Mémoire présenté devant le jury de l'EURIA en vue de l'obtention du Diplôme d'Actuaire EURIA et de l'admission à l'Institut des Actuaires

le 18 septembre 2024

Par : Paule FOMEN
Titre : Analyse du taux de surprime du régime des Catastrophes Naturelles et impact des tempêtes sur la tarification en assurance MRH

Confidentialité : Non

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

Membre présent du jury de l'Institut des Actuaires :
Filipe GOMES
Romain LAILY
Samuel STOCKSIEKER
Signature :

Entreprise :
Actuelia
Signature :

Membres présents du jury de l'EURIA :
Franck VERMET

Directeur de mémoire en entreprise :
Mordehaï ROOS
Signature :

Invité :

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 :

EURIA
EURo Institut d'Actuariat

6, avenue le Gorgeu
CS 93837
29238 Brest Cedex 3

T +33 (0)2 98 01 66 55
euria@univ-brest.fr

Remerciements

Je souhaite tout d'abord exprimer ma profonde gratitude à David FITOUCHI, Frank BOUKOBZA, Louis-Anselme DE LAMAZE et Camille BLANC-VANNET, associé.es au sein du cabinet Actuelia, pour m'avoir accueillie chaleureusement au sein de leur équipe.

Mes sincères remerciements vont à Maryam HARIKI et Mordehaï ROOS, mes tuteurs en entreprise, ainsi qu'à Franck VERMET, mon tuteur académique, pour leur disponibilité, leurs précieux conseils, et leurs relectures attentives tout au long de la rédaction de ce mémoire. Leur guidance a été d'une grande aide pour la réalisation de ce travail.

Je suis également reconnaissante envers l'ensemble des consultants du cabinet Actuelia pour leur bonne humeur contagieuse et leurs encouragements quotidiens, qui ont rendu cette expérience plus belle.

Je tiens aussi à remercier le corps éducatif de l'EURIA et de l'ESILV pour la qualité des enseignements dispensés, qui ont fourni une base solide de connaissances et de compétences nécessaires à la réalisation de ce mémoire.

Je remercie également Boris NOUMEDEM, actuaire externe au cabinet Actuelia, pour avoir relu ce mémoire, apporté un regard critique et partagé des idées et pistes de réflexion précieuses qui ont grandement enrichi ce travail.

Enfin, je remercie du fond du cœur ma famille et mes ami.es pour leur soutien inconditionnel et leurs encouragements constants. Leur présence et leur soutien ont été des piliers essentiels tout au long de la réalisation de ce mémoire.

Liste des abréviations

ACP Analyse en Composantes Principale

AIC Akaike's Information Criterion

ARMA Modèles autoregressifs et moyenne mobile

BIC Bayesian Information Criterion

CAH Classification Ascendante Hiérarchique

Cat-Nat Catastrophes Naturelles

CCR Caisse Centrale de Réassurance

GASPAR Gestion Assistée des Procédures Administratives relatives aux Risques

GEV Generalized Extreme Value

GIEC Groupe d'Experts Intergouvernemental sur l'Evolution du climat

GLM Modèles Linéaires Généralisés

GPD Loi de Pareto Généralisée

MAE Mean Absolute Error

MRH Multirisques Habitation

PIB Produit Intérieur Brut

POT Peak Over Threshold

RGA Retrait-Gonflement des sols Argileux

RMSE Root Mean Squared Error

TGN Tempête, Grêle et Neige

Résumé

Dans un contexte où les effets du changement climatique s'intensifient, les assureurs font face à des défis croissants. Le nombre de catastrophes naturelles a triplé au cours des trente dernières années, entraînant une augmentation des coûts d'environ 5,7 % par an. En France, les catastrophes naturelles sont réassurées par l'État via la CCR, dans le cadre du régime d'indemnisation des catastrophes naturelles établi par la Loi du 13 juillet 1982. La viabilité de ce régime est challengée par la fréquence croissante des catastrophes naturelles. Entre 2016 et 2022, le ratio de sinistralité de ce régime était en moyenne de 126 %, contre 95 % en moyenne depuis 2010. En réponse à cette problématique, le taux de surprime payé à la CCR par les assureurs passera dès l'année 2025 de :

— 6 % à 9 % de la prime afférente aux garanties vol et incendie pour les véhicules à moteur
— 12 % à 20 % de la prime afférente aux garanties dommages pour les biens autres que les véhicules à moteur.

Ce mémoire vise à évaluer l'impact de l'augmentation de la fréquence et du coût des événements climatiques d'inondation, de sécheresse et de tempêtes sur les portefeuilles des assureurs. L'objectif est de challenger la suffisance et la pérennité des nouveaux taux de surprime du régime Cat-Nat, sur un portefeuille MRH. De plus, l'impact de la dérive de sinistralité liée aux tempêtes, non-couvertes par ce régime, sera également analysé en rapport à la prime payée par les assurés.

Pour ce faire, des modèles de machine learning sont mis en œuvre pour modéliser la dérive de sinistralité liée aux risques d'inondation et de sécheresse. Cette modélisation est réalisée à partir de données météorologiques de température, humidité, précipitations et pression issues de la base de données SYNOP mise à disposition en open data par Météo France. Une dérive de sinistralité pour le risque de tempête est également calibrée en appliquant la théorie des valeurs extrêmes. Ces dérives sont ensuite prises en compte dans la tarification d'un portefeuille Multirisque Habitation afin de déterminer leurs impacts sur les niveaux de primes. Enfin, la viabilité des nouveaux taux de surprime sera discutée en fonction des dérives modélisées.

Mots clefs: Risque climatique, Open data, Séries temporelles, Forêt aléatoire, Théorie des valeurs extrêmes, Copules, GLM, Tarification

Abstract

As the effects of climate change intensify, insurers are encountering increasing challenges. The number of natural disasters has tripled over the past thirty years, leading to an annual cost increase of approximately 5.7 %. In France, natural disasters are reinsured by the State via the CCR, under the natural disaster compensation regime established by the Law of July 13, 1982. The viability of this regime is threatened by the increasing frequency of natural disasters. Between 2016 and 2022, the claims-to-premium ratio of this regime averaged 126 %, compared to an average of 95 % since 2010. To address this issue, the percentage of the premium paid to the CCR by insurers will increase from 2025 onwards :

— From 6 % to 9 % of the premium related to theft and fire coverage for motor vehicles
— From 12 % to 20 % of the premium related to damage coverage for property other than motor vehicles

This thesis aims to assess the impact of the increased frequency and cost of flood, drought, and storm events on insurers' portfolios. The objective is to evaluate the adequacy and sustainability of the new premium rates of the natural disasters regime on a home insurance portfolio. Additionally, the impact of the claims drift related to storms, which are not covered by this regime, will also be determined in regards to the premium paid by policyholders.

To achieve this, machine learning models are implemented to model the claims drift associated with flood and drought risks. This modeling is based on meteorological data such as temperature, humidity, precipitation and pressure from the SYNOP database made available as open data by Météo France. A claims drift for storm risk is also calibrated using extreme value theory. These drifts are the taken into consideration in the pricing of a home insurance portfolio to determine their impact on premium levels. Finally, the viability of the new premium rates will be discussed based on the modeled drifts.

Keywords: Climate risk, Open data, Time series, Random forest, Extreme Value Theory, Copulas, GLM, Pricing

Note de Synthèse

Contexte et problématique

Au cours des trente dernières années, le nombre de catastrophes naturelles a triplé, entraînant une augmentation annuelle des coûts d'environ 5,7 %. Cette hausse, tant en fréquence qu'en coût, a des répercussions significatives, notamment pour les assureurs. En France, les catastrophes naturelles sont réassurées par l'État via la Caisse Centrale de Réassurance (CCR), conformément au régime d'indemnisation établi par la Loi du 13 juillet 1982. Ce régime couvre les catastrophes naturelles telles que les inondations, la sécheresse, les mouvements de terrain, les ouragans, et les tornades. Toutefois, les tempêtes, la grêle et la neige ne sont pas incluses dans ce régime, car elles sont considérées comme assurables par les assureurs.

Entre 2016 et 2022, le ratio de sinistralité de ce régime a atteint en moyenne 126 %, contre 95 % en moyenne depuis 2010, ce qui remet en question sa viabilité face à l'augmentation des catastrophes naturelles. Pour faire face à cette problématique, le taux de surprime payé à la CCR par les assureurs augmentera dès 2025, passant de :

- **6 % à 9 %** de la prime afférente aux garanties vol et incendie pour les véhicules à moteur,
- **12 % à 20 %** de la prime afférente aux garanties dommages pour les biens autres que les véhicules à moteur.

Ce mémoire vise à challenger la suffisance et la viabilité des nouveaux taux de surprime appliqués à un portefeuille multirisque habitation (MRH). Il s'agit également d'analyser l'impact des sinistres liés aux tempêtes futures, qui ne sont pas couvertes par le régime des catastrophes naturelles, sur les primes payées par les assureurs. Pour cela, une projection de la dérive de sinistralité à l'horizon 2028 est établie pour les risques liés aux inondations, à la sécheresse et aux tempêtes. Ces dérives seront ensuite prises en compte dans la tarification d'un portefeuille MRH afin d'en évaluer les impacts.

Création de la base de données

Pour déterminer la dérive de fréquence des inondations, de la sécheresse et des tempêtes, il était nécessaire de disposer de données observées des paramètres météorologiques

ainsi que celle sur l'occurrence des catastrophes naturelles. À cette fin, des données issues de sources « open data » ont été utilisées, notamment les bases **SYNOP** et **GASPAR**.

La base SYNOP, mise à disposition par Météo France, contient des observations journalières recueillies par stations météorologiques de la **température, l'humidité, la vitesse du vent, la vitesse des rafales, la pression atmosphérique et la hauteur des précipitations**, qui sont essentielles pour les différents modèles. La base GASPAR (Gestion Assistée des Procédures Administratives relatives aux Risques) quant à elle, recense, pour chaque commune, le nombre de catastrophes naturelles reconnues par un état de catastrophe naturelle, publié au Journal Officiel. Étant donné l'augmentation de la sinistralité liée aux événements climatiques depuis 2016, un historique de données couvrant la période de **2016 à 2023** a été retenu. L'analyse a été limitée à la **France Métropolitaine**.

Enfin, une base de données géographiques comportant les coordonnées de chaque département en France a également été utilisée. Les différentes bases : SYNOP, GASPAR et Géographique ont été fusionnées grâce à la clé commune : le département, afin d'obtenir une base appelée « base climatique ». Cette base contient, pour chaque jour entre le 1er janvier 2016 et le 31 décembre 2023, et pour chaque département de la France Métropolitaine, les paramètres météorologiques observés, indispensables pour les modèles qui seront construits par la suite. Les valeurs manquantes ont été remplacées par la **moyenne mensuelle**.

Cette base permettra de déterminer la dérive de fréquence pour les inondations, la sécheresse et les tempêtes. Le graphique suivant illustre le processus de construction des bases et d'obtention de la dérive.

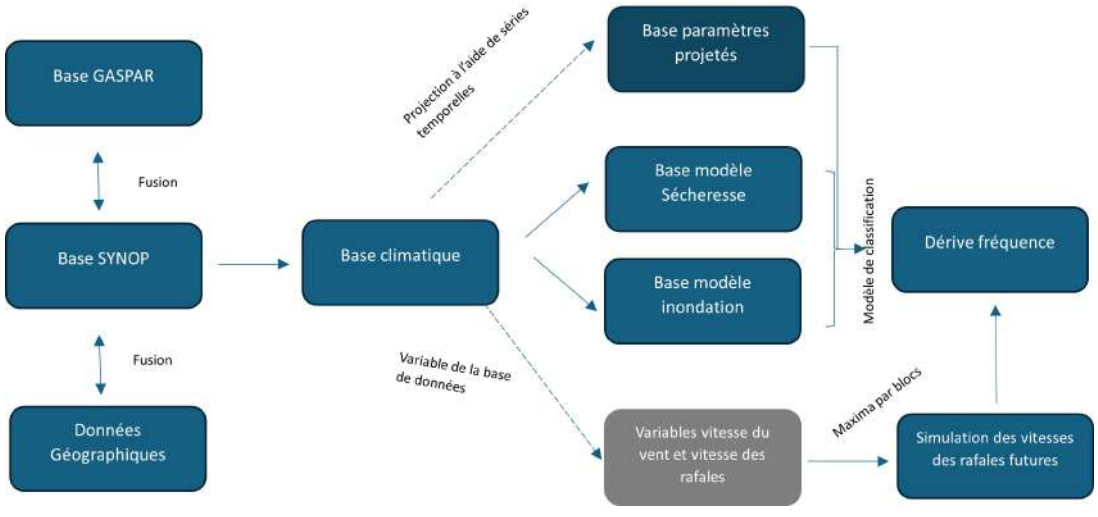


FIGURE 1 – Méthodologie d'obtention des dérives

Obtention des dérives

I- Dérive de fréquence des risques inondations et sécheresse

1- Séries temporelles

Pour estimer le nombre d'arrêtés de catastrophes naturelles liés aux inondations et à la sécheresse, il est crucial de projeter les paramètres météorologiques dans le futur. Dans cette optique, la base climatique a été segmentée en **cinq** zones de risques homogènes, grâce à l'Analyse en Composantes Principale (ACP) et la Classification Ascendante Hiérarchique (CAH) afin de minimiser le coût de la modélisation.

Les données ont été scindées en deux ensembles :
- une base d'entraînement comprenant les observations de **2016 à 2022**, et
- une base de test contenant les observations de **2023**.

Pour chaque zone, la modélisation de la température, de l'humidité, des précipitations et de la pression a été réalisée à partir des moyennes observées. Cette modélisation s'appuie sur des **séries temporelles**. Les **composantes non-stationnaires**, à savoir la tendance et la saisonnalité, ont été modélisées par un **modèle paramétrique**. La tendance a été ajustée par un polynôme de degré r, tandis que la saisonnalité a été modélisée par un polynôme trigonométrique. Les résultats de la modélisation de la température sont présentés dans le tableau suivant :

	Zone 1	Zone 2	Zone 3	Zone 4	Zone 5
R^2	75,78 %	81,04 %	77,21 %	76,72 %	90,83 %
MAE	1,40	1,33	1,17	1,44	0,84
RMSE	1,80	1,73	1,52	1,85	1,11

TABLE 1 – Evaluation de la qualité d'ajustement des composantes non-stationnaires de la température sur la base d'entraînement

Les **composantes stationnaires**, c'est-à-dire les résidus, ont été modélisés à l'aide du modèle **ARMA**, comme illustré dans le tableau suivant :

	Zone 1	Zone 2	Zone 3	Zone 4	Zone 5
Modèle ajusté	ARMA(1,1)	ARMA(3,0)	ARMA(1,1)	ARMA(4,0)	ARMA(1,3)

TABLE 2 – Modèles ARMA ajustés par zone pour la température

L'ajustement de la tendance, de la saisonnalité et des résidus pour chacun des paramètres a permis de simuler 10 000 trajectoires de chacun de ces paramètres pour les années 2024 à 2028.

Les résultats de ces simulations pour la température sont synthétisés dans le tableau suivant :

	Zone 1	Zone 2	Zone 3	Zone 4	Zone 5
Température moyenne 2016-2023 (°C)	13,17	13,93	14,42	14,76	17,35
Température moyenne 2024-2028 (°C)	14,00	14,68	15,19	15,44	17,59
Ecart en %	5,93	5,11	5,06	4,42	1,39

TABLE 3 – Comparaison des températures moyennes observées et celle prédites entre 2024 et 2028

2- Modèles de classification

Les bases de données : *base modèle sécheresse* et *base modèle inondation*, ont ensuite été utilisées pour ajuster un modèle de classification. Ces bases incluent les variables suivantes : département, date d'observation, mois d'observation, température, humidité, pression, précipitation sur les dernières 24 heures, et une variable sinistre (valant 1 en cas de déclaration d'état de catastrophe naturelle et 0 sinon). Les données ont été séparées en deux : **75 %** des données pour entrainer le modèle et **25 %** pour le valider.

La modélisation a été réalisée à l'aide de la **régression logistique**, **des arbres de décision** et de la **forêt aléatoire**, en utilisant un historique de 2016 à 2023 pour le modèle d'inondation, et de 2016 à 2022 pour le modèle de sécheresse. Cette différence dans les périodes d'historique vise à éviter tout biais dû au délai de parution des arrêtés de catastrophes naturelles pour la sécheresse (souvent entre un et deux ans). Le modèle de **forêt aléatoire** a été retenu en raison de sa meilleure capacité à détecter les sinistres comparé à la régression logistique.
Les matrices de confusion associées à ces modèles sont présentées ci-après.

	Prédit Négatif	Prédit Positif
Réel Négatif	83 735	27
Réel Positif	45	6 332

TABLE 4 – Matrice de confusion pour le risque de sécheresse : base de test

	Prédit Négatif	Prédit Positif
Réel Négatif	83 251	387
Réel Positif	133	4 105

TABLE 5 – Matrice de confusion pour le risque d'inondation : base de test

La base de données des paramètres simulés a ensuite été utilisée comme entrée pour le modèle de classification afin d'estimer le nombre d'arrêtés de catastrophes naturelles

entre 2024 et 2028 pour les inondations, et de 2023 à 2028 pour la sécheresse. Cette approche a permis de calculer une dérive de sinistralité future, estimée à **9 %** pour le risque de sécheresse et à **23 %** pour le risque d'inondation.

II- Dérive de fréquence du risque tempête

Dans cette modélisation, les variables d'intérêt sont la **vitesse du vent** et la **vitesse des rafales**. Pour estimer le nombre de tempêtes futures, il est crucial de simuler les vitesses de rafales à venir, car les assureurs n'indemnisent les sinistrés que lorsque ces vitesses dépassent **100 km/h**. La modélisation a été effectuée sur la vitesse du vent en utilisant la **théorie des valeurs extrêmes**, plus précisément la méthode des **maxima par blocs** avec des blocs mensuels. Ce choix s'explique par la qualité des données disponibles : la vitesse du vent présentait un taux de valeurs manquantes nettement inférieur (0,7 %) à celui des données sur les rafales (18,6 %). Les paramètres obtenus à partir de cette modélisation permettent ensuite de simuler les vitesses de vent futures. Ces vitesses sont transformées en vitesses de rafales via une régression linéaire.

Pour diminuer le coût de la modélisation, les départements ont été regroupés en cinq zones de risques homogènes. Une distribution GEV (Generalized Extreme Value) a été ajustée pour chaque zone afin de déterminer la distribution la plus adaptée à la modélisation de la vitesse du vent, en se basant sur le paramètre de forme. Les zones 1, 2, 3 et 4 ont montré un paramètre de forme ayant « zéro » dans leur intervalle de confiance, ce qui indique que la distribution de Gumbel est adéquate pour ces zones. En revanche, pour la zone 5, qui représente le sud-est de la France, un modèle GEV complet est nécessaire, comme confirmé par une analyse ANOVA.

	Zone 1	Zone 2	Zone 3	Zone 4	Zone 5
Paramètre de position	44,25	44,52	43,27	58,48	63,84
Ecart-type	0,83	0,88	0,75	1,28	1,04
Paramètre d'échelle	7,76	8,17	7,02	11,97	9,65
Ecart-type	0,61	0,63	0,55	0,95	0,73
Quantile associé à la période de retour	76,81	76,02	73,61	117,74	92,45
Résultat ANOVA	0,53	0,19	0,63	0,47	0,001234

TABLE 6 – Paramètres de la loi Gumbel ajusté

La simulation des vitesses maximales du vent atteintes chaque mois à l'aide de ces paramètres a révélé l'absence de prise en compte de la dépendance entre les différentes zones géographiques. Pour remédier à cela, une copule de **Frank** a été ajustée afin de modéliser cette dépendance. En intégrant à la fois les paramètres spécifiques à chaque zone et le paramètre de la copule de Frank, il a été possible de simuler les vitesses maximales du vent mensuelles pour la période 2024-2028 de manière plus adéquate.

Une régression linéaire a été appliquée pour transformer les simulations de la vitesse du vent en vitesses des rafales.

$$\text{vitesse\_des\_rafales} = 1{,}50 \times \text{vitesse\_du\_vent} + 5{,}94$$

L'équation précédente, validée par des tests appropriés, a été utilisée pour effectuer cette transformation.
Les métriques suivantes ont été obtenues pour ce modèle :

	R^2	RMSE	MAE
Base d'apprentissage	84,52 %	6,51	4,71
Base de test	84,29 %	6,58	4,70

TABLE 7 – Evaluation du modèle de régression

Cette méthode a permis de calculer une dérive de fréquence des tempêtes, estimée à **5 %** pour la période 2024-2028.

III- Dérives de coût :

La dérive de coût a été estimée à partir d'un benchmark d'articles publiés, car les données relatives aux coûts sont considérées comme confidentielles et ne sont pas disponibles en open data. Par principe de prudence, une croissance exponentielle a été utilisée pour estimer la dérive des coûts. La dérive estimée est la suivante :

Risque	Horizon	Évolution du coût
Inondation	2028	[3,6 ; 7,9[%
Sécheresse	2028	[2,6 ; 4,7[%
Tempête	2028	[0,0 ; 5,3[%

TABLE 8 – Hypothèse de dérive du coût des sinistres à horizon 2028

Mise en place de la tarification

Une base de données regroupant les informations sur le portefeuille et les sinistres d'un assureur MRH entre **2016 et 2022** a été utilisée pour évaluer l'impact des dérives estimées sur le portefeuille.

1- Suffisance et viabilité du nouveau taux de surprime du régime des catastrophes naturelles

Actuellement, le taux de surprime Cat-Nat est calculé comme **12 %** de la prime des garanties dommages. La première étape consiste à déterminer, en fonction de la sinistralité du portefeuille, la proportion que représente la prime pure des catastrophes naturelles

par rapport à la prime dommages, dans le cas où cette prime serait calculée par l'assureur.

Pour cela, la **prime pure dommages** a été modélisée à l'aide de **modèles linéaires généralisés (GLM) fréquence X sévérité**, avec une loi de **Poisson dispersée** pour les modèles de fréquence et une loi **Gamma** pour les modèles de sévérité, appliqués à chacune des garanties. Les données ont été divisées en deux ensembles :

— une base d'entraînement contenant **80 %** des données, et
— une base de test contenant les **20 %** restants.

La sélection des variables s'est faite selon le critère d'**AIC**, et les modèles ont été validés par un test de **déviance**.
Les résultats obtenus sur la base de test pour les différentes garanties sont présentés dans le tableau suivant :

	Coût Moyen			Fréquence		
Garantie	**Réel**	**Prédit**	**Ecart**	**Réel**	**Prédit**	**Ecart**
Dégâts des eaux	1 836,22	1 841,22	-0,69 %	0,0350	0,0338	-3,55 %
Vol	1 564,94	1 562,13	-0,18 %	0,0079	0,0076	-4,65 %
Incendie	8 240,66	8 236,83	-0,05 %	0,0043	0,0041	-2,85 %
Bris de glace	691,59	690,03	-0,23 %	0,0070	0,0068	-3,05 %
Dégâts électriques	1 572,30	1 568,66	-0,23 %	0,0098	0,0094	-3,22 %
Tempête Grêle Neige	4 572,53	4 566,89	-0,12 %	0,0115	0,0112	-3,04 %

TABLE 9 – Résultats du modèle Fréquence X Coût sur les garanties dommages sur la base de test

Pour modéliser la **prime pure des catastrophes naturelles**, un modèle **Tweedie** a été choisi, en raison de sa capacité à gérer des données comportant de nombreux zéros. Le paramètre de puissance de ce modèle a été déterminé à 1,37 pour le modèle d'inondation et à 1,35 pour celui de sécheresse. La prime pure a ensuite été modélisée via un GLM Tweedie pour ces deux risques. La validation du modèle a été réalisée à l'aide de la méthode de validation croisée k-fold, avec 5 plis, en utilisant le critère du **gini**. La stabilité du gini à chaque pli a confirmé la validité du modèle.

La prime pure dommages a été obtenue en additionnant la prime pure des garanties dommages et la prime pure des catastrophes naturelles, elle-même calculée comme la somme des primes pures des risques inondation et sécheresse. La proportion de la prime pure des catastrophes naturelles par rapport à la prime dommages a ainsi été calculée à **14,68 %** comme le montre le tableau suivant :

		Proportion
Prime pure dommages	185,34 €	14,68 %
Prime pure catastrophes naturelles	27,21 €	

TABLE 10 – Proportion de la prime pure catastrophe naturelle en fonction de la prime pure dommages

Elle est au dessus du taux légal actuel de 12 % contribuant ainsi à la fragilisation du régime Cat-Nat.

Les dérives de fréquence et de coût estimées pour les inondations et la sécheresse ont permis de simuler de nouveaux sinistres et d'évaluer leur impact sur la charge de sinistralité de l'assureur. À horizon 2028, l'augmentation de la charge simulée est entre **11,36 %** et **20,81 %**, ce qui correspond à un taux de surprime compris entre **16,90 %** et **17,74 %**.

Le graphique suivant montre l'évolution de cette sinistralité entre 2025 et 2028 et sa projection linéaire.

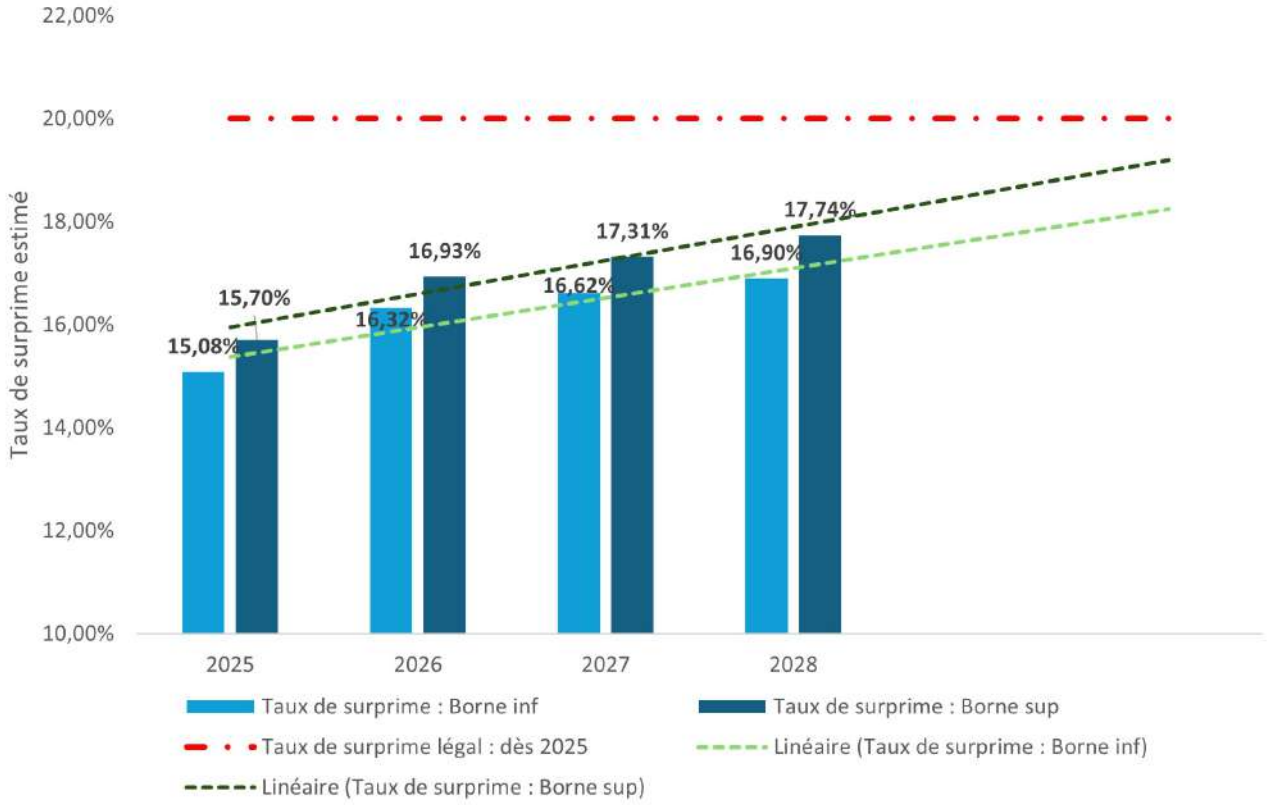


FIGURE 2 – Taux de surprime estimé entre 2025 et 2028

Indépendamment du scénario de dérive de coût, une tendance à la hausse de l'estimation du taux de surprime est observée pour les années à venir. La prévision linéaire suggère qu'à long terme (les 10 à 15 prochaines années), un taux de surprime de 20 % pourrait devenir insuffisant pour couvrir la sinistralité du portefeuille si les conditions actuelles persistent. Un ajustement régulier et une réévaluation du taux de surprime seront donc nécessaires pour garantir une couverture adéquate du risque.

2- Impact des dérives sur la prime tempête

Les dérives de fréquence et de coût des tempêtes ont ensuite été utilisées pour simuler de nouveaux sinistres et estimer leur impact sur la prime pure payée par les assurés. Cette analyse indique une augmentation de la prime pure de **2,97 %** à **6,63 %** pour la garantie Tempête, Grêle et Neige (TGN) en raison du risque tempête d'ici 2028.

Prime	Prime avec dérive : borne inférieure	Prime avec dérive : borne supérieure
53,25	54,83	56,77
	2,97 %	**6,63 %**

TABLE 11 – Estimation de l'augmentation de la prime pure TGN du fait des dérives de sinistralité estimées

Conclusion

Le réchauffement climatique, caractérisé par l'augmentation des événements climatiques extrêmes, impose des défis majeurs au secteur de l'assurance. Ce mémoire se concentre sur l'évaluation de l'impact de ces changements sur la sinistralité et la surprime du régime Cat-Nat, ainsi que sur la prime liée à la garantie TGN d'un portefeuille MRH.

En utilisant des séries temporelles et un modèle de classification à l'aide d'une forêt aléatoire, l'étude prévoit une hausse de la fréquence des sinistres de **23 %** pour les inondations et de **9 %** pour la sécheresse d'ici **2028**. Cela entraîne une surprime Cat-Nat de **16,90 %** à **17,74 %** pour le portefeuille MRH étudié, avec une augmentation continue à long terme. Bien que ces chiffres soient spécifiques au portefeuille MRH étudié, ils reflètent une tendance générale à la hausse applicable à d'autres portefeuilles MRH, avec des variations du taux de surprime selon l'exposition propre à chaque portefeuille.

Les résultats montrent que, sous ces hypothèses, le taux de surprime de **20 %** pourrait être suffisant à court ou moyen terme (entre 7 et 9 ans, par exemple, pour ce portefeuille). Dans les années où la sinistralité serait inférieure à la prime collectée, ces excédents pourraient être utilisés par la CCR pour alimenter les provisions d'égalisation. Cependant, à

plus long terme (entre 10 et 15 ans), si la tendance à la hausse de la fréquence et de la gravité des sinistres se poursuit, ce taux deviendrait insuffisant pour couvrir les risques liés à ce portefeuille, rendant nécessaire une revalorisation continue pour maintenir l'équilibre du régime. La projection linéaire suggère que cette surprime Cat-Nat pourrait atteindre entre **29,87 %** et **32,20 %** d'ici **2050** soulevant des inquiétudes sur l'accessibilité financière à la couverture pour les assurés. L'étude propose plusieurs solutions pour maintenir l'équilibre du régime Cat-Nat dans les années à venir : ajuster annuellement le taux de surprime en fonction de l'évolution de la sinistralité, réduire les risques couverts par le régime, ou passer d'un taux de surprime mutualisé à un taux segmenté en fonction de l'exposition au risque du logement assuré.

En ce qui concerne les tempêtes, une hausse de la sinistralité de **5 %** est anticipée, avec une augmentation de la prime TGN de **2,97 %** à **6,63 %** d'ici **2028**. En considérant l'augmentation de la surprime Cat-Nat et cette augmentation de la prime TGN, la prime totale payée par les assurés pourrait ainsi croître de **10,97 % à 14,63 %** à l'horizon 2028.

Synthesis Note

Context and Problem Statement

Over the past thirty years, the number of natural disasters has tripled, leading to an annual increase in costs of approximately 5.7 %. This rise in both frequency and cost has significant implications, particularly for insurers. In France, natural disasters are reinsured by the State through the Caisse Centrale de Réassurance (CCR), in accordance with the compensation regime established by the Law of July 13, 1982. This regime covers natural disasters such as floods, droughts, landslides, hurricanes, and tornadoes. However, storms, hail, and snow are not included in this regime, as they are considered insurable by private insurers.

Between 2016 and 2022, the claims to premium ratio of the natural disasters regime averaged 126 %, compared to an average of 95 % since 2010, which raises concerns about its viability considering the continual increase of natural disasters. To address this issue, the percentage of premium paid to the CCR by insurers will increase starting in 2025, from :

- **6 % to 9 %** of the premium related to theft and fire coverage for motor vehicles,
- **12 % to 20 %** of the premium related to damage coverage for property other than motor vehicles.

This thesis aims to challenge the adequacy and viability of the new premium rates applied to a home insurance portfolio. It also seeks to analyze the impact of claims related to future storms, which are not covered by the natural disasters regime, on the premiums paid by insurers. To achieve this, a projection of the drift for 2028 will be established for risks related to floods, droughts, and storms. These drifts will then be factored into the pricing of a house insurance portfolio to evaluate their impacts.

Dataset Creation

To determine the frequency drift of floods, droughts, and storms, it was necessary to have observed data on meteorological parameters as well as the occurrence of natural disasters. For this purpose, data from *open data* sources were used, including the **SYNOP** and **GASPAR** datasets.

The SYNOP dataset, provided by Météo France, contains daily observations collected per weather stations on temperature, humidity, wind speed, gust speed, atmospheric pressure, and precipitation levels, which are essential for the various models. The GASPAR dataset (Gestion Assistée des Procédures Administratives relatives aux Risques) records, for each municipality, the number of natural disasters recognized by a natural disaster decree published in the Official Journal. Given the increase in claims related to climatic events since 2016, a data history covering the period from **2016 to 2023** was selected. The analysis was limited to **Metropolitan France**.

Finally, a geographical dataset containing the coordinates of each department in France was also used. The different datasets : SYNOP, GASPAR, and Geographic, were merged using the common key : the department, to obtain a dataset called the "climatic dataset." This dataset contains, for each day between January 1, 2016, and December 31, 2023, and for each department of Metropolitan France, the observed meteorological parameters necessary for the models to be constructed later. Missing values were imputed by replacing them with the **monthly mean**.

This dataset will allow for the determination of the frequency drift for floods, droughts, and storms.
The following graph illustrates the process of constructing the datasets and obtaining the drift.

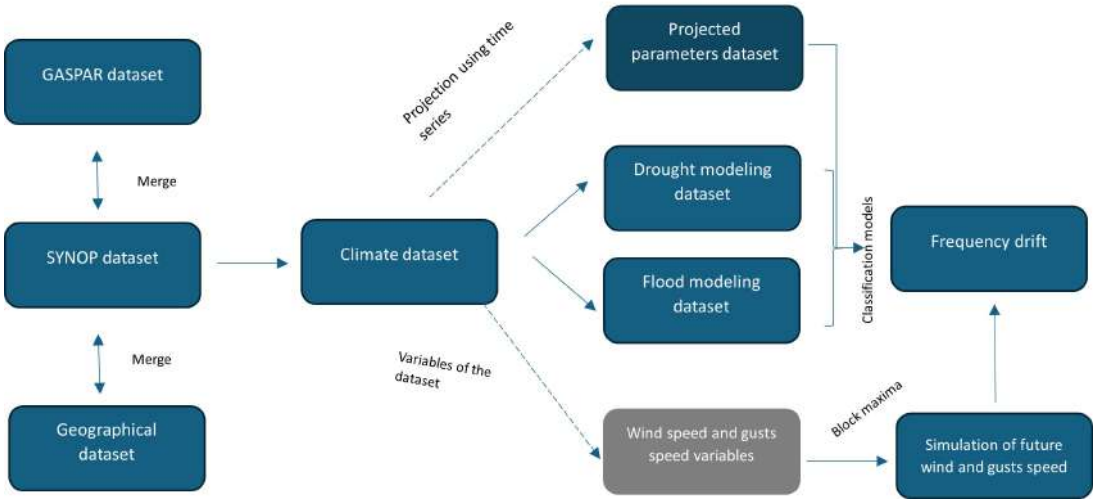


FIGURE 3 – Methodology for Obtaining Drifts

Obtaining the Drifts

I- Frequency Drift of Flood and Drought Risks

1- Time Series

To estimate the number of natural disaster decrees related to floods and droughts, it is crucial to project meteorological parameters in the future. With this in mind, the climatic dataset was segmented into **five** homogeneous risk zones, using Principal Component Analysis (PCA) and Agglomerative Hierarchical Clustering (AHC) to minimize modeling costs.

The data was divided into two sets : a training set comprising observations from **2016 to 2022**, and a test set containing observations from **2023**. For each zone, the modeling of temperature, humidity, precipitation, and pressure was done using time series, based on observed zone averages. The **non-stationary components**, namely trend and seasonality, were modeled using a **parametric model**. The trend was adjusted using a polynomial of degree r, while seasonality was modeled with a trigonometric polynomial. The results of the temperature modeling are presented in the following table :

	Zone 1	Zone 2	Zone 3	Zone 4	Zone 5
R^2	75.78 %	81.04 %	77.21 %	76.72 %	90.83 %
MAE	1.40	1.33	1.17	1.44	0.84
RMSE	1.80	1.73	1.52	1.85	1.11

TABLE 12 – Evaluation of the Fit Quality for Non-Stationary Components of Temperature : Train dataset

The **stationary components**, that is, the residuals, were modeled using the **ARMA model**, as shown in the following table :

	Zone 1	Zone 2	Zone 3	Zone 4	Zone 5
Modèle ajusté	ARMA(1,1)	ARMA(3,0)	ARMA(1,1)	ARMA(4,0)	ARMA(1,3)

TABLE 13 – ARMA Models Fitted per Zone for Temperature

By fitting the trend, seasonality, and residuals for each parameter, 10,000 trajectories were simulated for the years 2024 to 2028. The results of these simulations for temperature are summarized in the following table :

	Zone 1	Zone 2	Zone 3	Zone 4	Zone 5
Average Temperature 2016-2023 (°C)	13.17	13.93	14.42	14.76	17.35
Average Temperature 2024-2028 (°C)	14.00	14.68	15.19	15.44	17.59
Difference in %	5.93	5.11	5.06	4.42	1.39

TABLE 14 – Comparison of Observed and Predicted Average Temperature between 2024 and 2028

2- Classification Models

The datasets : drought model dataset and the flood model dataset, were then used to fit a classification model. These datasets include the following variables : **department, observation date, observation month, temperature, humidity, pressure, precipitation over the last 24 hours, and a claim variable (equal to 1 in the case of a declared natural disaster and 0 otherwise)**.

The modeling was performed using **logistic regression**, **decision tree** and **random forest**, using historical data from 2016 to 2023 for the flood model and from 2016 to 2022 for the drought model. This difference in period was intended to avoid bias due to the delay in the publication of natural disaster decrees for droughts, which usually takes one to two years. The **random forest** model was selected for its ability to better detect claims compared to logistic regression. The confusion matrices associated with these models are presented below.

	Predicted Negative	Predicted Positive
Actual Negative	83,735	27
Actual Positive	45	6,332

TABLE 15 – Confusion Matrix for Drought Risk : Test dataset

	Predicted Negative	Predicted Positive
Actual Negative	83,251	387
Actual Positive	133	4,105

TABLE 16 – Confusion Matrix for Flood Risk : Test dataset

The simulated parameter dataset was then used as input for the classification model to estimate the number of natural disaster decrees between 2024 and 2028 for floods and between 2023 and 2028 for droughts. This approach allowed for the calculation of future claim frequency drift, estimated at **9 %** for drought risk and **23 %** for flood risk.

II- Frequency Drift of Storm Risk

In this modeling, the variables of interest are **wind speed** and **gust speed**. To estimate the number of future storms, it is crucial to simulate future gust speeds since insurers only compensate for damages when these speeds exceed **100 km/h**. The modeling was conducted on wind speed using **extreme value theory**, specifically the **block maxima** method with monthly blocks. This approach was chosen due to the quality of available data : wind speed data had a significantly lower rate of missing values (0.7 %) compared to gust speed data (18.6 %). The parameters obtained from this modeling were then used to simulate future wind speeds, which were subsequently converted into gust speeds via linear regression.

To reduce the modeling cost, the departments were grouped into five homogeneous risk zones. A Generalized Extreme Value (GEV) distribution was fitted for each zone to determine the most appropriate distribution for modeling wind speed, based on the shape parameter. Zones 1, 2, 3, and 4 showed a shape parameter with "zero" in their confidence intervals, indicating that the Gumbel distribution is suitable for these zones. However, for Zone 5, which represents the southeast of France, a full GEV model is necessary, as confirmed by ANOVA analysis.

	Zone 1	Zone 2	Zone 3	Zone 4	Zone 5
Location Parameter	44.25	44.52	43.27	58.48	63.84
Standard Deviation	0.83	0.88	0.75	1.28	1.04
Scale Parameter	7.76	8.17	7.02	11.97	9.65
Standard Deviation	0.61	0.63	0.55	0.95	0.73
Quantile Associated with Return Period	76.81	76.02	73.61	117.74	92.45
ANOVA Result	0.53	0.19	0.63	0.47	0.001234

TABLE 17 – Parameters of the Fitted Gumbel Distribution

The simulation of maximum wind speeds reached each month using these parameters revealed a lack of consideration for the dependence between different geographic zones. To address this, a **Frank copula** was fitted to model this dependence. By incorporating both the specific parameters for each zone and the Frank copula parameter, it was possible to simulate monthly maximum wind speeds for the period 2024-2028 more accurately.

A linear regression was applied to convert the wind speed simulations into gust speeds. The equation obtained, validated by appropriate tests, was used for this conversion :

$$\text{gust\_speed} = 1.50 \times \text{wind\_speed} + 5.94$$

The evaluation metrics obtained for this model are as follows :

	R^2	RMSE	MAE
Training Set	84.52 %	6.51	4.71
Test Set	84.29 %	6.58	4.70

TABLE 18 – Evaluation of the Regression Model

This method allowed for the calculation of a storm frequency drift, estimated at **5 %** for the period 2024-2028.

III- Cost drifts :

The cost drift was estimated based on a benchmark of published articles, as cost-related data is considered confidential and is not available in open data. By principle of prudence, an exponential growth model was used to estimate the cost drift. The estimated drift is as follows :

Risk	Horizon	Cost Evolution
Flood	2028	[3.6 ; 7.9[%
Drought	2028	[2.6 ; 4.7[%
Storm	2028	[0.0 ; 5.3[%

TABLE 19 – Assumed Cost Drift for Claims by 2028

Implementation of Pricing

A dataset containing information on a home insurer's portfolio from **2016 to 2022** was used to assess the impact of the estimated drifts on the portfolio.

1- Adequacy and Viability of the New Premium Rate for the Natural Disaster Scheme

Currently, the premium paid is calculated as **12 %** of the premium of property damage coverage. The first step was to determine, based on the portfolio's claims experience, the proportion of the pure premium for natural disasters relative to the property damage premium, assuming this premium was calculated by the insurer.

In order to achieve this, the **property damage premium** was modeled using **Generalized Linear Models (GLM) frequency X severity**, with a **Poisson** distribution for frequency models and a **Gamma** distribution for severity models, applied to each coverage. The data was split into a training set containing **80 %** of the data and a test set containing the remaining **20 %**. Variable selection was based on the **AIC** criterion, and the models were validated by a **deviance** test.

The results obtained on the test set for the different coverages are presented in the table below :

	Average Cost			Frequency		
Coverage	**Actual**	**Predicted**	**Difference**	**Actual**	**Predicted**	**Difference**
Water Damage	1 836.22	1 841.22	-0.69 %	0.0350	0.0338	-3,55 %
Theft	1 564.94	1 562.13	-0.18 %	0.0079	0.0076	-4.65 %
Fire	8 240.66	8 236.83	-0.05 %	0.0043	0.0041	-2.85 %
Glass Breakage	691.59	690.03	-0.23 %	0.0070	0.0068	-3.05 %
Electrical Damage	1 572.30	1 568.66	-0.23 %	0.0098	0.0094	-3.22 %
Storm Hail Snow	4 572.53	4 566.89	-0.12 %	0.0115	0.0112	-3.04 %

TABLE 20 – Results of Frequency X Cost Model on Property Damage Coverages on the test dataset

To model the **pure premium for natural disasters**, a **Tweedie** model was chosen due to its ability to handle data with many zeros. The power parameter of this model was determined to be 1.37 for the flood model and 1.35 for the drought model. The pure premium was then modeled via a GLM Tweedie for these two risks. Model validation was performed using the k-fold cross-validation method, with 5 folds, using the **gini** criterion. The stability of the gini across folds confirmed the model's validity.

The property damage pure premium was obtained by adding the pure premium of each property damage coverage and the pure premium for natural disasters, was calculated as the sum of the pure premiums for flood and drought risks. The proportion of the pure premium for natural disasters relative to the property damage premium was thus calculated as **14.68 %**, which is above the actual legal rate of 12 %.

		Proportion
Property damage pure premium	185.34 €	14.68 %
Natural disaster pure premium	27.21 €	

TABLE 21 – Proportion of Natural Disaster Pure Premium Relative to Property Damage Pure Premium

The estimated frequency and cost drifts for floods and droughts were used to simulate new claims and evaluate their impact on the insurer's claim amount. By 2028, this amount is expected to increase between **11.36 %** and **20.80 %**, corresponding to a premium rate between **16.90 %** and **17.74 %**.

The following chart shows the evolution of this rate between 2025 and 2028 and its linear projection.

FIGURE 4 – Estimated Surcharge Rate Between 2025 and 2028

Regardless of the cost drift scenario, an upward trend in the estimated surcharge rate is observed for the coming years. The linear forecast suggests that in the long term (10 to 15 years), a surcharge rate of 20 % might become insufficient to cover the portfolio's claims burden if current conditions persist. Regular adjustment and re-evaluation of the surcharge rate will therefore be necessary to ensure adequate risk coverage.

2- Storm Drift

The estimated frequency and cost drifts for storms were then used to simulate new claims and estimate their impact on the pure premium paid by policyholders. This simulation estimated an increase in the pure premium between **2.97 %** and **6.63 %** for the Storm, Hail, and Snow coverage due to the storm risk by 2028.

Premium	Premium with Drift : Lower Bound	Premium with Drift : Upper Bound
53.25	54.83	56.77
	2.97 %	**6.63 %**

TABLE 22 – Estimated Increase in Pure Premium for Storm, Hail, and Snow Coverage Due to Estimated Claims Drift

Conclusion

Climate change, characterized by the increase in extreme weather events, presents significant challenges to the insurance sector. This thesis focuses on assessing the impact of these changes on the premium rate of the natural disasters scheme, as well as the premium related to the Storm, Hail and Snow guarantee of a home insurance portfolio.

Using time series analysis and a random forest classification model, the study predicts a **23 %** increase in claims frequency for floods and a **9 %** increase for droughts by **2028**. This will result in a natural disaster premium rate ranging from **16.90 % to 17.74 %** for the housing insurance portfolio studied, with a continuous increase in the long term. Although these figures are specific to the portfolio analyzed, they reflect a general upward trend applicable to other home insurance portfolios, with variations in the premium rate depending on the specific exposure of each portfolio.

The results show that, under these assumptions, the natural disasters premium rate of 20 % could be sufficient in the short to medium term (between 7 and 9 years, for example, for this portfolio). In years when claims amount is lower than the premiums collected, these surpluses could be used by CCR to fund equalization reserves. However, in the long term (between 10 and 15 years), if the trend of increasing claims frequency and severity persists, this rate would become insufficient to cover the risks associated with this portfolio, making it necessary to continuously reassess the rate to maintain the regime's balance. A linear projection suggests that this premium rate could reach between **29.87 %** and **32.20 %** by **2050**, raising concerns about financial accessibility of the coverage for policyholders.

The study suggests several options to ensure the balance of the Cat-Nat regime in the upcoming years : adjusting the natural disasters premium rate annually based on the evolution of claims frequency, reducing the risks covered by the regime, or transitioning from a unique premium rate to a segmented one based on the risk exposure of the insured property.

Regarding storms, a **5 %** increase in claims frequency is anticipated, with an increase in the Storm, Hail and Snow premium of **2.97 % to 6.63 %** by **2028**. Considering the increase in the natural disasters premium rate and this rise in the Storm, Hail and Snow premium, the total premium paid by policyholders could thus grow by **10.97 %** to **14.63 %** by **2028**.

Table des matières

I L'assurance et les risques climatiques **3**

1 Les catastrophes naturelles en France **5**
1.1 Définition des catastrophes naturelles 5
1.2 Lien entre le réchauffement climatique et l'occurrence des catastrophes naturelles 6
1.3 Le régime d'indemnisation des Catastrophes Naturelles 9
1.3.1 Principes généraux du régime Cat-Nat 9
1.3.2 Périmètre de couverture 11
1.3.3 Les franchises 12
1.3.4 Mécanisme d'indemnisation des Cat-Nat 12
1.3.5 Sinistralité observée sur le régime Cat-Nat 13

2 Les tempêtes **19**
2.1 Définition d'une tempête 19
2.2 Les tempêtes en France 20
2.3 La garantie TGN MRH en quelques chiffres 21
2.4 Spécificité de la garantie TGN en assurance MRH 22
2.5 Synthèse de la première partie 24

II Evaluation de la dérive de sinistralité des événements climatiques **25**

3 Base de données **29**
3.1 Open Data et Risques Climatiques 29
3.2 Traitement des données 30
3.2.1 Choix de la plage temporelle 30
3.2.2 Traitement des données SYNOP 30
3.2.3 Traitement des données GASPAR 32

4 Dérive Inondation et Sécheresse **39**
4.1 Construction du zonier . . . 40
4.1.1 Analyse en Composantes Principales (ACP) . . . 40
4.1.2 Classification Ascendante Hiérarchique (CAH) . . . 42
4.2 Séries Temporelles . . . 44
4.2.1 Modélisation de la température . . . 46
4.3 Classification des sinistres . . . 55
4.3.1 Modèles Linéaires Généralisés . . . 55
4.3.2 Régression Logistique . . . 56
4.3.3 Les arbres de décision . . . 60
4.3.4 Forêts aléatoires . . . 62
4.4 Dérive de sinistralité de la fréquence . . . 64
4.5 Critiques des modèles . . . 66

5 Dérive tempête **67**
5.1 Construction du zonier . . . 67
5.2 Modélisation de la vitesse du vent par zone. . . . 70
5.2.1 Théorie des valeurs extrêmes . . . 70
5.2.2 Théorie des Copules . . . 76
5.2.3 Modélisation de la relation entre la vitesse du vent et la vitesse des rafales . . . 81
5.2.4 Détermination de la dérive de sinistralité . . . 84
5.3 Dérive de sinistralité du coût des sinistres . . . 86
5.4 Limites de la méthodologie . . . 87

III Impact de la prise en compte des dérives dans le tarif **89**

6 Tarification **93**
6.1 Assurance Multirisques Habitation . . . 93
6.1.1 Les principaux indicateurs de sinistralité . . . 94
6.1.2 Quelques chiffres du marché MRH en France . . . 94
6.2 Présentation des bases de données . . . 95
6.2.1 Ajout de variables . . . 95
6.2.2 Détermination des seuils des sinistres graves . . . 96
6.2.3 Analyses descriptives . . . 98
6.3 Détermination de la suffisance et viabilité du taux de surprime de 20 % . 101
6.3.1 Détermination de la prime pure dommages . . . 101
6.3.2 Détermination de la prime pure catastrophes naturelles . . . 107
6.3.3 Prise en compte des dérives inondation et sécheresse . . . 110
6.4 Estimation de l'impact de la dérive de sinistralité liée aux tempêtes . . . 114
6.5 Impact global des événements climatiques sur le tarif à l'horizon 2028 . . . 114
6.6 Critiques des modèles . . . 115

Conclusion **117**

A Chapitre 4 : Dérive de sinistralité des inondations et de la sécheresse **123**

B Chapitre 5 : Evaluation de la dérive de sinistralité tempête **127**

C Chapitre 6 : Tarification **129**

Bibliographie **144**

Introduction

Le réchauffement climatique représente un enjeu de plus en plus critique pour les assureurs du monde entier. En effet, la fréquence et le coût des sinistres liés à ce phénomène augmentent de manière alarmante. Au cours des trente dernières années, le nombre de catastrophes naturelles et climatiques a triplé, entraînant une hausse des coûts d'environ 5,7 %, alors que la croissance moyenne du PIB n'est que de 2,7 %. Les projections du rapport du Groupe d'Experts Intergouvernemental sur l'Evolution du climat (GIEC) sont préoccupantes : la température moyenne mondiale augmente d'environ 0,2 °C tous les dix ans, avec une hausse potentielle de 1,5 °C d'ici 2100 dans le meilleur des scénarios, et jusqu'à 4,8 °C dans le pire des cas. Ces perspectives laissent présager un impact croissant pour les assureurs face à l'augmentation des risques climatiques.

En France, le régime d'indemnisation des catastrophes naturelles, appelé « régime Cat-Nat », indemnise les assurés lorsque l'État décrète un état de catastrophe naturelle. Ce régime repose sur la garantie Cat-Nat, obligatoire pour les contrats d'assurance dommages, et les assureurs ont la possibilité de se réassurer auprès de la Caisse Centrale de Réassurance (CCR) moyennant une surprime. Depuis 1999, ce taux de surprime est fixé à 12 % de la prime afférente aux garanties dommages pour un contrat Multirisques Habitation (MRH). Actuellement, 95 % des assureurs sont réassurés par la CCR, car la CCR bénéficie de la réassurance illimitée auprès de l'État. La viabilité de ce régime est menacée par l'augmentation tant en fréquence qu'en coût des catastrophes naturelles. Le ratio de sinistralité des huit dernières années atteint 126 %, contre 95 % sur les vingt dernières années. Pour faire face à cette hausse, le taux de surprime Cat-Nat pour l'assurance MRH passera à 20 % de la prime afférente aux garanties dommages dès 2025.

Cette augmentation du taux de surprime soulève plusieurs questions pour les assureurs, notamment en termes de **suffisance** et de **viabilité**, surtout face aux projections de hausse continue des températures. De plus, le régime Cat-Nat ne couvre pas les risques liés aux tempêtes, à la grêle et à la neige, considérés comme des dommages assurables. Or, avec le réchauffement climatique, la garantie Tempête, Grêle et Neige (TGN) pourrait également être impactée. Il est donc essentiel de déterminer quelle augmentation de prime sera nécessaire pour couvrir la sinistralité croissante associée à cette garantie.

Ce mémoire propose d'intégrer les dérives de sinistralité, en termes de fréquence et de coût des événements climatiques, dans la tarification d'un portefeuille MRH afin de répondre à ces enjeux.

La première partie du mémoire reviendra sur la définition des catastrophes naturelles et climatiques, ainsi que sur l'impact du réchauffement climatique sur celles-ci. Le fonctionnement du régime Cat-Nat et de la garantie TGN y sera également explicité.

La deuxième partie du mémoire sera consacrée à l'obtention des dérives de sinistralité pour les risques d'inondation, de sécheresse et de tempête à l'horizon 2028. Tout d'abord, la construction de la base de données à partir de données publiques issues des bases SYNOP, GASPAR et de données géographiques sera présentée. Ensuite, la méthodologie pour l'obtention des dérives, ainsi que la théorie sous-jacente, seront exposées et mises en œuvre.

Pour les risques d'inondation et de sécheresse, cette méthodologie consistera à prédire le nombre de déclarations d'état de catastrophe naturelle entre 2024 et 2028 pour les inondations, et entre 2023 et 2028 pour la sécheresse. Cette prédiction sera réalisée à l'aide d'un modèle de classification ajusté sur les paramètres météorologiques journaliers (température, humidité, précipitations, pression) de l'historique 2016 à 2023, ces paramètres étant eux-mêmes projetés via des séries temporelles pour obtenir les prévisions 2024 à 2028. Pour le risque tempête, la dérive de sinistralité sera estimée en utilisant la théorie des valeurs extrêmes, plus précisément la méthode des maxima par blocs, appliquée aux variables de vitesse du vent et des rafales. La dérive des coûts sera, quant à elle, obtenue via un benchmark d'articles scientifiques publiés par différents auteurs.

La troisième partie utilisera les dérives de sinistralité obtenues pour évaluer la suffisance et la viabilité du nouveau taux de surprime Cat-Nat sur un portefeuille MRH. L'analyse débutera par la détermination de la prime pure dommages du portefeuille MRH, suivie par l'estimation de la prime pure pour les catastrophes naturelles, ce qui permettra de calculer la surprime nécessaire pour couvrir la sinistralité actuelle du portefeuille. Ensuite, les dérives de sinistralité permettront de simuler de nouveaux sinistres et d'évaluer leur impact sur la charge de sinistralité, ainsi que sur le taux de surprime requis à l'horizon 2028 et au-delà. De plus, la dérive liée au risque tempête sera utilisée pour simuler de nouveaux sinistres et évaluer l'impact sur la prime de la garantie TGN. Cette démarche globale permettra de déterminer l'influence des événements climatiques sur les tarifs d'un assureur MRH.

Première partie

L'assurance et les risques climatiques

Chapitre 1

Les catastrophes naturelles en France

1.1 Définition des catastrophes naturelles

Une catastrophe naturelle se manifeste par un événement d'origine naturelle, survenant soudainement et violemment, causé par l'intensité anormale d'un phénomène naturel. Ses conséquences engendrent des perturbations majeures susceptibles d'occasionner d'importants dégâts matériels et humains . *[Bolluze, 2024]*

Les différentes formes de catastrophes naturelles au sens géophysique sont :

- la sécheresse
- les inondations
- les tempêtes
- les cyclones
- les séismes
- les avalanches
- les glissements de terrain et coulées de boue
- les éruptions volcaniques

D'après la répartition des catastrophes naturelles dans le monde en 2022, les inondations et les tempêtes sont les événements les plus fréquents représentant plus de 70% des catastrophes naturelles survenues.

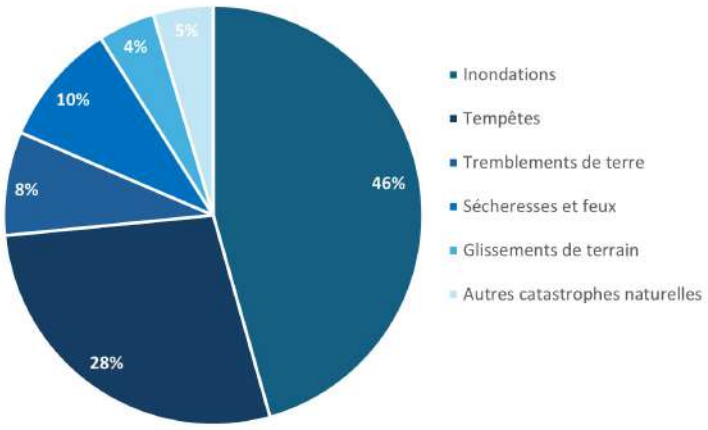


FIGURE 1.1 – Répartition des catastrophes naturelles dans le monde en 2022 (Source : *Statista*)

1.2 Lien entre le réchauffement climatique et l'occurrence des catastrophes naturelles

D'après les Nations Unies, le réchauffement climatique se définit comme une variation à long terme de la température moyenne de la terre et des modèles météorologiques. Cette variation peut résulter de facteurs naturels que les cycles solaires ou les éruptions volcaniques massives. Toutefois, depuis le 19ème siècle, l'activité humaine est la principale cause de cette variation de température au travers l'émission de gaz à effet de serre tels que le dioxyde de carbone et la méthane. Ainsi, la température moyenne de la surface de la terre a augmenté de manière significative, dépassant de 1,1°C celle enregistrée en 1800. Les données révèlent que la décennie récente (2011-2020) a été la plus chaude jamais observée sachant que depuis 1850 chaque décennie a connu une augmentation thermique progressive. *[ONU, 2023]*

Selon le rapport du GIEC de mars 2023, le réchauffement climatique va se poursuivre avec une projection de 1,5°C d'augmentation par rapport à la période de référence 1850-1900, envisagée d'ici le début des années 2030. Les projections du GIEC ont identifié divers scénarios qui dépendent en grande partie des émissions de gaz à effet de serre. Dans le cas de faibles émissions, il est envisageable de contenir le réchauffement en dessous de 2°C, voire 1,5°C en cas de réduction significative (SSP-11.9). Cependant, en se basant sur les politiques actuelles, la trajectoire du réchauffement climatique serait de 3,2°C d'ici 2100. *[Climat, 2023]*

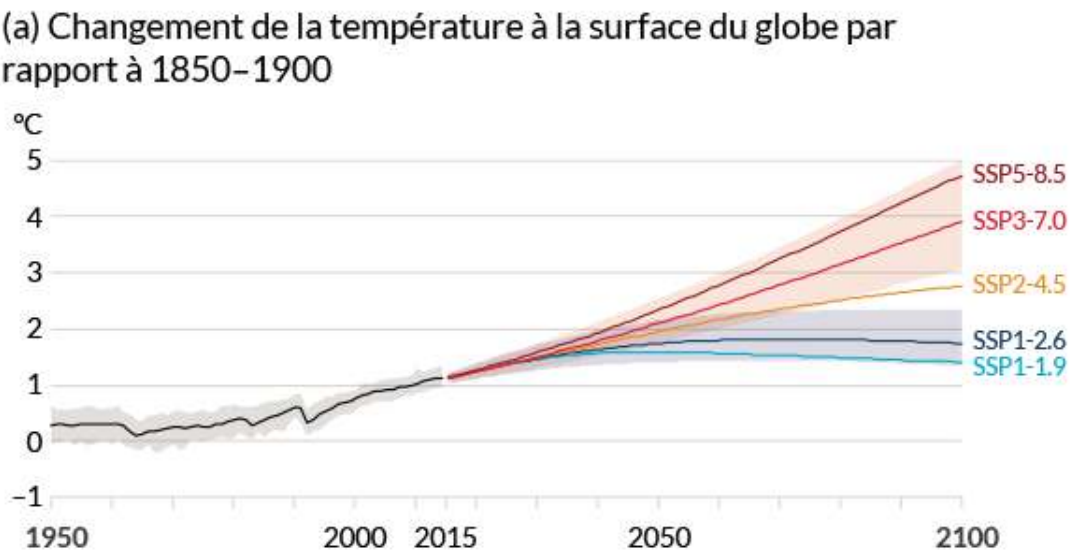


FIGURE 1.2 – Changement de la température à la surface du globe par rapport à 1850 - 1900 (Source : *Rapport du GIEC*)

Le réchauffement climatique a engendré une série d'effets dommageables qui se manifestent à l'échelle mondiale, dont un étant la hausse de la fréquence et l'intensité des catastrophes naturelles. En effet, la hausse des températures atmosphériques et océaniques conduit à une hausse du niveau des mers ce qui amplifie l'intensité des phénomènes tels que les tempêtes, les vents, les sécheresses, les incendies ainsi que les précipitations et les inondations qui ont une durée plus longues. Les données révèlent des faits alarmants, le nombre de catastrophes naturelles a triplé au cours des trente dernières années générant de ce fait un coût plus élevé avec une croissance d'environ 5,7 % par an alors que la croissance du taux de Produit Intérieur Brut (PIB) est d'environ 2,7 %.

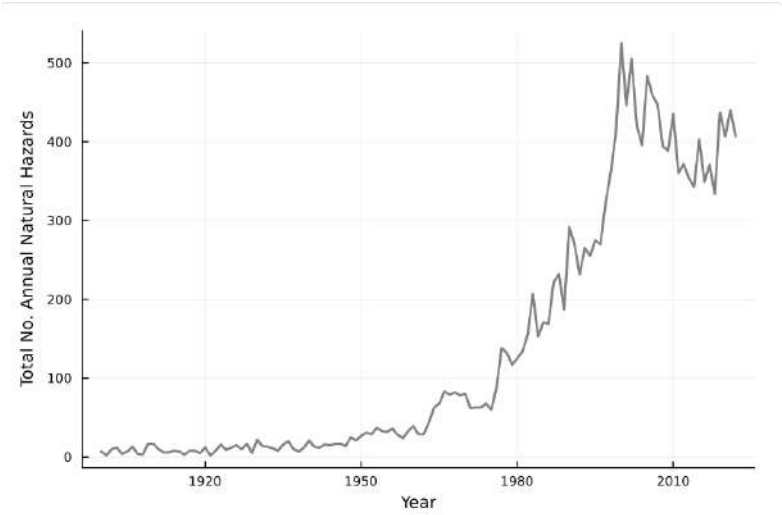


FIGURE 1.3 – Evolution du nombre de sinistres dus aux catastrophes naturelles dans le monde entre 1900 et 2022 (Source : *EM-DAT*)

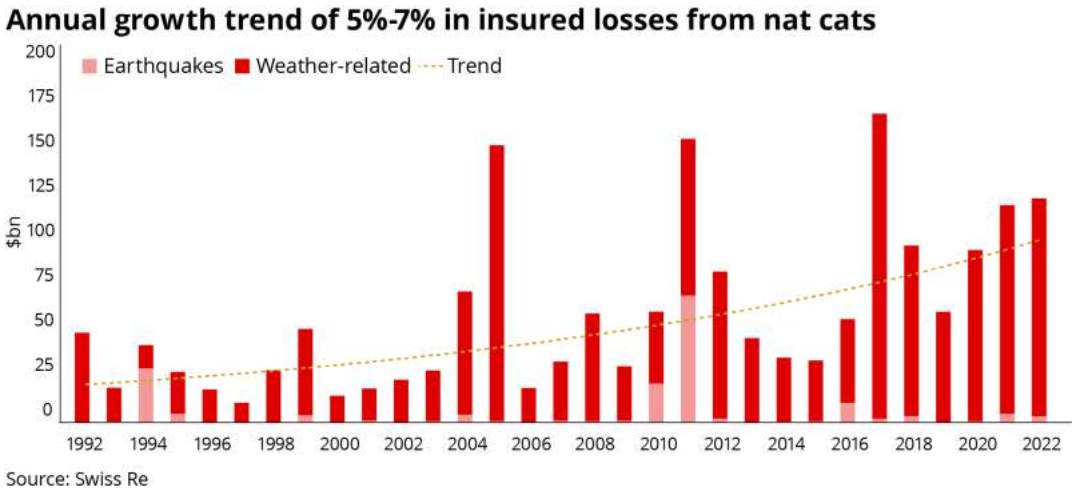


FIGURE 1.4 – Évolution du coût des sinistres dus aux catastrophes naturelles dans le monde de 1992 à 2022 (Source : *Swiss Re*)

La France, comme le reste du monde, n'est pas à l'abri de ces défis. En 2023, par exemple, le pays a été confronté à 14 inondations, 15 épisodes venteux avec des rafales dépassant les 150 km/h, et 2 tempêtes touchant le nord-ouest. Le coût financier de ces catastrophes s'est élevé à 6,5 milliards d'euros, plaçant ainsi 2023 au troisième rang des années les plus sinistrées, après 1999 et 2022.

FIGURE 1.5 – Evolution de la fréquence des sinistres tous périls (Source : *Bilan CatNat CCR*)

Cette augmentation tant de la fréquence que du coût des événements naturels soulève la **question de leur impact sur le secteur de l'assurance, notamment sur les portefeuilles des assureurs et les primes à percevoir pour couvrir le risque associé aux catastrophes naturelles**.

Avant d'approfondir cette problématique, il est essentiel de rappeler le fonctionnement de l'indemnisation des Catastrophes Naturelles (Cat-Nat) en France ainsi que l'évolution de leur sinistralité.

1.3 Le régime d'indemnisation des Catastrophes Naturelles

1.3.1 Principes généraux du régime Cat-Nat

La France a mis en place un dispositif spécifique garantissant à tous ses citoyens une indemnisation adéquate en cas de sinistre causé par un phénomène naturel. Ce régime d'indemnisation des catastrophes naturelles communément appelé « régime Cat-Nat » a été instauré par la Loi du 13 juillet 1982 pour remédier à une lacune en matière de couverture des risques naturels, qui étaient largement sous-assurés jusqu'alors.

L'objectif était de concevoir un système répondant à plusieurs exigences : *[CCR, 2022]*

— Assurer une couverture étendue, à la fois pour les particuliers, les entreprises et les collectivités territoriales, afin de protéger contre tous les types de risques naturels.

— Fonctionner sur le principe de solidarité en assurant une prise en charge des dommages matériels à un coût supportable pour l'ensemble de la société. Cette solidarité se matérialise par l'application de taux de surprime uniques fixés par l'État, qui s'élèvent actuellement à 12 % de la prime pour les biens autres que les véhicules à moteur, et à 6 % des primes pour le vol et l'incendie (ou à défaut 0,50 % de la prime dommage) pour les véhicules terrestres à moteur.

— Optimiser l'efficacité du dispositif en combinant les compétences des acteurs publics et privés. Ce partenariat se traduit par des conditions nécessaires pour déclencher le mécanisme d'indemnisation, notamment la publication d'un arrêté constatant l'état de catastrophe naturelle dans le Journal Officiel pour la condition d'ordre public, et la souscription d'un contrat d'assurance dommages pour la couverture du bien endommagé pour la condition d'ordre privé. Les critères d'appréciation de l'état de catastrophes naturelles comprennent divers éléments tels que la durée de retour pour les inondations, les caractéristiques des mouvements de terrain, le bilan hydrique et la nature des sols pour la sécheresse et la réhydratation, l'origine, la localisation et l'ancienneté des bâtiments touchés pour les avalanches, ainsi que la magnitude et les résultats de l'enquête macrosismique pour les séismes.

— Assurer la solvabilité et la durabilité du système sur le long terme.

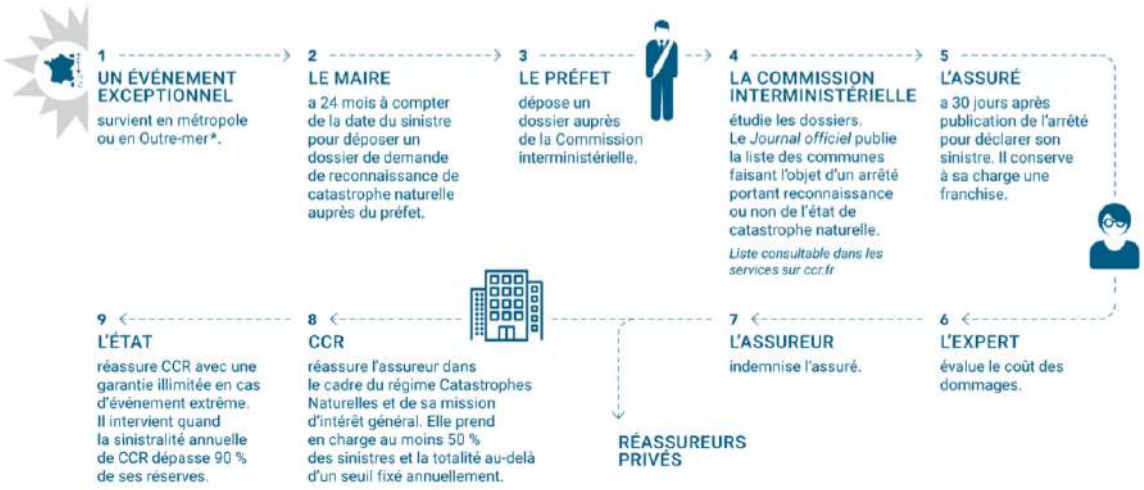


FIGURE 1.6 – Schéma d'indemnisation des catastrophes naturelles en France (Source : *CCR*)

Les fondements du régime des Catastrophes Naturelles reposent sur l'assurance et la réassurance publique. Les assureurs sont chargés de diffuser et de mutualiser largement la protection légale à travers leurs contrats, de collecter la surprime légale, et d'évaluer et d'indemniser rapidement les sinistrés conformément aux exigences réglementaires. La

réassurance publique, assurée par la Caisse Centrale de Réassurance (CCR), a pour mission de soutenir tout assureur qui en fait la demande dans le cadre légal peu importe les caractéristiques de leur portefeuille, assurer une mutualisation nationale de tous les risques en couvrant les portefeuilles des différents assureurs et créer des couvertures de réassurance robustes et durables, tout en évitant un transfert excessif des risques vers le réassureur et, indirectement, vers l'État.

Les dégâts causés par les catastrophes naturelles représentent une charge de sinistralité considérable qui dépasse les capacités du marché de l'assurance et de la réassurance. Pour éviter toute défaillance du système, une intervention de l'État en dernier recours est prévue. Cette garantie de l'État est accordée à la CCR afin de lui permettre d'accomplir sa mission d'intérêt général. Elle constitue ainsi le dernier rempart de l'architecture du régime Cat-Nat. Ce dispositif apparaît comme une structure cohérente et robuste, assurant une indemnisation proportionnelle à l'ampleur des dommages. Les événements de gravité moyenne sont gérés conjointement par l'assurance et la réassurance publique, tandis que les événements plus graves bénéficient d'une prise en charge accrue de la réassurance publique. Enfin, les événements majeurs mobilisent l'ensemble des acteurs du régime : l'assurance, la réassurance publique et l'État.

1.3.2 Périmètre de couverture

L'assurance catastrophe naturelle constitue une extension de garantie incluse dans tous les contrats d'assurance dommages (multirisque habitation, tous risques auto, local professionnel). Conformément à l'article L125-1 du Code des Assurances, cette garantie prend en charge les dommages matériels directs non assurables, causés principalement par l'intensité anormale d'un phénomène naturel, lorsque les mesures préventives habituelles se révèlent inefficaces. Parmi les périls généralement couverts, on trouve notamment :*[CCR, 2024b]*

- les inondations
- la sécheresse
- les mouvements de terrain
- les cyclones et les ouragans
- les séismes
- les avalanches
- le volcanisme
- les tsunamis.

Pour ces périls, la liste non-exhaustive des dommages matériels directs pris en charge par la garantie catastrophe naturelle sont :

- Les dommages matériels directs aux bâtiments, au matériel et au mobilier, incluant la couverture de la valeur à neuf si spécifiée dans le contrat
- Les honoraires d'architecte, de décorateur et de contrôle technique
- Les frais de démolition et de déblai des biens assurés sinistrés

— Les dommages causés par l'humidité ou la condensation consécutive à la stagnation de l'eau dans les locaux
— Les frais de pompage, de nettoyage et de désinfection des locaux sinistrés, ainsi que toute mesure de sauvetage
— Les frais d'études géotechniques nécessaires à la remise en état des biens garantis
— Les véhicules couverts par une assurance dommages (la responsabilité civile obligatoire seule ne couvre pas ce type de sinistre).

Toujours conformément à l'article L125-1 du Code des Assurances, les dommages résultants de phénomènes naturels assurables sont exclus du périmètre de couverture de la garantie catastrophe naturelle. Les périls concernés par cette exclusion sont :

— les tempêtes
— la grêle
— la neige
— le gel.

1.3.3 Les franchises

Le mécanisme des franchises est déterminé par l'État, et elles sont à la fois obligatoires et non rachetables. Depuis le 1er janvier 2001, leur établissement se fait comme suit :

Biens à usage d'habitation et autres biens non professionnels	Dommages directs	380€	Sécheresse 1 520€
Biens à usage professionnel	Dommages directs	10% (Minimum 1 140€)	Sécheresse 10% (Minimum 3 050€)
	Pertes d'exploitation	3 jours ouvrés (Minimum 1 140€)	

TABLE 1.1 – Franchises fixées par l'État (Source : *CCR*)

En cas de plusieurs déclarations d'état de catastrophe naturelle sur une même commune sur une période de 5 ans, et en absence d'un Plan de Prévention des Risques Naturels, la franchise peut être adaptée. *[CCR, 2024b]*

1.3.4 Mécanisme d'indemnisation des Cat-Nat

Le dispositif de réassurance proposé par la CCR se compose d'une couverture proportionnelle en quote-part à hauteur de 50 % ainsi que d'une couverture non proportionnelle sur rétention. Concernant la couverture proportionnelle en quote-part, les compagnies d'assurance transfèrent 50 % de leurs primes liées aux risques de catastrophes naturelles à la CCR, qui en retour prend en charge 50 % des sinistres relevant du régime Cat-Nat. Ce partage de 50 % entre les assureurs et la CCR facilite une répartition équitable des

risques et permet également à la CCR de garantir une mutualisation solidaire entre différents portefeuilles, dont les niveaux d'exposition varient considérablement. En moyenne, sur la période allant de 1982 à 2022, la CCR a couvert 51 % des sinistres liés aux catastrophes naturelles.

Pour les 50 % restants sous la responsabilité des assureurs, la CCR intervient (par le biais de la couverture non proportionnelle en rétention) lorsque la sinistralité dépasse un certain seuil, souvent fixé au montant des primes Cat-Nat de l'assureur. Par ailleurs, dans le cadre du régime Cat-Nat, l'État soutient la CCR dès que le niveau de sinistralité annuel excède un certain seuil, actuellement fixé à 90 % des réserves de la CCR. L'État offre une garantie illimitée à la CCR, laquelle n'a été mobilisée qu'une seule fois, en 2000 pour l'exercice 1999, en réponse aux tempêtes Lothar et Martin. Actuellement, la CCR couvre environ 95 % du marché. Le schéma de réassurance fait l'objet de renégociations régulières entre l'État et les représentants du secteur, afin de garantir son adéquation avec les besoins et les risques actuels. *[SENAT, 2024]*

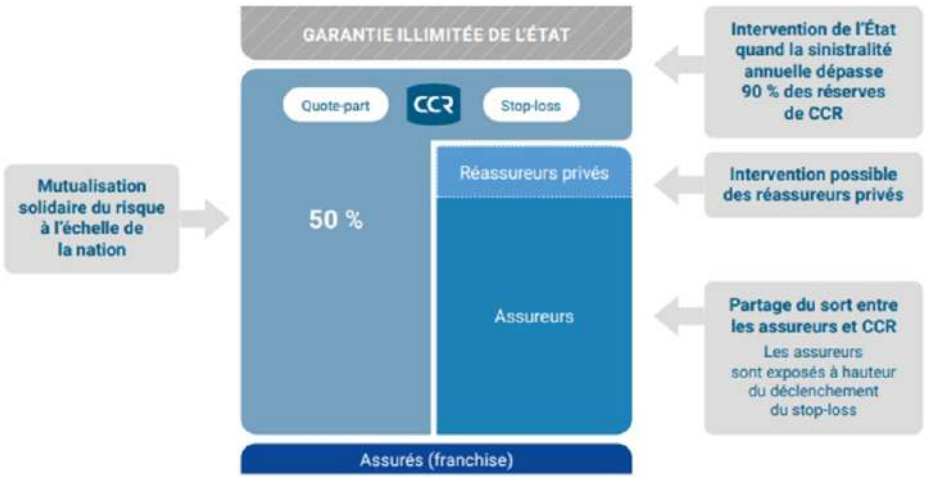


FIGURE 1.7 – Mécanisme d'indemenisation des catastrophes naturelles en France (Source : *SENAT*)

1.3.5 Sinistralité observée sur le régime Cat-Nat

Entre 1982 et 2022, le montant des sinistres couverts par la garantie catastrophes naturelles s'élève à 49,9 milliards d'euros. Parmi ceux-ci, les dommages dus à la sécheresse et aux inondations représentent **91 %**, tandis que les autres risques ne comptent que pour 9 %. La part de la sécheresse a connu une hausse considérable au cours des dernières années. Entre 2010 et 2016, elle ne représentait que 25 % à 35 % de la sinistralité. Cependant, depuis 2017, cette part augmente rapidement. Sur les cinq dernières années, la sécheresse représente 70 % des dommages survenus.

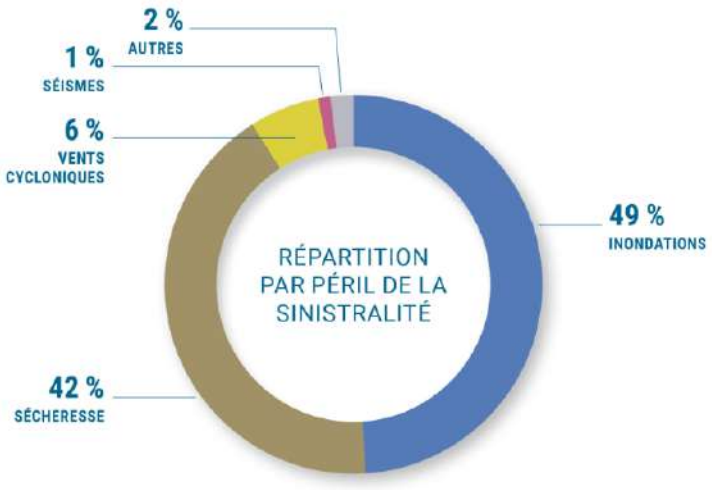


FIGURE 1.8 – Répartition de la sinistralité par péril (Source : *Bilan CatNat CCR*)

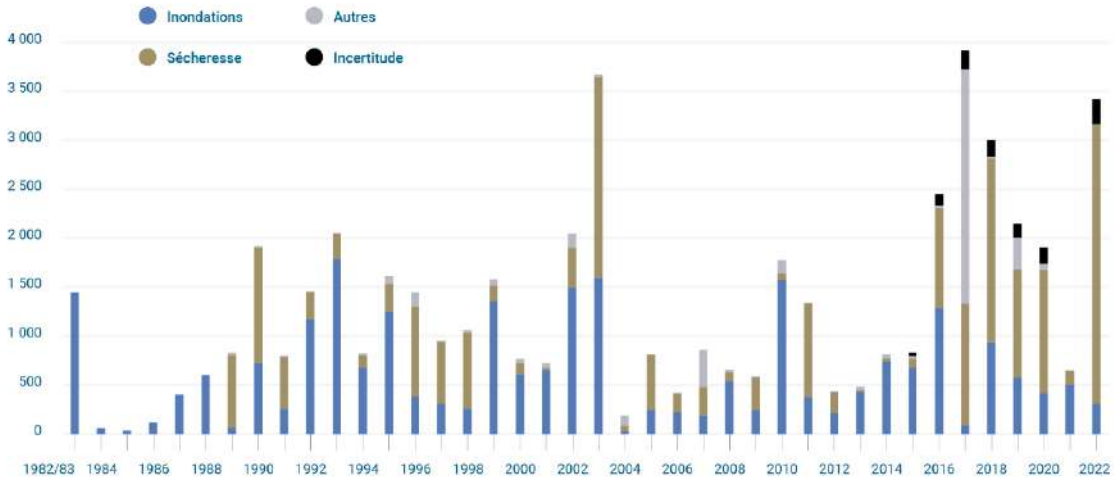


FIGURE 1.9 – Sinistralité des catastrophes naturelles non-auto entre 1982 et 2022 (Source : *Bilan CatNat CCR*)

Depuis 2016, le régime Cat-Nat connaît une hausse structurelle tant de la fréquence que de l'intensité des sinistres, marquée notamment par des épisodes de sécheresse et d'inondations, à l'exception de l'année 2021. L'évolution du ratio sinistres sur primes depuis 2016 révèle que, sur les six années considérées, quatre affichent un ratio supérieur à 100 %, avec une moyenne de **126 %** entre 2016 et 2022 indiquant que les primes collectées par la CCR ne sont pas suffisantes pour couvrir l'ensemble des sinistres liés aux catastrophes naturelles. Les ratios S/P les plus élevés se retrouvent dans le sud de la France, frappé à la fois par les inondations et la sécheresse. La branche auto est moins affectée par la sinistralité croissante que la branche non-auto celle-ci n'étant pas soumise au risque de sécheresse.

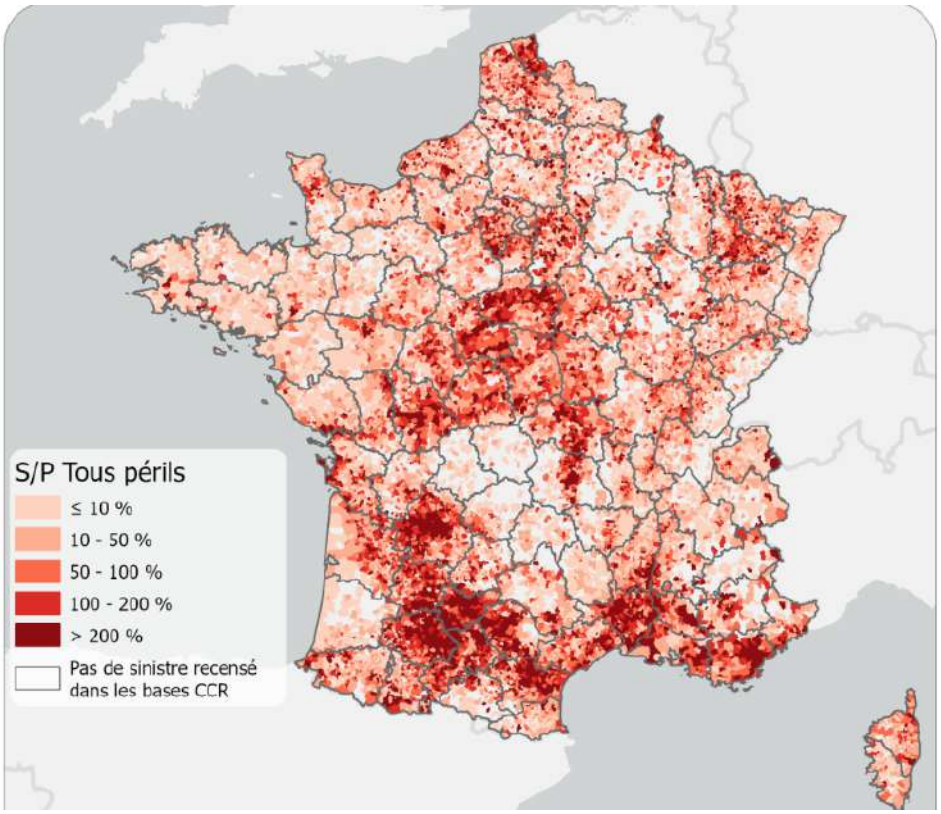


FIGURE 1.10 – Sinistralité observée tous périls confondus entre 1982 et 2022 (Source : *Bilan CatNat CCR*)

Malgré l'augmentation structurelle de la fréquence et de l'intensité des catastrophes naturelles, le taux de la surprime n'a pas évolué depuis l'année 2000. Cette situation crée un déséquilibre menaçant la pérennité du régime Cat-Nat. En effet, l'augmentation des sinistres est l'une des principales raisons de la réduction de la provision d'égalisation Cat-Nat de la CCR, un phénomène principalement attribué à la sinistralité liée aux épisodes de sécheresse. La charge moyenne a atteint environ un milliard d'euros par an entre 2015 et 2023, contre 445 millions d'euros en moyenne depuis 1982. *[SENAT, 2024]*

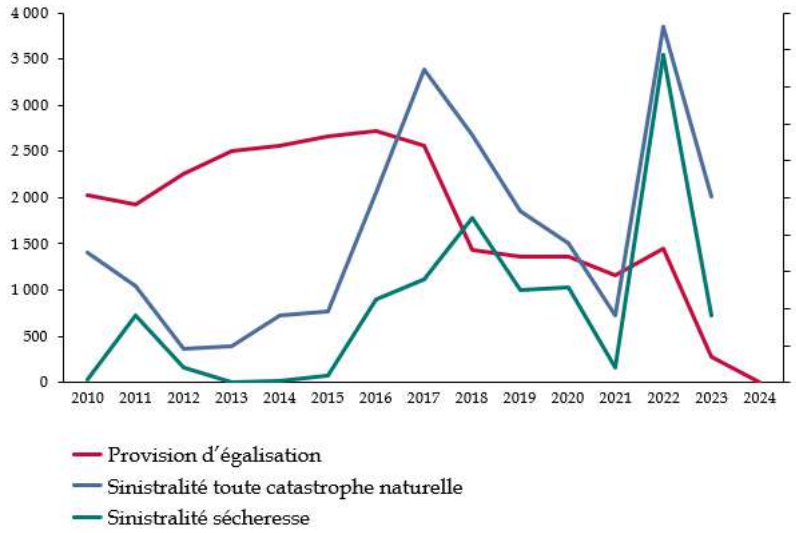


FIGURE 1.11 – Evolution de la provision d'égalisation de la CCR et de la sinistralité depuis 2010 (Source : *SENAT*)

Afin de remédier à cette situation, le taux de surprime a été réévalué et ses évolutions entreront en vigueur dès 2025.

Il passera de :

— **12 % à 20 %** pour les contrats d'assurance de dommages aux biens résidentiels et professionnels, et

— **6 % à 9 %** pour les garanties vol et incendie des contrats automobiles (ou à défaut, de 0.50 % à 0.75 % de la prime dommage). *[Maire, 2023]*

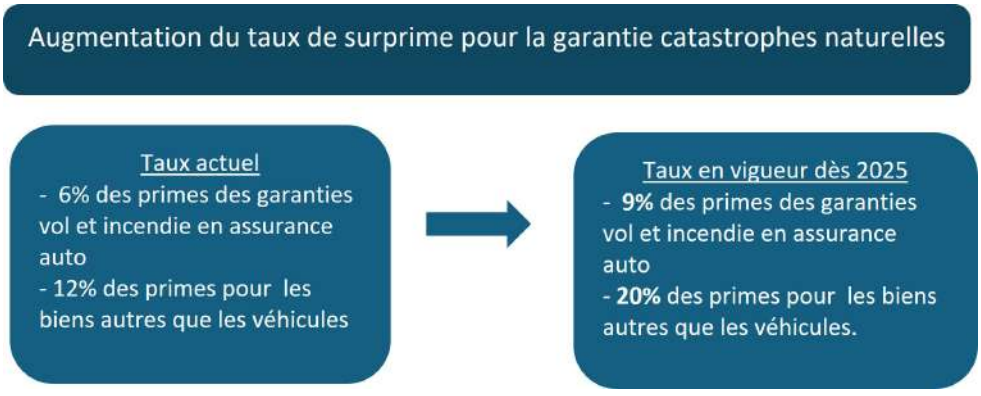


FIGURE 1.12 – Augmentation du taux de surprime Cat-Nat (Source : *Légifrance*)

La branche non-auto subit une plus grande incidence de la recrudescence des catastrophes naturelles, notamment causée par la sécheresse. Cela se traduit par une augmentation beaucoup plus significative du taux de surprime par rapport à la branche auto (8 points contre 3 points). Pour cette raison, nous nous concentrerons sur l'assurance **MRH** dans ce mémoire.

Dans un contexte où les primes payées par les assurés connaissent une hausse constante, attribuable à l'inflation, aux changements législatifs et aux effets des changements climatiques, cette augmentation du taux de surprime soulève plusieurs interrogations pour les assureurs :

1. Quelle est la viabilité à long terme des nouveaux taux, notamment face aux projections de hausse continue des températures ?
2. Comment ajuster la prime des assurés pour refléter adéquatement le niveau de risque encouru ?
3. Quel impact la sinistralité liée aux tempêtes, à la grêle et à la neige, qui ne relèvent pas du régime des catastrophes naturelles, a-t-elle sur leur portefeuille ?

Chapitre 2

Les tempêtes

Dans le premier chapitre, il a été vu que les dommages causés par les tempêtes ne sont pas inclus dans le régime Cat-Nat car ils sont considérés comme assurables. Cependant, du fait de leur nature climatique, leur fréquence et leur coût sont également influencés par le réchauffement climatique. Il semble donc important de revenir sur la définition des tempêtes et du coût engendré par les tempêtes majeures.

2.1 Définition d'une tempête

Une tempête se caractérise par des vents violents produits par une dépression barométrique marquée. Ce phénomène météorologique se produit lorsque des courants d'air chaud venant de la mer rencontrent des courants d'air froid provenant de la terre. Le choc entre ces masses d'air, aux températures et à la teneur en eau différentes, engendre des vents violents, parfois accompagnés de pluie, de neige, de grêle ou de sable. L'air chaud, plus léger, s'élève au-dessus de l'air froid, formant ainsi un front dont la perturbation est appelée dépression. Une définition universelle de la tempête n'étant définie, l'échelle de Beaufort créée au 19ème siècle par un amiral britannique initialement à usage marine demeure la meilleure référence. Selon cette échelle, on parle de tempête lorsque la vitesse du vent est située entre 89 km/h et 117 km/h.

En assurance, les dommages causés par une tempête ne sont pris en compte que si la vitesse des rafales dépasse les **100 km/h**. Ces dommages peuvent avoir des répercussions humaines, économiques ou environnementales. Les conséquences économiques des tempêtes sont significatives. Dans ce mémoire, seuls les dommages causés aux véhicules seront examinés.

Echelle de BEAUFORT

Beaufort	Description	Noeuds de	Noeuds à	Km / h de	Km / h à	Vagues (m) de	Vagues (m) à
0	Calme	0		0		0.0	
1	Très légère brise	1	3	1	5	0,1	
2	Légère brise	4	6	6	11	0.2	0.5
3	Petite brise	7	10	12	19	0.6	0.9
4	Jolie brise	11	16	20	28	1.0	1.5
5	Bonne brise	17	21	29	38	2.0	2.5
6	Vent frais	22	27	39	49	3.0	4.0
7	Grand frais	28	33	50	61	4.0	5.5
8	Coup de vent	34	40	62	74	5.5	7.0
9	Fort coup de vent	41	47	75	88	7.5	10.0
10	Tempête	48	55	89	102	10.0	12.5
11	Violente tempête	56	63	103	117	12.5	14.0
12	Ouragan	64	>	118	>	16.0	>

FIGURE 2.1 – Echelle de Beaufort (Source : *Ouranos*)

2.2 Les tempêtes en France

Depuis 1980, la France a été frappée par plusieurs tempêtes remarquables. Parmi celles-ci, on peut citer :

— La tempête Lothar survenue en 1999, caractérisée par des rafales dépassant les 140 km/h, touchant plus de 56 % du territoire.

— La tempête Martin, survenue peu après la tempête Lothar, en 1999, affectant environ 50 % du territoire. Ces deux tempêtes ont entraîné la perte de 92 vies humaines et des dommages évalués à 5,9 milliards d'euros.

— La tempête Klaus, en 2009, a été la plus violente, avec des pointes de vent atteignant 191 km/h. Elle a causé la mort de 12 personnes et les dommages sont estimés à 1.2 milliards d'euros.

— La tempête Xynthia, en 2010, a provoqué de nombreuses submersions avec comme bilan 53 morts et plus de 6000 maisons endommagées.

— Enfin, la tempête Alex, en 2020, a été particulièrement pluvieuse. *[GALL, 2023]*

Pour évaluer la sévérité des tempêtes de manière comparative, un indice appelé *SSI*[1]a été élaboré. Cet indice se base uniquement sur les surfaces relatives de zones touchées par des rafales dépassant les 100, 120, 140 ou 160 km/h, multipliées par ces valeurs de vent au cube pour refléter leur potentiel destructeur. Une tempête est considérée :

1. Storm Severity Index

— exceptionnelle si son indice SSI dépasse 12,

— forte si son indice SSI se situe entre 4 et 12, et

— modérée si son indice SSI est inférieur à 4.

Parmi les tempêtes recensées depuis 1980, neuf sont classées comme exceptionnelles, trente-six comme fortes et trois cent trente comme modérées. *[France, 2023]*

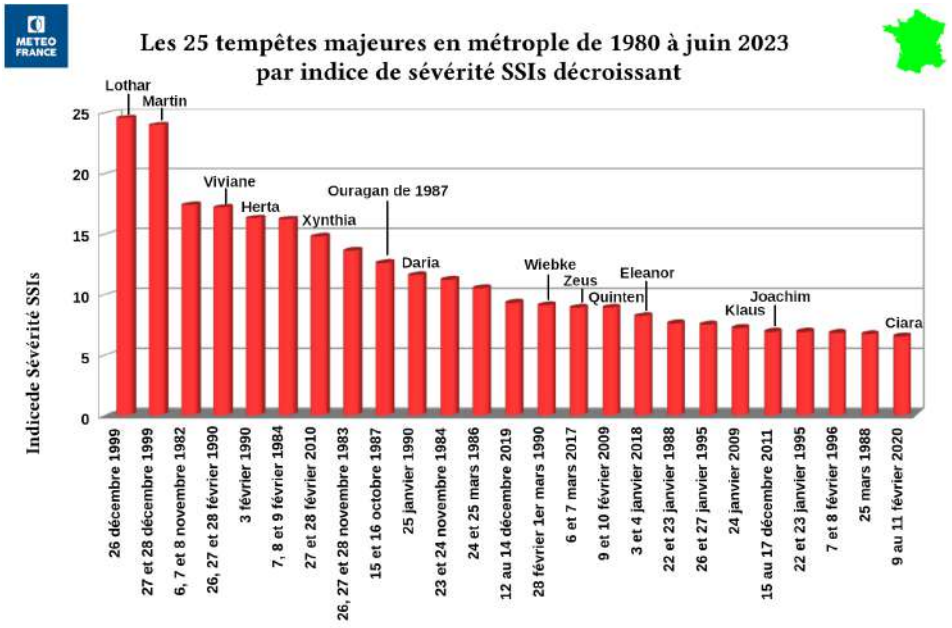


FIGURE 2.2 – Les 25 tempêtes majeures en métropole de 1980 à juin 2023 par indice de sévérité décroissant (Source : *Météo France*)

2.3 La garantie TGN MRH en quelques chiffres

Le graphique suivant présente l'évolution du sinistre moyen (en euros courants) et de la fréquence de la garantie Tempête Grêle Neige (TGN) en France sur une période allant de 1990 à 2022.

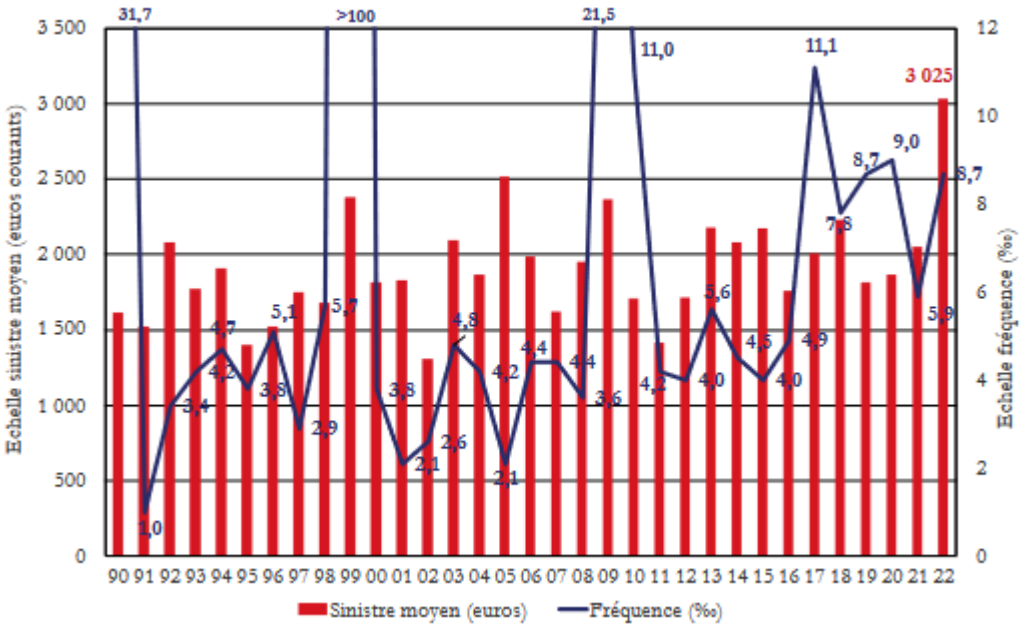


FIGURE 2.3 – Historique de la fréquence et du coût moyen tempête depuis 1990 (Source : *France Assureurs*)

Le graphique souligne l'imprévisibilité des tempêtes en France, tant en termes de fréquence que de coût. Les années marquées par des événements météorologiques extrêmes, comme 1999 et 2022, ont des impacts financiers particulièrement lourds. La tendance récente à la hausse dans la fréquence et le coût des sinistres pourrait signaler une vulnérabilité accrue face aux tempêtes, à la grêle ou la neige, possiblement en lien avec le changement climatique.

Tout comme pour la garantie Cat-Nat, la fréquence et le coût des dommages causés par les tempêtes, la grêle et la neige sont en constante augmentation. Il devient donc crucial, pour obtenir une vision globale des risques climatiques, de prendre en compte non seulement les risques couverts par le régime Cat-Nat, mais également ceux engendrés par ces phénomènes météorologiques.

2.4 Spécificité de la garantie TGN en assurance MRH

Les dommages causés par les tempêtes, la grêle et la neige sont pris en charge lorsque l'assuré a souscrit une garantie dommages aux biens. En cas de survenance de l'un de ces événements, l'assureur est tenu d'indemniser l'assuré si son habitation a subi des dégâts. Les modalités de cette garantie sont définies dans le contrat d'assurance. En règle générale, pour que l'indemnisation soit accordée, la vitesse des rafales doit dépasser 100

km/h et une étendue significative des dégâts doit être constatée.

Les cotisations perçues pour la garantie TGN MRH en 2022 se sont élevées à 1 860M€, représentant ainsi **38 %** des cotisations de la couverture contre les événements naturels.

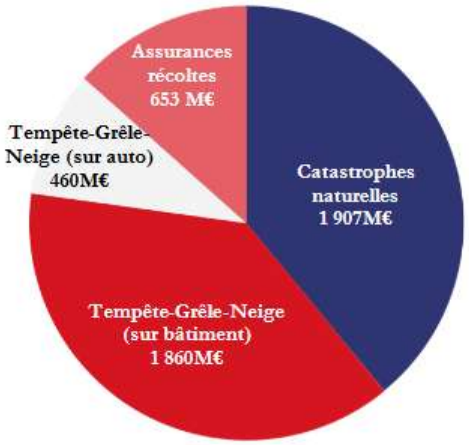


FIGURE 2.4 – Cotisations perçues en 2022 au titre des événements naturels (Source : *France Assureurs*)

Historiquement, les tempêtes sont responsables pour environ **72 %** des sinistres de la garantie TGN en assurance MRH. **Pour cette raison nous nous intéresserons uniquement à la sinistralité liée aux tempêtes dans la suite du mémoire**.

Année	Montant des indemnités (en M€) (estimation marché national)				
	Tempête	Grêle	Neige sur toitures	Total en euros courants	*Total en euros constants 2022 **
2018	810	390	75	1 275	1 459
2019	810	530	30	1 370	1 556
2020	884	97	9	990	1 121
2021	631	286	43	960	1 038
2022	**1 350**	**3 317**	**3**	**4 670**	**4 670**
Total 2018 – 2022	**4 485**	**4 620**	**160**	**9 265**	**9 845**
Total depuis 1990	**23 085**	**7 845**	**1 170**	**32 100**	**49 103**
Total depuis 1984	**24 320**	**8 040**	**1 300**	**33 660**	**53 686**

FIGURE 2.5 – Evolution de la sinistralité de la garantie TGN (Source : *France Assureurs*)

2.5 Synthèse de la première partie

Cette première partie a mis en évidence la croissance continue de la fréquence et du coût des événements climatiques en raison du changement climatique. Les projections de température pour les prochaines décennies indiquent une augmentation potentielle de 3,2 °C d'ici 2100, ce qui renforce l'exposition des assureurs aux risques de survenance d'événements climatiques majeurs. Il devient donc impératif pour eux de mieux pour les assureurs de mieux comprendre et gérer ces risques afin de garantir l'indemnisation des assurés en cas de sinistre, tout en préservant leur solvabilité.

Dans cette perspective, il est proposé d'intégrer la dérive de sinistralité, causée par les événements climatiques d'**inondation**, de **sécheresse** et de **tempêtes** dans la tarification des contrats d'assurance. Une fois cette dérive de sinistralité obtenue, elle sera appliquée à un portefeuille MRH, lequel est particulièrement vulnérable aux événements climatiques, notamment en raison du risque accru de sécheresse. L'objectif est d'évaluer la suffisance et la pérennité des nouveaux taux de surprime du régime Cat-Nat, qui entreront en vigueur en janvier 2025. De plus, l'impact de la dérive de sinistralité liée aux tempêtes, qui ne sont pas couvertes par le régime Cat-Nat, sera également analysé en ce qui concerne la prime payée par les assurés.

Méthodologie d'obtention des dérives de sinistralité

L'obtention des dérives de sinistralité repose sur l'utilisation de données publiques relatives aux paramètres météorologiques, notamment la température, l'humidité, les précipitations, la pression, ainsi que la vitesse du vent et des rafales.

Pour les risques d'**inondation** et de **sécheresse**, un processus en plusieurs étapes sera suivi. Dans un premier temps, des **séries temporelles** seront employées pour projeter les différents paramètres météorologiques (température, humidité, précipitation, pression) à l'horizon 2028. Ensuite, un **modèle de classification** sera ajusté pour déterminer la dérive de sinistralité associée aux paramètres projetés.

Concernant le risque de **tempêtes**, la dérive de sinistralité sera déterminée à l'aide de la théorie des valeurs extrêmes, plus précisément par la méthode des **maxima par blocs**, en utilisant le paramètres météorologique de vitesse du vent. Cette approche permettra d'obtenir les paramètres nécessaires à la simulation de la vitesse du vent et des rafales pour les années à venir.

La construction des bases de données et la méthodologie détaillée pour l'obtention de ces dérives sont explicitées dans la partie suivante.

Deuxième partie

Evaluation de la dérive de sinistralité des événements climatiques

Cette partie a pour objectif de déterminer la dérive de sinistralité future pour les risques d'inondation, de sécheresse et de tempêtes. La détermination des dérives de sinistralité pour les risques d'**inondation** et de **sécheresse** s'appuie sur les travaux réalisés par Inès BOUCHOUCHI dans son mémoire intitulé : « *Défi climatique et durabilité, vers les limites de l'assurabilité ?* » [BOUCHOUCHI, 2024]. Le délai de parution d'un arrêté de catastrophe naturelle étant généralement long pour le risque de sécheresse, l'évolution de la sinistralité entre deux années peut être significative. Il était donc utile d'actualiser les modèles avec des données plus récentes pour obtenir une évaluation précise de la dérive de sinistralité.

Les travaux relatifs à la dérive de sinistralité pour le risque de **tempête** s'inspirent des travaux réalisés Nathalie BEDI dans son mémoire intitulé : « *Modélisation du risque tempête en France Métropolitaine* » [BEDI, 2018]. Depuis 2016, les effets du réchauffement climatique sont devenus plus apparents pour les assureurs. La réexamination des travaux était pertinente pour obtenir une dérive qui appréhende l'évolution récente de sinistralité.

Chapitre 3

Construction des bases de données

3.1 Open Data et Risques Climatiques

Le terme « open data » désigne les données disponibles pour tous, pouvant être librement utilisées et partagées. Pour qu'une donnée soit qualifiée d'ouverte, elle doit satisfaire à trois critères fondamentaux :*[ods, 2024]*

- Les données doivent être facilement disponibles dans un format pratique et leur accès doit être gratuit.
- Ceux ayant accès aux données doivent avoir la liberté de les réutiliser et de les redistribuer, y compris en les combinant avec d'autres ensembles de données.
- Les données doivent être accessibles à tous, qu'il s'agisse d'entreprises privées, de chercheurs ou d'organismes publics.

Les différentes sources de données utilisées pour la construction de la base de données sont les suivantes :

Base SYNOP

La base SYNOP est une base de données mise à disposition du public par Météo France. La base de données contient des observations ayant une fréquence de trois heures, allant de l'année 1996 jusqu'à ce jour. Les données sont récoltées par station météorologiques circulant sur le système mondial de télécommunication (SMT) de l'Organisation Météorologique Mondiale (OMM).

Ces observations sont aussi bien des paramètres atmosphériques mesurés tels que la température, l'humidité, la direction et la force du vent, pression atmosphérique, hauteur de précipitations que des paramètres observés (description des nuages, visibilité) depuis la surface terrestre. *[France, 2024]*

Base de données géographique des départements

Cette base de données contient, pour chaque département de France, le code postal, la région, ainsi que les coordonnées géographiques (latitude et longitude). Elle est disponible sur la plateforme *data.gouv*, une plateforme ouverte et communautaire dédiée à la centralisation et à la structuration des données ouvertes en France.

Base GASPAR

La base de Gestion Assistée des Procédures Administratives relatives aux Risques (GASPAR) est une base de données disponible sur *data.gouv*. Elle recense, pour chaque commune, le nombre de catastrophes naturelles reconnues par un état de catastrophe naturelle, avec publication au Journal Officiel. *[data.gouv, 2024]*

3.2 Traitement des données

3.2.1 Choix de la plage temporelle

Dans le but de fournir des analyses pertinentes sur la sinistralité liée aux événements climatiques, la période d'observation de **2016 à 2023** a été choisie. Cette plage temporelle permet d'observer les tendances et variations récentes, car les effets du changement climatique y sont plus apparents. De plus, cette période offre des données récentes et fiables, essentielles pour prédire l'occurrence future de ces événements climatiques.

3.2.2 Traitement des données SYNOP

La base de données SYNOP contient les paramètres météorologiques observés pour chaque station météorologique en France. Les paramètres retenus dans le cas de cette étude sont :

- l'indicatif OMM de la station
- la date d'observation
- la température en °C
- l'humidité en %
- la pression au niveau de la station
- les précipitations les dernières 24h en mm
- la vitesse du vent en km/h
- la vitesse des rafales en km/h.

Les données sont collectées quotidiennement toutes les trois heures et répertoriées mensuellement. La base finale a été obtenue en calculant pour chaque paramètre météorologique la moyenne journalière des observations et en concaténant toutes les données mensuelles de 2016 et 2023.

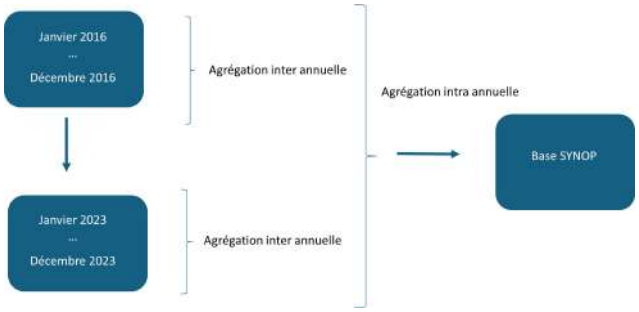


FIGURE 3.1 – Agrégation des données SYNOP mensuelles pour l'obtention de la base finale

La base SYNOP ainsi obtenue contient les données météorologiques de 62 stations (identifiées par l'indicatif OMM) situées en France métropolitaine et dans les régions et collectivités françaises en dehors du continent européen. Pour cette étude, les analyses ont été restreintes à la **France métropolitaine**, englobant ainsi **42 stations**. Cette restriction a été faite car les paramètres météorologiques des régions et collectivités françaises d'outre-mer, tels que la température, présentent des valeurs extrêmes qui ne sont pas représentatives de l'ensemble du territoire métropolitain.

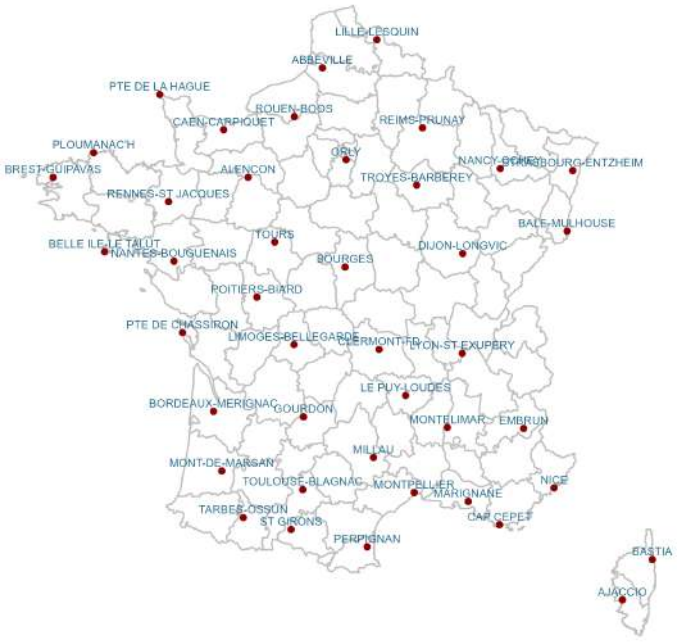


FIGURE 3.2 – Localisation des Stations Météorologiques en France Métropolitaine

Cette base présente néanmoins une limite. En effet, les données sont collectées par station météorologique, et la France métropolitaine ne compte que 42 stations pour 97 départements. Pour pallier ce manque de données, le **centroïde** de chaque département a été déterminé, et pour chaque département, la **moyenne** des paramètres météorologiques des **trois stations les plus proches** a été attribuée.

Il est essentiel d'évaluer la complétude des données en examinant les valeurs manquantes. Le pourcentage de valeurs manquantes pour chaque paramètre est généralement faible, à l'exception de la vitesse des rafales qui présente un taux de 18,6% de données manquantes. Pour la température, l'humidité, la pression et la vitesse du vent, une imputation par la **moyenne mensuelle** est effectuée. En ce qui concerne la vitesse des rafales, cruciale pour déterminer la dérive des tempêtes, les valeurs manquantes sont conservées comme telles pour une analyse ultérieure dans le cadre de la détermination de cette dérive.

Paramètre	Température	Humidité	Pression	Précipitation	Vitesse du vent	Vitesse des rafales
% de valeurs manquantes	1,0%	1,4%	1,6%	4,4%	0,7%	18,6%

TABLE 3.1 – Pourcentage de valeurs manquantes par paramètre météorologique

Après ces traitements, la base de données SYNOP comprend 280 320 observations journalières de paramètres météorologiques pour la période allant du 1er janvier 2016 au 31 décembre 2023. Ces observations couvrent les départements de la France métropolitaine et se caractérisent par les éléments suivants :

Paramètre	Minimum	1er Quartile	Médian	Moyenne	3ème Quartile	Maximum
Température (°C)	-10,8	8,5	12,9	13,3	18,2	34,0
Humidité (%)	1,0	66,1	76,1	74,5	84,6	100,0
Pression (Pa)	88868,0	98774,0	100330,0	99624,0	101255,0	104611,0
Précipitation (mm)	0,0	0,0	0,2	1,9	1,9	105,7
Vitesse du vent (km/h)	0,0	14,4	19,4	21,5	26,3	169,9
Vitesse des rafales (km/h)	0,0	26,3	34,6	38,1	46,1	278,3

TABLE 3.2 – Statistiques descriptives des paramètres météorologiques

3.2.3 Traitement des données GASPAR

La base de données GASPAR téléchargée répertorie les arrêtés de catastrophes naturelles de 1982 à juillet 2024, mais seuls ceux repertoriés de **2016 à 2023** seront pris en compte dans ce mémoire. Les variables présentes dans cette base sont :

- le code de la catastrophe naturelle
- le libellé et code postale des communes touchées
- le type de catastrophe naturelle survenue : inondation, sécheresse, mouvement de terrain etc.

— les dates de début, fin, publication des arrêtés et de leur mise à jour et mise à jour.

Aucune valeur manquante n'a été identifiée dans la base de données, cependant, elle contenait 1 016 lignes en doublon qui ont été supprimées. Afin de faciliter les analyses, une agrégation des libellés de risque a été effectuée suivant la *liste des aléas GASPAR* disponible sur la plateforme *Géorisques.gouv [Géorisques, 2012]*.

Type de risque	Description
Inondation	Inondations et/ou coulées de boues, Inondations remontée nappe, Inondations par choc mécanique lié à l'action des vagues.
Sécheresse	Sécheresses
Mouvement de terrain	Mouvement de terrain, Glissement de terrain, Effondrement et/ou affaissement, Mouvements de terrain différentiels consécutifs à la sécheresse et à la réhydratation des sols
Autres	Secousse sismique, Séismes, Avalanche, Vents cycloniques

TABLE 3.3 – Regroupement des catastrophes naturelles en famille de risque

Statistiques descriptives

Une fois ces traitements effectués la base GASPAR contient 38 974 déclarations dont 38 657 en France Métropolitaine et 317 en Outre-mer. La répartition de ces catastrophes naturelles par libellé de risques révèle que **93 %** des catastrophes naturelles de ce régime sont dues aux **inondations** et à la **sécheresse**. En conséquence, **seules ces deux catastrophes naturelles seront étudiées dans la suite des travaux, pour les risques relatifs aux régime Cat-Nat.**

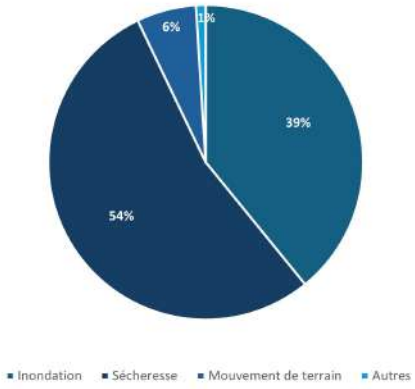


FIGURE 3.3 – Répartition du nombre d'arrêtés de catastrophes naturelles par risque

Inondations

Une inondation est une submersion rapide ou lente d'une zone habituellement hors d'eau. Il existe trois types d'inondation : *[Pyrénées-Orientales.gouv, 2022]*

1. La montée lente des eaux en région de plaine, par débordement d'un cours d'eau ou remontée de la nappe phréatique
2. La formation rapide de crues torrentielles
3. Le ruissellement pluvial

L'évolution du nombre d'arrêtés entre 2016 et 2023 montre une sinistralité élevée en 2016 et 2018, due respectivement aux inondations des bassins de la Seine moyenne et de la Loire, ainsi qu'aux pluies torrentielles dans le sud de la France. En revanche, l'année 2017 a été la moins sinistrée, avec seulement 407 arrêtés.

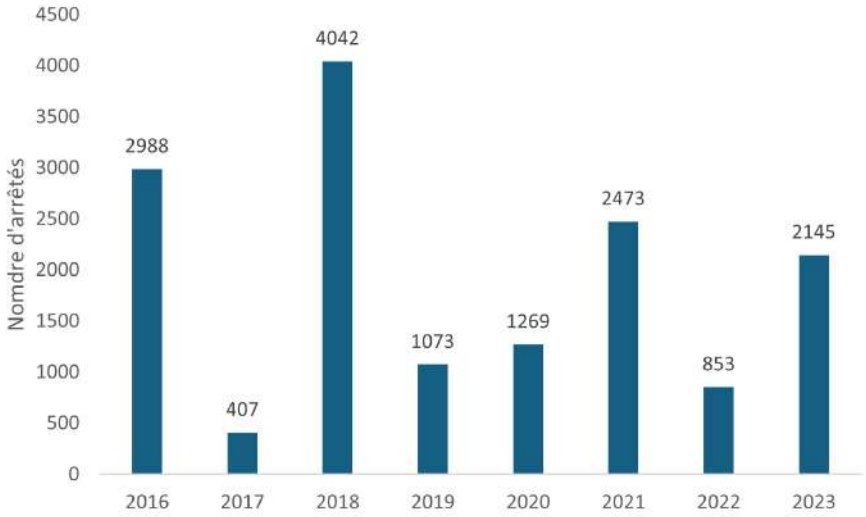


FIGURE 3.4 – Evolution du nombre d'arrêtés du risque inondation entre 2016 et 2023

Une représentation de ces arrêtés par département montre que les départements les plus touchés se trouvent en Hauts-de-France, Île-de-France, et dans le Sud-Ouest de la France.

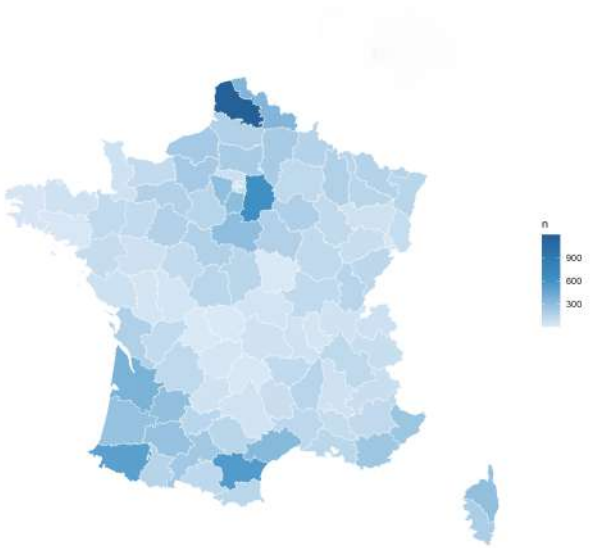


FIGURE 3.5 – Répartition des sinistres inondation par département entre 2016 et 2023

Sécheresse

La sécheresse se caractérise par une période prolongée de pénurie d'eau suffisamment grave pour affecter la végétation, les animaux et les activités humaines. Elle est généralement causée par une combinaison de facteurs : une insuffisance de précipitations, un manque d'eau dans les sols ou dans les cours d'eau, des températures élevées, et une consommation excessive d'eau. *[Manche.gouv, 2022]*

Contrairement aux inondations qui ont une durée courte : entre un et trois jours, la sécheresse s'étend souvent sur une période longue, dépassant fréquemment les trois mois. Cette différence de durée a un impact sur la reconnaissance de l'état de catastrophe naturelle, comme l'illustrent les sinistres de 2023, dont les premiers arrêtés n'ont été publiés qu'en juillet 2024. L'évolution du nombre d'arrêtés entre 2020 et 2022, observée à janvier 2023, utilisé dans le mémoire d'Inès Bouchouchi, et celle présentée dans cette étude, justifie la nécessité de réévaluer les dérives.

	2020	2021	2022
Nombre d'arrêtés sécheresse au 31/01/2023	2 588	647	2 845
Nombre d'arrêtés sécheresse au 01/07/2024	2 597	1 114	6 429
Evolution	**0,35 %**	**72,18 %**	**125,98** %

L'évolution du nombre d'arrêtés montre une sinistralité fortement marquée en 2022. Ceci est causé par le phénomène de **Retrait-Gonflement des sols Argileux (RGA)** touchant principalement les département du Var, de la Haute-Vienne, du Vaucluse, Tarn, Gers et des Bouches-du-Rhône. Le RGA se réfère aux mouvements alternatifs, souvent récurrents, de contraction et d'expansion du sol. Ces mouvements sont liés aux périodes de sécheresse, où le sol se rétracte, et de réhydratation, où le sol dit « gonflant » ou « expansif » se dilate.

FIGURE 3.6 – Evolution du nombre d'arrêtés du risque sécheresse entre 2016 et 2022

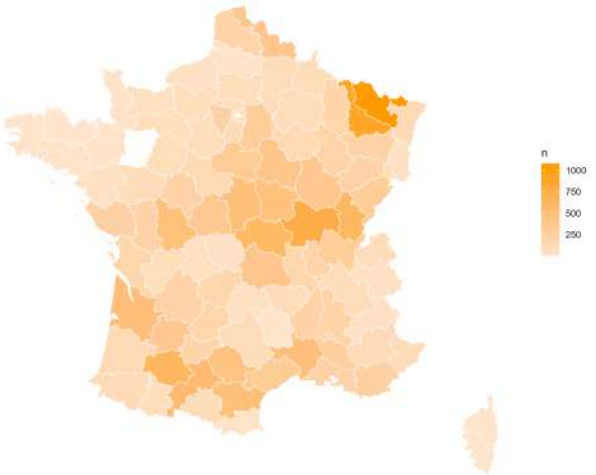


FIGURE 3.7 – Répartition des sinistres sécheresse par département entre 2016 et 2022

Cette représentation du nombre d'arrêtés par département montre que le centre de la France est la zone la plus touchée par le risque de sécheresse. On observe également une forte sinistralité au Nord-Est et dans le Sud-Ouest.

Construction des bases pour le modèle de classification

Dans la suite des travaux, un modèle de classification sera ajusté pour les risques d'inondations et de sécheresse. Afin de construire ces bases, une variable « **sinistre** » valant **1** pour toutes les lignes de la base GASPAR a été rajoutée à celle-ci. Ensuite, la base GASPAR a été fusionnée avec la base SYNOP, permettant ainsi d'obtenir les bases de données nécessaires pour le modèle de classification.

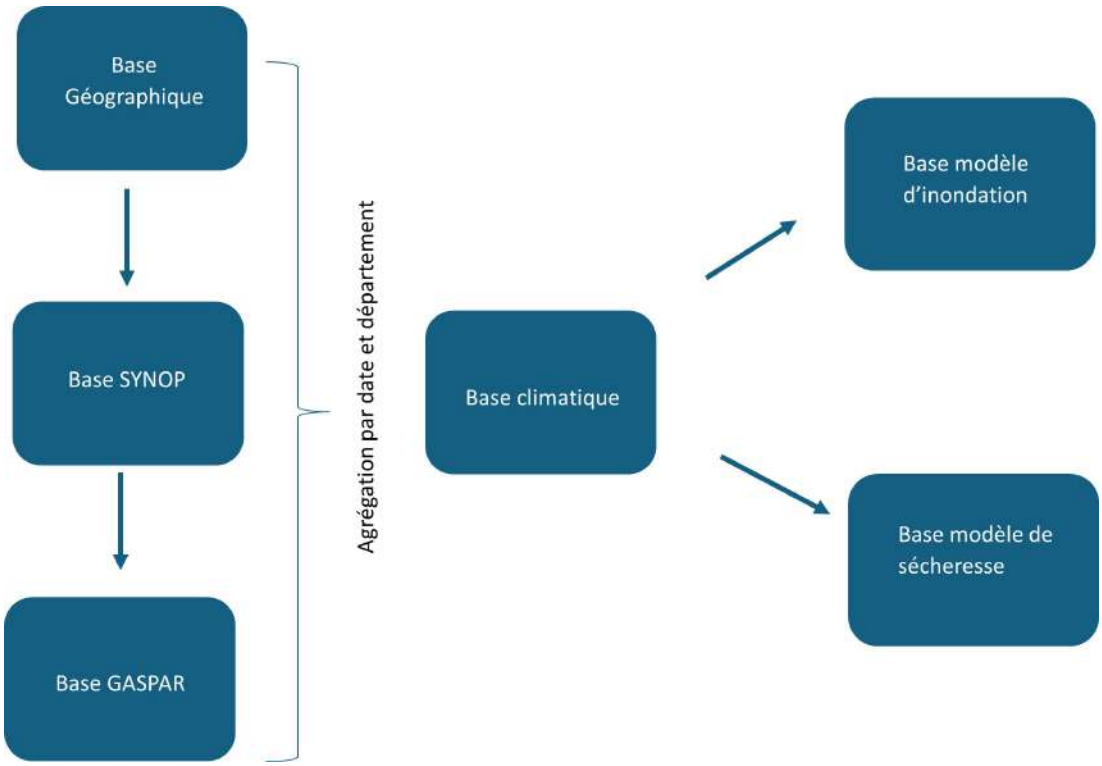


FIGURE 3.8 – Construction des bases pour la classification

Les bases de données étant maintenant constituées, les modèles permettant de calculer les dérives de sinistralité pour les risques d'inondation, de sécheresse et de tempêtes peuvent désormais être ajustés. Dans un premier temps, les dérives de fréquence pour les risques d'inondation et de sécheresse seront déterminées, suivies par celle des tempêtes. Ensuite, la dérive de coût pour chacun de ces sinistres sera également évaluée.

Chapitre 4

Evaluation de la dérive de sinistralité des inondations et de la sécheresse

En première partie, il a été établi que **93 %** de la sinistralité du régime des catastrophes naturelles était due à la sécheresse et aux inondations. Par conséquent, seuls ces deux risques seront pris en compte pour déterminer la dérive de sinistralité future des risques du régime Cat-Nat. Il sera donc question dans ce chapitre de déterminer les dérives de fréquence pour les inondations et la sécheresse.

Pour ce faire,

1. Une **prévision des paramètres météorologiques** de température, d'humidité, de pression et de précipitations présents dans la base SYNOP sera réalisée à l'aide de **séries temporelles** pour les années allant de **2024 à 2028**.

2. Deux **modèles de classification** seront construits pour déterminer l'occurrence ou non d'un arrêté de catastrophe naturelles : l'un pour le risque d'inondation et l'autre pour le risque de sécheresse, en utilisant les données historiques des bases « Base-sécheresse » et « Base-inondation » construites dans le chapitre précédent à partir des bases GASPAR et SYNOP.

3. Une estimation de la sinistralité future pour les risques d'inondations et de sécheresse sera effectuée en utilisant les données projetées du premier point et les modèles construits au deuxième point. Cette prédiction permettra ensuite de déterminer la dérive de sinistralité future.

Avant de procéder à la modélisation, il a été décidé de regrouper la France en zones de risques homogènes, afin de réduire le coût de la modélisation. Ceci permet de réaliser le premier point, à savoir la prévision des paramètres météorologiques, par zone. Par la

suite, les variations moyennes observées au sein de chaque zone sont utilisées pour passer des prévisions par zone aux prévisions par département.
Le regroupement a été réalisé en utilisant l'**Analyse en Composantes Principales (ACP)** et la **Classification Ascendante Hiérarchique (CAH)**, sur les données de **températures journalières** observées au sein de chaque département de la France métropolitaine entre 2016 et 2023.

4.1 Construction du zonier

4.1.1 Analyse en Composantes Principales (ACP)

L'ACP est une technique fondamentale en statistiques exploratoires multidimensionnelles. Son but est de résumer l'information provenant de nombreuses variables en révélant des liens entre elles et en créant des groupes d'individus similaires. En ACP, les données sont numériques et se présentent sous forme de matrice avec n lignes et p colonnes. Dans notre contexte, les lignes correspondent aux départements et les colonnes aux dates d'observation. *[N.Jégou, 2020]*

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ x_{21} & \cdots & x_{2p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix}$$

où x_{np} représente la température observée au sein du département i pour la date p.

La matrice X est centrée et réduite, puis la matrice des variances-covariances qui lui est associée est calculée. Grâce à cette matrice des variances-covariances, la dispersion des données est analysée, et les facteurs permettant de réduire l'espace à une dimension plus petite, tout en préservant au maximum la configuration globale des individus, sont extraits.

La représentation des départements sur le plan factoriel offre la meilleure visualisation plane du nuage d'individus.

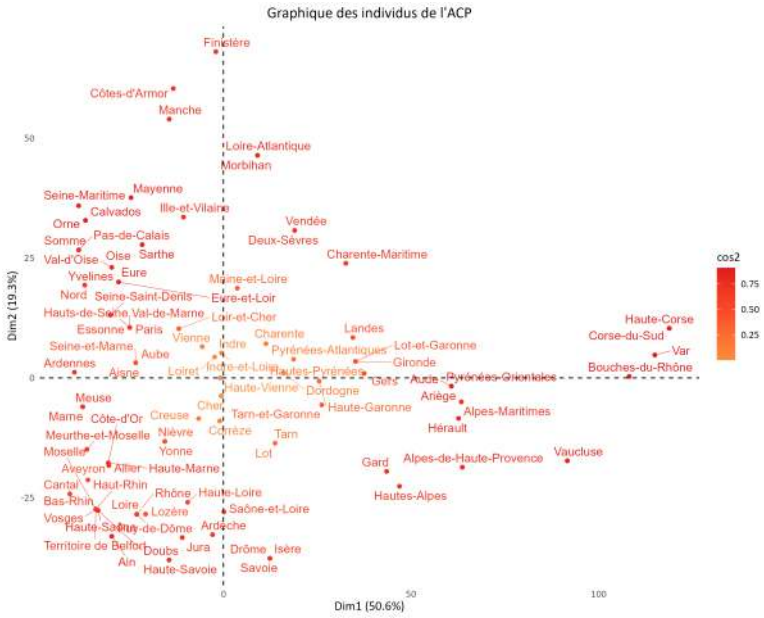


FIGURE 4.1 – Graphique des individus de l'ACP

Les axes factoriels sont des combinaisons linéaires des colonnes de X et sont orthogonaux entre eux. Les deux premiers axes expliquent **69,9 %** de l'inertie totale. Le premier axe semble opposer les départements en fonction de leur climat général avec à gauche de l'axe les départements de la façade atlantique et du nord de la France, connus pour des températures relativement modérées, avec peu de variations extrêmes, et à droite de l'axe les départements du sud-est de la France caractérisés par des étés chauds et secs et des hivers doux. Le second axe en revanche pourrait représenter des différences saisonnières ou d'altitude dans les températures.

L'un des critères permettant de déterminer le nombre de dimensions optimal pour représenter les départements est la *règle de coude*. Cette méthode consiste à identifier le point où le pourcentage d'inertie commence à diminuer beaucoup plus lentement. Le graphique suivant permet de déterminer ce nombre.

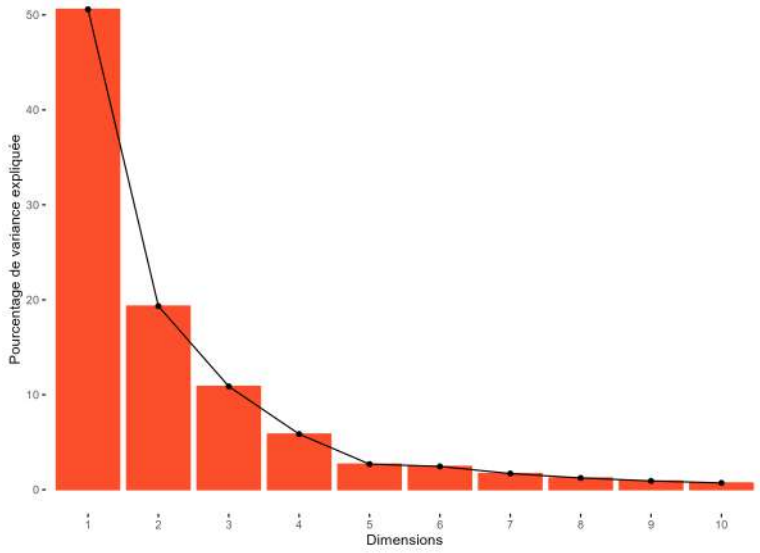


FIGURE 4.2 – Graphique des éboulis pour déterminer le nombre de dimensions optimal

D'après la règle du coude, le nombre de dimensions à prendre en compte pour ces données est de **deux**. Dans le désir de capturer une plus grande variabilité des données, notamment pour la suite des analyses, le choix de conserver **cinq** dimensions a été fait.

Les résultats de cette ACP sont ensuite utilisés pour effectuer une Classification Ascendante Hiérarchique (CAH), afin de regrouper les départements en zones de risques homogènes.

4.1.2 Classification Ascendante Hiérarchique (CAH)

La classification ascendante hiérarchique est une méthode itérative de classification non-supervisée. Elle se fonde sur le calcul de la dissimilarité entre N objets (dans ce cas, des départements) et procède par regroupement des deux objets qui minimisent un critère d'agrégation défini, formant ainsi une nouvelle classe. Ensuite, la dissimilarité entre cette nouvelle classe et les $N-2$ autres objets est recalculée en utilisant le même critère d'agrégation. Le processus continue en regroupant les deux objets ou classes dont l'union minimise à nouveau le critère d'agrégation. Ces étapes se répètent jusqu'à ce que tous les objets soient regroupés. *[Lumivero, 2024]*

La méthode d'agrégation utilisée dans le cadre de ce mémoire est la méthode de **Ward**. Cette technique consiste à regrouper, à chaque itération, les deux objets qui minimisent la variance intra-groupe, ce qui équivaut à réduire l'inertie inter-groupe. *[JMP, 2018]*

$$D_{KL} = \frac{\left\|\overline{\mathbf{x}}_K - \overline{\mathbf{x}}_L\right\|^2}{\frac{1}{N_K} + \frac{1}{N_L}}$$

— n est le nombre d'observations
— v est le nombre de variables
— x_i est la i-ème observation
— C_K est le K-ième cluster, sous-ensemble de $\{1, 2, \ldots, n\}$
— N_K est le nombre d'observations dans C_K
— $\mathbf{x}$ est le vecteur moyen de l'échantillon
— $\mathbf{x}_K$ est le vecteur moyen pour le cluster C_K
— $\|\mathbf{x}\|$ est la racine carrée de la somme des carrés des éléments de $\mathbf{x}$ (la longueur euclidienne du vecteur $\mathbf{x}$)
— $d(x_i, x_j)$ est $\|\mathbf{x}_i - \mathbf{x}_j\|^2$

Le processus successif de regroupements produit un dendrogramme, dont la racine représente la classe englobant tous les individus. Le dendrogramme illustre une hiérarchie de partitions, permettant de choisir une partition en coupant l'arbre à un certain niveau. La classification obtenue grâce aux résultats de l'ACP donne le partitionnement suivant :

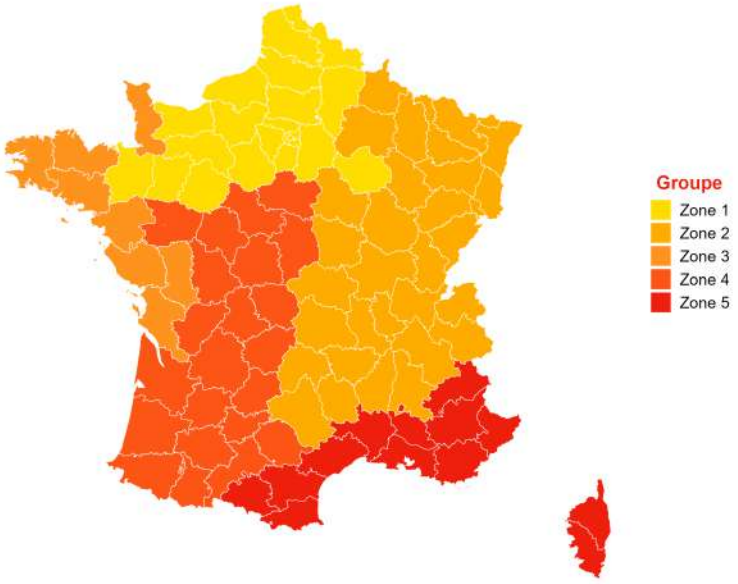


FIGURE 4.3 – Zonier pour l'obtention de la dérive du régime Cat-Nat

Ayant défini les différentes zones, les paramètres météorologiques (température, humidité, pression et précipitations) sont calculés pour chacune de ces zones en prenant la **moyenne** des mesures des différents départements qui les composent. Cela permet de constituer une base de données contenant les observations météorologiques journalières pour chaque zone, couvrant la période de 2016 à 2023.

Il sera maintenant question d'aborder le premier point mentionné en introduction de ce chapitre, à savoir la prévision des paramètres météorologiques à l'aide de séries temporelles.
Avant cela, revenons sur la théorie des séries temporelles.

4.2 Séries Temporelles

Une série temporelle est une suite d'observations indexée par le temps. Plusieurs composantes peuvent généralement être identifiées dans une série temporelle :*[AILLIOT, 2024]*

- La **tendance** qui correspond à l'évolution à long terme des observations
- Les **composantes saisonnières**, qui représentent les variations périodiques.
- Un **résidu** qui semble avoir un comportement aléatoire

Pour modéliser ces différentes composantes, un **modèle additif** est souvent utilisé :

$$y_t = T_t + S_t + x_t \tag{4.1}$$

pour $t \in \{1, \dots, T\}$ avec

- $\{T_t\}$ la tendance,
- $\{S_t\}$ la composante saisonnière et
- $\{x_t\}$ le résidu sans tendance ni saisonnalité.

Un **modèle multiplicatif** peut également être utilisé. Il a pour forme

$$y_t = T_t * S_t * x_t$$

Les composantes $\{T_t\}$ et $\{S_t\}$ sont généralement supposées **déterministes** (c'est-à-dire sans composante aléatoire) tandis que $\{x_t\}$ est la réalisation d'un processus aléatoire $\{X_t\}$. Comme $\{x_t\}$ est une série temporelle sans tendance ni saisonnalité, le processus $\{X_t\}$ est supposé *stationnaire*. Avec ces hypothèses, $\{y_t\}$ est également la réalisation d'un processus $\{Y_t\}$ qui n'est généralement pas stationnaire.

Un processus $X = (X_t)_{t \in I}$ est ***stationnaire*** (au second ordre) si

- $\forall t \in \mathbb{N}, \mathbb{E}(X_t) = \mu$, l'espérance n'évolue pas au cours du temps,
- $\forall (t, h) \in \mathbb{N}^2, \text{cov}(X_t, X_{t+h}) = \gamma(h)$, les covariances n'évoluent pas au cours du temps.

Étapes dans la modélisation d'une série temporelle

1. Modélisation des composantes non-stationnaires.

La modélisation d'une série temporelle commence par la modélisation des composantes non-stationnaires. Cette étape vise en particulier à « stationnariser » la série temporelle initiale $\{y_t\}$. Si on suppose que l'équation 4.1 est valide, on cherchera à estimer les fonctions $\{T_t\}$ et $\{S_t\}$.
Il existe différentes manières pour stationnariser une série temporelle :

- Modèle paramétrique
- Différentiation et modèle SARIMA
- Lissage paramétrique

La méthode présentée et utilisée dans ce mémoire est celle du Modèle paramétrique.

Modèle Paramétrique.

Une forme paramétrique est choisie pour les fonctions $\{T_t\}$ et $\{S_t\}$.

- La modélisation de la tendance $\{T_t\}$ peut être faite au moyen d'un **polynôme de degré** r c'est-à-dire en supposant que

$$T_t = \nu_0 + \nu_1 t + \ldots + \nu_r t^r.$$

- La saisonnalité $\{S_t\}$ est généralement représentée en utilisant un **polynôme trigonométrique** de la forme

$$S_t = \kappa_0 + \kappa_1 \cos\left(\frac{2\pi t}{D}\right) + \lambda_1 \sin\left(\frac{2\pi t}{D}\right) + \kappa_2 \cos\left(\frac{4\pi t}{D}\right) + \lambda_2 \sin\left(\frac{4\pi t}{D}\right) + \ldots$$

Les différents paramètres peuvent être estimés par la **méthode des moindres carrés**.

En retirant les composantes non-stationnaires estimées de la série temporelle initiale, on obtient une estimation de $\{x_t\}$.

2. Modélisation des composantes stationnaires.

Par la suite, la composante stationnaire $\{X_t\}$ est modélisée à l'aide des **Modèles autoregressifs et moyenne mobile (ARMA)**. Un processus *stationnaire* X suit un modèle ARMA(p,q) s'il vérifie l'équation suivante :

$$X_t - \alpha_1 X_{t-1} - \ldots - \alpha_p X_{t-p} = \epsilon_t + \beta_1 \epsilon_{t-1} + \ldots + \beta_q \epsilon_{t-q} \tag{4.2}$$

pour $t \in \mathbb{Z}$ avec ϵ_t un *bruit blanc* tel que $\text{var}(\epsilon_t) = \sigma^2$ et $(\alpha_1, \ldots, \alpha_p, \beta_1, \ldots, \beta_q, \sigma)$ les paramètres du modèle. Lorsque $q = 0$, on obtient le modèle AR(p) [1] et lorsque $p = 0$, on

1. Autorégressif d'ordre p

obtient le modèle MA(q) [2].

On appelle ***bruit blanc*** un processus stationnaire centré $(\epsilon_t)_{t\in\mathbb{Z}}$ tel que la fonction d'autocovariance γ vérifie
- $\gamma(0) = \text{var}(\epsilon_t) < +\infty$ et
- $\gamma(h) = 0$ si $h \neq 0$.

En particulier, si $(\epsilon_t)_{t\in\mathbb{Z}}$ est une suite de variables aléatoires i.i.d [3] avec
- $\mathbb{E}[\epsilon_t] = 0$ et
- $\text{var}(\epsilon_t) < +\infty$, alors $(\epsilon_t)_{t\in\mathbb{Z}}$ est un bruit blanc (ou *bruit blanc fort*). Si les variables aléatoires sont gaussiennes, on parle de *bruit blanc gaussien.*

4.2.1 Modélisation de la température

Afin de faciliter la lisibilité des résultats, seuls les graphiques de la zone 5 seront présentés dans cette section, la même méthode a été appliquée aux autres zones. La série temporelle a été séparée en :
- une base d'**apprentissage** qui contient les observations de **2016 à 2022**,
- une base de **test** contenant les observations de **2023**.

Dans un premier temps, les données ont été représentées graphiquement afin d'identifier une éventuelle tendance ou saisonnalité. Le graphique montre une saisonnalité de forme **trigonométrique**. Cependant, il est difficile de discerner une tendance claire des températures dans cette zone à partir de ce seul graphique.

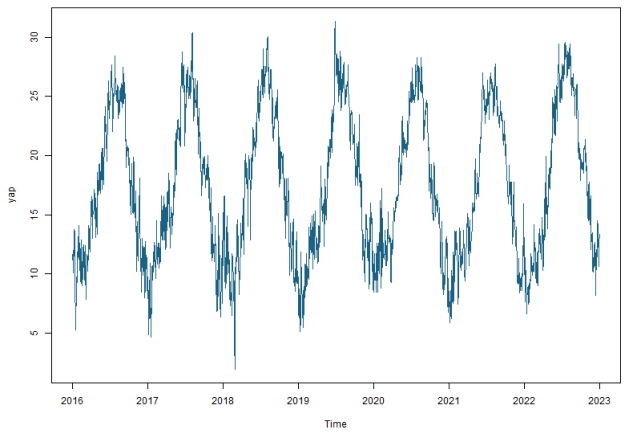


FIGURE 4.4 – Températures observées au sein de la zone 5

Afin d'obtenir une meilleure vision de la série temporelle, la fonction *stl* de R a

2. Moyenne mobile d'ordre q
3. indépendantes et identiquement distribuées

été utilisée pour décomposer la série sous la forme d'un modèle additif. Malgré cette décomposition, il reste difficile de définir une fonction précise pour la tendance.

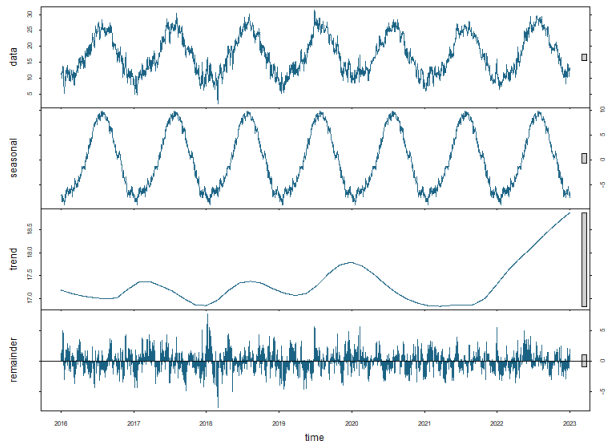


FIGURE 4.5 – Décomposition de la série temporelle initiale

Divers tests ont été effectués pour déterminer quel ordre de polynôme permettrait d'obtenir des coefficients de tendance significatifs. Ceci a été réalisé en ajustant des modèles linéaires avec des ordres de polynômes différents. À l'issue de ces tests, il a été conclu que la **tendance linéaire** serait la plus appropriée pour modéliser la tendance de la température de la zone 5, ainsi que celle des autres zones, car seule la p-valeur du coefficient du polynôme d'ordre 1 était significative .

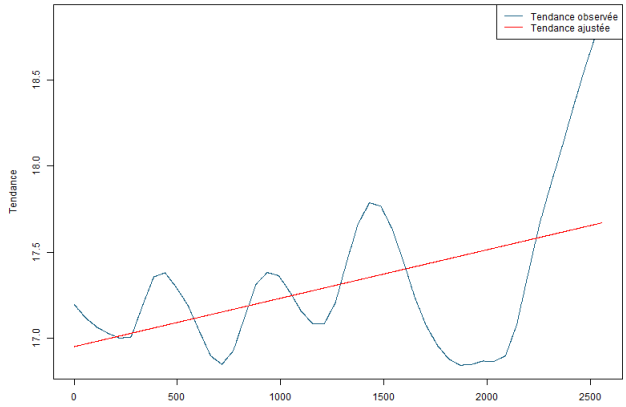


FIGURE 4.6 – Tendance ajustée sur les températures de la zone 5

La modélisation de la composante non-stationnaire de la série temporelle a été effectuée à l'aide d'une régression linéaire, intégrant une **tendance linéaire** et une **saison-**

nalité trigonométrique, en utilisant un polynôme d'ordre 2. Le modèle obtenu est le suivant :

$$T_t + S_t = 16,95 + 0,90t - 7,48\cos(2\pi t) - 3,37\sin(2\pi t) + 0,44\cos(4\pi t) + 0,99\sin(4\pi t) \tag{4.3}$$

Le coefficient de détermination (R^2) de ce modèle s'élève à **90,83 %**, indiquant que la variance de la tendance et de la saisonnalité des températures est bien capturée par le modèle. Le premier graphique de la figure 4.7 montre que la tendance et la saisonnalité s'ajustent bien aux données.

	Zone 1	Zone 2	Zone 3	Zone 4	Zone 5
R^2	75,78 %	81,04 %	77,21 %	76,72 %	90,83 %
MAE	1,40	1,33	1,17	1,44	0,84
RMSE	1,80	1,73	1,52	1,85	1,11

TABLE 4.1 – Evaluation de la qualité d'ajustement des composantes non-stationnaires sur la base d'entraînement

La partie non-stationnaire étant modélisée, il est désormais possible de modéliser la partie stationnaire, c'est-à-dire les **résidus**. Pour obtenir ces résidus, la tendance et la saisonnalité sont soustraites de la série temporelle originale :

$$X_t = Y_t - T_t - S_t$$

où Y_t représente les données de températures
La seconde partie du graphique suivant montre l'allure des résidus.

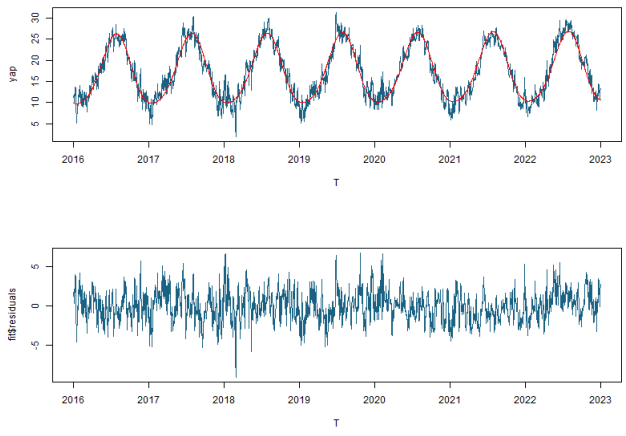
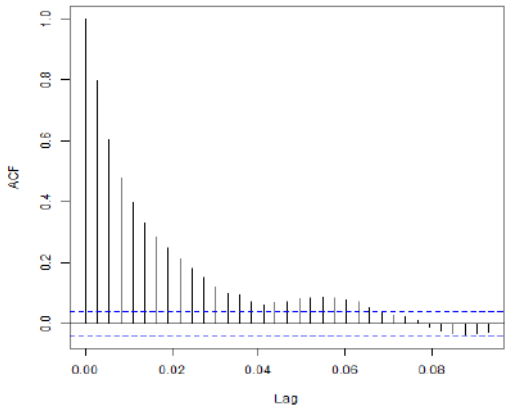
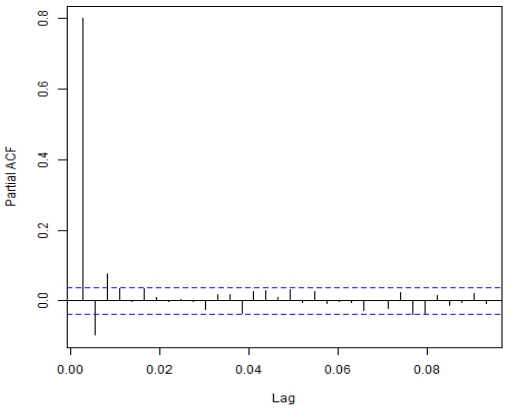


FIGURE 4.7 – Modélisation de la tendance et la saisonnalité à l'aide d'un modèle paramétrique

La fonction d'auto-corrélation empirique montre que les résidus X_t ne peuvent être assimilés à un bruit blanc. En revanche, elle décroît très rapidement et la petitesse de la fonction d'auto-corrélation partielle suggère que X_t peut être interprété comme une réalisation d'un modèle ARMA.



(a) ACF des résidus de la zone 5



(b) PACF des résidus de la zone 5

En supposant que X est un processus stationnaire gaussien, la fonction *auto.arima* de R a permis d'obtenir comme modèle pour X_t le modèle $ARMA(1,3)$.

$$X_t = 0,84X_{t-1} + \epsilon_t + 0,04\epsilon_{t-1} - 0,13\epsilon_{t-2} - 0,08\epsilon_{t-3}$$

	Zone 1	Zone 2	Zone 3	Zone 4	Zone 5
Modèle ajusté	ARMA(1,1)	ARMA(3,0)	ARMA(1,1)	ARMA(4,0)	ARMA(1,3)

TABLE 4.2 – Modèles ARMA ajustés par zone

Pour valider ce modèle, il est nécessaire d'étudier ses résidus. En effet, il est préférable que ϵ_t soit un bruit blanc gaussien. Il faut donc vérifier que les ϵ_i sont **indépendants** et que $\forall t,\ \varepsilon_t \sim \mathcal{N}(0, \sigma^2)$.

1. **Hypothèse d'indépendance :**
   Le **test de Portmanteau** (dont la théorie est disponible en annexes) avec la statistique de Ljung-Box conduit à ne pas rejeter l'hypothèse d'indépendance des résidus, avec une p-valeur de $0,999$.
2. **Hypothèse de normalité :**
   Une représentation graphique quantile-quantile montre que les queues de la distribution des résidus ϵ obtenus sont plus lourdes que celles d'une loi normale.

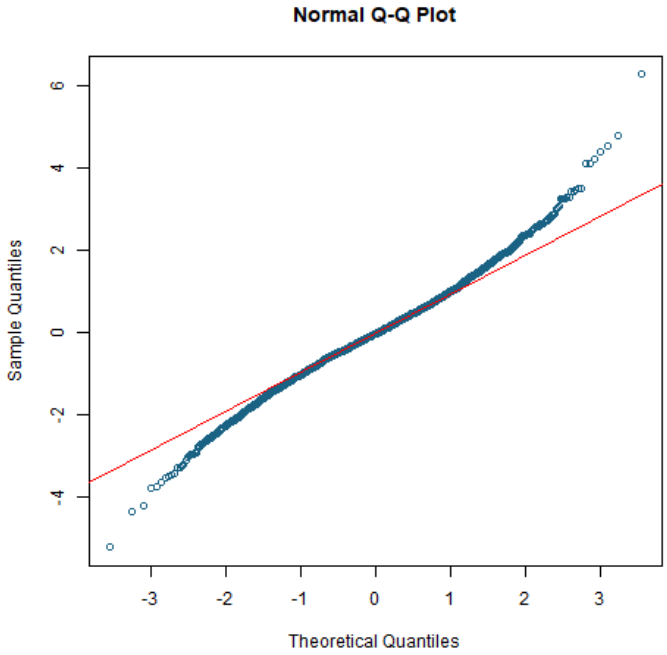


FIGURE 4.9 – QQ-plot de la distribution des résidus du modèle ARMA(1,3)

Ce constat est confirmé par le test de **Shapiro-Wilk**, qui rejette l'hypothèse de normalité avec une p-valeur de 4.515×10^{-13}. Divers modèles ont été envisagés pour pallier ce problème de non-normalité, notamment l'utilisation de la distribution de Student pour les résidus ainsi que les modèles GARCH [4]. Cependant, ces approches n'ont pas été concluantes. Cette non-normalité pourrait affecter la fiabilité des prévisions du modèle.

Validation du modèle

Avant de générer des prévisions pour les températures futures, il est essentiel d'évaluer la capacité prédictive du modèle sur de nouvelles données. À cet effet, une validation a été réalisée en utilisant les données de **2023**. La tendance et la saisonnalité ont été ajustées à l'aide de l'équation 4.3, suivie par l'utilisation du modèle ARMA pour générer 10 000 trajectoires des résidus, permettant ainsi la création de 10 000 simulations des températures journalières. Ces simulations ont servi à construire un intervalle de confiance.

Les résultats des métriques de performance révèlent un écart entre celles obtenues sur l'ensemble d'apprentissage (Tableau 4.1) et celles obtenues sur l'ensemble de test. Les métriques sont plus élevées sur la base de test, ce qui pourrait être attribué à des variations non capturées de la tendance, la saisonnalité ou des résidus par le modèle,

4. Generalized Autoregressive Conditional Heteroskedasticity

ou encore à des différences intrinsèques entre les données de test et celles utilisées pour l'entraînement.

	Zone 1	Zone 2	Zone 3	Zone 4	Zone 5
MAE	2,05	2,13	1,66	2,25	1,11
RMSE	2,54	2,67	2,21	2,83	1,52

TABLE 4.3 – Métriques pour la validation du modèle sur les données 2023

Le graphique comparatif entre les températures observées en 2023 et celles simulées par le modèle montre que la majorité des observations se situent à l'intérieur de l'intervalle de confiance à 97,5 %, représenté en gris. La courbe rouge, qui représente la moyenne des simulations, suit bien les données observées, suggérant une bonne adéquation entre les prédictions du modèle et les températures réelles. Cette analyse graphique indique que, malgré une surestimation occasionnelle des erreurs individuelles, le modèle capture globalement la tendance des données

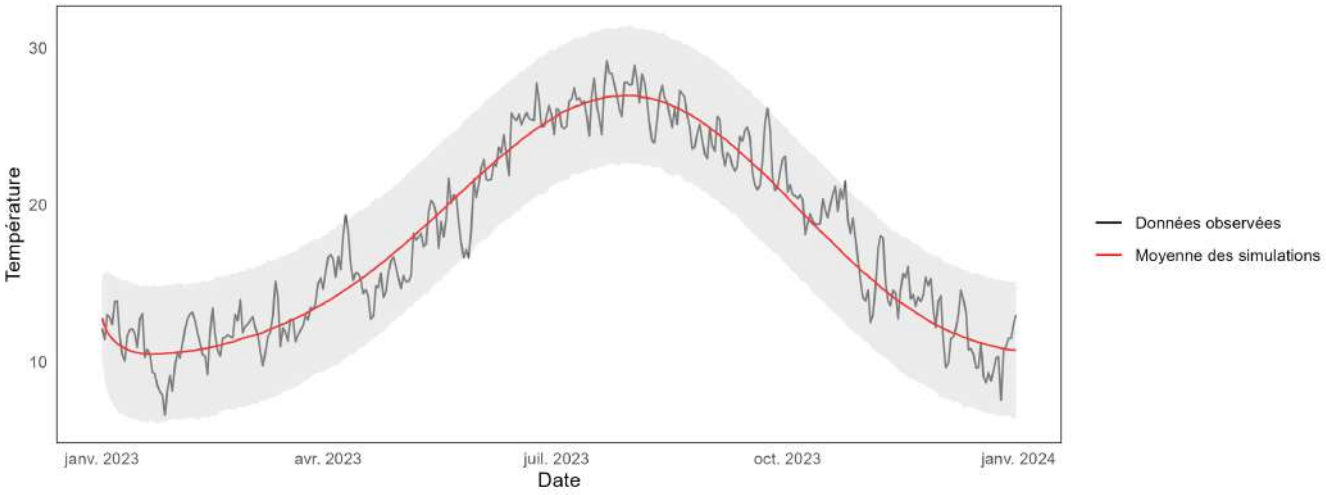


FIGURE 4.10 – Validation du modèle sur les données observées de 2023

Prévision des températures futures à l'horizon 2028 par zone

La méthode utilisée pour obtenir une simulation des températures de 2023 explicitée dans la partie précédente a été utilisée sur une durée plus longue cette fois ci pour obtenir une prévision des températures de 2024 à 2028. Le graphique suivant montre les prévisions obtenues pour la zone 5.

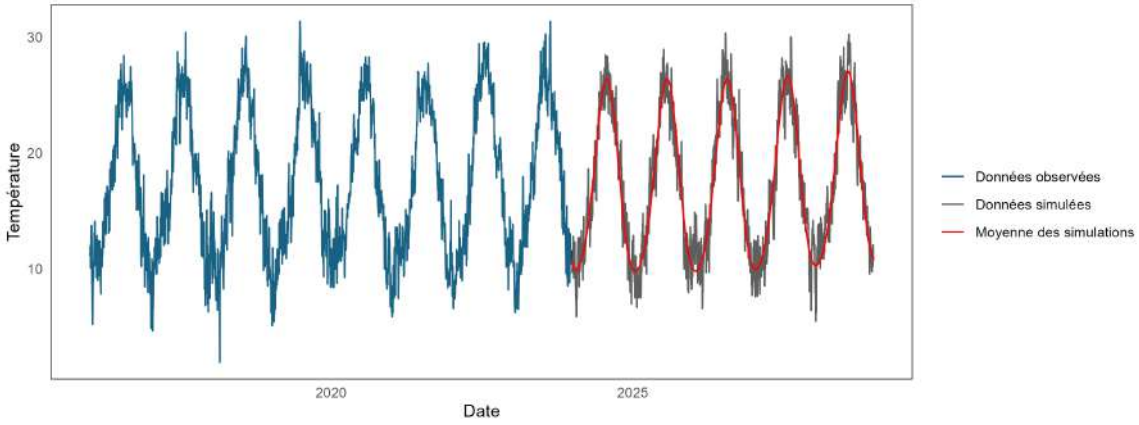


FIGURE 4.11 – Prévision de la température future dans la zone 5

Comme le montre le tableau suivant, les températures moyennes prévues pour la période 2024-2028 indiquent une augmentation progressive de la température par rapport aux années précédentes.

	Zone 1	Zone 2	Zone 3	Zone 4	Zone 5
Température moyenne 2016-2023 (°C)	13,17	13,93	14,42	14,76	17,35
Température moyenne 2024-2028 (°C)	14,00	14,68	15,19	15,44	17,59
Ecart en %	**5,93**	**5,11**	**5,06**	**4,42**	**1,39**

TABLE 4.4 – Comparaison de la température moyenne observée et celle prédite entre 2024 et 2028

Une projection jusqu'à 2050 de ces températures montre une augmentation moyenne de 2,5°C de la température d'ici là. Ce résultat est en ligne avec le scénario SSP3-7.0 du GIEC (1.2), qui prévoit une augmentation de la température de 4°C d'ici 2100.

	Zone 1	Zone 2	Zone 3	Zone 4	Zone 5
Augmentation à horizon 2050 (°C)	1,71	1,71	1,74	1,89	1,11

TABLE 4.5 – Augmentation de température à horizon 2050

La même méthodologie a été utilisée pour obtenir une prévision pour **l'humidité**, la **pression** et la **précipitation**.
Les résultats sont présentés dans les tableaux suivants :

	Zone 1	Zone 2	Zone 3	Zone 4	Zone 5
Humidité moyenne 2016-2023 (%)	77,83	72,92	79,10	75,23	67,94
Humidité moyenne 2024-2028 (%)	76,47	71,50	78,38	74,41	69,95
Ecart en %	**-1,78**	**-1,98**	**-0,91**	**-1,10**	**2,87**

TABLE 4.6 – Comparaison de l'humidité moyenne observée et celle prédite entre 2024 et 2028

	Zone 1	Zone 2	Zone 3	Zone 4	Zone 5
Pression moyenne 2016-2023 (Pa)	100 413,68	97 836,19	101 112,91	99 171,18	99 618,80
Pression moyenne 2024-2028 (Pa)	100 469,27	97 838,18	101 180,67	99 224,20	99 600,58
Ecart en %	**0,06**	**0,00**	**0,07**	**0,02**	**-0,02**

TABLE 4.7 – Comparaison de la pression moyenne observée et celle prédite entre 2024 et 2028

	Zone 1	Zone 2	Zone 3	Zone 4	Zone 5
Précipitation moyenne 2016-2023 (mm)	1,91	1,86	1,99	2,14	1,66
Précipitation moyenne 2024-2028 (mm)	1,87	1,93	1,95	2,12	1,72
Ecart en %	**-2,17**	**3,91**	**-2,51**	**-1,10**	**3,23**

TABLE 4.8 – Comparaison de la précipitation moyenne observée et celle prédite entre 2024 et 2028

Prévision à l'horizon 2028 par département

L'hypothèse selon laquelle la moyenne des variations de chaque paramètre pour chaque zone est représentative des variations observées au sein de chaque département est posée. Cette hypothèse permet d'utiliser les écarts obtenus au sein de chaque zone pour déterminer les paramètres météorologiques de chaque département.

Une fois ce passage effectué, passant des prévisions futures par zones aux prévisions futures par département, il est important de comparer la corrélation entre les différents paramètres sur les deux bases. Les tableaux suivants montrent les corrélations entre les différents paramètres météorologiques sur les données historiques dans un premier temps, et sur les données projetées par la suite.

	Température	Humidité	Précipitation	Pression
Température	1,00	-0,52	-0,11	-0,02
Humidité	-0,52	1,00	0,49	0,10
Précipitation	-0,11	0,49	1,00	-0,16
Pression	-0,02	0,10	-0,16	1,00

TABLE 4.9 – Corrélation de Pearson observée entre les paramètres météorologiques : **Données historiques**

	Température	Humidité	Précipitation	Pression
Température	1,00	-0,47	-0,03	-0,04
Humidité	-0,47	1,00	0,47	0,11
Précipitation	-0,03	0,47	1,00	-0,18
Pression	-0,04	0,11	-0,18	1,00

TABLE 4.10 – Corrélation de Pearson observée entre les paramètres météorologiques : **Prévisions 2024 à 2028**

Les écarts observés entre les données réelles et les données projetées sont généralement faibles, ce qui suggère que les données projetées capturent bien les relations entre les différentes variables climatiques. Cependant, certaines différences peuvent être notées : notamment la relation entre la température et la précipitation qui est plus faible dans les données projetées.

Ces données seront ensuite utilisées dans le cadre d'un modèle de classification pour estimer le nombre d'arrêtés de catastrophes naturelles relatifs aux risques d'inondation et de sécheresse au cours des cinq prochaines années. Cela permettra d'obtenir la dérive de sinistralité pour ces risques. Avant de procéder à la classification, il est important de revoir la théorie qui la sous-tend.

4.3 Classification des sinistres

4.3.1 Modèles Linéaires Généralisés

La régression est une technique statistique utilisée pour estimer un vecteur de paramètres θ à partir d'un ensemble de données observées Y en minimisant l'erreur quadratique. Dans le cas où la variable observée suit une distribution normale, cette minimisation de l'erreur quadratique équivaut à maximiser la fonction de vraisemblance, faisant ainsi de la régression une application de la méthode du maximum de vraisemblance.

Les Modèles Linéaires Généralisés (GLM), introduits par John Nelder et Robert Wedderburn, étendent l'estimation par maximum de vraisemblance à des distributions autres que la loi normale, regroupées sous le terme de famille exponentielle. *[LAILY, 2023]*

$$g\left(\mathbb{E}(Y|X)\right) = \beta_0 + \beta_1 \cdot X_1 + \ldots + \beta_p \cdot X_p + \epsilon$$

Les GLM sont définis par trois éléments :

— **Une composante aléatoire** : Il s'agit de la loi de probabilité adossée à la variable cible Y. Elle doit être une loi de la famille exponentielle. La densité d'une loi appartenant à la famille exponentielle s'écrit sous la forme :

$$f_{\theta,\phi}(y) = \exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right]$$

Avec $a(.), b(.)$ et $c(.)$ des fonctions, θ le paramètre naturel de la famille exponentielle et ϕ le paramètre de dispersion

— **Une composante déterministe** : Elle permet d'associer à chaque observation un prédicteur linéaire noté η_i, telle que

$$\forall i \in \{1, \ldots, n\}, \eta_i = \beta_0 + \beta_1 \cdot x_{i,1} + \ldots + \beta_p \cdot x_{i,p}$$

— **Une fonction lien g** : Elle permet de relier la variable cible à l'ensemble des variables explicatives telle que

$$g\left(\mathbb{E}(Y|X)\right) = \beta_0 + \beta_1 \cdot X_1 + \ldots + \beta_p \cdot X_p$$

Quelques GLM fréquemment utilisés sont :

Loi	Support	Fonction de lien	Expression du modèle
Normale	$\mathbb{R}$	$g(x) = x$	$y_i = \beta_0 + \beta_1 \cdot x_{i,1} + \cdots + \beta_p \cdot x_{i,p} + \epsilon$
Bernoulli	$\{0,1\}$	$g(x) = \ln\left(\frac{x}{1-x}\right)$	$y_i = \left(1 + \exp\left(-[\beta_0 + \beta_1 \cdot x_{i,1} + \cdots + \beta_p \cdot x_{i,p} + \epsilon]\right)\right)^{-1}$
Poisson	$\mathbb{N}$	$g(x) = \ln(x)$	$y_i = \exp(\beta_0 + \beta_1 \cdot x_{i,1} + \cdots + \beta_p \cdot x_{i,p} + \epsilon)$
Gamma	$\mathbb{R}^+$	$g(x) = \frac{1}{x}$	$y_i = \left(\beta_0 + \beta_1 \cdot x_{i,1} + \cdots + \beta_p \cdot x_{i,p} + \epsilon\right)^{-1}$

TABLE 4.11 – Exemples de GLM

La classification à l'aide de GLM est faite au travers du modèle de **régression logistique**.

4.3.2 Régression Logistique

Le modèle de régression logistique, également connu sous le nom de modèle GLM de **Bernoulli**, est un modèle linéaire généralisé utilisé pour les **variables binaires**. L'expression de ce modèle est présentée dans le tableau 4.11. Dans le cadre de ce mémoire, ce modèle est utilisé pour prédire, en fonction des paramètres météorologiques journaliers, l'occurrence d'un arrêté de catastrophe naturelle lié pour les risques de sécheresse et d'inondation.

Afin d'évaluer la performance du modèle de régression logistique, les métriques usuelles sont : *[ICHI.PRO, 2020]*

1. **Accuracy** : Mesure la proportion de prédictions correctes parmi toutes les prédictions faites. C'est une mesure globale de la performance du modèle.

$$\text{Accuracy} = \frac{\text{Nombre de prédictions correctes}}{\text{Nombre total de prédictions}}$$

2. **Précision** : Évalue la proportion des prédictions positives correctes parmi toutes les prédictions positives.

$$\text{Précision} = \frac{\text{Vrais positifs}}{\text{Vrais positifs} + \text{Faux positifs}}$$

3. **Rappel** : Mesure la proportion des vrais positifs capturés parmi tous les éléments positifs réels.

$$\text{Rappel} = \frac{\text{Vrais positifs}}{\text{Vrais positifs} + \text{Faux négatifs}}$$

4. **Spécificité** : Mesure la proportion de vrais négatifs correctement identifiés parmi les vrais négatifs réels

$$\text{Rappel} = \frac{\text{Vrais négatifs}}{\text{Vrais négatifs} + \text{Faux positifs}}$$

5. **F1-Score** : Moyenne harmonique de la précision et du rappel.

$$F1 = 2 \cdot \frac{\text{Précision} \cdot \text{Rappel}}{\text{Précision} + \text{Rappel}}$$

6. **AUC-ROC (Area Under the Receiver Operating Characteristic Curve)** : L'AUC-ROC mesure la performance globale d'un modèle de classification en considérant toutes les valeurs possibles du seuil de décision. Une AUC proche de 1 indique un excellent modèle, tandis qu'une AUC de 0,5 indique un modèle sans pouvoir discriminant.

$$\text{AUC-ROC} = \int_0^1 TPR \cdot d(FPR)$$

Ces métriques sont calculées à partir de la matrice de confusion. La **matrice de confusion** est un outil permettant de visualiser la performance d'un algorithme de classification en comptant les occurrences des vrais positifs (TP), faux positifs (FP), vrais négatifs (TN) et faux négatifs (FN).

	Prédit Négatif	Prédit Positif
Réel Négatif	TN	FP
Réel Positif	FN	TP

Modélisation

Les bases de données utilisées pour cette modélisation sont intitulées « base sécheresse » et « base inondation ». Elles résultent de la fusion des bases SYNOP et GASPAR et contiennent les observations journalières des paramètres météorologiques entre 2016 et 2023. Les variables présentes dans ces bases incluent :

- le département
- la date d'observation
- la température en °C
- l'humidité en %
- la pression au niveau de la station en Pa
- les précipitations des dernières 24 heures en mm
- la variable *Sinistre*, qui vaut 1 lorsqu'un arrêté de catastrophe naturelle est déclaré, et 0 sinon

Les statistiques descriptives des paramètres météorologiques sont résumées dans le tableau ci-dessous :

Paramètre	Minimum	1er Quartile	Médian	Moyenne	3ème Quartile	Maximum
Température (°C)	-10,8	8,5	12,9	13,3	18,2	34,0
Humidité (%)	1,0	66,1	76,1	74,5	84,6	100,0
Pression (Pa)	88868,0	98774,0	100330,0	99624,0	101255,0	104611,0
Précipitation (mm)	0,0	0,0	0,2	1,9	1,9	105,7

TABLE 4.12 – Statistiques descriptives des paramètres météorologiques

Les données initiales présentant des échelles variées, pouvant introduire des biais dans la modélisation, ont été **normalisées** à l'aide de la méthode **min-max**, alignant ainsi toutes les variables sur une échelle commune. L'analyse de la **corrélation de Spearman** entre les variables quantitatives a été effectuée.

Les résultats n'indiquent pas de corrélations fortes entre les variables, suggérant qu'elles peuvent toutes être intégrées dans le modèle sans risque de multicolinéarité significative.

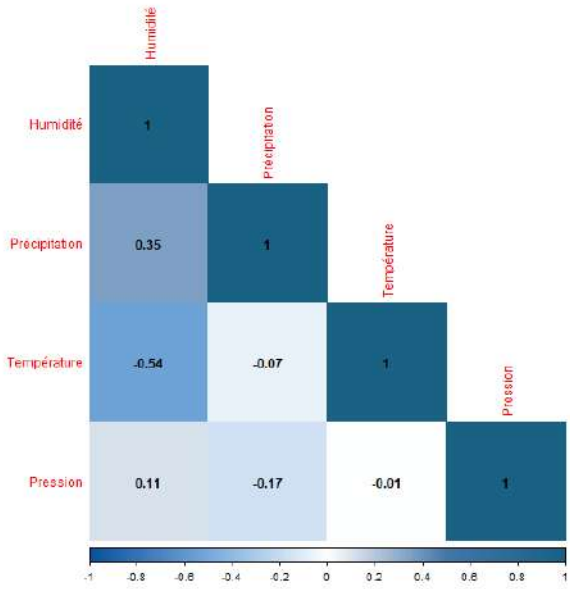


FIGURE 4.12 – Corrélation entre les variables numériques

À l'exception de la date d'observation, toutes les autres variables ont été intégrées dans la modélisation. La sélection des variables a été faite à l'aide de la fonction *stepAIC* sur R. Celle-ci permet d'identifier un sous-ensemble de variables explicatives qui donne le meilleur modèle en termes de parsimonie et de qualité d'ajustement grâce au critère d'AIC.

Risque de sécheresse

Dans la première partie, il a été établi que la sécheresse, étant un risque de longue durée, influence la déclaration de l'état de catastrophe naturelle. En raison de cette caractéristique, l'année 2023 ne sera pas incluse dans la modélisation afin d'éviter tout biais, étant donné que le nombre d'arrêtés pour 2023 est susceptible de changer considérablement au cours de l'année. Les données utilisées pour évaluer ce risque couvriront donc la période de **2016 à 2022**.

La variable cible *Sinistre* présente un déséquilibre significatif, où seulement **7 %** des données correspondent à la déclaration d'un arrêté de catastrophe naturelle. Ce déséquilibre est considéré comme représentatif de la rareté des événements climatiques extrêmes et n'a pas été corrigé dans le modèle. Cette décision vise à assurer que les prédictions reflètent fidèlement la réalité observée.

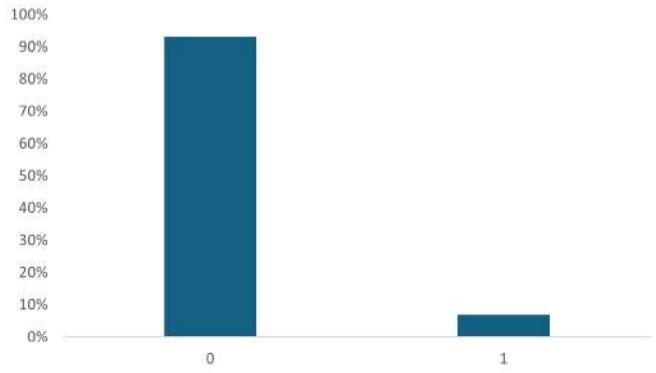


FIGURE 4.13 – Répartition de la variable cible *Sinistre*

Les données ont été séparées en deux : 75 % pour entraîner le modèle et 25 % pour le valider. Le modèle de régression logistique a été ajusté aux données d'entraînement. Les résultats obtenus sur l'ensemble de test sont présentés dans le tableau ci-dessous :

Métrique	Valeur
Accuracy	93,06 %
Balanced accuracy	53,41 %
F1-score	12,97 %
Recall	7,37 %
Précision	53,17 %
AUC	88,28 %

TABLE 4.13 – Résultats des métriques de performance du modèle

Le taux de précision globale (*accuracy*) du modèle est de 93,06 %, mais cette métrique est influencée par le déséquilibre de classe. En effet, la *balanced accuracy*, qui corrige cet effet de déséquilibre, est beaucoup plus basse, à 53,41 %, ce qui met en évidence les difficultés du modèle à bien identifier les deux classes.

Le rappel (*recall*) est particulièrement bas à 7,37 %, ce qui révèle la difficulté du modèle à détecter la classe minoritaire. La précision de 53,17 % indique que, bien que le modèle soit capable de prédire la classe majoritaire, il échoue souvent à identifier correctement les exemples de la classe minoritaire. Ceci conduit à un F1-score de 12,97 % soulignant une faible performance du modèle dans l'équilibre entre la précision et le rappel. Malgré ces performances mitigées, l'AUC de 88,28 % montre une bonne capacité de discrimination globale entre les deux classes, bien que cette performance ne se traduise pas dans les autres métriques. La matrice de confusion illustre en détail ces performances contrastées du modèle.

La matrice de confusion montre un nombre élevé de vrais positifs (83 351) par rapport

aux faux positifs (411), mais aussi un nombre non négligeable de faux négatifs (5 911), crucial dans notre étude axée sur la détection précise des sinistres.

	Prédit Négatif	Prédit Positif
Réel Négatif	83 351	411
Réel Positif	5 911	466

TABLE 4.14 – Matrice de confusion de la régression logistique

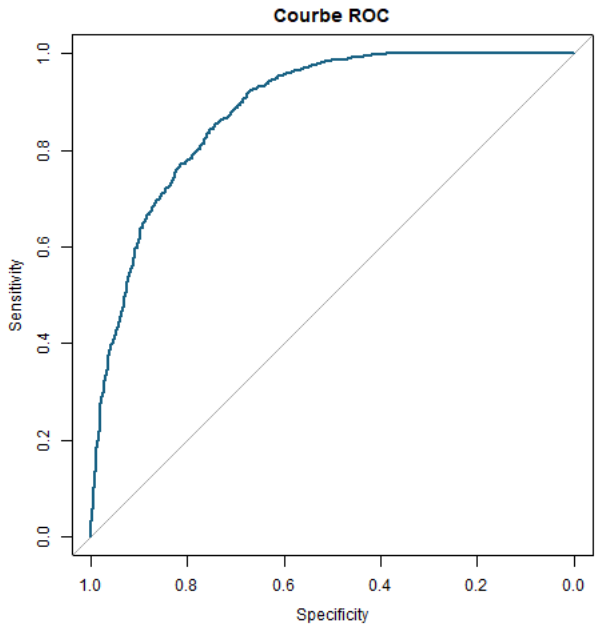


FIGURE 4.14 – Courbe ROC Régression Logistique pour le risque de sécheresse

Bien que le modèle actuel démontre des résultats prometteurs sur la classe majoritaire, l'exploration de méthodes alternatives comme les arbres de décision et les forêts aléatoires pourrait améliorer la sensibilité aux sinistres tout en maintenant une précision globale adéquate.

4.3.3 Les arbres de décision

Les arbres de décision font partie des modèles supervisés non-paramétriques. Introduits par Breiman et ses collaborateurs en 1984 sous l'acronyme CART (Classification and Regression Trees), ces modèles visent à diviser les individus en groupes homogènes par rapport à la variable à prédire. Le processus de regroupement est récursif jusqu'à ce que des nœuds terminaux soient atteints. Durant la phase d'apprentissage, l'objectif est de trouver les questions sur la variable explicative qui permettront de segmenter au mieux les données. Ensuite, le modèle est appliqué sur un ensemble de test. Le défi ré-

side dans la recherche du modèle optimal qui minimise les critères de segmentation pour permettre une bonne généralisation du modèle. *[VERMET, 2023]*

La méthode se décrit de la manière suivante : soit (X, Y) un vecteur aléatoire prenant des valeurs dans $\mathbb{R}^p \times \mathcal{G}$, où $\mathcal{G} = \{1, \ldots, K\}$ dans le cas de la classification. Considérons un ensemble d'apprentissage $L_n = \{(x_i, y_i) \in \mathbb{R}^p \times \mathcal{G}, i = 1, \ldots, n\}$, où les (x_i, y_i) sont des réalisations indépendantes de variables aléatoires de même loi que (X, Y). À chaque étape du partitionnement, l'espace des entrées est divisé en deux sous-ensembles, et un arbre binaire est naturellement associé à cette partition.

La construction de l'arbre de décision se déroule en deux étapes : d'abord la détermination de l'arbre maximal ou saturé sur l'ensemble d'apprentissage, puis l'élagage pour obtenir l'arbre optimal.

1. **Détermination de l'arbre saturé** :
   Pour obtenir des arbres de décision binaires, l'ensemble de données est subdivisé en deux classes. La première division est obtenue en sélectionnant la variable explicative qui permet la meilleure séparation des individus. Cette subdivision est de la forme : $\{X^j \leq s\} \cup \{X^j > s\}$, où $j \in \{1, \ldots, p\}$, $X = (X^1, \ldots, X^p)$ et $s \in \mathbb{R}$. Cela signifie que toutes les observations dont la valeur de la $j^{\text{ème}}$ variable est inférieure à s sont affectées au sous-arbre de gauche, et les autres au sous-arbre de droite.

$$n_{1,-}(j, s) = \{i \in \{1, \ldots, n\} : x_i^j \leq s\}$$

   et :

$$n_{1,+}(j, s) = \{i \in \{1, \ldots, n\} : x_i^j > s\}$$

   Pour que cette division soit pertinente, la méthode choisit la meilleure coupure (j, s) possible, en minimisant une *fonction de coût* $C(j, s)$. Soit $x_i = (x_i^1, \ldots, x_i^p)$, la fonction de coût à minimiser dans le cas de la classification est :

$$C(s, j) = \sum_{k=1}^{K} \hat{p}^k_{n_{1,-}(j,s)} \left(1 - \hat{p}^k_{n_{1,-}(j,s)}\right) + \sum_{k=1}^{K} \hat{p}^k_{n_{1,+}(j,s)} \left(1 - \hat{p}^k_{n_{1,+}(j,s)}\right),$$

   où :

   $\hat{p}^k_{n_{1,\pm}(j,s)}$ représente la proportion d'observations de la classe k dans l'ensemble $n_{1,\pm}(j, s)$.

   Cette division crée des sous-populations correspondant aux premiers nœuds de l'arbre, et se fait en fonction de l'hétérogénéité ou de l'impureté. Cette impureté peut être mesurée à l'aide de l'indice de Gini, de l'entropie, ou de l'indice d'erreur,

etc. L'opération est ensuite répétée sur les deux nouvelles classes obtenues. L'algorithme s'arrête lorsque le nœud est homogène (aucune subdivision n'est possible) ou lorsque le nombre d'observations dans le nœud est inférieur à un seuil fixé.

2. **Élagage et construction de l'arbre optimal** : Un des inconvénients des arbres de décision est le surapprentissage. En effet, ils peuvent bien performer sur l'ensemble d'apprentissage mais avoir une fragilité de prédiction sur l'ensemble de test. L'objectif de l'élagage est donc de réduire la complexité de l'arbre en obtenant un modèle moins complexe et plus général sur l'ensemble de test. La technique de l'élagage consiste à éliminer les nœuds qui n'apportent pas de contribution significative à l'arbre de décision.

4.3.4 Forêts aléatoires

Les forêts aléatoires représentent une amélioration des techniques de *bagging* par l'ajout de la randomisation. Le terme *bagging* est une fusion des mots *bootstrap* et *aggregating*. En partant d'un échantillon d'apprentissage L_n et d'une méthode de base, le *bagging* consiste à tirer indépendamment B échantillons *bootstrap* $L_{n,1}, \ldots, L_{n,B}$ dans L_n, puis à calibrer la méthode de base sur chacun d'eux pour obtenir B prédicteurs $\hat{\Phi}_1(\cdot), \ldots, \hat{\Phi}_B(\cdot)$. *[VERMET, 2023]*

Les forêts aléatoires, proposées par Breiman en 2001, visent à réduire la corrélation entre les modèles de base en introduisant une part d'aléatoire dans le choix des variables. Cette technique permet de diminuer la variance du prédicteur obtenu par agrégation. Pour ce faire, Breiman suggère de sélectionner aléatoirement, à chaque étape de la construction de l'arbre, m variables parmi les p variables explicatives et de déterminer la coupure en utilisant uniquement ces variables. Une valeur couramment utilisée pour m dans les problèmes de classification est $m = \sqrt{p}$.

Considérons l'ensemble d'apprentissage $L_n = \{(x_i, y_i) \in \mathbb{R}^p \times \mathcal{G}, i = 1, \ldots, n\}$, où $\mathcal{G} = \{1, \ldots, K\}$ dans le cas de la classification. L'algorithme se déroule comme suit :

Algorithme (Forêts aléatoires) :

1. Choisir $B, m \in \mathbb{N}^*$.

2. Pour $b = 1$ à B :
— Tirer un échantillon Bootstrap $L_{n,b}$ à partir de L_n ;
— Avec cet échantillon $L_{n,b}$, construire un arbre CART qui définit un prédicteur $\hat{\Phi}_b(\cdot)$. Chaque coupure de l'arbre est déterminée en se limitant à un ensemble de m variables sélectionnées aléatoirement parmi les p variables explicatives, ces tirages étant indépendants pour chaque coupure.

3. Définir le prédicteur final $\hat{\Phi}(\cdot)$ comme la moyenne des B prédicteurs individuels : $\hat{\Phi}(\cdot) = \frac{1}{B} \sum_{b=1}^{B} \hat{\Phi}_b(\cdot)$ (ou $\hat{\Phi}(\cdot) = \arg\max_j \text{card}\{b : \hat{\Phi}_b(\cdot) = j\}$).

Comparaison des modèles d'arbre de décision et forêt aléatoire

Les modèles d'arbre de décision et de forêt aléatoire montrent des performances significativement améliorées. Les deux modèles montrent une amélioration de l'accuracy par rapport au modèle logistique initial, avec la forêt aléatoire atteignant une précision globale de 99,89 %. La *balanced accuracy* de l'arbre de décision reste toutefois plus faible. Celle de la forêt aléatoire en revanche est très élevée : 99,40 % indiquant que ce modèle identifie très bien la déclaration d'arrêté de catastrophes naturelles. Le F1-score, *recall* et précision confirment la capacité des modèles à maintenir un équilibre entre la précision et le *recall*, avec des scores élevés pour la forêt aléatoire, indiquant une capacité supérieure à détecter à la fois les sinistres et les non-sinistres.

	Arbre de décision	Forêt aléatoire
Accuracy	94,19 %	99,89 %
Balanced accuracy	63,97 %	99,40 %
F1-score	38,96 %	99,44 %
Recall	26,39 %	99,29 %
Précision	74,42 %	99,58 %
AUC	71,14 %	99,88 %

TABLE 4.15 – Comparaison des métriques obtenues

Les matrices de confusion pour l'arbre de décision et la forêt aléatoire détaillent la répartition des prédictions par rapport aux valeurs réelles

	Prédit Négatif	Prédit Positif
Réel Négatif	83 185	577
Réel Positif	4 697	1 680

TABLE 4.16 – Arbre de décision

	Prédit Négatif	Prédit Positif
Réel Négatif	83 735	27
Réel Positif	45	6 332

TABLE 4.17 – Forêt aléatoire

Les modèles d'arbre de décision et de forêt aléatoire surpassent le modèle de régression logistique initial en termes de performances prédictives, en particulier dans la détection des sinistres, avec la forêt aléatoire ayant les meilleures métriques. Le graphique suivant montre l'importance des variables du modèle de forêt aléatoire définit selon le critère de Gini.

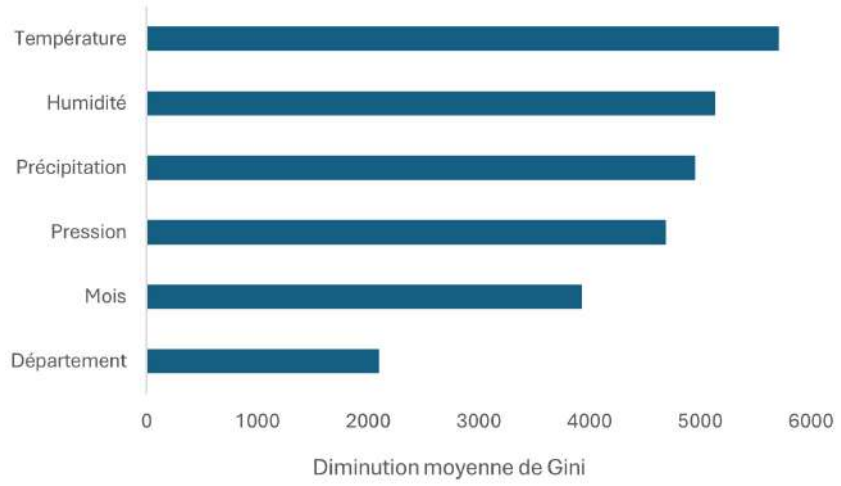


FIGURE 4.15 – Importance des variables de la forêt aléatoire

Selon le critère de diminution de Gini, la température est la variable la plus importante du modèle suivi par l'humidité et la précipitation.

Le modèle de **forêt aléatoire** sera donc celui utilisé pour obtenir une prédiction du nombre d'arrêtés de catastrophe naturelles sur les cinq années à venir.

La même méthodologie a été utilisée pour évaluer le nombre d'arrêtés de catastrophes naturelles lié au risque **inondation**, et les résultats correspondants sont disponibles en annexe (**voir annexe**).

4.4 Dérive de sinistralité de la fréquence

Les données journalières météorologiques futures pour chaque département obtenues après détermination de l'évolution de ces paramètres (température, humidité, précipitation, pression) au sein de chaque zone via projection à l'aide de séries temporelles sont utilisées comme entrées du modèle de forêt aléatoire et permettent de déterminer une estimation du nombre d'arrêtés de catastrophes naturelles pour les années à venir.

Le modèle de forêt aléatoire pour le risque de sécheresse a été ajusté sur les données de 2016 à 2022 en raison du délai dans la parution des arrêtés relatifs à ce risque. Le nombre d'arrêtés de catastrophes naturelles pour les années 2023 à 2028 pour le risque de sécheresse est présenté dans le graphique ci-dessous.

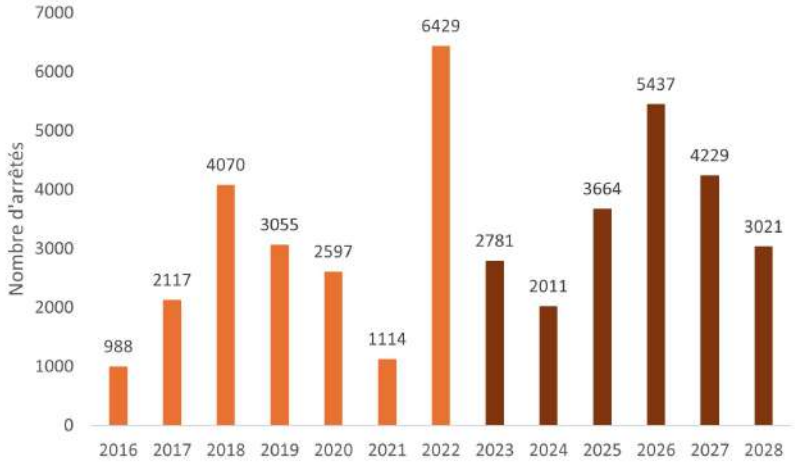


FIGURE 4.16 – Projection du nombre d'arrêtés futurs pour le risque de sécheresse

Une comparaison de la moyenne du nombre d'arrêtés entre 2017 et 2022 et celle obtenue entre 2023 et 2028 révèle une augmentation de la sinistralité de **9 %** pour le risque de sécheresse.

Le modèle de forêt aléatoire pour le risque d'inondation a, quant à lui, été ajusté sur l'ensemble des données historiques, soit de 2016 à 2023. Le nombre d'arrêtés projeté pour les années 2024 à 2028 est présenté dans le graphique ci-dessous.

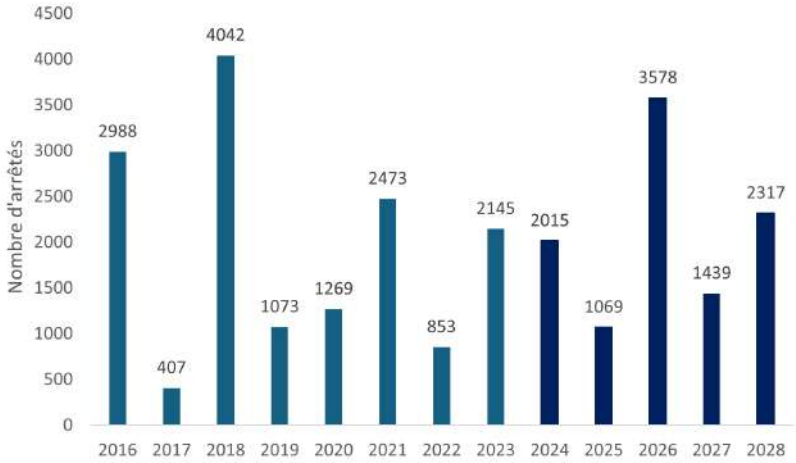


FIGURE 4.17 – Projection du nombre d'arrêtés futurs pour le risque d'inondation

Une comparaison de la moyenne du nombre d'arrêtés entre 2018 et 2023 et celle entre 2024 et 2028 montre une augmentation de la sinistralité de **23 %**.

4.5 Critiques des modèles

Bien que les différents modèles ajustés aient été validés, il est essentiel de souligner leurs limites.

1. **Modélisation des Paramètres Météorologiques** :
   La prévision des paramètres météorologiques a été réalisée en modélisant une tendance linéaire, une saisonnalité trigonométrique avec un polynôme d'ordre 2 et les résidus avec un modèle ARMA. Cependant, cette approche présente certaines limites : Supposer une **tendance linéaire** pour la température par exemple, peut ne pas être adaptée, car les tendances climatiques peuvent suivre des schémas non linéaires en raison des influences anthropiques et naturelles. Aussi, utiliser un polynôme trigonométrique pour modéliser la **saisonnalité** peut ne pas capturer toutes les nuances des variations saisonnières, surtout si ces variations changent de manière significative à cause du changement climatique. Enfin, la modélisation des **précipitations** est particulièrement complexe, et la méthodologie employée peut ne pas refléter fidèlement les fluctuations et extrêmes observés dans les données historiques et futures.

2. **Modèle de Classification** :
   Bien que les forêts aléatoires soient robustes, elles peuvent parfois surajuster les données historiques, ce qui les empêche de bien se généraliser aux nouvelles conditions climatiques futures. Aussi, un changement dans les caractéristiques des régimes de catastrophes naturelles pourrait affecter les performances du modèle, rendant les prédictions moins fiables.

Chapitre 5

Evaluation de la dérive de sinistralité du risque tempêtes

Pour déterminer l'occurrence des tempêtes à venir, la simulation des vitesses des rafales est nécessaire, étant donné que les assureurs n'indemnisent les tempêtes que lorsque ces vitesses dépassent les **100 km/h**. L'analyse antérieure de la base de données SYNOP a révélé que la variable de *vitesse des rafales* présente des valeurs manquantes à hauteur de 18,6 %, avec des départements n'ayant pas d'observations pendant des mois entiers. Afin d'éviter tout biais potentiel résultant du remplacement de ces valeurs, la modélisation de la dérive de sinistralité du risque tempête sera faite en utilisant la variable de *vitesse du vent*, qui présente une corrélation très forte et positive avec la vitesse des rafales. Seulement 0,7 % des valeurs de la *vitesse du vent* étaient manquantes, et elles ont été remplacées par la **moyenne**. La transformation de la vitesse du vent en vitesse des rafales est réalisée par le biais d'une **régression linéaire** appliquée aux données non manquantes de la base SYNOP.

Avant de procéder à la modélisation, il a été décidé de regrouper la France Métropolitaine en zones de risques homogènes afin de réduire le coût de la modélisation. Tout comme pour les risques inondation et sécheresse, ce regroupement a été réalisé en utilisant l'Analyse en Composantes Principales (ACP) et la Classification Ascendante Hiérarchique (CAH), cette fois-ci sur les données de vitesse du vent observées au sein de chaque département de la France métropolitaine.

5.1 Construction du zonier

Les différents départements sont tout d'abord représentés sur le plan factoriel à travers l'ACP.

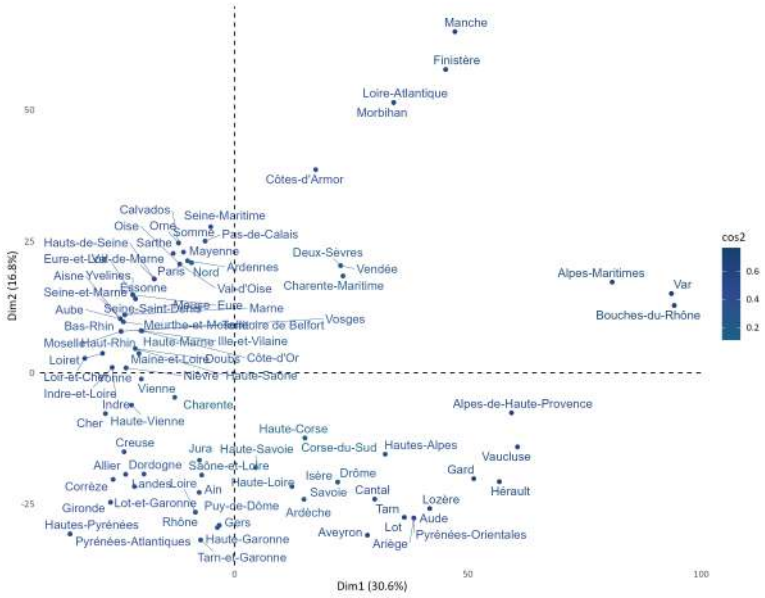


FIGURE 5.1 – Graphique des individus de l'ACP

Il apparaît que la première et la deuxième dimension expliquent **46,4 %** de l'inertie totale. Le premier axe est associé à l'intensité de la vitesse du vent, les départements situées au sud de la France étant les plus corrélées à cet axe. Le second axe représente la localisation géographique, avec les départements du nord de la France se trouvant dans la partie supérieure du plan factoriel et celles du sud dans la partie inférieure.

La règle du coude est par la suite utilisée pour déterminer le nombre de classe optimale pour le zonage.

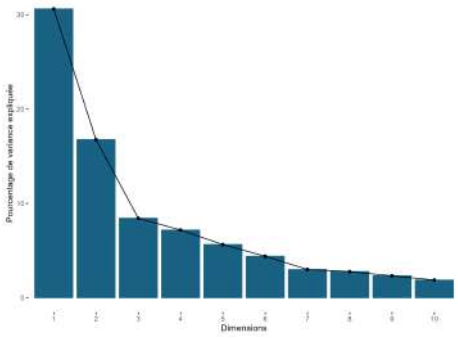


FIGURE 5.2 – Graphique des éboulis

Le nombre optimal de dimension selon cette règle est de trois, mais le choix de conserver les **cinq** premières dimensions a été fait car un déséquilibre significatif entre les classes

était observé avec un partitionnement en trois classes.

La Classification Ascendante Hiérarchique est par la suite effectuée sur les résultats de l'ACP, ce qui permet d'obtenir les zones représentées sur la figure suivante.

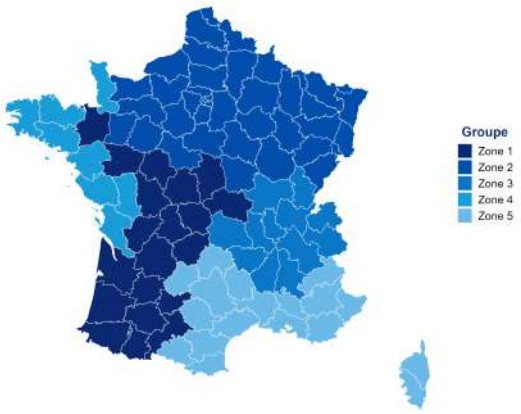


FIGURE 5.3 – Zonier du risque tempête

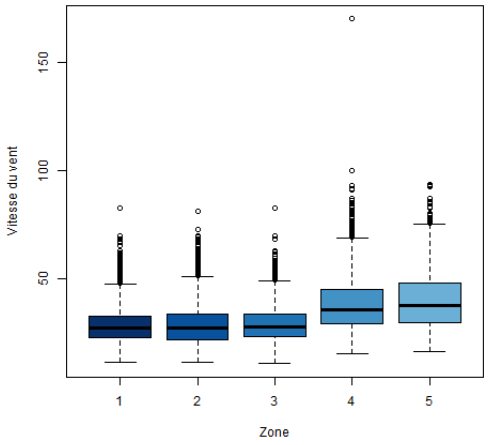


FIGURE 5.4 – Boîte à moustache de la vitesse du vent par zone

Les zones 4 et 5, correspondant au Nord-Ouest et au Sud-Ouest de la France, connaissent les vitesses de vent les plus élevées en moyenne, avec une plus grande dispersion des données, indiquant des variations plus importantes. Les zones 1, 2 et 3 montrent des vitesses de vent moyennes similaires, mais la distribution des vitesses de vent diffère dans le troisième quartile. La présence de nombreuses valeurs aberrantes dans toutes les zones suggère des occurrences sporadiques de vitesses de vent extrêmement élevées.

5.2 Modélisation de la vitesse du vent par zone.

La modélisation des risques extrêmes nécessite des outils pour estimer la loi des observations au-delà d'un seuil donné et pour calculer les quantiles extrêmes à l'aide de cette loi estimée. C'est pourquoi la théorie des valeurs extrêmes sera utilisée pour modéliser la vitesse du vent dans chaque zone, car elle permet d'obtenir la queue de distribution, qui est l'élément d'intérêt.

5.2.1 Théorie des valeurs extrêmes

La statistique classique se concentre principalement sur la partie centrale de la loi modélisant au mieux le phénomène considéré, en utilisant des outils tels que le calcul de l'espérance, la médiane, la variance, et le théorème central limite. Dans ce contexte, l'étude se porte sur les queues de distribution de la loi c'est-à-dire des grandes déviations par rapport à la médiane des distributions de probabilité. La théorie des valeurs extrêmes fournit un cadre théorique solide pour analyser ces valeurs dites extrêmes.

Deux approches sont utilisées pour appliquer la théorie des valeurs extrêmes. La première consiste à générer des séries de maxima (ou minima) par blocs, tandis que la seconde implique l'extraction des valeurs de pointe dépassant (ou étant inférieures à) un certain seuil à partir d'un enregistrement continu. Dans le cadre de cette modélisation, la méthode utilisée est celle des **maxima par blocs**. Bien que la méthode de dépassement de seuil permette d'obtenir la distribution des tempêtes qui seront indemnisées l'assureur, elle ne permettra pas de quantifier la tendance des vitesses de vent et potentiellement déterminer si elles sont plus fréquentes avec le temps. Ainsi, l'utilisation des maxima par blocs apparaît comme une approche plus adéquate pour étudier cette tendance. *[RIEDER, 2014]*

La modélisation des queues de distribution en utilisant la méthode des maxima par blocs repose sur le théorème de **Fisher-Tippet**. Il est supposé que l'échantillon des maxima suit précisément une distribution Generalized Extreme Value (GEV), ayant la forme suivante :

$$\mathrm{GEV}(x) = \begin{cases} \exp\left(-\left[1+\xi\frac{x-\mu}{\sigma}\right]^{-\frac{1}{\xi}}\right) & \text{si } \xi \neq 0 \\ \exp\left(-\exp\left(-\frac{x-\mu}{\sigma}\right)\right) & \text{si } \xi = 0 \end{cases} \tag{5.1}$$

où $\left\{x \in \mathbb{R}, 1+\xi\left(\frac{x-\mu}{\sigma}\right) > 0\right\}$ et $\mu \in \mathbb{R}, \sigma > 0, \xi \in \mathbb{R}$.

Pour chaque fonction de GEV, il existe trois paramètres :
μ, **paramètre de position** s'assimilant à la moyenne pour une loi normale
σ, **paramètre d'échelle** s'assimilant à l'écart-type
ϵ, **paramètre de forme**, qui forme la distribution *[Numérique, 2013]*

Il existe **trois** types de distributions GEV selon la valeur de ξ. Pour $\xi = 0$, il s'agit de la distribution de **Gumbel** ; pour $\xi > 0$, de la distribution de **Fréchet** ; et pour $\xi < 0$, de la distribution de **Weibull**.

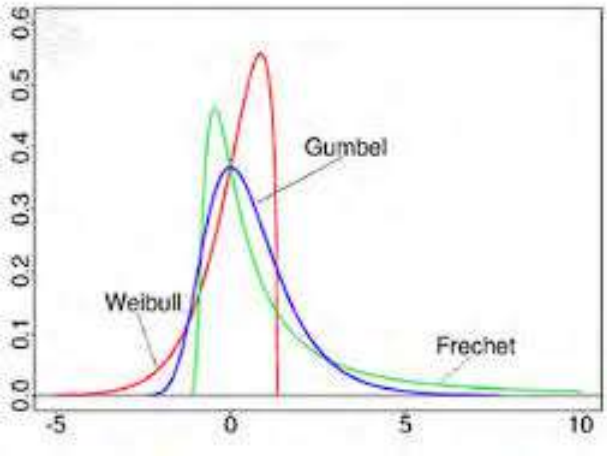


FIGURE 5.5 – Distribution GEV (Source : *Google*)

Ces trois lois correspondent aux trois types de comportement possibles de la queue de distribution :

— la distribution de Gumbel correspond a une décroissance exponentielle de la queue de la distribution
— la distribution de Fréchet correspond a une décroissance polynômiale
— la distribution de Weibull correspond a une densité supérieure bornée

La méthode des maxima par blocs consiste à une séparation des données en blocs dont les maxima sont supposés suivre une distribution GEV. Il est donc nécessaire de déterminer au préalable la taille des blocs.

Pour cette étude des blocs de taille **mensuels** ont été choisis. Il sera donc question de modéliser les vitesses de vent maximales atteintes chaque mois. Afin de faciliter la lisibilité des résultats, seuls les graphiques de la zone 4 représentant le Nord-Ouest de la France seront discutés ici, tandis que ceux des autres zones sont disponibles en annexe.

Dans un premier temps, les histogrammes des maxima de chaque zone ainsi que leur ECDF [1] ont été représentés dans le but de visualiser la distribution des données et d'identifier les seuils des risques.

1. Empirical Cummulative Distribution Function

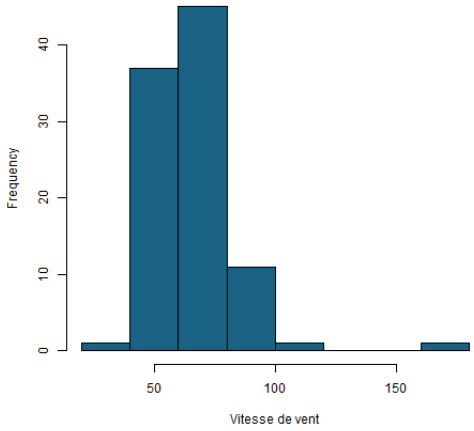


FIGURE 5.6 – Maxima mensuels pour la zone 4

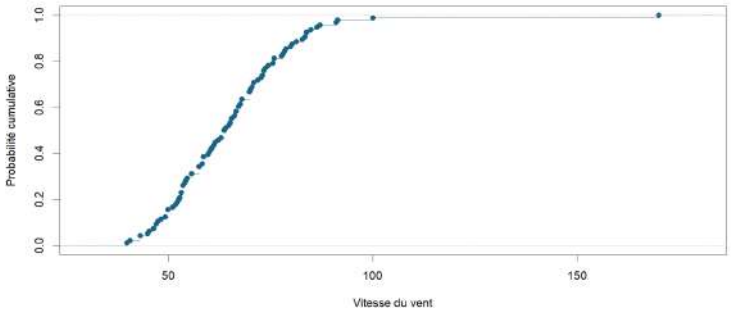


FIGURE 5.7 – ECDF zone 4

L'analyse visuelle de l'histogramme révèle une asymétrie positive légèrement marquée dans la distribution des maxima mensuels, avec une queue supérieure lourde, comme confirmée par l'ECDF.

Estimation des paramètres de la distribution GEV

La maximisation de la vraisemblance sous l'hypothèse d'une loi GEV avec des paramètres $\theta = (\mu, \sigma, \xi)$ permet d'obtenir l'estimateur du maximum de vraisemblance pour les distributions appartenant à la famille des distributions GEV. Toutefois, cette maximisation nécessite des méthodes numériques car il n'existe pas de solution analytique. Des problèmes peuvent apparaître si les conditions de régularité [2] typiquement requises pour les estimateurs du maximum de vraisemblance ne sont pas respectées. Smith (1985) a présenté les résultats suivants :

— si $\xi > -1/2$, l'estimateur de vraisemblance est régulier
— si $-1 < \xi < -1/2$, l'estimateur est super-efficace
— si $\xi < -1$, l'estimateur du point extrêmal est donné par la plus grande des observations

La distribution GEV a été ajustée à l'aide du package *fevd* de R. Une analyse du paramètre de forme ξ sur chacune des zones montre que l'estimateur de vraisemblance est régulier sur chacune des zones ($\xi > -1/2$) et donc un intervalle de confiance peut être définit comme le montre le tableau suivant :

2. Les conditions de régularité incluent : (1) L'information de Fisher I_n pour θ doit exister et être inversible. (2) La condition $E_\theta[\nabla \ln L_n(.;\theta)] = 0$ doit être satisfaite, et $I_n(\theta) = -E_\theta[\nabla^2 \ln L_n(.;\theta)]$.

	Zone 1	Zone 2	Zone 3	Zone 4	Zone 5
Paramètre de position	44,47	44,95	43,40	58,21	65,17
Intervalle de confiance	[42,65 ; 46,29]	[43,09 ; 46,82]	[41,82 ; 44,99]	[55,60 ; 60,82]	[62,96 ; 67,38]
Paramètre d'échelle	7,89	8,36	7,09	11,80	10,03
Intervalle de confiance	[6,59 ; 9,20]	[7,04 ; 9,67]	[5,95 ; 8,22]	[9,90 ; 13,70]	[8,48 ; 11,58]
Paramètre de forme	-0,05	-0,09	-0,03	0,03	-0,25
Intervalle de confiance	[-0,21 ; 0,10]	[-0,22 ; 0,03]	[-0,16 ; 0,10]	[-0,07 ; 0,15]	[-0,37 ; -0,12]

TABLE 5.1 – Paramètres de la distribution GEV ajustée

La valeur du paramètre de forme ξ permet également de déterminer à quelle famille de distribution GEV les maxima appartiennent. Pour les zones 1 à 4, l'analyse de ce paramètre révèle que **0** est inclus dans l'intervalle de confiance, suggérant que la distribution de **Gumbel** est la plus appropriée pour modéliser les maxima mensuels dans ces régions. En revanche, pour la zone 5 qui représente le Sud de la France, le paramètre de forme ξ est négatif et 0 n'est pas inclus dans l'intervalle de confiance. Cela indique que la distribution de Weibull pourrait être plus adaptée pour modéliser les maxima mensuels dans cette zone spécifique. Cependant, la distribution de Weibull est souvent utilisée pour modéliser des phénomènes avec des densités supérieures bornées, ce qui n'est pas typiquement le cas pour la vitesse du vent.

Dans l'étude des valeurs extrêmes, une méthode efficace pour valider les modèles est l'analyse graphique. Pour illustrer cette validation, les résultats obtenus après la modélisation pour la zone 4 sont présentés dans le graphique 5.8 :

1. **Probability plot**
   Ce graphique compare la fonction de répartition empirique en abcisse avec celle de la loi GEV ajustée en ordonnée. La fonction de répartition de la la distribution GEV est donnée par l'équation 5.1. Un bon ajustement est indiqué par les points étant proche de la diagonale. Le premier graphique montre que les points relativement alignés sur la diagonale, vérifiant ainsi cette hypothèse par le modèle.

2. **Quantile plot**
   Ce graphique compare la fonction quantile de la loi GEV ajustée en abcisse et la fonction quantile empirique en ordonnée. La fonction quantile est l'inverse de la fonction de répartition et est donnée par :

$$\mathrm{Q}(p) = \begin{cases} \mu - \sigma\xi^{-1}\left[1 - (-\ln(1-p))^{-\xi}\right] & \text{si } \xi \neq 0 \\ \mu - \sigma\ln(-\ln(p)) & \text{si } \xi = 0 \end{cases} \tag{5.2}$$

   où p est la probabilité.
   Tout comme pour le probability plot, un bon ajustement est indiqué par des points alignés sur la diagonale. Le second graphique montre que cette hypothèse est également vérifiée.

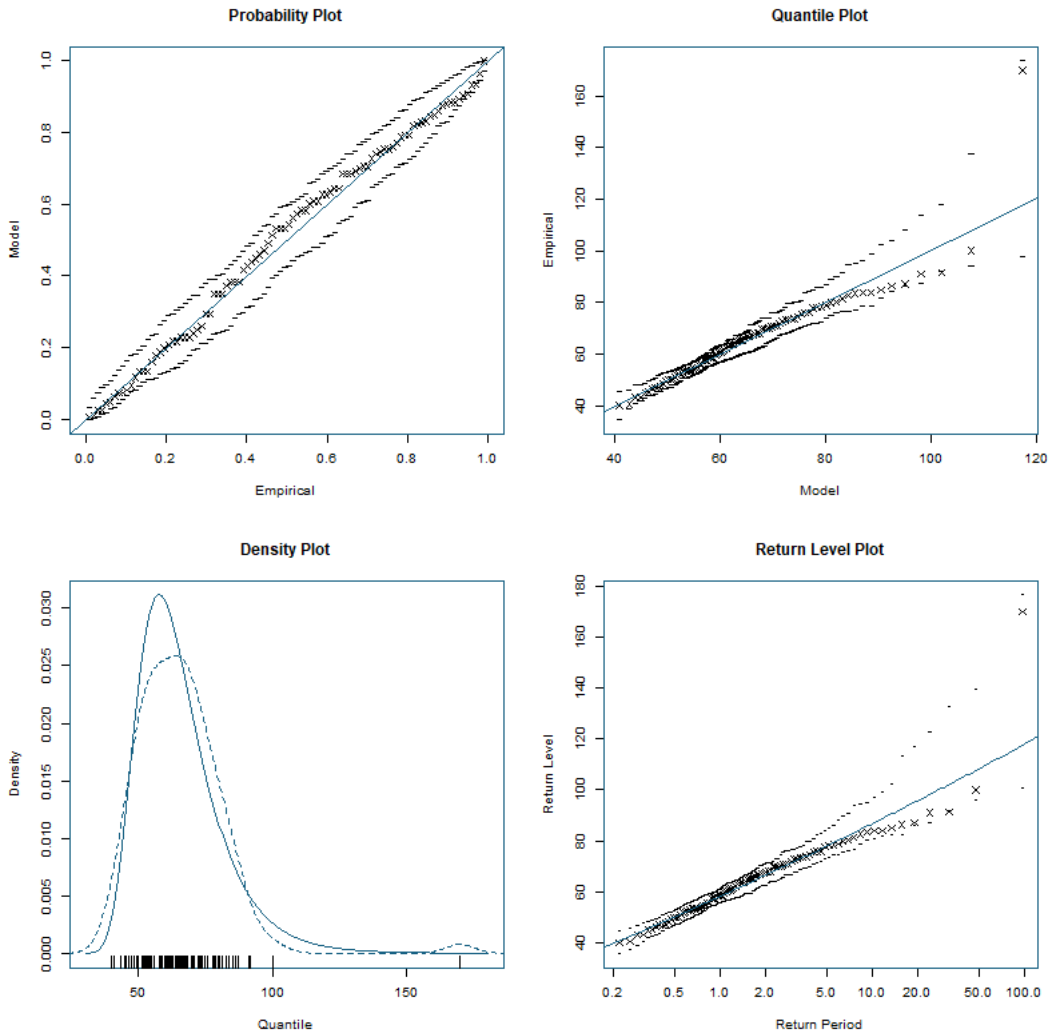


FIGURE 5.8 – Validation du modèle pour la zone 4

3. **Density plot**
   Ce graphique montre la densité empirique des données superposée à la densité théorique de la loi de Gumbel ajustée. Il permet de visualiser la correspondance entre la densité observée et celle prévue par le modèle. Le troisième graphique montre que la distribution GEV s'ajuste plutôt bien aux données.

4. **Return level plot**
   Ce graphique indique les périodes de retour en abcisse et les niveaux de retour associés en ordonnée. La vitesse de vent associée à une période de retour de 100 ans pour cette zone est de **117,7 km/h**.

Cette validation par analyse graphique est aussi vérifiée pour les autres zones (**voir annexe**).

La loi de Gumbel a par la suite été ajustée aux données. L'ANOVA [3] a été utilisée pour effectuer un test de rapport de vraisemblance entre la distribution GEV ajustée et la distribution de Gumbel avec $\xi = 0$.

— **Hypothèse nulle (H_0)** : Le modèle de Gumbel est suffisant pour décrire les données (c'est-à-dire $\xi = 0$).

— **Hypothèse alternative (H_a)** : Un modèle GEV est nécessaire pour décrire les données (c'est-à-dire $\xi \neq 0$).

Les résultats de l'ajustement de la loi de Gumbel et de l'ANOVA sont présentés dans le tableau suivant :

	Zone 1	Zone 2	Zone 3	Zone 4	Zone 5
Paramètre de position	44,25	44,52	43,27	58,48	63,84
Ecart-type	0,83	0,88	0,75	1,28	1,04
Paramètre d'échelle	7,76	8,17	7,02	11,97	9,65
Ecart-type	0,61	0,63	0,55	0,95	0,73
Quantile associé à la période de retour	76,81	76,02	73,61	117,74	92,45
Résultat ANOVA	0,53	0,19	0,63	0,47	0,001234

TABLE 5.2 – Paramètres de la loi Gumbel ajustée et résultats de l'ANOVA

Ces résultats indiquent que pour les zones 1, 2, 3 et 4, avec un niveau de confiance de 95 %, l'hypothèse selon laquelle la distribution de Gumbel est adéquate pour modéliser les maxima mensuels de vitesse du vent n'est pas rejetée. Cela suggère que la distribution de Gumbel est suffisante pour ces zones. Cependant, pour la zone 5, l'hypothèse nulle est rejetée, ce qui est attribué à la valeur négative du paramètre de forme ξ dans cette

3. Analysis of Variance

zone. Cette observation indique que la distribution de Gumbel n'est pas appropriée pour modéliser les maxima mensuels de vitesse du vent dans cette région comme l'avait révélé l'analyse initiale. Pour cette zone, un modèle GEV est donc nécessaire.

Pour la suite des travaux, les paramètres de la loi Gumbel seront utilisés pour les zones 1 à 4, et les paramètres de la distribution GEV ajustée pour la zone 5.
Ces paramètres ont été utilisés pour simuler les vitesses de vent maximales atteintes dans chacune des zones sur l'historique d'observation c'est à dire 8 ans (2016 à 2023) et la corrélation entre les différentes zones a été déterminée.

	Zone 1	Zone 2	Zone 3	Zone 4	Zone 5
Zone 1	1	0,05	-0,17	-0,04	0,01
Zone 2	0,05	1	-0,21	-0,19	-0,15
Zone 3	-0,17	-0,21	1	-0,08	0,02
Zone 4	-0,04	-0,19	-0,08	1	-0,03
Zone 5	0,01	-0,15	0,02	-0,03	1

TABLE 5.3 – Tableau des corrélations entre les zones sur les données simulées

On se rend compte que la corrélation entre les différentes zones est généralement faible et souvent négative, ce qui est contraire aux caractéristiques du vent. En effet le vent étant un phénomène en mouvement constant, des vents forts dans une zone devraient influencer les zones avoisinantes. Cette relation entre les zones est bien reflétée dans les données réelles, dont les corrélations sont présentées dans le tableau suivant :

	Zone 1	Zone 2	Zone 3	Zone 4	Zone 5
Zone 1	1	0,61	0,55	0,61	0,43
Zone 2	0,61	1	0,49	0,72	0,35
Zone 3	0,55	0,49	1	0,37	0,48
Zone 4	0,61	0,72	0,37	1	0,28
Zone 5	0,43	0,35	0,48	0,28	1

TABLE 5.4 – Tableau des corrélations entre les zones sur les données réelles

Afin de modéliser la dépendance entre les différentes zones, la théorie des **copules** sera utilisée.

5.2.2 Théorie des Copules

Une mesure de dépendance couramment utilisée en statistique est le coefficient de corrélation de Pearson, qui évalue la relation linéaire entre deux variables aléatoires quantitatives. Cependant, lorsqu'il s'agit de variables aléatoires continues qui ne suivent pas

une distribution normale, les concepts basés sur la linéarité peuvent ne pas être adaptés. *[MAO, 2022]*

Une copule est un outil statistique qui modélise la dépendance entre des variables aléatoires, permettant de distinguer la structure de dépendance décrite par la fonction de répartition conjointe et le comportement marginal des variables considérées. En utilisant les copules, il est possible de résumer la structure de dépendance d'une distribution conjointe en la séparant des comportements marginaux.

La distribution uniforme est intimement liée aux copules. La fonction de répartition univariée d'une variable aléatoire de loi uniforme sur l'intervalle $I = [0, 1]$. Soit $U \sim \mathcal{U}(0, 1)$. Alors, on a :

$$P(U \leq u) = \begin{cases} 0 & \text{si } u \leq 0 \\ u & \text{si } 0 \leq u \leq 1 \\ 1 & \text{si } u \geq 1 \end{cases}$$

Une copule bivariée est une fonction $C : I^2 \to I$ qui est définie par les conditions suivantes :

1. C est attachée, c'est-à-dire

$$C(u, 0) = 0 = C(0, v), \quad \text{pour tout } u, v \in I;$$

2. Les marges sont uniformes, c'est-à-dire

$$C(u, 1) = u \quad \text{et} \quad C(1, v) = v, \quad \text{pour tout } u, v \in I;$$

3. C est une fonction 2-croissante sur I^2, c'est-à-dire pour tout $u_1, u_2, v_1, v_2 \in I$ tels que $u_1 \leq u_2$ et $v_1 \leq v_2$, on a

$$C(u_2, v_2) - C(u_2, v_1) - C(u_1, v_2) + C(u_1, v_1) \geq 0.$$

Il est possible de construire une copule en s'assurant que les trois conditions de la définition précédente soient respectées. Le théorème de Sklar montre comment relier le principe de la copule à la fonction de répartition bivariée et aux lois marginales.

Théorème de Sklar :
Soit F une fonction de répartition conjointe de marges F_1 et F_2. Alors, il existe une copule $C : I^2 \to I$ telle que pour tout $(x, y) \in \mathbb{R}^2$, on a :

$$F(x, y) = C(F_1(x), F_2(y))$$

De plus, si F_1 et F_2 sont continues, alors C est unique.

Il existe une variété de copules adaptées à différentes situations, chacune exprimant une structure de dépendance distincte : *[Marie-Christine BRASSIER, 2010]*

— dépendance dans les petites valeurs
— dépendance dans les valeurs extrêmes
— dépendance de queue
— dépendance positive ou négative

Les copules paramétriques se divisent en deux grandes familles : les copules elliptiques et les copules archimédiennes.

Famille des copules elliptiques

Appliquées aux distributions symétriques on y trouve :

— **Copule Gaussienne** : Elle ne montre pas de dépendance de queue, ce qui la rend inadaptée aux valeurs extrêmes. Sa pertinence réside dans son lien avec la distribution normale multivariée. La structure de dépendance qu'elle modélise est compatible avec le coefficient de corrélation linéaire. *[PLANCHET, 2024]*

— **Copule de student** : Associée à une distribution multivariée de Student, cette copule capture les dépendances extrêmes, tant positives que négatives.

Famille des copules archimédiennes

Ces copules sont avantageuses pour modéliser diverses structures de dépendance, notamment les dépendances asymétriques, où les coefficients de queue inférieure et supérieure sont différents.

— **Copule de Gumbel** : Capte les dépendances positives, surtout marquées sur la queue supérieure.

— **Copule de Frank** : Capte à la fois les dépendances positives et négatives.

— **Copule de Clayton** : Met en évidence des dépendances positives, en particulier pour des événements à faible intensité.

— **Copule HRT** : Modélise la dépendance lors d'événements extrêmes de forte intensité, avec une structure de dépendance inversée par rapport à la copule de Clayton. *[Marie-Christine BRASSIER, 2010]*

Copule	Fonction génératrice $\varphi(u)$	$C(u,v)$
Gumbel	$(-\ln u)^a,\ a \geq 1$	$\exp\left(-\left[(-\ln u)^a + (-\ln v)^a\right]^{1/a}\right)$
Frank	$-\ln\left(\frac{e^{-au}-1}{e^{-a}-1}\right),\ a \neq 0$	$\frac{1}{a}\ln\left[1+\frac{(e^{-au}-1)(e^{-av}-1)}{(e^{-a}-1)}\right]$
Clayton	$\frac{u^{-a}-1}{a},\ a > 0$	$\left(u^{-a}+v^{-a}-1\right)^{-1/a}$

TABLE 5.5 – Formules des différentes copules

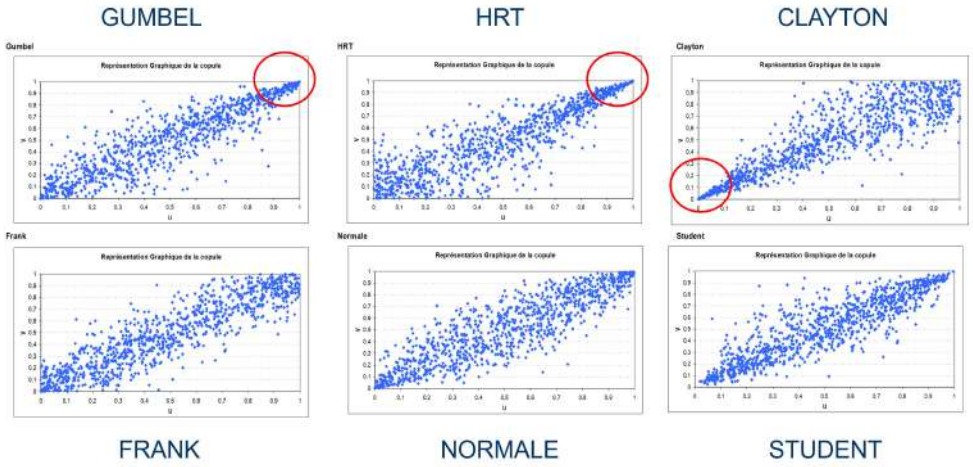


FIGURE 5.9 – Représentation graphique des copules en deux dimensions (Source : *Altia*)

Modélisation du paramètre de la copule

Le nuage de points des copules a permis de déterminer que la famille de **copule archimédienne** est plus adaptée pour modéliser la dépendance entre les différentes zones.

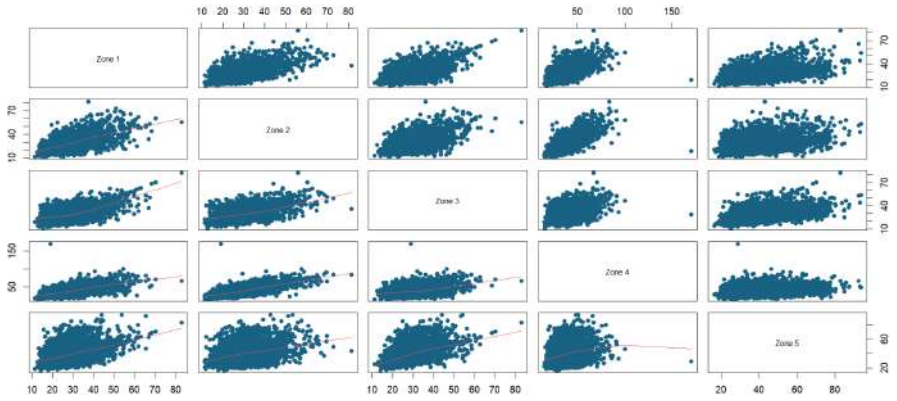


FIGURE 5.10 – Nuage de points de la vitesse du vent par zones

Les copules de Gumbel, Frank et Clayton ont été ajustées et le paramètre de chacune des copules a été estimé à l'aide de la méthode du maximum de vraisemblance. En particulier, la méthode du **Maximum de Vraisemblance Canonique (CML)** a été appliquée car elle permet une estimation simultanée des lois marginales et de la copule. Cette approche a été préférée car elle maximise la vraisemblance conjointe des données, ce qui la rend plus précise et robuste par rapport à d'autres méthodes telles que l'Inversion of

Margins (IFM), où les lois marginales sont ajustées séparément avant d'ajuster la copule.

Le choix de la copule a été fait en utilisant les critères de sélection de vraisemblance, Akaike's Information Criterion (AIC) et Bayesian Information Criterion (BIC) qui jouent un rôle important dans la sélection du modèle, en équilibrant la complexité du modèle et l'ajustement aux données observées.

— $\text{AIC} = 2k - 2\ell(\hat{\theta}; \mathbf{u})$
— $\text{BIC} = k\log(n) - 2\ell(\hat{\theta}; \mathbf{u})$
où k est le nombre de paramètres, n est la taille de l'échantillon et $\ell(\hat{\theta}; \mathbf{u})$ la log de vraisemblance maximisée

La copule de **Frank** a été choisie car elle maximise la vraisemblance des données, mais minimise également l'AIC et le BIC.

Copule	Paramètre	Vraisemblance	AIC	BIC
Gumbel	1,36	1853,21	-3704,42	-3698,44
Frank	2,68	1855,5	-3708,98	-3702,99
Clayton	0,52	1536,69	-3071,37	-3065,39

TABLE 5.6 – Choix de la copule

Cette copule a donc été utilisée pour simuler les vitesses de vent maximales atteintes dans chaque zone. Le nuage de points montre que la dépendance entre les différentes zones est bien capturée par la copule de Frank car les points des données observées en bleu se superposent plutôt bien aux points simulées en rouge.

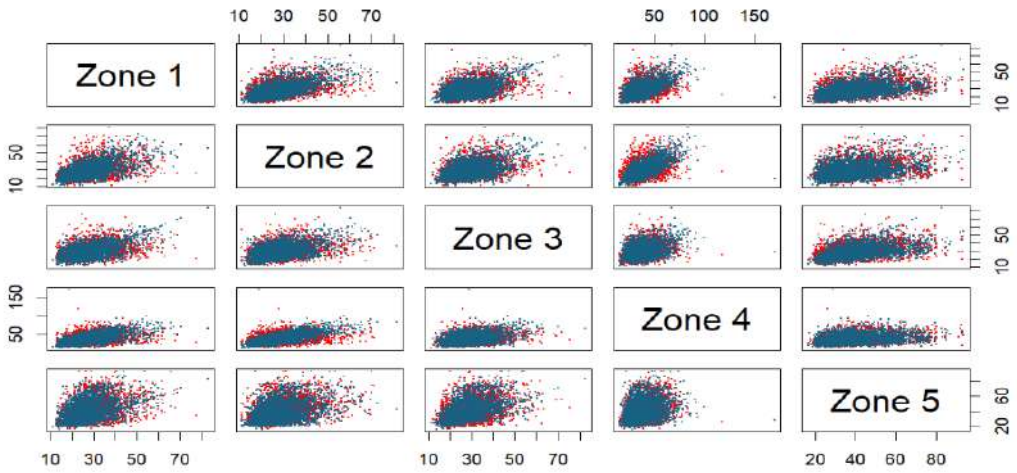


FIGURE 5.11 – Comparaison des données réels et de la copule de Frank

Pour obtenir les vitesses de vent maximales mensuelles atteintes dans chacune des zones, les paramètres de la loi de Gumbel pour les zones 1 à 4 et de la distribution GEV pour la zone 5 ont été utilisés pour calibrer les lois marginales. La copule de Frank précédemment calibrée a été utilisée pour générer des vecteurs de valeurs uniformes dépendantes représentant les relations de dépendance entre les zones. Ces valeurs uniformes ont ensuite été transformées en vitesses de vent en appliquant l'inverse des fonctions de répartition des lois marginales calibrées, afin que les simulations respectent les distributions marginales historiquement observées pour chaque zone. Le tableau suivant montre la vitesse de vent maximale simulée par zone et par mois pour l'année 2024.

	Zone 1	Zone 2	Zone 3	Zone 4	Zone 5
Janvier	41,39	38,17	38,75	76,88	72,98
Février	46,42	44,76	51,10	45,56	51,36
Mars	64,19	57,74	41,29	103,80	78,81
Avril	39,26	41,98	44,46	45,82	82,03
Mai	36,04	48,25	50,48	51,63	67,38
Juin	65,04	50,91	38,07	50,82	71,36
Juillet	43,32	53,09	46,98	51,11	63,26
Août	50,57	56,49	52,01	53,95	70,93
Septembre	45,97	30,76	45,87	57,05	83,90
Octobre	40,83	45,21	70,84	78,86	59,28
Novembre	34,54	42,22	44,29	48,85	68,48
Décembre	37,82	49,76	46,43	85,99	68,15

TABLE 5.7 – Vitesse de vent maximales simulées par zones pour 2024

Après avoir obtenu la vitesse du vent, il convient maintenant de déterminer la vitesse maximale des rafales atteinte chaque mois.

5.2.3 Modélisation de la relation entre la vitesse du vent et la vitesse des rafales

En introduction de ce chapitre, il a été rappelé que la variable d'intérêt pour les assureurs, à savoir la vitesse des rafales, contenait 18,6 % de valeurs manquantes, avec certains départements n'ayant pas d'observations pour des mois entiers. Pour éviter le biais potentiel causé par le remplacement de ces données, les différents modèles ont été ajustés en utilisant la vitesse du vent. Cette section vise à déterminer la relation entre la vitesse du vent et la vitesse des rafales à l'aide d'une **régression linéaire**, afin de convertir les simulations du tableau 5.10 en simulations de la vitesse des rafales pour les différentes zones.

Régression linéaire simple

La régression linéaire simple vise à expliquer une variable Y, qui dans ce contexte est la vitesse des rafales, en utilisant une variable X, qui représente la vitesse du vent.

On suppose que pour tout i :

$$Y_i = a + bx_i + \epsilon_i \quad \text{avec} \quad \{\epsilon_i\} \text{ i.i.d. et } \sim \mathcal{N}(0, \sigma^2).$$

Les hypothèses de ce modèle sont : *[MARTIN, 2021]*

— L'espérance de Y_i dépend linéairement de x_i : $\mathbb{E}(Y_i) = a + bx_i$.
— La variance des Y_i est constante : $\mathbb{V}(Y_i) = \mathbb{V}(\epsilon_i) = \sigma^2$.
— Les réponses et termes résiduels sont gaussiens et indépendants.

Les métriques usuelles d'évaluation de ce modèle de régression sont :

1. **Coefficient de Détermination (R^2)** : Le R^2 mesure la proportion de la variance totale des valeurs observées y_i expliquée par les valeurs prédites $\hat{y}_i$. Il varie de 0 à 1, où 1 indique que le modèle explique toute la variabilité des données.

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

2. **Erreur Quadratique Moyenne Racine** : La Root Mean Squared Error (RMSE) mesure la racine carrée de la moyenne des carrés des erreurs entre les valeurs réelles y_i et les valeurs prédites $\hat{y}_i$.

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

3. **Erreur Absolue Moyenne** : La Mean Absolute Error (MAE) calcule la moyenne des valeurs absolues des différences entre les valeurs réelles et les valeurs prédites. Elle est moins sensible aux valeurs aberrantes que la RMSE.

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

La base de données utilisée est celle hors valeurs manquantes des vitesses de rafales. Celle-ci est subdivisée en
— une base d'**apprentissage** qui contient **80 %** des données, et
— une base de **test** qui contient **20 %** des données.

L'ajustement du modèle linéaire sur la base d'entraînement donne comme relation entre la vitesse des rafales et la vitesse du vent :

$$\text{vitesse\_des\_rafales} = 1{,}50 \times \text{vitesse\_du\_vent} + 5{,}94$$

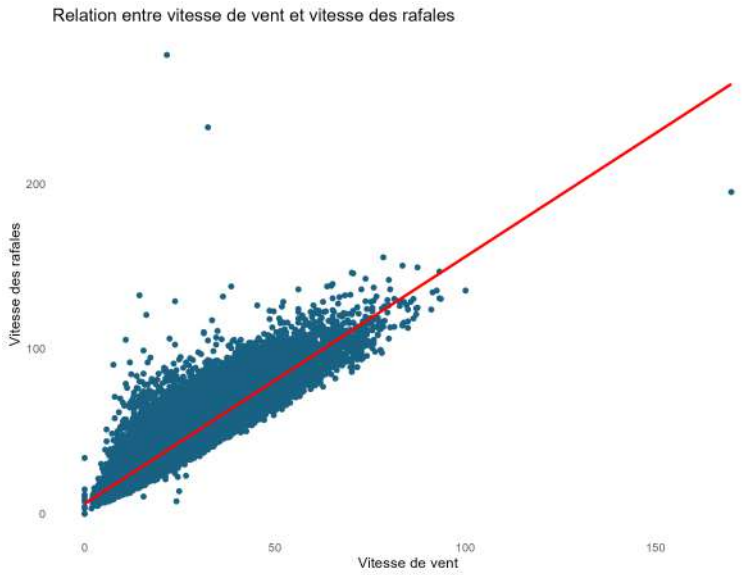


FIGURE 5.12 – Relation entre la vitesse des rafales et la vitesse du vent

Le modèle ne montre pas de signe de surapprentissage car les métriques sur les ensembles d'apprentissage et de test présentent des valeurs similaires comme l'indique le tableau suivant. Le coefficient de détermination, R^2 de **84,5 %** indique que le modèle explique une grande partie de la variance de la vitesse du vent, ce qui suggère qu'il s'ajuste bien aux données.

	R^2	RMSE	MAE
Base d'apprentissage	84,52 %	6,51	4,71
Base de test	84,29 %	6,58	4,70

TABLE 5.8 – Evaluation du modèle de régression

Grâce à cette relation, les vitesses des rafales pour chacune des zones peuvent être estimées à partir des données de vitesse du vent simulées. On obtient les résultats suivants pour l'année 2024 :

	Zone 1	Zone 2	Zone 3	Zone 4	Zone 5
Janvier	67,96	65,99	64,00	121,11	115,27
Février	75,49	79,26	82,50	74,21	82,89
Mars	102,15	76,32	67,82	161,43	124,00
Avril	64,78	66,75	72,57	74,60	128,83
Mai	59,95	80,18	81,58	83,29	106,88
Juin	103,37	82,43	62,99	82,09	112,84
Juillet	70,86	83,99	76,33	82,52	94,75
Août	81,71	67,48	83,87	86,77	112,20
Septembre	74,82	103,68	74,67	91,42	131,63
Octobre	67,13	82,65	112,06	124,07	101,61
Novembre	57,70	56,04	72,30	79,13	108,53
Décembre	62,61	75,83	75,51	134,75	108,03

TABLE 5.9 – Vitesse de rafales simulées par zones pour 2024

Les cellules en rouge clair indiquent les mois où une tempête est prévue. La cellule en rouge foncé représente un ouragan, car les rafales de vent y dépassent les 145 km/h. Les ouragans sont classés parmi les catastrophes naturelles réassurées par l'État via la CCR.

Il en ressort que la zone 5 est la plus à risque, avec une estimation de dix tempêtes pour l'année 2024, suivie de la zone 4 avec trois tempêtes. Les zones 2 et 3 sont les moins à risque, avec une seule tempête estimée chacune.

5.2.4 Détermination de la dérive de sinistralité

La dérive de sinistralité est mesurée en comparant le pourcentage de variation entre la moyenne historique du nombre de tempêtes et les projections obtenues par simulation pour les années futures.

Les tableaux suivants présentent le nombre de fois où la vitesse maximale mensuelle des rafales a dépassé les 100 km/h, tant dans les données historiques que dans les projections pour les années à venir.

	Zone 1	Zone 2	Zone 3	Zone 4	Zone 5
2016	2,00	2,00	1,00	4,00	9,00
2017	4,00	2,00	3,00	5,00	8,00
2018	3,00	2,00	2,00	7,00	8,00
2019	4,00	2,00	1,00	8,00	8,00
2020	2,00	3,00	3,00	4,00	9,00
2021	1,00	3,00	0,00	5,00	7,00
2022	2,00	3,00	2,00	5,00	6,00
2023	3,00	6,00	4,00	4,00	10,00
Moyenne 2019-2023	**2,40**	**2,43**	**2,00**	**5,20**	**8,00**

TABLE 5.10 – Nombre de tempêtes annuelles sur l'historique 2016 à 2023

	Zone 1	Zone 2	Zone 3	Zone 4	Zone 5
2024	2,00	1,00	1,00	3,00	10,00
2025	2,00	0,00	1,00	8,00	11,00
2026	4,00	3,00	2,00	8,00	10,00
2027	1,00	0,00	2,00	6,00	8,00
2028	3,00	4,00	2,00	9,00	9,00
Moyenne 2024-2028	**2,40**	**1,60**	**1,60**	**6,80**	**9,60**

TABLE 5.11 – Nombre estimé de tempêtes de 2024 à 2028

Une comparaison des prévisions et des données historiques, indépendamment de la zone, montre une augmentation prévue d'environ **5 %** du nombre de tempêtes à **l'horizon 2028**. Cette prévision est cohérente avec les chiffres publiés par l'ARGUS de l'assurance dans l'article intitulé *« Garanties tempête, grêle, neige : vers une rentabilité nulle ? »*. Cet article rapporte que Weather Claim Control et RiskWeatherTech estiment une dérive de sinistralité pour les garanties tempête, grêle et neige entre **1,4 % et 2,5 % par an**, en raison de l'augmentation des phénomènes orageux, des épisodes de grêle et des vents violents très localisés. *[L'ARGUS, 2022]*
La décomposition par zone révèle que le modèle prédit : une sinistralité inchangée pour la zone 1, une baisse de sinistralité pour les zones 2 et 3 et une hausse de sinistralité pour les zones 4 et 5.

	Zone 1	Zone 2	Zone 3	Zone 4	Zone 5
Dérive à l'horizon 2028	**0 %**	**-34 %**	**-20 %**	**31 %**	**20 %**

TABLE 5.12 – Dérive de la fréquence de sinistralité

5.3 Dérive de sinistralité du coût des sinistres

L'obtention d'une dérive de coût a été abordée par Inès BOUCHOUCHI dans son mémoire intitulé « *Défi climatique et durabilité, vers les limites de l'assurabilité ?* ». *[BOUCHOUCHI, 2024]*

La difficulté de la modélisation du coût des sinistres provient du fait que ces montants sont souvent considérés comme confidentiels. Pour contourner ce problème, une approche par benchmark est utilisée, en s'appuyant sur des publications de la CCR et France Assureurs. En particulier, les rapports intitulés *Conséquences du changement climatique sur le coût des catastrophes naturelles en France à horizon 2050 [CCR, 2024a]* et *Impact du changement climatique sur l'assurance à horizon 2050 [Assureurs, 2022]*, qui fournissent une évaluation de l'impact financier du réchauffement climatique à l'horizon 2050.

Risque	France Assureurs	CCR : Scénario 4.5	CCR : Scénario 8.5
Inondation	30 %	24 %	21 %
Sécheresse	17 %	15 %	28 %
Tempête	32 %	-	-

TABLE 5.13 – Hypothèses d'évolution du coût des sinistres à horizon 2050

Ces publications permettent d'estimer, en utilisant une **croissance exponentielle**, l'évolution des coûts des sinistres à l'horizon **2028**. Les prévisions sont résumées dans le tableau suivant :

Risque	Horizon	Évolution du coût
Inondation	2028	[3,6 ; 7,9[%
Sécheresse	2028	[2,6 ; 4,7[%
Tempête	2028	[0,0 ; 5,3[%

TABLE 5.14 – Hypothèse de dérive du coût des sinistres à horizon 2028

Le recours à une croissance exponentielle pour modéliser ces prévisions repose dans un premier temps sur un principe de prudence, et également sur l'observation faite dans la première partie du mémoire (Figure 1.3) : le graphique montrant l'évolution du nombre d'événements climatiques entre 1900 et 2022 révèle une augmentation exponentielle. Il semble donc raisonnable de supposer que les coûts associés aux événements climatiques suivront une tendance similaire, étant donné la corrélation probable entre la fréquence des sinistres et les coûts engendrés.

5.4 Limites de la méthodologie

— La principale limite de cette méthodologie réside dans le fait qu'elle ne modélise que la vitesse maximale du vent atteinte par mois. Cela implique qu'au sein d'un mois donné, pour une zone définie, une seule valeur de vitesse maximale est retenue. En conséquence, la méthodologie suppose implicitement qu'une seule tempête peut se produire par mois, ce qui limite le nombre total de tempêtes à un maximum de 12 par an. Cette hypothèse est simplificatrice et ne reflète pas la réalité, où plusieurs tempêtes peuvent survenir au cours d'un même mois.

— Aussi, en ne considérant que la vitesse maximale mensuelle, les variations intra-mensuelles et les événements extrêmes qui peuvent se produire à différentes périodes d'un même mois sont négligés.

— Les tempêtes successives peuvent être corrélées, en ne modélisant que les maxima mensuels, la dépendance temporelle entre les tempêtes n'est pas prise en compte.

— **Dérive des Coûts** :
La dérive des coûts a été estimée en utilisant un benchmark d'articles publiés. Cependant, ces chiffres sont susceptibles d'évoluer avec le temps, reflétant les nouvelles données et les changements économiques.

Troisième partie

Impact de la prise en compte des dérives dans le tarif

Cette partie vise à évaluer l'impact des dérives obtenues dans la partie précédente sur le tarif d'un portefeuille MRH. Plus précisément, elle a pour objectif de déterminer la suffisance et la viabilité des nouveaux taux de surprime, ainsi que d'analyser l'impact de la dérive de sinistralité des tempêtes sur la prime pure payée par les assurés. Pour rappel, les dérives de fréquence observées pour les risques d'inondation, de sécheresse et de tempêtes sont respectivement de **23 %**, **9 %** et **5 %**. De plus, un intervalle de confiance pour les dérives de coût a été établi à partir d'une comparaison avec des articles publiés.

Dans un premier temps, il sera question de revenir sur le secteur de l'assurance MRH, en présentant les différentes garanties ainsi que les principaux indicateurs du marché. Par la suite, une analyse descriptive de la base de données utilisée sera réalisée. La méthodologie utilisée pour évaluer les différents impacts des dérives sera alors explicitée, mise en œuvre, puis les résultats obtenus seront discutés.

Chapitre 6

Tarification

6.1 Assurance Multirisques Habitation

L'assurance multirisques habitation, aussi appelée « assurance MRH », a pour but de couvrir les coûts liés à un sinistre touchant un logement et son contenu. En France, cette assurance est obligatoire pour les locataires et les copropriétaires. Pour ces derniers, l'obligation se limite à la responsabilité civile envers la copropriété, les voisins, les tiers, et les locataires éventuels. Les garanties principales proposées sont :*[InfoService, 2024]*

- **Responsabilité civile** : Indemnise les tiers pour les dommages dont l'assuré est responsable.
- **Dégâts des eaux** : Couvre les dommages causés par des incidents tels que fuites d'eau, ruptures de conduites et infiltrations d'eau.
- **Incendie** : Protège le logement, les biens assurés et couvre les dommages corporels causés par le feu et la fumée.
- **Bris de glace** : Couvre les éléments en verre de l'habitation de l'assuré.
- **Vol** : Prend en charge les conséquences d'un vol ou d'un cambriolage.
- **Événements climatique (tempête, grêle, neige)** : Couvre les biens de l'assuré contre les dommages liés au vent, à la grêle et à la neige.
- **Protection juridique**
- **Dégâts électriques**
- **Attentats et actes de terrorisme** : Couvre les dommages dus à un attentat, sabotage, émeute, acte de terrorisme ou mouvement populaire.
- **Catastrophes naturelles** : Couvre les dommages causés par les catastrophes naturelles. Elle découle de la souscription à une assurance dommages aux biens, comme les incendies, dégâts des eaux, bris de glace, etc. Les catastrophes naturelles sont réassurées par la CCR, avec une surprime de 12 % des primes dommages

reversée par les assureurs à la CCR pour cette couverture. *[TAMET, 2024]*

6.1.1 Les principaux indicateurs de sinistralité

— **Fréquence** : La fréquence correspond au nombre de sinistres observés par unité d'exposition :

$$\text{Fréquence} = \frac{\text{Nombre de sinistres}}{\text{Exposition}}$$

— **Exposition** : L'exposition au risque est la période de temps, sur une durée donnée, pour laquelle l'assuré était couvert par l'assurance.

— **Coût Moyen** : Le coût moyen correspond à la charge totale divisée par le nombre de sinistres :

$$\text{Coût Moyen} = \frac{\text{Charge totale}}{\text{Nombre de réclamations}}$$

— **Prime pure** : La prime pure correspond au coût total divisée par l'exposition au risque.

$$\text{Prime pure} = \frac{\text{Charge totale}}{\text{Exposition totale}}$$

— **S/P** : Le S/P est la charge totale divisée par le total des primes acquises.

$$\text{S/P} = \frac{\text{Charge totale}}{\text{Prime acquise}}$$

— **Ratio combiné** : Le ratio combiné est la somme des frais de gestion et de la charge totale sur le total des primes encaissées. Un ratio supérieur à 100 % signifie que le produit n'est pas rentable.

6.1.2 Quelques chiffres du marché MRH en France

France Assureurs publie chaque année des données sur l'assurance habitation. En 2022, la prime moyenne de tous les contrats MRH s'élevait à 268 euros hors taxe, et les cotisations pour toutes les garanties confondues (y compris les catastrophes naturelles) atteignaient 12,2 milliards d'euros, soit une augmentation de 4,1 % par rapport à 2021. Cette augmentation des primes de l'ensemble des dommages aux biens des particuliers n'a pas été suffisante pour compenser la hausse notable des charges de prestations (+29,1 %). Cette évolution a contribué à la détérioration du ratio combiné comptable net de réassurance, qui a atteint 102,7 % en 2022. En incluant les catastrophes naturelles, ce ratio s'élève à 105,3 % des primes, soit une dégradation de 5,6 points sur un an, principalement en raison d'une augmentation significative des coûts climatiques sur l'année (doublement des charges de prestations liées aux catastrophes naturelles en 2022). *[Assureurs, 2023]*

6.2 Présentation des bases de données

Les bases de données utilisées sont les bases polices et sinistres d'une compagnie d'assurance pour la période **2016 à 2022**. Ces bases avaient déjà été exploitées en interne dans le cadre d'une mission et un pré-traitement avait donc déjà été effectué. La base police contient les détails des contrats. Les variables présentes dans cette base sont :

— Numéro de police
— Type de résidence : principale ou secondaire
— Type d'habitation : maison ou appartement
— Nombre de pièces
— Situation juridique : locataire ou propriétaire
— Valeur des objets
— Zone géographique
— Formule de garantie : quatre niveaux existent : Essentiel, Medium, Confort, Tous risques
— Exposition

La base des sinistres contient les détails relatifs aux sinistres, le lien entre les deux bases est établi via le numéro de police. Les variables présentes dans cette base sont :

— Numéro de police
— Nombre de sinistres par garanties (incendie, vol, dégâts des eaux, dégâts électriques, responsabilité civile, protection juridique, catastrophes naturelles)
— Coût des sinistres par garanties (incendie, vol, dégâts des eaux, dégâts électriques, responsabilité civile, protection juridique, catastrophes naturelles)

6.2.1 Ajout de variables

En plus de ces variables, trois autres variables ont été rajoutées.
— Le zonier du risque tempête construit en partie II (Figure 5.3)
— Les zoniers inondation et sécheresse construits à partir du nombre d'arrêtés de catastrophes naturelles par département (**voir annexe**).

Mise en « AS-IF » des coûts

La base de données couvrant les années 2016 à 2022, il est essentiel d'ajuster les diverses variables de coût en fonction de l'inflation pour permettre une comparaison pertinente. À cette fin, l'indice du coût de la construction fourni par la Fédération Française du Bâtiment (ICC-FFB) est utilisé. Cet indice, mis à jour trimestriellement par la FFB, est calculé sur la base du coût de construction d'un immeuble à Paris, en excluant le prix du terrain. La référence de cet indice est fixée à 1 au 1er janvier 1941. L'ICC-FFB permet de suivre les variations des différents composants (matériaux, main-d'œuvre, services divers, etc.) impliqués dans la construction d'un bâtiment. Cet indice est ainsi employé

pour ajuster les coûts, c'est-à-dire pour recalculer les montants des sinistres aux coûts qu'ils auraient engendrés s'ils étaient survenus durant l'année de référence, qui est 2022 dans ce cas.

$$\text{Coût 2016 indexé} = \frac{\text{Indice 2022}}{\text{Indice 2016}} \times \text{Coût 2022}$$

6.2.2 Détermination des seuils des sinistres graves

Un sinistre grave est un sinistre avec une fréquence faible mais qui engendre des coûts au dessus de la normale. Ces coûts quand pris en compte dans un modèle peuvent biaiser les résultats et doivent être étudiés séparément. Pour identifier de tels sinistres, la théorie des valeurs extrêmes est employée via la méthode des dépassements de seuil.

Soit X une variable aléatoire. Considérons $X_1, \ldots, X_n$ une suite de variables aléatoires indépendantes et identiquement distribuées, et (u_n) une suite de réels. On s'intéresse aux dépassements du seuil u, c'est à dire aux observations $(X_i - u)_+$. Pour une variable aléatoire X ayant une fonction de répartition F, *la distribution des excès au dessus d'un seuil* u est définie par :

$$F_u(x) = P[X - u \leq x \mid X > u] \tag{2.3}$$

On pose $Y_{u,i} = X_i - u$.

Loi de Pareto Généralisée (GPD)

La distribution de Pareto généralisée $GPD(\beta, \xi)$ se définit comme suit :

$$G^p_{\xi,\beta}(x) = \begin{cases} 1 - \left(1 + \xi \frac{x}{\beta}\right)^{-\frac{1}{\xi}}, & \text{si } \xi \neq 0 \\ 1 - \exp\left(-\frac{x}{\beta}\right), & \text{si } \xi = 0 \end{cases}$$

où

$$x \geq 0 \quad \text{si } \xi \geq 0$$

$$0 \leq x \leq -\frac{\beta}{\xi} \quad \text{si } \xi < 0$$

On suppose que les dépassements $Y_{u,1}, \ldots, Y_{u,n}$ au delà du seuil u suivent une loi $GPD(\sigma_u, \xi)$ telle que

$$P(Y_u < y \mid Y_u > 0) = 1 - \left[1 + \xi \left(\frac{y}{\sigma_u}\right)\right]^{-\frac{1}{\xi}}_+$$

La difficulté de modélisation à l'aide de la méthode des seuils réside dans le choix d'un seuil approprié. En effet, plus le seuil est élevé, plus l'approximation asymptotique de la distribution des excédents par la Loi de Pareto Généralisée (GPD) est juste et plus le biais diminue. À l'inverse, un seuil plus bas augmente le nombre d'excédents pour l'estimation des paramètres de la GPD et réduit ainsi la variance de ces estimations. Le choix du seuil peut être fait à l'aide d'outils graphiques

Graphique de Hill

Considérons $X_1 > \ldots > X_n$ les statistiques d'ordre de variables aléatoires indépendantes et identiquement distribuées. L'estimateur de Hill pour l'indice de queue ξ est défini par :

$$H_{k,n} = \frac{1}{k}\sum_{i=1}^{k}(\ln X_i - \ln X_{k+1})$$

Si $\xi > 0$ alors $H_{k,n} \xrightarrow{P} \alpha = \frac{1}{\xi}$. Le graphe de Hill représente l'ensemble des points $\{(k, H_{k,n}^{-1}), 1 \leq k \leq n-1\}$

L'objectif est de déterminer la valeur de k à partir de laquelle l'estimation se stabilise, c'est-à-dire lorsque la courbe devient une droite horizontale. Cette valeur indique le nombre d'observations à considérer comme extrêmes et, par conséquent, le seuil à utiliser. Le graphique ci-dessous illustre cette application pour la garantie ***dégâts des eaux***.

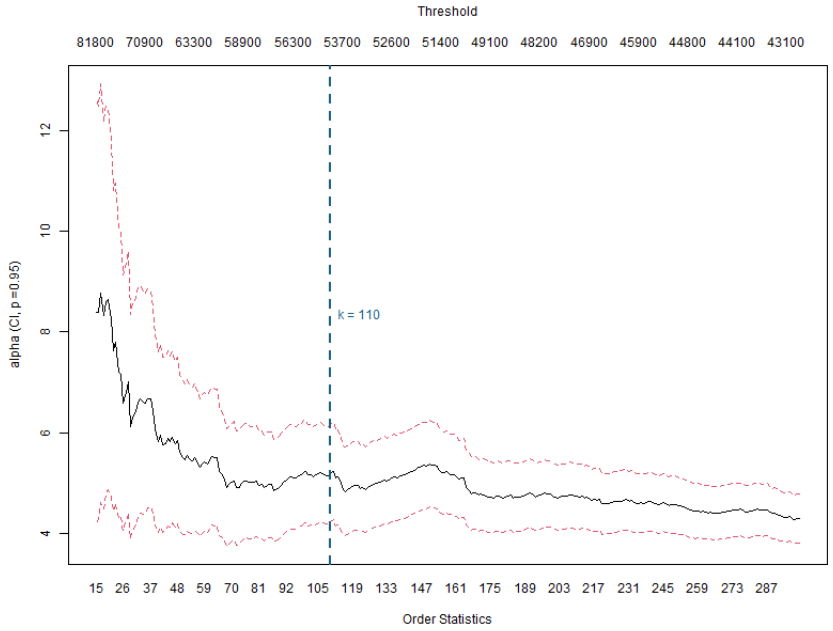


FIGURE 6.1 – Graphique de hill : garantie dégâts des eaux

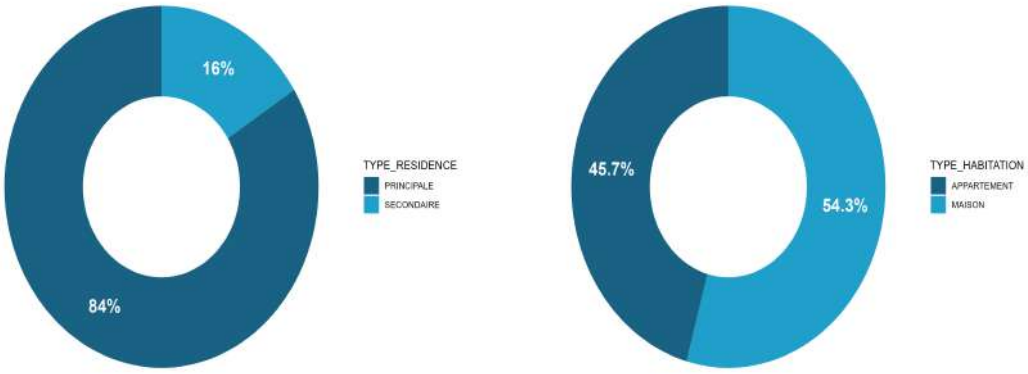
Le graphique de Hill indique une valeur de k égale à **110**, correspondant à un seuil de 54 000 euros. Au-delà de ce seuil, le coût des sinistres est considéré comme extrême pour cette garantie. La même méthode a été utilisée pour les autres garanties dommages.
Les résultats obtenus sont résumés dans le tableau suivant :

Garantie	k	Seuil
Vol	170	48 500
Incendie	71	190 000
Bris de glaces	29	15 000
Dégâts électriques	37	30 500
Tempête, Grêle, Neige	75	98 000

TABLE 6.1 – Seuil des valeurs extrêmes des garanties dommages

6.2.3 Analyses descriptives

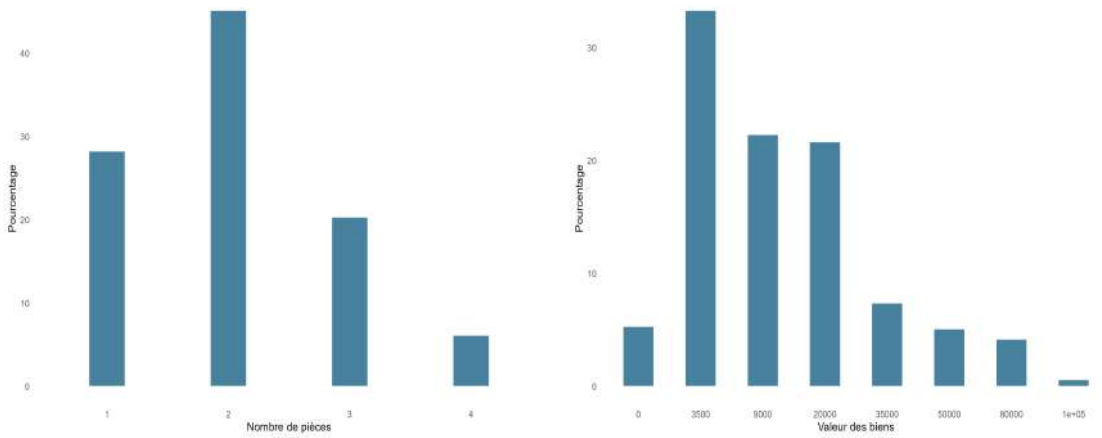
Avant toute modélisation, il est important d'analyser les variables et en déduire les relations entre celles-ci. Les logements assurés sont majoritairement des résidences principales, représentant 84 % du total des logements. En ce qui concerne le type de logement, la répartition entre appartements et maisons est proche de l'équilibre, avec 45,7 % des logements étant des appartements et 54,3 % des maisons. La répartition entre propriétaires et locataires est également proche de l'équilibre avec 55,5 % pour les locataires et 44,5 % pour les propriétaires.



(a) Type de résidence (b) Type d'habitation

FIGURE 6.2 – Répartition des variables dans la base de données

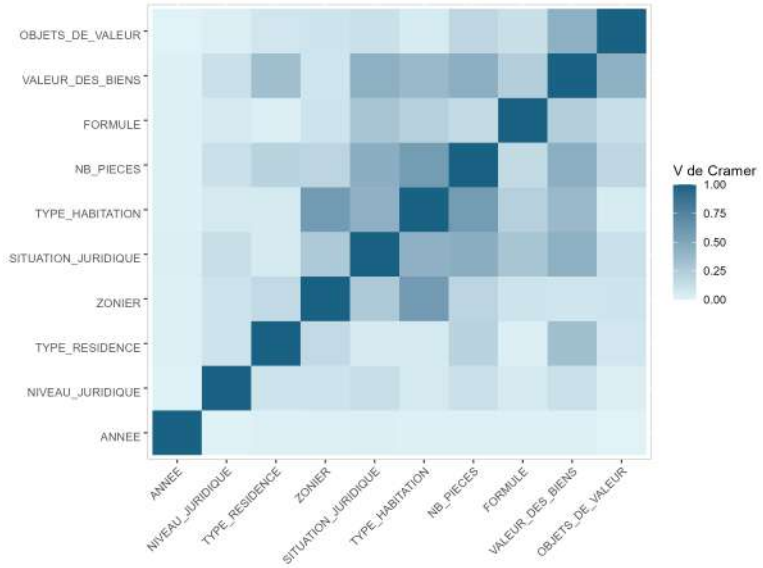
La répartition des logements selon le nombre de pièces présentée dans le graphique suivant révèle que la majorité des logements dispose de deux pièces ou moins. De plus, la valeur des objets présents dans ces logements est en majeur partie entre 3 500 et 20 000 euros.



(a) Répartition du nombre de pièces (b) Répartition de la valeur des objets

FIGURE 6.3 – Répartition des variables dans la base de données

Les variables étant essentiellement catégorielles, le **coefficient V de Cramer** est utilisé pour mesurer la force de l'association entre celles-ci. Le V de Cramer est une adaptation du test du chi-carré et est particulièrement utile pour évaluer la force de l'association lorsque l'une ou les deux variables ont plus de deux catégories. Le graphique suivant montre les corrélations entre les différentes variables :



Les résultats montrent qu'il n'y a pas de corrélation forte entre les variables, mais on observe quelques associations modérées, notamment entre la situation juridique et le nombre de pièces, le nombre de pièces et le type d'habitation, le type d'habitation

et le zonier, etc. Aucune corrélation forte n'ayant été identifiée, il n'y a pas de redondance d'information significative entre ces variables. Par conséquent, toutes ces variables peuvent potentiellement être intégrées dans les modèles sans risque de multicolinéarité excessive.

La répartition de la fréquence de sinistralité par garantie est la suivante :

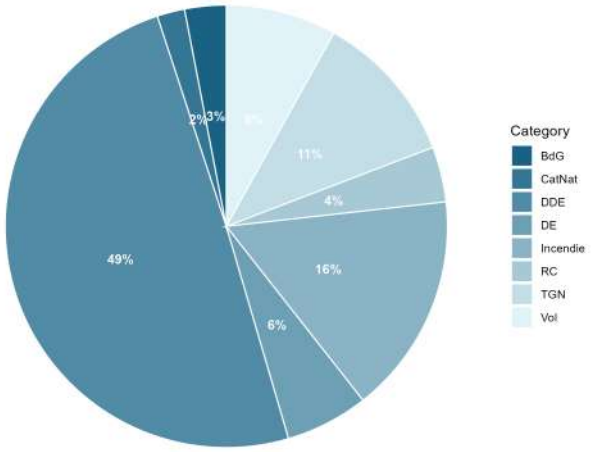


FIGURE 6.4 – Répartition de la sinistralité par garantie

Les **dégâts des eaux** sont les plus courants, représentant près de la moitié des sinistres. Ils sont suivis par les incendies et les événements climatiques tels que les tempêtes, la grêle et la neige. Bien que les catastrophes naturelles soient les moins fréquentes, ne représentant que 2 % de la fréquence totale des sinistres du portefeuille, elles se classent néanmoins en quatrième position parmi les garanties les plus coûteuses. Cela souligne que malgré leur rareté, les catastrophes naturelles engendrent des coûts de sinistralité élevés.

Méthodologie

Dans la première partie, il a été établi que la surprime Cat-Nat reversée à la CCR dans le cadre du régime des catastrophes naturelles est calculée comme **12 % de la prime afférente aux garanties dommages** pour un contrat MRH. Cette surprime Cat-Nat augmentera à **20 %** à partir de 2025. Afin d'évaluer la suffisance et viabilité de ces 20 % au regard des dérives de sinistralité précédemment obtenues les étapes suivantes seront suivies :

1. **Détermination de la prime pure des garanties dommages :**

La prime pure des garanties dommages sera déterminée à l'aide des GLM **Fréquence**× **Coût**

2. **Modélisation de la prime pure des catastrophes naturelles :**
   Une fois la prime pure des garanties dommages déterminée, la prime pure Cat-Nat sera modélisée à l'aide du modèle de Tweedie. Ceci permettra de déterminer quelle proportion cette prime représente par rapport à la prime dommages totale et évaluer dans quelle mesure celle-ci s'écarte de la surprime Cat-Nat actuelle de 12 %.
3. **Prise en compte des dérives dans la tarification future :**
   Les dérives obtenues dans la partie précédente seront prises en compte dans la tarification afin de fournir une estimation de la surprime nécessaire pour couvrir la sinistralité des années à venir et d'évaluer la suffisance de la surprime de 20 % prévus pour les prochaines années, sur le portefeuille. Pour rappel, les dérives de fréquence obtenus pour les risques d'inondation, de sécheresse et de tempêtes à l'horizon 2028 sont respectivement **23 %, 9 % et 5 %**.

6.3 Détermination de la suffisance et viabilité du taux de surprime de 20 %

6.3.1 Détermination de la prime pure dommages

Soit N, une variable aléatoire à valeurs dans $\mathbb{N}$ représentant le nombre total de sinistres et $Y = (Y_i)_{i>0}$, une suite de variables aléatoires à valeurs dans $\mathbb{R}^+$ telle que pour tout i, Y_i représente le coût du sinistre i. La charge totale de sinistres S s'écrit :

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

Hypothèses

— Pour tout i strictement positif, les variables aléatoires N et Y_i sont supposées indépendantes ;
— La suite de variables aléatoires $Y = (Y_i)_{i>0}$ est supposée iid [1].

Grâce aux hypothèses précédentes, la prime pure est définie comme l'espérance de la charge totale des sinistres.

$$E(S) = E\left(\sum_{i=1}^{N} Y_i\right) = E(N)E(Y_1) = \text{Fréquence} \times \text{Coût Moyen}$$

Pour modéliser ces composantes, les GLM avec une distribution de **Poisson** ou **Binomiale Négative** sont couramment utilisés pour la fréquence des sinistres. Pour le

1. indépendantes et identiquement distribuées

coût moyen des sinistres, les distributions **Gamma** ou **Lognormale** sont souvent appropriées. La théorie sur les GLM, détaillée dans la section précédente dédiée aux modèles de classification (4.3.1), fournit un cadre pour ces estimations.

La prime pure est calculée pour chaque garantie, et la prime pure dommages est obtenue comme étant la somme des primes pures des différentes garanties :

$$E(S) = E(S_{\text{dégâts des eaux}}) + E(S_{\text{incendie}}) + E(S_{\text{vol}}) + E(S_{\text{bris de glaces}}) + E(S_{\text{dégâts électriques}}) + \cdots$$

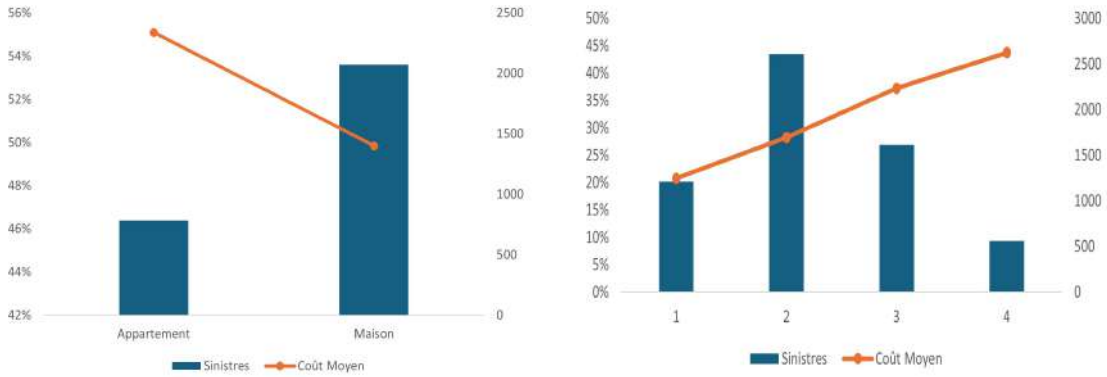
$$E(S) = E(S_{\text{attriti\_DDE}}) + E(S_{\text{grave\_DDE}}) + E(S_{\text{attriti\_INC}}) + E(S_{\text{grave\_INC}}) + \cdots$$

La détermination de la prime pure sera effectuée en deux étapes. Dans un premier temps, la prime pure des **sinistres attritionnels** sera calculée. Ensuite, celle des **sinistres graves** sera déterminée, et ce pour chaque garantie. Étant donné que la même méthodologie est appliquée à toutes les garanties dommages, seul le modèle des **dégâts des eaux** sera présenté, étant le risque le plus fréquent.

Modèle GLM pour la garantie dégâts des eaux

Analyses statistiques préliminaires

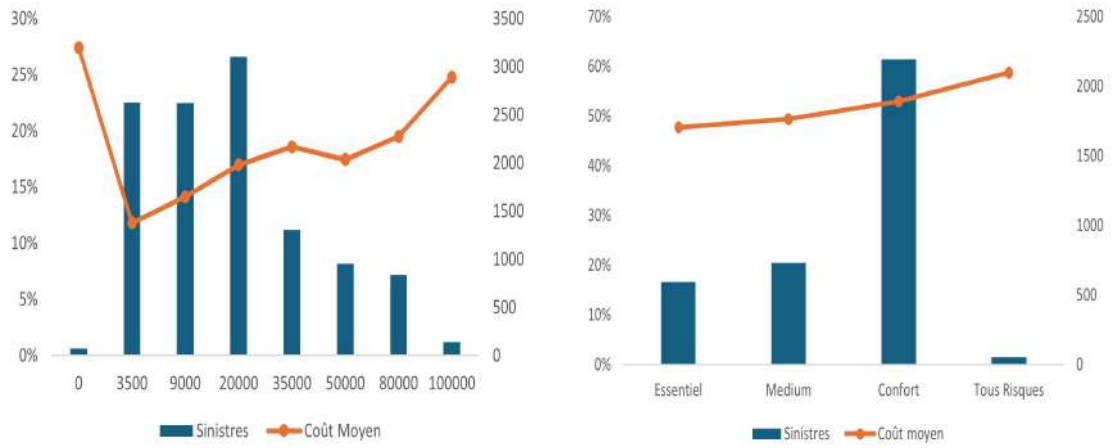
Quelques analyses du coût moyen des sinistres en fonction des variables présentes dans la base de données montre que la répartition des sinistres selon le type d'habitation est proportionnelle à la répartition des maisons et appartements dans notre base de données. Cependant, le coût moyen des sinistres est plus élevé pour les appartements que pour les maisons. En ce qui concerne la répartition du coût moyen par nombre de pièces, une tendance à la hausse est observée. Plus le logement comporte de pièces, plus le coût moyen des sinistres augmente.



(a) Coût moyen par type d'habitation (b) Coût moyen par nombre de pièces

FIGURE 6.5 – Comparaison des coûts moyens

Les logements ayant des objets de valeur inférieure à 3 500 euros présentent un coût moyen de sinistre le plus élevé. Cette anomalie s'explique par le faible nombre de sinistres dans cette catégorie, ce qui biaise la moyenne. En dehors des objets de valeur inférieure à 3 500 euros, le coût moyen augmente proportionnellement à la valeur des objets.
Le coût moyen des sinistres augmente également en fonction de la formule de garantie souscrite. Plus la formule est étendue, plus le coût moyen des sinistres est élevé.



(a) Coût moyen par valeur des objets

(b) Coût moyen par formule

FIGURE 6.6 – Comparaison des coûts moyens

1 - Détermination de la prime pure attritionnelle

Choix de loi pour le modèle de fréquence

Les deux lois couramment utilisées pour la calibration d'un GLM fréquence sont les lois de **Poisson** et **Binomiale Négative**. Afin de déterminer laquelle de ces lois est la plus appropriée pour modéliser le nombre de sinistres, le package *fitdistrplus* de R a été utilisé pour ajuster ces deux distributions aux données. Le critère d'AIC a été utilisé pour comparer les modèles. La loi de **Poisson** a été retenue, car elle a présenté l'AIC le plus bas parmi les deux lois ajustées.

	AIC
Poisson	**27 926,38**
Binomiale Négative	28 006,71

TABLE 6.2 – AIC obtenu en fonction de la loi de fréquence ajustée

Cependant, la loi de Poisson repose sur l'hypothèse que l'espérance et la variance sont égales, ce qui n'est pas toujours le cas dans la pratique. Dans ce modèle, la variance

dépasse l'espérance de 4 %. En raison de cette surdispersion, la loi de **Poisson surdispersée** sera utilisée pour modéliser la fréquence des sinistres.

Choix de loi pour le modèle de coût

Le critère d'AIC a également été utilisé pour sélectionner la loi la mieux adaptée aux données. Le modèle dont l'AIC est le plus faible a été choisi pour la modélisation. En l'occurrence, la distribution **Gamma** a présenté le plus petit AIC et a donc été retenue pour la modélisation.

	AIC
Gamma	**43 828,49**
Log-Normale	43 978,86
Weibull	44 826,42

TABLE 6.3 – AIC obtenu en fonction de la loi de coût ajustée

Sélection des variables

Pour identifier les variables les plus pertinentes à inclure dans le modèle, les critères d'**AIC** et de **déviance** ont été utilisés. Chaque variable a d'abord été intégrée individuellement dans un GLM, puis les autres variables ont été rajoutées progressivement pour évaluer leur contribution à la performance globale du modèle. L'importance d'une variable est déterminée par la réduction de l'AIC ou de la déviance qu'elle entraîne. Plus une variable réduit l'AIC ou la déviance du modèle, plus elle est considérée comme significative.
Les résultats de cette analyse sont explicités dans le tableau suivant :

Variable	Déviance	AIC	Importance (%)
Zonier	2 094,05	1 544,68	76,25
Nombre de pièces	222,12	194,99	8,08
Type de résidence	151,59	144,80	5,52
Formule	107,16	86,81	3,90
Valeur des biens	99,70	51,88	3,63
Type d'habitation	38,66	32,22	1,41
Objets de valeur	20,01	13,23	0,73
Situation juridique	8,14	1,36	0,30
Niveau juridique	4,75	-2,03	0,17

TABLE 6.4 – Sélection de variables par méthode forward

Le *zonier* est la variable la plus influente, avec une importance de 76 %, suivie du nombre de pièces et du type de résidence. Seules les cinq variables les plus significatives seront retenues pour le modèle final afin de garantir la robustesse et parsimonie du

modèle. Une analyse des interactions entre les variables a également été réalisée, mais aucune interaction significative n'a été détectée.

Validation du modèle

Pour valider le modèle, la base de données a été divisée en deux :
— une base d'apprentissage contenant **80 %** des données, et
— une base de test contenant **20 %** des données.

Le **test d'adéquation de déviance** permet de mesurer la qualité d'ajustement du modèle. Si la déviance du modèle est inférieure au quantile 0,95 d'une loi du khi-deux avec les degrés de liberté correspondants, la qualité d'ajustement du modèle est considérée comme satisfaisante. Les résultats suivants ont été obtenus suite à la modélisation :

Modèle	Degré de liberté	Quantile 0,95 de la loi χ^2	Déviance résiduelle
Fréquence	422 174	423 686	57 200
Coût	14 709	14 993	13 494

TABLE 6.5 – Validation des modèles par test de déviance

La déviance résiduelle des modèles de fréquence et de coût est bien inférieure au quantile 0,95 de la loi de χ^2, indiquant que l'ajustement des modèles est satisfaisant.
Les résultats des prédictions pour ces modèles sont les suivants :

Modèle	Réel	Prédit	Ecart
Fréquence moyenne	0,0350	0,0338	-3,55 %
Coût moyen	1 836,22	1 821,42	-0,69 %

TABLE 6.6 – Résultat du modèle Fréquence X Coût pour les dégâts des eaux

Les écarts absolus étant très faibles (inférieurs à 5 %), il est possible de conclure que les modèles sont satisfaisants.

La même méthodologie a été appliquée pour modéliser les autres garanties dommages. Les résultats de ces prédictions sont également satisfaisants et sont présentés dans le tableau suivant :

	Coût Moyen			Fréquence		
Garantie	**Réel**	**Prédit**	**Ecart**	**Réel**	**Prédit**	**Ecart**
Vol	1 564,94	1 562,13	-0,18 %	0,0079	0,0076	-4,65 %
Incendie	8 240,66	8 236,83	-0,05 %	0,0043	0,0041	-2,85 %
Bris de glace	691,59	690,03	-0,23 %	0,0070	0,0068	-3,05 %
Dégâts électriques	1 572,30	1 568,66	-0,23 %	0,0098	0,0094	-3,22 %
Tempête Grêle Neige	4 572,53	4 566,89	-0,12 %	0,0115	0,0112	-3,04 %

TABLE 6.7 – Résultat du modèle Fréquence X Coût sur les autres garanties dommages

b - Détermination de la prime pure des sinistres extrêmes

Pour modéliser la prime pure des sinistres extrêmes, la méthode **Peak Over Threshold (POT)** a été utilisée. Cette méthode suppose que les excès au-dessus d'un seuil défini suivent une distribution de Pareto. Le seuil dans ce cas est celui obtenu précédemment à l'aide du graphique de Hill (Tableau 6.1). Pour un seuil u, l'espérance des excès est donnée par :

$$E(X - u \mid X > u) = \frac{\sigma}{1 - \xi}$$

où

- σ est le paramètre d'échelle donné pour le seuil u et
- ξ est le paramètre de forme tel que $\xi < 1$ afin de garantir que la moyenne existe. *[Yue et co, 2016]*

Soit λ_u la fréquence des sinistres dépassant le seuil u. La prime pure est donnée par

$$P = E(S) = \frac{\lambda_u \sigma_u}{1 - \xi}$$

Le modèle POT est ajusté en utilisant la fonction *fevd* du package `extRemes` sur R. Les paramètres estimés du modèle pour chaque garantie sont présentés dans le tableau ci-dessous :

Garantie	ξ	σ	λ
Dégats des eaux	0,21	6 909,13	$2,08 \times 10^{-}4$
Vol	0,05	10 418,41	$3,22 \times 10^{-}4$
Incendie	0,04	37 116,54	$1,35 \times 10^{-}4$
Bris de glace	0,21	1 349,74	$5,44 \times 10^{-}5$
Dégâts électriques	0,17	3 650,08	$7,01 \times 10^{-}5$
Tempête, Grêle, Neige	0,05	13 878,85	$1,42 \times 10^{-}4$

TABLE 6.8 – Paramètres du modèle POT ajusté par garantie

La validation du modèle a été effectuée à l'aide du test de Kolmogorov-Smirnov. Les p-valeurs obtenues pour chacune des garanties sont toutes supérieures à 5 %, indiquant un bon ajustement du modèle.

6.3.2 Détermination de la prime pure catastrophes naturelles

Les catastrophes naturelles étant peu fréquentes en comparaison aux autres risques d'une assurance MRH, le choix du modèle pour la détermination de la prime pure s'est orienté vers les modèles de Tweedie. En effet, le modèle de Tweedie permet de modéliser simultanément la fréquence et le coût moyen des sinistres, ce qui équivaut à modéliser directement la prime pure. Ce modèle est particulièrement adapté lorsque les données présentent une forte asymétrie ou contiennent de nombreuses valeurs nulles, comme c'est souvent le cas pour les dommages liés aux catastrophes naturelles. La prime pure catastrophe naturelle est calculée comme étant la somme de la prime pure inondation et la prime pure sécheresse

Modèles de Tweedie

La famille de Tweedie est une sous-classe des lois exponentielles. Contrairement à de nombreuses distributions classiques, la loi de Tweedie n'a pas de formule fermée pour sa densité. Toutefois, elle est définie par une fonction de variance spécifique, ce qui la rend utile dans de nombreux contextes pratiques. Pour une variable aléatoire Y ayant une distribution de Tweedie, sa variance peut être exprimée en fonction de sa moyenne μ par :

$$Var(Y) = \phi\mu^p$$

où
— ϕ est le paramètre de dispersion et
— p le paramètre de puissance, qui détermine la forme de la distribution

Valeur de p	Distribution
$p = 0$	Normale
$p = 1$	Poisson
$p \in]1, 2[$	Poisson-Gamma composé
$p = 2$	Gamma
$p = 3$	Normale inverse

TABLE 6.9 – Distributions usuelles retrouvées dans la famille Tweedie

Afin de déterminer la valeur de p la plus adéquate pour les modèles, la fonction *tweedie.profile* de R a été utilisée. Elle permet d'obtenir une valeur de p de **1,37** pour le modèle d'inondation et **1,35** pour le modèle de sécheresse. **Seuls les résultats de la modélisation de la prime pure sécheresse seront explicités**. La même méthodologie a été utilisée pour déterminer la prime pure inondation.

Sélection des variables

Tout comme pour les GLM Fréquence X Coût, les critères d'AIC et de déviance ont été utilisés pour la sélection des variables. Le **zonier sécheresse** se démarque comme

étant la variable la plus importante du modèle avec une importance de **90 %**, suivi par le **type d'habitation**.

Validation des modèles

Tout comme pour le modèle précédent, le critère de déviance a été utilisé pour valider le modèle. Cependant, ce critère n'a pas été respecté, car la déviance résiduelle était supérieure au quantile 0,95 d'une loi de χ^2. Cette non-conformité pourrait être attribuée à la rareté des sinistres ou une insuffisance des variables explicatives pertinentes. Malgré cela, l'écart de prédiction moyenne absolue est de **0,53 %**. La prime moyenne observée est de 20,18 euros, tandis que la prime prédite par le modèle est de 20,07 euros. Cet écart étant faible, et étant donné l'intérêt pour la prime moyenne afin de déterminer la proportion que celle-ci représente en fonction de la prime pure des dommages, ce modèle a été pris en compte.

Pour vérifier la robustesse du modèle et sa capacité à identifier les zones les plus risquées à partir du zonier, une validation par k-fold (avec 5 blocs) a été réalisée en utilisant le **coefficient de Gini** comme métrique. Le coefficient de Gini mesure l'inégalité de la distribution d'une variable, variant entre 0 (égalité parfaite) et 1 (inégalité maximale). Dans le cadre de ce modèle, le coefficient de Gini mesure l'inégalité dans les prédictions. Le graphique suivant montre le coefficient de gini obtenu par pli :

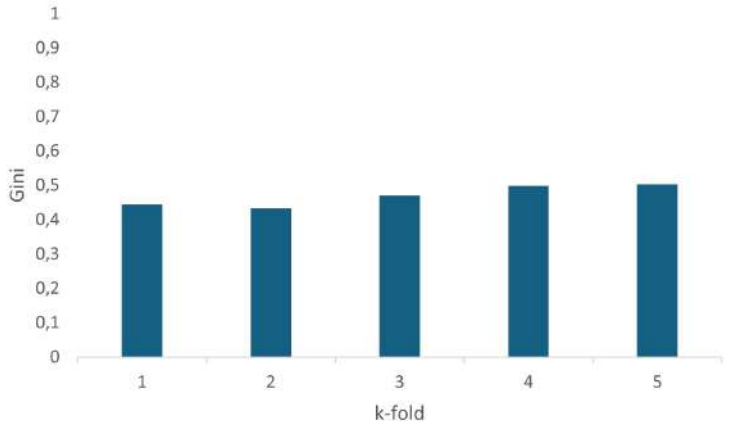


FIGURE 6.7 – Validation croisée du modèle Tweedie : Sécheresse

La valeur du coefficient de Gini est relativement constante entre les plis (entre 0,44 et 0,50), indiquant que le modèle est robuste et ne présente pas de problèmes de surapprentissage. Le coefficient de Gini moyen est de **0,47**, suggérant une répartition modérément inégale des prédictions et indiquant que le modèle capture tant bien que mal les différences de risque entre les assurés. En effet, la prime pure moyenne pour les zones 4 et 5, qui sont les zones les plus à risque, est nettement plus grande que celle des zones 1, 2 et

3, ce qui confirme son utilité malgré ses limitations initiales.

Le modèle de détermination de la prime pure inondation n'a pas non plus satisfait le test de déviance. Tout comme le modèle de détermination de la prime pure sécheresse, l'écart de prédiction moyenne était très faible. La validation croisée utilisant comme métrique le coefficient de Gini a démontré une stabilité de celui-ci, indiquant ainsi la robustesse du modèle. Le coefficient de Gini moyen était de **0,53**, ce qui montre une capacité moyenne du modèle à distinguer le risque porté par les assurés. (**voir annexe**)

Le tableau suivant montre la prime pure moyenne estimée par zone pour les risques d'inondation et de sécheresse :

	Tarif Sécheresse		Tarif Inondation	
Zones	**Prime pure (€)**	**% de la PP moyenne**	**Prime pure (€)**	**% de la PP moyenne**
Zone 1	4,75	23,67 %	1,29	18,35 %
Zone 2	10,37	51,65 %	3,06	43,58 %
Zone 3	14,80	73,68 %	5,92	84,46 %
Zone 4	27,71	138,00 %	19,15	273,23 %
Zone 5	33,75	168,06 %	20,97	299,12 %

TABLE 6.10 – Prime pure par zone pour les modèles de sécheresse et inondation

Détermination de la proportion de la prime pure Cat-Nat

Les différents GLM Fréquence X Coût ont permis de déterminer la prime pure dommages en faisant la somme de la prime pure moyenne obtenue pour chaque garantie. Les modèles tweedie ont permis de déterminer la prime pure moyenne catastrophe naturelle, si elle devait être calculée à partir du risque porté par l'assureur. La prime pure moyenne globale pour les catastrophes naturelles s'élève à **27,21 euros**, représentant **14,68 %** de la prime pure dommages. Cette proportion est supérieure aux 12 %, soit en moyenne par contrat 22,24 euros reversés comme surprime à la CCR par l'assureur, contribuant ainsi à la fragilisation du régime des catastrophes naturelles et donc la nécessité de l'augmentation du taux de surprime par la CCR.

		Proportion
Prime pure moyenne dommages	185,34 €	14,68 %
Prime pure moyenne catastrophes naturelles	27,21 €	

TABLE 6.11 – Proportion de la prime pure catastrophe naturelle en fonction de la prime pure dommages

Le tableau suivant présente la décomposition du taux de surprime par zone, en fonction de la prime pure moyenne des catastrophes naturelles pour chacune de ces zones. Ces valeurs sont calculées à partir des primes pures par zone et par type de risque, comme indiqué dans le tableau 6.10, en supposant une équivalence entre les zones de

risques d'inondation et de sécheresse. Cette approche offre une vue d'ensemble du taux de surprime en fonction du niveau d'exposition au risque combiné d'inondation et de sécheresse, la zone 1 étant la moins exposée et la zone 5 la plus exposée.

	Zone 1	Zone 2	Zone 3	Zone 4	Zone 5
Prime pure moyenne Cat-Nat	6,04 €	13,43 €	20,72 €	46,86 €	54,71 €
Taux de surprime	3,26 %	7,24 %	11,18 %	25,29 %	29,52 %

TABLE 6.12 – Taux de surprime estimé par zone

On observe une variation significative du taux de surprime selon le niveau d'exposition :

- Les zones 1 et 2 sont les moins exposées, avec des taux de surprime de respectivement 3,26 % et 7,24 %. Ces taux reflètent une faible probabilité de sinistres majeurs, ce qui se traduit par une surprime relativement basse avec des expositions plutôt faibles.
- La zone 3 représente une exposition moyenne, la surprime estimée pour cette zone est de 11,18 %, proche du taux légal actuel de 12 %. Cela indique que pour les zones avec une exposition intermédiaire, le taux de surprime reste aligné avec le taux actuel.
- Les zones 4 et 5 sont les plus exposées aux risques, avec des taux de surprime respectivement de 25,29 % et 29,52 %. Ces valeurs sont plus de deux fois supérieures au taux légal actuel, et bien au-delà du taux de 20 % qui sera applicable à partir de 2025.

Cette observation soulève la question de savoir si la mutualisation des risques pourra être maintenue à long terme, compte tenu de la disparité de l'exposition aux risques entre les différentes zones. En particulier, pour les zones les plus à risque, il est possible que l'occurrence de ces événements devienne presque certaine dans les années à venir, ce qui pourrait remettre en cause la viabilité de cette mutualisation.

Les analyses présentées dans la prochaine partie continuent de supposer une mutualisation des risques à l'échelle nationale, comme c'est actuellement le cas.

6.3.3 Prise en compte des dérives inondation et sécheresse

Dès 2025, le taux de surprime passera à **20 % de la prime pure des garanties dommages**. Pour évaluer la suffisance et la viabilité de ce taux dans les années à venir, les dérives de fréquence et de coût obtenues précédemment ont été utilisées.

- Une dérive de hausse de fréquence du risque inondation à hauteur de **23 %** a été obtenu précédemment. Pour la prendre en compte dans le tarif, une simulation d'une augmentation de 23 % du nombre de sinistres inondation futurs a été réalisée, en tenant compte de l'exposition au risque des différentes zones et du

nombre de personnes assurées dans ces zones : plus de sinistres affectés aux zones les plus exposées et ayant un plus grand nombre d'assurés et vice versa. La même procédure a été employée pour les sinistres liés à la sécheresse, pour lesquels une dérive de **9 %** avait été obtenue.

— Une fois les « nouveaux » sinistres simulés, le coût de ceux ci a été déterminé en utilisant la loi Gamma, avec pour moyenne la moyenne des sinistres observés dans la base de données initiale augmentée de la dérive coût estimée précédemment (Tableau 5.14), et pour écart-type celui des observations de la base de données. Le coût des « anciens » sinistres a également été revalorisé en fonction de la dérive de coût. La dérive coût étant un intervalle de confiance, les bornes supérieure et inférieure ont été utilisées afin d'obtenir un intervalle de confiance du taux de surprime à horizon 2028.

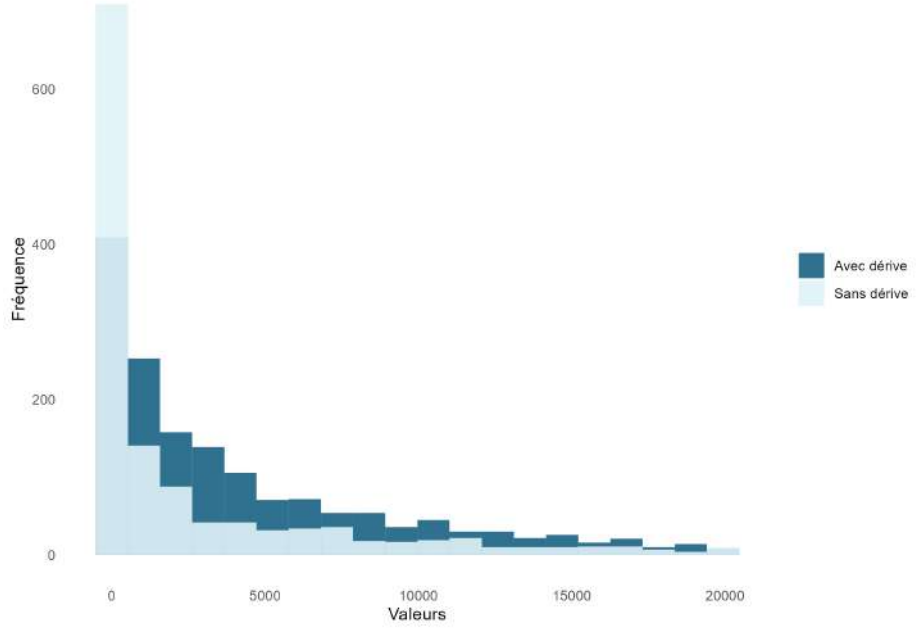


FIGURE 6.8 – Histogramme du coût des sinistres avec et sans dérive

Cela a permis d'estimer la charge de sinistralité future attendue entre **111,36 %** et **120,81 %** en **2028**, soit une augmentation de sinistralité entre **11,36 %** et **20,81 %** par rapport à la sinistralité actuelle. Cette charge de sinistralité future correspond à un taux de surprime moyen compris entre **16,90 %** et **17,74 %**.

Une estimation de la **progression annuelle** du taux de surprime nécessaire pour couvrir la sinistralité du portefeuille a été réalisée, permettant ainsi d'obtenir une prévision linéaire de l'évolution du taux de surprime au-delà de 2028 présentée dans le graphique

suivant :

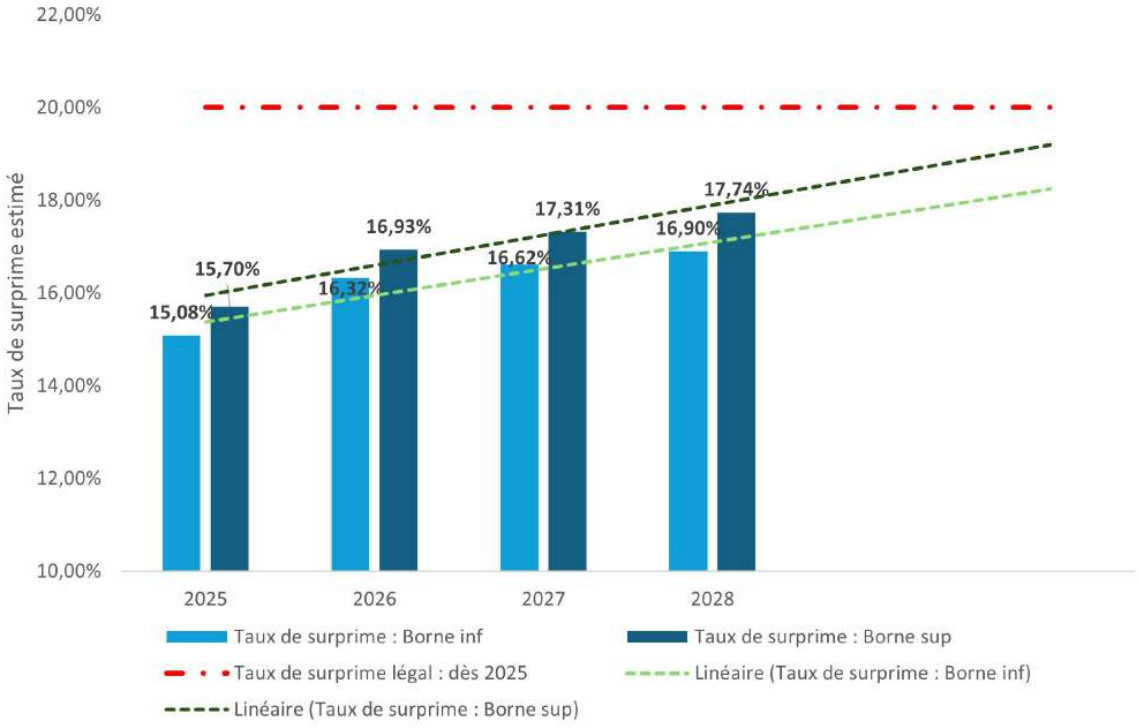


FIGURE 6.9 – Taux de surprime estimé entre 2025 et 2028

L'analyse de ce graphique selon les différents scénarios est comme suit :

— **Scénario de la borne inférieure de la dérive de coût** :
Si la dérive de coût se maintient au niveau de la borne inférieure, un taux de surprime de 20 % serait suffisant pour couvrir le risque lié à ce portefeuille jusqu'en 2028 et les cinq années suivantes. Les projections montrent qu'en 2028, le taux de surprime estimé pour la borne inférieure atteint 16,90 %, restant ainsi en deçà du seuil de 20 %. Si cette situation est similaire chez plusieurs assureurs, cela permettrait à la CCR de constituer des provisions d'égalisation pour compenser les années déficitaires.

— **Scénario de la borne supérieure de la dérive de coût** :
Pour la borne supérieure, une tendance similaire à celle observée pour la borne inférieure est constatée. Cependant, dans ce cas le taux de surprime permettrait de couvrir le risque lié au portefeuille jusqu'en 2031. Au-delà de cette date, le taux de surprime deviendrait insuffisant.

Analyse des résultats

Indépendamment du scénario de dérive de coût, une tendance à la hausse de l'estimation du taux de surprime est observée pour les années à venir. Comme l'indique la prévision linéaire, si les conditions actuelles persistent, ce taux de surprime estimé pourrait atteindre ou même dépasser le taux légal de 20 % à long terme (les 10 à 15 prochaines années).

Cette tendance à la hausse du taux de surprime est due à l'intégration des dérives de sinistralité pour les risques d'inondation et de sécheresse et pourrait donc se manifester chez d'autres assureurs MRH, avec des variations en fonction de l'exposition spécifique de chaque portefeuille.

Face à cette tendance, une question clé se pose : **est-il toujours pertinent de maintenir un taux de surprime fixe dans un contexte où la sinistralité est en augmentation continue** ?

1. Une option envisageable serait de **revaloriser annuellement le taux de surprime**, afin de l'ajuster en fonction de l'évolution de la sinistralité. Cependant, cette solution soulève deux enjeux majeurs :
   — **Jusqu'où pourrait s'élever le taux de surprime ?** :
   La prévision linéaire à l'horizon **2050** pour ce portefeuille montre un taux de surprime estimé entre **29,87 %** et **32,20 %,** ce qui représente une hausse significative par rapport aux niveaux actuels.

	Borne inférieure	Borne supérieure
Taux de surprime estimé à l'horizon 2050	**29,87 %**	**32,20 %**

TABLE 6.13 – Prévision linéaire du taux de surprime à l'horizon 2050

   — **Quelle est la capacité des assurés à supporter cette hausse ?** :
   Un tel accroissement du taux de surprime pourrait devenir financièrement difficile à assumer pour les assurés, remettant en question la viabilité de la couverture assurantielle dans un tel scénario.
2. Une **réduction des dépenses du régime Cat-Nat** pourrait également être envisagée. Cela pourrait se concrétiser par une diminution du nombre de risques couverts, ou encore par une rétrocession de ces risques à des réassureurs du secteur privé.
3. Une autre piste à envisager serait de **passer d'un taux de surprime mutualisé à un taux segmenté**, ajusté en fonction de l'exposition spécifique au risque du bien assuré. Comme le montre le tableau 6.12, le taux de surprime varie sensiblement en fonction de l'exposition. Un tel ajustement permettrait de mieux refléter la réalité du risque en fonction de la localisation du logement, offrant ainsi une tarification plus juste et équitable. Toutefois, cette approche pourrait égale-

ment soulever des défis importants, notamment le risque de rendre la couverture assurantielle financièrement inaccessible pour les zones les plus exposées.

6.4 Estimation de l'impact de la dérive de sinistralité liée aux tempêtes

La même méthodologie que celle utilisée précédemment pour prendre en compte la dérive de sinistralité liée à la fréquence et au coût des inondations et de la sécheresse a été appliquée pour évaluer l'impact de la dérive de sinistralité en termes de fréquence et de coût des sinistres tempête sur la prime pure payée par les assurés. La dérive de fréquence a été estimée par zone à partir des dérives précédemment obtenues, présentées dans le **tableau 5.12**. Cette dérive de fréquence a permis de simuler de nouveaux sinistres ou réduire le nombre de sinistres existants en fonction de la zone, tandis que la dérive de hausse de coût estimée entre **0 % et 9 %** a permis d'ajuster le coût des sinistres existants et de simuler le coût des nouveaux sinistres en utilisant une loi **Gamma**. Ces simulations ont permis de déterminer l'évolution de la prime pure pour les sinistres relevant de la garantie tempête, grêle et neige, dont les résultats sont présentés dans le tableau suivant :

Prime Pure actuelle	Prime avec dérive : borne inférieure	Prime avec dérive : borne supérieure
53,25	54,83	56,77
	2,97 %	**6,63 %**

TABLE 6.14 – Estimation de l'augmentation de la prime pure TGN du fait des dérives de sinistralité estimés à l'horizon 2028

Les scénarios de dérives, prenant en compte la borne inférieure et supérieure, prédisent une augmentation de la prime pure pour cette garantie comprise entre **2,97 %** et **6,63 %** à l'horizon 2028.

6.5 Impact global des événements climatiques sur le tarif à l'horizon 2028

Le taux de surprime du régime Cat-Nat étant réglementaire, si aucune revalorisation annuelle n'est faite, son augmentation à l'horizon 2028 par rapport à la situation en 2024 sera de **8 %** pour l'assurance MRH. À cela s'ajoute la hausse estimée de la garantie TGN, liée aux tempêtes, comprise entre 2,97 % et 6,63 %.

Le tableau suivant récapitule la hausse de prime estimée à l'horizon 2028, pour le risque lié aux événements climatiques :

	Borne inférieure	Borne supérieure
Hausse de prime estimée	**10,97 %**	**14,63 %**

TABLE 6.15 – Augmentation estimée de la prime pure à l'horizon 2028 en raison des événements climatiques

Par conséquent, l'augmentation totale de la prime pure payée par les assurés pourrait se situer entre **10,97 %** et **14,63 %** à l'horizon 2028, en raison des événements climatiques.

6.6 Critiques des modèles

Les limites identifiées pour les différents modèles ajustés sont :

— **Variables Explicatives de la Prime Pure des Catastrophes Naturelles** :
La prime pure des catastrophes naturelles a été déterminée en utilisant les variables suivantes : zonier, type d'habitation et nombre de pièces. Bien que ces variables fournissent une base utile pour l'évaluation des risques, d'autres facteurs importants, tels que l'année de construction, les matériaux de construction, et la distance par rapport aux sources d'eau (notamment pour les risques d'inondation), n'ont pas été intégrés en raison d'un manque d'information. Etant donné que les résultats dépendent fortement de la qualité des données utilisées, ils peuvent évoluer avec l'amélioration ou l'enrichissement des données disponibles.

— **Dérives de Fréquence et de Coût** :
Les dérives de fréquence et de coût utilisées pour simuler les sinistres futurs reposent sur des hypothèses qui peuvent ne pas se concrétiser comme prévu. Les facteurs de dérive peuvent évoluer de manière imprévisible en raison de facteurs externes tels que les changements climatiques, les avancées technologiques, et les mesures de prévention mises en place. Par conséquent, les prévisions basées sur ces dérives doivent être interprétées avec prudence.

— **Taux de Surprime Obtenu à Travers les Différents Scénarios** :
Les taux de surprime calculés sont spécifiques au portefeuille MRH étudié. Les résultats peuvent varier considérablement en fonction des caractéristiques spécifiques du portefeuille et de son exposition au risque.

Conclusion

Le réchauffement climatique, marqué par une augmentation de la fréquence et de la gravité des événements climatiques extrêmes, pose des défis considérables pour le secteur de l'assurance. Ce mémoire s'est concentré sur l'évaluation de l'impact de ces changements sur la sinistralité et la surprime du régime Cat-Nat, ainsi que sur la prime liée à la garantie TGN d'un portefeuille MRH.

Pour estimer la dérive de la fréquence de sinistralité liée aux inondations et à la sécheresse, qui représentent 93 % de la sinistralité du régime Cat-Nat, des séries temporelles ont été utilisées pour obtenir une prévision des paramètres météorologiques (température, humidité, précipitations, pression) à l'horizon 2028 à partir de la base SYNOP. Grâce à un modèle de classification basé sur une forêt aléatoire, il a été possible d'estimer le nombre de déclarations d'état de catastrophes naturelles entre 2024 et 2028. Cette approche a révélé une dérive de fréquence de sinistralité de **23 %** pour les inondations et de **9 %** pour la sécheresse, en comparant les sinistres futurs prévus avec ceux observés historiquement.

En projetant une augmentation de 23 % du nombre de sinistres liés aux inondations et de 9 % pour les sécheresses d'ici 2028, et une augmentation de coût entre 3,6 % et 7,9 % pour les inondations et 2,6 % et 4,7 % pour la sécheresse, les résultats indiquent que ces dérives entraîneraient un taux de surprime nécessaire pour couvrir ces risques, entre **16,90 %** et **17,74 %** en **2028**, avec une tendance annuelle à la hausse. Bien que ces chiffres soient spécifiques au portefeuille MRH étudié, ils reflètent une tendance générale à la hausse applicable à d'autres portefeuilles, avec des variations du taux de surprime selon l'exposition propre à chaque portefeuille.

Les résultats montrent que, sous ces hypothèses, le taux de surprime de **20 %** pourrait être suffisant à court ou moyen terme (entre 7 et 9 ans, par exemple, pour ce portefeuille). Dans les années où la sinistralité serait inférieure à la prime collectée, ces excédents pourraient être utilisés par la CCR pour alimenter les provisions d'égalisation. Cependant, à plus long terme (entre 10 et 15 ans), si la tendance à l'augmentation de la fréquence et de la sévérité des sinistres se poursuit, ce taux deviendrait insuffisant pour couvrir les risques liés à ce portefeuille, rendant nécessaire une revalorisation continue pour maintenir l'équilibre du régime.

En réponse aux résultats de cette étude, plusieurs recommandations ont été formulées, notamment l'idée de procéder à une revalorisation annuelle du taux de surprime, ajustée en fonction de l'évolution de la sinistralité. Cependant, cette approche présente des défis importants. Les projections linéaires montrent que le taux de surprime pourrait augmenter de manière significative, atteignant entre **29,87 %** et **32,20 %** d'ici **2050**. Une telle hausse, bien supérieure aux niveaux actuels, soulève des préoccupations quant à la capacité des assurés à absorber cette augmentation. En effet, une surprime trop élevée pourrait rendre la couverture assurantielle financièrement inaccessible pour de nombreux assurés, compromettant l'accès à l'assurance et menaçant l'équilibre financier du régime Cat-Nat. D'autres pistes ont également été suggérées, comme la réduction des dépenses du régime Cat-Nat à travers une révision des risques couverts ou la rétrocession de certains risques à des réassureurs privés, ou encore le passage d'un taux de surprime mutualisé à un taux segmenté en fonction de l'exposition du logement assuré.

Concernant les sinistres liés aux tempêtes, l'utilisation de la méthode des maxima par blocs combinée à des copules a permis de simuler la vitesse du vent à l'horizon 2028. Cette approche a révélé une dérive de fréquence de sinistralité de 5 %. Cette dérive entraînerait une augmentation de la prime de la garantie TGN comprise entre **2,97 %** et **6,63 %** d'ici 2028. Lorsqu'on combine cette hausse avec l'augmentation du taux de surprime Cat-Nat, l'étude estime que la prime pure totale payée par les assurés pourrait augmenter de **10,97 % à 14,63 %** d'ici **2028**.

Pour approfondir et enrichir ce mémoire, il serait pertinent d'améliorer les prévisions des paramètres météorologiques en intégrant des modèles plus sophistiqués, qui tiennent compte des facteurs exogènes et des dynamiques spatio-temporelles, cela permettrait d'affiner les estimations des dérives de fréquence. En ce qui concerne l'évaluation de la vulnérabilité du portefeuille, il serait intéressant de recourir à des modèles de simulation de sinistres, qui permettent de quantifier les dommages attendus pour différents types d'événements climatiques. Les fonctions de dommage, qui relient l'intensité d'un événement (comme la hauteur d'une inondation ou la vitesse du vent) aux dommages subis par les biens assurés, peuvent jouer un rôle crucial. Ces fonctions permettent d'évaluer l'impact potentiel sur des bâtiments en tenant compte de leurs caractéristiques spécifiques, telles que l'année de construction, la localisation, et la résistance structurelle. Les courbes AEP (Annual Exceedance Probability) et OEP (Occurring Exceedance Probability) sont également des outils essentiels pour quantifier le risque. Elles fournissent une estimation de la probabilité annuelle qu'un certain niveau de pertes soit dépassé (AEP) ou de la probabilité qu'un événement spécifique entraîne des pertes excédant un certain seuil (OEP). Ces courbes permettent à l'assureur de mieux comprendre la distribution des risques extrêmes, aidant ainsi à la gestion des réserves et à la prise de décision stratégique en matière de tarification et de couverture des risques. En utilisant ces outils, l'assureur peut non seulement évaluer plus précisément les risques associés à chaque type d'événement climatique, mais aussi optimiser la gestion du portefeuille, en s'assurant

que les primes sont ajustées en fonction de la vulnérabilité réelle des biens assurés. Cela contribuerait à une meilleure anticipation des sinistres potentiels et à une allocation plus efficace des ressources pour faire face aux défis posés par le changement climatique.

En conclusion, le réchauffement climatique, en augmentant la fréquence et la sévérité des événements extrêmes, impose au secteur de l'assurance des ajustements constants. Cette étude révèle la nécessité d'adapter régulièrement le régime Cat-Nat pour maintenir son équilibre financier. Des revalorisations de surprimes et des stratégies de segmentation tarifaire pourront ainsi mieux répartir les coûts induits par les catastrophes naturelles, assurant la pérennité de la couverture pour les assurés face aux risques croissants.

Annexes

Annexe A

Chapitre 4 : Dérive de sinistralité des inondations et de la sécheresse

A - Test de Ljung-Box

Le test de Ljung-Box est un test statistique conçu pour déterminer si un ensemble d'autocorrélations d'une série temporelle est significativement différent de zéro. Dans le cadre de séries temporelles, il est appliqué aux résidus d'un modèle ARIMA ajusté. L'hypothèse testée est que les résidus du modèle ARIMA n'ont pas d'autocorrélation [Wikipedia, 2011].

- **Hypothèse nulle** : Les données sont indépendamment distribuées : il n'y a pas auto-corrélation des erreurs.
- **Hypothèse alternative** : Les données ne sont pas indépendamment distribuées : il y a auto-corrélation des erreurs.

La statistique de test est la suivante :

$$Q = n(n+2)\sum_{k=1}^{h}\frac{\hat{\rho}_k^2}{n-k}$$

où n est la taille de l'échantillon, $\hat{\rho}_k$ est l'autocorrélation de l'échantillon au lag k, et h est le nombre de lags testés. Sous H_0, la statistique Q suit asymptotiquement une distribution $\chi^2_{(h)}$. Pour un niveau de confiance α, la région critique pour rejeter l'hypothèse de hasard est :

$$Q > \chi^2_{1-\alpha,h}$$

où $\chi^2_{1-\alpha,h}$ est le quantile $(1-\alpha)$ de la distribution khi-deux avec h degrés de liberté.

B - Test de Shapiro-wilk

Le test de Shapiro-Wilk teste l'hypothèse nulle stipulant qu'un échantillon $x_1, ..., x_n$ provient d'une population qui suit une distribution normale [Wikipedia, 2024].

— **Hypothèse nulle** : La population suit une distribution normale.
— **Hypothèse alternative** : La population ne suit pas une distribution normale

C - Résultat de la modélisation de la température sur les différentes zones

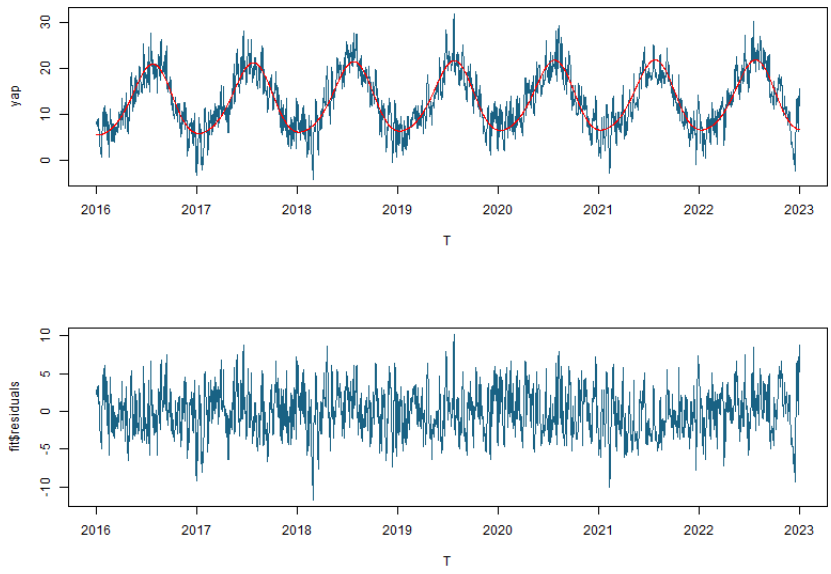


FIGURE A.1 – Zone 1

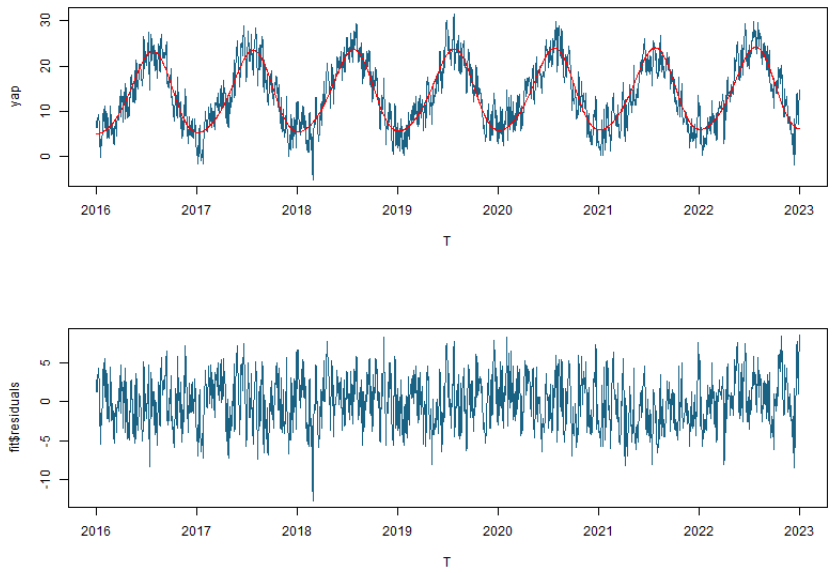


FIGURE A.2 – Zone 2

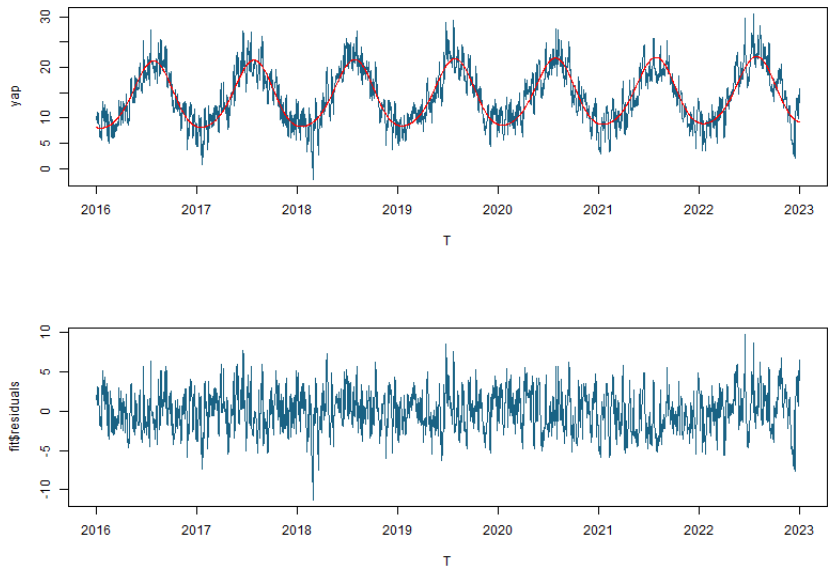


FIGURE A.3 – Zone 3

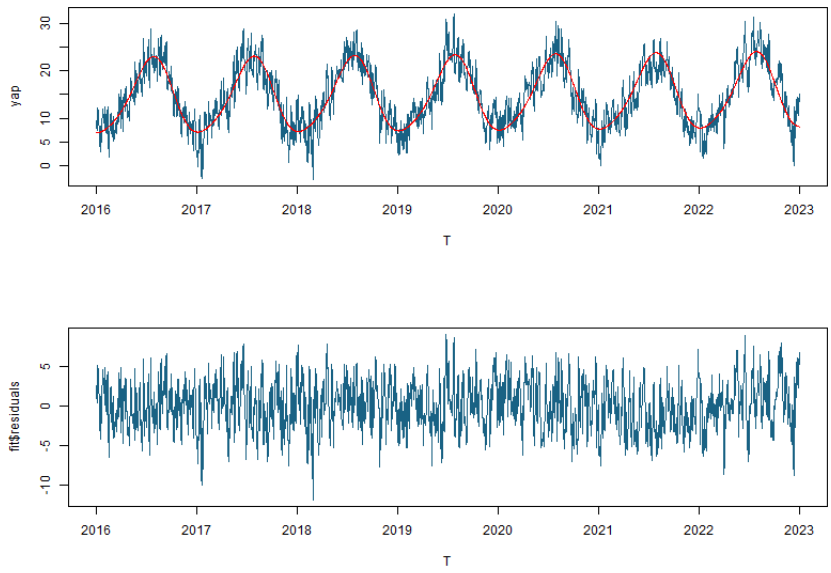


FIGURE A.4 – Zone 4

D - Métriques obtenues pour le modèle d'inondation

Retour mémoire

	Régression Logistique	Arbre de décision	Forêt aléatoire
Accuracy	95,60 %	96,30 %	99,41 %
Balanced accuracy	64,71 %	70,24 %	95,61 %
F1-score	97,71 %	98,07 %	99,69 %
Recall	99,10 %	99,27 %	99,84 %
Précision	96,30 %	96,91 %	99,54 %
AUC	92,69 %	81,14 %	98,67 %

TABLE A.1 – Comparaison des modèles de classification pour le risque inondation

Annexe B

Chapitre 5 : Evaluation de la dérive de sinistralité tempête

A - Mesures de dépendance

— **Coefficient de corrélation de Pearson** :
Ce coefficient est utilisé pour quantifier la relation entre deux variables quantitatives. Il est basé sur l'hypothèse que les deux échantillons suivent une distribution normale. Pour deux variables aléatoires réelles X et Y ayant des écart-types respectifs σ_X et σ_Y, la corrélation de Pearson se calcule comme suit :

$$\text{Coefficient de corrélation de Pearson} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

— **Corrélation de Spearman** :
Cette mesure évalue la dépendance statistique entre deux variables aléatoires qui n'ont pas nécessairement une relation linéaire. Le test se base sur les rangs des échantillons X et Y pour calculer la corrélation :

$$\text{Coefficient de corrélation de Spearman} = \frac{Cov(rg_X, rg_Y)}{\sigma_{rg_X} \sigma_{rg_Y}}$$

B - Validation du modèle GEV

Retour mémoire

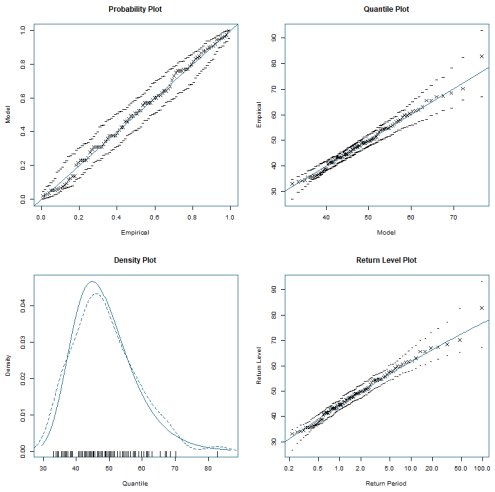


FIGURE B.1 – Zone 1

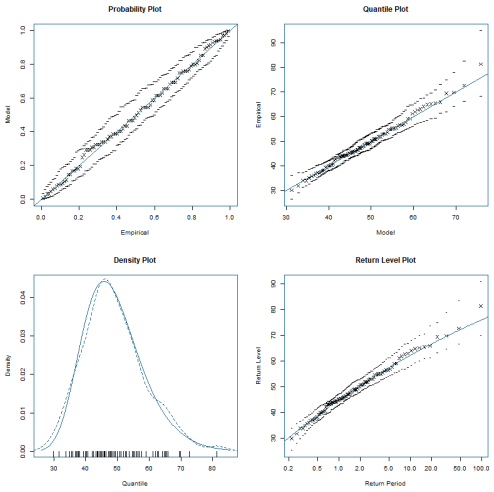


FIGURE B.2 – Zone 2

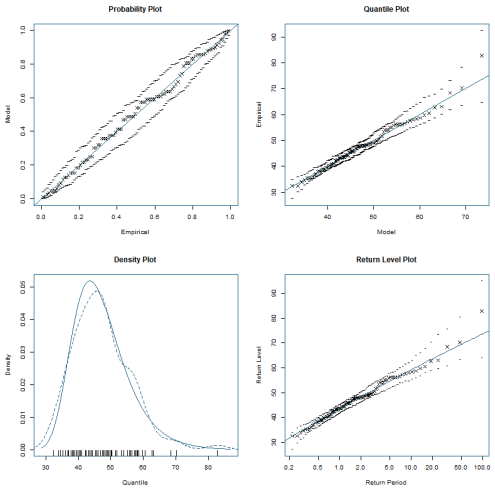


FIGURE B.3 – Zone 3

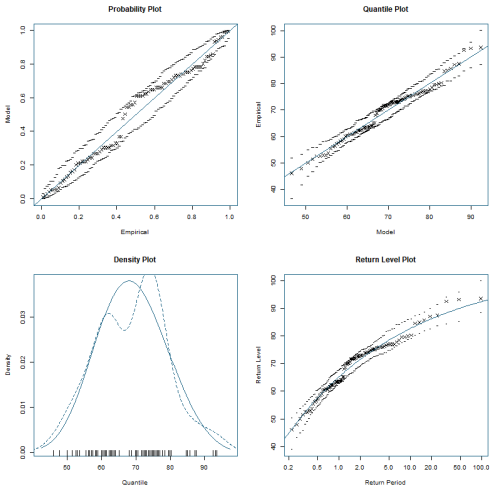


FIGURE B.4 – Zone 5

Annexe C

Chapitre 6 : Tarification

A - Zoniers inondation et séchéresse construits

Retour mémoire

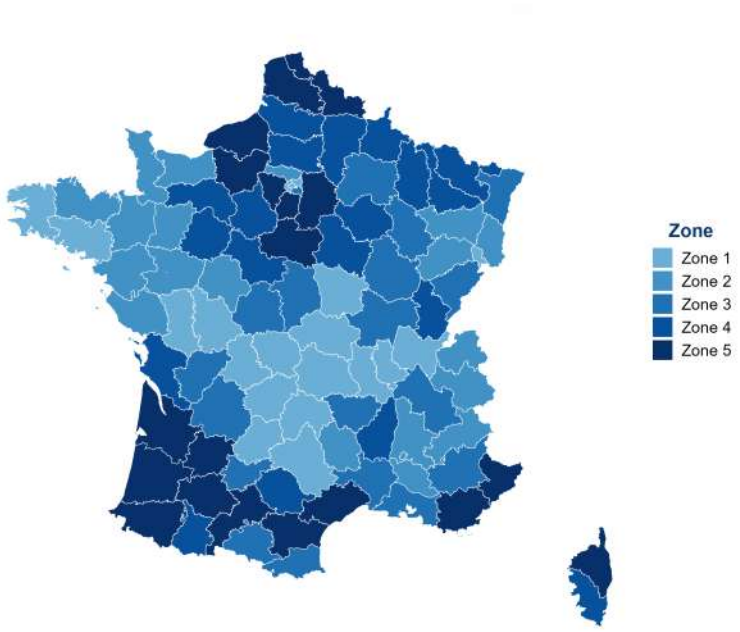


FIGURE C.1 – Zonier pour le risque d'inondation

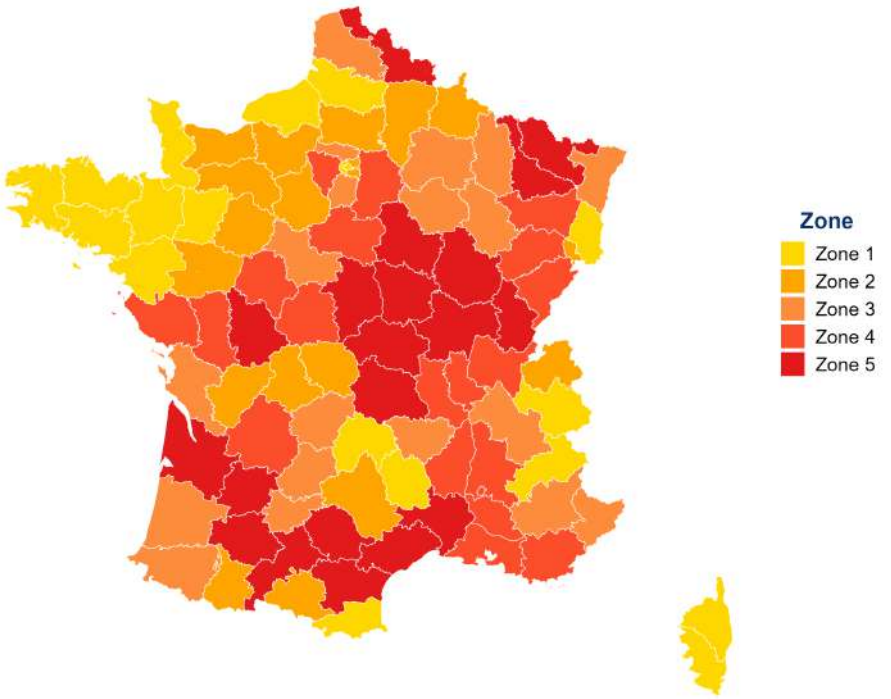


FIGURE C.2 – Zonier pour le risque de sécheresse

B - Lois testées pour les modèles de Fréquence et Coût

Fréquence

— **Loi de Poisson :**
Soit λ le nombre de moyen d'occurrences dans un intervalle de temps fixé

$$f(k) = \frac{\lambda^k}{k!} e^{-\lambda} \quad \forall k = 0, 1, \dots$$

— **Loi Binomiale Négative :**
Soient n le nombre de succès pour une série d'expériences indépendantes, avec p la probabilité de succès

$$f(k; n, p) = \binom{k+n-1}{k} p^n q^k \quad \forall k = 0, 1, \dots$$

où $\binom{k+n-1}{k}$ est un coefficient binomial.

Coût Moyen

— **Loi Gamma :**
Soient $k > 0$ paramètre de forme et $\theta > 0$ paramètre d'échelle.

$$f(x; k, \theta) = \frac{x^{k-1} e^{-x/\theta}}{\Gamma(k) \theta^k}$$

— **Loi Log-normale :**
Soient μ réel paramètre d'espérance et $\sigma > 0$ paramètre d'écart-type du logarithme de la variable.

$$f_Y(x; \mu, \sigma) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right)$$

— **Loi de Weibull :**
Soient $k > 0$ paramètre de forme et $\lambda > 0$ paramètre d'échelle.

$$f(x; k, \lambda) = \frac{k}{\lambda} \left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k}$$

C - Méthode de maximum de vraisemblance

Loi de Poisson

Soient, $x_1, x_2, \ldots, x_n$ un échantillon d'une loi de Poisson avec paramètre λ. La fonction de probabilité de la loi de Poisson est donnée par :

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}, \quad \text{où } x \in \{0, 1, 2, \ldots\}.$$

La fonction de vraisemblance pour un échantillon est :

$$L(\lambda) = \prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!}.$$

$$\log L(\lambda) = \sum_{i=1}^{n} \left(x_i \log \lambda - \lambda - \log x_i! \right).$$

Comme $\log x_i!$ ne dépend pas de λ, il peut être ignoré :

$$\log L(\lambda) = \left(\sum_{i=1}^{n} x_i \right) \log \lambda - n\lambda.$$

Pour maximiser la log-vraisemblance, la dérive par rapport à λ est réalisée et égalée à zéro :

$$\frac{d}{d\lambda} \log L(\lambda) = \frac{\sum_{i=1}^{n} x_i}{\lambda} - n = 0.$$

$$\hat{\lambda} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Ainsi, l'estimateur du maximum de vraisemblance pour le paramètre λ d'une loi de Poisson est la moyenne des observations.

Loi Gamma

Soient, $x_1, x_2, \ldots, x_n$ un échantillon d'une loi Gamma avec α le paramètre de forme et β le paramètre d'échelle. La fonction de densité de probabilité de la loi Gamma est donnée par :

$$f(x; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, \quad \text{où } x > 0.$$

$$L(\alpha, \beta) = \prod_{i=1}^{n} \frac{\beta^\alpha}{\Gamma(\alpha)} x_i^{\alpha-1} e^{-\beta x_i}.$$

$$\log L(\alpha, \beta) = n\alpha \log \beta - n \log \Gamma(\alpha) + (\alpha - 1) \sum_{i=1}^{n} \log x_i - \beta \sum_{i=1}^{n} x_i.$$

Pour maximiser la log-vraisemblance par rapport à β, la dérive est prise et égalée à zéro :

$$\frac{\partial \log L(\alpha, \beta)}{\partial \beta} = \frac{n\alpha}{\beta} - \sum_{i=1}^{n} x_i = 0,$$

ce qui donne l'estimateur $\hat{\beta}$:

$$\hat{\beta} = \frac{n\alpha}{\sum_{i=1}^{n} x_i}.$$

Pour obtenir α, la dérivée par rapport à α est utilisée :

$$\frac{\partial \log L(\alpha, \beta)}{\partial \alpha} = n \log \hat{\beta} - n\psi(\alpha) + \sum_{i=1}^{n} \log x_i,$$

où $\psi(\alpha)$ est la fonction digamma. Cette équation nécessite généralement une résolution numérique pour obtenir α.

D - Mise en AS-IF des coûts

Année	Trimestre 1	Trimestre 2	Trimestre 3	Trimestre 4
2016	929,5	931,2	935,9	942,0
2017	955,8	960,1	965,6	974,8
2018	981,8	988,1	987,5	988,2
2019	993,5	994,5	994,2	994,3
2020	995,1	995,2	996,8	1000,5
2021	1022,3	1033,4	1055,2	1066,4
2022	1101	1135,5	1142,8	1137

TABLE C.1 – Indice ICC-FFB entre 2016 et 2022

E - Validation croisée du modèle d'inondation

Retour mémoire

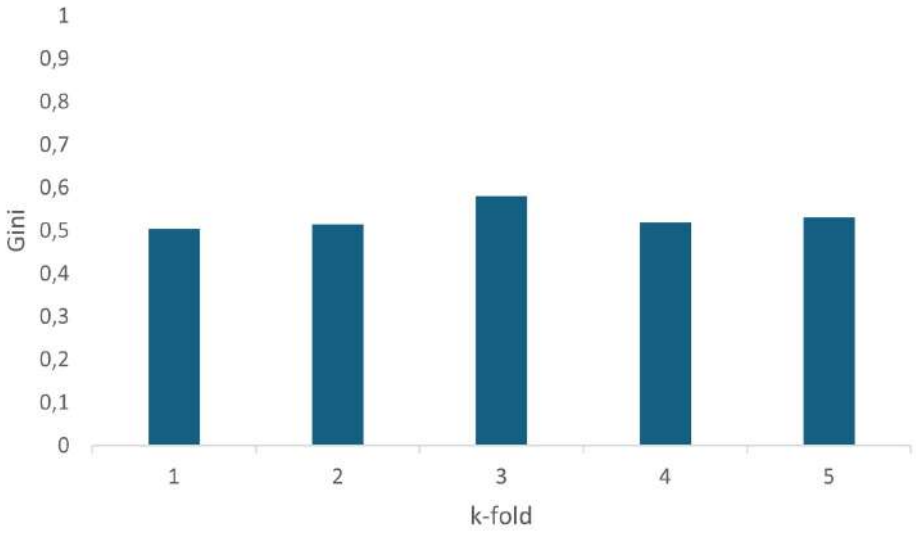


FIGURE C.3 – Validation croisée du modèle Tweedie : Inondation

Table des figures

1 Méthodologie d'obtention des dérives iv
2 Taux de surprime estimé entre 2025 et 2028 x
3 Methodology for Obtaining Drifts xiv
4 Estimated Surcharge Rate Between 2025 and 2028 xx

1.1 Répartition des catastrophes naturelles dans le monde en 2022 (Source : *Statista*) 6
1.2 Changement de la température à la surface du globe par rapport à 1850 - 1900 (Source : *Rapport du GIEC*) 7
1.3 Evolution du nombre de sinistres dus aux catastrophes naturelles dans le monde entre 1900 et 2022 (Source : *EM-DAT*) 8
1.4 Évolution du coût des sinistres dus aux catastrophes naturelles dans le monde de 1992 à 2022 (Source : *Swiss Re*) 8
1.5 Evolution de la fréquence des sinistres tous périls (Source : *Bilan CatNat CCR*) 9
1.6 Schéma d'indemnisation des catastrophes naturelles en France (Source : *CCR*) 10
1.7 Mécanisme d'indemenisation des catastrophes naturelles en France (Source : *SENAT*) 13
1.8 Répartition de la sinistralité par péril (Source : *Bilan CatNat CCR*) 14
1.9 Sinistralité des catastrophes naturelles non-auto entre 1982 et 2022 (Source : *Bilan CatNat CCR*) 14
1.10 Sinistralité observée tous périls confondus entre 1982 et 2022 (Source : *Bilan CatNat CCR*) 15
1.11 Evolution de la provision d'égalisation de la CCR et de la sinistralité depuis 2010 (Source : *SENAT*) 16
1.12 Augmentation du taux de surprime Cat-Nat (Source : *Légifrance*) 16

2.1 Echelle de Beaufort (Source : *Ouranos*) 20
2.2 Les 25 tempêtes majeures en métropole de 1980 à juin 2023 par indice de sévérité décroissant (Source : *Météo France*) 21
2.3 Historique de la fréquence et du coût moyen tempête depuis 1990 (Source : *France Assureurs*) 22

2.4 Cotisations perçues en 2022 au titre des événements naturels (Source : *France Assureurs*) . . . 23
2.5 Evolution de la sinistralité de la garantie TGN (Source : *France Assureurs*) 23

3.1 Agrégation des données SYNOP mensuelles pour l'obtention de la base finale . . . 31
3.2 Localisation des Stations Météorologiques en France Métropolitaine . . . 31
3.3 Répartition du nombre d'arrêtés de catastrophes naturelles par risque . . . 33
3.4 Evolution du nombre d'arrêtés du risque inondation entre 2016 et 2023 . . 34
3.5 Répartition des sinistres inondation par département entre 2016 et 2023 . 35
3.6 Evolution du nombre d'arrêtés du risque sécheresse entre 2016 et 2022 . . 36
3.7 Répartition des sinistres sécheresse par département entre 2016 et 2022 . . 36
3.8 Construction des bases pour la classification . . . 37

4.1 Graphique des individus de l'ACP . . . 41
4.2 Graphique des éboulis pour déterminer le nombre de dimensions optimal . 42
4.3 Zonier pour l'obtention de la dérive du régime Cat-Nat . . . 43
4.4 Températures observées au sein de la zone 5 . . . 46
4.5 Décomposition de la série temporelle initiale . . . 47
4.6 Tendance ajustée sur les températures de la zone 5 . . . 47
4.7 Modélisation de la tendance et la saisonnalité à l'aide d'un modèle paramétrique . . . 48
4.9 QQ-plot de la distribution des résidus du modèle ARMA(1,3) . . . 50
4.10 Validation du modèle sur les données observées de 2023 . . . 51
4.11 Prévision de la température future dans la zone 5 . . . 52
4.12 Corrélation entre les variables numériques . . . 58
4.13 Répartition de la variable cible *Sinistre* . . . 59
4.14 Courbe ROC Régression Logistique pour le risque de sécheresse . . . 60
4.15 Importance des variables de la forêt aléatoire . . . 64
4.16 Projection du nombre d'arrêtés futurs pour le risque de sécheresse . . . 65
4.17 Projection du nombre d'arrêtés futurs pour le risque d'inondation . . . 65

5.1 Graphique des individus de l'ACP . . . 68
5.2 Graphique des éboulis . . . 68
5.3 Zonier du risque tempête . . . 69
5.4 Boîte à moustache de la vitesse du vent par zone . . . 69
5.5 Distribution GEV (Source : *Google*) . . . 71
5.6 Maxima mensuels pour la zone 4 . . . 72
5.7 ECDF zone 4 . . . 72
5.8 Validation du modèle pour la zone 4 . . . 74
5.9 Représentation graphique des copules en deux dimensions (Source : *Altia*) 79
5.10 Nuage de points de la vitesse du vent par zones . . . 79
5.11 Comparaison des données réels et de la copule de Frank . . . 80
5.12 Relation entre la vitesse des rafales et la vitesse du vent . . . 83

6.1 Graphique de hill : garantie dégâts des eaux 97
6.2 Répartition des variables dans la base de données 98
6.3 Répartition des variables dans la base de données 99
6.4 Répartition de la sinistralité par garantie 100
6.5 Comparaison des coûts moyens 102
6.6 Comparaison des coûts moyens 103
6.7 Validation croisée du modèle Tweedie : Sécheresse 108
6.8 Histogramme du coût des sinistres avec et sans dérive 111
6.9 Taux de surprime estimé entre 2025 et 2028 112

A.1 Zone 1 124
A.2 Zone 2 125
A.3 Zone 3 125
A.4 Zone 4 126

B.1 Zone 1 128
B.2 Zone 2 128
B.3 Zone 3 128
B.4 Zone 5 128

C.1 Zonier pour le risque d'inondation 129
C.2 Zonier pour le risque de sécheresse 130
C.3 Validation croisée du modèle Tweedie : Inondation 134

Liste des tableaux

1 Evaluation de la qualité d'ajustement des composantes non-stationnaires de la température sur la base d'entraînement v
2 Modèles ARMA ajustés par zone pour la température v
3 Comparaison des températures moyennes observées et celle prédites entre 2024 et 2028 vi
4 Matrice de confusion pour le risque de sécheresse : base de test vi
5 Matrice de confusion pour le risque d'inondation : base de test vi
6 Paramètres de la loi Gumbel ajusté vii
7 Evaluation du modèle de régression viii
8 Hypothèse de dérive du coût des sinistres à horizon 2028 viii
9 Résultats du modèle Fréquence X Coût sur les garanties dommages sur la base de test ix
10 Proportion de la prime pure catastrophe naturelle en fonction de la prime pure dommages x
11 Estimation de l'augmentation de la prime pure TGN du fait des dérives de sinistralité estimées xi
12 Evaluation of the Fit Quality for Non-Stationary Components of Temperature : Train dataset xv
13 ARMA Models Fitted per Zone for Temperature xv
14 Comparison of Observed and Predicted Average Temperature between 2024 and 2028 xvi
15 Confusion Matrix for Drought Risk : Test dataset xvi
16 Confusion Matrix for Flood Risk : Test dataset xvi
17 Parameters of the Fitted Gumbel Distribution xvii
18 Evaluation of the Regression Model xviii
19 Assumed Cost Drift for Claims by 2028 xviii
20 Results of Frequency X Cost Model on Property Damage Coverages on the test dataset xix
21 Proportion of Natural Disaster Pure Premium Relative to Property Damage Pure Premium xix
22 Estimated Increase in Pure Premium for Storm, Hail, and Snow Coverage Due to Estimated Claims Drift xx

1.1 Franchises fixées par l'État (Source : *CCR*) 12

3.1 Pourcentage de valeurs manquantes par paramètre météorologique 32
3.2 Statistiques descriptives des paramètres météorologiques 32
3.3 Regroupement des catastrophes naturelles en famille de risque 33

4.1 Evaluation de la qualité d'ajustement des composantes non-stationnaires sur la base d'entraînement 48
4.2 Modèles ARMA ajustés par zone 49
4.3 Métriques pour la validation du modèle sur les données 2023 51
4.4 Comparaison de la température moyenne observée et celle prédite entre 2024 et 2028 52
4.5 Augmentation de température à horizon 2050 52
4.6 Comparaison de l'humidité moyenne observée et celle prédite entre 2024 et 2028 53
4.7 Comparaison de la pression moyenne observée et celle prédite entre 2024 et 2028 53
4.8 Comparaison de la précipitation moyenne observée et celle prédite entre 2024 et 2028 53
4.9 Corrélation de Pearson observée entre les paramètres météorologiques : **Données historiques** 54
4.10 Corrélation de Pearson observée entre les paramètres météorologiques : **Prévisions 2024 à 2028** 54
4.11 Exemples de GLM 55
4.12 Statistiques descriptives des paramètres météorologiques 57
4.13 Résultats des métriques de performance du modèle 59
4.14 Matrice de confusion de la régression logistique 60
4.15 Comparaison des métriques obtenues 63
4.16 Arbre de décision 63
4.17 Forêt aléatoire 63

5.1 Paramètres de la distribution GEV ajustée 73
5.2 Paramètres de la loi Gumbel ajustée et résultats de l'ANOVA 75
5.3 Tableau des corrélations entre les zones sur les données simulées 76
5.4 Tableau des corrélations entre les zones sur les données réelles 76
5.5 Formules des différentes copules 78
5.6 Choix de la copule 80
5.7 Vitesse de vent maximales simulées par zones pour 2024 81
5.8 Evaluation du modèle de régression 83
5.9 Vitesse de rafales simulées par zones pour 2024 84
5.10 Nombre de tempêtes annuelles sur l'historique 2016 à 2023 85
5.11 Nombre estimé de tempêtes de 2024 à 2028 85
5.12 Dérive de la fréquence de sinistralité 85
5.13 Hypothèses d'évolution du coût des sinistres à horizon 2050 86

5.14 Hypothèse de dérive du coût des sinistres à horizon 2028 86

6.1 Seuil des valeurs extrêmes des garanties dommages 98
6.2 AIC obtenu en fonction de la loi de fréquence ajustée 103
6.3 AIC obtenu en fonction de la loi de coût ajustée 104
6.4 Sélection de variables par méthode forward 104
6.5 Validation des modèles par test de déviance 105
6.6 Résultat du modèle Fréquence X Coût pour les dégâts des eaux 105
6.7 Résultat du modèle Fréquence X Coût sur les autres garanties dommages 105
6.8 Paramètres du modèle POT ajusté par garantie 106
6.9 Distributions usuelles retrouvées dans la famille Tweedie 107
6.10 Prime pure par zone pour les modèles de sécheresse et inondation 109
6.11 Proportion de la prime pure catastrophe naturelle en fonction de la prime pure dommages 109
6.12 Taux de surprime estimé par zone 110
6.13 Prévision linéaire du taux de surprime à l'horizon 2050 113
6.14 Estimation de l'augmentation de la prime pure TGN du fait des dérives de sinistralité estimés à l'horizon 2028 114
6.15 Augmentation estimée de la prime pure à l'horizon 2028 en raison des événements climatiques 115

A.1 Comparaison des modèles de classification pour le risque inondation . . . 126

C.1 Indice ICC-FFB entre 2016 et 2022 133

Bibliographie

[AILLIOT, 2024] AILLIOT, P. (2024). *Cours de séries temporelles*. EURIA.

[ARNAUD, 2016] ARNAUD, E. (2016). *Modélisation du risque sécheresse en France*. Institut des Actuaires.

[Assureurs, 2022] Assureurs, F. (2022). Impact du changement climatique sur l'assurance à horizon 2050.

[Assureurs, 2023] Assureurs, F. (2023). L'assurance habitation en 2022.

[BEDI, 2018] BEDI, N. (2018). *Modélisation du risque tempête*. Institut des Actuaires.

[Bolluze, 2024] Bolluze, L. (2024). Catastrophes naturelles : Définition et indemnisation.

[BOUCHOUCHI, 2024] BOUCHOUCHI, I. (2024). Défi climatique et durabilité, vers les limites de l'assurabilité ?

[BREHIN, 2021] BREHIN, F. (2021). *Risque de crue de la Seine sur le bassin parisien*. Institut des Actuaires.

[CCR, 2022] CCR (2022). Indemnisation des catastrophes naturelles en france.

[CCR, 2024a] CCR (2024a). Conséquences du changement climatique sur le coût des catastrophes naturelles en france à horizon 2050.

[CCR, 2024b] CCR (2024b). Garantie catastrophe naturelle.

[Climat, 2023] Climat, R. A. (2023). Synthèse du 6e rapport du giec : l'urgence climatique est là, les solutions aussi.

[data.gouv, 2024] data.gouv (2024). Base nationale de gestion assistée des procédures administratives relatives aux risques (gaspar).

[France, 2023] France, M. (2023). Caractérisation de la sévérité des tempêtes.

[France, 2024] France, M. (2024). Données synop essentielles omm.

[GAHBICHE, 2017] GAHBICHE, M. (2017). *Estimation de la prime pure catastrophes naturelles au travers de données géographiques*. Institut des Actuaires.

[GALL, 2023] GALL, A. L. (2023). Voici les cinq plus grosses tempêtes qui ont frappé la france en 30 ans, avant l'arrivée de ciaran.

[Géorisques, 2012] Géorisques (2012). Aléas de la base gaspar.

[ICHI.PRO, 2020] ICHI.PRO (2020). Démystifier les métriques d'évaluation de la classification : exactitude, précision, rappel, etc.

[InfoService, 2024] InfoService, A. (2024). Que faut-il savoir sur l'assurance multirisques habitation ?

[JMP, 2018] JMP (2018). Distance method formulas.

[LAILY, 2023] LAILY, R. (2023). *Cours : Application sur R en Tarification non-vie.* EURIA.

[L'ARGUS, 2022] L'ARGUS (2022). Garanties tempête, grêle, neige : vers une rentabilité nulle ?

[Lumivero, 2024] Lumivero (2024). Classification ascendante hiérarchique (cah).

[Maire, 2023] Maire, B. L. (2023). Publication des arrêtés renforçant les moyens d'action du régime d'indemnisation des catastrophes naturelles et du fonds de garantie des victimes.

[Manche.gouv, 2022] Manche.gouv (2022). Risque inondation : Définition.

[MAO, 2022] MAO, G. (2022). *Théorie des valeurs extrêmes.* ESILV.

[Marie-Christine BRASSIER, 2010] Marie-Christine BRASSIER, Gilles DEPOMMIER, P. S. (2010). Copules et dépendance entre les risques.

[MARTIN, 2021] MARTIN, O. (2021). Introduction à la régression linéaire.

[N.Jégou, 2020] N.Jégou (2020). *Analyse en Composantes Principales.* Université de Rennes.

[Numérique, 2013] Numérique, I. (2013). Méthodes.

[ods, 2024] ods, A. (2024). Qu'est-ce que l'open data ? guide pratique.

[ONU, 2023] ONU (2023). En quoi consistent les changements climatiques.

[PLANCHET, 2024] PLANCHET, F. (2024). *Modèles Financiers en Assurance et Analyses Dynamiques : Introduction à la théorie des copules.* ISFA.

[Pyrénées-Orientales.gouv, 2022] Pyrénées-Orientales.gouv (2022). Risque inondation : Définition.

[RIEDER, 2014] RIEDER, H. E. (2014). *Extreme Value Theory : A primer.* Lamont-Doherty Earth Observatory.

[SENAT, 2024] SENAT (2024). Régime d'indemnisation des catastrophes naturelles.

[TAMET, 2024] TAMET, E. (2024). Quelles sont les garanties d'une assurance habitation ?

[VERMET, 2022] VERMET, F. (2022). *Cours d'apprentissage statistique.* EURIA.

[VERMET, 2023] VERMET, F. (2023). *Cours d'arbres de décision et méthodes ensemblistes.* EURIA.

[Wikipedia, 2011] Wikipedia (2011). Ljung-box test.

[Wikipedia, 2024] Wikipedia (2024). Test de shapiro-wilk.

[Yue et co, 2016] Yue, H. et co (2016). Evt and its application to pricing reinsurance.