources différentes :
Premièrement, l'**ECAD** (European Climate Assessment Dataset) a extrapolé des relevés issus de 106 stations Météo-France afin d'obtenir quotidiennement ces informations à une maille plus fine correspondant à des points espacés chacun de 0,5 degrés sur l'ensemble du territoire métropolitain. Il en résulte 252 stations extrapolées sur un historique d'une cinquantaine d'années et contenant quotidiennement les températures maximales quotidiennes (en Degrés), les températures minimales quotidiennes (en Degrés), les températures moyennes quotidiennes (en Degrés), les pressions moyennes quotidiennes (en Bars) et la somme des précipitations quotidiennes (en Millimètres).

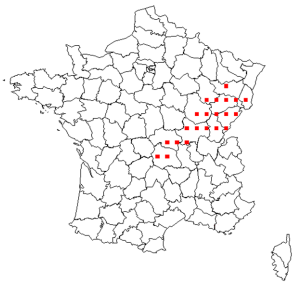
Deuxièmement, les données **SYNOP de Météo-France** sont constituées de relevés effectués **toutes les 3 heures** sur 52 stations en France Métropolitaine, ce qui est un point positif majeur car elles permettent de challenger les observations de l'Anelfa qui indiquent que les chutes de grêle ont tendances à avoir lieu dans l'après-midi. Les données SYNOP présentent 4 variables fiables : la pression, la pluviométrie, les températures, la direction du vent et la vitesse du vent.

La grêle étant un phénomène très localisé, nous décidons donc de l'étudier à la maille la plus fine disponible, c'est à dire à partir des stations fictives extrapolées de l'ECAD. Nous attribuons alors à chaque station extrapolée de l'ECAD les relevés issus de la station SYNOP la plus proche.

Croisement des données

L'objectif est d'associer l'ensemble des variables climatiques explicatives X_i à une variable cible Y_i booléenne (0 ou 1) : **la Survenance**.
Pour ce faire, à partir de leurs géolocalisations, nous associons chaque code postal de France métropolitaine à la station ECAD la plus proche. Ainsi, si à partir des cartographies issues des données internes de Pacifica nous observons des sinistres dans divers codes postaux dépendant d'une station ECAD, cette dernière se voit attribuer une survenance pour le jour considéré. Nous relions ainsi la sinistralité à notre base de données climatiques.





De plus, nous ne considérons que les grêle d'été et décidons donc de ne conserver que les données des mois de Mai, Juin, Juillet et Août. Finalement, nous obtenons les relevés des variables climatiques sur 20 années de 123 jours d'été.

Prédiction de la grêle :

Le premier objectif de ce mémoire est de construire un modèle prédictif de la survenance de grêle, à partir de l'ensemble des variables explicatives constituées précédemment.
Après plusieurs comparaisons, nous portons notre choix sur un **Random Forest**, un algorithme de clas-

sification (ou de regression) s'appuyant sur 2 méthodes :

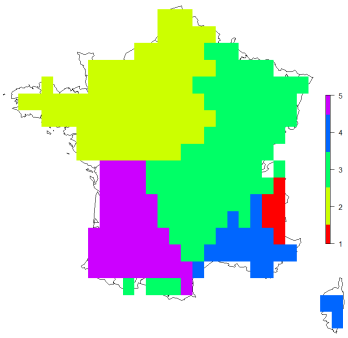
— La création d'arbres de décisions basés sur l'algorithme de **CART** : A chaque itération, l'algorithme sépare chaque noeud de l'arbre afin de maximiser le gain d'information en s'appuyant sur une mesure d'impureté, **l'indice de Gini**.
— **Le bagging** : un procédé d'agrégation de modèles sur échantillons bootstrappés permettant ainsi de réduire la variance.

Nous sélectionnons ensuite les 6 variables jugées les plus significatives par notre modèle, c'est-à-dire celles qui maximisent le gain d'information en séparant au mieux les classes selon l'indice de Gini : la température maximale, la variation entre la température maximale du jour et de la veille, la variation de température entre 15h et 18h, la pression et les précipitations.

En sortie, le Random Forest prédit correctement 68,2% des survenances historiques tout en ne générant quasiment pas de fausses survenances (0,58% des prédictions), ce qui sous-estime le risque mais reste tout de même une bonne qualité de prédiction comparativement aux autres méthodes étudiées.

Module Aléa :

Pour modéliser nos variables explicatives, nous devons réduire la complexité de nos données tout en tenant compte de l'hétérogénéité spatiale. Pour cela, nous regroupons les stations à partir de méthodes d'analyse factorielle et de classification que sont l'**Analyse en Composantes Principales (ACP)** et la **Classification Ascendante Hiérarchique (CAH)** dans le but d'obtenir des zones de risque géographiques homogènes.



Pour chacune de ces zones globales, nous modélisons nos variables explicatives en nous basant sur les distributions empiriques et en utilisant un outil statistique qui permet d'introduire des structures de dépendance complexes : **les copules**. Nous générons ensuite 10 000 scénarios de 123 jours d'été à horizon 1 an pour nos 5 zones globales.

La prédiction de la Survenance étant caractérisée dans notre modèle au niveau de granularité des stations extrapolées ECAD, nous descendons à ce niveau d'échelle en utilisant des régressions linéaires entre les séries moyennes de chaque zone globale et les série des stations extrapolées de l'ECAD appartenant à la zone globale considérée pour chacune des variables du modèle.

Les résultats obtenus suite à l'application de notre Random Forest sur les simulations sont contrastés. Des évènements de grêle sont simulés tous les ans mais leurs dimensions sont très largement inférieures aux observations historiques. Ces différences peuvent s'expliquer par la perte de qualité engendrée par

un Random Forest qui sous-estime les survenances (30% de moins) mais également par certaines hypothèses prises en compte comme l'indépendance de la variable « précipitations » ou par le regroupement en zones globales basé sur des moyennes pouvant potentiellement induire des circonstances extrêmes plus rares et donc des empreintes peu étendues.

Modules Vulnérabilité et Financier :

Le module Vulnérabilité a pour objectif d'associer une perte aux différents biens sinistrés (automobiles et habitations) lors des survenances simulées dans le module Aléa. Pour cela, il doit associer l'**intensité** des survenances au nombre de biens touchés et aux taux de destruction pour un évènement de grêle donné.

Nous définissons la **fréquence de sinistralité** comme le nombre de sinistres rapporté au nombre de polices d'assurance sur une même zone : Fréquence de sinistralité = $\frac{\text{Nombre de sinistres sur la zone}}{\text{Nombre de polices sur la zone}}$

Lors d'un orage, la grêle ne frappe pas uniformément sur toutes les zones de survenance. En effet, il existe une variabilité importante de la fréquence de sinistralité en fonction des zones de survenance durant un même orage grêligène, traduisant une différence d'intensité des chutes.

Ainsi, il ne faut pas seulement associer les variables climatiques à la survenance de grêle mais également à une classe d'intensité. Pour cela, nous utilisons les fréquences de sinistralité historiques en assurance habitation en les discrétisant afin de classer les survenances de grêle en 4 classes d'intensité homogènes (1 : Intensité modérée, 2 : Intensité moyenne, 3 : Intensité forte, 4 : Intensité très importante).

De la même manière que pour la survenance, nous utilisons un Random Forest afin de prédire la classe d'intensité de chaque zone de survenance.
Sur notre base de test, le Random Forest prédit la bonne classe d'intensité dans 87,5% des cas mais a une légère tendance à sous-évaluer les classes d'intensité.

Nous modélisons les Fréquences de sinistralité automobiles et habitations en calibrant diverses lois continues (Log-normale, Pareto, Gamma, Beta) **conditionnellement aux labels d'intensité**. Au regard du critère du test de Kolmogorov Smirnov, pour l'ensemble des classes d'intensité, la loi Log Normale est celle qui modélise le mieux les Fréquences de sinistralité automobiles et habitation.

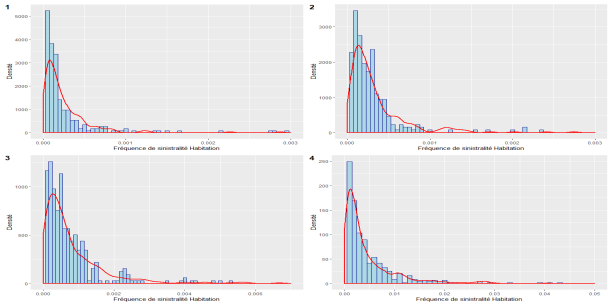


FIGURE 1 – Fréquence de sinistralité habitation - Distribution empirique (histogramme) et Simulations (rouge)

Pour chaque survenance simulée précédemment, nous prédisons un label d'intensité puis nous générons des fréquences de sinistralité automobiles et habitations en caractérisant leur dépendance par une copule de Gumbel. Cela indique que la survenance d'évènements majeurs de grêle entraîne générale-

ment un nombre important d'habitations et de voitures sinistrées.

Après l'obtention des fréquences de sinistralité pour chaque survenance, les nombres de sinistres NS affectant les automobiles ou les maisons s'obtiennent pour chacun en multipliant la fréquence de sinistralité FR par le nombre biens assurées N : $NS = FR \times N$.

Le **taux de destruction** d'un sinistre est égale à sa charge divisée par la somme assurée du bien sinistré. Le comportement des taux de destruction diffère grandement en fonction du type de bien. En effet, la charge associée à un sinistre grêle touchant une maison sera généralement plus importante que celle d'un appartement du fait de l'exposition directe de sa toiture. De plus, la **somme assurée** représente la valeur théorique du bien. En habitation, cette valeur est également très influencée par les caractéristiques des biens assurés. Par exemple, un assuré locataire n'a à sa charge que le contenu de son habitation tandis qu'un propriétaire doit prendre en charge le contenu et le bâti, lui conférant ainsi une somme assurée plus importante.

Nous séparons donc les habitations en 4 classes distinctes : les propriétaires de maisons, les propriétaires d'appartements, les locataires de maison et les locataires d'appartements. Pour chacune de ces classes, les sommes assurées sont considérées indépendamment de manière à modéliser les taux de destruction de la manière la plus fine possible. Les taux de destruction automobiles sont pour leur part considérés globalement.

De la même manière que pour les fréquences de sinistralité, nous modélisons les taux de destruction pour chaque classe en calibrant divers lois continues (Log-normale, Pareto, Gamma, Beta). Nous générons ainsi un taux de destruction pour chaque sinistre simulé.

La **charge finale** associée à chaque survenance est donnée par la formule suivante : $C = NS_i \times TD_i \times SA_i$ où NS_i correspond au nombre de sinistres de la classe i, TD_i correspond au taux de destruction de la classe i et SA_i correspond à la somme assurée moyenne de la classe i sur la zone.

A partir de ces informations, nous estimons les charges par évènement en agrégeant l'ensemble des survenances et nous en déduisons la charge annuelle. Nous pouvons ensuite tracer les courbes **OEP et AEP** qui associent une période de retour respectivement à chaque perte et à chaque année.

Conclusion :

En comparant les basses périodes de retour de notre modèle et de l'historique, nous remarquons que notre modèle sous-performe. Pour une période de retour à 20 ans, correspondant à la profondeur d'historique disponible, le modèle est inférieur de 27% à l'historique.
L'objectif était de mettre en place un modèle permettant à un assureur de calculer les fonds propres nécessaires pour faire face à des évènements de grande ampleur comme la grêle de la Pentecôte de 2014. Notre approche ne retourne pas des résultats assez sévères pour répondre à la problématique initiale. En revanche, certains axes d'amélioration, notamment lors de la modélisation des variables explicatives, pourraient rendre cette étude pertinente.

Remerciements

Je tiens à remercier Valery CAUSSARIEU, Manager de l'équipe Modélisation actuarielle au sein de la Direction de la Solvabilité et de l'Actuariat pour l'encadrement de mon mémoire.

J'aimerais également remercier Pierre BERTHOU et Anaïs BELABED pour leurs conseils aussi bien sur la construction que sur la réalisation technique nécessaires à l'aboutissement de ce mémoire.

Enfin, je remercie Monsieur Olivier Lopez, tuteur école et directeur de l'ISUP, pour son apport sur la rigueur technique et théorique de mon mémoire.

Table des matières

Première partie

Introduction

1 Présentation du phénomène de grêle

1.1 Présentation générale

Les épisodes de grêle se produisent en général lors d'orages violents, au sein de **cumulonimbus**. Moins de 10 % des cumulonimbus donnent de la grêle atteignant le sol.

Les grêlons se forment au sein des courants ascendants et descendants entre la base, chaude et humide, et le sommet très froid du nuage. La grêle se forme dans cette "colonne d'ascendance" par dépôts successifs de glace autour de petites particules solides appelées **noyaux glaçogènes** avant de tomber au sol sous forme d'averses de grêle.
Les noyaux glaçogènes proviennent d'éléments soit naturels (poussière, suie volcanique...), soit artificiels (rejets des réacteurs d'avion...). Il arrive souvent que les grêlons fusionnent entre eux pour donner des particules encore plus grosses : on parle **d'accrétion**.
Si les ascendances sont faibles, la particule qui grossit va très rapidement descendre et être éjectée du nuage pour atteindre le sol sous forme de pluie.
Si les ascendances sont fortes, la particule reste longtemps en suspension dans le nuage en collectant l'eau surfondue (eau restée en phase liquide alors que les températures sont négatives en raison de l'absence d'impuretés dans l'air, jusqu'à –40 degrés Celsius). Ce sont donc les plus gros cumulonimbus (les orages les plus violents) qui génèrent les plus gros grêlons.

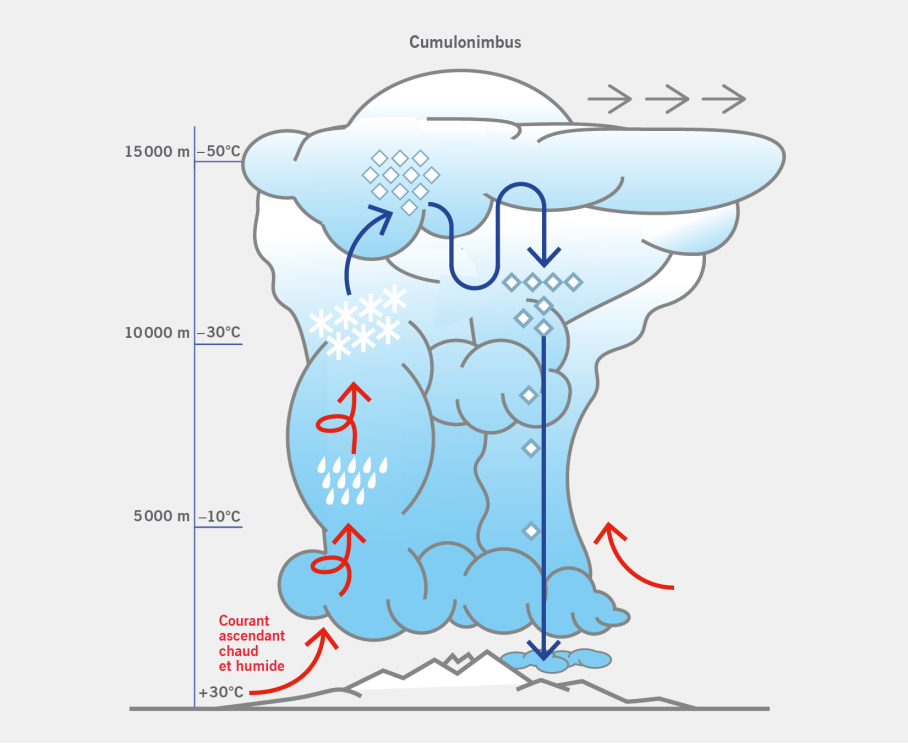


FIGURE 2 – Nuage de grêle

1.2 Le grêlon

Les noyaux glaçogènes présents dans les nuages permettent la formation de la glace par 2 mécanismes :

— **Condensation solide** : Le noyau glaçogène, souvent très haut dans le nuage (donc à température très négative) va grossir progressivement par dépôt de la vapeur d'eau. On parle alors de croissance sèche (formation d'une couche de givre blanc opaque).

— **Congélation de l'eau surfondue** présente dans le nuage : en grossissant, la particule ainsi formée va descendre un peu plus vite dans le nuage que les gouttelettes d'eau surfondue. Elle va les collecter et les faire geler instantanément. Elle devient alors opaque, de forme circulaire avec un diamètre de quelques millimètres en grossissant encore, la particule descend de plus en plus vite au regard des gouttelettes d'eau surfondue qu'elle collecte encore plus rapidement. Elle se réchauffe progressivement et l'eau gèle plus lentement à sa surface. La particule prend alors une apparence transparente et sa densité augmente. On parle de croissance humide (formation d'une couche translucide par capture des gouttelettes).

FIGURE 3 – Grêlon

Au cours de son évolution, un grêlon va alterner des phases de croissance sèche et de croissance humide en fonction de sa vitesse de chute et des caractéristiques de son environnement (température et concentration en eau surfondue). Enfin, lorsque la particule est suffisamment grosse au regard des courants ascendants du nuage, elle finit par tomber au sol : c'est l'averse de grêle.

1.3 Formation d'un orage

Les orages sont formés par des blocs de constructions appelés **cellules** induits par des phénomènes thermodynamiques et de cinématique fondamentale.
En effet, une cellule est une unité dynamique caractérisée par des régions compactes de courants ascendants (identifiables par écho-radar) provoqués par la **convection atmosphérique** (mouvements internes de l'atmosphère terrestre résultant d'une instabilité de l'air due à une différence de température verticale ou horizontale). Cela se produit par le fait de **l'énergie potentielle de convection** disponible

(**EPCD ou CAPE**) qui est l'énergie potentielle par unité de masse qu'a une parcelle d'air plus chaude que son environnement se traduisant par une poussée d'Archimède ascensionnelle [1].

La vie d'une cellule dure approximativement 30 minutes et peut être synthétisée par 3 étapes :

— Un stade de **cumulus** caractérisé uniquement par des courants ascendants (echo radar de 3 à 6 kms de hauteur).

— Un stade de **maturité** avec courants ascendants et descendants combinés. Les nuages sont plus hauts et les radars identifient une réflectivité plus forte.

— Un stade de **dissipation** où les courants descendants prédominent et où le niveau d'intensité de l'orage diminue avec la chute de précipitations.

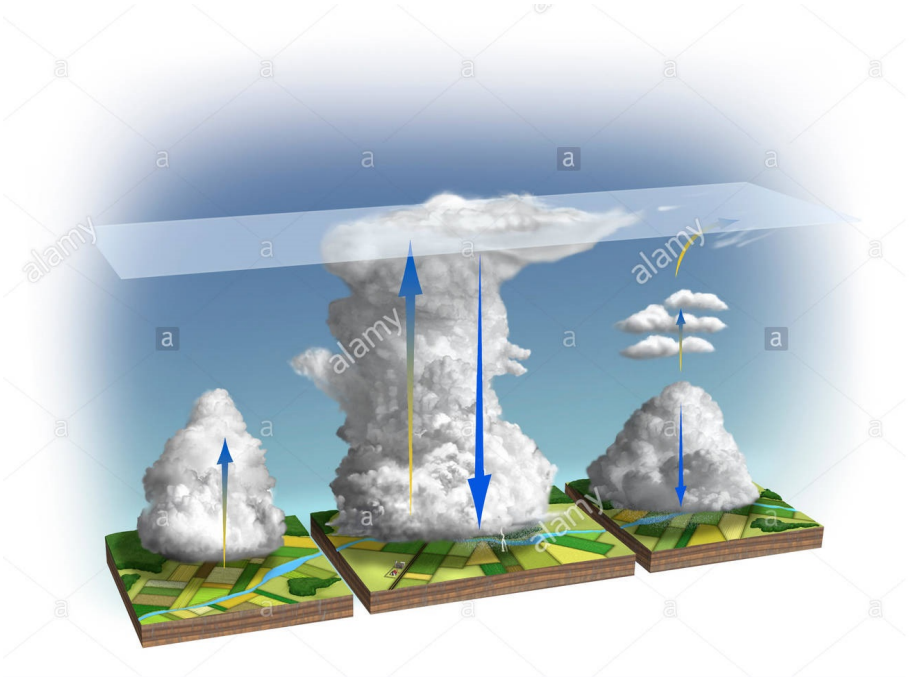


FIGURE 4 – Cycle d'une cellule orageuse

De nouvelles cellules orageuses se forment généralement sur le flanc des précédentes. Les orages ordinaires sont largement majoritaires et sont la base sur laquelle se forment les orages violents.

1.4 Orages violents

Les orages violents nécessitent 3 ingrédients :

— **Une couche humide** de profondeur suffisante dans la moyenne ou basse troposphère.

— **Un taux de changement** de température assez élevé pour permettre une flottabilité positive substantielle ou une énergie potentielle disponible par convection.

— Le **soulèvement suffisant d'une parcelle de la couche humide** pour permettre au nuage d'atteindre son niveau de convection libre, où il peut remonter uniquement avec sa flottabilité.

Les fronts froids aident à déstabiliser les structures thermodynamiques.

La structure dynamique et la sévérité des tempêtes sont fortement influencées par l'environnement, notamment les vents relatifs ou cisaillants et leur nature (les changements de directions et de vitesse du vent avec l'altitude).

Les radars identifient les orages violents en 3 catégories :

— **Les orages à cellule unique** : Ils diffèrent des orages simples car les courants ascendants ont une durée de vie courte et le premier écho-radar se positionne à plus haute altitude (6-9 kms). Au stade mature, les zones de réflexion fortes (>50dBZ) restent continues, des rafales de vents ainsi que de brèves apparitions de grêle peuvent apparaître ;

— **Les orages multicellulaires** : C'est une séquence organisée de plusieurs cellules à des stades de développement différents avec création périodique de nouvelles cellules. Avec le développement de nouvelles cellules, une région d'écho se crée à mi-niveau dans l'orage en dessous de laquelle une zone « d'écho faible » (WER) est visible. Ce type d'orage peut durer plusieurs heures et être accompagné de rafales et de grêles importantes ;

— **Les orages supercellulaires** : Ce sont des évènements rares qui se développent à partir des orages multicellulaires caractérisés par des courants ascendants et descendants très importants pouvant coexister durant plusieurs heures. La zone d'écho faible est persistante et devient bornée avec les courants ascendants. Lorsque celle-ci commence à descendre, l'apparition de fortes grêles, de tornades ainsi que de rafales de vents est plus que probable.

1.5 Formation de la grêle

La première condition à la formation de la grêle est un courant ascendant assez fort capable de supporter la formation de grêlons.
La puissance des courants ascendants est principalement influencée par **l'énergie convective potentielle (CAPE)**. Plus la CAPE est grande, plus la probabilité de développement de grêle importante est forte au cœur de l'orage. La taille des grêlons peut donc être estimée grâce à la puissance des courants. Cependant, des courants trop forts (supérieurs à 40 mètres par seconde) ne permettent pas la formation de grêlon important car les particules ne restent pas suffisamment longtemps dans la zone de croissance optimale. La vitesse optimale des courants doit être comprise entre 20 et 40 mètres par seconde de manière à avoir des chutes de grêle significatives. La croissance de la grêle dans les nuages est décrite

en 3 étapes :

— **De petites particules grandissent dans des courants relativement faibles** sur le flanc du courant principal par accrétion de gouttelettes d'eau surfondue ;

— Certaines de ces particules voyagent dans les courants faibles autour du bord avant du courant principal avant **d'entrer à l'intérieur du courant principal** en tant qu'embryon avec un diamètre de quelques millimètres ;

— Même si il y a présence d'oscillations négligeables amenant à de petites fluctuations dans l'intensité des courants, ces embryons vont graduellement se transformer en grêlon (diamètre supérieur à 5 mm). Il en résulte des grêlons à structure d'oignon avec une alternance de couches sèches et humides en fonction du fait que les gouttelettes d'eau surfondue soient ou ne soient pas congelées par le processus de conduction et d'évaporation à la surface des grêlons.

A partir du moment où la grêle sort du courant et commence à tomber vers le sol, la fonte commence en dessous de 0 degrés Celsius sous l'influence de la distance entre le niveau de congélation et le sol,

de la température moyenne des courants descendants et de la vitesse de chute (dépendant de la taille et de la densité des grêlons). La quantité de grêlons peut avoir une forte variabilité en fonction du type d'orage, de l'environnement et de la saison.

Généralement lors d'un orage, les grêlons les plus importants tombent en premier car ils sont les premiers à sortir du courant ascendant et tombent plus vite du fait de leur masse plus importante. Au sein d'une même chute de grêle il peut donc il y avoir une **variabilité de sinistralité entre des zones touchées relativement proches**. Certaines prendront l'ensemble des gros grêlons tandis que les autres prendront ceux de tailles petites et intermédiaires.

2 La grêle en France

La grêle est un phénomène climatique qui peut toucher tout type de biens. Cependant, en France, il n'est pas considéré comme une "catastrophe naturelle" malgré les débats existant sur ce sujet et est donc classé dans les risques assurables. Ces évènements sont pris en compte dans la **garantie TGN** (Tempête, Grêle, Neige) des contrats d'assurances.
Pour un assureur, la compréhension du phénomène de grêle doit prendre en considération plusieurs facteurs :

— La répartition géographique ;

— Les différences propres au phénomène en fonction des saisons ;

— La fréquence des évènements ;

— L'intensité des évènements. Cette intensité varie en fonction du nombre de grêlons par unité de temps, la taille des grêlons ainsi que la vitesse de chute.

Les chutes de grêle présentent la particularité d'être très localisées, ce qui complique les recensements. Cela en fait un phénomène particulièrement difficile à étudier pour les assureurs dans l'optique d'identifier le risque et son évolution sur le territoire français.
Aujourd'hui en France, **l'Anelfa (Association Nationale d'Etude et de Lutte contre les Fléaux Atmosphériques)** est la seule association à avoir effectué des relevés au sol des chutes historiques de grêle afin de les étudier. Ces études ont d'ailleurs été reprises par **Freddy Vinet** il y a une vingtaine d'années pour réaliser ce qui est encore à ce jour la seule étude complète du phénomène de grêle sur le territoire français [2].

2.1 La garantie TGN

Dans la législation française, les évènements issus des risques Tempête, Grêle et Neige sont couverts par la **garantie TGN**. Cela veut dire qu'ils ne font pas partie des Catastrophes Naturelles et sont donc considérés comme assurables.
En 1990, l'article L. 122-7 a été inséré au livre Ier du code des assurances ainsi rédigé : « Les contrats d'assurance garantissant les dommages d'incendie à des biens situés en France ainsi qu'aux corps de véhicules terrestres à moteur ouvrent droit à la garantie de l'assuré contre les effets du vent dus aux tempêtes, ouragans ou cyclones, sur les biens faisant l'objet de tels contrats.
En outre, si l'assuré est couvert contre les pertes d'exploitation après incendie, cette garantie est étendue aux effets du vent dus aux tempêtes, ouragans ou cyclones.»

A titre d'exemple, en 2016, La FFA (Fédération Française de l'Assurance) [4] indiquait que les cotisations afférentes à la garantie tempête grêle neige (TGN) sur bâtiment sont estimées à 1,54 Milliards d'euros pour l'ensemble du marché national (hors assurances automobile et récoltes) ; ce qui représente 8,8 % de l'ensemble des cotisations en assurances de dommages aux biens. La répartition des cotisations est de 70 % pour les risques des particuliers, 18 % pour ceux des entreprises et 12 % pour les risques de dommages aux biens agricoles. La sinistralité est en hausse de 31 % en 2016 par rapport à 2015, avec 635 Millions d'euros d'indemnités versées par les assureurs. Cependant, l'exercice 2016 reste relativement clément avec un ratio S/P de 41%, en progression de 9 points de pourcentage par rapport à 2015.

2.2 Etude des chutes de grêle sur le territoire français

2.2.1 L'Anelfa

L'Anelfa est une association loi 1901 fondée en 1951 dont les deux objectifs majeurs sont :

— Développer les recherches scientifiques dans le domaine de la physique des nuages et de la modification du temps ;

— Perfectionner une méthode de traitement des orages afin de réduire les dégâts causés par la grêle.

L'Anelfa dispose de grêlimètres (un appareil mis au point par des chercheurs canadiens permettant d'enregistrer la trace d'impacts de grêlons au sol) répartis à une maille très fine sur certains départements partenaires français afin d'étudier les chutes de grêles.[5]

FIGURE 5 – Grêlimètres - Source Anelfa

Après la chute de grêle, les impacts de grêlons sont rendus visibles par un encrage en noir de la plaque au rouleau d'imprimerie.

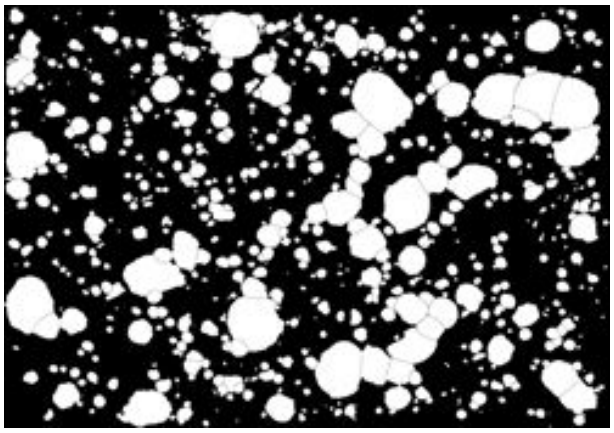


FIGURE 6 – Grêlimètres - Source Anelfa

Le diamètre de chaque grêlon à l'origine d'un impact et le nombre de grêlons tombés sont déterminés dans différentes gammes de dimension supérieurs à 5 mm de diamètre.

Selon l'Anelfa, les paramètres les plus significatifs d'une chute de grêle sont le **diamètre** des plus gros grêlons, le **nombre total de grêlons** et l'**énergie cinétique globale** de la chute de grêle.

Les mesures ne sont effectuées que dans les départements partenaires, globalement dans le Sud-Ouest, le Sud et le Centre de la France.

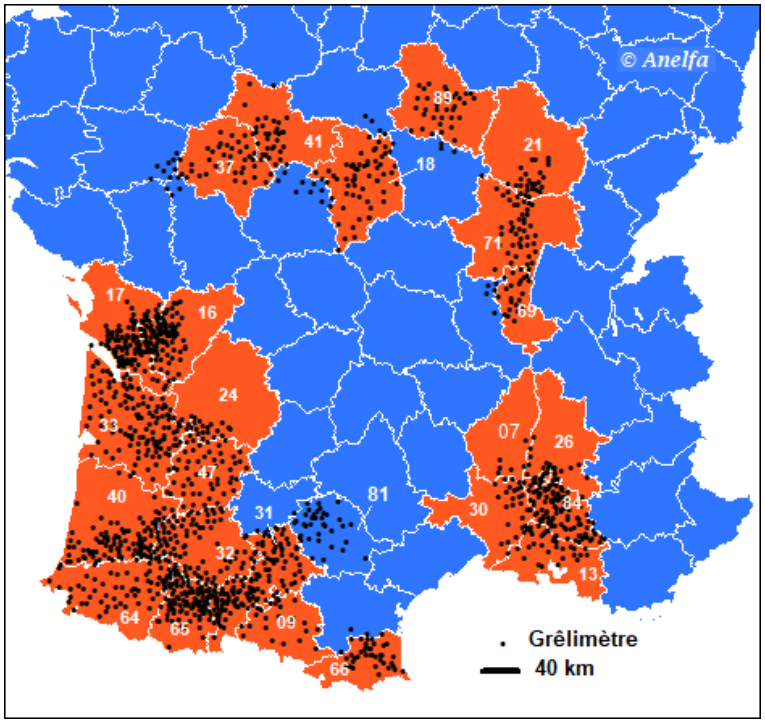


FIGURE 7 – Position des grêlimètres - Source Anelfa

L'échelle étant relative à une chute de grêle en un point, un même orage peut produire des chutes de différentes classes en différents endroits, de même qu'un cyclone ou une tornade peut changer d'intensité en cours de vie.

CLASSE ANELFA ▸	A0	A1	A2	A3	A4	A5
Diamètre maximal des grêlons	<1	1-1.9	2-2.9	3-3.9	4-4.9	>=5
Terme usuel	grésil, petit pois	bille, grain de raisin, cerise	oeufs de pigeon, pièce de 2 €	noix, balle de ping-pong	oeuf de poule, balle de Golf	pêche, pomme orange, balle de tennis
Energie cinétique moyenne	10 $J.m^{-2}$	50 $J.m^{-2}$	200 $J.m^{-2}$	500 $J.m^{-2}$	800 $J.m^{-2}$	
dommages types	accidents de la route, fleurs coupées	dommages aux vignes vergers, tabac	dommages importants aux céréales légumes, arbres	dommages à 100% sur toute culture, vitres cassées, voitures endommagées	paysage d'hiver, animaux tués personnes blessées, avions au sol endommagés	évènement extrêmement dangereux, risque mortel
couverture moyenne du sol	0.1	0.15		0.35	0.35	

FIGURE 8 – Échelle d'intensité - Source Anelfa

Répartition mensuelle historique :

Le graphique ci-dessous représente le nombre de chutes enregistrées par mois, celles-ci étant classées selon l'échelle relative au diamètre des grêlons.

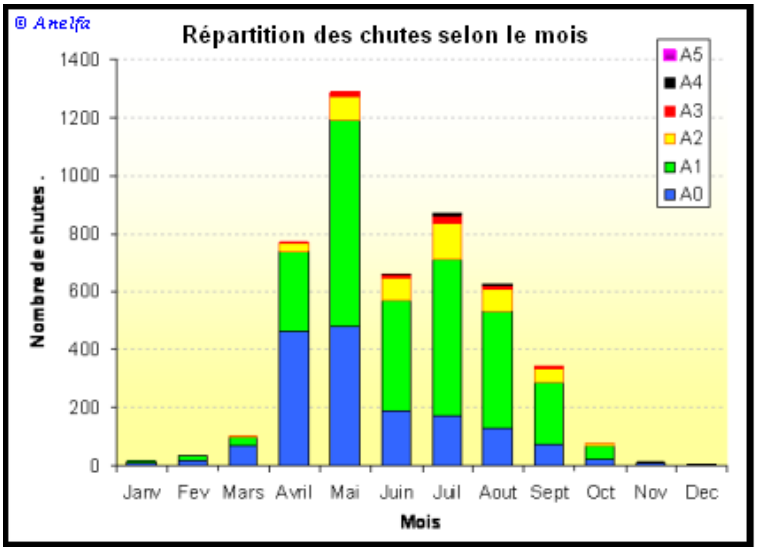


FIGURE 9 – Répartition mensuelle des chutes de grêle - Source Anelfa

Répartition horaire historique :

L'énergie cinétique mesurée à l'aide des grêlimètres est un paramètre permettant d'évaluer l'intensité des chutes de grêle. En prenant en compte le nombre de grêlons au mètre carré, leur diamètre et la

vitesse de chutes, c'est un indicateur complet car fortement corrélé aux dégâts sur végétaux.
Le graphique ci-dessous représente la répartition de l'intensité des chutes dans la journée.

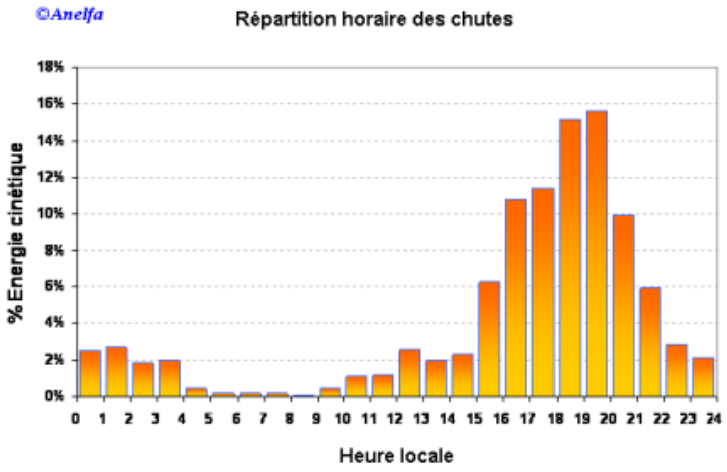


FIGURE 10 – Répartition horaire des chutes de grêle - Source Anelfa

Les études réalisées par l'Anelfa sur l'ensemble des départements couverts par l'association entre 1989 et 2019 ont menées à la conclusion que l'intensité des chutes de grêle a sensiblement augmenté sur les vingt années considérées mais que la fréquence d'occurrence de ces chutes restait stable au cours du temps. La charge des assureurs associée aux évènements de grêle devrait ainsi être amplifiée dans les années à venir.
Le travail réalisé par l'Anelfa est cependant effectué sur un nombre restreint de départements et nous n'obtenons donc pas une cartographie fiable du risque sur le territoire français.

2.3 Thèse de Freddy Vinet

Au-delà des relevés effectués par l'Anelfa, la seule étude complète réalisée sur le sujet est une thèse intitulée **"Le risque-grêle en France étude géographique"** (Vinet, 1999) [2] effectuée par Freddy Vinet en 1999. Pour mener à bien son étude, il a eu accès aux relevés des grêlimètres de l'Anelfa, aux données recueillies sur 104 stations Météo France, aux données de sinistralité issues du portefeuille de Groupama ainsi qu'à de multiples relevés de presse.
Les cartes suivantes indiquent respectivement les répartitions mensuelles et géographiques des chutes de grêle en France sur l'historique d'informations recensé par Freddy Vinet. [2]

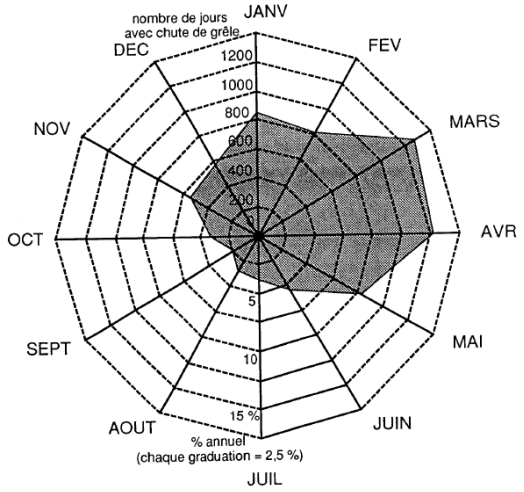


FIGURE 11 – Répartition mensuelle de l'ensemble des chutes recensées [2]

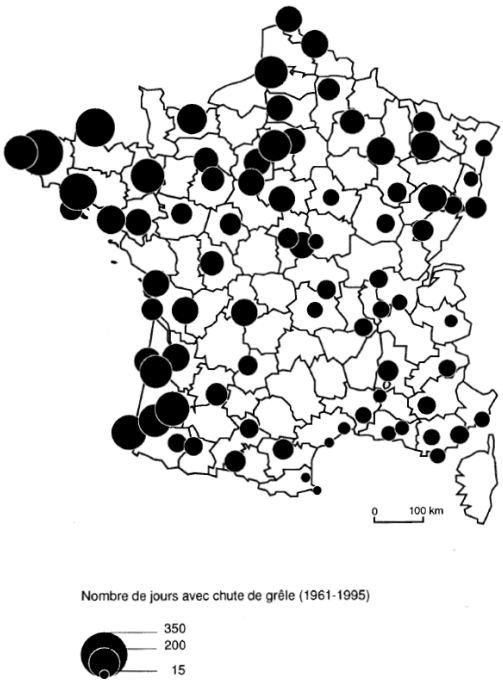


FIGURE 12 – Nombre de jours avec chute de grêle entre 1961 et 1995 [2]

Les observations effectuées par Vinet indiquent que la majorité des épisodes grêles ont lieu entre les mois de Mars et Avril.
De plus, la majorité de ces chutes de grêle touchent la côte Atlantique du pays. Cependant, Freddy Vinet indique dans sa thèse que « le véritable risque réside dans la grêle d'été » par opposition aux grêles de printemps et d'Hiver qui "ne participent pas à la définition du risque-grêle" car ces dernières sont négligeables sur la masse de sinistralité observée. Nous verrons ultérieurement que les dégâts ne sont engendrés quasiment que par les **grêles d'été**, qui ne représentent pourtant que 27% du total des grêles annuelles.

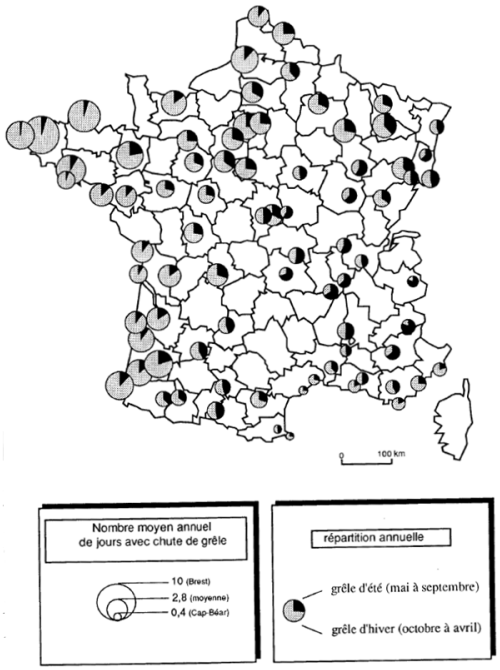


FIGURE 13 – Répartition entre grêle d'été et grêle d'hiver par station de recensement [2]

La proportion de grêles d'été diffère en fonction de la zone du territoire caractérisée. En effet, l'ouest du territoire est en réalité touché par beaucoup de petites grêles d'hiver et n'est donc que très peu exposé au risque.
Une carte ne prenant en compte que les grêles d'été peut donc être établie à partir de l'historique des relevés.

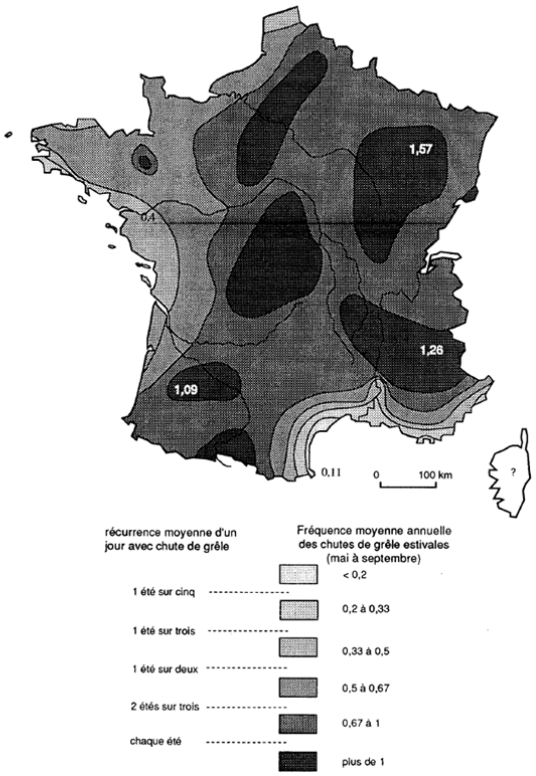


FIGURE 14 – Cartographie des zones à risque [2]

Les observations confirment que les **zones côtières** (zones méditerranéennes et atlantiques) sont globalement épargnées tandis que les zones de **plaines et montagneuses** au coeur des terres sont particulièrement exposées au risque de grêle.

2.4 Ce qui existe sur la grêle

2.4.1 Études

Les études portant sur la grêle existent, mais celles-ci sont rarement appropriées quand il s'agit d'étudier la modélisation du phénomène. En effet, on apprend dans l'article *"Munich Re - Severe thunderstorms in Europe"*[23] que le **réchauffement climatique** engendre sur les océans une évaporation plus importante de l'eau en surface. Lors de l'apparition de phénomènes de convection, cette eau devient potentiellement une source importante de grêle. Le réchauffement climatique ne devrait donc pas augmenter la fréquence d'occurrence mais devrait être un vecteur **d'aggravation de la sévérité des évènements futurs** du fait de l'accroissement de l'énergie potentielle de convection (**CAPE**) mais également par l'augmentation de la quantité d'eau à l'intérieur des orages grêligènes.

Par ailleurs, certaines études témoignent d'approches alternatives de modélisation de la grêle. En 2015, Susanna MOHR et Michael KUNZ dans ***"Hail potential in Europe based on a regional climate model hindcast"*** [3] ont ajusté les paramètres d'un modèle logistique afin d'estimer le nombre de jours pour lesquels une augmentation significative de la probabilité de grêle était présente. Cet indice est le ***PHI (Potential Hail Index)***, dont les paramètres d'entrée sont : l'**indice de relief**, la **température minimale** et la **température maximale**.
L'étude se restreint aux 3 mois d'été et se base sur des données climatiques de très bonne qualité et

disponibles à une maille très fine grâce au modèle numérique de prévisions météorologiques COSMO. In fine, les résultats obtenus n'étaient pas suffisants pour mettre en place un modèle prédictif efficace. Cependant, le PHI est un indicateur fiable des zones à risque.

De manière plus générale, la plupart des études réalisées sur le marché expliquent que les facteurs climatiques classiques (température, pression ...) sont corrélés à la présence de grêle mais que ces informations doivent être extrêmement fines [15] (données en altitude, au coeur des orages ...) afin de pouvoir réellement interpréter une causalité. Il est donc relativement fréquent que les conditions climatiques de base nécessaires à l'occurrence de grêle soient vérifiées mais qu'aucun évènement ne soit observé. Une vision plus fiable est donnée par les informations internes aux nuages (température, quantité d'eau, vitesse des courants ascendants ...).

2.4.2 Modèles

Parmi les divers modèles existants[8][13][17] sur le marché, les deux modèles les plus connus sont HailCalc de RMS et GCat de Guy Carpenter. Ce dernier utilise des procédés complexes basés sur des données issues **d'imagerie radars et satellites**. Le modèle se décompose en module Aléa, Vulnérabilité, Exposition et Financier.
Il caractérise tout d'abord l'aléa en partant des données satellites pour observer la durée, la dimension spatiale ainsi que les trajectoires des orages greligènes. Les images radars permettent ensuite d'identifier la **réflectivité lumineuse** au sein du nuage. L'intensité de ces signaux traduit ensuite la présence ou non de grêlons. En fonction de cette intensité, les systèmes récents sont capables d'identifier le moment, le lieu et l'intensité des chutes de grêle par unité de surface (dépendant de la résolution de l'imagerie disponible).

L'image suivante montre la réflectivité lumineuse au sein d'un orage grêligène à partir de l'imagerie radar. On remarque notamment en rouge l'évolution de la trajectoire des chutes et le degré d'intensité de celles-ci.

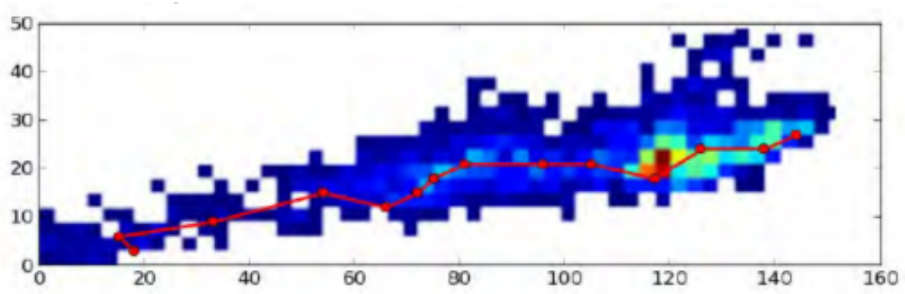


FIGURE 15 – Orage grêligène par imagerie radar

A partir de ces imageries, un catalogue complet d'historique d'orages (date, position, intensité) est constitué. Les trajectoires, les empreintes et l'intensité par unité de surface sont ainsi simulées pour un nombre important d'années à partir de cet historique.

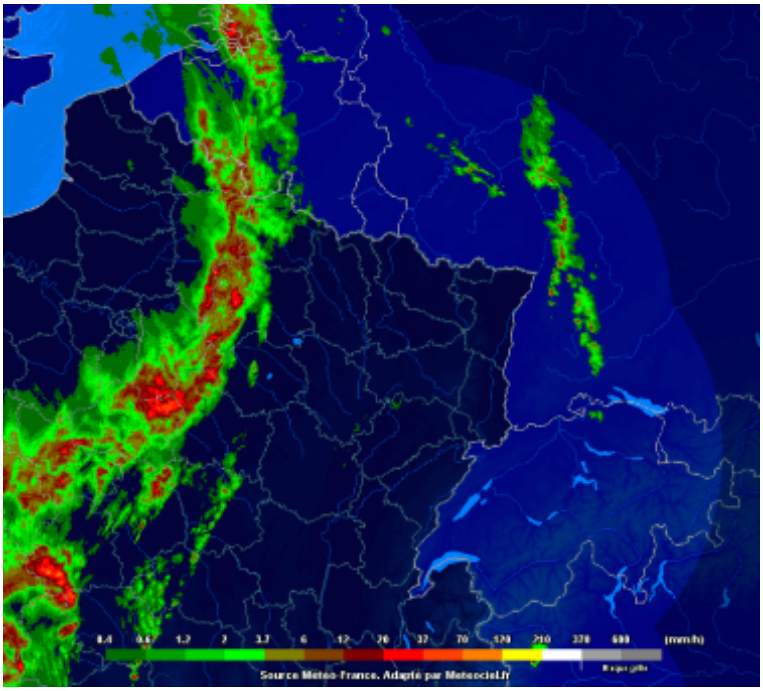


FIGURE 16 – Orage grêligène par imageries radar et satellite

Chaque zone est ensuite caractérisée par son exposition (informations sur le parc immobilier de la zone) afin d'obtenir une redistribution individuelle du risque. Le calcul des dommages s'effectue ensuite en se basant sur l'historique de sinistralité, les pertes financières sont ensuite déduites des dommages obtenus.

L'aléa du péril de grêle nécessite donc des données à une maille très fine afin d'être correctement étudié. La disponibilité et le coût de ces données ainsi que la complexité des modèles sous-jacents sont un frein rendant difficilement envisageables leurs utilisations dans le cadre de ce mémoire et nous contraint donc à étudier une approche plus simple.

2.4.3 Étude de la sinistralité grêle Pacifica

Vision annuelle historique :

Le nombre d'évènements majeurs (supérieurs à un certain montant de charge redressée) est en moyenne de 2 par an. On observe une pointe en 2013 à 5 évènements tandis que plusieurs années ne présentent aucun évènement de cette ampleur.

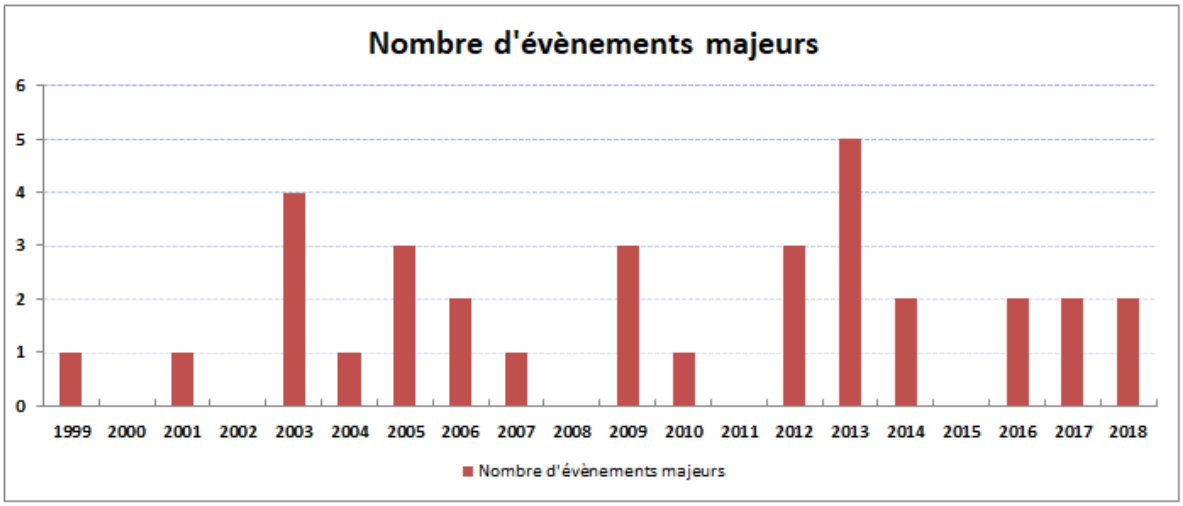


FIGURE 17 – Nombre d'évènements majeurs

De la même manière, la charge annuelle atteint des valeurs importantes en 2013 et 2014, tandis que les années 2000, 2002, 2008, 2011 et 2015 sont épargnées.

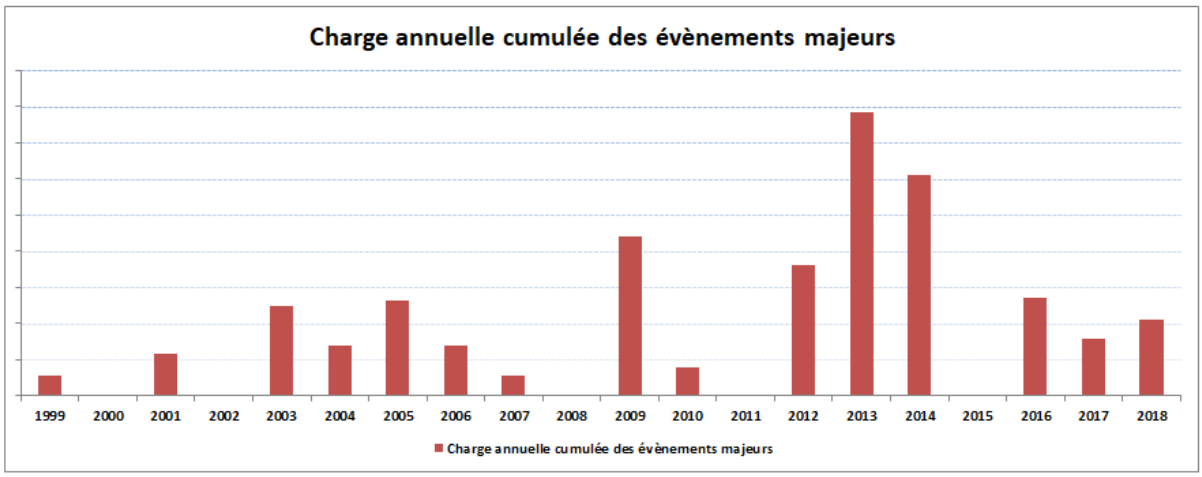


FIGURE 18 – Charge annuelle historique des évènements majeurs

Nous observons une importante volatilité de la sinistralité caractéristique des périls naturels. En effet, les dommages du quotidien comme les accidents de voitures ou les incendies sont courants, et présentent donc une certaine stabilité, tandis que les périls naturels sont généralement engendrés par des phénomènes météorologiques extrêmes dont la fréquence est généralement faible.
La charge engendrée par ces phénomènes peu fréquents est donc très variable et dépend de plusieurs paramètres. La zone touchée ainsi que sa dimension présentent une importance capitale car l'exposition du portefeuille de l'assureur (nombre de polices assurées, caractéristiques des biens) n'est pas uniforme sur l'ensemble du territoire français. Cependant, la grêle touche des zones préférentielles, généralement au cœur des terres et de nombreux départements ne sont ainsi que peu exposés au risque. De plus, l'intensité inhérente à un évènement de grêle, pouvant par exemple être caractérisée par une taille de grêlon, joue un rôle important par la quantité de biens impactés et donc sur la charge globale

de l'évènement.
La sinistralité inhérente à la grêle présente donc une hétérogénéité en vision annuelle.

Vision mensuelle historique :

Les évènements de grêle présentant une charge importante se concentrent sur 4 mois de l'année correspondant à la période estivale élargie : **Mai, Juin, Juillet et Août**. Cela vient corroborer les travaux de Freddy Vinet considérant l'automne et l'hiver comme des périodes de grêle "douce" car ne provoquant que peu de dommages. En effet, les épisodes de grêles importants ont besoin de phénomènes de convection intenses pour générer des flux orageux d'intensité suffisante. Cela n'est possible que lors de présence d'air chaud en basse altitude et donc durant la période estivale.

La figure suivante présente le nombre total d'évènements majeurs de grêle par mois entre 1999 et 2018.

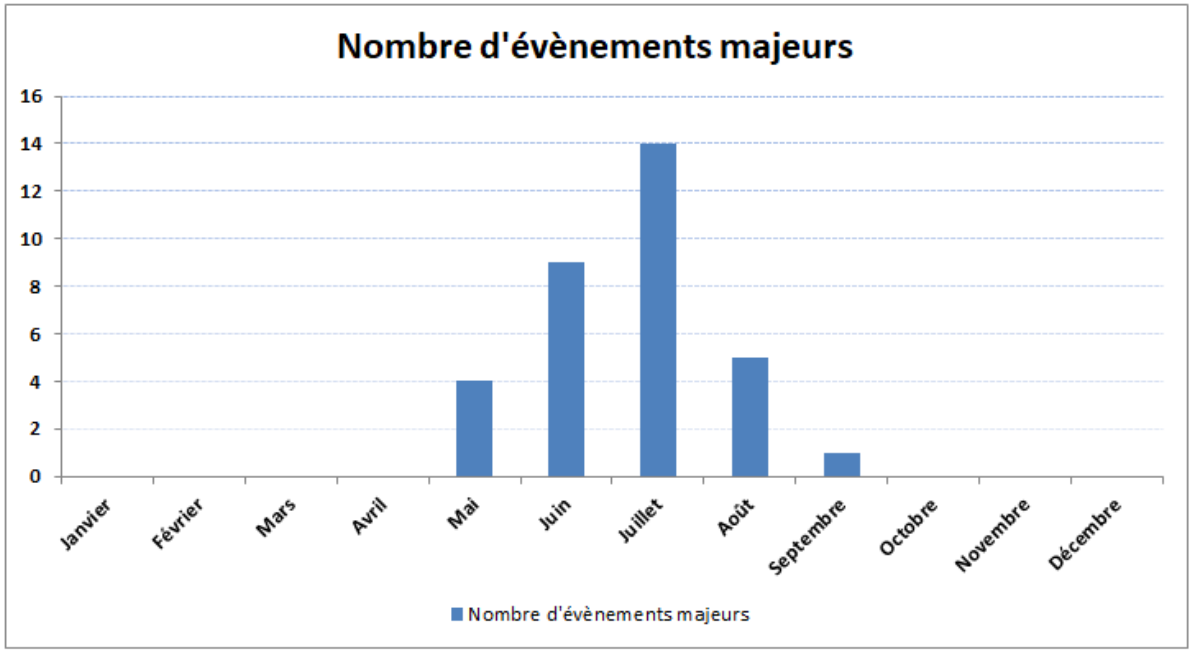


FIGURE 19 – Nombre d'évènements historiques supérieurs à 5 Millions d'euros mensuellement

Les mois de Juin et Juillet concentrent 77% de la charge historique des évènements majeurs.

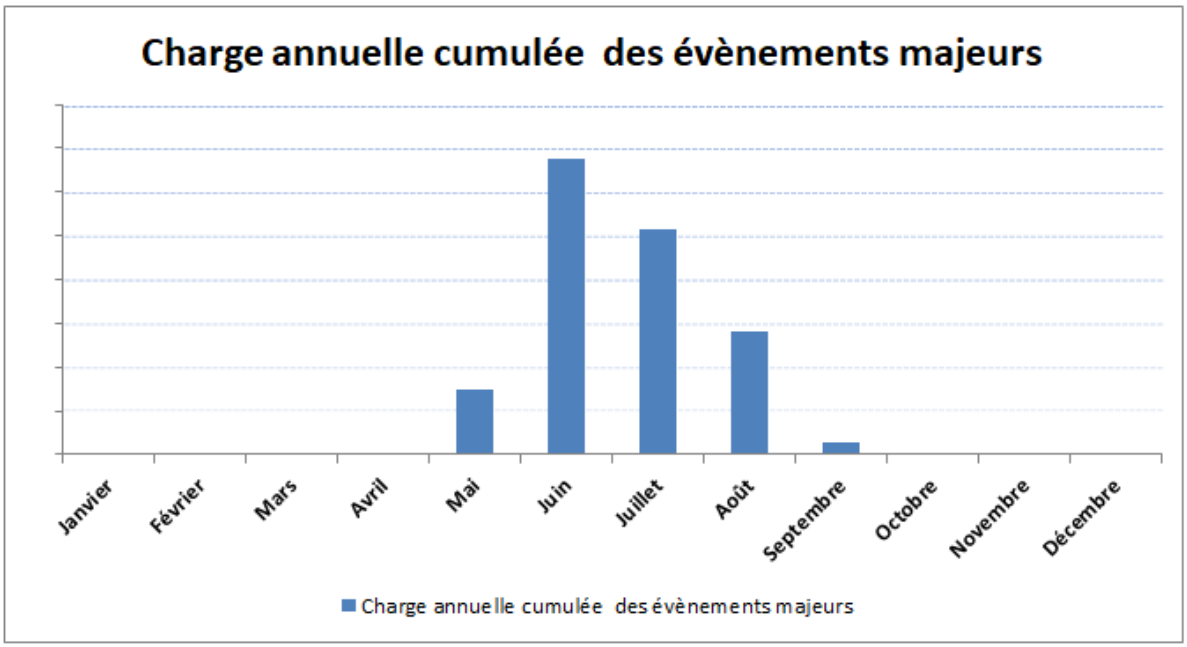
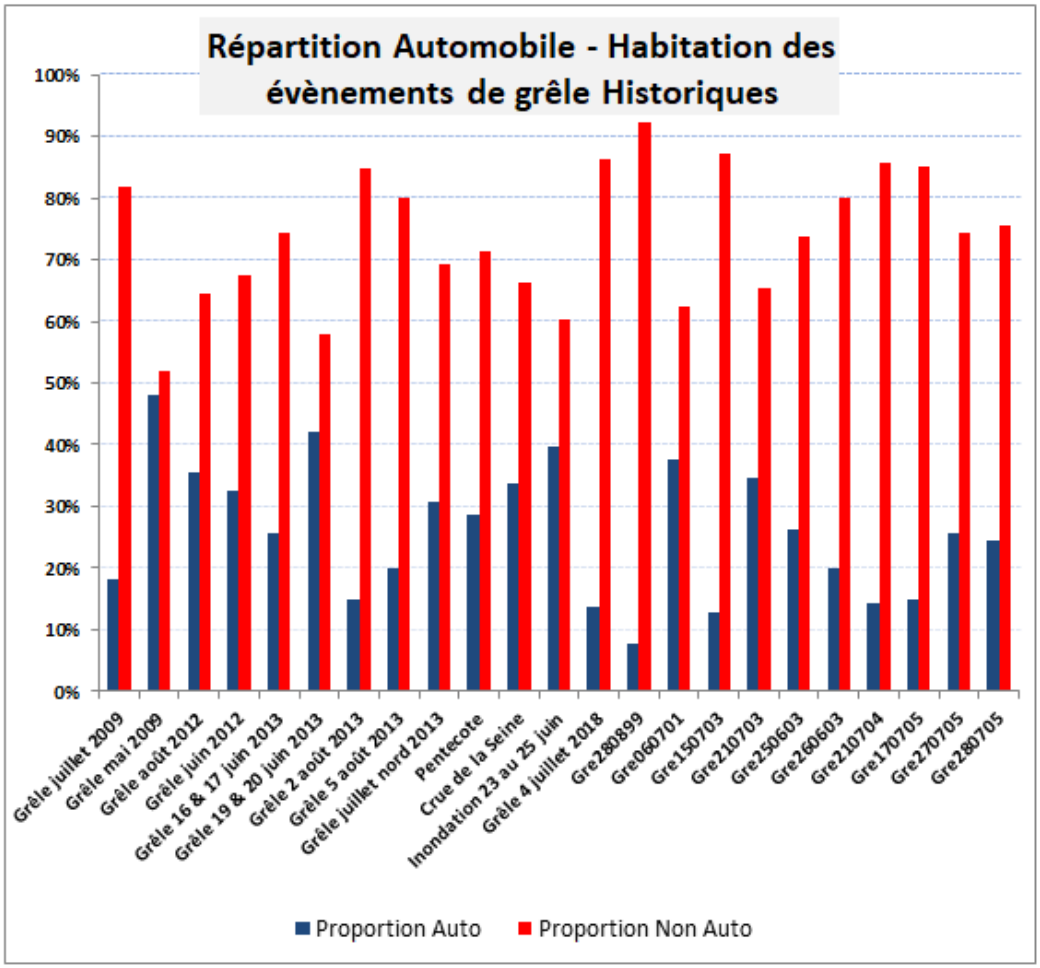


FIGURE 20 – Charge mensuelle historique des évènements supérieurs à 5 Millions d'euros

Répartition entre la charge automobile et habitation :

A l'échelle nationale, les proportions de sinistres automobiles et habitations sont quasiment équivalentes avec respectivement 48,78% et 51,22% du nombre total de sinistres. Cependant, la charge des sinistres due à la grêle est majoritairement portée par les habitations avec 71,51% du total.



Répartition des évènements majeurs sur le territoire français :

Les évènements majeurs ont lieu durant la période estivale. Ils sont engendrés par des fronts orageux venant du sud traversant la France en passant par le sud-ouest puis en rejoignant le coeur des terres.

Voici la présentation de 3 évènements majeurs ayant touché le portefeuille Pacifica. Chacun d'entre eux est décrit par une carte présentant les départements les plus touchés par l'évènement climatique et une carte affichant le point de survenance de chaque sinistre engendré par ce même évènement.

Grêle de Mai 2009 :

La Grêle de Mai 2009, engendrée par un front Sud-Nord présente la particularité d'avoir touché 3 zones distinctes pendant son passage. En effet, les départements représentant les proportions les plus importantes de la charge totale sont le Gard (30%), l'Oise (23%) et le Cher (11%) respectivement au Sud, au Centre et au Nord de la France.

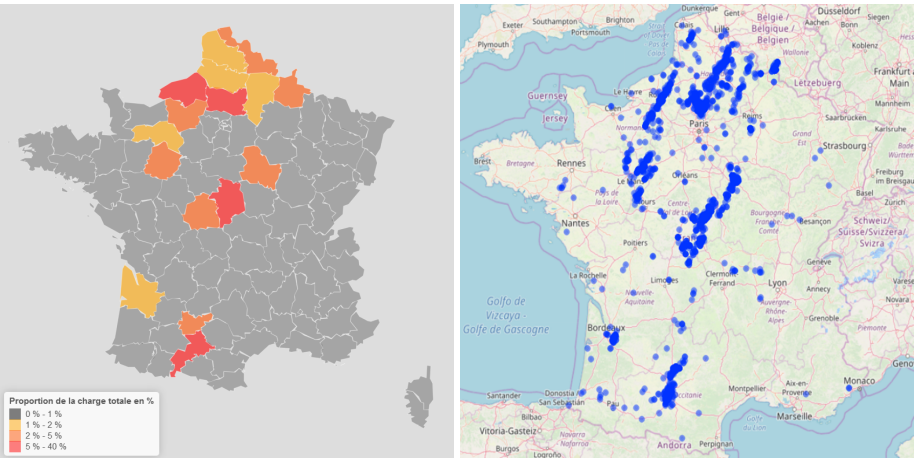


FIGURE 21 – Grêle de Mai 2009

Un même évènement peut donc présenter plusieurs "empreintes" de grêle distinctes.

Grêle du 2 Août 2013 :

La grêle du 2 Août 2013 a la particularité d'être un des cinq évènements les plus destructeurs des 20 dernières années mais dont 87% de la charge se concentre sur un seul département : la Dordogne. Les départements frontaliers que sont la Gironde et la Corrèze prennent une part relativement significative avec respectivement 5% et 6% de la charge tandis que la part du reste du territoire est négligeable.

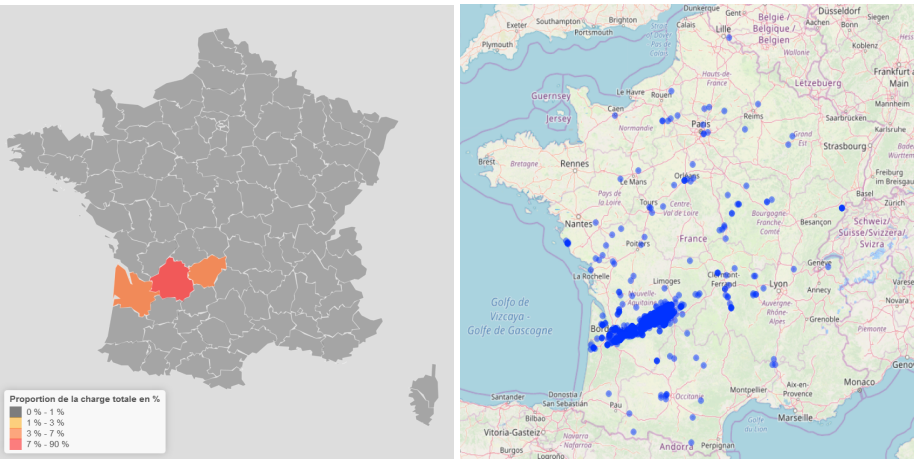


FIGURE 22 – Grêle du 2 août 2013

Pentecôte 2014 :

La Pentecôte de 2014 est l'évènement le plus important des 20 dernières années. Ce dernier a concentré sa sinistralité sur la région parisienne où l'on remarque une empreinte très marquée. Les départements les plus touchés sont la Seine-et-Marne (20%), le Val d'Oise (17%) et le Loiret (14%). Malgré une empreinte très marquée en Ile-de-France, cet évènement présente la particularité de très peu toucher la

ville de Paris. En effet, les habitations parisiennes sont majoritairement constituées d'appartements, des biens extrêmement peu sensibles à la grêle.

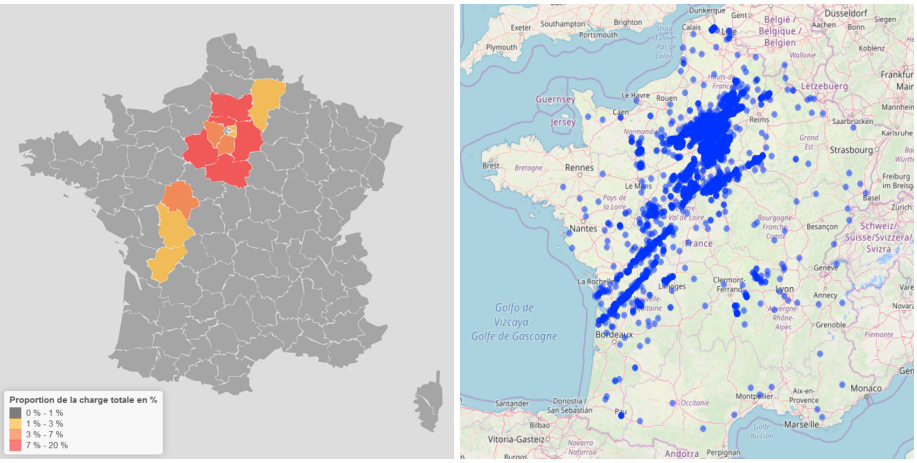


FIGURE 23 – Pentecôte 2014

La grêle s'oppose donc à des périls comme la sécheresse ou le gel par son hétérogénéité. En effet, ces derniers touchent des zones larges tandis que les chutes inhérentes à la grêle sont généralement plus localisées.

3 Les données disponibles

Précédemment, nous avons vu que les études récentes lient la présence de grêle à des conditions climatiques particulières.
Voici les principales données climatiques que nous avons retenues pour la modélisation du péril grêle dans le cadre de ce mémoire.

<u>Les données ECAD :</u>

L'European Climate Assessment & Dataset (ECAD) est une association regroupant les relevés de plusieurs instituts météorologiques européens afin de réaliser divers travaux (études, reconstitution de données manquantes ...).
En France, l'ECAD utilise les relevés issus de 106 stations Météo-France. Ces relevés sont **extrapolés** afin d'obtenir quotidiennement ces informations à une maille plus fine correspondant à des points espacés chacun de 0.5 degrés sur l'ensemble du territoire métropolitain français.
Certaines incohérences sont observées sur les données brutes de Météo-France du fait, par exemple, d'erreurs lors de certains relevés ou encore de l'évolution des techniques de recensement au fil du temps. Ces relevés sont donc traités en amont dans l'optique d'optimiser au maximum la qualité des données en entrée. Les extrapolations tiennent ensuite compte de plusieurs paramètres tels que l'altitude, les vents, le climat de la zone ...

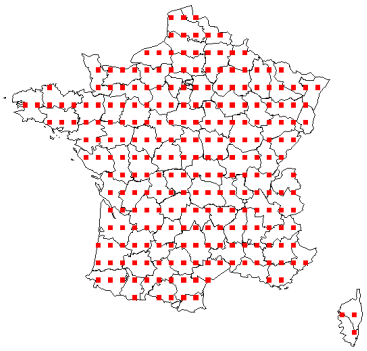


FIGURE 24 – Positionnement des 252 stations extrapolées de l'ECAD

A l'issue de ce retraitement, nous obtenons finalement 252 stations extrapolées ne présentant pas de données manquantes sur un historique d'une cinquantaine d'années et contenant quotidiennement les variables suivantes :

- Températures maximales quotidiennes (en Degrés) ;
- Températures minimales quotidiennes (en Degrés) ;
- Températures moyennes quotidiennes (en Degrés) ;
- Pressions moyennes quotidiennes (en Bars) ;
- Sommes des précipitations quotidiennes (en Millimètres).

Les données SYNOP :

Les données SYNOP de Météo-France sont constituées de relevés effectués **toutes les 3 heures** sur 52 stations en France Métropolitaine.
Officiellement, plus d'une trentaine de variables sont observées pour chaque station (pression, pluviométrie, températures, direction et vitesse du vent ...). Cependant, ces informations présentent une forte problématique de qualité de données. De nombreuses variables ne sont pas renseignées sur plusieurs stations, certaines stations disparaissent au fil du temps (car trop anciennes ou endommagées), d'autres apparaissent plus tard mais surtout, beaucoup de relevés ne sont tout simplement pas effectués sur plusieurs périodes. Certaines variables présentant trop de pertes ne sont donc pas exploitables.

FIGURE 25 – Positionnement des 52 stations SYNOP

Données exogènes :

Grâce aux informations disponibles sur le site data.gouv, nous avons accès aux caractéristiques géographiques de chaque commune : zone fluviale, littoral, plaine, reliefs ...
Freddy Vinet indique dans son étude que les reliefs sont en partie responsables du climat continental de l'Est de la France et jouent un rôle majeur dans l'apparition des grêles d'été. D'autre part, Freddy Vinet indique que plusieurs éléments montrent que les chutes de grêle sont sensibles aux « phénomènes de surfaces » et que « Le rythme des chutes de grêle prouve le rôle de la thermoconvection ». Ces données pourraient ainsi permettre d'étudier l'impact des caractéristiques géographiques sur la survenance de grêle.

4 La méthode de modélisation retenue

4.1 Description d'un modèle CAT

Les évènements naturels se distinguent généralement d'autres périls par leurs liens avec des phénomènes physiques. En effet, ils dépendent directement de phénomènes physiques caractérisés par leurs raretés et de leurs intensités, pour lesquelles l'approche statistique traditionnelle n'est pas toujours la plus adaptée. Les méthodes couramment mises en oeuvre pour modéliser des risques plus "classiques", comme l'approche dite "fréquence-sévérité", basées sur l'historique de distribution des pertes sont rigides car ne tenant pas compte de variables explicatives, notamment climatiques, et peuvent donc être challengés par des méthodes hybrides.

L'utilisation de modèles Catastrophes, basés sur les caractéristiques scientifiques du risque assuré en supplément des notions statistiques basées sur les données historiques de la survenance des sinistres, est donc devenu d'une importance capitale dans l'optique de générer au mieux une distribution des pertes.
Un modèle Catastrophes permet de modéliser l'impact d'un risque spécifique sur le portefeuille d'une compagnie d'assurance. L'objectif final est de simuler le risque sur un nombre de périodes (années) conséquent et ainsi d'obtenir des résultats chiffrés (charge annuelle, charge par évènement, périodes de retours, sensibilité géographique du portefeuille). L'entreprise est ainsi bien mieux informée sur les dégâts potentiels du risque sur son portefeuille et se trouve donc dans les meilleures dispositions pour mettre en place sa structure de réassurance ou encore constituer les provisions adéquates.

Les modèles Catastrophes se décomposent généralement en 3 modules : Aléa, Vulnérabilité et Financier.

Le module Aléa

Le module Aléa sert à reproduire les caractéristiques du phénomène physique inhérentes au risque. Il permet de simuler stochastiquement un nombre important d'évènements à partir des données physiques et statistiques. Généralement, des zones homogènes sont définies afin de modéliser la fréquence d'occurrence ainsi que l'intensité de l'évènement du risque climatique considéré (chute de grêle ou de neige, gel ...), une dépendance spatiale étant mise en place afin d'assurer la cohérence des simulations entre les zones.

Le module Vulnérabilité

Le module Vulnérabilité permet de lier les évènements simulés par le module Aléa à l'exposition du portefeuille de l'entreprise en mettant en perspective la géolocalisation du risque avec la nature de biens ou encore les sommes assurées. Une maison aura généralement une exposition plus grande à l'inondation qu'un appartement situé au troisième étage d'un immeuble. De nombreuses caractéristiques permettent de distinguer les biens assurés : le type de bien, les matériaux de construction, l'ancienneté du bien, la position géographique ... Cette connaissance de la structure du portefeuille permet une quantification précise des pertes associées à chaque évènement simulé en associant les taux de destruction correspondants.

Le module Financier

Le module Financier permet de calculer le montant des pertes finales de l'assureur. Il peut tenir compte de différents facteurs comme par exemple la sensibilité des pertes par rapport à la structure de réassurance, les franchises, la coassurance ...

En sortie de ce module, deux courbes sont généralement associées :

- La courbe OEP (*Occurrence Exceedance Probability*) qui associe à une période de retour donnée, la perte maximale annuelle correspondante ;
- La courbe AEP (*Aggregate Exceedance Probability*) qui associe une période de retour à la perte totale engendrée par une année de sinistralité simulée.

4.2 Présentation générale du modèle

Notre modélisation du risque de grêle est premièrement basée sur l'utilisation de la sinistralité Pacifica. Avec cette dernière, nous définissons précisément la localisation et l'intensité des chutes de grêle historiques. Dans un second temps, nous associons ces résultats avec les données météorologiques disponibles afin d'établir un potentiel lien de causalité. L'objectif final est de simuler 10 000 années de grêle, en estimer la sinistralité compte tenu de la structure du portefeuille de Pacifica, puis observer l'impact financier sur ce même portefeuille. Ainsi, l'entreprise sera en mesure d'apprécier au mieux son risque et de prendre les mesures adéquates afin de s'en protéger.

Traitement initial

Initialement, la mise en place du modèle se fera en plusieurs étapes. Tout d'abord, nous avons besoin d'identifier précisément la localisation et l'intensité des chutes de grêle historiques. Pour cela, nous utiliserons les données issues de la sinistralité historique de Pacifica afin d'établir une cartographie de ces chutes. Par la suite, dans l'optique d'établir une relation de cause à effet entre les variables disponibles (températures, pression ...) et la grêle, nous agrégerons l'ensemble des données existantes à la maille la plus fine possible. En croisant ces données, nous seront en mesure d'utiliser un algorithme prédictif afin de déterminer les variables responsables de la survenance de grêle.

Premièrement, afin d'obtenir la granularité la plus fine l'ensemble des données météorologiques disponibles sur un historique de vingt ans vont être regroupées sur un ensemble de 252 stations fictives correspondant à des zones géographiques réparties uniformément sur l'ensemble du territoire de France métropolitaine. Ces stations fictives sont issues de l'extrapolation de relevés météorologiques effectués sur 52 stations réelles Météo France.

Deuxièmement, les sinistres de Pacifica en France métropolitaine sur les 20 dernières années vont être regroupés par commune pour chaque journée de sinistralité. Ces sinistres vont ensuite être associés à la station fictive la plus proche afin de lui attribuer un label de survenance. La survenance (ou la non survenance) pourra ainsi être associée aux caractéristiques météorologiques de la zone correspondante sur cette journée.

Finalement, un algorithme de classification, le ***Random Forest***, sera lancé sur la base de données créée afin de réaliser deux objectifs : sélectionner les variables les plus significatives du modèle et être en mesure d'associer des survenances de grêle aux simulations des variables explicatives effectuées dans le module Aléa.

Module Aléa

Tout d'abord, le module Aléa modélisera les variables explicatives sélectionnées précédemment. Pour cela, nous utiliserons une méthode de réduction de dimension associant les stations météorologiques à des zones globales. Nous utilisons ensuite des descentes d'échelles afin de restituer artificiellement une granularité géographique plus fine. Nous calibrerons ensuite des copules et nous choisirons la plus adaptée afin de modéliser au mieux la dépendance entre les stations et les variables.

Ensuite, à partir de l'estimation des paramètres de la copule et des distributions historiques, nous effectuons des générations aléatoires afin d'obtenir 10 000 années de simulations.

Finalement, pour chaque journée de simulation, nous lancerons notre algorithme prédictif afin de déterminer les zones de survenance et de constituer les évènements de grêle.

Modules Vulnérabilité et Financier

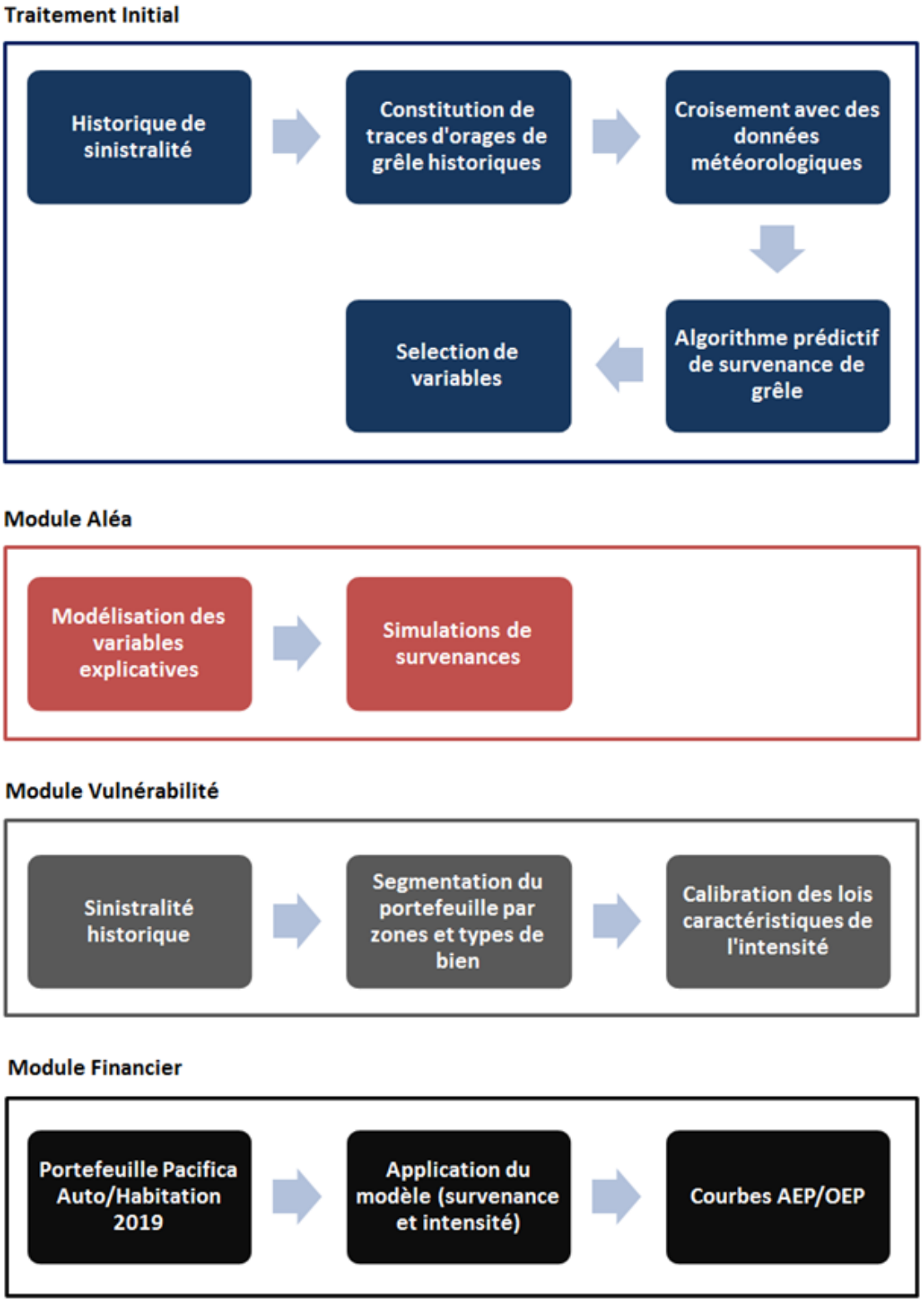
A partir de nos évènements simulés, nous devons désormais calculer la charge associée à chaque zone de survenance. Les zones de survenance étant les 252 stations fictives précédemment évoquées.

Tout d'abord, nous observons les caractéristiques du portefeuille Pacifica (automobiles, propriétaires de maisons, propriétaires d'appartements, locataires de maisons, locataires d'appartements ...). Pour chaque groupe de caractéristiques (ex : propriétaires de maisons, locataires d'appartements ...), nous observons la distribution des fréquences d'occurrence (à partir de laquelle nous définissons l'intensité des chutes de grêle) et des taux de destruction, qui nous permettront de déterminer respectivement le nombre de biens touchés et d'associer une charge à chacun de ces biens pour une zone donnée à partir du nombre de biens assurés et des sommes assurées pour chaque groupe de caractéristiques par zone de survenance.

Puis, nous calibrons une loi pour chaque paramètre sélectionné : fréquence d'occurrence pour les habitations, fréquence d'occurrence pour les automobiles, taux de destruction pour les automobiles, taux de destruction pour les propriétaires de maison, taux de destruction pour les propriétaires d'appartements, taux de destruction pour les locataires de maisons, taux de destruction pour les locataires d'appartements.
Ensuite, nous calibrons à nouveau des copules pour caractériser la dépendance entre la fréquence de biens touchés et les taux de destructions pour chaque classe.

Finalement, pour chaque zone de survenance de nos 10 000 années simulées, nous associons un nombre de biens touchés ainsi que des taux de destruction à chaque groupe de caractéristique pour chaque zone de survenance. Les sommes assurées par département nous permettent finalement d'associer un coût à chaque évènement.

Afin d'étudier l'exposition de l'assureur au risque de grêle, nous étudierons la distribution des pertes de l'assureur. Nous associons donc des périodes de retour aux années de sinistralités et aux occurrences afin respectivement de déterminer les courbes AEP et OEP.



4.3 L'intérêt de la modélisation du phénomène

Un assureur modélise généralement les risques liés aux périls naturels car il cherche à analyser et à évaluer le risque auquel il est exposé. Il se met ainsi dans les meilleures dispositions dans le cadre de

l'optimisation de son capital réglementaire et l'optimisation des traités de réassurance.

Optimisation du capital réglementaire

Suite à la réforme de **Solvabilité 2**, le calcul des fonds propres réglementaires pour chaque compagnie d'assurance a évolué. La directive est basée sur 3 piliers encadrant les exigences quantitatives et qualitatives pour la maîtrise de leurs besoins en fonds propres.
Solvabilité 2 impose notamment le calcul du **SCR (Solvency Capital Requirement)** qui est le capital nécessaire afin que la probabilité de défaut à horizon 1 an de l'assureur soit inférieure à 0.5%. Le SCR peut être calculé à partir de la formule standard [18] ou à partir d'un modèle interne (total ou partiel).
Dans le cas de la formule Standard, le calcul du SCR se fait par niveau de risque avant d'être agrégé à partir d'une matrice de corrélation à chaque étape.
Nous observons dans la figure suivante les différents modules de risques pris en compte dans la formule standard. On retrouve d'ailleurs la grêle comme sous-risque du risque Catastrophe du module Non-vie.

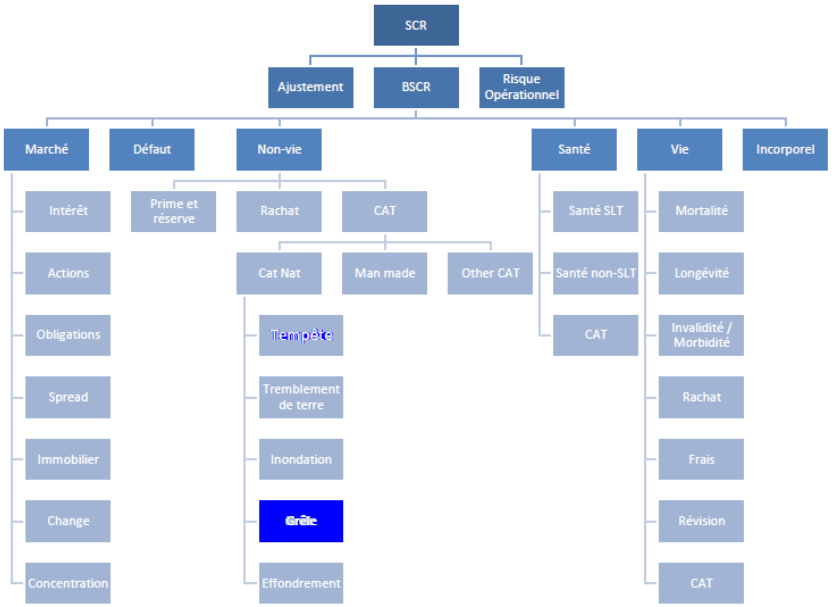


FIGURE 26 – Répartition des risques couverts par la formule standard pour le calcul du SCR

L'intérêt serait de comparer les résultats de la formule standard avec ceux d'un modèle CAT basé sur des variables climatiques représentatives du phénomène physique. L'approche par le modèle CAT est la plus précise, car elle est la seule à refléter le profil de risque réel de l'assureur et à s'adapter à des changements de structure de portefeuille (géographique, type de risque assurés,...). Cependant, cette méthode étant spécifique à chaque assureur, la validation d'un modèle catastrophe fait l'objet d'un contrôle pointu de la part du régulateur.

Optimisation de la Réassurance

La grêle est un péril pouvant connaître des pics de sinistralité. Cela peut se caractériser par une augmentation importante de la fréquence des évènements ou encore par un évènement d'une intensité

exceptionnelle. Dans ce genre de cas, le capital de solvabilité réglementaire n'est pas toujours suffisant pour que l'assureur puisse honorer ses engagements envers les assurés, il se tourne donc généralement vers la réassurance.
La réassurance est un contrat par lequel un assureur cède une partie de son risque à un réassureur en échange d'une rémunération.
Il existe deux types de réassurance : proportionnelle et non proportionnelle.

La réassurance proportionnelle, peu adaptée aux périls climatiques, consiste à définir la part de prime et de charge de sinistralité qui sera cédée au réassureur. Il en existe deux types : la Quote-Part, qui définit une part fixe des primes et de la charge de sinistralité qui sera cédée au réassureur, et l'Excédent de plein, qui va fixer un « plein de rétention », montant en dessous duquel le réassureur ne prend pas en charge le risque. L'assureur ne cédera donc que si la charge associée au risque dépasse ce montant.

La réassurance non proportionnelle consiste à transférer au réassureur les sinistres au-delà d'un montant spécifié (la priorité ou rétention) et jusqu'à un certain montant défini (la portée). La priorité est en général suffisamment élevée pour que le traité ne se déclenche qu'en cas d'évènement extrême. La réassurance non proportionnelle est donc particulièrement adaptée aux périls naturels.
La modélisation du risque de grêle permettra donc de déterminer la structure de réassurance la plus adaptée.

Deuxième partie

Traitement initial

5 Survenance de grêle et sélection de variables

5.1 Traitement des données

Chez Pacifica, les évènements de grêle sont définis sur des plages de 1 jour. Sur une journée, un évènement de grêle peut toucher des zones très vastes comme très restreintes que l'on appelle des « empreintes ». L'objectif du traitement des données est d'utiliser la sinistralité de Pacifica pour déterminer l'étendue des empreintes pour chacun des évènements historiques. Dès lors que les zones de survenance sont définies, le but est de pouvoir associer l'ensemble des variables disponibles (ECAD, Météo-France, reliefs ...) à une survenance ou une non-survenance de grêle.
In fine, cette préparation permettra la mise en place d'un modèle prédictif de la survenance de grêle à partir de l'ensemble des données à disposition et ainsi de déterminer les variables significatives pour cette survenance.
Lors de la simulation de nos 10 000 scénarios horizon 1 an, nous serons donc en mesure de déterminer les zones de survenances pour chaque évènement de grêle. Il restera par la suite à associer une intensité et un coût à chaque évènement simulé.

5.1.1 Données internes

La grêle est un phénomène ponctuel et très localisé. Les épisodes de grêle peuvent avoir lieu sur une période de quelques secondes jusqu'à quelques dizaines de minutes mais rarement plus et toucher des zones très restreintes.
Ces caractéristiques différencient structurellement la grêle d'autres périls comme le gel ou la subsidence, beaucoup plus homogènes et durables.
Afin de pouvoir associer une variable explicative à une survenance de grêle, nous commençons par caractériser les survenances de grêle historiques. Aucune base de données ou cartographie ne restitue quotidiennement l'historique des chutes de grêle en France. Nous décidons alors de partir de l'historique de sinistralité des évènements majeurs de Pacifica sur la période 1999 - 2018 et de réaliser nous-même cette cartographie en prenant en compte les zones référencées pour chaque sinistre. Nous prenons en compte les années de l'historique à partir de 1999 car c'est à partir de cette année que le portefeuille de l'entreprise devient suffisamment conséquent pour que la masse des sinistres par évènement soit significative.
Pour caractériser les évènements, nous observons les bases de sinistre de l'ensemble des produits **automobiles et habitations** sur la période définie. Nous extrayons de cet historique tous les évènements jugés majeurs (charge redressée supérieure à un certain seuil).
La sinistralité attritionnelle par opposition aux évènements majeurs correspond aux évènements de grêle dont la charge totale est inférieure au seuil choisi. La sinistralité attritionnelle se caractérise par sa stabilité dans le temps. Elle est relativement simple à modéliser en se basant sur des approches classiques et ne sera pas traitée dans ce mémoire.
Au sein de l'entreprise, les évènements de grêle sont par défaut considérés sur des plages de 1 jour. Une journée comporte un évènement de grêle à partir du moment où au moins une zone de survenance a été observée sur la journée.

Sur la période 1999 - 2018, 33 évènements majeurs sont observés sur la France métropolitaine.

Les paramètres suivants sont sélectionnés dans les bases internes pour chaque sinistre :

- Date ;
- Nom et Numéro de l'évènement
- Code postal ;
- Code INSEE ;
- Charge nette.

Les charges engendrées par les sinistres ne sont pas comparables au cours du temps. Il faut tenir compte de l'évolution de cette valeur pour avoir une vision actualisée de la charge sinistre.
Les charges nettes sont redressées en tenant compte de 2 facteurs :

- **l'inflation** : Indice de la Fédération Française du Bâtiment (FFB) du coût de la construction pour les sinistres habitation, et Indice INSEE du coût de la réparation automobile pour les sinistres automobiles.
- **l'index années d'assurance** : Indice d'évolution de la taille du portefeuille au cours du temps. Dans le cas de Pacifica, l'index année d'assurance est très important car l'entreprise a réalisé une forte croissance sur les 20 dernières années et la sinsitralité entre 2 années de l'historique n'est donc évidemment pas comparable.

Nous regroupons les sinistres par date et par code postal. Nous choisissons le code postal comme zone de référencement au dépend du code INSEE, pourtant plus précis, car la perte d'information sur le portefeuille est respectivement de 2% contre 10%.
Finalement, nous obtenons une liste contenant pour chaque journée de survenance historique l'ensemble des codes postaux touchés ainsi que le nombre de sinistres et la charge redressée cumulée pour chacun de ces codes postaux.

5.1.2 Traitement des données climatiques

Nous avons précédemment vu que la grêle était un phénomène très localisé, nous décidons donc de l'étudier à la maille la plus fine disponible, c'est à dire à partir des stations fictives extrapolées de l'ECAD. Nous récupérons donc l'ensemble des relevés des 5 variables disponibles pour les 252 stations sur la période 1999 - 2018 :

- Températures maximales quotidiennes (en Degrés) ;
- Températures minimales quotidiennes (en Degrés) ;
- Températures moyennes quotidiennes (en Degrés) ;
- Pressions moyennes quotidiennes (en Bars) ;
- Sommes des précipitations quotidiennes (en Millimètres).

Les données climatiques SYNOP existent à une maille moins fine (52 stations) mais sont constituées d'un grand nombre de variables intéressantes et **relevées toutes les 3 heures**, ce qui est un point positif majeur car nous avons vu précédemment grâce aux relevés de l'Anelfa que les chutes de grêle ont tendances à avoir lieu dans l'après-midi. Un pas de 3h permet donc d'observer les différences de valeurs en cours de journée pouvant être liées à la survenance de grêle. Nous excluons les variables présentant trop de pertes et nous gardons finalement les variables suivantes :

- Pressions (00h, 3h, 6h, 9h, 12h, 15h, 18h, 21h) en hecto Pascal.

— Températures (00h, 3h, 6h, 9h, 12h, 15h, 18h, 21h) en Kelvin.

— Vitesse des vents (00h, 3h, 6h, 9h, 12h, 15h, 18h, 21h) en mètres par secondes.

— Direction du vent (00h, 3h, 6h, 9h, 12h, 15h, 18h, 21h) en degrés.

Afin d'observer l'importance qu'un moment de la journée à dans l'apparition de grêle, ces variables sont décomposées chacune en 8 variables horaires, "Pression à minuit", "Pression à 3h", ..., "Pression à 21h", "Température à minuit", ..., "Température à 21h" ...
Pour agréger les données ECAD et SYNOP, nous attribuons à chaque relevé quotidien des stations extrapolées de l'ECAD les relevés issus de la station SYNOP la plus proche. Un axe d'amélioration aurait été d'extrapoler les relevés SYNOP en tenant compte de l'altitude, de la distance ou encore des vents pour améliorer la qualité de ces variables.
La grêle étant un phénomène ponctuel, il est intéressant d'étudier les variations climatiques au cours de la journée. Nous créons donc des variables supplémentaires correspondant aux deltas entre les différents horaires (exemple : différence de pression entre 15 heures et 18 heures, différence de température entre 3 heures et 6 heures ...).
Nous obtenons finalement 92 variables explicatives climatiques journalières sur un historique de 20 ans pour chacune des 252 stations extrapolées de l'ECAD.

5.1.3 Regroupement des données

L'objectif de ce procédé est de caractériser dans cette nouvelle base de données la survenance ou la non survenance par une nouvelle variable booléenne (1 ou 0).

Pour cela, nous commençons par associer chaque station extrapolée de l'ECAD aux codes postaux les plus proches. Un carré de 0,5 degrés de côté est donc tracé autour de chaque station de manière à ce que chacune d'entre elles occupe une zone unique et que l'ensemble du territoire français soit quadrillé. Ensuite, nous obtenons la position de chaque code postal à partir des données fournies par data.gouv. Nous associons chaque code postal à une station extrapolée de l'ECAD si son centroïde se trouve dans la zone d'une station extrapolée de l'ECAD.

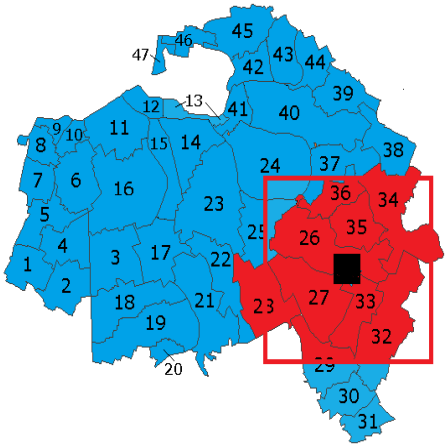


FIGURE 27 – Exemple d'association de codes postaux à une station fictive

A partir de notre historique de survenances constitué à l'aide des données internes de Pacifica (nombre

de sinistres et charges cumulées par code postal quotidiennement depuis 20 ans), nous sommes en mesure de relier la sinistralité historique aux stations ECAD et donc à notre base de données climatiques.

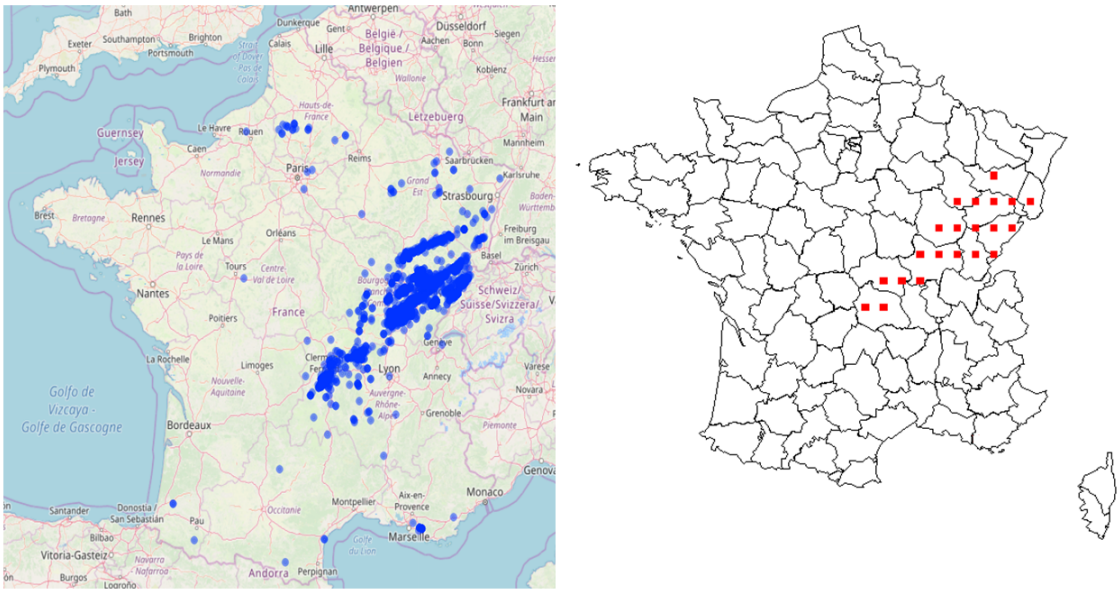


FIGURE 28 – Stations ECAD activées pour l'évènement de grêle de Juin 2012

Nous filtrons finalement cette base en attribuant une survenance uniquement si le nombre de sinistres par code postal dépasse un certain seuil. De cette manière, nous excluons les sinistres à retardement (ayant eu lieu auparavant mais déclarés au moment de la découverte du sinistre) ou encore les sinistre automobiles déclarés dans la zone de résidence du propriétaire mais s'étant par exemple déroulés sur le lieu de vacances de l'assuré. Les produits d'assurance automobile présentent la particularité de pouvoir se déplacer. Le code postal associé à un sinistre Auto peut donc ne pas correspondre au Code postal où le sinistre a eu lieu (ex : suite à un trajet de retour de vacances de l'assuré). Nous considérons dans notre modèle que ce biais est négligeable.

Finalement, les données exogènes "type de zone" (relief, fluvial, maritime, plaine) et "altitude de la station" sont ajoutées afin d'observer l'impact de la topologie sur l'apparition de grêle. Le nombre de variables explicatives X_i en entrée du modèle est de 59 auxquelles s'ajoute notre variable cible booléenne Y : la survenance.

5.1.4 Caractéristiques de la base finale

Nous observons que la proportion d'occurrences de grêle par jour et par station est extrêmement faible.

Proportion de Non survenances	Proportion de survenances
99,002%	0,998%

Ce phénomène est donc rare et géographiquement restreint, ce qui confirme les dires des différentes études existant sur le sujet. En effet, les évènements majeurs de grêle sont rares (0 à 5 par an) et ne vont chacun activer qu'un nombre restreint de stations.

5.2 Détermination de la survenance

Nous allons maintenant utiliser deux algorithmes prédictifs afin d'identifier la survenance de grêle. La survenance étant une variable booléenne, nous utilisons une régression logistique que nous mettons en parallèle d'un algorithme plus complexe mais très efficace dans ce cadre : le *Random Forest*. L'algorithme retournant la meilleure qualité de prédiction sera choisi.
Nous commençons par séparer nos données en sous-base d'entraînement et en sous-base de test afin de mettre en place une procédure d'apprentissage-validation.

Procédure d'apprentissage/validation : Cette procédure évalue la performance d'un modèle en terme de prévision. On sépare de manière aléatoire les données en deux parties distinctes :

— $D_A = (y_i, x_i), i \in I_A$ un échantillon d'apprentissage de taille A ;

— $D_V = (y_i, x_i), i \in I_V$ un échantillon de validation de taille V.

tels que $A + V = n$, $I_A \cap I_V = \emptyset$ et $I_A \cup I_V = \{1, ..., n\}$. L'échantillon d'apprentissage est utilisé pour estimer les paramètres du modèle, celui de validation pour estimer l'erreur ou la probabilité d'erreur. Une fois le modèle construit sur l'ensemble des observations, les paramètres θ du modèle sont réestimés sur D_A. Puis, la variable réponse est prédite sur D_V.

La quantité de données nécessaire à l'apprentissage est supérieur à la quantité de données nécessaire pour effectuer nos test. Nous décidons donc de séparer nos bases d'entraînement et de test respectivement en 2/3 et 1/3 de la base initiale.

5.2.1 Régression logistique

Notations [25] :

— $Y = (Y_1, ..., Y_n)^T$: Vecteur colonne des variables réponses ;
— $X_i = (X_{i1}, ..., X_{ip})$: Vecteur colonne des caractéristiques de Y_i.
— X matrice de taille n x p dont les lignes sont les vecteurs lignes X_i^T.

$$X = \begin{pmatrix} X_1^T \\ X_2^T \\ ... \\ X_p^T \end{pmatrix} = \begin{pmatrix} X_{1,1}...X_{1,p} \\ X_{2,1}...X_{2,p} \\ ... \\ X_{n,1}...X_{n,p} \end{pmatrix}$$

.

— $\beta = (\beta_1, ..., \beta_p)^T$: Vecteur colonne des paramètres

Définition : Modèle Linéaire Généralisé

Un modèle est un GLM si :

— L'espérance conditionnelle $Y|X$ suit une certaine loi issue d'une famille exponentielle, $Y|X = x \sim \underset{(\theta,\alpha)}{P}$:

$$f(y_i, \theta_i, \alpha) = exp\left(\frac{y_i\theta_i - b(\theta_i)}{\alpha} + c(y_i, \alpha)\right)$$

où α est le paramètre de dispersion.

— $Y|X$ s'exprime selon une relation linéaire à une transformation près (caractérisée par la fonction de lien).

$$g(E(Y|X)) = g(\mu(x)) = X_1\beta_1 + X_2\beta_2 + ... + X_p\beta_p$$

où g est la fonction lien canonique (celle qui est asymptotiquement la plus efficace) tel que $g = (b')^{-1}$.

Estimation de β :

Les paramètres β sont estimés par la méthode du maximum de vraisemblance. On maximise alors la log-vraisemblance :

$$L_{ln}(\theta, \alpha; y) = ln\left(\prod_{i=1}^{n} f(y_i; \theta_i, \alpha)\right) \tag{1}$$

$$= \sum_{i=1}^{n} ln\left(f(y_i; \theta_i, \alpha)\right) \tag{2}$$

$$= \sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{\alpha} + c(y_i, \alpha) \tag{3}$$

Comparaison des modèles :

En présence de plusieurs modèles, il est nécessaire de déterminer quel est le meilleur. La vraisemblance permet de mesurer l'adéquation du modèle aux données, il est alors possible de comparer la log-vraisemblance, et donc la déviance des modèles afin de détecter le plus performants. Néanmoins, le critère de la vraisemblance ne permet pas toujours de choisir le meilleur modèle car la log-vraisemblance croît avec le nombre de paramètres estimés. Il convient donc d'utiliser des critères de vraisemblance pénalisés comme l'AIC ou le BIC.

Définitions :

— Le critère d'information d'Akaike AIC s'écrit comme suit :

$$AIC = 2k - 2ln(L)$$

.

— Le critère de Schwartz BIC s'écrit comme suit :

$$BIC = kln(n) - 2ln(L)$$

.

où k est le nombre de paramètres à estimer du modèle et L est le maximum de la fonction de vraisemblance du modèle.
Le meilleur modèle minimisera donc un de ces critères.

Déviance :

La déviance permet de mesurer l'ajustement du modèle en comparant le modèle saturé au modèle considéré selon la formule suivante :

$$\Delta = 2(l_s - l)$$

où l_s la log vraissemblance du modèle saturé et l la log vraisemblance du modèle initial.

Cas de la régression logistique :

Le Modèle linéaire généralisé ou GLM est un modèle où l'on essaye d'expliquer une variable Y en fonction des variables explicatives $(X_1, ..., X_n)$.
La survenance peut être exprimée par une variable booléenne prenant la valeur 1 lorsqu'on observe un évènement majeur de grêle et de 0 lorsque ce n'est pas le cas.
L'objectif est donc de pouvoir probabiliser cette survenance en utilisant une régression logistique.

La régression logistique est un cas particulier du modèle linéaire généralisé.
Dans le cas de la régression logistique, nous supposons que $Y_i|X$ suit une loi de Bernoulli et la fonction de lien canonique utilisée est la fonction "logit" :

$$g^{-1}(x) = \frac{1}{1 + exp(-x)}$$

Cette dernière va permettre de restituer une probabilité de survenance pour chaque individu en gardant la valeur de sortie dans l'intervalle [0,1].

Application de la régression logistique à l'historique de survenance :

En entrée, nous disposons de 59 variables explicatives. Bien évidemment, nous ne voulons intégrer au modèle que les variables significatives, pour cela, nous devons supprimer celles qui n'apportent aucune information au modèle (trop corrélées avec certaines variables explicatives ou tout simplement inadaptées).
Nous décidons donc de sélectionner les variables les plus significatives à l'aide de la *"forward stepwise selection"*.

Procédure de sélection de variables par méthode ascendante (*forward stepwise selection)*

— L'algorithme débute en entrée avec le modèle M_0, c'est à dire le modèle ne contenant que la constante ;

— Il effectue ensuite les k régressions possibles (k étant le nombre de variables explicatives) avec une seule variable explicative. Pour chacune d'elles, le test choisi est appliqué (dans notre cas l'AIC). Le modèle retenu est celui pour lequel la variable explicative choisie est la plus significative ;

— Appliquer cette démarche pour chaque itération i, en effectuant les (ki) régressions possibles, effectuer le test et retenir le modèle pour lequel la variable est la plus significative. Si on ne peut plus introduire de variables significatives dans le modèle (exemple : l'AIC ne baisse pas), alors elles ne sont pas retenues et le processus s'arrête.

La *"forward stepwise selection"* retourne 30 variables jugées significatives dans notre modèle (température maximale, pression, précipitations ...).

Nous souhaitons désormais évaluer la qualité de prédiction de notre Régression logistique. Pour cela, nous mettons en perspective le nombre de bonnes prédictions avec le nombre de mauvaises prédictions pour les survenances et les non survenances.

La régression logistique détermine ce que l'on peut considérer comme une "probabilité de survenance", un réel compris entre 0 et 1. Il reste ensuite à choisir le seuil au dessus duquel nous attribuons une "survenance" et en dessous duquel nous attribuons une "non survenance". Nous balayons donc les différents seuils pas à pas afin de déterminer lequel retourne la meilleure précision.

Pour témoigner des différences engendrées par le choix des seuils, nous observons les différences de résultats lorsque l'on place notre seuil respectivement à 10% et 30% de chances de survenance.

Seuil à 30% de chances de survenance :

Les 33 évènements majeurs de grêle historiques comportent en totalité 1505 zones de survenance. Sur ces 1505 zones de survenances historiques observées, nous en prédisons 70, ce qui représente une précision de 4.6%. De plus, nous prédisons 165 survenances supplémentaires sur des zones de l'historique ne présentant pas de survenance de grêle (235% de plus que les prédictions correctes).

Nous présentons dans les cartes suivantes quelques exemples de nos prédictions sur des journées de notre historique. Premièrement, nous prenons des journées ayant enregistré une survenance réelle, afin de comparer les traces réelles aux prédictions. Puis, nous regardons quelques journées de non survenances afin d'observer si l'algorithme n'a pas tendance à associer la grêle à d'autres phénomènes (pluie, neige ...).

Exemples de survenance réelles (en bleu) et de survenances simulées (en rouge) :

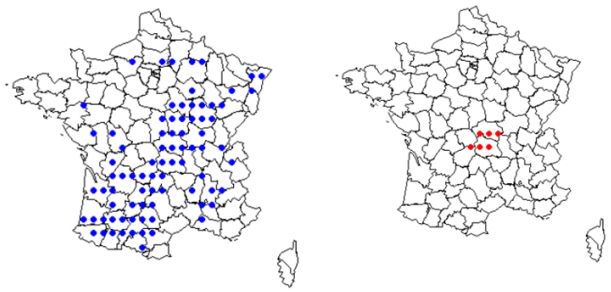


FIGURE 29 – Évènement du 6 Août 2013

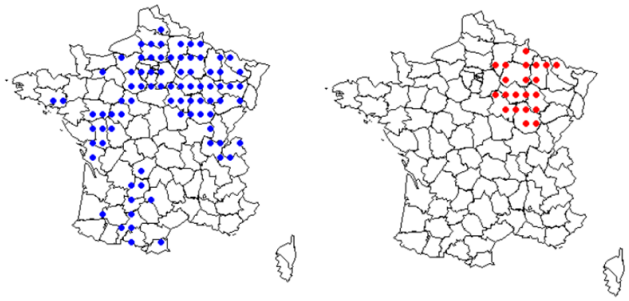


FIGURE 30 – Évènement du 19 Juin 2013

Nous observons ensuite quelques exemples de journées de l'historique n'ayant aucune survenance de grêle réelle contrairement à la prédiction du modèle.

FIGURE 31 – Journée simulée du 4 Juillet 1999

FIGURE 32 – Journée simulée du 9 Juin 2000

Les évènements historiques ne sont pas entièrement restitués, seul le coeur des empreintes (les zones où la grêle a été particulièrement intense) apparaît comme sinistré en sortie du modèle. Un seuil à 30% implique donc la perte des zones les moins intenses d'un orage grêligène. De plus, les jours de non survenances ayant des traces simulées sont caractérisés par des formes similaires aux orages grêligène historiques. La régression logistique capte donc de manière globale des phénomènes climatiques similaires à la grêle (pluies éparses, orages de pluie, ...).

Seuil à 10% de chances de survenances :

Sur les 1505 survenances historiques observées, nous en prédisons 340, ce qui représente une précision de 22%. De plus, nous prédions 1482 survenances supplémentaires sur des points de l'historique ne présentant pas de survenances de grêle (435% fois plus que les prédictions correctes).

De même que précédemment, nous observons nos prédictions en face de nos journées historiques.

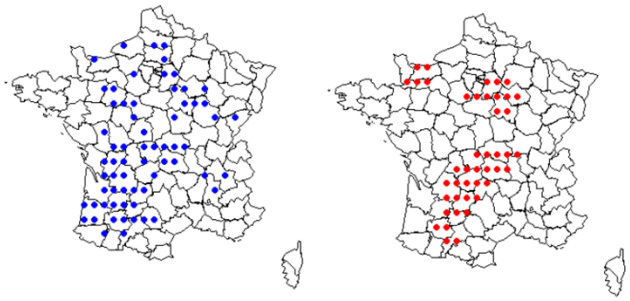


FIGURE 33 – Évènement du 4 Juillet 2006

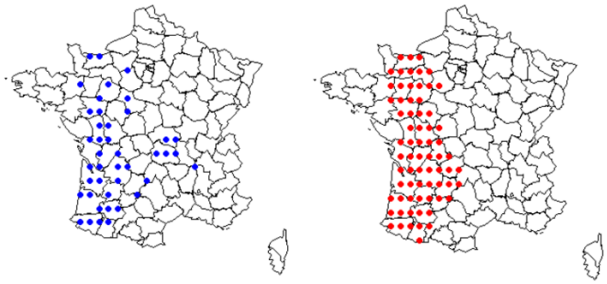


FIGURE 34 – Évènement du 15 Juillet 2003

Exemples de survenances simulées sans évènement réel (en rouge) :

FIGURE 35 – Journée simulée du 4 Août 1999

FIGURE 36 – Journée simulée du 6 Août 1999

Les évènements historiques sont entièrement restitués, voir sur-restitués mais la sortie contient un nombre trop important de mauvaises prédictions. Cela témoigne de l'incapacité du modèle à introduire de la subtilité dans le traitement des variables. Ce dernier n'est par exemple pas capable de différencier les orages de grêles des orages de pluie. La variable « somme des précipitation » est un paramètre important dans la survenance de grêle, elle se voit donc attribuée un coefficient significatif. Cependant, la « somme des précipitations » capte aussi bien les chutes de grêle, de neige ou de pluie. Des pluies abondantes auront donc de fortes chances d'être assimilés à de la grêle dans ce type de modèle.

Conclusion :

La qualité de prédiction des modèles est faible. Les résultats obtenus ne sont donc pas satisfaisants. En effet, en attribuant un coefficient à chaque variable, la régression logistique va lui donner une importance dans le modèle. Cependant, cette méthodologie ne capte pas les subtilités inhérentes à la survenance de grêle. Par exemple, si la température maximale est élevée, la probabilité d'occurrence de grêle va augmenter significativement. Or, au-dessus d'un certain seuil, la température est beaucoup trop élevée pour que la grêle puisse atteindre le sol sous la forme solide. De la même manière, si les conditions de température sont réunies mais que la pression n'est pas suffisamment basse alors il n'y aura pas de survenance de grêle.

Nous avons donc besoin d'une méthodologie plus complexe, capable de nuancer l'importance d'une variable en fonction des différentes situations. Les "Arbres de décisions" et notamment le "*Random Forest*" présentent ces caractéristiques.

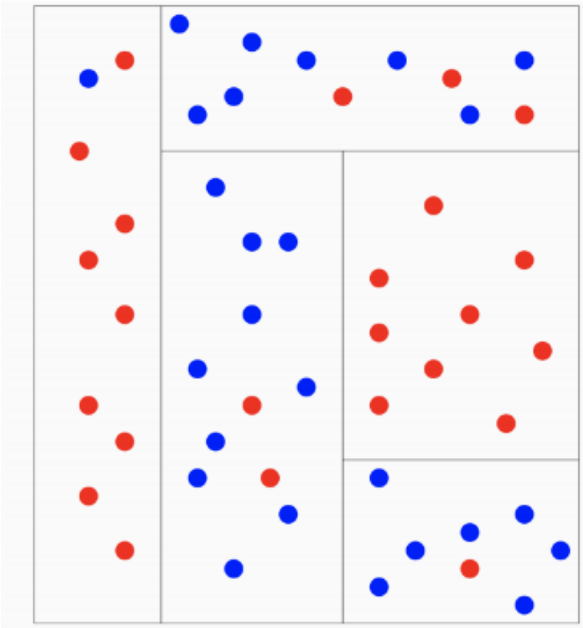
5.3 *Le Random Forest*

Le *Random Forest* [26] est un algorithme basé sur 2 méthodes : la création d'arbres de décisions basés sur l'algorithme de **CART** et la méthode du **bagging**.

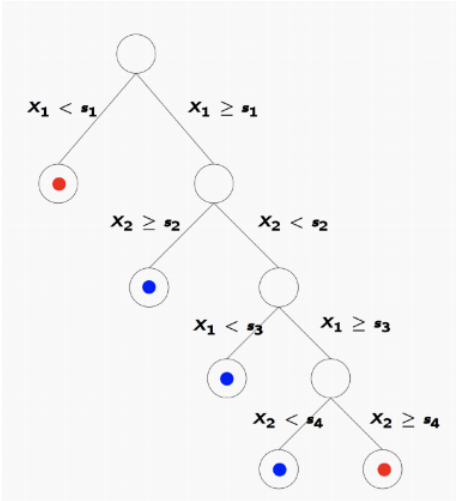
5.3.1 Les arbres de décision

Les arbres de décisions, basés sur l'algorithme de **CART (Classification And Regression Trees)** sont définis de telle sorte qu'on estime une variable Y qui peut être catégorielle (classification) ou continue (régression) à partir des variables d'entrée $X, ..., X_n$ (catégorielle ou continue également).

L'algorithme de CART propose de restreindre son attention aux partitions binaires récursives. A chaque étape, la méthode sépare les données en deux régions en fonction d'une variable de séparation et d'un point de séparation.



A la fin du processus, la règle de classification fait que la classe la plus représentée dans chaque noeud final est considérée comme la valeur de sortie.



Vocabulaire :

- Un noeud a exactement 0 ou 2 "enfants" ;
- Chaque séparation définit 2 noeuds "enfants", le gauche et le droit ;
- Un noeud sans enfant est appelé un noeud terminal ;
- Le premier noeud est appelé racine ;
- Nous passons d'un noeud à son enfant droit ou gauche en répondant à une question simple : X_j est-il supérieur ou égal à α ?

Comment séparer les noeuds :

Nous voulons séparer un noeud N en un noeud enfant gauche N_L et un noeud enfant droit N_R. L'enfant dépend donc de la séparation définie par une caractéristique et un seuil (j, t).

$$N_L(j,t) = x \in N : x_j < t \text{ et } N_R(j,t) = x \in N : x_j \geq t$$

.

Pour trouver la meilleure séparation, il faut ainsi trouver la meilleure paire (j, t) en deux étapes :

- Comparer l'impureté du noeud N avec celles de $N_L(j,t)$ et $N_R(j,t)$ pour chaque paire (j,t) ;
- Utiliser le gain d'information apporté pour chaque paire (j,t).

Dans les cas de classification, nous calculons d'abord la distribution des classes, puis on considère une mesure d'impureté. Nous aborderons ici **l'indice de Gini** mais il en existe d'autres comme l'indice d'entropie.

Indice de Gini :

$$G(N) = G(p_N) = \sum_{k=1}^{K} p_{N,k}(1 - p_{N,k})$$

Nous notons que si N est pur, $G(N) = 0$.

On considère ensuite le gain d'information :

$$IG(j,t) = G(N) - \frac{|N_L(j,t)|}{|N|} G(N_L(j,t)) - \frac{N_R(j,t)}{N} G(N_R(j,t)$$

.

L'algorithme de CART :

L'algorithme de CART va créer son arbre itérativement : pour chaque noeud N de l'arbre, il va trouver le meilleur couple (j,t) qui maximise le gain d'information $IG(j,t)$. Cela engendre la création de 2 noeuds enfants. Cette action est répétée pour chaque noeud jusqu'à ce qu'une condition d'arrêt soit atteinte.

Quelques exemples de conditions d'arrêt :

— La profondeur maximale de l'arbre est atteinte ;

— tous les noeuds ont un échantillon inférieur à un certain nombre ;

— L'impureté dans chaque noeud est suffisamment petite.

Dans notre cas, la variable réponse, la survenance, est un label binaire (0 ou 1). L'arbre va donc séparer les données en 2 de manières successives afin d'isoler les survenances des non-survenances au maximum en bloc. On part du nœud central contenant toute l'information puis on va séparer ces nœuds en blocs rectangulaires selon une certaine dimension (variable choisie pour effectuer la séparation du nœud) et selon un certain seuil (position de la coupure sur l'échelle de l'axe de la variable de séparation). Ces paramètres sont définis en comparant les mesures d'impureté entre les différentes combinaisons de nœuds enfants avec le parent (indice de Gini). L'arbre s'arrête si une condition d'arrêt est atteinte (longueur maximale, nombre d'individus minimum atteint ...).

Les arbres de décisions ont tendance à *"overfiter"*, c'est à dire à réaliser un sur-apprentissage, notamment lorsque la profondeur de l'arbre est importante.
Cette profondeur est un paramètre clef car si l'arbre n'est pas assez long on va perdre en qualité de prédiction (petite variance mais biais fort). En revanche, s'il est trop long, les cuts seront beaucoup trop adaptés à nos données. Par conséquent, cela provoquera un *"overfit"* (petit biais mais grosse variance). Les résultats sont généralement peu concluants et faibles mais facilement interprétables.

Le *Random Forest* part donc de ces arbres de décisions, mais va utiliser la méthode du bagging afin de perfectionner les résultats.

5.3.2 Le bagging

Les arbres de décision souffrent d'une forte variance.
Pour répondre à cette problématique, l'idée est assez simple : privilégier l'agrégation de modèles simples plutôt que d'utiliser un modèle très sophistiqué.

Le bagging est donc une méthode ayant pour objectif de réduire la variance de différentes méthodes d'apprentissage statistique.

Le bagging est un procédé d'agrégation de modèles sur échantillons bootstrappés.

Cela signifie que l'on applique le même algorithme sur plusieurs sous-échantillons de données. Ces échantillons sont créés par bootstrap, un tirage aléatoire avec remise (on peut donc avoir plusieurs fois la même valeur). Le résultat final sera ensuite obtenu en faisant la moyenne des différentes prédictions. Dans le cas binaire, la classe choisie est celle ayant le plus de chances d'apparaître.

Le *Random Forest* va donc agréger l'ensemble des arbres qu'il a utilisés pour réaliser ses prédictions. Dans notre cas, cette méthode est caractérisée par une longueur (nombre de noeuds) ainsi que par un nombre d'arbres.

Application du *Random Forest* à nos données :

Le *Random Forest* détermine à chaque noeud la variable et le seuil permettant le mieux de séparer les classes selon l'indice de Gini. Nous appliquons donc cet algorithme sur notre base d'entraînement, issue de notre base d'historique avec l'ensemble des variables explicatives.

Nous observons ensuite dans la figure suivante l'indice de Gini cumulé moyen pour les 6 variables jugées les plus significatives par le modèle.

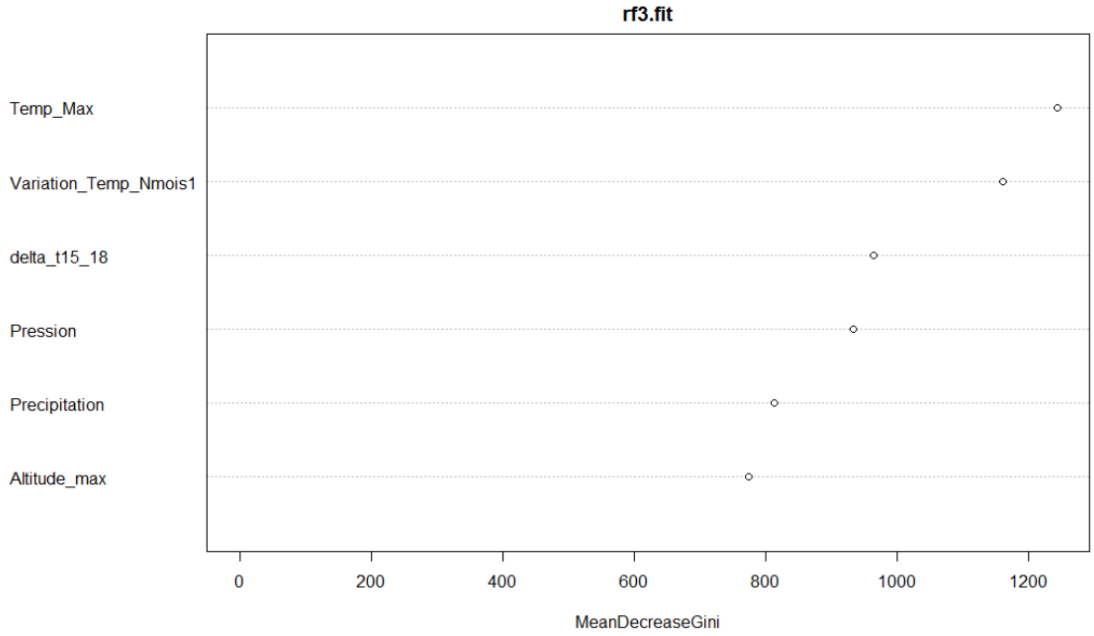


FIGURE 37 – Indice de Gini cumulé moyen pour chaque variable

Nous remarquons que parmi les 6 variables les plus significatives, 4 sont issues de l'ECAD (températures maximales, Variation de température moyenne entre le jour même et la veille, pression et somme des précipitations), ce qui témoigne de l'importance de la finesse géographique des informations, ce qui n'est pas le cas des variables SYNOP. A cela, on ajoute tout de même une variable SYNOP : La variation de températures entre 15h et 18h et 1 variable exogène : l'altitude de la zone.

Choix du nombre d'arbres :

L'out of bag error est une mesure de l'erreur de prédiction de *Random Forest*. Chaque arbre de la forêt est construit sur une fraction des données ("in bag"), la fraction qui sert à l'entraînement de l'algorithme. Ensuite, pour chacun des individus de la fraction restante ("out of bag"), l'arbre prédit une classe. L'Out of Bag error est donc l'estimation de l'erreur de prédiction sur une sous-base de test. Nous cherchons à minimiser cette erreur grâce au choix du nombre d'arbres.
Le *Random Forest* réalise ses séparations suite aux moyennes obtenues sur un groupe d'arbres. Or, à partir d'un certain nombre d'arbres, le gain est nul et le poids des calculs ne fait qu'augmenter.

La figure suivante indique "l'Out of Bag Error" en fonction du nombre d'arbres.

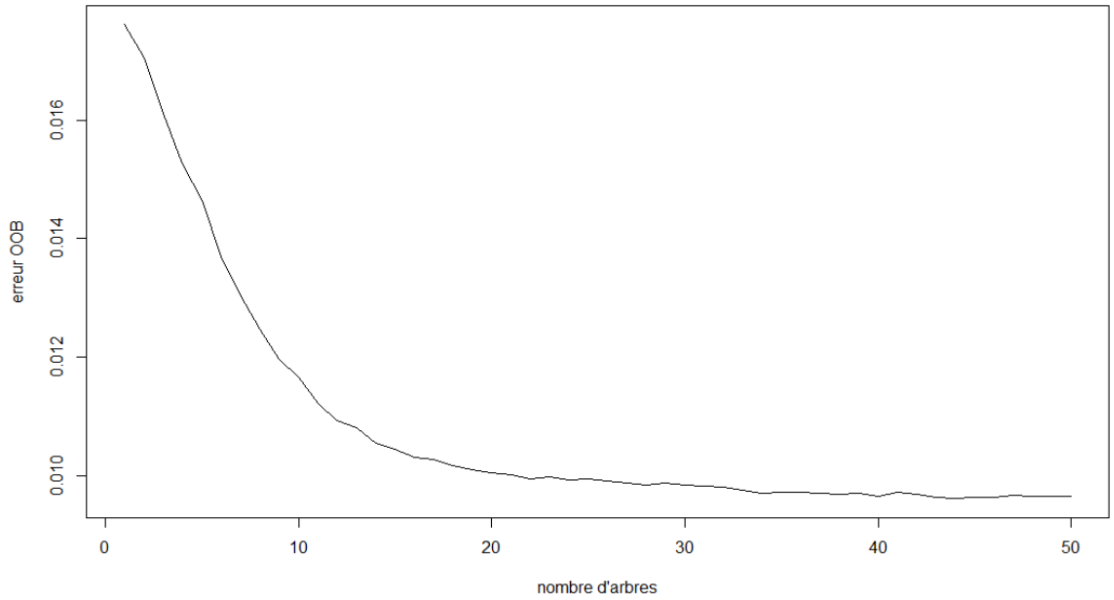


FIGURE 38 – Evolution de l'erreur OOB en fonction du nombre d'arbres dans le modèle.

La qualité du modèle ne s'améliore plus à partir d'une quarantaine d'arbres. Nous fixons donc le nombre d'arbres maximal à 40.

Résultats :

Sur les 1505 survenances historiques observées, nous en prédisons 1068, ce qui représente une précision de 70,9%. De plus nous attribuons 6 survenances supplémentaires sur des points de l'historique ne présentant pas de survenance réelle de grêle, ce qui est négligeable (0,55% des prédictions)

De la même manière que lors de notre régression logistique, nous observons dans les cartes suivantes 2 journées de notre historique ayant enregistré une survenance de grêle afin de comparer les traces réelles aux prédictions.

Exemples de survenances réelles (en bleu) et de survenances simulées (en rouge) :

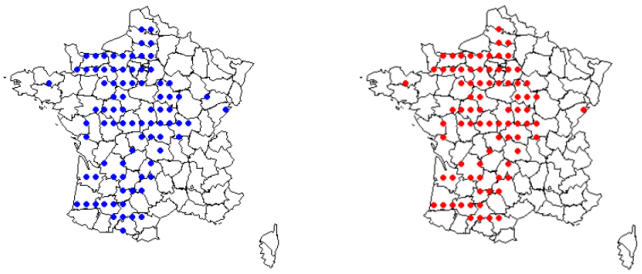


FIGURE 39 – Évènement du 15 Juillet 2003

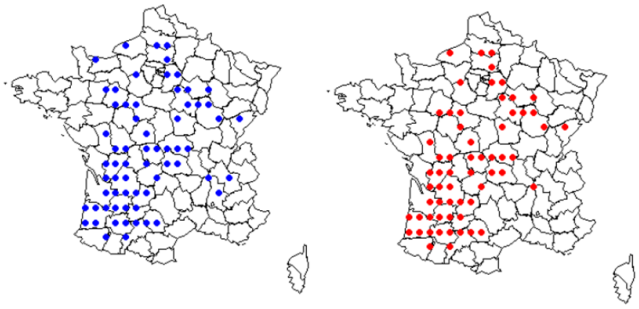


FIGURE 40 – Évènement du 4 Juillet 2006

Conclusion :

La restitution des évènements est très précise, seules quelques zones de survenances de faible intensité ne sont pas correctement prédites. En plus d'une meilleure qualité globale de prédiction, l'avancée principale de cette approche par rapport à l'utilisation d'une régression logistique est la finesse des séparation de variables permettant que la pluie et la grêle ne soient plus confondues. En effet, la grêle n'apparaît pas lorsque les températures sont trop élevées ou la pression trop basse. Dans ces cas, les précipitations observées sont généralement pluvieuses. Le *Random Forest* va donc exclure ces conditions de non survenances afin d'éviter de réaliser de mauvaises prédictions.
L'algorithme semble donc bien capter les conditions climatiques engendrant les chutes de grêle.

Nous retenons donc l'approche par *Random Forest* afin de modéliser la survenance de grêle.

5.3.3 Sélection de variables

Dans la suite du processus nous devrons simuler les variables retenues par notre modèle. Cependant, pour des raisons de complexité, nous ne pouvons pas simuler plusieurs dizaines de variables et tenir compte de l'ensemble des dépendances. Nous décidons alors de ne retenir que les 5 variables jugées les plus significatives par notre modèle (ainsi que l'altitude maximale de la zone qui est une valeur fixe par station).
Comme nous l'avons vu dans la figure 30, dans un *Random Forest*, l'importance d'une variable est déterminée par la baisse d'impureté moyenne qu'elle engendre pour chaque arbre.

Les variables permettant d'obtenir la meilleure séparation des classes sont par ordre d'importance :

— La Température maximale ;

— La variation entre la Température maximale du jour et de la veille ;

— La variation de Température entre 15h et 18h ;

— La Pression ;

— Les Précipitations.

Nous décidons d'observer la qualité de nos résultats à la suite de cette sélection de variables. Avec notre modèle simplifié, composé de nos 5 variables climatiques et de l'altitude.

Sur les 1505 survenances historiques observées, nous en prédisons 1026, ce qui représente une précision de 68,2%. De plus, nous attribuons 6 survenances supplémentaires sur des points de l'historique ne présentant pas de survenance de grêle, ce qui est négligeable (0,58% des prédictions). En diminuant notre nombre de variables explicatives à 6, nous abaissons donc la qualité de notre prédiction de 2,7%, ce qui est un compromis acceptable pour une baisse aussi significative du niveau de complexité du modèle.

Troisième partie

Module aléa

Afin de simuler plusieurs scénarios d'évènements de grêle, nous allons modéliser nos variables explicatives (température maximale, variation de température maximale entre 2 jours, pression moyenne, somme des précipitations, différence de température entre 15h et 18h et altitude maximale de la zone). Nous simulerons ensuite 10 000 scénarios à horizon 1 an et nous prédirons l'ensemble des survenances de grêle sur une année. Nous avons initialement à disposition nos données croisées sur 252 stations entre 1998 et 2018. Pour mener à bien cette modélisation, il nous faut diminuer le coût de la modélisation. Pour cela, il faut réduire la quantité d'informations pour une partie de nos données tout en prenant en compte au maximum l'hétérogénéité spatiale. Nous allons donc regrouper les stations à partir de méthodes d'analyse factorielle et de classification dans le but d'obtenir des zones de risque homogène face au risque de grêle. Dans un premier temps, chacune des zones sera constituée par la température maximale, la pression moyenne et la somme des précipitations quotidiennes relevées sur l'ensemble des stations de la zone.

Nous décidons de ne pas considérer la variation de température maximale et la différence de température entre 15h et 18h dans la constitution de nos zones. Nous faisons ce choix d'une part car la variation de température maximale sera obtenue dans un second temps à partir des températures maximales quotidiennes. D'autre part, la variation de température entre 15h et 18h est issue des relevés SYNOP, cela veut dire que ces relevés sont réalisés à une maille moins fine que les autres variables, issues des relevés ECAD. Son utilisation dans le cadre d'une classification à l'échelle des variables ECAD entrainerait un biais. Ces 2 variables seront donc considérées dans un second temps.
Par la suite, les valeurs historiques associées à chaque zone pour chacune de ces 3 variables seront obtenues en faisant la moyenne de l'ensemble des stations constituant ces zones. A la suite des simulations, les valeurs de chacune des zones seront extrapolées de manière à obtenir des résultats pour chaque station.

6 Classification en zones de risque homogènes

6.1 Analyse en Composantes Principales (ACP)

L'Analyse en Composantes Principales est une méthode multifactorielle d'analyse de données. Elle consiste à réaliser une analyse descriptive sur un tableau de données avec des variables quantitatives afin de déterminer un espace de dimension réduite conservant le plus d'informations possibles tout en expliquant au mieux la variabilité et dont les axes sont des combinaisons linéaires des variables initiales. Ici nous cherchons à regrouper les 252 stations pour former des zones de risque homogène face à la grêle.

Pour des besoins d'optimisation dans les calculs, nous nous ramenons à un tableau de données à deux dimensions où les relevés quotidiens de chaque station sont observés sur la totalité de l'historique. Cette nouvelle matrice de données est notée $X(t,s) = (x_{ijk})\ i = 1,\dots,n;\ j = 1,..,p;\ k = 1,\dots,q$.
Ainsi, comme nous avons 252 points géographiques avec 20 années de 123 jours de données de température maximales, de pression et de précipitations, notre matrice de données transposées contient $s = 2460$ colonnes et $t = 252$ lignes.

$$X = \begin{bmatrix} X_{11} & \dots & X_{1n} & \dots & X_{1p} & \dots & X_{1q} \\ \vdots & & \vdots & & \vdots & & \vdots \\ X_{t1} & \dots & X_{tn} & \dots & X_{tp} & \dots & X_{tq} \end{bmatrix}$$

avec $m = 252, l = 2460, n = 2460 \times 3 = 7380$.

Une ligne x_i représente alors l'évolution temporelle d'une station et une colonne x_j représente les observations de températures maximales ($1 \leq j \leq n$) ou de précipitations ($(n+1) \leq j \leq p$) ou de pression moyenne ($(n+p+1) \leq j \leq q$) sur une journée pour toutes les stations. Chaque ligne de la matrice représente une évolution temporelle de la localisation x_j et chaque colonne i de la matrice représente une vision de l'ensemble des stations pour la journée t_i.

Les données doivent être centrées à l'aide du vecteur des moyennes en chaque localisation $\bar{X} = (\bar{X_1}, ..., \bar{X_q})$ avec $\bar{X_j} = \frac{1}{n} \sum X_{ijk}$ et sont ensuite réduites :

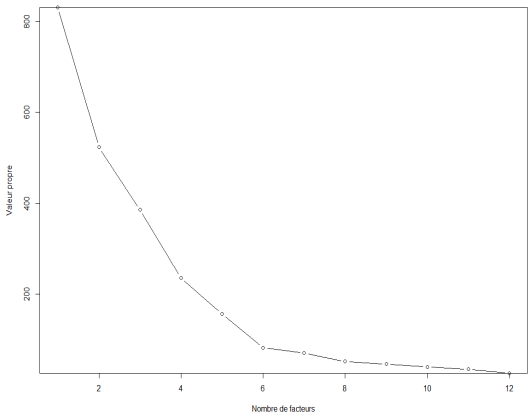
$$X = \begin{bmatrix} \frac{X_{11} - \bar{X_1}}{\sigma(X_1)} & \dots & \frac{X_{1q} - \bar{X_q}}{\sigma(X_q)} \\ \vdots & & \vdots \\ \frac{X_{t1} - \bar{X_1}}{\sigma(X_1)} & \dots & \frac{X_{tq} - \bar{X_q}}{\sigma(X_q)} \end{bmatrix}$$

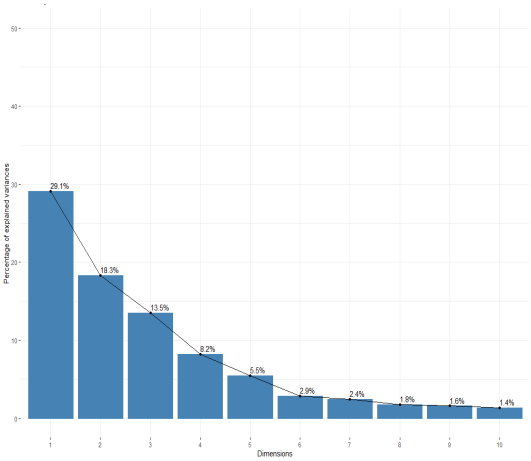
Dans le cadre de l'ACP, le but est de projeter le nuage de points sur un espace de plus faible dimension en minimisant les déformations des distances inhérentes à la projection. Le sous-espace recherché est tel que la moyenne des carrés des distances entre points projetés soit maximale.

Un premier axe u_1, est une combinaison linéaire des X_n, tel que la variance du nuage autour de cet axe soit maximale. Puis, on continue la recherche des axes de projections sur le même principe en imposant que le nouvel axe soit orthogonal aux précédents. Pour trouver u_1, il s'agit ainsi de résoudre le problème d'optimisation $\max\left(u_1^T \Sigma u_1\right)$, avec $u_1^T u_1 = 1$ et Σ une matrice symétrique définie positive. Il s'agit d'un problème de maximisation d'une forme quadratique et la solution est le vecteur propre associé à la plus grande valeur propre de la matrice Sigma. La matrice Sigma usuellement utilisée pour une ACP normée est la matrice de covariance entre les variables initiales. Les vecteurs propres de la matrice Sigma sont appelés composantes principales.

Nous obtenons une matrice de données contenant les axes principaux en colonnes, orthogonaux deux à deux. Ils sont obtenus par combinaison linéaire des variables initiales. Chaque axe contient plus d'information que le suivant. A partir d'un certain axe, l'information contenue devient négligeable. Nous focalisons donc notre attention sur les axes principaux contenant le plus d'informations.

La figure suivante représente les valeurs propres et les pourcentages d'inerties obtenues en réalisant l'Analyse en Composantes Principales sur nos données. Une grande partie de l'information est contenue dans les 5 ou 6 premiers axes.





Sur notre premier graphique, la règle du coude revient à choisir les 5 ou 6 premières valeurs propres. Ensuite, à l'aide du second graphique, nous constatons que l'inertie expliquée par les 5 premiers axes principaux est de 75 %.

La figure suivante montre, pour le premier axe, la mise en opposition des individus qui sont les mieux représentés. Cela nous permet d'observer les zones s'opposant grâce à cet axe et ainsi de déterminer rapidement quel type d'information apporte le premier axe.

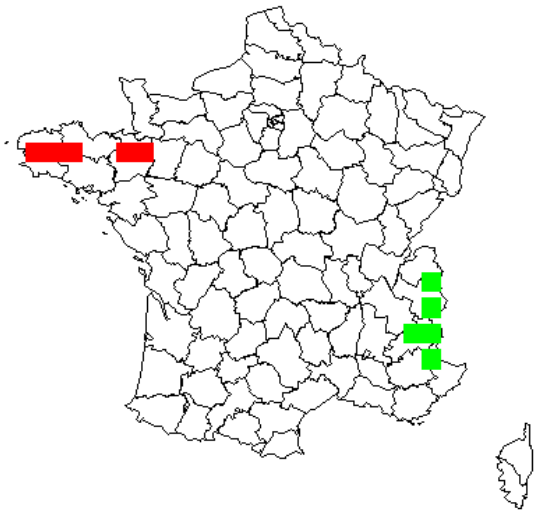


FIGURE 41 – Zones les mieux représentées sur le premier axe principal

Nous remarquons que cet axe permet de différencier le climat de montagne caractérisé par les Alpes (en vert) de la zone bretonne, particulièrement peu touchée historiquement par les grêles d'été.

6.2 La Classification Ascendante Hiérarchique (CAH)

La Classification Ascendante Hiérarchique est une méthode visant à séparer un ensemble d'individus en sous-groupes en fonction de la distance mesurée entre chacun d'eux. La méthode suppose que

chaque groupe peut être différencié à l'aide d'une métrique.

La Classification Ascendante Hiérarchique part du cas où il existe autant de classes que d'individus. Ensuite, à chaque itération, les individus présentant les plus fortes similarités sont regroupés jusqu'à ce qu'il n'y ait plus qu'une seule entité regroupant tous les individus. A chaque étape, toutes les combinaisons possibles de deux classes sont effectuées. La variance intraclasse est calculée pour le nouveau groupe formé et la combinaison présentant la variance intraclasse minimale est choisie. Le regroupement des objets dépend donc de la mesure de similarité/dissimilarité utilisée.

La Classification Ascendante Hiérarchique permet donc, grâce à un critère de similarité choisi préalablement, de regrouper les individus afin qu'au sein d'une même classe ils soient le plus similaires possible et inversement entre les classes.

6.3 Notion de distance

Une distance est une application de E × E dans $\mathbb{R}$ telle que, $\forall i, j, k \in E$:

- $d(i, j) = d(j, i)$;
- $d(i, j) \geq 0$;
- $d(i, j) = 0 \iff i = j$;
- $d(i, j) \leq d(i, k) + d(k, j)$.

On parle de dissimilarité lorsque :

- $d(i, j) = d(j, i)$;
- $d(i, j) \geq 0$;
- $d(i, i) = 0$.

La similarité est une application s de $E \times E$ dans $\mathbb{R}$ telle que :

- $s(i, j) = s(j, i)$;
- $s(i, j) \geq 0$;
- $s(i, i) \geq s(i, j)$.

6.3.1 Distance euclidienne

De nombreuses distances existent, cependant, lors de notre classification nous nous concentrerons sur la distance euclidienne car très utilisée et très simple à appliquer. Soient deux points $x = (x_1, \ldots, x_n)$ et $y = (y_1, \ldots, y_n)$. La distance euclidienne est définie par :

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Une classification hiérarchique nécessite de définir, en plus de la distance entre les individus, une stratégie d'agrégation pour mesurer la distance entre les groupes. La méthode la plus couramment utilisée est la distance de Ward, qui consiste à regrouper les classes minimisant l'inertie intra classe.

6.4 Méthode de Ward

La méthode de Ward regroupe les individus faisant le moins varier l'inertie intra-classe ou maximise l'inertie inter-classe. L'indice de dissimilarité entre deux classes correspond à la perte d'inertie inter-classe résultant de leur regroupement.

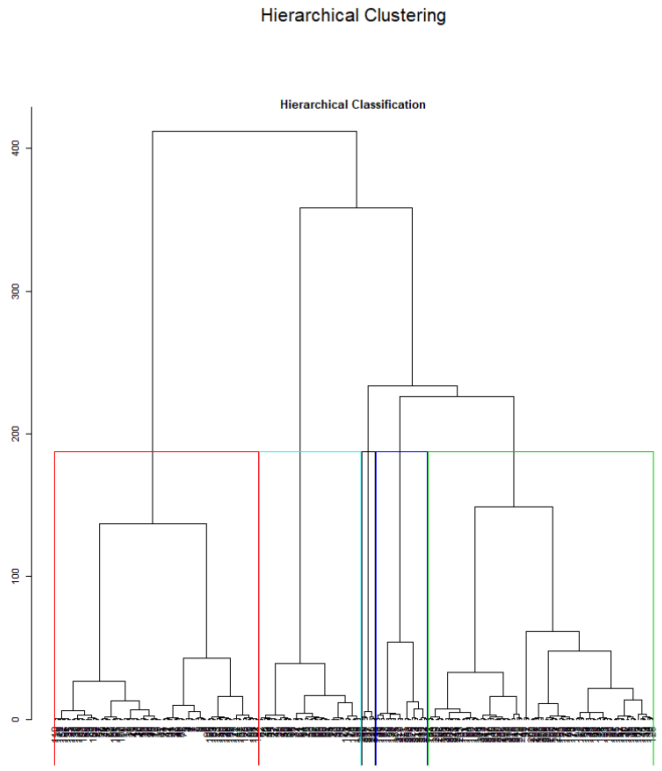
Nous considérons maintenant deux classes A et B, g_A et g_B leurs centres de gravité respectifs. Le p_A et p_B leurs poids respectifs, le centre de gravité résultant de la réunion de ces deux classes, noté g_{AB}, s'obtient par la formule suivante :

$$g_{AB} = \frac{p_A g_A + p_B g_B}{p_A + p_B}$$

L'inertie interclasse est la moyenne des carrés des distances des centres de classe au centre de gravité total. La variation d'inertie entre les classes A et B est donnée par :

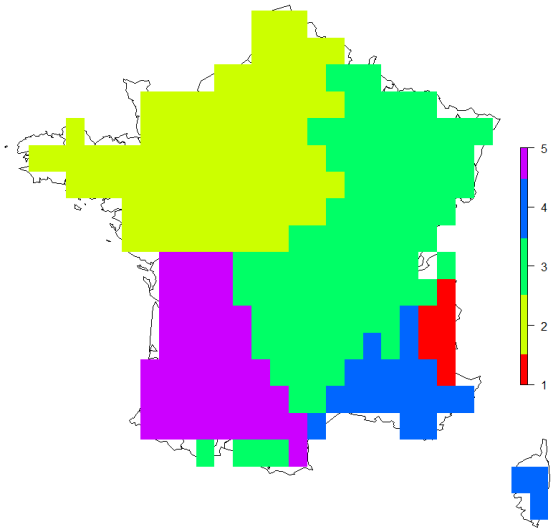
$$\delta(A,B) = p_A d^2(g_A, g) + p_B d^2(g_B, g) - (p_A + p_B) d^2(g_{AB}, g)$$

Où g est le centre de gravité du nuage de points de E.
Nous réalisons une classification sur les résultats de l'ACP, ce qui nous donne le dendrogramme suivant :



En observant le gain d'inertie, le partitionnement optimal est de 3 classes. Cependant, 3 classes ne sont pas suffisantes pour différencier les zones de survenances de grêle. D'une part, Freddy Vinet indique régulièrement dans son étude l'importance des zones côtières et des reliefs afin de distinguer les zones sensibles à la survenance de grêle. La façade atlantique est par exemple beaucoup moins sujette aux

grêles de forte intensité que les zones continentales et les reliefs jouent un rôle majeur dans l'apparition de micro climats multiples. D'autre part, nous avons vu précédemment que le critère du coude nous indiquait que le nombre de classes optimales se trouvait entre 5 et 6. Pour simplifier l'utilisation des calculs dans la suite du module aléa, nous choisissons de garder 5 classes homogènes.



La figure précédente, nous montre les 5 zones globales conservées.
La première zone en rouge caractérise le territoire d'influence des reliefs des Alpes.
La deuxième zone, en jaune, englobe le Nord-Ouest de la France. Elle est caractérisée par l'influence des océans, ce qui entraine des étés cléments et des chutes de grêle rares.
La troisième zone, caractérise le grand Est, une zone au climat continentale propice aux fortes chaleurs et aux chutes de grêle d'été.
La quatrième zone, caractérise le climat méditerranéen du Sud Est et de la Corse.
Enfin, la cinquième zone caractérise le Sud-Ouest, où un climat océanique prédomine mais qui se distingue de la deuxième zone par des étés plus chauds et par une proximité avec les reliefs pyrénéens.
Nous considérons dans la suite de ce module que chaque zone se voit associée la moyenne des températures maximales, pressions moyennes et des précipitations des stations qu'elle contient.

7 Modélisation des variables climatiques à partir des copules

Avant de simuler nos variables explicatives, nous allons modéliser la dépendance des variables entre elles et entre les régions. En effet, les phénomènes climatiques ne sont pas statiques et il y a donc de fortes chances qu'il existe une dépendance entre les zones globales. Afin de modéliser cette dépendance, nous utilisons la théorie des copules qui permet d'introduire des structures de dépendance complexes.

7.1 Les copules

La complexité croissante des produits d'assurance et l'obligation de couverture d'événements ont récemment mis en valeur l'importance de la prise en compte de structures de dépendances complexes entre les variables aléatoires.

Pendant très longtemps, la loi normale multivariée fut le seul outil utilisé pour rendre compte d'une certaine dépendance. Cette méthode a ensuite été vivement critiquée car modélisant mal la dépendance dans les queues de distribution.
Un outil statistique plus adapté a été introduit pour caractériser cette dépendance : **les copules**.

Les copules [6] [7] ont été initialement étudiées en mathématiques par Sklar en 1959 puis en statistiques dans l'analyse des données bivariées et multivariées. La première application en actuariat a été faite au début des années 1990 notamment en assurance de biens. L'utilisation des copules en Finance s'est généralisée dans les années 2000 et notamment après la crise de 2008.

Par la suite, nous allons étudier les propriétés ainsi que les différentes structures de copules afin de sélectionner la copule qui permet de modéliser au mieux la dépendance entre les différentes zones.

Définition :
Une copule est une fonction de répartition multivariée.
$C : [0,1]^d \longrightarrow [0,1]$ avec des marginales uniformes, c'est à dire $C(1,...,1,u_i,1,...,1) = u_i, \forall i \in 1,...,d$

Théorème de Sklar :
Soit F une fonction de répartition multivariée du vecteur $X = (X_1,...,X_d)$ de lois marginales $(F_1,...,F_d)$. Alors il existe une copule C telle que $\forall x_1,...,x_d : F(X_1,...,X_d) = C(F_1(x_1),...,F_d(x_d))$ Si, en plus, les lois marginales $F_1,...,F_d$ sont continues, alors C est unique.

Ce théorème est fondamental car il permet de modéliser la fonction de répartition multivariée d'un vecteur de variables aléatoires et de modéliser ainsi la dépendance en deux étapes : modéliser d'abord les lois marginales, puis définir la structure de dépendance pour les lier.
Dans la pratique, on utilise des copules paramétriques pour approcher la copule C dans l'optique de simplifier le problème. En effet, les copules paramétriques permettent de décrire de nombreuses formes de dépendance et sont suffisantes dans de nombreux cas.

Il existe de multiples structures de copules mais les copules paramétriques présentent l'avantage d'être définies par un nombre fini de paramètres. Il suffit donc de choisir la copule paramétrique la mieux adaptée puis de calibrer les paramètres associés pour la définir.

Au sein des copules paramétriques, on observe deux familles : les Elliptiques et les Archimédiennes.

Les copules elliptiques :

Une copule est elliptique lorsqu'elle est générée par un vecteur ayant une distribution elliptique. La famille des copules elliptiques comprend notamment les copules gaussiennes et de student parmi les plus fréquemment utilisées.

La copule gaussienne :

La copule gaussienne est la copule sous-jacente à la loi normale multivariée. La copule gaussienne ne présente pas de dépendance de queue et n'est donc pas adaptée afin de capter une structure de dépendance des valeurs extrêmes.

La copule gaussienne peut être modélisée avec la matrice de corrélation linéaire.

Soit θ le coefficient de corrélation linéaire et ϕ^{-1} l'inverse d'une fonction de répartition gaussienne centrée réduite, elle est définie par :

$$C(u,v) = \frac{1}{2\pi\sqrt{1-\theta^2}} \int_{-\infty}^{\phi^{-1}(u)} \int_{-\infty}^{\phi^{-1}(v)} \exp\left(\frac{-(x^2-2\theta xy+y^2)}{2(1-\theta^2)}\right) \mathrm{d}x\mathrm{d}y$$

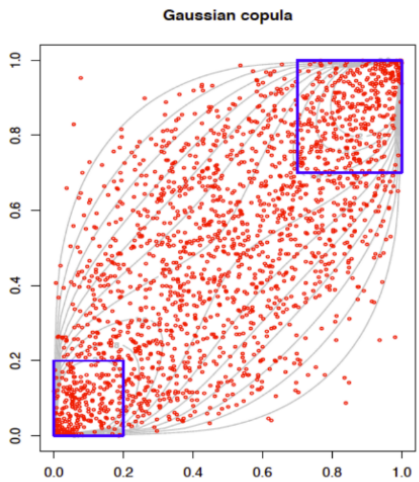


FIGURE 42 – Exemple de copule gaussienne

La copule de Student :

La copule de Student est la copule sous-jacente à une distribution multivariée de Student. Contrairement à la copule Gaussienne, elle capte les dépendances extrêmes de toutes sortes. Elle est construite de la même manière que la copule gaussienne mais à partir de la distribution de Student centrée réduite. La copule de Student tend vers la copule normale quand $\nu \longrightarrow \infty$.

$$C(u,v) = \frac{1}{2\pi\sqrt{1-\theta^2}} \int_{-\infty}^{t_\nu^{-1}(u)} \int_{-\infty}^{t_\nu^{-1}(v)} (1+\frac{(x^2-2\theta xy+y^2)}{2(1-\theta^2)})^{-((\nu+2)/2)} \mathrm{d}x\mathrm{d}y$$

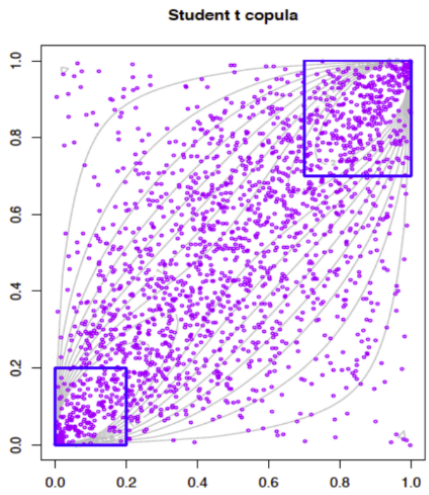


FIGURE 43 – Exemple de copule de Student

Les copules Archimédiennes :

Les copules Archimédiennes constituent une famille importante de copules. On définit une copule Archimédienne par le biais d'un générateur ϕ :

$$C(u_1, u_2) = \phi^{-1}\left\{\phi(u_1) + \phi(u_2)\right\}, \forall u_1, u_2$$

Le générateur ϕ doit posséder les propriétés suivantes :

- Application de la forme : $\phi : [0,1] \longrightarrow [0, +\infty)$, continue, strictement décroissante et convexe ;
- $\phi(1) = 0$;
- $lim(\phi_{(t\to 0)}(t)) = +\infty$.

Les copules Archimédiennes les plus communément utilisées sont :

- **La copule de Clayton :**

  Fonction génératrice $\phi(t) = \frac{t^{-\theta}-1}{\theta}$ avec en dimension 2,

$$C_\theta(u, v) = (u^{-\theta} + v^{-\theta} - 1)^{-1/\theta}$$

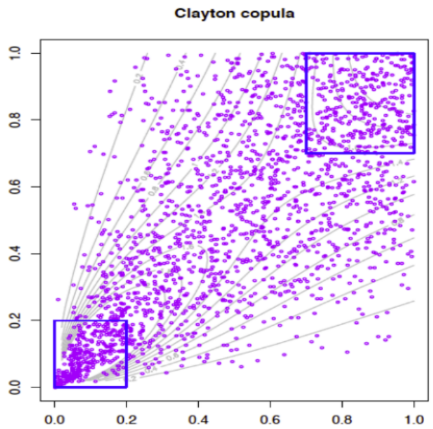


FIGURE 44 – Exemple de copule de Clayton

— **La copule de Frank :**

Fonction génératrice $\phi(t) = ln\left(\frac{e^{-\theta}-1}{e^{-\theta}t-1}\right), \theta \neq 0$ avec en dimension 2,

$$C_\theta(u,v) = -\frac{1}{\theta} ln\left(1 + \frac{(e^{-\theta u}-1)(e^{-\theta v}-1)}{e^{-\theta}-1}\right)$$

— **La copule de Gumbel :**

Fonction génératrice $\phi(t) = (-ln(t))^\theta, \theta \geq 1$ avec en dimension 2,

$$C_\theta(u,v) = exp\left(-((-ln(u))^\theta + (-ln(v)^\theta)^{1/\theta})\right)$$

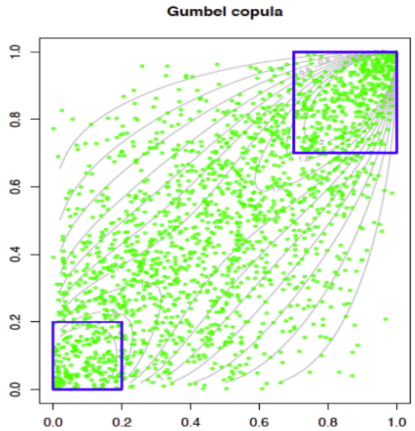


FIGURE 45 – Exemple de copule de Gumbel

Méthode d'estimation par Maximum de vraissemblance

Sous des conditions de continuité, la densité jointe de la distribution F s'écrit :

$$f(x,y) = C(F_1(x), F_2(y)) f_1(x) f_2(y)$$

Soit $(x_j, y_j)_{\{0 \leq j \leq N\}}$ l'échantillon d'observations, $\theta = (\alpha, \theta_1, \theta_2)$ le vecteur des paramètres à estimer et Θ l'espace dans lequel θ prend ses valeurs. Alors la log-vraisemblance s'exprime alors :

$$ln(L(\alpha, \theta_1, \theta_2, x_1, ..., x_N, y_1, ..., y_N) = \sum_{i=1}^{n} ln(c(F_1(x_i, \theta_1), F_2(x_i, \theta_2), \alpha)) + \sum_{i=1}^{n} (ln(f_1(x_i, \theta_1)) + ln(f_2(x_i, \theta_2), \alpha)))$$

où θ est le vecteur des paramètres de la copule et des marginales. Ainsi, en spécifiant les lois marginales et le type de copule, l'estimateur du maximum de vraisemblance de θ est :

$$\hat{\theta} = argmax_{\theta \in \Theta}(ln(L(\theta))$$

Copule empirique et Goodness of fit

Le "Goodness of fit test" est un critère quantitatif pour choisir la meilleure copule en se basant sur la distance entre notre estimateur paramétrique et la copule empirique .
Soit $C_\Theta = \{C_\theta, \theta \in \Theta\}$ une classe paramétrique de copules et GF_n le processus empirique :

$$GF_n(u) = \sqrt{n}(C_n(u) - C_{\theta_n}(u) \text{ , pour } u \in [0,1]^d$$

où :

— C_n est la copule empirique calculée à partir des pseudos observations.

$$C_n(u) = \frac{1}{n} \sum_{i=1}^{n} 1_{\{\hat{U}_i \leq u\}} \text{ pour } u \in [0,1]^d$$

où $\hat{U}_i = (\hat{U}_{i1}, ..., \hat{U}_{ip}), i \in 1, ..., n$, sont les pseudo-observations, c'est à dire les observations $X_i = (X_{i1}, ..., X_{ip})$ transformées en uniformes : $\hat{U}_{ij} = \frac{R_{ij}}{n+1}$ avec R_{ij} le rang de X_{ij} parmi $(X_{1j}, ..., (X_{nj}), j \in \{1, ..., d\}$

— $C_{\theta_n}(u)$ est un estimateur de C sous l'hypothèse $H_0 : C \in C_\Theta$

Soit S_n la statistique de test

$$S_n = \int_{[0,1]^d} GF_n(u)^2 dC_n(u) = \sum_{i=1}^{n} (C_n(\hat{U}_i) - C_{\theta_n}(\hat{U}_i))^2$$

Nous pouvons obtenir une p-value approximative de test pour la statistique S_n grâce à la procédure de bootstrap suivante :

— On calcule C_n et estime θ_n à partir des pseudo-observations $\hat{U}_1, ..., \hat{U}_n$;
— On calcule la statistique $S_n = \sum_{i=1}^{n} (C_n(\hat{U}_i) - C_{\theta_n}(\hat{U}_i))^2$;
— Pour N suffisamment grand, on répète les étapes suivantes pour tout $k \in \{1, ..., N\}$:
  — On génère un échantillon aléatoire $X_1^{(k)}, ..., X_n^{(k)}$ à partir de la copule C_{θ_n} et on calcule les pseudo-observations associées $\hat{U}_1^{(k)}, ..., \hat{U}_n^{(k)}$
  — Soit

$$C_n^{(k)}(u) = \frac{1}{n} \sum_{i=1}^{n} 1_{\{\hat{U}_i^{(k)} \leq u\}} \text{ pour } u \in [0,1]^d$$

  — On calcule une réalisation indépendante de S_n sous H_0 par :

$$S_n^{(k)}(u) = \sum_{i=1}^{n} (C_n^{(k)}(\hat{U}_i^{(k)}) - C_{\theta_n}^{(k)}(\hat{U}_i^{(k)}))^2$$

— Une p-value approximative $\hat{\alpha}$ pour le test est donnée par $\frac{1}{N}\sum_{i=1}^{n} 1_{\{S_i^{(k)} \geq S_i\}}$, On rejette l'hypothèse d'adéquation de la copule paramétrique si $\hat{\alpha} \geq 0{,}05$.

Parmi les copules acceptées, on sélectionne celle qui maximise la p-value.

Fonction de Kendall :

La fonction de Kendall[6] est définie comme suit :

$$K(t) = \mathbb{P}(C(U,V) \leq t) = \mathbb{P}(Z \leq t)$$

Soient $U = (U_1, ..., U_n)$ et $V = (V_1, ..., V_n)$ un couple de variables aléatoires. La fonction de Kendall peut alors être estimée de la manière suivante :

$$\hat{K}(t) = \frac{1}{n}\sum_{i=1}^{n} 1(Z_i \leq t)$$

$$\text{où } Z_i = \frac{1}{n-1}\sum_{i \neq j} 1(U_i \leq U_j, V_i \leq V_j)$$

La fonction de Kendall est un critère graphique. Les fonctions de Kendall empiriques et théoriques sont tracées sur un même graphique. Plus les deux courbes sont proches, plus la copule sera adaptée pour modéliser la dépendance.

7.2 Modélisation des températures maximales, des pressions moyennes et des précipitations

Les figures suivantes représentent respectivement les nuages de points des pseudo-observations des différentes zones pour les températures maximales, les pressions moyennes et les précipitations.

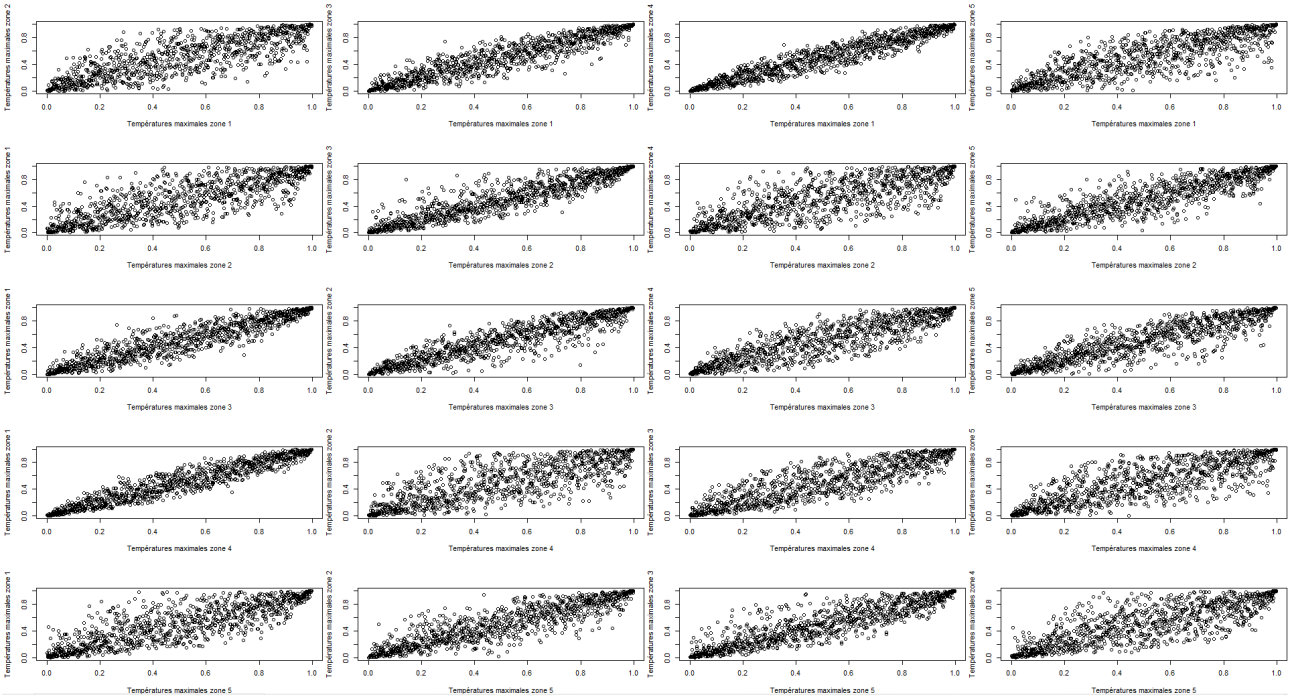


FIGURE 46 – Dépendances entre les températures maximales quotidiennes

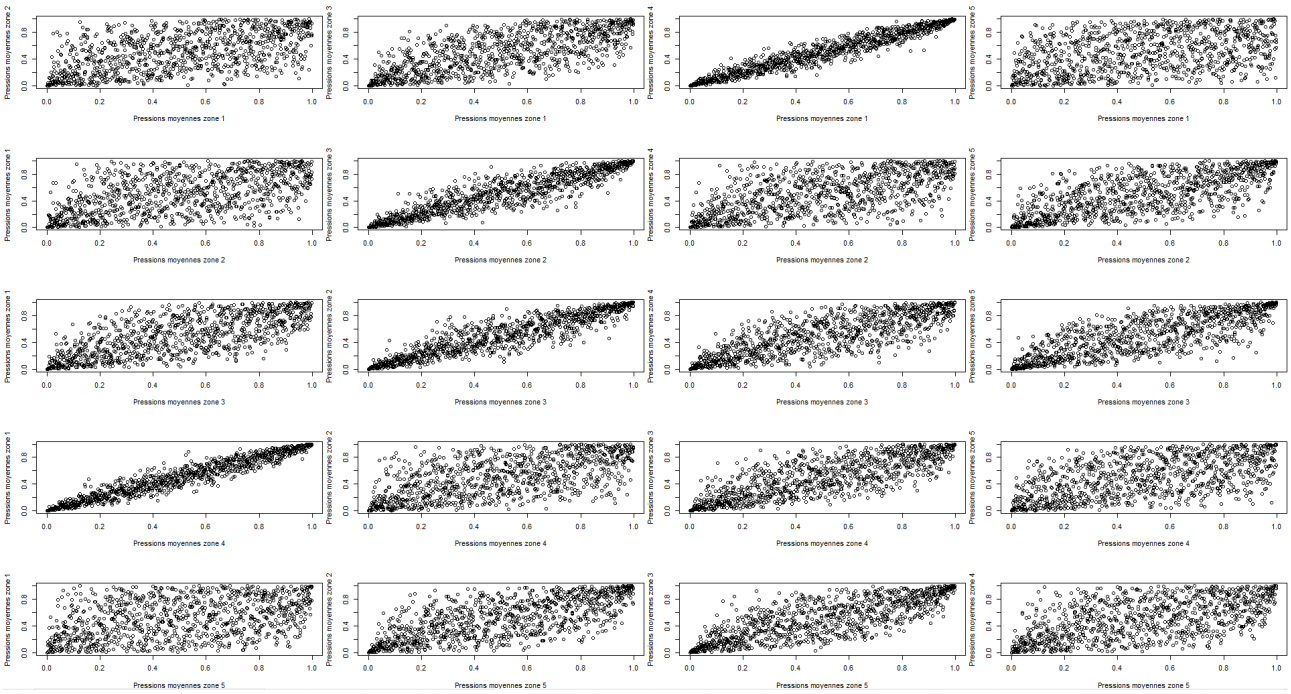


FIGURE 47 – Dépendances entre les pressions moyennes quotidiennes

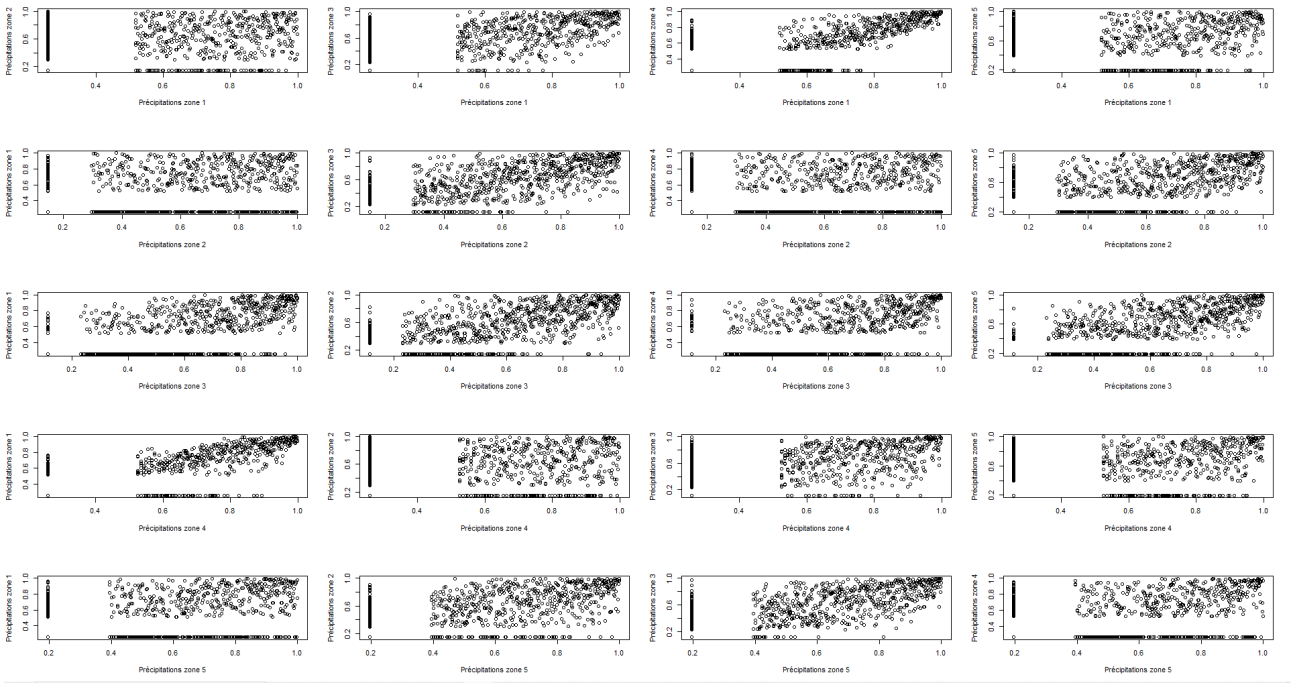


FIGURE 48 – Dépendances entre les précipitations

Nous remarquons que les copules les plus adaptées pour étudier la dépendance entre les zones globales sont les copules elliptiques du fait de la symétrie des nuages de points.

Par la suite, nous caractérisons la dépendance entre les pressions moyennes, les précipitations et les régions en estimant les paramètres des copules gaussiennes et de Student par maximum de vraisemblance.

Pour des raisons de complexité, les précipitations sont modélisées indépendamment des 2 autres variables globales. Les résultats de l'étude de la dépendance entre les précipitations et les zones globales sont fournis en annexe.

Les résultats pour les copules Gaussiennes et de Student sont donnés ci-dessous.

	Températures maximales -zone 1	Températures maximales -zone 2	Températures maximales -zone 3	Températures maximales -zone 4	Températures maximales -zone 5	Pressions moyennes - zone 1	Pressions moyennes - zone 2	Pressions moyennes - zone 3	Pressions moyennes - zone 4	Pressions moyennes - zone 5
Températures maximales -zone 1	1	0,82	0,92	0,83	0,05	0,79	0,22	0,04	0,13	0,73
Températures maximales -zone 2		1	0,95	0,23	0,17	0,89	- 0,13	0,22	- 0,03	0,96
Températures maximales -zone 3			1	0,05	- 0,11	0,30	0,88	- 0,19	0,31	0,43
Températures maximales -zone 4				1	0,92	0,14	0,89	0,82	0,19	0,92
Températures maximales -zone 5					1	0,12	0,32	0,17	0,17	0,64
Pressions moyennes - zone 1						1	0,04	0,09	0,24	0,73
Pressions moyennes - zone 2							1	0,08	- 0,14	0,82
Pressions moyennes - zone 3								1	0,57	0,82
Pressions moyennes - zone 4									1	0,61
Pressions moyennes - zone 5										1

FIGURE 49 – Paramètres de la copule Gaussienne

Degrés de liberté : 34

	Températures maximales -zone 1	Températures maximales -zone 2	Températures maximales -zone 3	Températures maximales -zone 4	Températures maximales -zone 5	Pressions moyennes - zone 1	Pressions moyennes - zone 2	Pressions moyennes - zone 3	Pressions moyennes - zone 4	Pressions moyennes - zone 5
Températures maximales -zone 1	1	0,82	0,92	0,83	0,05	0,79	0,21	0,04	0,13	0,73
Températures maximales -zone 2		1	0,95	0,23	0,17	0,89	- 0,14	0,22	- 0,03	0,96
Températures maximales -zone 3			1	0,05	- 0,12	0,30	0,88	- 0,20	0,31	0,42
Températures maximales -zone 4				1	0,92	0,14	0,89	0,82	0,20	0,92
Températures maximales -zone 5					1	0,12	0,32	0,17	0,17	0,64
Pressions moyennes - zone 1						1	0,05	0,09	0,23	0,73
Pressions moyennes - zone 2							1	0,07	- 0,15	0,81
Pressions moyennes - zone 3								1	0,57	0,81
Pressions moyennes - zone 4									1	0,60
Pressions moyennes - zone 5										1

FIGURE 50 – Paramètres de la copule de Student

Afin de déterminer la copule la mieux adaptée, nous utilisons 2 critères : la fonction de Kendall et la Goodness of Fit. Les fonctions de Kendall empiriques et théoriques sont affichées pour les 2 copules étudiées dans la figure suivante.

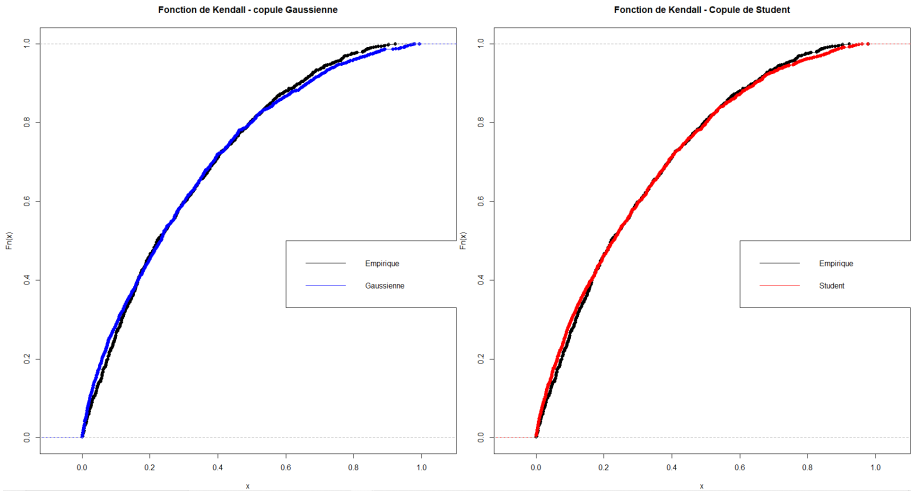


FIGURE 51 – Fonction de Kendall

A première vue, nous ne sommes pas en mesure de faire un choix entre nos 2 copules. Nous sélectionnons ensuite la meilleure copule à l'aide du critère quantitatif de la **"Goodness of fit"**. Les p-values obtenues pour chaque copule sont données dans le tableau suivant :

Copule	**p-value**
Gaussienne	0.81
Student	0.86

Au regard du critère de Goodness of fit, la copule de Student est la meilleure copule pour exprimer la dépendance entre les températures maximales, les pressions et les zones globales.

Pour simuler des réalisations du vecteur aléatoire $X = (X_1, ..., X_n)$ dont la distribution est $F(x_1, ..., x_n) = C(F_1(x_1), ..., F_n(x_n))$, il faut simuler le vecteur aléatoire $U = (U_1, ..., U_n)$ dont la fonction de distribution est la copule C. En effet, si $U \sim [0,1]$ et X est une variable aléatoire de fonction de répartition F, alors $F^{-1}(U)$ est égale en loi à X et $F(X) \sim U[0,1]$. On aura donc $X = (F_1^{-1}(U1), ..., F_n^{-1}(Un))$

Nous effectuons donc 10 000 années de simulations de 123 jours correspondant aux mois de Mai, Juin, Juillet et Août pour les températures maximales, pressions moyennes et précipitations pour chacun des 5 zones globales. Pour chaque simulation, la valeur correspondante est obtenue à partir de la distribution empirique historique inhérente à la série.

7.3 Descente d'échelles

La prédiction de survenance de grêle à partir de données météorologiques implique de disposer de ces données à une maille plus fine que nos 5 régions homogènes actuelles. La survenance étant caractérisée dans notre modèle par station extrapolée, nous souhaitons descendre à ce niveau de granularité afin d'améliorer la précision.

Les méthodes de descente d'échelles sont basées sur l'existence d'une relation entre la grande et la petite échelle. Dans notre étude, nous allons nous focaliser sur l'approche statistique. Cette approche est basée sur la recherche d'un lien statistique entre les variables locales et les variables globales du modèle. Cette méthode est tout d'abord basée sur le calibrage de modèles décrivant la relation entre la variable globale et les variables locales.

Dans notre cas nous disposons de 5 zones globales et de 252 stations extrapolées. Il nous faut donc pour chaque station et chaque variable, un modèle unique permettant d'obtenir une valeur à partir des observations issues de la zone globale.
Nous avons les données quotidiennes de températures maximales (TMax), pression moyenne (PMoy), et somme des précipitations (SP) pour les mois de Mai, Juin, Juillet et Aout de 1999 à 2017 pour chaque station S_s (s appartient à $1, \ldots, 252$). La température maximale pour la station s au jour i se note $\{TMax_i, S_s\}$. A la suite de la classification, ces données ont été agrégées en zone homogène avec celles de toutes les autres stations, qui font partie de la même zone notée Z_t (t appartient à {1, …,5}) en calculant leurs moyennes. Nous obtenons donc des paires entre variables locales et globales pour l'ensemble des journées d'été de l'historique et pour chacune de nos 3 variables explicatives $\{(TMax_i S_s, TMax_i Z_t)\}$. Nous estimons les relations linéaires entre les séries moyennes de chaque zone et les série des stations extrapolées de l'ECAD appartenant à la zone considérée.

Dans les figures suivantes nous observons respectivement la relation entre les températures maximale de la station Loc471 (appartenant à la zone 3) et de la zone 3 ainsi que la relation entre les pressions moyennes de la station Loc471 et de la zone 3.

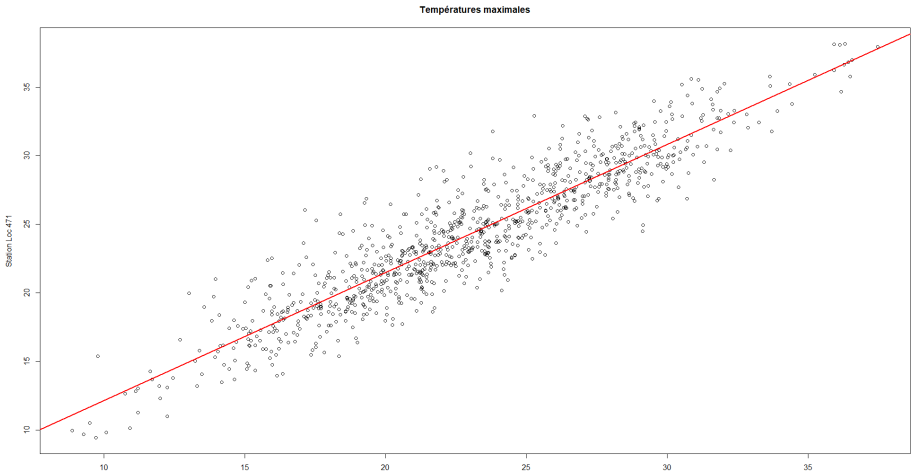


FIGURE 52 – Relation entre les températures maximales de la station Loc471 et de la zone globale 3

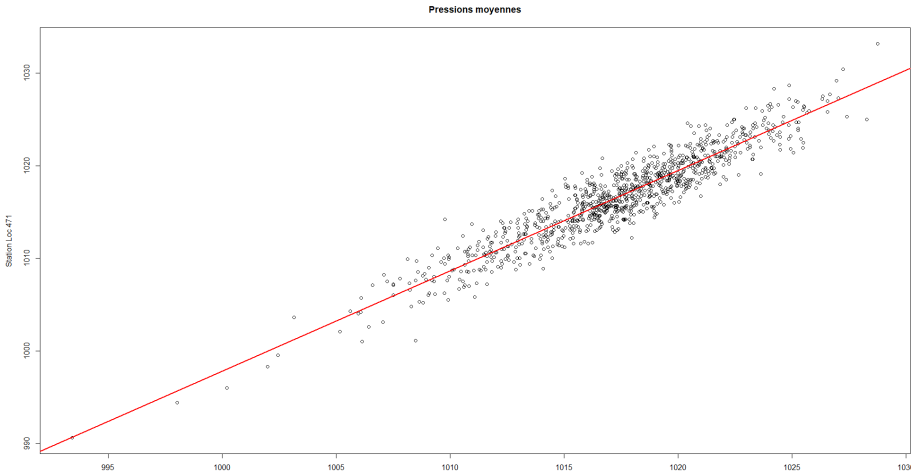


FIGURE 53 – Relation entre les pressions moyennes de la station Loc471 et de la zone globale 3

Nous créons donc 252 × 3 modèles linéaires de ce type entre les zones globales et les stations extrapolées. Ensuite, ces modèles sont appliqués aux simulations effectuées sur les variables globales.
Nous obtenons finalement 123 valeurs quotidiennes de températures maximales, pression moyenne et précipitations pour les 252 stations extrapolées de l'ECAD pour chacune des 10.000 années simulées.

7.4 Constitution de la base finale

Afin de constituer notre base de données finale, nous devons encore modéliser deux variables explicatives :

— La Variation de température maximale entre deux journées consécutives ;
— La variation de température entre 15h et 18h.

La principale difficulté réside dans l'ajout de ces variables par l'utilisation de méthodologies n'altérant pas la structure de dépendance de l'ensemble de variables de la base.

7.4.1 Variation de températures Maximales

La "Variation de températures maximales entre 2 journées" n'a pas été incluse dans la méthodologie de simulation par copule et loi empirique car cette variable est générée directement à partir des "températures maximales".

$$\Delta\text{température Maximale}_i = \text{température Maximale}_i - \text{température Maximale}_{i-1}$$

La problématique lors de l'application de cette méthode est la cohérence temporelle de nos observations. En effet, dans ce mémoire nous considérons que les évènements de grêle sont caractérisés quotidiennement. La cohérence temporelle lors de la simulation de nos variables n'est donc pas sensée avoir un impact sur la qualité de notre modèle. Cependant, en introduisant la variable "Différence de température maximale entre 2 jours", il aurait été préférable d'observer l'évolution des températures maximales au cours du temps.
Dans un premier temps, une approche consistant à modéliser les températures maximales pour chaque station à l'aide de séries temporelles a été implémentée pour tenir compte de cette problématique (voir annexe). Cependant, cette méthode ne permettait pas de modéliser correctement les dépendances entre les stations et elle n'a donc pas été conservée.
L'utilisation d'une approche permettant de tenir compte de la temporalité de cette variable constituerait donc un axe d'amélioration important.

Les Variations de température sont ensuite constituées selon la formule précédemment citée.

7.4.2 Variation de température entre 15h et 18h

La variation de température entre 15h et 18h est simulée par tirage aléatoire conditionnel au sein de l'historique. Pour chaque journée simulée nous établissons des intervalles autour des valeurs observées (0.5 degrés pour la température, 1 hecto Pascal pour la pression et 5 mm pour les précipitations).
Les variations de températures historiques observées conditionnellement à ces paramètres climatiques sont stockées dans une liste. Le delta de pression est ensuite choisi aléatoirement parmi ces valeurs historiques. Cette opération est effectuée pour chaque jour et chaque station.

8 Simulation de la grêle

Suite à la modélisation des variables climatiques, 10 000 années composées uniquement des mois de Mai, Juin, Juillet et Août sont simulées. Nous appliquons le *Random Forest* (précédemment calibré sur l'historique de survenance) sur la base simulée.
Pour chaque journée simulée, l'algorithme active les stations qu'il considère comme répondant aux caractéristiques de l'apparition de la grêle.

Voici quelques exemples d'évènements simulés :

FIGURE 54 – Exemple d'évènement simulé

FIGURE 55 – Exemple d'évènement simulé

Résultats :

Historiquement, les évènements de grêle majeurs ont eu lieu en moyenne 2 fois par an et ces grêles activent en moyenne des zones couvertes par environ 11 stations de l'ECAD.
Le nombre d'évènements maximal observé a été atteint sur l'année 2013 avec 5 évènements et l'évènement le plus étendu a activé des zones couvertes par 54 stations de l'ECAD.

Sur nos 10 000 années de simulations, les évènements de grêle majeurs ont lieu en moyenne tous les

ans. Cependant, l'évènement simulé le plus étendu a pour sa part touché 4 stations, soit 13,5 fois moins que l'évènement historique le plus important.

Les simulations prédisent donc des grêles dont les étendues spatiales sont très largement inférieures aux observations historiques.

Ces différences peuvent s'expliquer par la perte de qualité engendrée par un Random Forest qui sous-estime les survenances (30% de moins) mais également par certaines hypothèses prises en compte comme l'indépendance des précipitations ou par le regroupement en zones homogènes puis l'utilisation de descentes d'échelles, une méthode basée sur des moyennes pouvant potentiellement induire des circonstances extrêmes plus rares et donc des empreintes peu étendues.

Quatrième partie

Modules Vulnérabilité et Financier

Le module Aléa nous a permis de générer nos simulations de survenance de grêle pour l'année 2019 sur chacune de nos 252 stations. Nous voulons à présent associer une perte aux différents biens assurés lors de la survenance de grêle.

Le module Vulnérabilité a pour objectif de modéliser l'impact des sinistres simulés sur le portefeuille de l'entreprise. Pour cela, ce module doit associer l'intensité de l'évènement simulé au nombre de biens touchés et aux taux de destruction pour un évènement de grêle donné.

De la même manière que dans le module Aléa, nous considérerons le risque à l'échelle des stations extrapolées de l'ECAD dans le module vulnérabilité en leur associant la strucutre du portefeuille du département dans lequel les stations sont situées.

Nous nous baserons sur le même procédé que lors de la prédiction de la survenance, en utilisant un *Random Forest* afin d'associer nos données climatiques à une classe d'intensité. En prenant notre historique de sinistralité, nous calibrerons ensuite des lois de probabilité conditionnées à l'intensité pour caractériser la fréquence de sinistralité et les taux de destruction.

La charge associée à un évènement sera donc modélisée à partir de 2 paramètres :
- Le nombre de sinistres ;
- Les taux de destruction.

Chacune de ces 2 variables sera étudiée pour chaque classe de bien et chaque intensité. Pour chaque survenance simulée, chaque classe aura un nombre de sinistres et un taux de destruction associés.

Nous commencerons par étudier les caractéristiques des biens assurés (propriétaire, locataire, maison, appartement) pour capter ce qui pourrait permettre d'affiner la modélisation du risque. Nous verrons que les taux de destructions sont fortement liés à ces caractéristiques lors de la survenance d'un évènement de grêle. En effet, la grêle ne dégradant pas tous les types de biens de la même manière, l'intensité associée à un évènement voit donc son impact nuancé par des caractéristiques de ces biens ainsi que de leur répartition sur la zone de survenance.

Nous considérons donc dans notre modèle que les paramètres responsables des montants de pertes sont l'intensité des chutes de grêle et les caractéristiques des biens assurés.

9 Caractéristiques des bien assurés

En habitation, certaines caractéristiques permettent de différencier efficacement les biens vis-à-vis du risque de grêle. En effet, un assuré locataire n'a à sa charge que le contenu de son habitation. Or, le contenu est rarement touché lors d'un épisode de grêle. Par opposition, un propriétaire doit prendre en charge le contenu et le bâti. De la même manière, un habitant du dernier étage est bien plus exposé aux chutes de grêle qu'un assuré habitant dans les étages intermédiaires.

Les caractéristiques sélectionnées sont les suivantes :

— **Le type d'assuré** : Propriétaire, Locataire et Propriétaire Non Occupant.
Comme évoqué précédemment, le locataire n'assure que le contenu de son bien tandis que le propriétaire doit prendre en compte le bâti et le contenu. Le propriétaire non occupant qui, comme son nom l'indique, n'habite pas son bien, n'a donc pas de contenu à charge. Il peut se différencier du propriétaire par les délais avant déclaration de sinistre ou par sa charge généralement beaucoup moins élevée car correspondant à un bien de moindre valeur.

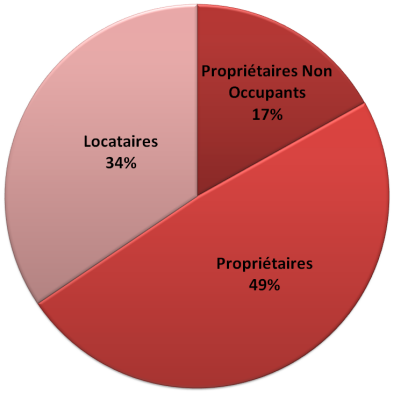


FIGURE 56 – Répartition entre Propriétaires et Locataires sur le portefeuille national de Pacifica

— **Le type de logement** : Appartements et Maisons.
La charge varie beaucoup entre ces 2 types de biens. Dans un premier temps, les maisons se distinguent par des sommes assurées plus importantes du fait de plusieurs facteurs comme la superficie (souvent bien plus importante), ou la présence d'une dépendance, ce qui est relativement courant pour les maisons. Dans un second temps, les maisons sont beaucoup plus vulnérables à la grêle à cause des toitures (taule, gouttières . . .) qui peuvent être beaucoup plus fragiles que pour un immeuble.

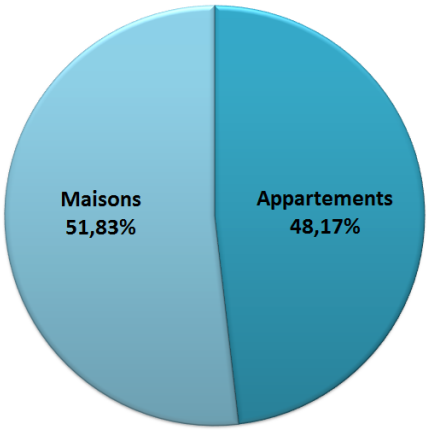


FIGURE 57 – Répartition entre maisons et appartements sur le portefeuille national de Pacifica

De nombreuses variables peuvent être pertinentes afin d'améliorer la qualité du modèle. Le type de toiture ou l'étage du logement (dans le cas d'un apartement) sont par exemple des paramètres majeurs pouvant jouer fortement sur les dégâts associés aux biens. Cependant, ces variables ne sont pas assez souvent renseignées ou le sont sur un historique insuffisant pour que nous en tenions compte dans le modèle.

Les 4 classes d'études distinctes sont décrites dans le tableau suivant :

Type de logement	Type d'assuré	Eléments assurés
Maison	Propriétaire/Propriétaire non occupant	Bâti et Contenu
Maison	Locataire	Contenu
Appartement	Propriétaire/Propriétaire non occupant	Bâti et Contenu
Appartement	Locataire	Contenu

Les biens du portefeuille automobile 4 roues n'ayant pas de distinction particulière, on se concentre à présent sur le portefeuille habitation de Pacifica.

10 La fréquence de sinistralité : Une mesure fiable de l'intensité

Lors d'un épisode de grêle, sur une zone définie (Région, Département, Code postal ...), un certains nombre de biens est touché. Ce nombre de sinistres est à mettre en perspective de la structure du portefeuille dans la zone considérée. En effet, 1000 sinistres dans une zone où le portefeuille Pacifica comprend 50 000 polices d'assurances ne traduit pas la même intensité des chutes de grêle que 1000 sinistres dans une zone où le portefeuille Pacifica comprend 10 000 polices d'assurances.

Pour avoir un indicateur fiable de **l'intensité d'une chute de grêle**, nous définissons alors **la fréquence de sinistralité** comme le nombre de sinistres rapporté à l'exposition (au nombre de polices d'assurance sur une même zone) :

$$\text{Fréquence de sinistralité} = \frac{\text{Nombre de sinistres sur la zone}}{\text{Nombre de polices sur la zone}}$$

En effet, la fréquence de sinistralité est le marqueur le plus fiable de l'intensité d'un évènement de grêle. La charge importante d'évènements historiques comme la Pentecôte de 2014 s'explique principalement par une quantité importante de biens touchés. Contre intuitivement, les taux de destruction sont plus liés au type de biens impactés qu'à l'intensité des chutes de grêle.

De plus, la grêle n'a pas le même impact en fonction des biens considérés. Les voitures sont beaucoup plus sensibles aux petites grêles que les maisons ou les appartements. Les tuiles de maisons nécessitent une force d'impact plus importante qu'un pare-brise pour subir des dégâts. Il est donc important de distinguer la fréquence de sinistralité des biens automobiles de celle des habitations.

Dans le cas des habitations, l'historique de sinistralité engendré par la grêle pour Pacifica est composé à 95% de maisons. Nous considérons donc que le risque est porté par les maisons plutôt que par l'ensemble des bâtiments. La prise en compte des appartements dans la fréquence constituerait un biais puisque cette partie du portefeuille n'est que marginalement exposée à ce risque.

La fréquence pour le portefeuille habitation est donc définie ainsi :

$$\text{Fréquence de sinistralité habitation} = \frac{\text{Nombre de sinistres habitation sur la zone}}{\text{Nombre de maisons sur la zone}}$$

L'objectif final pour chaque survenance simulée de notre catalogue est de pouvoir établir un nombre de sinistres automobiles et habitations. Afin d'obtenir un modèle complet, nous verrons comment différencier les appartements et les maisons à partir de l'historique de sinistralité et de la structure du portefeuille sur la zone considérée.

10.1 Association de l'intensité des chutes de grêle à la fréquence

Lors d'un orage grêligène, la grêle ne frappe pas uniformément sur toute la zone de survenance. En effet, nous observons concrètement une variabilité importante de la fréquence de sinistralité en fonction des zones durant un même évènement. Cela s'explique par la différence d'intensité des chutes de grêle entre les zones de survenance au sein d'un même orage grêligène. Ainsi, il ne faut pas seulement associer les variables climatiques à la survenance de grêle mais également à une classe d'intensité. Pour cela, nous utilisons les fréquences de sinistralité historiques en assurance habitation afin de classer les survenances de grêle en 4 classes d'intensité homogènes (1 : Intensité modérée, 2 : Intensité moyenne, 3 : Intensité forte, 4 : Intensité très importante) :

$$I_{i,d} \in \{1,2,3,4\}$$

où $I_{i,d}$ est l'indice d'intensité de la grêle pour la zone i à la date d.

Nous observons ainsi l'intensité des chutes de grêle par zone de survenance pour certains évènements historiques.

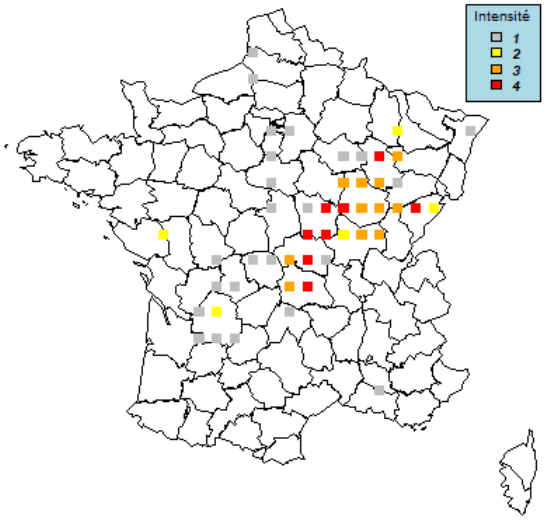


FIGURE 58 – Évènement du 6 Juillet 2001

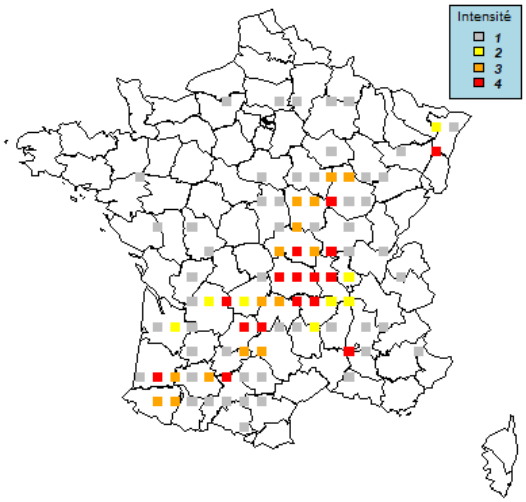


FIGURE 59 – Évènement du 6 Août 2013

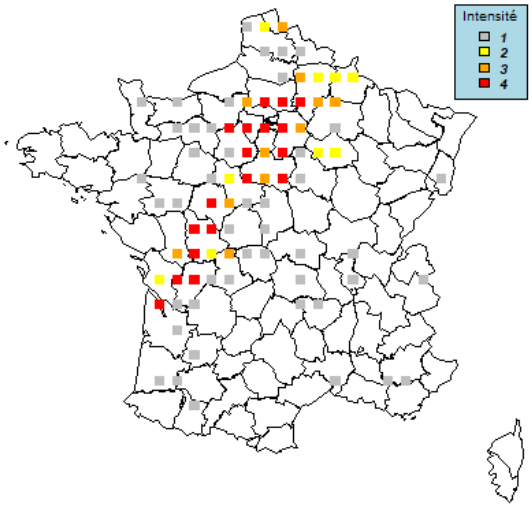


FIGURE 60 – Évènement du 8 Août 2014

Nous observons une intensité importante des chutes au coeur des orages de grêle.

Prédiction de l'intensité par *Random Forest* :

De même que pour la survenance, nous utilisons un *Random Forest* afin de prédire la classe d'intensité de chaque zone de survenance en séparant l'échantillon des données historiques de survenance en base d'entraînement et en base de test.

Nous observons dans la figure suivante l'indice de Gini cumulé moyen pour nos variables explicatives.

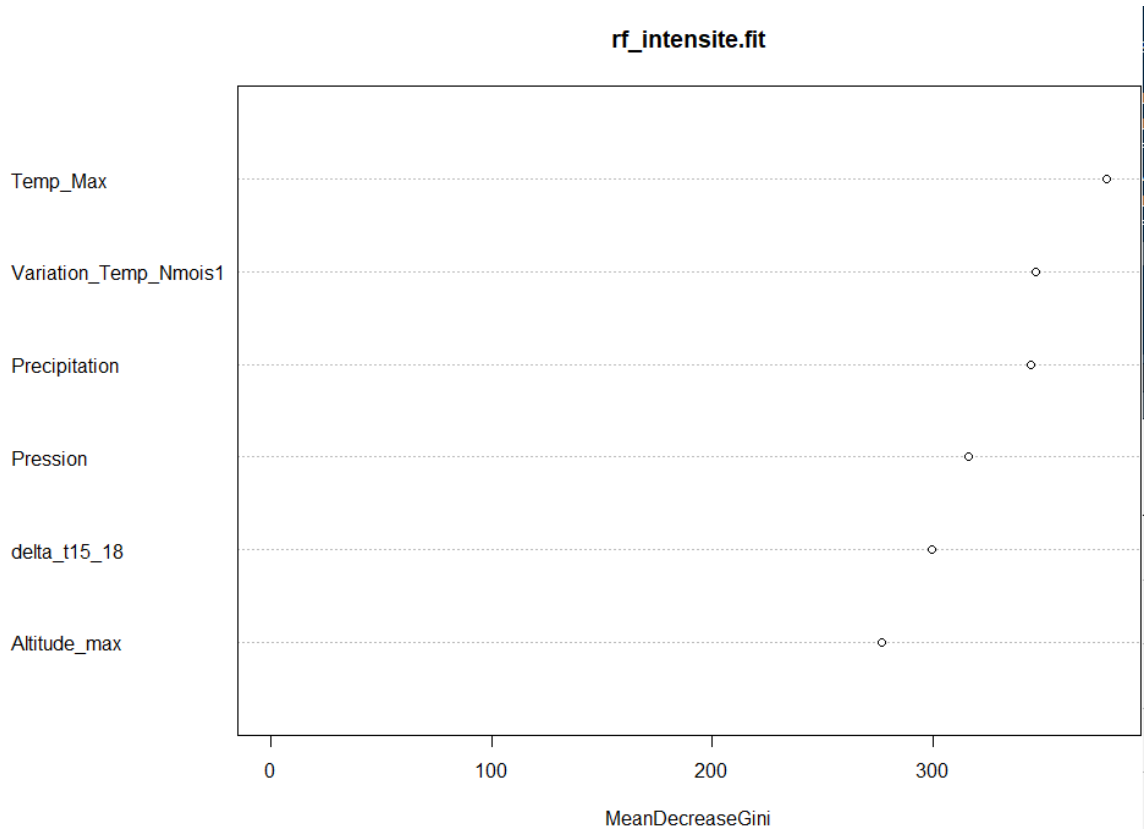


FIGURE 61 – Indice de Gini cumulé moyen pour chaque variable

Les 2 variables les plus importantes du modèle sont la température Maximale et la Variation de température maximale.
L'intensité des chutes s'explique donc principalement par des variations importantes de températures dans un contexte de fortes chaleurs.

Résultats :

Sur notre base de test, le *Random Forest* restitue la bonne classe d'intensité dans 87.5% des cas. La qualité de la prédiction est donc bonne.

Les répartitions des classes d'intensités historiques et de nos prédictions sont présentées dans les tableaux suivant :

	Intensité 1	**Intensité 2**	**Intensité 3**	**Intensité 4**
Historique	64.18%	11.90%	10.75%	13.17%
Prédictions	72.61%	9.47%	7.6%	10.32%

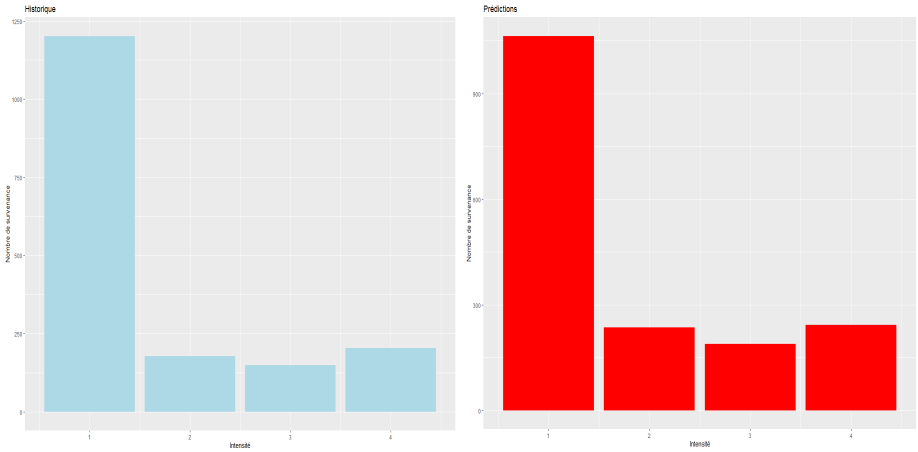


FIGURE 62 – Répartition des classes d'intensité

Le modèle a donc tendance à légèrement sous-évaluer l'intensité de la grêle.

10.2 Distributions des fréquences de sinistralité

Nous avons précédemment été en mesure de prédire un label d'intensité des chutes de grêle. Afin de tenir compte de la corrélation entre nos variables climatiques et nos fréquences de sinistralité, nous calibrons des lois conditionnellement aux labels d'intensité.

Lors de la calibration de lois statistiques, on peut s'assurer de l'adéquation de la loi sélectionnée à nos données à l'aide de test statistiques. Dans le cas continu, nous choisissons de mesurer cette adéquation à l'aide du test de Kolmogorov Smirnov.

Test de Kolmogorov Smirnov :

Dans le cas continu, afin de mesurer l'adéquation entre la fonction de répartition d'une loi théorique et la fonction de répartition empirique, le test de Kolmogorov Smirnov [**?**] est couramment utilisé.

Soit X la variable étudiée et $(x_i)_{1 \leq i \leq n}$les observations de X. Nous souhaitons tester :

H_0 : La distribution de X suit la loi théorique

H_1 : La distribution de X ne suit pas la loi théorique considérée.

Soient $\hat{F}_n$ la fonction de répartition empirique et F la fonction de répartition de la loi normale centrée réduite. On considère :

$$K = sup \mid \hat{F}_n(x) - F(x) \mid$$

Sous H_0, K suivra une distribution selon la fonction de Kolmogorov Smirnov. Soit D_α la valeur critique de la loi Kolmogorov Smirnov pour un seuil α. Alors, la zone de rejet de l'hypothèse nulle est :

$$K \leq D_\alpha$$

Nous fixons notre seuil à 5%, Si la p-value est supérieure à α, on considérera que le test est validé.

Nous calibrons plusieurs lois par maximum de vraisemblance pour chaque classe d'intensité pour les périmètres habitation et automobile et nous utilisons les test de Kolmogorov Smirnov afin de choisir laquelle décrit le mieux nos données.
Nous observons les résultats dans les tableaux suivants :

Habitation :

Intensité 1	Paramètres	KS test p-value
Pareto	lambda=0,4842 ; sigma=1.866e-05	< 7,329e-7
Gamma	shape=0,8366 ; rate=2807.4	0,0008436
Bêta	shape1=0,835 ; shape2=2802,15	5,09e-5
LGN	mu=-8,824 ; sigma=1,058	0,2368

Intensité 2	Paramètres	KS test p-value
Pareto	lambda=0,4878 ; sigma=2,867e-5	< 2,87e-7
Gamma	shape=0,8382 ; rate=1861,65	0,00116
Bêta	shape1=0,8362 ; shape2=1853,86	1,15e-5
LGN	mu=-8,40 ; sigma=0,961	0,634

Intensité 3	Paramètres	KS test p-value
Pareto	lambda=0,381 ; sigma=3,67e-5	< 4,468e-13
Gamma	shape=0,908 ; rate=947,98	0,00849
Bêta	shape1=0,907 ; shape2=944,70	0,00849
LGN	mu=-7,592 ; sigma=1,069	0.613

Intensité 4	Paramètres	KS test p-value
Pareto	lambda=0,298 ; sigma=9,23e-05	< 2,2e-16
Gamma	shape=0,832 ; rate=155,89	0,135
Bêta	shape1=0,823 ; shape2=152,78	0,033
LGN	mu=-5,94 ; sigma=1,172	0,2821

Automobile :

Intensité 1	Paramètres	KS test p-value
Pareto	lambda=0,665 ; sigma=2,002e-05	< 0,0012
Gamma	shape=0,923 ; rate=5468,78	2,001e-6
Bêta	shape1=0,923 ; shape2=5462,217	0,00011
LGN	mu=-9,316 ; sigma=0,961	0,158

Intensité 2	Paramètres	KS test p-value
Pareto	lambda=0,529 ; sigma=2,003e-5	< 6,35e-8
Gamma	shape=0,651 ; rate=1931,5	2,69e-5
Bêta	shape1=0,647 ; shape2=1909,3	2,69e-5
LGN	mu=-8,93 ; sigma=0,953	0,2179

Intensité 3	Paramètres	KS test p-value
Pareto	lambda=0,404 ; sigma=2,14e-5	< 8,19e-13
Gamma	shape=0,925 ; rate=1941,06	0,00653
Bêta	shape1=0,923 ; shape2=1936,09	0,018
LGN	mu=-8,278 ; sigma=1,025	0.922

Intensité 4	Paramètres	KS test p-value
Pareto	lambda=0,244 ; sigma=2,569e-05	< 2,2e-16
Gamma	shape=0,709 ; rate=198,37	0,0047
Bêta	shape1=0,704 ; shape2=195,53	0,00019
LGN	mu=-6,48 ; sigma=1,300	0,2821

Dans chaque cas, la loi log-normale répond le mieux au test et nous choisissons donc de modéliser l'ensemble de nos fréquences de sinistralité à partir de cette dernière.

Loi log-normale :

La loi log-normale de paramètres μ et σ admet pour densité de probabilité

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma x} e^{\frac{-(log(x)-\mu)^2}{2\sigma^2}}$$

avec μ la moyenne et σ l'écart-type.

Les distributions empiriques et modélisées sont affichées dans les figures suivantes :

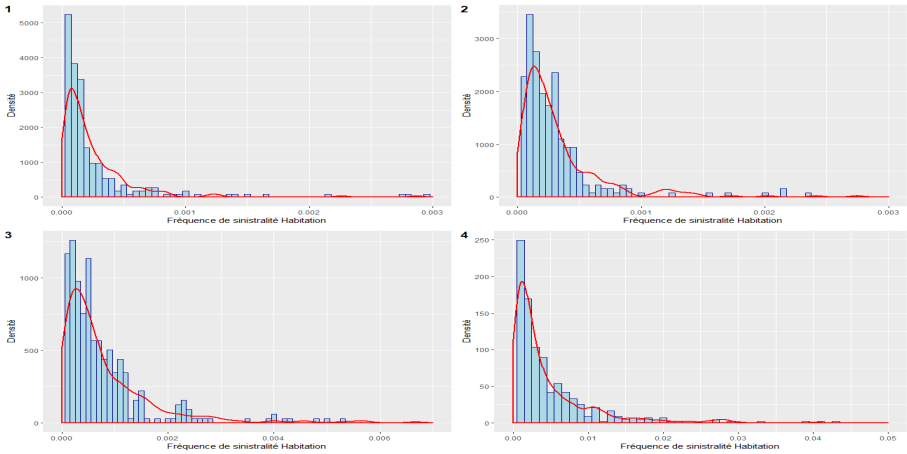


FIGURE 63 – Fréquence de sinistralité habitation - Distribution empirique (histogramme) et Simulations (rouge)

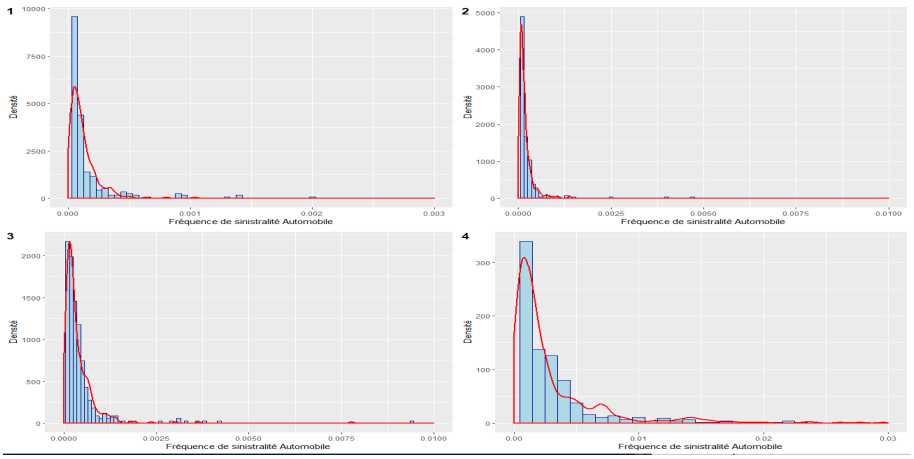


FIGURE 64 – Fréquence de sinistralité automobile - Distribution empirique (histogramme) et Simulations (rouge)

10.3 Dépendance entre fréquences de sinistralité automobile et habitation

Intuitivement, nous pouvons supposer qu'il existe une dépendance entre le nombre d'habitations et le nombre d'automobiles affectées par une même survenance de grêle. Cependant, certaines études ont par exemple démontrées qu'en dessous d'une certaine intensité, la grêle n'avait aucun impact sur les toitures de maisons mais qu'elle pouvait tout de même affecter les pares-brises de voitures. L'intensité des chutes de grêle impacte donc les habitations et les automobiles de manière différente. La structure de dépendance entre ces fréquences de sinistralité est donc complexe et demande ainsi l'utilisation de copules.

Les graphiques suivants montrent la distribution jointe entre les fréquences automobiles et habitations ainsi que la copule empirique. Cette dernière est obtenue en calculant les observations à partir des fonctions de répartition empiriques calculées ainsi :

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{X_i \leq x}$$

avec $X_1, ..., X_n$ un échantillon de variables iid à valeurs dans R

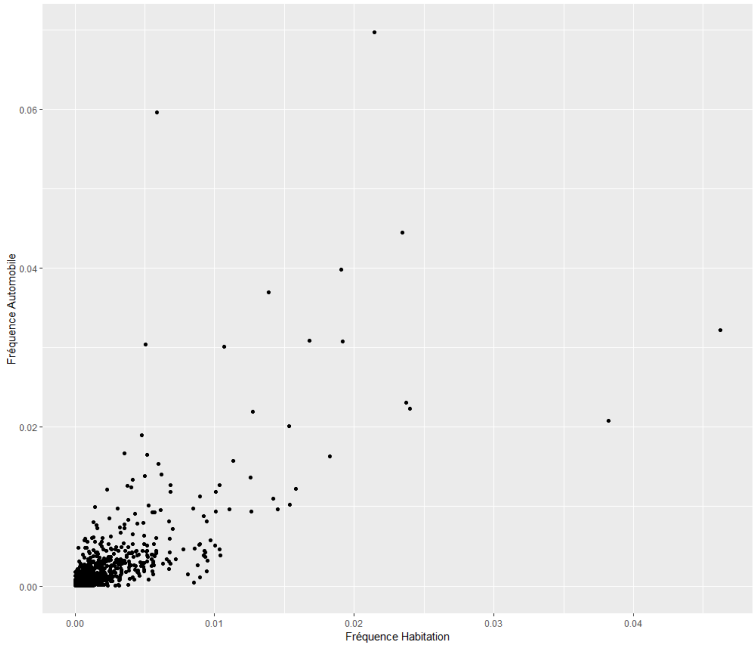


FIGURE 65 – Distribution Jointe

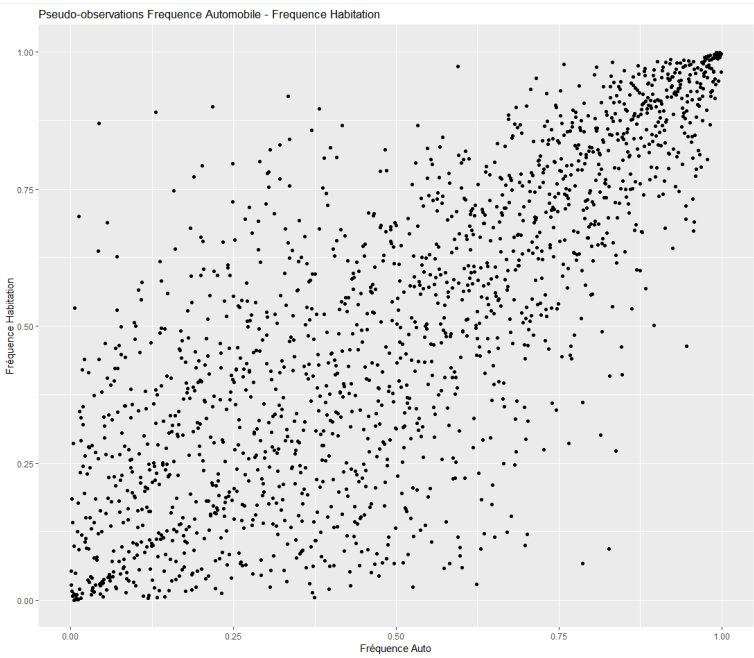


FIGURE 66 – Copule empirique

On observe graphiquement que les points ne sont pas uniformément répartis, une structure de dépendance semble donc évidente. La forte concentration observée en queue haute orienterait instinctivement notre choix vers une copule de Gumbel, qui est la copule caractérisant le mieux ce type de dépendance dans les copules usuelles utilisées.

Pour l'ensemble des 1505 zones de survenances sur la période 1999 - 2018, nous étudions la dépendance entre les fréquences habitations et automobiles en calibrant des copules Gaussienne, de Student, de Gumbel et de Clayton par maximum de vraisemblance (voir Chapitre Copules).
Les copules simulées sont affichées dans la figure suivante :

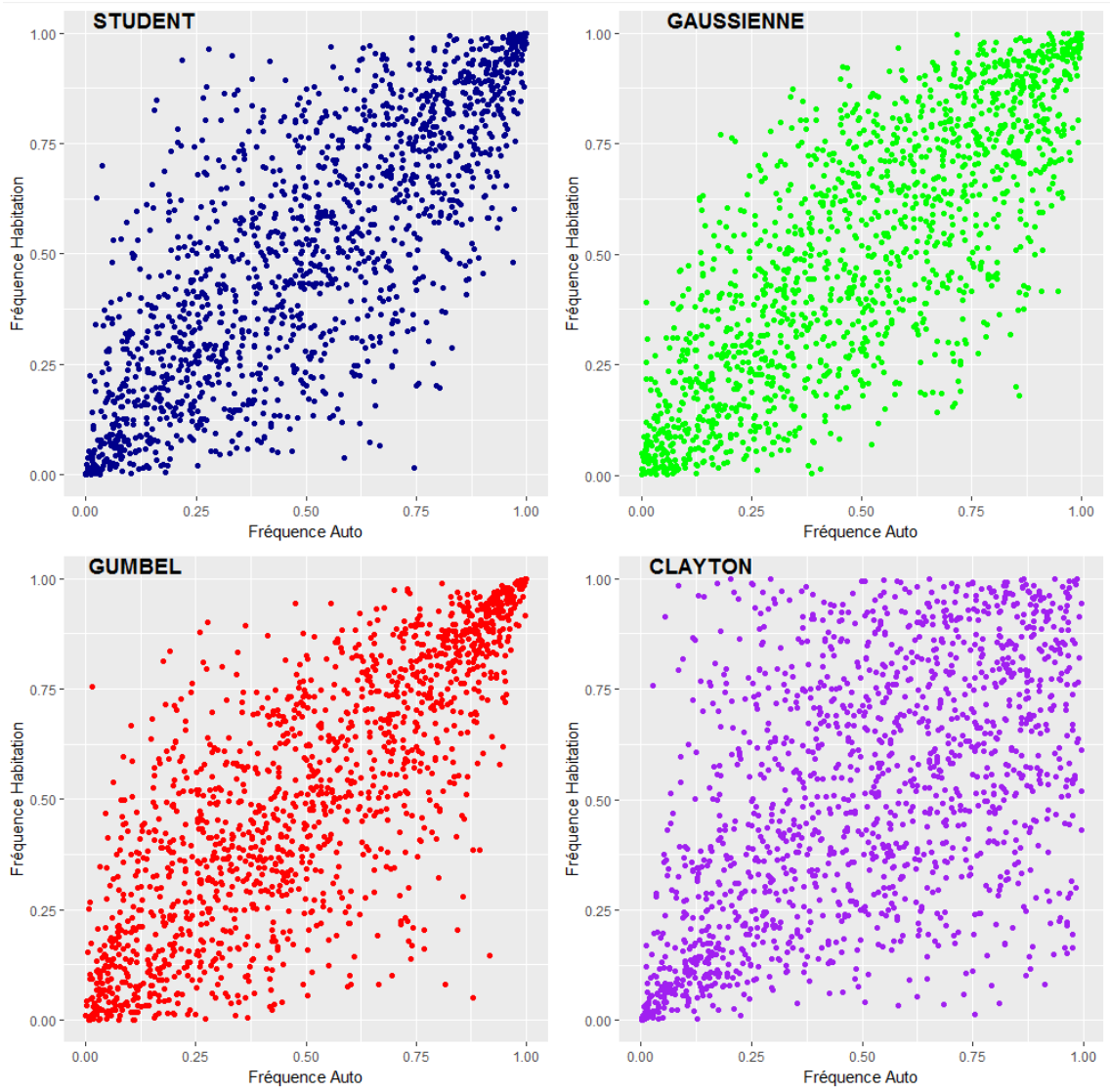


FIGURE 67 – Copules simulées

Afin de déterminer la copule la mieux adaptée, 2 critères sont utilisés : la fonction de Kendall et la Goodness of Fit (voir Chapitre Copules).

Les fonctions de Kendall empiriques et théoriques sont affichées pour l'ensemble des copules étudiées dans la figure suivante.

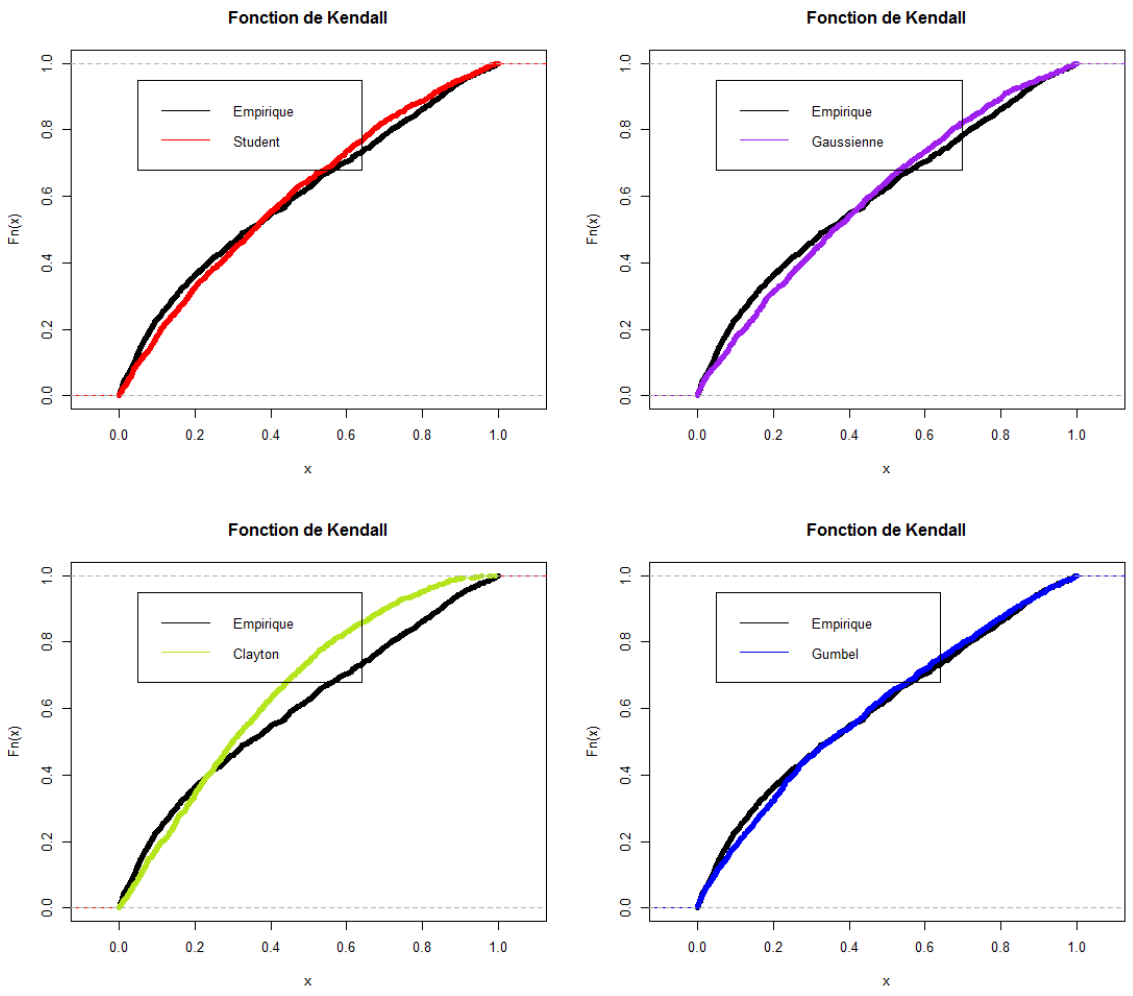


FIGURE 68 – Fonction de Kendall

Plus les courbes sont proches, plus la copule choisie approche correctement la structure de dépendance entre les 2 variables. Les fonctions de Kendall confirment nos premières observations. Les copules de Gumbel et de Student semblent être beaucoup mieux adaptées que les copules Gaussiennes et de Clayton. Cette dernière représente particulièrement mal la dépendance entre nos 2 variables.

La meilleure copule est ensuite sélectionnée à l'aide du critère quantitatif de la "Goodness of fit".
Les p-values obtenues pour chaque copule sont données dans le tableau suivant :

Copule	**p-value**
Student	0.84
Gaussienne	0.71
Gumbel	0.88
Clayton	0.21

Au regard du critère de Goodness of fit, la copule de Gumbel est la meilleure copule pour exprimer la

dépendance entre les fréquences de sinistralité des automobiles et des habitations.

Nous pouvons déduire de ces résultats que la survenance d'évènements majeurs de grêle entraîne de manière générale un nombre important d'habitations et de voitures sinistrées. Cependant, lors de la survenance d'évènements de moindre ampleur associant des fréquences de sinistralité moins importante, cette relation est moins forte.

10.4 Méthode de simulation

10.4.1 Obtention des fréquences automobiles et habitations

Nous voulons générer des fréquences de sinistralité pour notre échantillon de survenances tout en tenant compte de la dépendance (étudiée précédemment).

Pour chaque survenance de grêle, les fréquences sont générées selon la méthode suivante :

— Simulation d'un vecteur aléatoire (U_1, U_2) caractérisé par une structure de dépendance correspondant à la copule précédemment déterminée ;

— Soient F_A et F_H respectivement les fonctions de répartitions de la fréquence automobile et de la fréquence habitation, les fréquences simulées auront donc pour valeur : $(A, H) \simeq (F_A^{-1}(U), F_H^{-1}(V))$

10.4.2 Nombre de sinistres par survenance

Après avoir obtenu les fréquences de sinistralité pour chaque survenance, les nombres de sinistres affectant les automobiles et les maisons s'obtiennent par les calculs suivants :

$$NS_A = FR_A \times N_A$$

où NS_A le nombre de sinistres automobiles, FR_A fréquence de sinistralité automobile et N_A le nombre d'automobiles assurées.

$$NS_{Hm} = FR_{Hm} \times N_{Hm}$$

où NS_{Hm} le nombre de sinistres habitation impliquant des maisons, FR_{Hm} la fréquence de sinistralité maison et N_{Hm} le nombre de maisons assurées.

Précédemment, nous avons vu que le risque habitation est majoritairement porté par les maisons. Le nombre d'appartement sinistrés reste non négligeable et est évidemment corrélé au nombre de maisons sinistrées. Cependant, cette relation n'est pas identique sur l'ensemble du territoire à cause des différences de répartition du portefeuille en fonction des différentes zones.

La figure ci-dessous représente la proportion de maisons (et par opposition la proportion d'appartements) à l'échelle départementale en France en 2017 sur le portefeuille Pacifica.

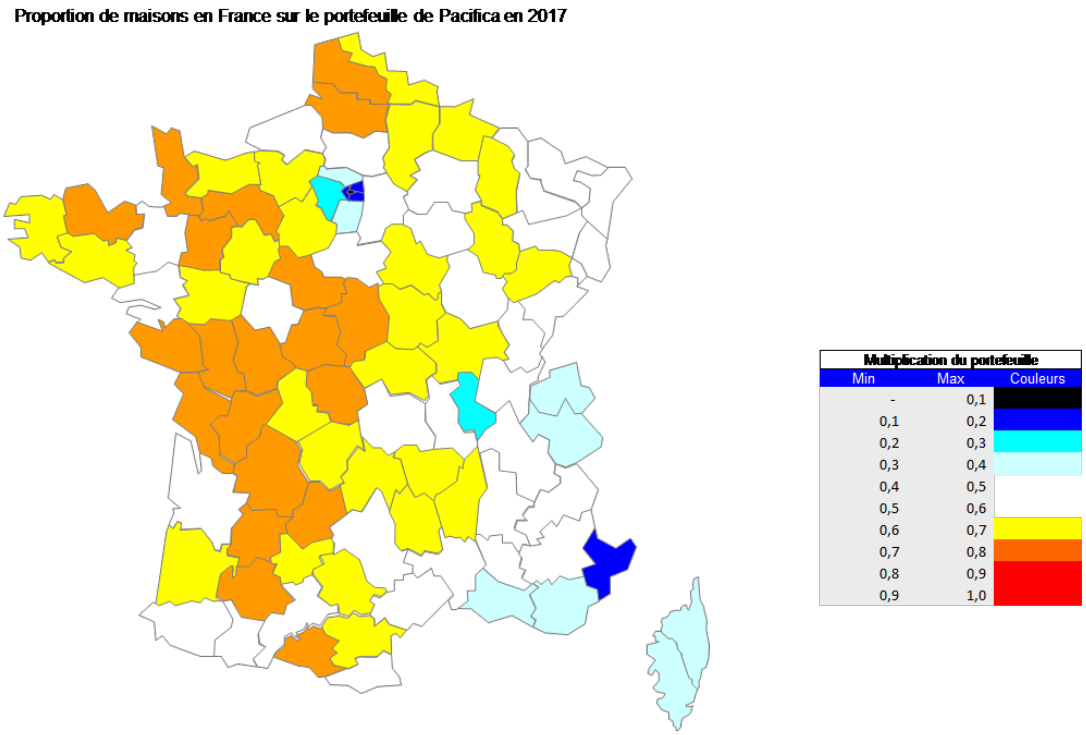


FIGURE 69 – Proportion de Maisons à l'échelle départementale en 2017

Le portefeuille Pacifica comporte à l'échelle nationale 52% de maisons. Or, 97,2% de la charge historique et 94,8% du nombre de sinistres habitation dus à la grêle concernent des maisons. Par ailleurs, une maison sinistrée coûtera en moyenne 1,87 fois plus chère qu'un appartement sinistré.

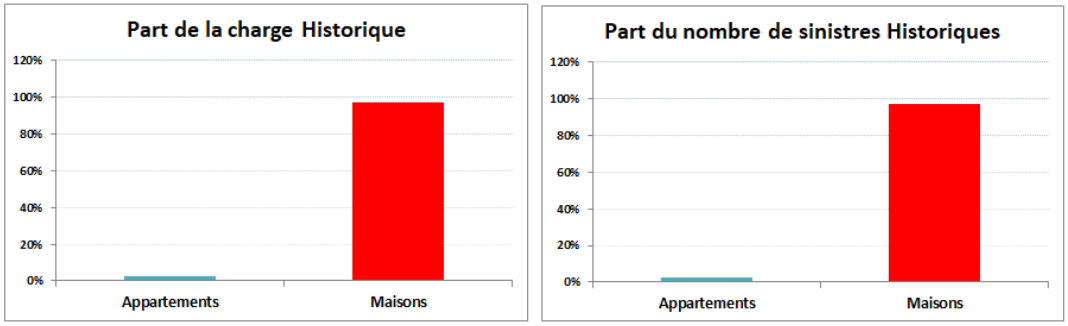


FIGURE 70 – Répartition des classes d'intensité

Il est possible que les copropriétés prennent en charge les sinistres affectant les immeubles. Ce type d'assurance n'étant pas considéré, c'est possiblement un facteur entraînant le nombre de sinistres d'appartements à la baisse. Néanmoins, les maisons restent donc beaucoup plus exposées à la grêle que les appartements. Cette tendance est d'autant plus criante lorsque l'on prend le cas de Paris. En effet, les maisons ne représentent que 7% des habitations de la capitale, mais historiquement 62% des sinistres

dus à la grêle touchent des maisons. Malgré la forte concentration au sein de la ville, l'exposition de Paris au risque de grêle reste limitée.
Le peu de sinistres concernant les appartements est problématique car certaines zones de survenances historique ne contiennent pas ce type de sinistres et il devient donc difficile d'étudier le comportement de ces sinistres.
Pour tenir compte de ces différences d'expositions, le nombre de sinistres affectant les appartements est reconstitué à partir de la répartition de la sinistralité historique entre les maisons et les appartements à l'échelle départementale pour tous nos évènements majeurs de grêle.

Cette répartition historique est appliquée brute directement sur nos survenances simulées.
Le nombre d'appartements touchés est alors caractérisé par la formule suivante :

$$NS_{Ha} = NS_H - NS_{Hm}$$
$$NS_{Ha} = \frac{NS_{Hm}}{\%S_{Hm}} - NS_{Hm}$$

où NS_{Ha} le nombre de sinistres habitation impliquant des appartements, NS_{Hm} le nombre de sinistres habitation impliquant des maisons, NS_H le nombre de sinistres habitation et $\%S_{Hm}$ le pourcentage de maisons sinistrées historiquement sur le département.

11 Taux de destruction

Afin d'étudier la vulnérabilité du portefeuille, on caractérise le taux de destruction d'un sinistre par la formule suivante :

$$\text{Taux de destruction} = \frac{\text{Charge du sinistre}}{\text{Somme assurée}}$$

11.1 Sommes assurées

La somme assurée est une caractéristique importante car elle définit l'engagement de l'assureur en cas de survenance d'un sinistre. Elle représente la valeur théorique du bien et donc le montant maximal qui pourra être versé en cas de sinistre (cas où le taux de destruction est de 100%).

En France, ce montant n'est pas publique et doit donc être estimé à partir des caractéristiques du bien et du contrat. En présence de contrats habitations, la somme assurée d'un bien est donnée par la formule suivante :

$$\text{Somme assurée}_i = \text{BSI}_i + \text{CSI}_i + \text{BSI\_D}_i + \text{CSI\_D}_i$$

avec

- BSI_i la valeur du bâti du logement principal ;
- CSI_i la valeur du contenu du logement principal ;
- BSI_D_i la valeur du bâti de la dépendance du logement ;
- CSI_D_i la valeur du contenu du logement principal.

Pour un locataire, la valeur du bâti (BSI) sera par exemple nulle car celui-ci ne possède pas les murs du logement dans lequel il habite. Cependant, les biens à l'intérieur du logement (CSI) lui appartiennent et sont donc pris en compte dans le calcul des sommes assurées. Par opposition, un propriétaire non occupant possède le logement mais ce dernier ne contient pas de biens lui appartenant. Le CSI sera donc nul.

Dans un premier temps, les sommes assurées sont donc basées sur les caractéristiques de l'assuré, mais la répartition de ces caractéristiques varie d'un département à l'autre.

Nous observons dans la figure suivante la proportion de propriétaire pour chaque département de France métropolitaine sur le portefeuille Pacifica en 2017.

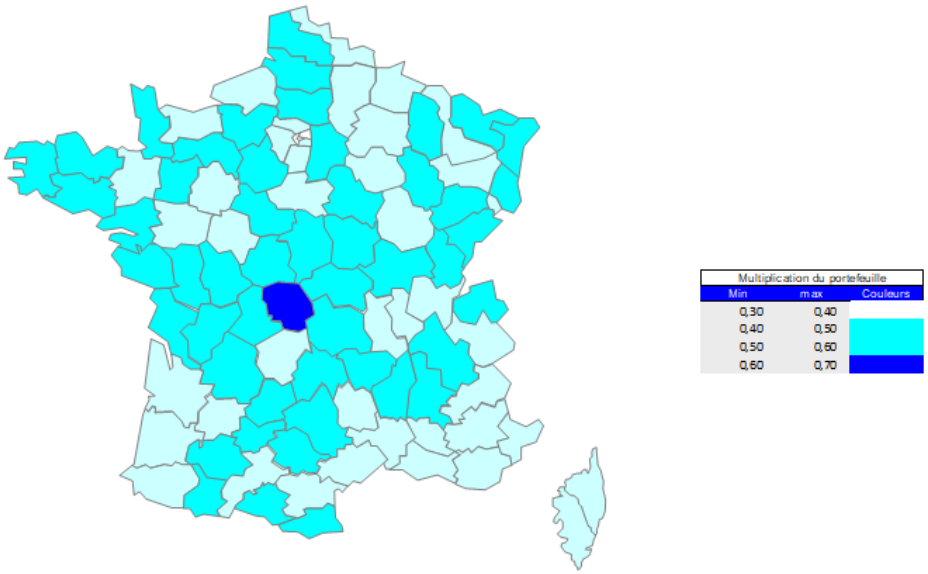


FIGURE 71 – Proportion de propriétaire en France sur le portefeuille Pacifica en 2017

De plus, le calcul des sommes assurées de chaque bien est également basé sur ces caractéristiques sous-jacentes. La valeur associée au bâti est caractérisée par plusieurs paramètres comme le nombre de pièces ou le type de logement (maison ou appartement). Ces caractéristiques sont observées indépendamment pour chaque département de manière à obtenir l'estimation la plus fine possible. En effet, la valeur associée au bâti d'un bien à Paris est sensiblement supérieure à la valeur associée au bâti d'un bien d'un département rural.

11.2 Redressement des sinistres

Dans l'optique du calcul des taux de destruction, la charge historique de chaque sinistre est disponible. Cependant, ces charges historiques ne sont plus d'actualité. En effet, les prix des biens sinistrés subissent avec le temps une certaine inflation dont il faut tenir compte au moment d'observer la charge à la vision de l'année 2019.

L'inflation inhérente au portefeuille Multirisque habitation est reconstituée à partir de l'indice FFB de la Fédération Française du Bâtiment, qui caractérise l'évolution des coûts de constructions des bâtiments. Cet indice est pertinent car il caractérise directement la nature des matériaux détruits par la grêle. Nous

calculons ensuite un coefficient de redressement à partir de cet indice en prenant pour année de base l'année 2018.

$$C_i = \frac{I_i}{I_{2018}}$$

où C_i le coefficient de redressement de l'année i, I_i l'indice FFB de la Fédération Française du Bâtiment de l'année i, I_{2018} l'indice FFB de la Fédération Française du Bâtiment de l'année 2018.

Le graphique suivant montre l'évolution de ce coefficient au cours du temps :

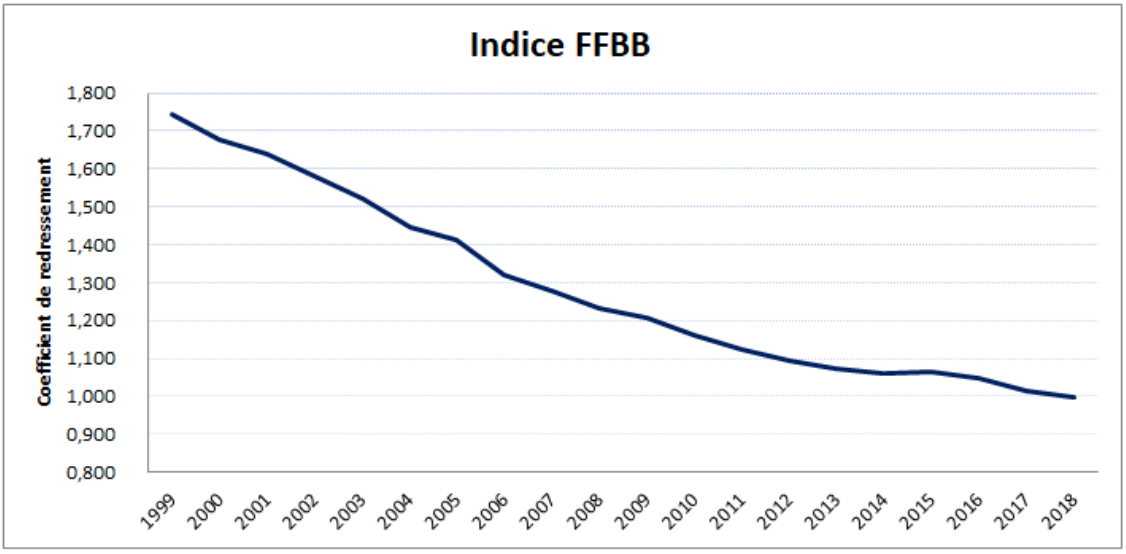


FIGURE 72 – Indice FFB

Pour le portefeuille automobile, l'inflation est caractérisée par l'indice INSEE du coût de la réparation automobile. Cet indice semble moins pertinent que l'indice FFB car la plupart des sinistres automobiles sont des bris de glace. Cependant, nous considérons que c'est la meilleure mesure disponible.
De la même manière, la figure suivante montre l'évolution de ce coefficient au cours du temps :

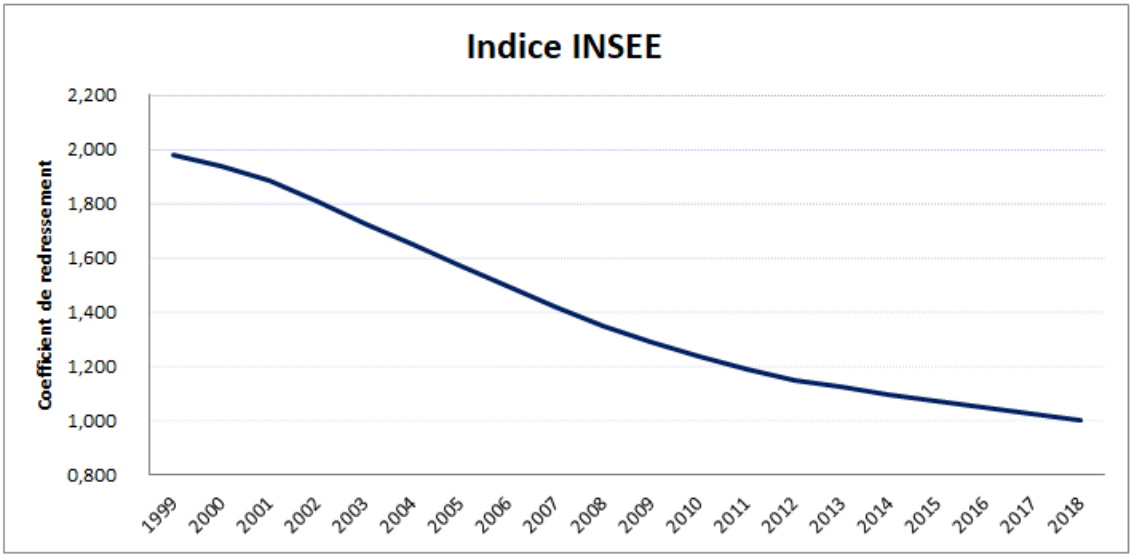


FIGURE 73 – Indice automobile

11.3 Relation entre taux de destruction et intensité

Intuitivement, on peut penser qu'un grand nombre de sinistres et des taux de destructions conséquents sont les caractéristiques d'un évènement majeur. Cependant, la littérature est assez floue sur le sujet et aucune étude n'a prouvée l'existence de cette relation.

Afin de simuler des fréquences et des taux de destruction cohérents, cette relation est étudiée sur l'historique du portefeuille Pacifica.

La figure suivante affiche les taux de destruction moyens en fonction des fréquences par station et par jour respectivement pour l'habitation et pour l'automobile.

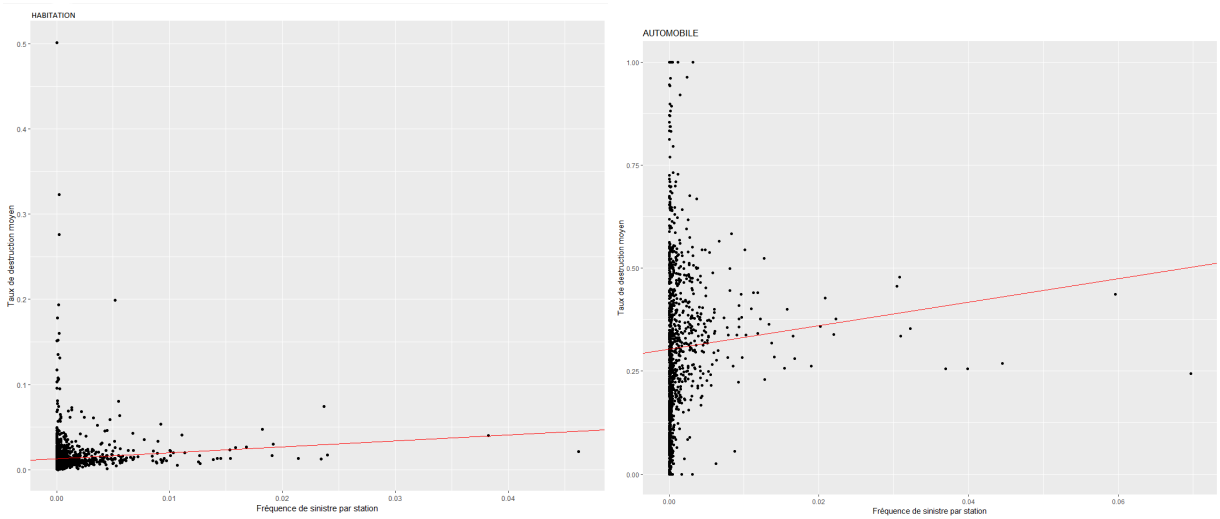


FIGURE 74

Nous observons une légère tendance croissante.

Nous calculons et affichons les pseudo-observations sur la figure suivante afin de distinguer une structure de dépendance sur les copules empiriques.

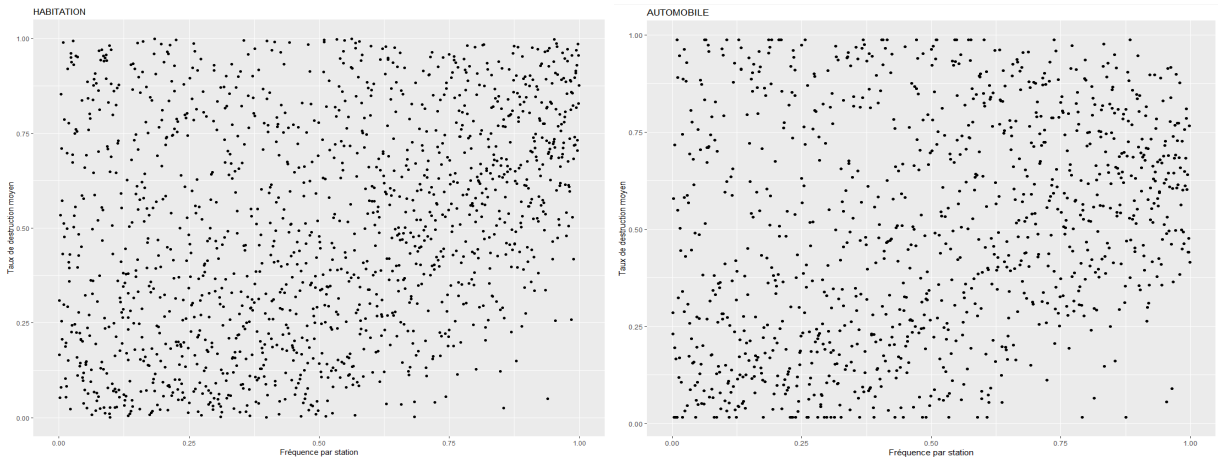


FIGURE 75 – Copules empiriques entre fréquence et taux de destruction pour habitation et automobile

Nous n'observons pas de structure de dépendance spécifique.
Nous remarquons cependant que les évènements provoquant un nombre conséquent de sinistres n'entraînent que rarement des taux de destruction faibles.

Nous observons dans les tableaux suivants les coefficients de Pearson et de Kendall entre les fréquences de sinistralité et les taux de destruction moyens par zone d'occurrence historique pour les automobiles et les habitations.

Habitation :

Coefficient de Pearson	**Coefficient de Kendall**
0.061	0.187

Automobile :

Coefficient de Pearson	**Coefficient de Kendall**
0.085	0.21

Nous observons des coefficients de corrélation faibles. Cependant, nous ne pouvons conclure à l'indépendance entre la fréquence de sinistralité et les taux de destructions et nous décidons donc d'étudier le sujet de manière plus précise en caractérisant les taux de destruction par classe d'intensité.

Le box plot est un outil permettant de résumer quelques caractéristiques de la série étudiée (moyenne, médiane, quartiles, minimum, maximum).

Nous observons dans la figure suivante le box plot des taux de destruction des propriétaires de maisons pour les 4 classes d'intensité.

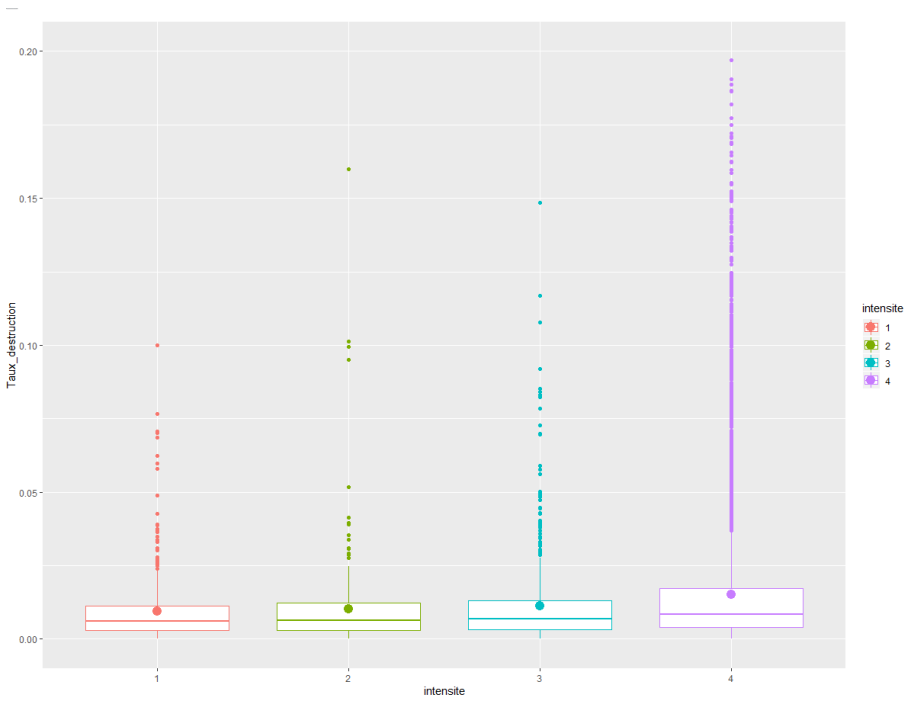


FIGURE 76 – Taux de destruction des propriétaires de maisons par classe d'intensité

Pour chaque classe d'intensité, les médianes, marquées par les barre horizontale au coeur des rectangles, sont comparables. Nous remarquons également que les 3 premières classes d'intensité ne se distinguent pas non plus par les valeurs de leurs moyennes (point majeure au coeur du rectangle) et du troisième quartile (troisième barre horizontale) ni par les valeurs à l'extérieur des moustaches, représentées par des points. Cependant, la classe d'intensité 4 semble se différencier des autres par sa moyenne et son 3ème quartile plus élevés, ainsi que des valeurs à l'extérieur des moustaches beaucoup plus nombreuses et beaucoup plus importantes. Cette dernière observation peut par ailleurs s'expliquer par la masse plus importante de sinistres lors d'une survenance d'intensité 4.

Compte tenu des différences observées sur les Box plot, nous décidons de modéliser nos taux de destruction de maisons conditionnellement aux labels d'intensité.

Nous calibrons plusieurs lois par maximum de vraisemblance pour chaque classe d'intensité et nous utilisons les test de Kolmogorov Smirnov afin de choisir laquelle décrit le mieux nos données.
Nous observons les résultats dans les tableaux suivants :

Intensité 1	**Paramètres**	**KS test p-value**
Pareto	lambda=0,2693 ; sigma=0,0001341	< 2,2e-16
Gamma	shape=1,048587 ; rate=110,41	0,01565
Bêta	shape1=1,036 ; shape2=107,816	0,2997
LGN	mu=-5,204 ; sigma=1,098	0.5792

Intensité 2	Paramètres	KS test p-value
Pareto	lambda=0,2926 ; sigma=0,0001879	< 2,831e-14
Gamma	shape=0,8938 ; rate=81,07	0,01352
Bêta	shape1=0,8654 ; shape2=76,2561	0,005955
LGN	mu=-5,1623 ; sigma=1,150	0.7676

Intensité 3	Paramètres	KS test p-value
Pareto	lambda=0,2767 ; sigma=0,0001716	< 2,2e-16
Gamma	shape=0,9545 ; rate=81,607	0,01834
Bêta	shape1=0,9353 ; shape2=78,431	0,0183
LGN	mu=-5,056 ; sigma=1,140	0.4333

Intensité 4	Paramètres	KS test p-value
Pareto	lambda=0,2066 ; sigma=6,44e-05	< 2,2e-16
Gamma	shape=0,9019 ; rate=57,842	2,2e-16
Bêta	shape1=0,8792 ; shape2=54,971	2,2e-16
LGN	mu=-4,809 ; sigma=1,163	0.00198

Pour les 3 premières classes d'intensité, la loi log-normale est la seule adaptée au regard du test de Kolmogorov Smirnov. Pour l'intensité 4, aucune ne valide l'hypothèse d'adéquation, mais la loi log-normale maximise la p-value et nous décidons donc également de la conserver pour cette classe d'intensité.

Les distributions empiriques et modélisées sont affichées dans les figures suivantes :

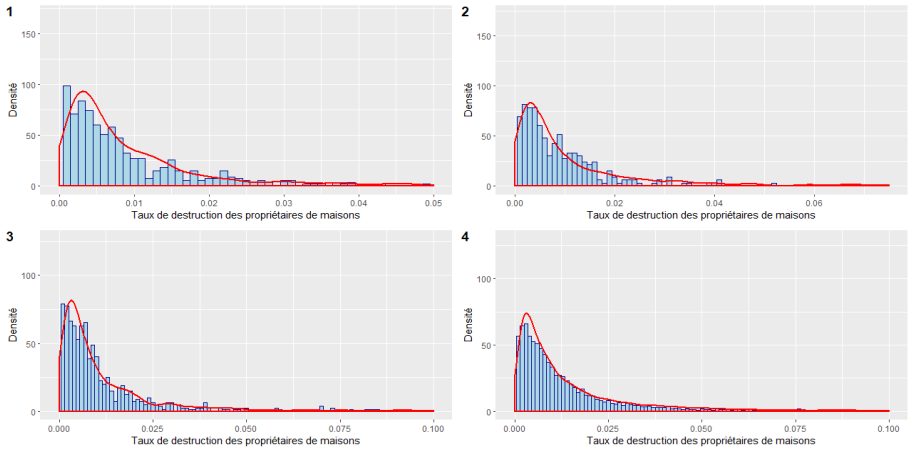


FIGURE 77 – Taux de destruction des propriétaires de maisons - Distribution empirique (histogramme) et Simulations (rouge)

Nous adoptons le même procédé pour les taux de destruction automobiles qui seront pour leurs parts considérés par des lois gamma. Ces résultats sont présentés en annexe.

Nous observons dans la figure suivante le box plot des taux de destruction des propriétaires d'appartements pour les 4 classes d'intensité.

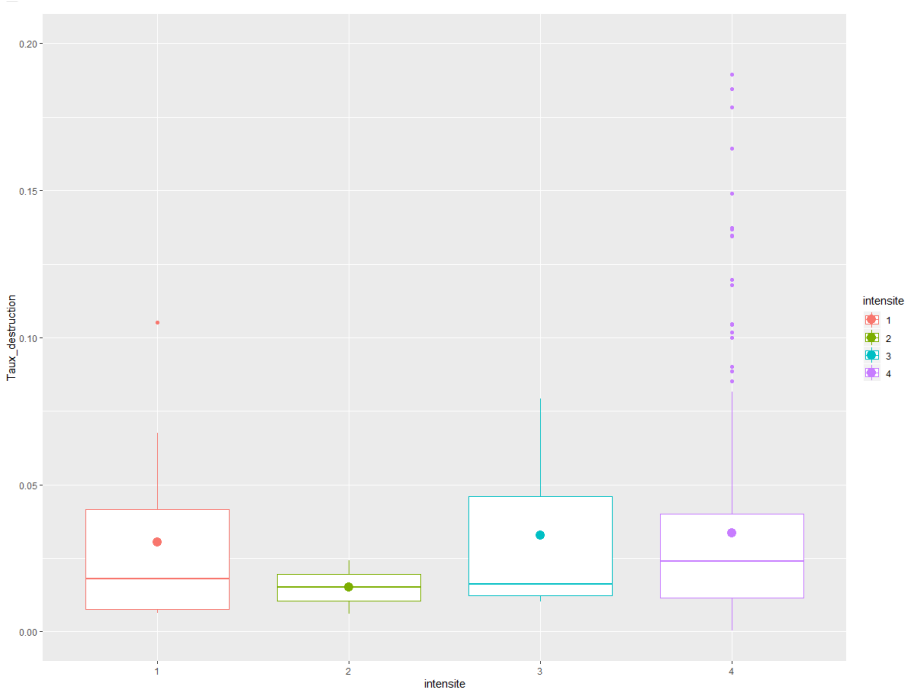


FIGURE 78 – Taux de destruction des propriétaires d'appartements par classe d'intensité

Nous ne remarquons pas de différences particulières pour les indicateurs de chaque classe d'intensité. De plus, la masse de sinistres concernant des appartements étant relativement faible, nous décidons de modéliser les taux de destruction des appartements indépendamment de l'intensité.

Conclusion :

Au sein de ce module vulnérabilité, nous avons pu caractériser le risque de grêle en fonction de divers paramètres.
Premièrement, nous définissons quelles caractéristiques propres aux biens assurés (habitations et automobiles) permettraient d'affiner au mieux la modélisation du risque et nous observons ensuite à l'échelle départementale la répartition de ces types de biens.
Deuxièmement, nous caractérisons la sinistralité par un nombre de biens touchés et un taux de destruction pour chaque zone et chaque type de bien à l'aide de la calibration de lois tout en considérant la dépendance entre les différents paramètres à l'aide des copules.
A partir de cela, nous associons à chaque survenance simulée dans le module Aléa un nombre d'appartements, de maisons et d'automobiles touchés ainsi qu'un taux de destruction associé à chaque caractéristique.

A l'aide de l'estimation des sommes assurées pour chaque zone, nous serons donc en mesure d'associer une charge à chaque survenance et donc à chaque évènement.

12 Module financier

Nous allons dans cette partie appliquer les résultats des 2 précédentes parties pour estimer la charge engendrée par le risque de grêle sur un portefeuille assurantiel.

Pour cela, nous partons du recensement des sommes assurées par département effectué précédemment dans le module Vulnérabilité.

Les évènements grêle, générés de façon quotidienne avec le module Aléa, sont appliqués au portefeuille pour traduire les fréquences et les taux de destruction du risque en une perte financière.

Notre catalogue d'évènement est désormais caractérisé par :

— Des stations de survenances ;

— Des nombres de sinistres pour les automobiles, les maisons et les appartements pour chaque station ;

— Des taux de destruction pour chaque sinistre.

A partir de ces informations, nous sommes capable de dresser un coût par station et ainsi, en regroupant l'ensemble des survenances, obtenir des charges par évènement et des charges annuelles.

Les polices du portefeuille ont été regroupées par département pour disposer des sommes assurées totales et de la somme assurée moyenne des biens de chacune des catégories (automobiles, propriétaires de maisons, propriétaires d'appartements, locataires de maisons, locataires d'appartement).

La charge est donc donnée pour chaque station de survenance par la formule suivante :

$$C = NS_i \times TD_i \times SA_i$$

où NS_i correspond au nombre de sinistres de la classe i, TD_i correspond au taux de destruction de la classe i et SA_i correspond à la somme assurée moyenne de la classe i sur la zone.

La charge associée à un évènement de grêle se calcule donc simplement en sommant la charge associée à chaque station.

12.1 Construction des courbes OEP - AEP

La courbe OEP représente la distribution de la perte maximale pour un seul évènement pendant une année tandis que la courbe AEP représente la distribution de la perte annuelle totale.

12.1.1 Courbes OEP et AEP

La courbe OEP présente les périodes de retour associées à l'évènement de perte maximale sur chaque année. Nous disposons ainsi de 10 000 pertes maximales annuelles. Après avoir ordonné ces pertes par ordre décroissant, nous pouvons tracer la courbe OEP en associant à chaque perte, une période de retour.

La courbe AEP représente les périodes de retour associées à la distribution des pertes annuelles. Pour obtenir la perte annuelle d'un scénario, nous sommons les pertes de l'ensemble des évènements pour

chaque année. Nous disposons ainsi de 10 000 pertes annuelles .
Pour tracer la courbe AEP, nous ordonnons les pertes des 10 000 scénarios par ordre croissant, ce qui permet de déterminer les périodes de retour associées de la manière suivante :

$$PR = \frac{1}{1 - Q_\alpha}$$

où PR est la Période de Retour et Q_α est le quantile d'ordre α.

Voici les courbes AEP et OEP obtenues avec 10 000 simulations, les pertes annuelles totales et maximales, en euros, sont représentées en fonction de la période de retour en années.

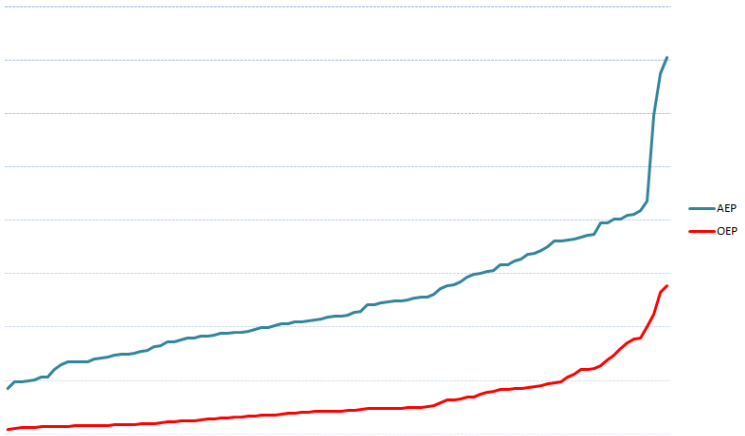


FIGURE 79 – Courbes AEP - OEP

12.1.2 Cohérence des Résultats

Afin de challenger notre modèle, nous comparons ce dernier à notre historique de sinistralité d'une durée de 20 ans (entre 1999 et 2018).

Dans la figure suivante, nous comparons la période de retour des charges redressées annuelles de notre historique et de notre modèle.

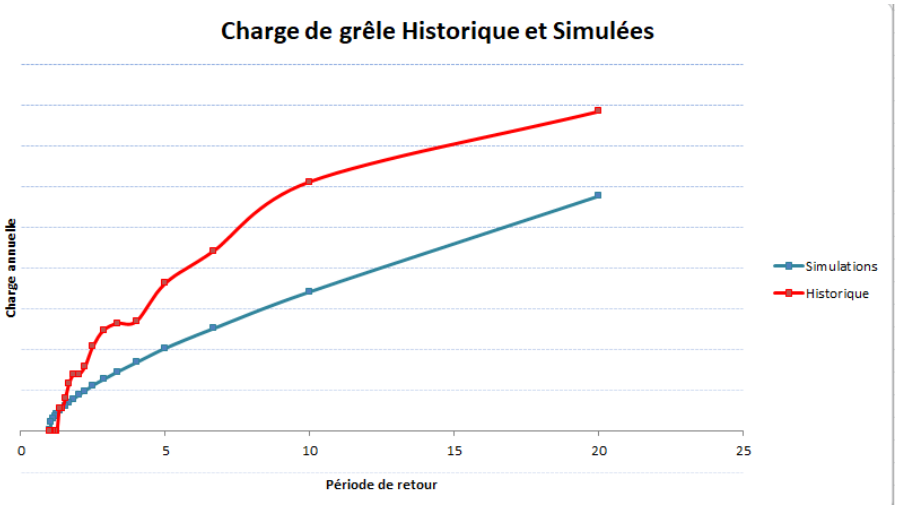


FIGURE 80 – Charges Historiques et Simulées

Tout d'abord, en observant les basses périodes de retour, nous remarquons que notre modèle simule un évènement de grêle chaque année tandis que l'historique démarre aux 1,33 ans. En effet, Pacifica a connu 5 années sans évènement majeur de grêle en 20 ans.
Par la suite, les courbes se croisent aux 1,5 ans. Suite à cela, le modèle sous-performe comparativement à l'historique. Pour une période de retour à 20 ans, le modèle est inférieur de 27% à l'historique.

Cinquième partie

Conclusion

Ce mémoire a pour objectif premier de modéliser les pertes engendrées par la grêle sur les habitations et les automobiles dans une compagnie d'assurance non vie à partir de données climatiques et en tenant compte de la structure spatiale du portefeuille.
Afin d'atteindre cet objectif, nous avons commencé par reconstituer les empreintes de chutes de grêle des orages grêligènes historiques à partir de la cartographie de la sinistralité historique de Pacifica. Ensuite, à l'aide de méthodes statistiques, nous avons mis en place un modèle prédictif de la survenance et de l'intensité de grêle en associant les survenances historiques aux données météorologiques disponibles. Ainsi, nous avons pu générer aléatoirement des scénarios de grêle potentiels. L'historique de sinistralité nous a permis de caractériser l'exposition du portefeuille Pacifica et ainsi de mettre en parallèle les simulations de grêle aux dégâts associés aux biens. Finalement, nous avons été en mesure d'estimer les pertes économiques associées à chaque évènement à partir de l'estimation des sommes assurées.

Ce modèle est basé sur une approche innovante. Cependant, il produit des résultats contrastés caractérisés par des évènements de grêle simulés de dimensions très largement inférieures aux observations historiques. Ces différences peuvent s'expliquer par la perte de qualité engendrée par un Random Forest qui sous-estime les survenances (30% de moins) mais également par certaines hypothèses prises en compte comme l'indépendance de la variable « précipitations » ou par le regroupement en zones globales basé sur des moyennes pouvant potentiellement induire des circonstances extrêmes plus rares et donc des empreintes peu étendues.
De plus, nous basons notre études sur des données climatiques recensées sur les 20 dernières années. Or, l'augmentation du nombre de catastrophes naturelles et le réchauffement climatique indiquent une évolution progressive des conditions climatiques. Cette tendance n'est pas prise en compte et son utilisation peut constituer un axe d'amélioration.
L'objectif était de mettre en place un modèle permettant à un assureur de calculer les fonds propres nécessaires pour faire face à des évènements de grande ampleur comme la grêle de la Pentecôte de 2014. Notre approche ne retourne pas des résultats assez sévères pour répondre à la problématique initiale. En revanche, certains axes d'amélioration, notamment lors de la modélisation des variables explicatives, pourraient rendre cette étude pertinente.

Finalement, nous avons réalisé une modélisation du péril grêle dont les utilisations sont multiples, comme l'approfondissement de la connaissance du risque ou encore l'optimisation de la réassurance. Ces travaux pourront d'ailleurs être potentiellement étendus à d'autres périls similaires mais également à des portefeuilles plus diversifiés.

Sixième partie

Annexes

A Modélisation des précipitations

	Températures maximales -zone 1	Températures maximales -zone 2	Températures maximales -zone 3	Températures maximales -zone 4	Températures maximales -zone 5
Températures maximales -zone 1	1	0,42	0,74	0,57	0,69
Températures maximales -zone 2		1	0,89	0,74	0,70
Températures maximales -zone 3			1	0,37	0,79
Températures maximales -zone 4				1	0,59
Températures maximales -zone 5					1

FIGURE 81 – Paramètres de la copule Gaussienne

Degrés de liberté : 8					
	Températures maximales -zone 1	Températures maximales -zone 2	Températures maximales -zone 3	Températures maximales -zone 4	Températures maximales -zone 5
Températures maximales -zone 1	1	0,40	0,72	0,57	0,67
Températures maximales -zone 2		1	0,90	0,74	0,69
Températures maximales -zone 3			1	0,36	0,78
Températures maximales -zone 4				1	0,59
Températures maximales -zone 5					1

FIGURE 82 – Paramètres de la copule de Student

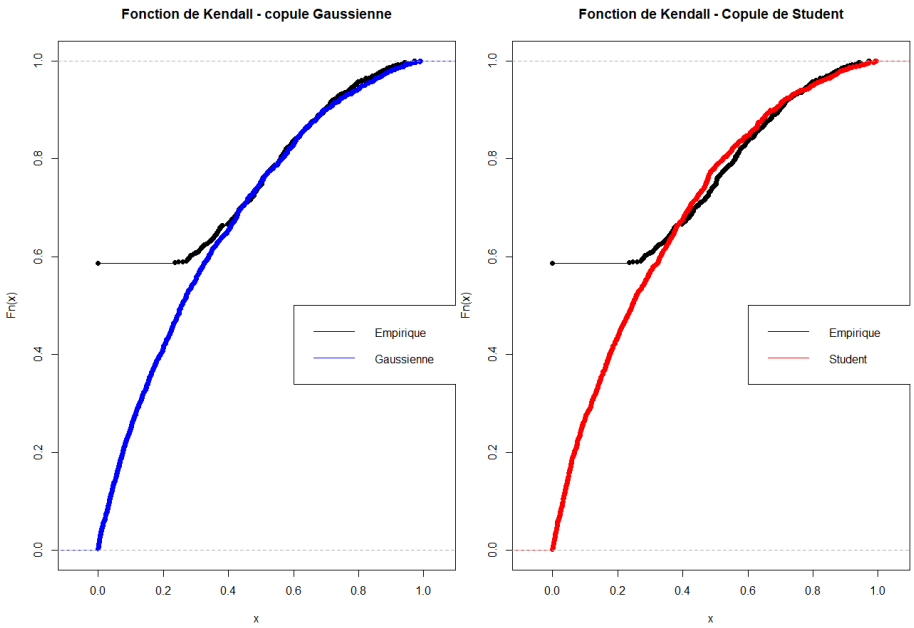


FIGURE 83 – Paramètres de la copule de Student

B Lois continues

B.1 Loi Gamma

Soit Y une variable aléatoire. Nous disons que Y suit une loi Gamma de paramètre r et β strictement positifs si la loi de probabilité s'écrit comme suit :

$$f(y;\beta,r) = \frac{\beta^r}{\Gamma(r)\,y^{r-1}exp(-\beta y)}$$

où $\Gamma(x) = \int_{y\leq 0} \exp(-y)\,y^{x-1}\,\mathrm{d}y$

L'espérance de cette loi vaut r/β et sa variance r/β^2.

B.2 Loi de Pareto

Soit Y une variable aléatoire. Nous disons que Y suit une loi de Pareto de paramètre k et x_m si la loi de probabilité s'écrit comme suit :

$$f(y;k,x_m) = k\frac{x_m^k}{x^{k+1}}$$

L'espérance de cette loi vaut $\frac{kx_m}{k-1}$ et sa variance $\left(\frac{x_m}{k-1}\right)^2 \frac{k}{k-2}$.

C Taux de destruction

C.1 Propriétaires d'appartements

	Paramètres	KS test p-value
Pareto	lambda=0,1335 ; sigma=1,413e-05	< 2,2e-16
Gamma	shape=0,793 ; rate=14,838	2,789e-06
Bêta	shape1=0,211 ; shape2=3,743	< 2,2e-16
LGN	mu=-3,677 ; sigma=1,218	0,117

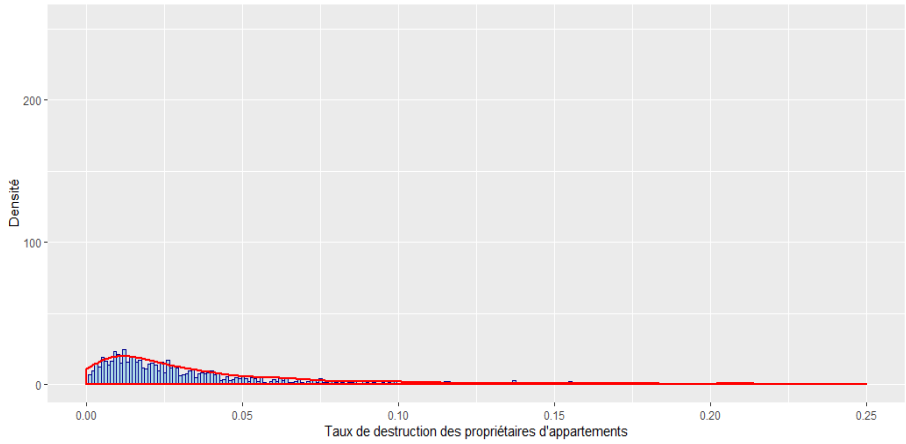


FIGURE 84 – Taux de destruction des propriétaires d'appartements - Distribution empirique (histogramme) et Simulations (rouge)

C.2 Locataires de maisons

	Paramètres	KS test p-value
Pareto	lambda=0,2728 ; sigma=0,0009368	1,474e-08
Gamma	shape=0,789 ; rate=10,16	0,0276
Bêta	shape1=0,646 ; shape2=6,710	0,0819
LGN	mu=-3,307 ; sigma=1,218	0,9818

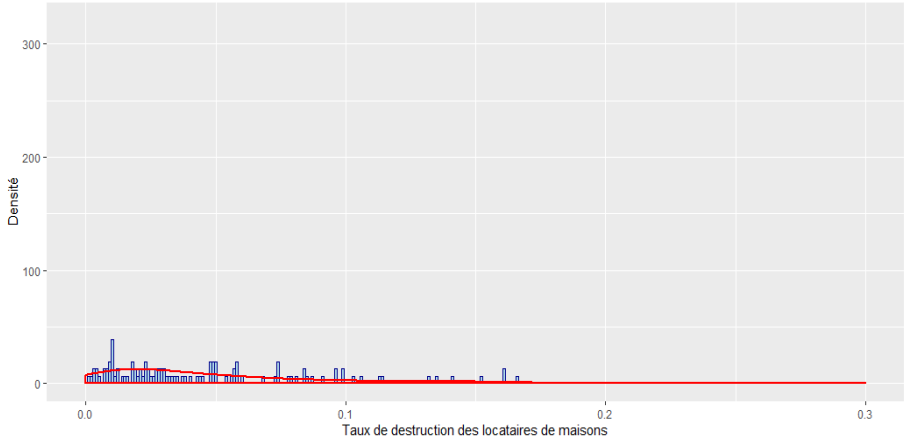


FIGURE 85 – Taux de destruction des locataires de maisons - Distribution empirique (histogramme) et Simulations (rouge)

C.3 Locataires d'appartements

	Paramètres	KS test p-value
Pareto	lambda=0,502 ; sigma=0,00306	0,1389
Gamma	shape=1,071 ; rate=28,078	0,9984
Bêta	shape1=1,017 ; shape2=25,297	0,7547
LGN	mu=-3,799 ; sigma=1,023	0,3127

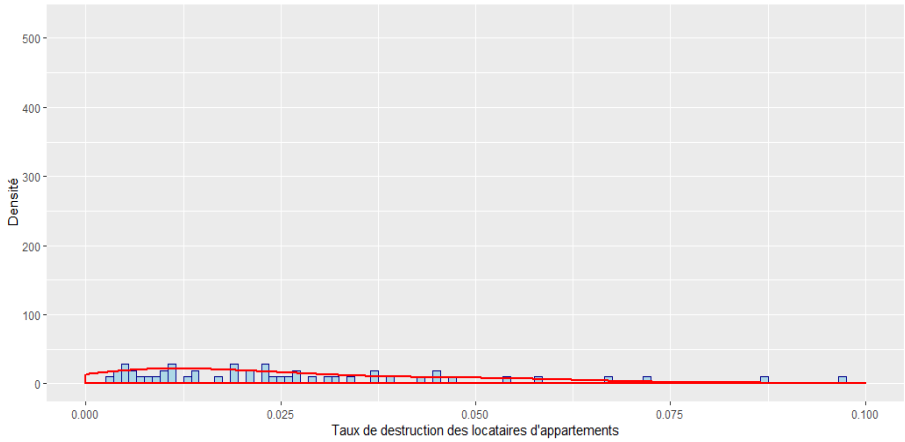


FIGURE 86 – Taux de destruction des locataires d'appartements - Distribution empirique (histogramme) et Simulations (rouge)

C.4 Automobiles

Intensité 1	Paramètres	KS test p-value
Pareto	lambda=0,1525 ; sigma=0,0001861	< 2,2e-16
Gamma	shape=0,721 ; rate=2,379	0,2008
Bêta	shape1=0,339 ; shape2=0,428	2,47e-12
LGN	mu=-2,026 ; sigma=1,624	0,00827

Intensité 2	Paramètres	KS test p-value
Pareto	lambda=0,186 ; sigma=0,000669	< 2,2e-16
Gamma	shape=0,761 ; rate=2,462	0,5996
Bêta	shape1=0,345 ; shape2=0,417	1,37e-08
LGN	mu=-1,957 ; sigma=1,608	0,0239

Intensité 3	Paramètres	KS test p-value
Pareto	lambda=0,178 ; sigma=0,0005595	< 2,2e-16
Gamma	shape=0,745 ; rate=2,226	0,0414
Bêta	shape1=0,353 ; shape2=0,413	3,16e-10
LGN	mu=-1,896 ; sigma=1,648	0,000253

Intensité 4	Paramètres	KS test p-value
Pareto	lambda=0,139 ; sigma=0,000122	< 2,2e-16
Gamma	shape=0,743 ; rate=2,102	< 2,2e-16
Bêta	shape1=0,338 ; shape2=0,355	< 2,2e-16
LGN	mu=-1,845 ; sigma=1,681	< 2,2e-16

Les distributions empiriques et modélisées sont affichées dans les figures suivantes :

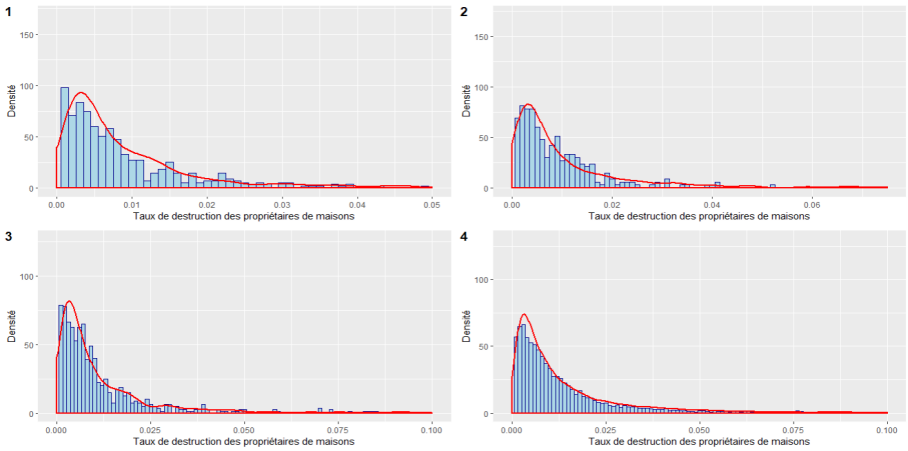


FIGURE 87 – Taux de destruction des automobiles - Distribution empirique (histogramme) et Simulations (rouge)

D Modélisation des températures maximales par série temporelles (méthode non conservée)

Les études précédemment effectuées ont montré que la survenance d'un évènement de grêle suppose généralement des températures observées élevées et des écarts importants entre les températures maximales du jour et de la veille. Il est donc nécessaire d'effectuer une modélisation de la température maximale quotidienne en France pour pouvoir en déduire l'apparition de grêle.
La localité très fine des évènements de grêle nous oblige à caractériser l'ensemble des stations de l'ECAD dans notre modélisation. Nous prenons donc les 252 séries historiques de températures maximales de 1950 à nos jours. Pour chacune de ces séries, plusieurs scénarios seront mis en oeuvre afin de simuler l'année suivante.

D.1 Présentation de l'historique des températures

Afin d'avoir une bonne visibilité, nous ne caractérisons qu'une seule station sur les 252 existantes. Cette démarche est appliquée à l'ensemble des stations.
Les températures maximales présentent une saisonnalité, conventionnelle pour les données de température.

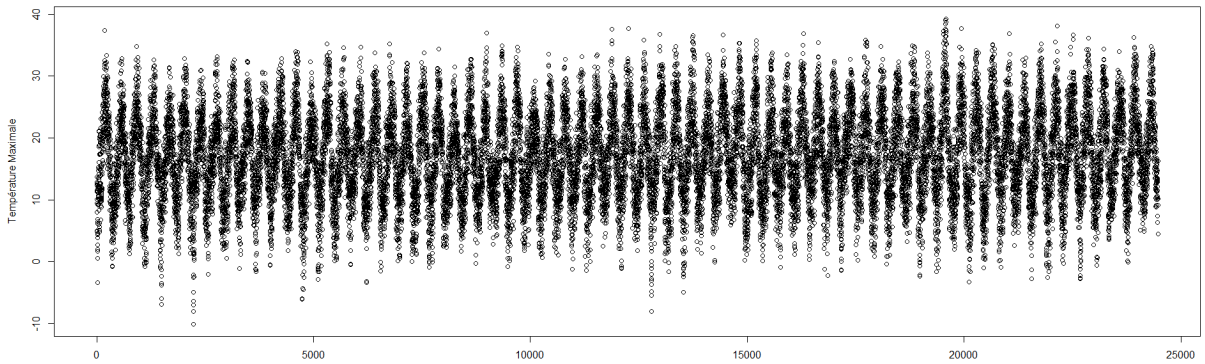


FIGURE 88 – températures Maximales journalières relevées sur la station 50

D.2 Séries temporelles et modèles ARMA

Avant de poursuivre, quelques rappels théoriques peuvent être nécessaires.

Définition : Stationnarité faible

Soit un processus temporel à valeurs réelles et en temps discret $Z_1, Z_2, ..., Z_t$. Il est dit stationnaire au sens faible (ou « de second ordre », ou « en covariance ») si :

- $E[Z_i] = \mu \quad \forall i = 1, ..., t$
- $Var[Z_i] = \sigma^2 \quad \forall i = 1, ..., t$
- $Cov[Z_i, Z_{i-k}] = f(k) = \rho_k \quad \forall i = 1, ..., t \quad \forall k = 1, ..., t$

Définition : Stationnarité forte

Soit un processus temporel à valeurs réelles et en temps discret $Z_1, Z_2, ..., Z_t$. Il est dit stationnaire au sens fort si pour toute fonction f mesurable :

$$f(Z_1, Z_2, ..., Z_t) \text{ et } f(Z_{1+k}, Z_{2+k}, ..., Z_{t+k}) \text{ ont la même loi.}$$

Définition : Bruit blanc

Une série est un bruit blanc faible si sa moyenne est nulle, sa variance constante et si la covariance de ses termes deux à deux est nulle.
Une série est un bruit blanc fort si ses termes sont indépendants et identiquement distribués (iid), de moyenne nulle et de variance constante.

Définition : ARMA

Soit ϵ_t un bruit blanc faible de variance σ^2. On dit qu'un processus (X_t) est un processus ARMA d'ordre (p,q), noté ARMA(p,q) si (X_t) est stationnaire et est la combinaison d'une composante Auto Régressive (AR) et d'une moyenne mobile (MA).

$$X_t = \sum_{i=1}^{p} \alpha_i X_{t-i} + \sum_{i=1}^{q} \theta_i \epsilon_{t-i} + \epsilon_t$$

avec respectivement ϵ_t bruit blanc, $\sum_{i=1}^{p} \alpha_i X_{t-i}$ un AR(p), $\sum_{i=1}^{q} \theta_i \epsilon_{t-i}$ un MA(q).

1) Stationnarisation

Pour travailler avec des modèles ARMA, nous devons avoir une série stationnaire, ce qui n'est pas le cas dans notre étude.
On peut obtenir une série stationnaire en différenciant ou en séparant la série selon l'approche Box jenkins (ou décomposition saisonnière). Nous retiendrons cette dernière méthode.

Nous allons dans un premier temps décomposer notre série de données en une réalisation de plusieurs phénomènes simultanés afin d'obtenir une série de la forme suivante :

$$X_t = M_t + S_t + Y_t$$

— M_t représente la **tendance** de la série. C'est l'évolution de la série au cours du temps, ce qui permet d'avoir une modélisation cohérente sur le long terme.

— S_t représente la **saisonnalité**, une composante se répétant à intervalle de temps égaux caractérisée par une forme constante.

— Y_t représente la **série résiduelle**, un processus stationnaire caractérisant la série sans les phénomènes précédents.

On observe cette décomposition grâce au logiciel R.
Pour cela, la tendance est estimée par une méthode de régression locale. La méthode consiste à déterminer localement un modèle de régression non paramétrique adapté aux données locales. La saisonnalité et la série résiduelle sont déduites de la série sans tendance $X_t - M_t$.

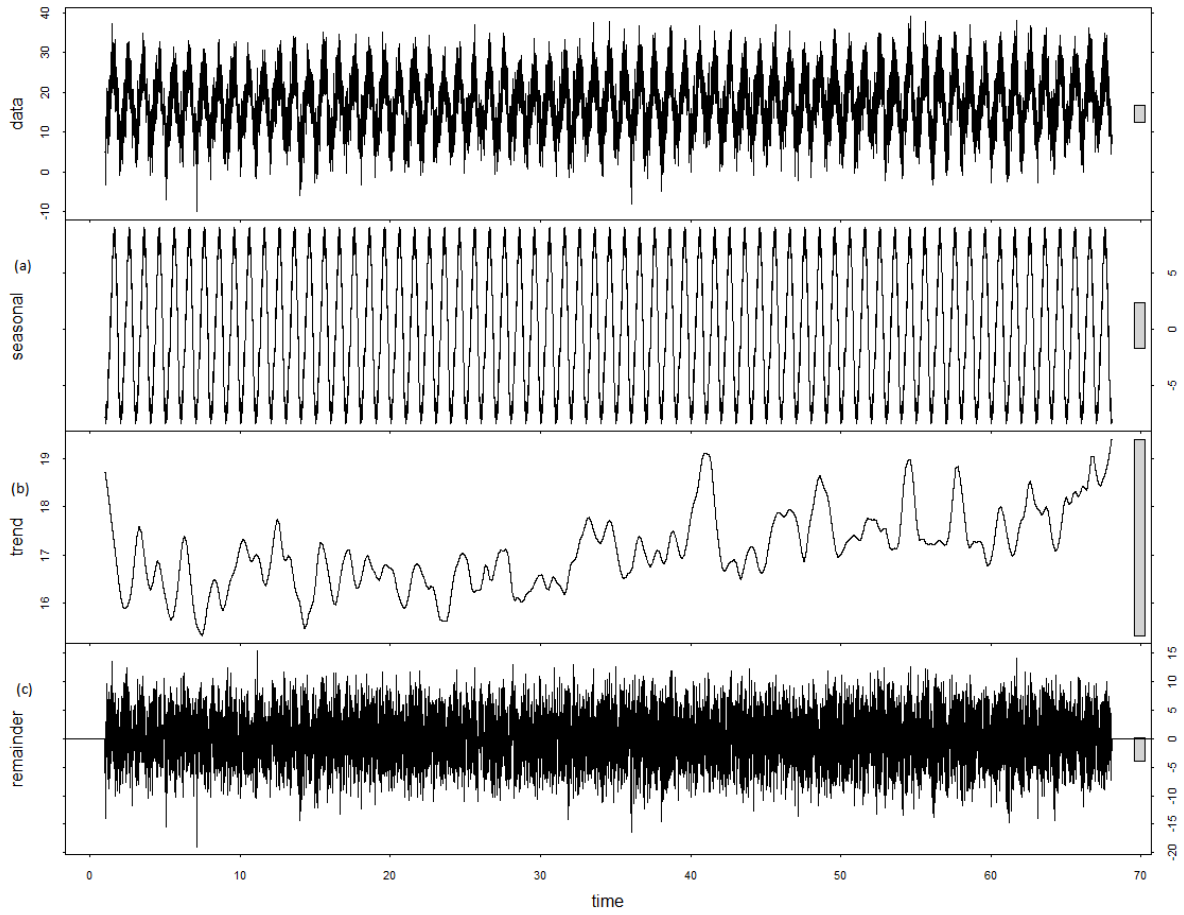
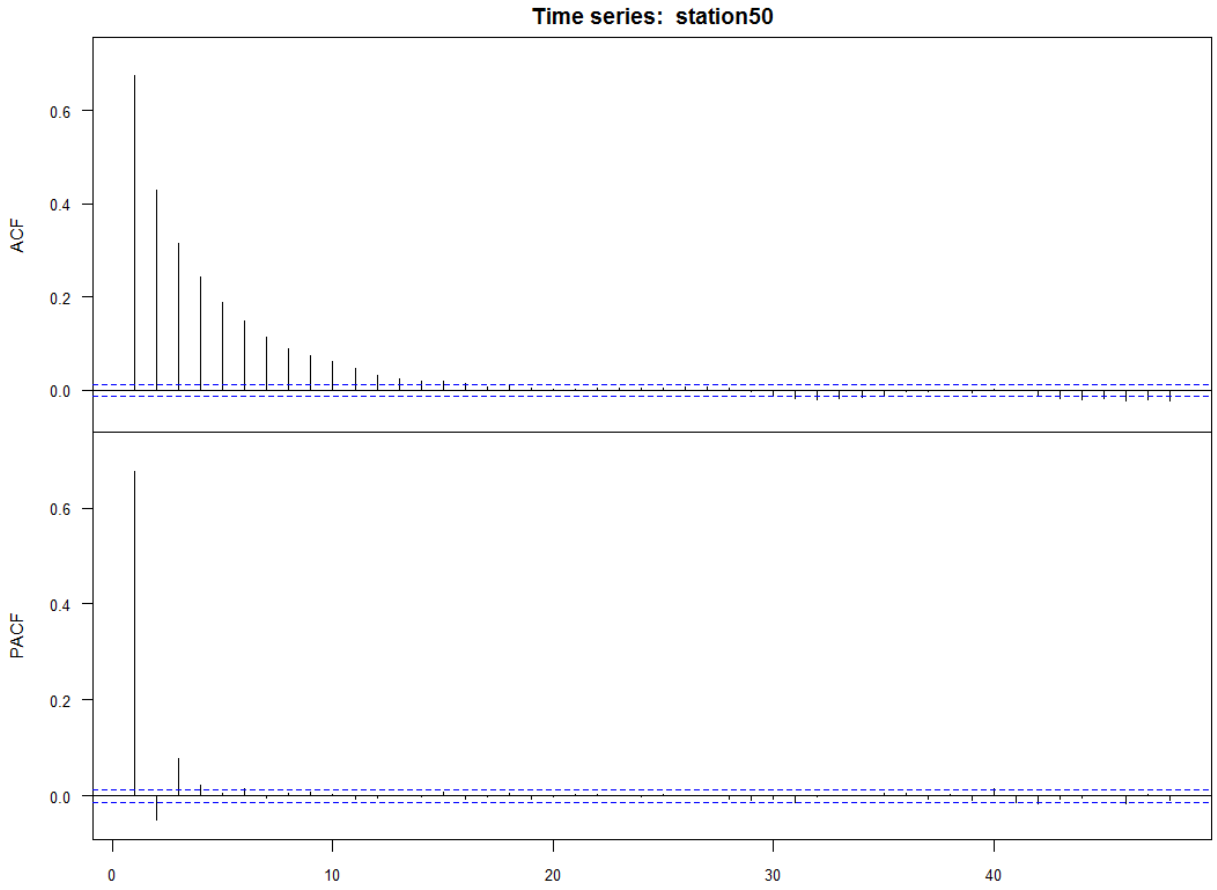


FIGURE 89 – Températures Maximales journalières relevées sur la station 50

On observe sur la série précédente une saisonnalité de période annuelle (figure a) et une tendance croissante (figure b) s'accentuant à partir des années 80. La série résiduelle (figure c) est dépourvue de tendance et de saisonnalité.

2) Modèles potentiels

Pour commencer, il faut déterminer les paramètres p (terme autorégressif) et q (terme de moyenne mobile) du modèle.
Les diagrammes des autocorrélations (ACF) et des autocorrélations partielles (PACF) permettent d'identifier une stationnarité, d'identifier un bruit blanc et de choisir les modèles adaptés à nos séries.
On observe le diagramme des autocorrélations (ACF) et le diagramme des autocorrélations partielles (PACF).



L'ACF décroît exponentiellement vers 0, ce qui témoigne d'une stationarité. Nous observons des autocorrélations simples et partielles explicitement différentes de 0, ce qui différencie notre série d'un bruit blanc.

l'ACF est caractérisé par un pic important au niveau du décalage 1 qui diminue après quelques décalages, ce qui témoigne de la présence d'un terme autorégressif dans les données. De plus, Le PACF témoigne de corrélations significatives au niveau du premier ou du deuxième décalage, suivies de corrélations non significatives, ce qui confirme la présence d'un terme autorégressif.

Le nombre de corrélations significatives indique l'ordre du terme autorégressif. Si le diagramme des autocorrélations partielles est nul à partir du (p+1)ème retard, alors un processus AR(p) doit être sélectionné. Pour cette série, un AR(2) ou un AR(3) semble adapté.

On estime nos coefficients à l'aide de la méthode du maximum de vraisemblance pour chacun des modèles présélectionnés.

Ces modèles doivent répondre à certains critères : les paramètres doivent être significatifs, les résidus doivent être Stationnaires (variance constante au cours du temps), Indépendants et Gaussiens.

Le test de Ljung-Box va permettre de vérifier si nous sommes bien en présence d'un bruit blanc. La statistique Ljung-Box permet de déterminer si les observations dans le temps sont aléatoires et indépendantes. Si les observations ne sont pas indépendantes, une observation peut être corrélée avec une

autre observation k unités de temps après, établissant ainsi une relation appelée auto-corrélation. L'hypothèse H0 indique que ϵ_t est un bruit blanc. Nous retenons tous les modèles dont les p-values ne nous permettent pas de rejeter l'hypothèse H0 (supérieures au seuil 0.05).
Nous pouvons également étudier à l'aide des QQ-plot la gaussianité des bruits blancs.

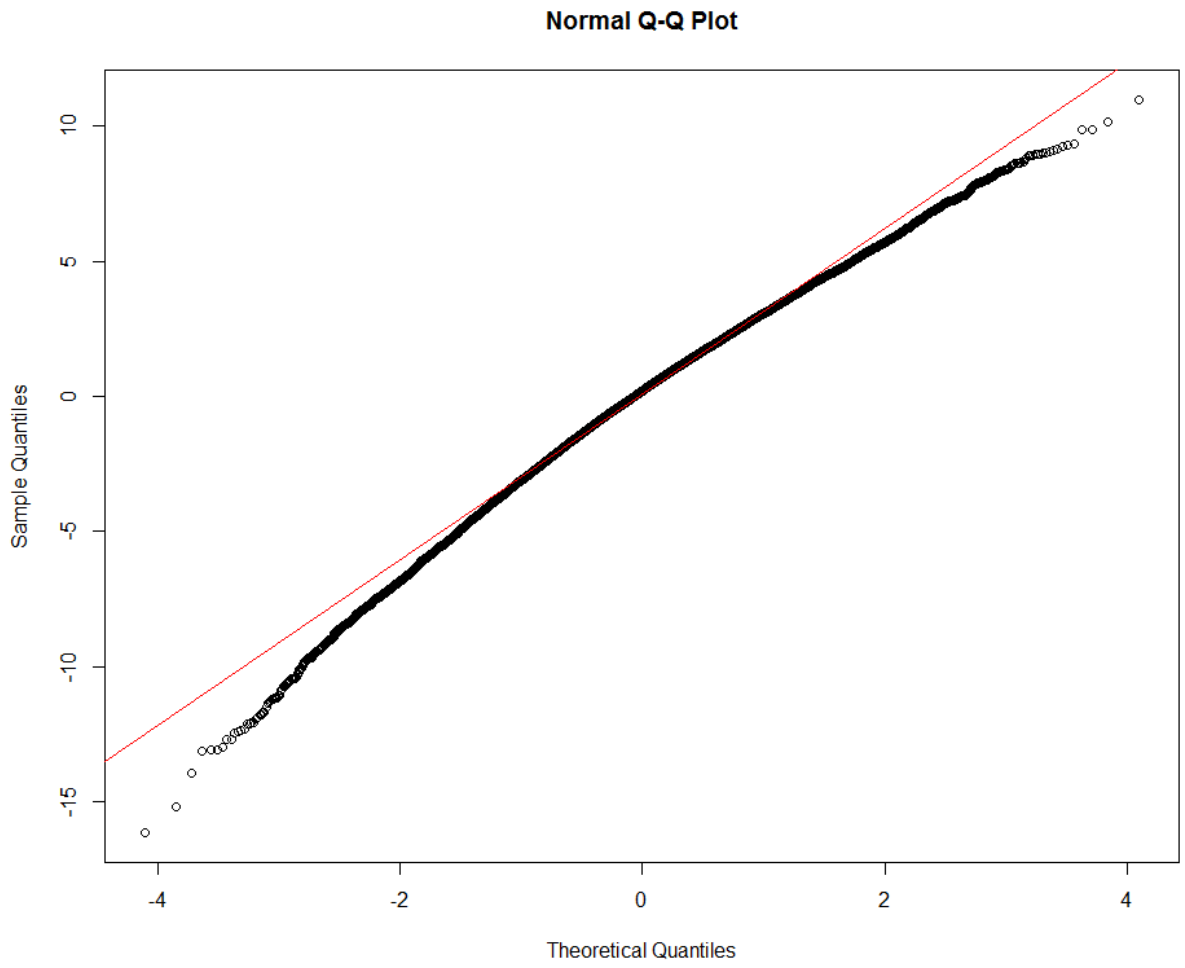


FIGURE 90 – Températures Maximales journalières relevées sur la station 50

3) <u>Choix du modèle :</u>

Définition : Le critère d'information d'Akaike s'écrit comme suit :

$$AIC = 2k - 2ln(L)$$

où k est le nombre de paramètres à estimer du modèle et L est le maximum de la fonction de vraisemblance du modèle.

Si l'on considère un ensemble de modèles candidats, le modèle choisi est celui qui aura l'AIC le plus faible. Ce critère repose donc sur un compromis entre la qualité de l'ajustement et la complexité du

modèle, en pénalisant les modèles ayant un grand nombre de paramètres, ce qui limite les effets de surajustement (augmenter le nombre de paramètre améliore nécessairement la qualité de l'ajustement).

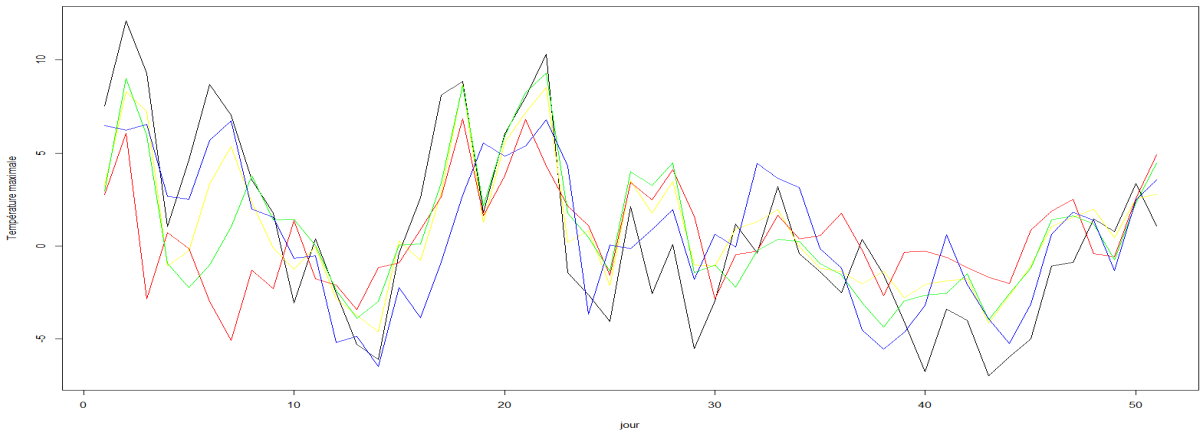
On choisit donc parmi tous les modèles satisfaisant l'ensemble des tests statistiques celui ayant l'AIC le plus faible pour chacune de nos séries de températures maximales.
On réalise cette procédure sur l'ensemble de nos séries de températures maximales pour obtenir un ensemble de modèles ARMA.

Cette partie nous a permis de déterminer une modélisation statistique de la température pour chacune des stations ECAD Françaises. La prochaine étape sera donc de simuler différents scénarios de température. Cependant, les températures ne peuvent pas être simulées indépendamment. En effet, la température est un phénomène homogène. Si un pic de température est observé sur une station, il est fortement probable que les stations situées aux alentours soient aussi affectées par ce pic.

Par conséquent, nous devons également modéliser la dépendance entre les stations pour pouvoir obtenir des prédictions cohérentes. Le seul terme non déterministe des modèles sélectionnés est la série résiduelle de l'ARMA. Nous allons donc modéliser les séries de résidus associées aux séries de température en prenant en compte la dépendance entre les stations par le biais de la théorie des copules.

D.3 Application des copules aux séries temporelles pour les températures maximales

On observe sur 5 stations choisies aléatoirement l'évolution des températures sur 50 jours.



La corrélation entre les températures est évidente. On a vu que nos résidus d'ARMA sont gaussiens, nos marginales sont donc connues. Au vu de la structure de nos copules empiriques, nous choisissons de modéliser notre dépendance par une copule elliptique. Nous sélectionnons la copule gaussienne afin de ne pas subir trop lourdement le grand nombre de dimensions à caractériser dans les calculs.

En raison du grand nombre de dimensions, nous portons notre choix sur une copule gaussienne. En effet, les copules fréquemment utilisées telles que les Gumbel ou Clayton n'ont qu'un seul paramètre. Elles ne sont donc pas adaptées au cas des grandes dimensions. Pour les mêmes raisons, nous n'utiliseront pas de test de Goodness of fit pour mesurer l'adéquation de nos copules à nos observations.

Nous simulons ensuite 10 000 scénarios à horizon 1 an pour chacune des 252 stations.

Nous observons ci-dessous les résidus empiriques et les résidus simulés par copule gaussienne entre 2 paires de stations choisies aléatoirement où la structure de dépendance semble conservée.

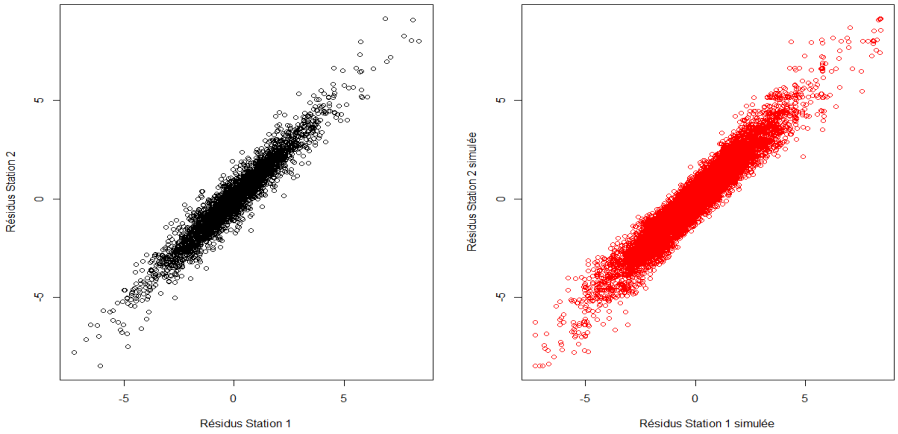


FIGURE 91 – Résidus de Températures Maximales : Station 1 vs Station 2 historiques et simulés

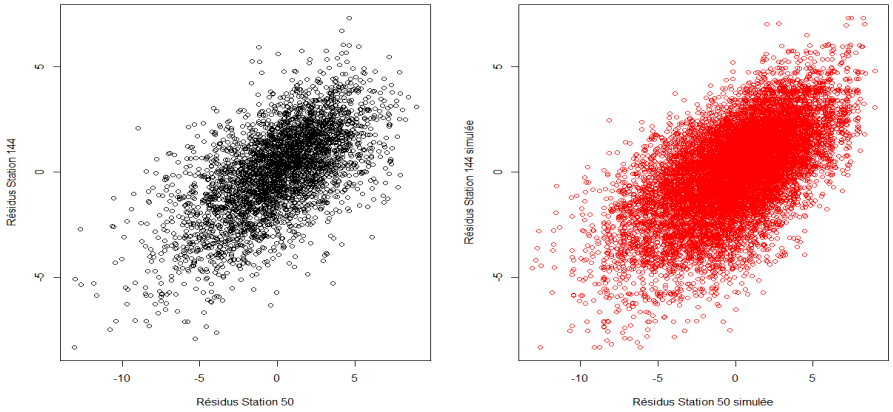


FIGURE 92 – Résidus de Températures Maximales : Station 50 vs Station 144 historiques et simulés

La dépendance entre les résidus semble correctement restituée. Cependant, la copule gaussienne ne permettra pas de restituer des dépendances extrêmes.

Les paramètres estimés par maximum de vraisemblance pour la copule gaussienne sur les résidus des 10 premières stations sont donnés ci-dessous :

	X1	**X2**	**X3**	**X4**	**X5**	**X6**	**X7**	**X8**	**X9**	**X10**
X1	1	0,95	0,90	0,90	0,82	0,86	0,83	0,78	0,75	0,66
X2		1	0,90	0,98	0,87	0,88	0,92	0,79	0,82	0,63
X3			1	0,90	0,80	0,96	0,88	0,90	0,82	0,78
X4				1	0,91	0,92	0,97	0,85	0,90	0,66
X5					1	0,83	0,92	0,80	0,89	0,62
X6						1	0,94	0,97	0,92	0,83
X7							1	0,90	0,97	0,70
X8								1	0,93	0,88
X9									1	0,72
X10										1

Les coefficients observées confirment la forte correlation entre les stations les plus proches.

En regardant la série dans son ensemble, une perte de qualité est également observée sur la modélisation des dépendances.

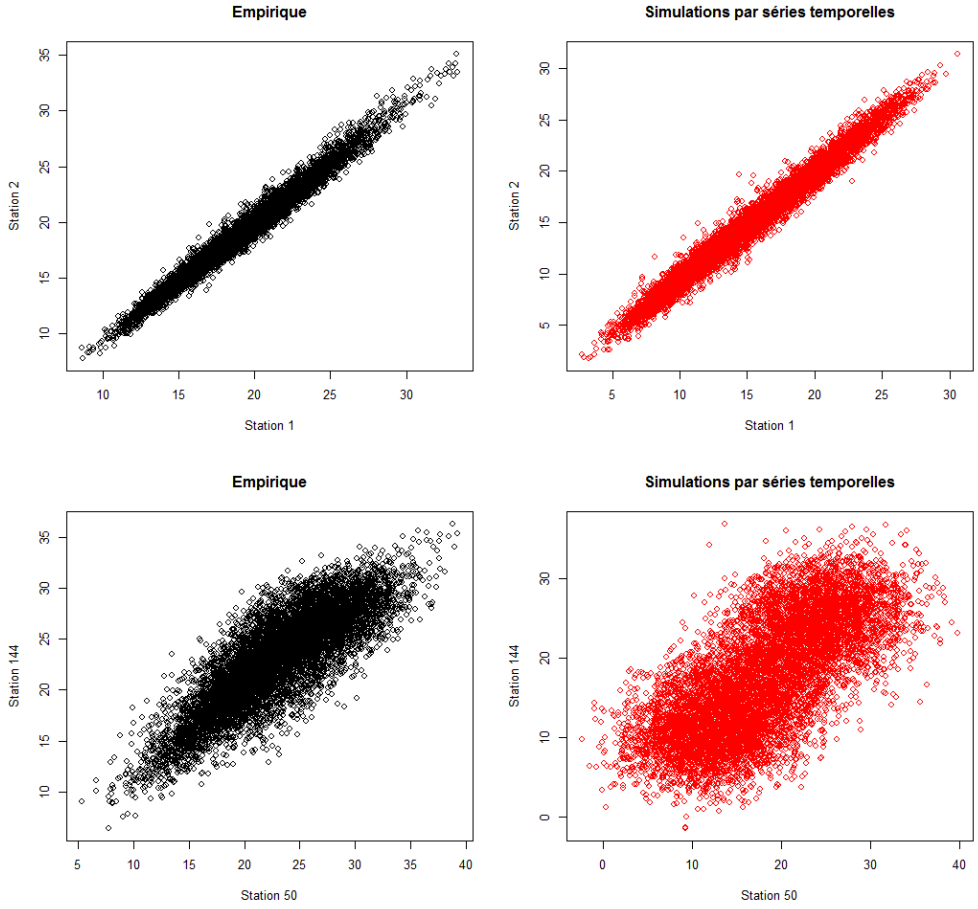


FIGURE 93 – Températures maximales empiriques et simulées par Séries temporelles

La fonction de Kendall cette témoigne de l'importance des écarts avec les observations historiques.

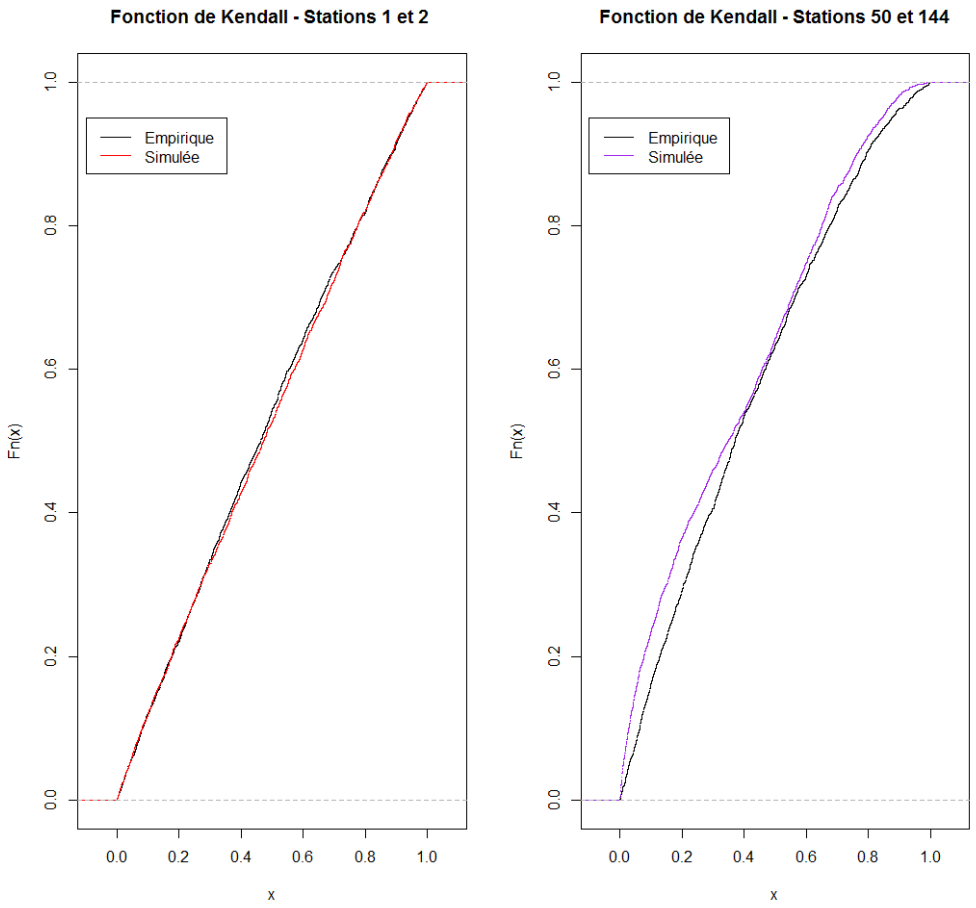


FIGURE 94 – Fonction de Kendall

Conclusion :

L'approche par séries temporelles présente de nombreux avantages tels que la présence d'une **tendance** et d'une **saisonnalité** mais également une **cohérences des valeurs au cours du temps**. Cette cohérence est par exemple nécessaire pour étudier les différences de températures entre 2 jours consécutifs (variable hautement significative dans notre modèle). Cependant, l'incapacité à restituer des dépendances cohérentes entre les stations et à corréler ces séries aux autres variables explicatives sont un handicap trop important pour que cette méthode soit conservée dans ce modèle.

En effet, dans le tableau suivant le test du Chi-deux démontre explicitement que les variables climatiques significatives ne sont pas indépendantes.

L'hypothèse H_0 d'indépendance entre les températures Maximales et la pression ainsi qu'entre la pression et les précipitations est rejetée.

	p-value
Températures Maximales/Pression	< 2,2e-16
Températures Maximales/Précipitations	0,235
Pression/Précipitations	< 2,2e-16

La mise en place d'une méthodologie plus large prenant en compte les dépendances entre les variables est donc nécessaire à la bonne simulation de survenance.

Références

[1] Marco Hohl. Relationship between Hailfall intensity and hail damage on ground. 2001

[2] F.Vinet. Le risque-grêle en France étude géographique. 1998

[3] Mohr, Kunz. Hail potential in Europe based on a regional climate model hindcast. 2015

[4] Fédération française de l'Assurance. La garantie tempête grêle neige. 2016

[5] Anelfa. La grêle. http ://www.anelfa.asso.fr.

[6] E. Di Bernardino. Théorie des copules. Support de cours ISUP. 2019

[7] Théorie des valeurs extrêmes. Support de cours ISUP, M. Kratz 2019.

[8] AIR. 2018. Crop Hail Model for the US

[9] B. BAILLEUL. Quelques données pour maitriser des engagements évènements naturels en France Allianz France. Mars 2013

[10] Olivia Martius, A. Hering, M. Kunz, A. Manzato, S. Mohr, L. Nisi, and S. Trefalt. Challenges and recent advances in hail research. Mars 2018

[11] Météo-France. LA GRÊLE

[12] Keraunos. Comprendre les orages géographique. 1998

[13] Jianming Yina, Yoshiaki Oganeb, Katsumasa Jinnaic. Modeling hail risk in the contiguous United States for insurance loss estimation. 2007.

[14] Rachid ABBOUBI. MODELISATION AROME Diagnostiques et Amélioration de la prévision de la grêle. 2010.

[15] Ubyrisk. Grêle Warning - Rapport annuel sur les chutes de grêle survenues en France et en Europe en 2018 et depuis 2000.

[16] Christine LAC. Peut-on prévoir les orages de grêle? - Centre national de recherches météorologiques. Novembre 2014.

[17] Martin Otto, Stefanie Busch. Modélisation le risque de grêle dans le Assurance multirisques allemande. 2009.

[18] Journal officiel de l'Union européenne. RÈGLEMENT DÉLÉGUÉ (UE) 2015/35 DE LA COMMISSION du 10 octobre 2014 complétant la directive 2009/138/CE du Parlement européen et du Conseil sur l'accès aux activités de l'assurance et de la réassurance et leur exercice (solvabilité II).

[19] Nathalie BEDI. Modélisation du risque de tempête en Frane Métropolitaine. 2018.

[20] Jana Friederike SCHULTE. Modélisation du risque subsidence en France métropolitaine. 2017.

[21] AURÉLIE NICOLAS. Risques climatiques : la grêle, l'aléa extrême. 2017.

[22] Michael Kunz, Heinz Jürgen Punge, Elody Fluck, Manuel Schmidberger, Susanna Mohr, David Piper, and Marc Puskeiler. Hail hazard and risk assessment in Europe and the relation to orographic and atmospheric characteristics. 2016.

[23] Eberhard Faust. Severe thunderstorms in Europe. 2016.

[24] Assurance Non Vie. Support de cours ISUP, Maud THOMAS. 2018.

[25] Apprentissage statistique. Support de cours ISUP, Claire BOYER. 2019.

[26] Statistique inférentielle. Support de cours ISUP, Olivier LOPEZ. 2016.

Table des figures