Mémoire présenté devant l'ENSAE Paris
pour l'obtention du diplôme de la filière Actuariat
et l'admission à l'Institut des Actuaires
le 05/03/2025

Par : **Sangeerthan KUNASEELAN**

Titre : **Modélisation des sinistres de grêle**

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

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

Membres présents du jury de la filière

Nom : Pierre PICARD

Membres présents du jury de l'Institut des Actuaires

Entreprise : AXA GIE

Signature :

Directeur du mémoire en entreprise :

Nom : Anthony NGO
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

Secrétariat:

Signature du candidat

Bibliothèque:

Ecole Nationale de la Statistique et de l'Administration Economique (ENSAE)
5, avenue Henry Le Chatelier - 91120 PALAISEAU, FRANCE

Résumé

Ce mémoire aborde la modélisation de la grêle, un péril naturel particulièrement préoccupant en raison des dommages conséquents qu'il peut occasionner, ainsi que de l'augmentation de sa fréquence. Les évènements récents, notamment la tempête de grêle QIARA en juin 2022, ont démontré l'ampleur des coûts associés, atteignant sur l'année 3 317 millions d'euros en France et dépassant largement le précédent record de 2014 évalué à 713 millions d'euros. Ce phénomène touche également d'autres pays européens, avec des grêlons exceptionnels observés en Italie en juillet 2023, atteignant un record européen de 19,46 cm de diamètre.

Dans un contexte de changement climatique et de l'absence de couverture de la grêle par le régime des catastrophes naturelles, il est crucial pour les assureurs et réassureurs de développer de nouveaux outils d'estimation de la sinistralité.

Ce mémoire, réalisé dans un contexte de réassurance, propose un modèle combinant données météorologiques et d'expositions pour identifier les variables majeures influençant les sinistres de grêle. Cette approche vise à fournir une estimation préliminaire des pertes rapidement après la connaissance d'un évènement de grêle susceptible d'impacter la réassurance d'une entité du groupe AXA, que ce soit à partir des actualités météorologiques, des premières déclarations de sinistres, ou du suivi en temps réel des évènements par exemple. Cela va ainsi faciliter la mise en place de stratégies adaptées et un provisionnement adéquat par les équipes en charge de la réassurance d'AXA.

Mots-clés : Grêle, Cartographie, Aléa, Vulnérabilité, Données météorologiques, Modèle Linéaire Généralisé, XG Boost, Échantillonnage

Abstract

This paper addresses the modeling of hail, a natural peril of particular concern due to the significant damages it can cause and its increasing frequency. Recent events, notably the QIARA hailstorm in June 2022, have demonstrated the magnitude of associated costs, reaching 3.317 million euros in France for the year and far exceeding the previous record of 713 million euros set in 2014. This phenomenon also affects other European countries, with exceptional hailstones observed in Italy in July 2023, reaching a European record of 19.46 centimeters in diameter.

In the context of climate change and the exclusion of hail from the natural disaster compensation scheme in France, it is crucial for insurers and reinsurers to develop new tools for estimating claims costs.

This paper proposes a model combining meteorological and exposure data to identify the major variables influencing claims related to hailstorms. This approach aims to provide a preliminary estimate of losses quickly after becoming aware of a hail event likely to impact the reinsurance of an AXA Group entity, whether based on weather news, initial claims reports, or real-time event monitoring, for example. This will facilitate the implementation of appropriate strategies and adequate provisioning by AXA's reinsurance teams.

Keywords : Hail, Mapping, Hazard, Vulnerability, Meteorological data, Generalized Linear Model, XG Boost, Sampling

Remerciements

Je tiens à remercier sincèrement mes tuteurs d'entreprise, Anthony NGO et Elie DADOUN pour votre encadrement, vos précieux conseils ainsi que pour votre confiance tout au long de ce stage au sein de AXA GIE.

Je remercie aussi Julien CARDINAL pour tes judicieuses recommandations sur les données de grêle mais aussi sur les meilleures boulangeries et pâtisseries de Paris.

Je remercie également l'ensemble des équipes d'AXA Group Ceded Re, et particulièrement l'équipe de Reserving pour votre accueil chaleureux, votre disponibilité, votre bienveillance et les belles discussions que nous avons pu avoir.

Je souhaite remercier l'ENSAE Paris et son corps professoral, en particulier mon tuteur académique Pierre PICARD, pour vos conseils.

Je réserve également des remerciements à mes amis, en particulier Neige, David, Noémie, Youssef, Salma, Sindy, Marine, Orianne et Pauline pour vos regards amusés lorsque je vous ai annoncé travailler 6 mois sur la grêle, ainsi que pour vos encouragements et votre soutien sans faille.

Enfin, je tiens à remercier ma famille et mes parents en particulier pour tous les sacrifices que vous avez faits jusqu'à ce jour pour me voir épanoui, ainsi que pour votre soutien précieux et inconditionnel.

Sangeerthan KUNASEELAN

Table des matières

Remerciements **4**

Note de Synthèse **9**

Executive Summary **13**

Introduction **17**

1 Présentation d'AXA **18**

2 Présentation du phénomène de grêle **20**

2.1 Formation de la grêle 20

2.1.1 Qu'est-ce que la grêle ? 20

2.1.2 Comment se forme la grêle ? 20

2.1.3 Quelles sont les précautions à prendre lors du traitement des données de grêle ? 21

2.2 La prise en compte de la grêle en assurance 22

2.3 Méthode actuelle d'estimation de la grêle 23

2.4 Etudes sur le sujet 24

2.4.1 Etude de F. VINET (1998) 24

2.4.2 Etude Keraunos 26

2.4.3 Etude de Pùčik et al. (2019) 28

3 Exploration des données **31**

3.1 Données d'exposition 31

3.1.1 Quelles LoB prendre en compte ? 31

3.1.2 Les données sont-elles pertinentes et de bonne qualité ? 32

3.1.3 Quel est le bon montant d'exposition à prendre en compte ? 34

3.1.4 À quel niveau géographique analyser les données ? 35

3.2 Données de sinistralité 35

3.2.1 Comment bien comprendre le montant de sinistre indiqué ? 35

3.2.2 Les données sont-elles pertinentes et de bonne qualité ? 37
3.2.3 Quel est le niveau géographique auquel réaliser l'étude ? 37
3.2.4 Comment lier les données d'exposition aux données de sinistres ? . 39
3.3 Données météorologiques 40
3.3.1 European Severe Weather Database (ESWD) 40
3.3.2 Meteomatics 44
3.3.3 Météo-France 46
3.3.4 GSOD 49
3.3.5 Comparaison des données 50
3.4 Aléa, vulnérabilité et risque 52
3.5 Données de la population 53

4 Cadre théorique 54
4.1 Mesures de risques 54
4.1.1 Les quantiles 54
4.1.2 Tail Value-at-Risk (TVaR) 55
4.2 Modèles linéaires généralisés 56
4.3 Méthodes de Machine Learning 60
4.3.1 Les arbres de décision 60
4.3.2 Gradient Boosting 62
4.3.3 Mesures d'erreurs 65

5 Modélisation et résultats 68
5.1 Agrégation des données Meteomatics 68
5.1.1 Sens de l'agrégation 70
5.1.2 Caractérisation des évènements de grêle 72
5.1.3 Fonctions retenues 76
5.2 Démarche pour la modélisation 83
5.3 Classification des sinistres 84
5.3.1 Identification des classes 84
5.3.2 Échantillonnage des évènements 88

5.3.3 Modélisation par Extreme Gradient Boosting (XG Boost) 95
5.4 Modélisation des taux de destruction 103
5.4.1 Modélisation appliquée aux évènements de classe 1 106
5.4.2 Modélisation appliquée aux évènements de classe 0 112
5.5 Modélisation croisée Classe de sinistre et Taux de destruction 114
5.6 Modélisation des classes 1 et 0 pour l'évènement 124

6 Limites et perspectives d'évolution du modèle 126
6.1 Problématique de la donnée disponible 126
6.2 Prise en compte changement climatique 126
6.3 Prise en compte d'autres sources de données 127

Conclusions 129

Bibliographie 137

Annexe 140

Note de Synthèse

Contexte de l'étude

Dans un contexte de changement climatique, il devient important de mieux appréhender les périls climatiques. Parmi eux, la grêle est un phénomène particulièrement intéressant dont les récents épisodes démontrent leur ampleur potentiel avec des montants de sinistres pouvant atteindre plusieurs milliards d'euros sur une année. Nous avons donc essayé de comprendre les conditions de formation de la grêle pendant les orages où des courants ascendants et descendants vont amplifier la taille des grêlons formés et ainsi, créer des grêlons dont les diamètres peuvent varier de quelques millimètres à plusieurs dizaines de cm.

Il semble ainsi essentiel de prendre en compte ces données pour assurer le risque de grêle, notamment dans le cadre de la garantie Tempête, Grêle, Neige (TGN) incluse dans les contrats d'assurance de biens tels que les multirisques habitation, entreprise, et automobile et qui ne nécessite pas la reconnaissance de l'état de catastrophe naturelle car ces périls sont considérés assurables et ainsi, non pris en compte dans le régime des Catastrophes Naturelles.

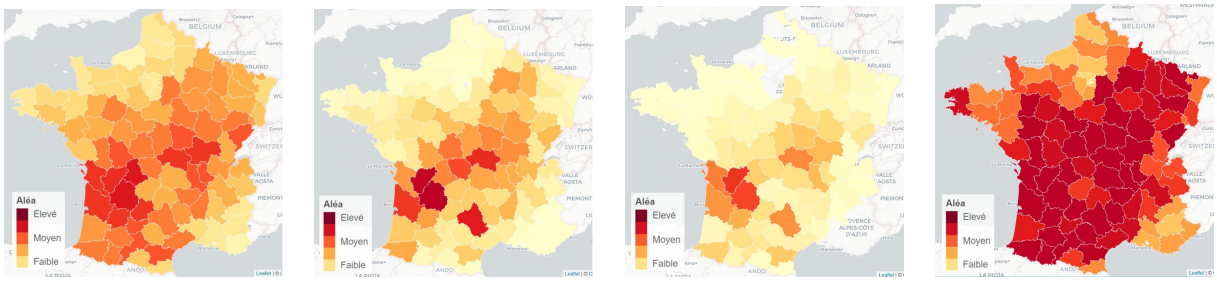
Les données disponibles

Nous disposons des données de sinistralité et d'exposition d'AXA, où les évènements de grêle sont définis sur une période spécifique selon le traité de réassurance. Cependant, ces données sont limitées car elles ne comprennent que les sinistres déclarés par les entités d'AXA lorsque le montant cumulé dépasse un seuil de déclaration précis. Nous identifions deux types d'évènements : ceux qui dépassent le seuil mais ne touchent pas la rétention (classe 0) et ceux qui dépassent la rétention et nécessitent une réassurance (classe 1). De plus, ces données seront agrégées à la maille du département.

L'observation des évènements de grêle présente des défis particuliers en raison de leur nature localisée et de leur courte durée. Nous décidons alors de croiser plusieurs sources de données :

- **Des données de terrain** : fournies par le laboratoire ESSL (*European Severe Storms Laboratory*)
- **Des données de stations météorologiques** : issues des stations de Météo-France et de GSOD (*Global Surface Summary of the Day*)
- **Données de Meteomatics** : une société de technologie météorologique qui nous fournit une modélisation de la taille des grêlons à une maille fine de 2 km par 2 km, avec une résolution temporelle horaire.

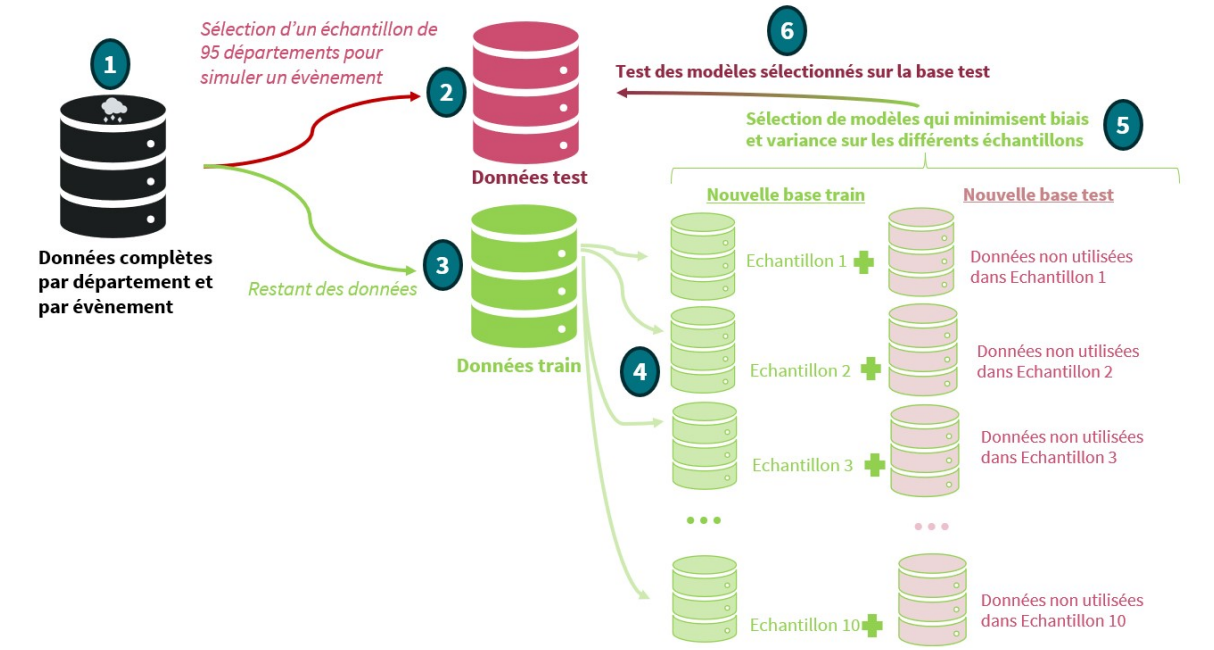
Une étape cruciale consiste à agréger ces données de Meteomatics pour les adapter à la maille du département et à la durée de l'évènement, afin de les aligner avec les données d'exposition et de sinistralité.



Graphique 1: Cartes de l'évènement QIARA de Juin 2022 en France selon différentes agrégations temporelles et géographiques des données Meteomatics

Modélisation

Pour compenser le déséquilibre des données due à la réassurance, nous avons développé un algorithme de ré-échantillonnage sous contrainte, permettant de générer de nouveaux évènements cohérents basés sur les caractéristiques des données observées. Notre base de données complète inclut les montants d'exposition, de sinistre et les variables climatiques agrégées par département et par évènement observé. Ensuite nous distinguons deux échantillons d'entraînement et de test comme indiqué dans le schéma 2



Graphique 2: Méthodologie d'entraînement et de test des données

En nous appuyant sur les échantillons obtenus par cette méthode, nous réalisons une modélisation en deux étapes :

1. Modélisation de la **classe de sinistralité**
2. Modélisation du **taux de destruction**

Modélisation de la classe de sinistralité

En ce qui concerne la première étape, nous classons les sinistres observés en **6 classes** allant de 0 à 5, la classe 0 indiquant une absence de sinistralité sur le département pour un évènement donné, 1 pour un sinistre qui ne coûte que quelques milliers d'euros jusqu'à 5 pour un sinistre à plusieurs dizaines de millions d'euros. Nous avons décidé de faire cette classification car nous considérons l'existence d'**effets de seuils** significatifs. Un évènement de grêle ayant touché un département à hauteur de plusieurs millions d'euros n'aura pas les mêmes caractéristiques qu'un évènement n'ayant coûté que quelques milliers d'euros. Par exemple, dans le cadre de contrat résidentiel, les classes les plus élevées laissent penser à une sinistralité ayant affectée la toiture mais aussi les biens contenus dans le logement, tandis qu'un sinistre de classe 1 peut laisser présager des dommages sur la vitre seulement.

Pour modéliser les classes de sinistres, après plusieurs comparaisons, notre choix s'est porté sur l'algorithme ***XG Boost***, s'appuyant sur une méthode de *boosting* dont l'approche séquentielle permet d'obtenir un modèle final corrigeant au mieux les erreurs à chaque étape. Étant donné le grand nombre de variables climatiques, environ 200, issues des agrégations des données de **Meteomatics**, nous décidons de sélectionner les 10 variables qui minimisent au mieux l'erreur de prédiction. Pour ce faire, nous effectuons une boucle sur ces variables et répétons le processus pour les évènements de classe 1 et 0. Les variables les plus pertinentes sont les agrégations basées sur les quantiles supérieurs à 80%, la durée maximale de l'épisode de grêle, ainsi que celles qui prennent en compte le nombre de grêlons de diamètres compris entre 2 et 4 cm et ceux de diamètre supérieur à 4 cm.

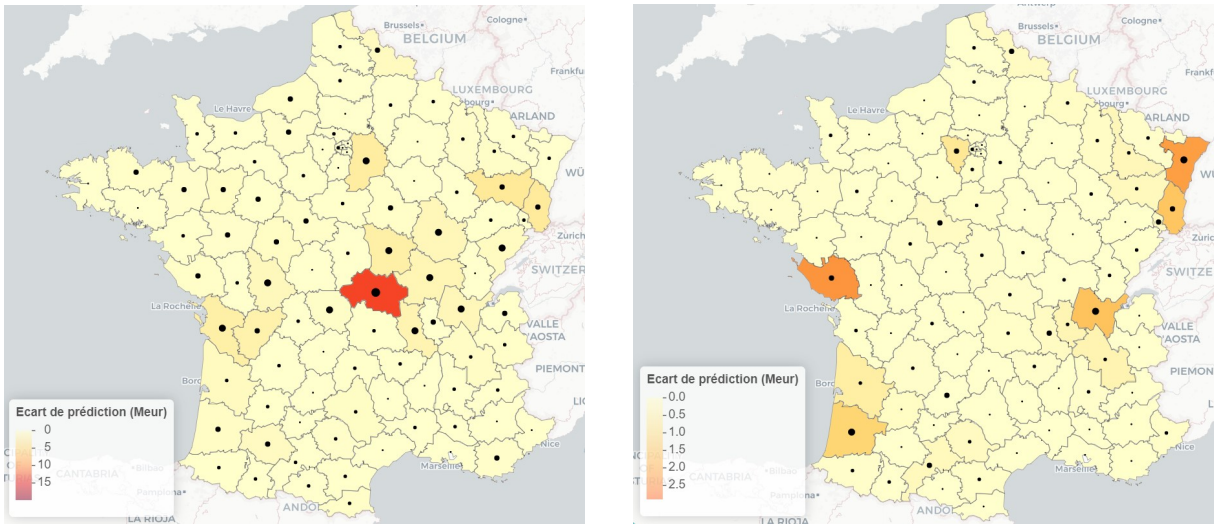
Modélisation du taux de destruction

Une fois ces premiers résultats obtenus, nous les combinons à nouveau aux variables climatiques agrégées pour déterminer le taux de destruction $= \frac{\text{Montant des sinistres}}{\text{Montant d'exposition total}}$ associé avec un modèle linéaire généralisé avec une famille de distribution gaussienne. Nous sélectionnons les 5 meilleures variables climatiques qui minimisent les erreurs de prédiction. Les premiers résultats étant peu satisfaisants, nous avons décidé de transformer par le logarithme ces taux de destruction, ce qui nous a donné de meilleurs résultats. Ensuite, nous appliquons le montant d'exposition du portefeuille d'AXA afin d'obtenir le montant du sinistre par département. Notre objectif est ainsi de concevoir un modèle qui permet à la fois d'estimer correctement la charge de sinistralité totale de l'évènement afin de savoir si l'évènement va toucher ou non la réassurance et dans les deux cas, estimer le montant de sinistre par département de manière la plus proche possible de ce qui va être réellement observé.

Résultats de la modélisation

La combinaison de ces modèles nous donne des résultats satisfaisants pour les évènements les plus importants, à savoir ceux qui ont dépassé la rétention de réassurance. Cette performance est illustrée par la carte de gauche de la figure 3. Nous y affichons les résul-

tats d'une combinaison de variables sélectionnés par les différents modèles présentés sur notre échantillon test. Nous observons que la plupart des écarts sont inférieurs à 2 millions d'euros par département avec cependant, un écart important pour un département dont la charge de plusieurs dizaines de millions d'euros est difficilement prévisible au million d'euros près. Sur la carte de droite, nous représentons un évènement test de classe 0, qui n'a donc pas touché la réassurance mais qui a été déclaré. L'échelle y est différente car l'ensemble des écarts sont limités à 3 millions d'euros avec la plupart des départements présentant un écart de moins d'un million d'écart.



Graphique 3: Carte des écarts de prédictions pour un évènement de classe 1 (gauche) et de classe 0 (droite)

Les données ont été modifiées mais tiennent compte du bon ordre de grandeur

Conclusion

Ce mémoire nous a permis de développer un modèle innovant en 2 étapes pour estimer les sinistres causés par la grêle sur le portefeuille résidentiel d'AXA à l'échelle départementale. Nous avons intégré diverses sources de données météorologiques et les avons enrichies avec des données démographiques et d'exposition. Ce modèle permet ainsi d'estimer rapidement la charge associée au risque de grêle et de déterminer si un évènement touchera la rétention de réassurance, facilitant ainsi l'élaboration de stratégies adaptées par les équipes de réassurance d'AXA. Des pistes d'amélioration incluent l'intégration de nouvelles données, comme les caractéristiques des toitures, le changement climatiques avec les analyses du GIEC ou l'extension à d'autres portefeuilles en prenant en compte le portefeuille automobile ou la surface agricole par exemple.

Executive Summary

Context of the study

In the context of climate change, it is becoming important to better understand climate hazards.Among them, hail is a particularly interesting phenomenon, with recent episodes demonstrating their potential magnitude, resulting in claim amounts that can reach several billion euros in a single year. We have therefore tried to understand the conditions of hail formation during thunderstorms, where upward and downward air currents amplify the size of hailstones, resulting in hailstones with diameters ranging from a few millimeters to several dozen centimeters.

It therefore seems essential to take this data into account when insuring hail risk, particularly within the framework of the Storm, Hail, Snow (TGN in french) guarantee included in property insurance contracts such as multi-risk home, business, and automobile insurance, which does not require the recognition of a state of natural disaster as these perils are considered insurable and thus not covered by the Natural Disaster regime.

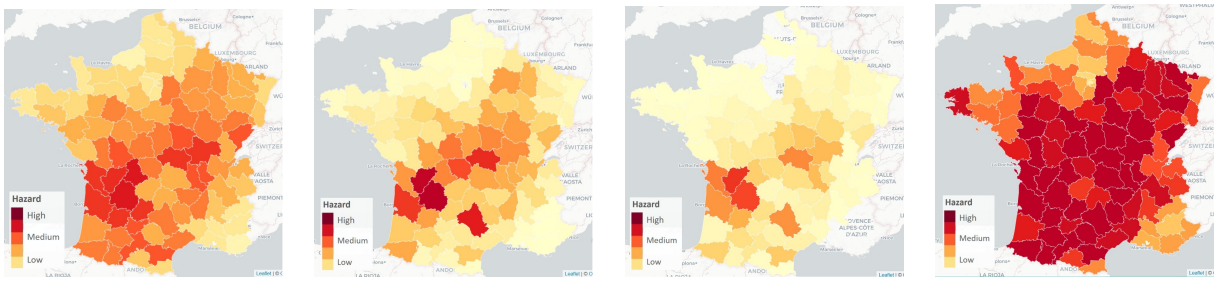
Available data

We have access to AXA's claims and exposure data, where hail events are defined over a specific period according to the reinsurance treaty. However, these data are limited as they only include claims reported by AXA entities when the cumulative amount exceeds a specific threshold. We identify two types of events: those that exceed the threshold but do not reach the retention level (class 0) and those that exceed the retention and require reinsurance (class 1). Furthermore, these data will be aggregated at the department level.

The observation of hail events presents particular challenges due to their localized nature and short duration. We therefore decided to combine multiple data sources:

- **Field data**: provided by the ESSL (*European Severe Storms Laboratory*)
- **Meteorological station data**: sourced from Météo-France stations and GSOD (*Global Surface Summary of the Day*)
- **Meteomatics data**: a meteorological technology company that provides us with hailstone size modeling at a fine 2 km by 2 km grid, with hourly temporal resolution.

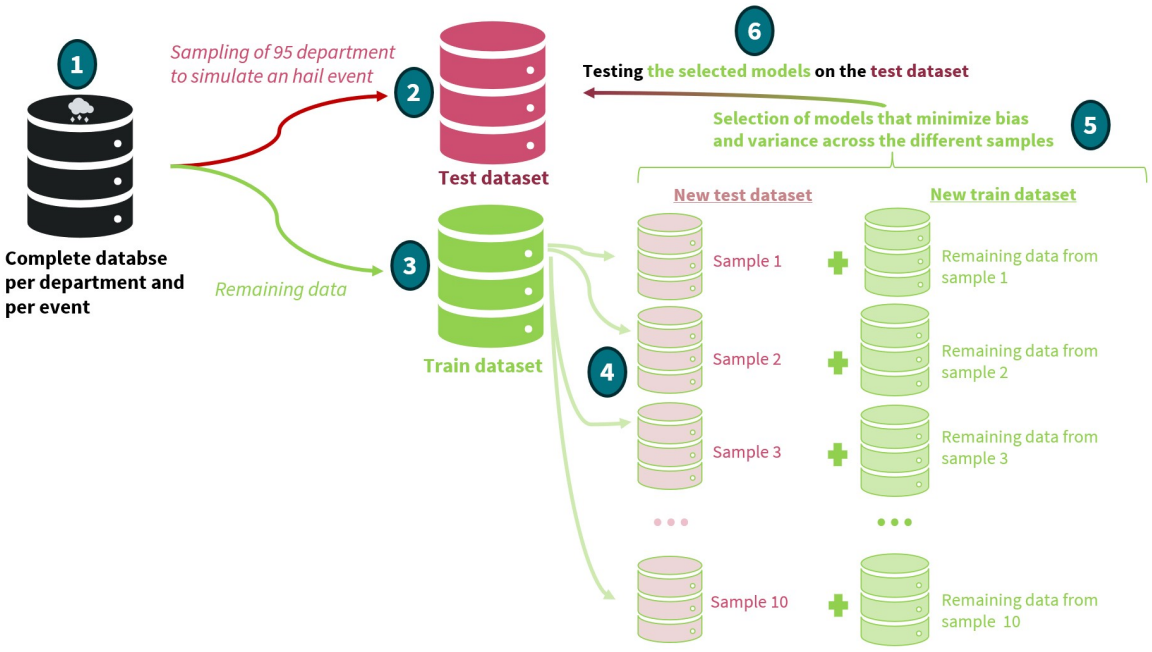
A crucial step involves aggregating this Meteomatics data to adapt it to the department level and event duration, in order to align it with exposure and claims data.



Graphique 4: Maps of the QIARA event in France (June 2022) based on different aggregations of Meteomatics data

Modeling

To compensate for the data imbalance due to reinsurance, we have developed a constrained resampling algorithm, allowing the generation of new coherent events based on the characteristics of the observed data. Our complete database includes exposure amounts, claims, and climate variables aggregated by department and observed event. We then distinguish two samples for training and testing as shown in the diagram 5.



Graphique 5: Methodology for Training and Testing Data

Using the samples obtained through this method, we perform a two-step modeling process:

1. Modeling of the **claims class**
2. Modeling of the **destruction rate**

Modeling the claims severity class

In the first step, we classify observed claims into **6 categories** ranging from 0 to 5. Class 0 indicates the absence of claims in a department for a given event, class 1 corresponds to a claim costing only a few thousand euros, while class 5 represents a claim amounting to several tens of millions of euros.We decided to implement this classification because we consider the existence of significant **threshold effects**. A hail event that has impacted a department with costs in the millions of euros will not have the same characteristics as an event that has only cost a few thousand euros. For example, in the context of residential contracts, the higher classes suggest claims affecting the roof as well as the contents within the dwelling, whereas a class 1 claim could imply damage limited to windows only.

To model the claims classes, after several comparisons, we chose the **XG Boost** algorithm, which relies on a *boosting* method. Its sequential approach allows the final model to correct errors as effectively as possible at each step. Given the large number of climate variables, approximately 200, derived from the aggregation of **Meteomatics** data, we decided to select the 10 variables that minimize prediction error. To achieve this, we perform a loop over these variables and repeat the process for class 1 and class 0 events. The most relevant variables are those based on aggregates with quantiles above 80%, the maximum duration of the hail episode, as well as those that account for the number of hailstones with diameters between 2 and 4 cm and those with diameters greater than 4 cm.

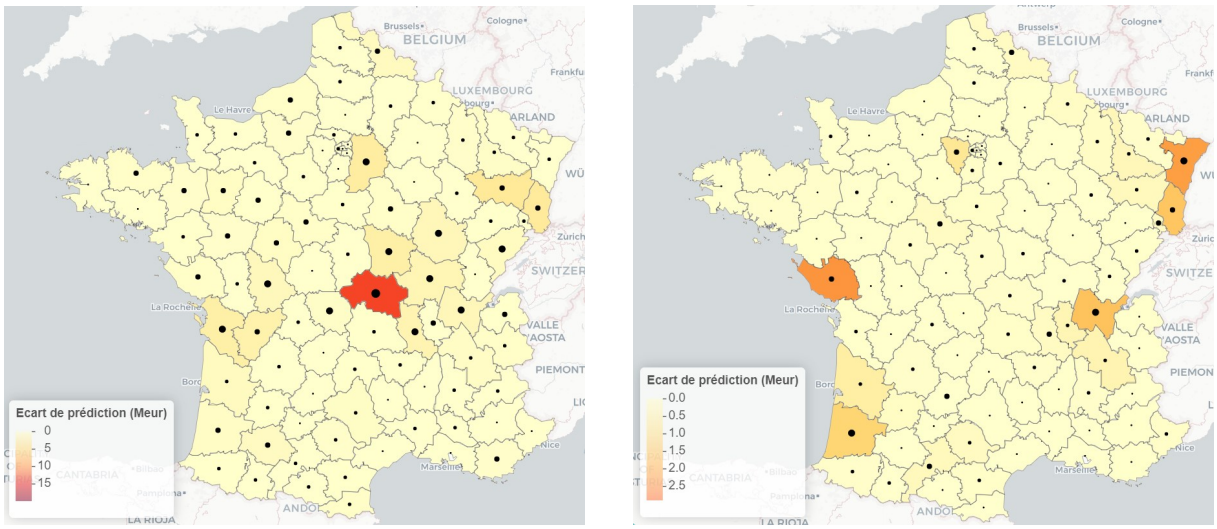
Modeling the destruction rate

Once these initial results are obtained, we combine them again with the aggregated climate variables to determine the destruction rate. This destruction rate is defined as $\text{destruction rate} = \frac{\text{Claim Amount}}{\text{Total Exposure Amount}}$ using a generalized linear model with a Gaussian family. We select the 5 best climate variables that minimize prediction errors.As the initial results were not satisfactory, we decided to apply a logarithmic transformation to these destruction rates, which yielded better results. We then applied AXA's portfolio exposure amounts to get the claim amount per department. Our objective is thus to build a model that accurately estimates the total claims amount of an event to determine whether it will impact reinsurance, and to estimate the claim amount per department as closely as possible to what will be observed.

Modeling Results

The combination of these models gives us satisfactory results for the most significant events, namely those that have exceeded the reinsurance retention, as shown by the left map in Figure 6. Here, we display the results of a combination of variables selected by the various models applied to our test sample. We observe that most discrepancies are below 2 million euros per department; however, there is a significant discrepancy for one department where the total amount of several tens of millions of euros is difficult to predict to the nearest million euros. In the right map, we represent a test event classified as class 0, which did not impact reinsurance but was declared nonetheless. The scale is

different here, as all discrepancies are limited to 3 million euros, with most departments showing a discrepancy of less than one million euros.



Graphique 6: Map of prediction discrepancies for a class 1 event (left) and a class 0 event (right)

The data has been modified but maintains the correct order of magnitude

Conclusion

This study has allowed us to develop an innovative two-step model to estimate the hail losses on AXA's residential portfolio at the departmental level. We integrated various meteorological data sources and enriched them with demographic and exposure data.This model enables a rapid estimation of the cost associated with hail risk and determines whether an event will impact reinsurance retention, thus facilitating the development of appropriate strategies by AXA's reinsurance teams. Potential improvements of the model include the integration of new data, such as roof characteristics, climate change analyses from the IPCC, or extending the model to other portfolios by considering the automobile portfolio or agricultural land, for example.

Introduction

10 milliards d'euros, il s'agit du coût estimé des aléas climatiques sur l'année 2022 par France Assureurs. A l'ère du changement climatique, la Fédération estime à 143 milliards d'euros le coût lié aux aléas climatiques à horizon 2050. Nous notons ainsi des conséquences majeures pour les assureurs et les réassureurs qui doivent prendre en compte une montée en fréquence et en intensité de ces évènements dans leurs modèles. Parmi les 10 milliards d'euros de coûts totaux, 3,3 milliards d'euros sont attribués aux dommages causés par les évènements de grêle sur les bâtiments, couverts par la garantie TGN, tandis que 1,6 milliard d'euros concernent les dommages sur les véhicules[1] alors que l'évènement extrême de grêle de référence jusque-là était la tempête ELA de 2014 avec un coût de 850 Millions d'euros sur le secteur automobile et du bâtiment [2].

Contrairement à d'autres aléas climatiques comme la tempête ou les inondations, la grêle ne se détecte que quelques heures avant l'évènement[3]. La mesure de son intensité est complexe due à son caractère très localisé et à la variabilité de ses impacts. Cette variabilité peut concerner par exemple la taille des grêlons, leur vitesse de chute ou encore l'orientation de leur impact. De plus, les appareils de relevé de grêle sont coûteux, contrairement aux stations de mesure de pluie qui sont plus abordables et plus répandues par exemple. Ce phénomène affecte également différents secteurs avec notamment l'immobilier, l'automobile et l'agriculture. La grêle est ainsi un défi significatif pour les actuaires afin d'évaluer et prévoir les coûts associés.

Les dommages causés par la grêle ne sont pas couverts pour le régime des catastrophes naturels (CatNat) en France comme peuvent l'être les inondations ou la sécheresse. Les sinistres, notamment dans le cadre des bâtiments relèvent davantage de la garantie Tempête Grêle Neige (TGN) des contrats multirisques habitation (MRH). Ainsi les assureurs de ce secteur ne peuvent bénéficier de la prise en charge des sinistres au-delà d'un certain seuil par la Caisse Centrale de Réassurance et cela pousse les assureurs à améliorer leurs modèles pour estimer ces sinistres et établir une tarification cohérente.

Il s'agit d'un enjeu essentiel notamment dans le cadre de Solvabilité II, qui établit les exigences de capital pour les assureurs en Europe. Les assureurs doivent posséder suffisamment de fonds propres pour faire face à des évènements climatiques extrêmes dont la grêle.

[1] L'assurance des évènements naturels en 2022
[2] Rapport de la grêle de 2017 du MRN
[3] Risques climatiques : la grêle, l'aléa extrême (argusdelassurance.com)

Ce mémoire a ainsi pour objectif de mettre en place un modèle grêle permettant d'établir une estimation de la charge de ces évènements. Dans le cadre de notre étude, nous allons nous focaliser sur le cas de la France ainsi que de l'Italie sévèrement touchés en 2022 et 2023 notamment.

Pour ce faire, nous commencerons par une présentation succincte du groupe AXA ainsi que l'équipe au sein de laquelle ce mémoire a été effectué pour expliquer les implications de cette étude pour son activité. Ensuite, nous aborderons les bases de données utilisées, qui comprennent à la fois des données internes, telles que la sinistralité et l'exposition d'AXA, mais aussi des données externes, incluant des informations météorologiques et géographiques. Nous présenterons également les outils théoriques qui nous seront utiles dans le cadre de notre modélisation du risque de grêle. Enfin, nous expliciterons la démarche adoptée pour arriver à notre modèle.

1 Présentation d'AXA

AXA est un groupe international français spécialisé dans l'assurance et la réassurance et qui opère à l'international. AXA SA en est la holding et gère au 31 décembre 2023 un total d'actifs de 644 milliards d'euros [7]. Le groupe AXA présente à cette même date un chiffre d'affaires de 103 milliards d'euros pour un résultat net de 7,2 milliards d'euros et un ratio Solvabilité II confortable de 227%.

Le Groupe AXA a été pionnier dans la création d'un véhicule de réassurance interne, AXA global Re (abrégé AGRe par la suite) il y a plus de 20 ans, devenant ainsi le premier groupe d'assurance à mettre en place une telle structure. Suite à une fusion d'AXA global P&C (AGPC) et d'AXA global Life (AGL), AGRe s'occupe de leurs activités respectives étant respectivement la réassurance non-vie et la réassurance vie

La mission d'AGRe consiste à contribuer à la stabilité globale du Groupe AXA à l'échelle mondiale, à soutenir le développement de ses entités en leur permettant de bénéficier de la diversification du capital, optimiser les programmes de réassurance et à garantir que l'appétit au risque est respecté. AXA global Re regroupe des fonctions opérationnelles et de supports. Les fonctions opérationnelles de réassurance incluent la relation client, la souscription, la gestion des sinistres et l'actuariat tandis que les fonctions de support comprennent la gestion des risques, les finances, les données de réassurance et les systèmes informatiques, les ressources humaines et la communication, la logistique, ainsi que les

affaires juridiques et la conformité.[5]

AGRe comptait 170 collaborateurs pour un chiffre d'affaires de 2,3 milliards d'euros et réassurait 60 entités à l'échelle mondiale à date du 30 juin 2022, date importante à laquelle AXA finalisé la transformation de sa holding AXA SA en réassureur interne du Groupe AXA à travers sa fusion avec AXA global Re après avoir obtenu l'agrément de l'Autorité de Contrôle Prudentiel et de Résolution (ACPR) pour exercer une activité de réassurance.[6]

Aujourd'hui, la direction qui centralise les activités de réassurance est AXA Group Ceded Re (AGCRe). Tout comme AGRe, cette direction permet 3 objectifs majeurs :

- **Le transfert de risque** : AGCRe peut décider de transférer une partie de ses risques issus des traités souscrits par les entités directement à d'autres organismes de réassurance (*back to back reinsurance*).
- **La mutualisation du risque** : AGCRe selon son niveau en capital et son appétit au risque peut également décider de retenir une partie du risque et de les mettre en commun au sein d'un *pool* (mise en commun des capacités du groupe pour couvrir certains risques) dont une partie pourra être redistribuée aux entités, une autre gardée par le groupe et une dernière partie couverte par des traités de réassurance.
- **L'optimisation du besoin en capital** : AGCRe peut élaborer des traités de réassurance spécifique pour chaque entité selon les besoins en capital qui peuvent varier selon les risques couverts et les zones d'exposition et peut répondre à un besoin en diversification des entités exposées à des risques peu diversifiées comme les catastrophes naturelles

Ce mémoire a été effectué dans cette direction, au sein de l'équipe de provisionnement, ce qui implique des spécificités particulières lors de l'analyse des données de grêle remontées par les différentes entités du Groupe AXA dans un contexte de réassurance.

2 Présentation du phénomène de grêle

2.1 Formation de la grêle

2.1.1 Qu'est-ce que la grêle ?

La grêle est phénomène météorologique caractérisé par des **précipitations solides** constituées de morceaux de glace, appelés grêlons qui se forment lors d'**orages violents**. Le diamètre de ces grêlons peut ainsi varier de 0,5 mm jusqu'à une vingtaine de cm dans les cas les plus extrêmes comme ce qui a été observé en juillet 2010 aux États-Unis. Dans le cas où le diamètre de ces morceaux de glace sont inférieurs à 0,5 mm, nous parlons alors de grésil.

2.1.2 Comment se forme la grêle ?

La grêle se forme dans les nuages lors des épisodes de grêle, plus particulièrement au sein des cumulonimbus qui sont des nuages larges de 5 à 15 km et qui peuvent s'élever jusqu'à 15 km en altitude. Ces cumulonimbus sont caractérisés par de très forts **courants ascendants et descendants** entre la base qui est chaude et humide et le sommet du nuage, à la limite de la stratosphère qui est très froid.

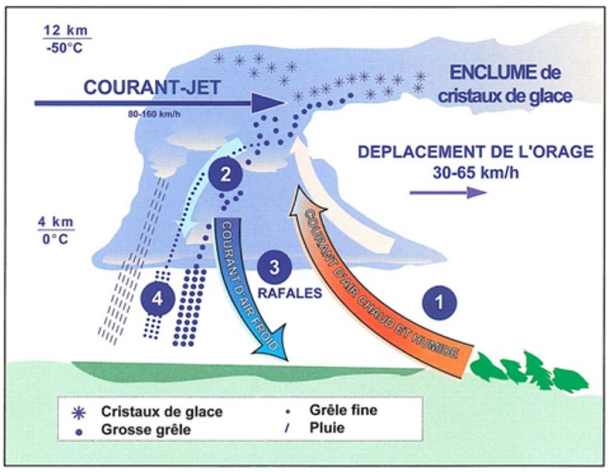
Lorsque de l'air chargé d'humidité s'élève rapidement dans ces nuages jusqu'à atteindre une altitude où la température devient inférieure à 0°C, l'air se refroidit et se condense en gouttelettes qui vont former de la glace autour de noyaux **glaçogènes**. Ces derniers sont issus des particules fines en suspension dans l'air provenant d'éléments naturels comme les cendres volcaniques ou alors de l'activité humaine comme les rejets des réacteurs d'avion. Une fois cette glace formée autour du noyau glaçogène, la particule plus dense va tomber.[22]

Dans le cas où les courants ascendants sont faibles, ces particules vont rapidement être éjectés du nuage et du fait de sa petite dimension, elle va fondre avant de toucher le sol donnant lieu à de la pluie. Dans le cas échéant, les courants ascendants vont faire monter les particules au sein du cumulonimbus. Lors de son ascension, les particules vont collecter de l'eau liquide présente dans le nuage et se **cumuler** à la glace déjà présente. Ces dépôts liquides vont à leur tour former une nouvelle couche de glace et ce cycle continue tant que le courant ascendant est suffisamment fort pour refaire monter ces particules.

Plus le cumulonimbus est violent, c'est-à-dire qu'il possède de forts courants ascen-

dants, plus la particule va rester longtemps au sein du nuage et aura ainsi de nombreux **dépôts successifs** autour des noyaux glaçogènes. C'est donc ce procédé qui donne naissance aux grêlons et qui explique ses couches de glaces concentriques. Il peut aussi arriver qu'au sein du nuage, les grêlons fusionnent entre eux et cela contribue à créer des particules d'un diamètre plus important. Nous utilisons alors d'**accrétion**.

Les averses de grêle sont généralement de courte durée et affectent des zones géographiques relativement restreintes. Toutefois, dans certaines conditions, plusieurs averses peuvent se produire simultanément au sein d'une même région. Par exemple, sur une superficie de 100 km^2, ces évènements peuvent engendrer la chute d'environ **300 milliards de grêlons** en l'espace d'une dizaine de minutes pour une masse totale d'environ **50 000 tonnes** d'après Météo-France.



Graphique 7: Schéma de la formation de la grêle
Source : ANELFA

2.1.3 Quelles sont les précautions à prendre lors du traitement des données de grêle ?

La grêle et les orages à l'origine de ces dernières sont des phénomènes atmosphériques complexes à modéliser à cause de plusieurs facteurs. Tout d'abord par le fait que la grêle soit un phénomène de courte durée, cela diminue la quantité d'information à analyser et nécessite une très bonne précision des instruments de mesure. Ensuite, il s'agit d'un phénomène à la fois local, mais qui peut également se produire dans des zones géographiques proches ou non. Ainsi, une partie d'une commune peut être touchée tandis qu'une zone située à quelques kilomètres restera intacte. Ce caractère local rend également plus complexe l'observation des données. En cas de grêles dans des zones montagneuses ou peu peuplées, nous aurons moins de remontées de données dans ces zones-là. De plus,

l'intérêt scientifique donné à ces études est relativement récent, les données et la littérature sur le sujet s'en retrouve réduites. L'ensemble de ces données viennent ainsi réduire la fiabilité des données.

2.2 La prise en compte de la grêle en assurance

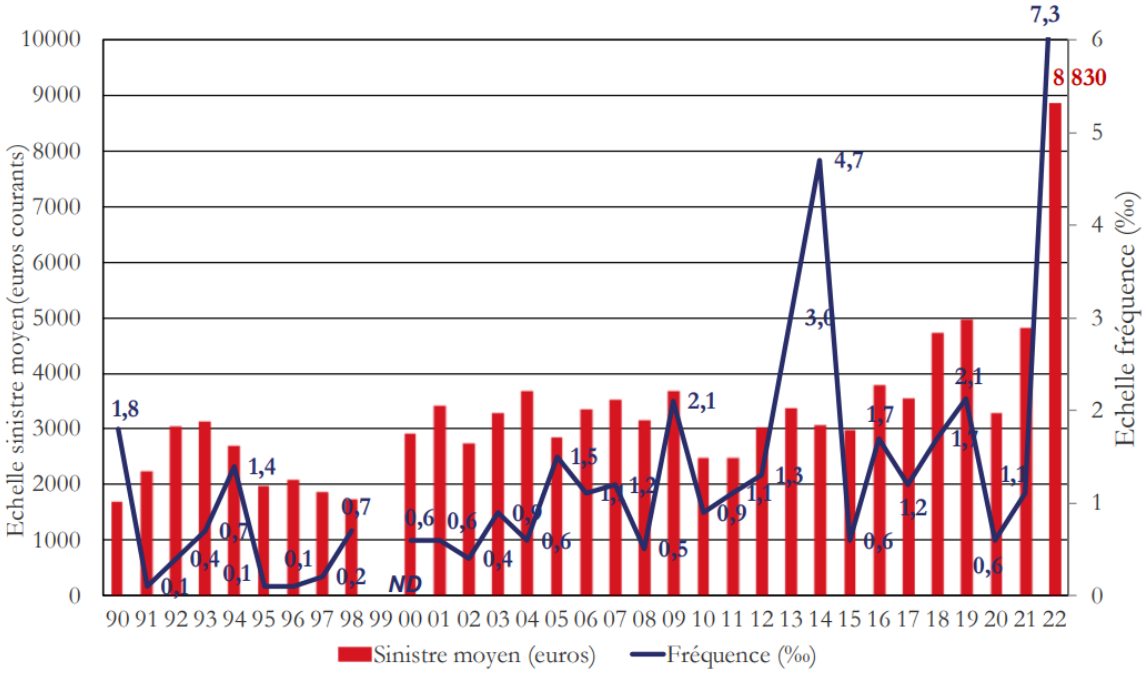
Les dommages causés par les tempêtes, la grêle et la neige sur des biens assurables tels que les habitations et les véhicules ne sont pas couverts par le régime des catastrophes naturelles défini par les articles L.125-1 et suivants du Code des Assurances qui ne s'applique qu'aux dommages matériels considéré comme non assurables et ayant eu pour «cause déterminante l'intensité anormale d'un agent naturel ». Or, les tempêtes, les grêles et les épisodes de neige ne sont pas considérés comme ayant une intensité anormale au sens de l'article L.125-1. Ces phénomènes sont donc considérés assurables et les dommages dues à ces aléas sont ainsi couverts dans le cadre de la garantie tempêtes, Grêle, Neige (dit **TGN** par la suite) d'après l'article L122-7 du Code des Assurances.

Cette garantie est incluse dans les contrats d'assurance de biens tels que les **multirisques habitation, multirisques entreprise, et multirisques automobile** et permet aux assurés d'être indemnisés en cas de dommages matériels causés par les effets de tempêtes (chute d'arbre par exemple), de neige (poids de la neige qui peut s'accumuler sur les toits et le déformer, voir le faire s'effondrer) et de grêle (impact sur la toiture, ou les vitres par exemple) sur les immeubles et les véhicules assurés en France. Elle prend également en charge les effets de l'infiltration d'eau causés par ces évènements climatiques.

Contrairement au régime de catastrophe naturelle, la garantie TGN ne nécessite pas une reconnaissance préalable de l'état de catastrophe naturelle pour être mise en œuvre. Les assureurs indemnisent directement les assurés «ayant souscrit un contrat d'assurance garantissant les dommages d'incendie ou tout autre dommage à des biens situés en France, ainsi que les dommages aux corps de véhicules terrestres à moteur »[10] pour ces types de dommages. Les personnes assurées victimes de dégâts doivent déclarer leur sinistre à leur assureur dans les meilleurs délais. De plus, en attendant l'intervention des experts, les assurés doivent prendre toutes les mesures conservatoires nécessaires, telles que la mise en place de bâches ou la couverture provisoire des biens endommagés pour éviter que les dégâts n'empirent.

La sinistralité ayant été particulièrement forte sur la grêle en 2022, nous allons reprendre quelques chiffres de cette année-là fournis par France Assureurs [16]. L'étude nous indique que les cotisations liées à la garantie TGN sur le bâtiment est évalué à **1,86**

milliards d'euros, ce qui représente 8,7% de l'ensemble des cotisations des assurances de dommages aux biens (en dehors des assurances automobiles et récoltes). Les cotisations de cette garantie se ventilent de la manière suivante : 70% pour les assurés couverts par le multirisque habitation, aussi appelés les particuliers, 19% par le multirisque professionnel, appelés les entreprises et enfin 11% pour les dommages aux biens agricoles. Contrairement au taux de surprime pour la couverture des catastrophes naturelles, qui est déterminé par les pouvoirs publics, le prix de la garantie TGN est fixé librement par les entreprises d'assurance.



Graphique 8: Historique de la fréquence et du coût moyen grêle depuis 1990
Source : France Assureurs (2022)

La figure 8 réalisée par France Assureurs en 2022 semble montrer une augmentation de la fréquence ainsi que de la sévérité des évènements de grêle en France. Alors que les sinistres moyens vacillaient aux alentours de 4 500€ jusqu'en 2021, l'année 2022 a atteint un record avec un coût moyen de 8 830€.

2.3 Méthode actuelle d'estimation de la grêle

Au sein du département *AXA Group Ceded Reinsurance*, l'estimation de la charge ultime pour les sinistres liés à la grêle se fait principalement à partir des données remontées par les différentes entités d'AXA dans le cadre d'un traité de réassurance. Nous réalisons ensuite des analyses de sinistralité en étudiant le coût moyen et la fréquence des sinistres,

puis nous comparons ces résultats aux données historiques. Lors de cette analyse, nous regroupons les sinistres par classes selon leur montant, afin d'identifier ceux qui présentent des comportements similaires. Cette approche nous permet de mieux comprendre la distribution des sinistres et d'affiner notre estimation en tenant compte des spécificités de chaque classe de sinistres. Ces informations sont complétées par des données externes fournies par des courtiers ainsi que des données météorologiques pouvant laisser présager des événements de forte intensité. Des hypothèses sur l'inflation sont également prises en compte. L'ensemble de ces éléments permet d'établir une estimation de la charge ultime.

Cependant, cette méthode présente plusieurs inconvénients. Tout d'abord, elle repose essentiellement sur les données des entités, ce qui peut entraîner un délai important entre la date d'occurrence de l'événement et la réception des premières données de sinistralité. De plus, les premières données reçues ne sont souvent pas fiables, avec de nombreux dossiers ouverts comportant des montants de sinistres forfaitaires inadaptés à l'événement par exemple.

L'objectif du modèle proposé dans ce mémoire est de permettre une estimation en *live cat*, c'est-à-dire de fournir une première estimation de la charge ultime à partir de diverses données météorologiques et d'exposition en temps réel, sans avoir à attendre les premières remontées d'information des entités. Cette approche permettra également de comparer les résultats de notre modèles aux analyses des entités. En anticipant les sinistres, le département sera en mesure de charger correctement les réserves et de mieux piloter le risque dès la connaissance de la survenance d'un évènement de grêle.

2.4 Etudes sur le sujet

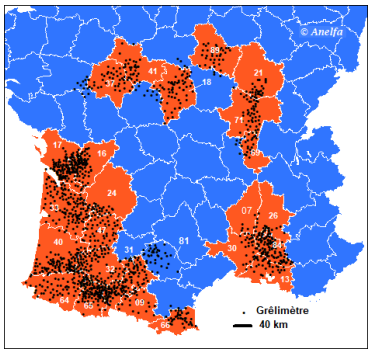
2.4.1 Etude de F. VINET (1998)

Le docteur Freddy VINET est professeur de géographie à l'université Paul Valéry à Montpellier et est responsable du master en gestion des risques de catastrophe. Il a participé à de nombreuses études dont le tsunami de 2004 en Indonésie, la tempête Xinthia en France en 2010 et la tempête Irma dans les Antilles en 2017. Dans le cadre de notre étude, nous allons nous intéresser à sa thèse soutenue en décembre 1998 sur le risque de grêle en France. Il s'appuie notamment sur des données de terrain fournies par l'ANELFA (Association Nationale d'Etude et de Lutte contre les Fléaux Atmosphériques).

L'ANELFA est une association régie par la loi de 1901, créée en 1951 par un groupe composé notamment d'agriculteurs, d'agronomes et de physiciens qui a deux objectifs majeurs [4] :

- Développer les recherches scientifiques dans le domaine de la physique des nuages et de la modification du temps
- Perfectionner une méthode de traitement des orages afin de réduire les dégâts causés par la grêle.

L'ANELFA a mis en place des grêlimètres à partir de 1988 notamment dans le Sud-Ouest de la France pour mesurer l'impact des grêlons. Ces grêlimètres sont des plaques en polystyrène qui se déforment sous la chute des grêles et qui en garde la forme. Ainsi, en étudiant les déformations de ces plaques, l'ANELFA est en mesure d'établir le diamètre des grêlons, le nombre de grêlons ainsi que leur énergie cinétique afin de caractériser leur intensité.

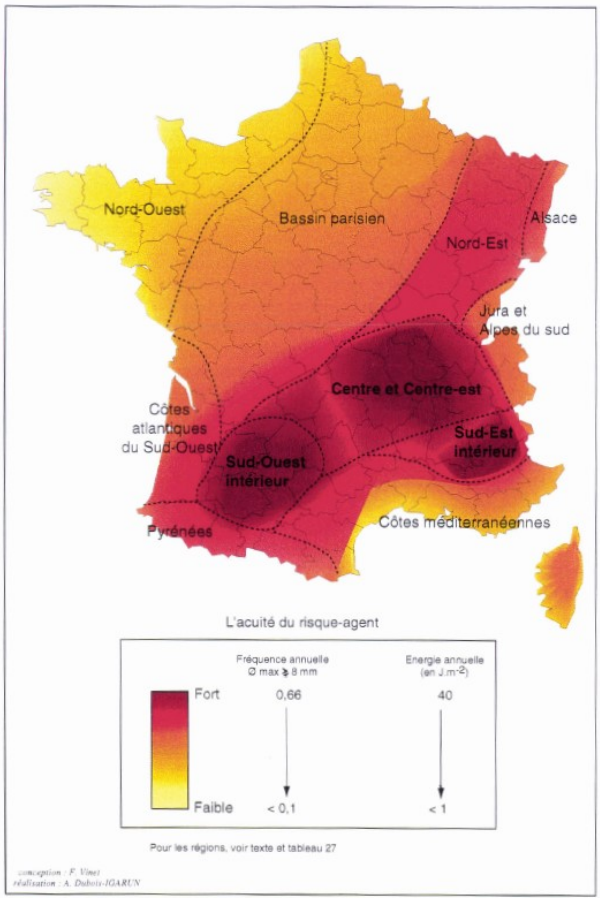


Graphique 9: Carte des grêlimètres en France
Source : ANELFA

Graphique 10: Grêlimètre
Source : ANELFA

Graphique 11: Plaque en polystyrène
Source : ANELFA

Grâce aux données de l'ANELFA ainsi que l'analyse des particularités topographiques des régions en France, VINET établit la carte 12 des zones en France les plus à risque d'être affectées par un évènement de grêle sur le territoire français. Cette carte montre un risque particulièrement important dans le Sud-Ouest, le Centre-Est ainsi que dans le Sud-Est de la France.



Graphique 12: Géographie du risque-agent grêle en France
Source : F. VINET (1998)

Nous allons par la suite nous appuyer sur cette idée de cartographie de la distribution des risques en France mais aussi en Italie pour mieux appréhender et mieux modéliser le risque de grêle.

2.4.2 Etude Keraunos

Keraunos est un observatoire français des tornades et des orages violents qui tire son origine du mot grec ancien signifiant foudre. Cet observatoire gère un site internet accessible à tous et sans frais permettant d'accéder à de nombreuses informations sur les phénomènes climatiques [1] ainsi qu'à des articles dont celui qui nous intéresse.

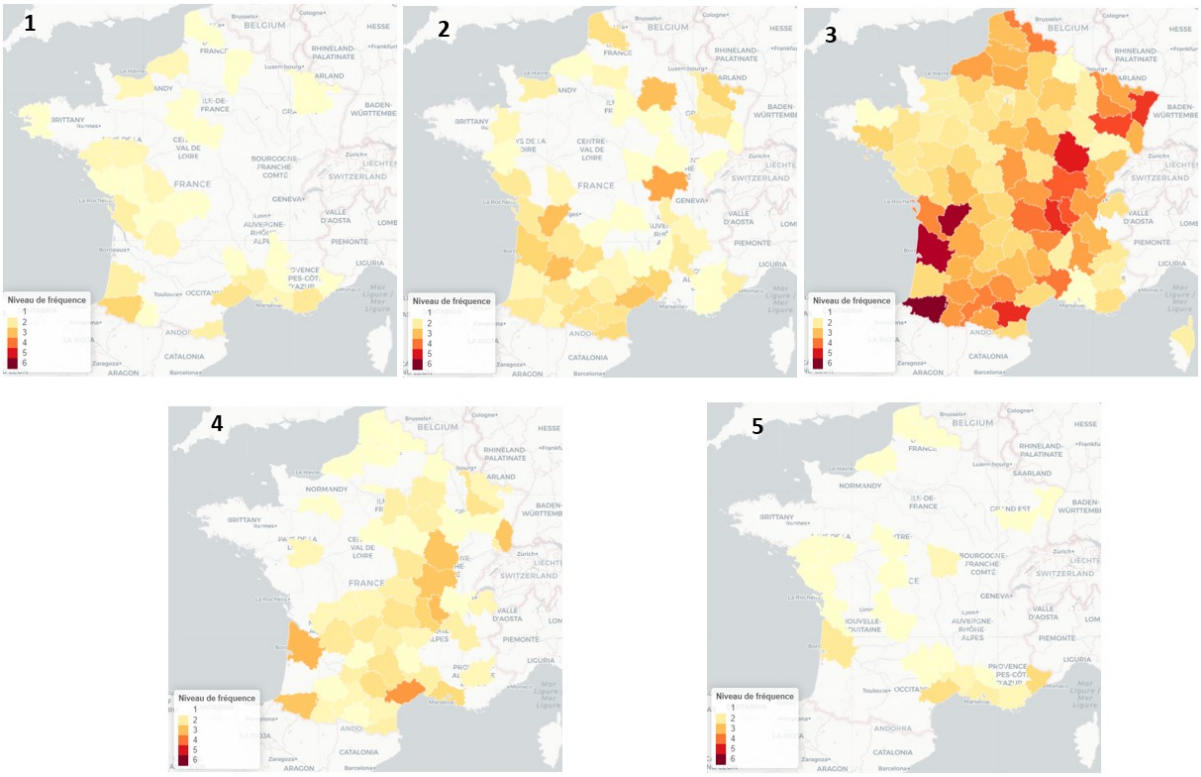
Dans son article, «La grêle en France »[21], Keraunos nous dresse la sévérité et saisonnalité de la grêle sur l'année. Keraunos nous indique une certaine classification des grêles en grêle fortes (diamètres de grêlons supérieurs à 2 cm), violentes (diamètre supérieur à 5 cm) et extrêmement sévères (diamètres supérieurs à 10 cm). Nous garderons en tête ces ordres de grandeurs lors de notre modélisation, notamment le seuil de 2 cm qui sera

[1] https://fr.wikipedia.org/wiki/Keraunos

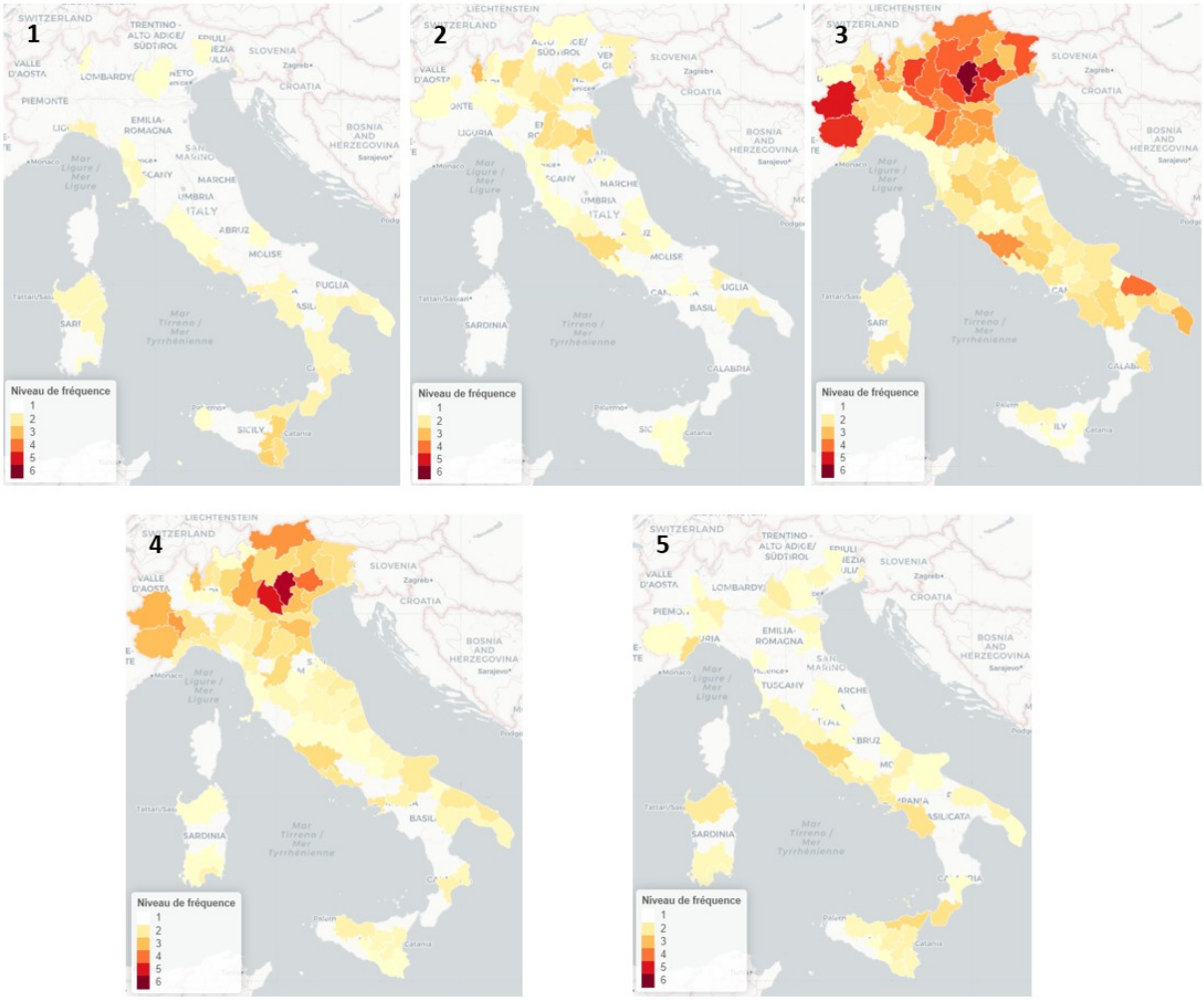
corroboré par le prochain article cité dans ce mémoire.

Ensuite, l'observatoire nous indique que de janvier à mars, la grêle est principalement observée près des côtes, sous formes d'averses courtes et des grêlons de petites tailles. Puis d'avril à mai, les chutes de grêle deviennent plus fréquentes à l'intérieur des terres en France avec des diamètres grêlons pouvant aller jusqu'à de 2 à 3 cm. Entre mi-mai et fin juillet, l'activité de la grêle atteint son apogée avec des chutes de grêlons particulièrement violents. À partir d'août et jusqu'en septembre, l'activité diminue progressivement et se concentre davantage dans les régions méditerranéennes. Le restant de l'année, les chutes se font de plus en plus rares.

Les figures 13 et 14 nous montrent que nous arrivons globalement à ces mêmes conclusions grâce à nos analyses réalisées à partir des données relevées entre 2006 et 2020 d'ESWD, données que nous présenterons en détail dans la section 3.3.1.



Graphique 13: Saisonnalité de la grêle en France
Source : ESWD (2006-2020)



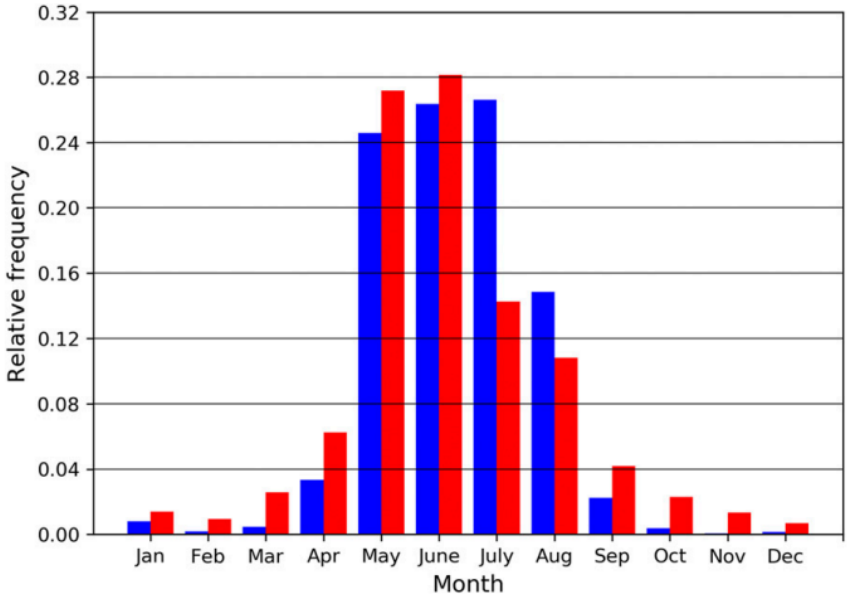
Graphique 14: Saisonnalité de la grêle en Italie
Source : ESWD (2006-2020)

Pour les figures 13 et 14, 1 correspond à la période janvier, février et mars, 2 correspond au mois d'avril, 3 correspond à la période intense de grêle de mai à juillet, ensuite nous avons la baisse de l'activité courant août et septembre et enfin sur le reste de l'année, les grêles sont peu nombreuses. Lors du pic d'activité de la grêle, nous constatons qu'en France, les départements les plus touchés sont ceux situés sur la diagonale allant du Sud-Ouest jusqu'au nord-est tandis qu'en Italie, l'activité se concentre au nord du pays, dans les zones davantage montagneuses.

2.4.3 Etude de Pùčik et al. (2019)

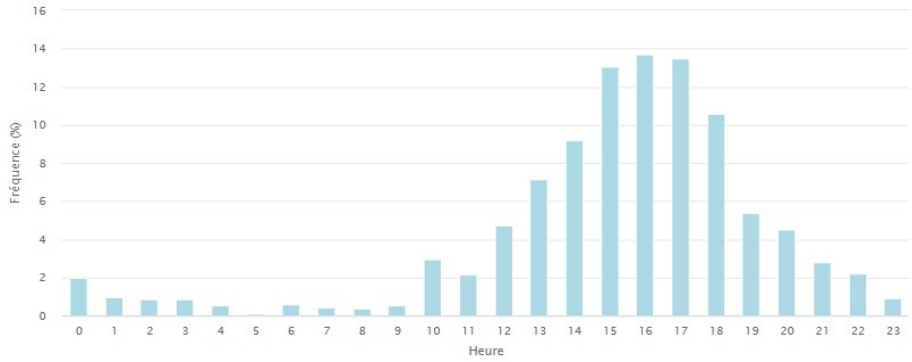
L'étude «Large Hail Incidence and Its Economic and Societal Impacts across Europe »réalisée par Pùčik et al. (2019) [24] s'appuie sur les données du laboratoire ESSL que nous présenterons en détail dans la section données que nous présenterons en détail dans la section 3.3.1. Elle présente différentes conséquences des grêles selon leurs caractéristiques.

Tout d'abord, cette étude corrobore les analyses réalisées par Keraunos concernant la fréquence mensuelle des grêles. Entre avril et mai, nous observons une augmentation rapide de l'activité de grêle, avec un pic généralement atteint en juin. Puis nous observons une forte décroissance à partir de fin août.



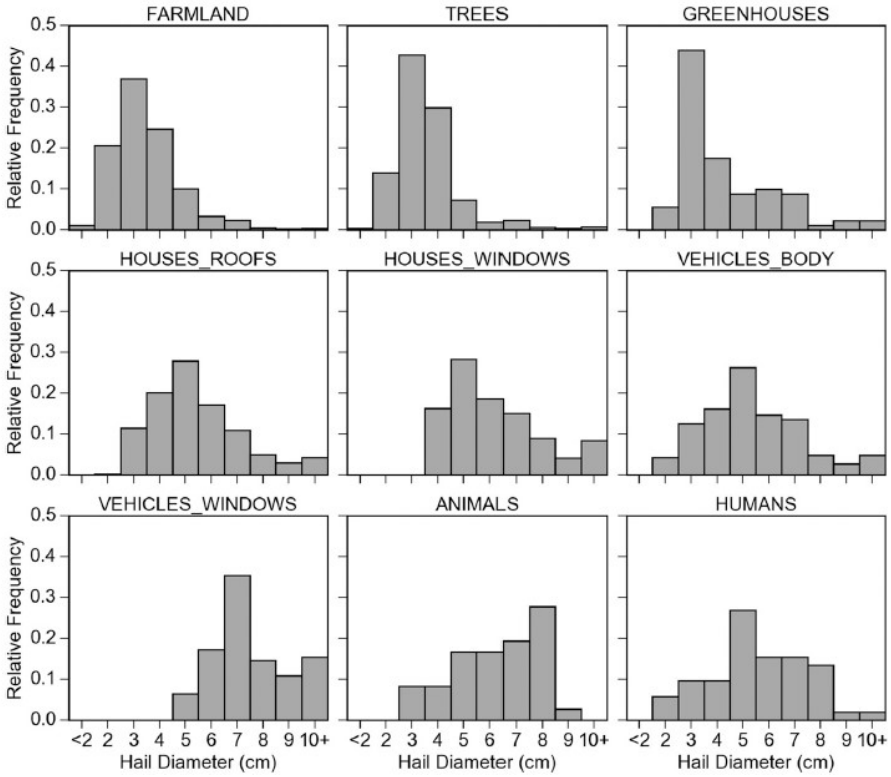
Graphique 15: Cycle annuel de la grêle en Europe en distinguant les régions situées au nord (en bleu) et au sud du 46e parallèle nord (en rouge)
Source : Pùčik et al (2019)

Les données de cette étude nous montre également un cycle journalier assez marqué comme nous le montre l'histogramme 16 où nous notons un pic en plein milieu de l'après-midi et d'après l'étude, il s'agit d'un phénomène qui se retrouve largement en Europe suggérant des conditions plus favorables à la formation d'orages de grêle à cette période-là.



Graphique 16: Fréquence horaire de la grêle en France
Source : ESWD (2006-2020)

L'étude nous indique à travers la figure 17, l'impact des grêlons selon leur taille sur différents éléments qui va particulièrement nous intéresser pour ce mémoire.



Graphique 17: Fréquence relative du type d'impact en fonction du diamètre maximal de la grêle, Source : Pùčik et al. (2019)

Nous avons en abscisse le diamètre des grêlons et en ordonnée la fréquence relative des types d'impact. Ainsi, nous observons dans le premier rectangle en haut à gauche que des dommages sont occasionnés à moins de 2 cm de diamètre tandis que pour les toitures des maisons (*houses_roofs*), nous observons des dommages dès 2 cm de diamètre et surtout à partir de 3 cm. Concernant les fenêtres des véhicules (*vehicles_windows*) ainsi que les

fenêtres de maisons (*houses_windows*), nous notons que nous avons davantage de dégâts à partir de 4 cm de diamètre. Ainsi, pour le reste de notre étude, nous allons garder ces seuils de **2 et 4 cm de diamètre**.

3 Exploration des données

3.1 Données d'exposition

La donnée de l'exposition d'AXA est essentielle pour comprendre la vulnérabilité du portefeuille de la compagnie d'assurance. Dans notre base de données, nous possédons plusieurs informations intéressantes comme le numéro de contrat, le numéro du site du bien assuré, la ligne d'activité (*Line of Business* en anglais et appelé *LoB* par la suite dans le mémoire), l'entité AXA concernée, des données sur la toiture du bien, son étage, l'année de création ou encore le type de bâtiment, à savoir maison ou immeuble notamment. Nous avons également des données géographiques comme la latitude, longitude, code postal ou encore le département ainsi que la qualité de résolution de l'information. Enfin nous avons différentes mesures du niveau d'exposition du contrat concerné. Nous avons par exemple une colonne TIV, pour *Total Insured Value* (somme totale assurée), SSIV pour *Standalone Sum Insured Value* (somme assurée autonome) ou encore la NIV pour *Net Insured Value* (Valeur nette assurée).

À partir de là, plusieurs questions se posent :

- Quelles LoB prendre en compte ?
- Les données sont-elles pertinentes et de bonne qualité ?
- Quel est le bon montant d'exposition à prendre en compte ?
- À quel niveau géographique souhaitons nous analyser les données ?

Nous allons ainsi essayer de répondre à ces questions et en tenir compte pour établir une nouvelle base de données retraitée

3.1.1 Quelles LoB prendre en compte ?

Dans notre base de données, nous possédons plusieurs lignes de métier : l'agriculture, le commercial, l'industriel, la marine, l'automobile, la municipalité et enfin le résidentiel. En termes d'exposition, le résidentiel - recouvrant les contrats Multirisque Habitation (MRH) - représente la plus grosse exposition suivie de l'industriel et du commercial.

Comme nous ne traitons pas les sinistres liés à la ligne marine, nous retirons les éléments associés dans notre nouvelle base retraitée. Enfin, nous considérons que les LoB associées au commercial, industriel et municipal auront des comportements similaires en termes de sinistralité car la majorité des expositions associées à ces LoB sont composées d'immeubles tandis que dans la LoB résidentiel, nous retrouvons à la fois des habitations dans des immeubles que dans des maisons. Enfin, nous laissons la ligne agricole à part car son comportement face à la chute de grêle sera différente comparée aux autres LoB et nous gardons la LoB automobile à part selon le même raisonnement. Nous nous retrouvons finalement avec 4 nouvelles LoB : la ligne résidentielle, la ligne affaire, la ligne automobile et enfin la ligne agricole.

3.1.2 Les données sont-elles pertinentes et de bonne qualité ?

La qualité des données est primordiale lors d'analyse actuarielle. Sans données fiables et qui peuvent être vérifiées, nous pourrions surexposer ou au contraire sous-exposer certaines zones géographiques, ce qui reviendrait à nuire la modélisation de la sinistralité de la grêle ainsi que l'interprétabilité des résultats. Dans cette partie, nous allons davantage nous pencher dans un premier temps à la qualité des variables géographiques. Nous avons ainsi le numéro de département, le code postal, la latitude, la longitude ainsi que la résolution des coordonnées précédemment citée.

Il existe plusieurs types de résolution au sein de notre base d'exposition, nous allons les citer du plus précis au moins précis : *COORDINATES*, *BUILDING*, *ADDRESS*, *STREET*, *CITY*, INSEE, *POSTCODE*, *SECTOR*, *DISTRICT*, *COUNTY*, *REGION*, *COUNTRY*, *UNKNOWN*. La première nous indique que les coordonnées en Latitude et Longitude sont précises au point près tandis que *INSEE* nous indique que nous sommes sûr de l'exposition liée au contrat à la commune selon les références de l'INSEE près tandis que *COUNTRY* nous indique que les points désignés par les coordonnées sont en réalité les coordonnées associées au centroïde du pays. Enfin, *UNKNOWN* nous montre l'absence de la précision de la qualité des données. A partir de cette connaissance de la qualité des données géographiques, nous allons garder l'information de la Latitude et Longitude associées aux lignes d'exposition avec *COORDINATES*, *BUILDING*, *ADDRESS*, *STREET*, *CITY*, INSEE, *POSTCODE*, *SECTOR*, *DISTRICT*, *COUNTY* en qualité de résolution des données.

Pour le reste des données, nous allons garder les données issues du code postal. Dans le cas où cette colonne n'est pas renseignée, nous allons utiliser les données du code de département. Il est également intéressant de noter que pour la LoB automobile, seule la

colonne du code de département est correctement renseignée. Lors de l'analyse de cette ligne de métier, il faudra ainsi se restreindre à une analyse au niveau du département. Nous avons décidé de ne pas simplement nous reposer sur la valeur indiquée par le code postal car ces lignes comportent un fort taux de données manquantes qui peut être dû à des erreurs opérationnelles, une migration des données d'un système à un autre ou un manque d'information certain par exemple. Dans le cas où le code département n'est pas renseigné, nous devons éliminer les lignes d'exposition associées mais ces cas restent marginaux au sein de notre base de données.

Maintenant que nous avons étudié le cas des données géographiques, occupons-nous des autres données qui pourraient s'avérer intéressante pour notre étude. À ce titre, nous retrouvons le type du bien, à savoir maison ou immeuble ou autres, nombre d'étage, l'étage du bien, l'année de construction, le matériau principal constituant les murs ainsi que le matériau principal de la toiture. La donnée de l'étage peut être une variable essentielle pour modéliser la grêle. En effet, un appartement situé au dernier étage pourra être plus facilement affecté par un épisode de grêle qu'un appartement au premier étage du fait de la proximité avec la toiture qui peut être endommagée par les grêlons. Il serait donc intéressant de retirer ou à minima d'accorder une pondération moindre aux expositions situées au premier étage d'un bâtiment de plusieurs étages en plus d'une pondération selon le type de toiture du bien.

Nous pouvons raisonnablement penser qu'une tuile en béton sera plus résistante qu'une toiture en tuile en terre cuite classique et ainsi ajouter une pondération sur l'exposition selon la résistance des matériaux à la grêle. Cependant ces données sont très peu remplies et sont donc en l'état inutilisables, tout comme l'année de création du bien ou le matériau principal composant les murs. Nous ne gardons de cette liste que le type de bien, à savoir immeuble ou maison. Une maison peut être davantage exposé qu'un immeuble pour différentes raisons.

En effet, la surface de la toiture d'une maison est souvent plus grande une fois rapporté à son volume total. La maison est ainsi plus exposée aux impacts de la grêle sur une proportion significative de sa structure comparée aux immeubles. De plus, les maisons sont souvent construites sur des terrains davantage ouverts sans forcément la protection d'autres grands bâtiments ou autres structures pouvant protéger des impacts directs ou indirects de la grêle. Les immeubles réduisent ainsi leur exposition directe aux grêles et aux vents violents qui pourraient rediriger ces grêles.

3.1.3 Quel est le bon montant d'exposition à prendre en compte ?

Nous pouvons désormais nous intéresser au montant d'exposition. Dans la base de données initiale, nous possédons l'information de la *TIV*, correspondant au montant total de l'exposition, la *NIV* qui correspond à l'exposition nette de conditions financières des traités de réassurance proportionnel et non proportionnelles avec notamment la priorité ou la limite par exemple ou encore la *SSIV* qui correspond au montant d'exposition ventilé par site d'exposition. Ces éléments prennent également en compte la valeur du contenu du bien assuré. Dans le cas de notre étude nous allons privilégier la *NIV* pour être cohérents avec le paiement des sinistres en face de ces contrats. Cependant, lors de l'analyse des bases de données, nous observons plusieurs lignes avec le même montant de *NIV*.

En réalisant une analyse plus poussée de ces éléments, nous nous sommes rendu compte qu'il s'agissait des contrats multi-sites. Il s'agit d'**un même contrat** qui couvre **différents sites** pouvant être sur des zones géographiques différentes. Dans ce cas, la *NIV* a été paramétrée de sorte à être une proportion du montant total d'exposition. Ainsi, si le contrat couvre 3 sites situés à Paris, Lyon et Marseille, avec une exposition respective de 7, 3 et 2 millions d'euros, l'exposition totale va être de 12 millions d'euros. La *NIV* associée à chacun de ces sites sera donc de 4 millions d'euros, ce qui sous-estime l'exposition à Paris et à Lyon tandis qu'elle le surestime à Marseille. La *SSIV* quant à elle donnera le bon montant d'exposition. Nous allons ainsi reprendre la valeur de la *NIV* pour les données en dehors des cas où nous observons un cas de multisites réperé par une duplication des numéros de contrats dans la base de données pour lesquels nous attribuons la valeur de la *SSIV* associée.

ContractID	SiteID	LoB	Département	Code Postal	Toit	TIV	NIV	SSIV
ID123904	S321678	Comm	75	75006	NA	6	4	7
ID123904	S321679	Comm	69	69006	NA	6	4	3
ID123904	S321680	Comm	13	13008	Béton	6	4	2

Table 1: Exemple de données d'exposition en Millions d'euros

De plus, le portefeuille d'exposition que nous utilisons est celui de 2024. Or, entre les dates de sinistres de grêle et 2024, le portefeuille assuré par AXA a évolué et n'est pas le même que celui de 2022 par exemple. Ainsi, nous prenons en compte un **indice d'évolution de la taille du portefeuille d'AXA** par ligne de métier pour réévaluer la taille du portefeuille. Nous faisons l'hypothèse simplificatrice que l'évolution du portefeuille reste similaire sur les différents départements d'un pays pour une même LoB.

3.1.4 À quel niveau géographique analyser les données ?

Grâce à l'analyse réalisée sur la qualité des données des coordonnées géographiques, nous avons pu cerner les données que nous allons pouvoir utiliser par la suite. Concernant la LoB automobile, nous sommes cantonnés à l'échelle du département en raison de l'absence d'une qualité de donnée plus fine à l'échelle du code postal ou du code commune de l'INSEE par exemple. Concernant les autres LoB de l'agriculture, des affaires ou du résidentiel, nous sommes libres à ce stade de choisir entre ces différentes échelles. La seule condition que nous devrons respecter sera la **cohérence** du niveau de données entre la base des expositions et la base des sinistres. En termes de pays, au vu des précédents épisodes de grêles qui ont particulièrement touché la France et l'Italie, nous allons nous concentrer sur ces deux pays.

3.2 Données de sinistralité

En face des données d'exposition, nous avons les données de la sinistralité du portefeuille d'AXA. Nous obtenons de la part de l'équipe chargée de l'analyse des sinistres en réassurance, les montants des **évènements de grêle** par **année** et par **pays** à différentes **dates d'extraction** de la sinistralité. Plusieurs colonnes sont intéressantes dans ces bases de données à savoir le numéro de contrat, le numéro de sinistre, la date d'extraction des données, l'entité d'AXA ayant remonté le sinistre, la date du sinistre, la LoB associée au sinistre, le code postal, le montant de la charge et sur certaines dates, nous avons également le montant de la première évaluation du sinistre. A partir de ces éléments, plusieurs questions se posent à nouveau :

- Comment bien comprendre le montant de sinistre indiqué ?
- Les données sont-elles pertinentes et de bonne qualité ?
- Quel est le niveau géographique auquel réaliser l'étude ?
- Comment lier les données d'exposition aux données de sinistres ?

3.2.1 Comment bien comprendre le montant de sinistre indiqué ?

L'équipe au sein de laquelle ce mémoire a été rédigé est l'équipe de **provisionnement** en **réassurance**. À ce titre, nous analysons les données qui nous sont fournies par les entités du groupe AXA.

Les traités de réassurance entre les entités du groupe AXA et AXA SA sont constitués principalement de traité en **excédant de sinistre** (abrégé XS par la suite). Un traité de

réassurance en XS est un accord où le réassureur prend en charge les sinistres dépassant un certain montant, appelé priorité, jusqu'à une limite définie par le traité. Deux types de traité XS sont appliqués de manière successive, à savoir tout d'abord un **XS par risque** et un traité **XS par évènement**. Nous pouvons également parler dans ce cas de traité XS par risque qui s'applique au bénéfice du traité XS par évènement (ou *XS per Risk treaty inures to the benefit of the XS per Event* en anglais). Ainsi, le traité en XS par risque s'applique d'abord à chaque sinistre de la cédante, limitant ainsi son exposition aux pertes majeures individuelles et ainsi écrêter ses sinistres. Ensuite, concernant les montants toujours à charge de la cédante, nous appliquons le traité en XS par évènement qui couvre les pertes cumulées résultant d'un même évènement de grêle.

La définition d'évènement dans le cadre des sinistres de grêle se limite au **cumul des sinistres** sur un ou plusieurs jours, durée définie de manière précise dans le traité de réassurance. La plupart du temps, les sinistres n'impactent pas le traité de réassurance qui couvre ainsi principalement les évènements de grêle importants. De plus les entités sont dans l'obligation de déclarer les sinistres qui atteignent selon leurs projections un certain pourcentage de la priorité, 80% de la priorité par exemple. Ce mécanisme fait qu'en tant que réassureur des entités, nous avons connaissance de tous les sinistres déclarés issus des évènements dont la charge de sinistre totale **dépasse le seuil de déclaration**. Cela explique notamment pourquoi nous n'avons pas de données de grêle en 2023 en France malgré différents évènements de grêle. Cela réduit donc la base de données à notre disposition et nous ne nous focaliserons ainsi que sur les plus gros évènements de grêle ayant dépassé ce seuil de déclaration. Il est également à noter que les montants remontés par les cédantes dans le cas où l'évènement dépasse le seuil, sont les montants sans prise en compte de la priorité du traité en XS par évènement, nous avons ainsi le montant total à charge de la cédante.

Ainsi, nous nous retrouvons en France avec 1 évènement en 2021, et 4 évènements en 2022 tandis qu'en Italie nous avons 4 évènements en 2022 et 2 en 2023. Pour chacun de ces évènements, nous prenons en compte les fichiers envoyés par l'équipe sinistre avec les dates d'extraction de données les plus récentes afin de travailler sur les sinistres les plus développés possible. De plus, pour prendre en compte l'inflation, nous allons redresser les charges de sinistres avec l'indice **FFB du coût de la construction** (ICC FFB) fourni par Fédération Française du Bâtiment.[14]

3.2.2 Les données sont-elles pertinentes et de bonne qualité ?

Le fait que ce soient différentes entités du groupe AXA qui nous remontent les données influent sur la cohérence des données. En effet, d'une entité à l'autre, l'agrégation des données n'est pas la même et le nom des colonnes ainsi que leur contenu peuvent être différents. Par exemple la colonne de la première estimation des données n'est présente que chez quelques entités et n'est donc pas exploitable au global. Ensuite, selon les années, nous n'avons pas des informations détaillées concernant les LoB des sinistres. Cela rend ces données inexploitables car non rattachables à la base d'exposition. Selon les entités, nous n'avons pas les mêmes dénominations pour ces LoB.

Ainsi ce qu'on appelait RESIDENTIEL pour l'exposition, cela correspond en Italie à *PROPERTY* pour une entité, *PROPERTY PERSONAL* ou *HOUSEHOLD* pour une autre. En France, l'information initialement présente dans les bases de donnée n'était détaillée qu'au niveau des LoB Automobile ou non Automobile. La catégorie non Automobile recouvrant aussi bien les contrats MRH que les contrats des lignes agricoles, commerciales ou industrielles. Différents échanges avec les filiales nous ont permis d'obtenir une nouvelle maille plus fine qui nous permet à présent de les distinguer pour les sinistres s'étant réalisés en 2022. Maintenant, la LoB RESIDENTIEL correspond à DomP (pour Dommage Particulier) pour les différentes entités en France par exemple, tandis que la ligne AUTOMOBILE correspond à deux catégories, *AUTOP* (Auto Particulier) et *AUTOE* (Auto Entreprise) côté sinistre.

Un traitement préalable a donc été réalisé afin d'établir une correspondance entre les informations des LoB contenu dans nos bases d'exposition et nos bases sinistres qui nous servira ensuite à joindre nos bases de données et effectuer les analyses pour permettant d'effectuer une modélisation des sinistres.

3.2.3 Quel est le niveau géographique auquel réaliser l'étude ?

Les données des codes postaux sont présentes dans les bases de données des entités en France pour les sinistres de 2021 et 2022, tandis que seul le code de la province est disponible dans les fichiers fournis par les entités en Italie, ce niveau de détail de la province étant l'équivalent du département en France. Ce code province se caractérise par deux lettres permettant d'identifier les provinces en Italie. Ces données géographiques pour l'Italie ne sont cependant disponibles qu'en 2023. Nous décidons alors de ne poursuivre l'analyse qu'avec les données en France étant donné qu'une modélisation des risques en Italie sur la base de 2 uniques évènements de 2023 ne seraient pas pertinents.

En France, nous pouvons ainsi étudier la grêle au niveau du code postal qui nous paraît judicieux étant donné le caractère local de ces évènements. Par la suite, nous aurons besoin d'utiliser des données météo sous forme de radar qui sont plus facilement exploitable avec des codes communes définies par l'INSEE. Ce code, également appelé code officiel géographique est composé de 5 chiffres : les 2 premiers représentent le numéro du département, et les 3 derniers désignent un code spécifique à la commune. Ces codes sont donc plus précis que les codes postaux et les deux sont différents la plupart du temps.

Nom Commune	**Code Postal**	**Code Commune**	**Code DPT***	**Nom DPT***
PARIS	75006	75106	75	PARIS
CHEVREMONT	90340	90026	90	BELFORT
THEREVAL	50180	50239	50	NORMANDIE
THEREVAL	50570	50239	50	NORMANDIE
CORDON	74700	74089	74	HAUTE-SAVOIE
DOMANCY	74700	74103	74	HAUTE-SAVOIE
SALLANCHES	74700	74256	74	HAUTE-SAVOIE
ST PIERRE LAVAL	42620	03250	03	ALLIER
LEYVAUX	43450	15105	15	CANTAL

Table 2: Exemple de correspondance code postal - code commune en France
*(*DPT : département)*

Différents cas sont indiqués dans la table 2. Tout d'abord, nous pouvons voir que la plupart des codes postaux sont bien différents des codes communes. Ensuite, les lignes 3 et 4 nous indiquent qu'un même code commune peut faire référence à plusieurs codes postaux tandis que les lignes 5 et 6 nous montrer le cas inverse où un même code postal fait référence à plusieurs codes communes. Enfin, les deux dernières lignes nous indiquent pourquoi il est préférable d'utiliser les codes communes pour la suite de notre analyse. En effet, nous voyons que les deux premiers caractères du code postal ne correspondent pas toujours au bon département, bien que ces cas soient minoritaires. Nous décidons alors de transformer les données géographiques des codes postaux en codes communes. Nous nous appuyons sur le tableau de correspondance fourni par data.gouv.fr afin d'établir cette nouvelle table.

Après avoir réalisé la jointure entre notre base sinistre et la table de correspondance précédemment citée, nous nous sommes aperçus de la présence de nombreuses **valeurs manquantes** alors que la base des codes postaux est censée être exhaustive. Après analyse des cas avec les valeurs manquantes, nous réalisons que des valeurs en code commune ont également été entrés dans la colonne CODE POSTAL. Les échanges avec les entités nous ont permis de comprendre que certaines entités remplissaient cette colonne avec en

priorité les codes postaux, et lorsque cette information venait à manquer, les entités indiquaient le code commune. D'autres indiquent en priorité les codes communes. Dans le cas où la jointure nous donnait une valeur manquante et que l'entité indique en priorité le code postal, nous refaisons alors une nouvelle jointure sur la base des codes communes, ce qui réduit fortement le nombre de jointures absentes. Les cas résiduels pouvant venir d'erreur de frappe opérationnelle ou encore des codes postaux qui n'existent plus en raison de réorganisation du territoire. Cependant, cette manipulation réduit fortement la fiabilité des données et nous pouvons nous retrouver avec des montants de sinistres affectés aux mauvaises communes. Par cette raison et par l'absence de données aussi fines au niveau de l'Italie, nous décidons de rester à l'échelle du département bien que cette échelle géographique soit moins fine pour l'analyse des données.

3.2.4 Comment lier les données d'exposition aux données de sinistres ?

La base de données comportant les données d'exposition contient des numéros de contrat et la plupart des fichiers de sinistres remontés par les entités contient également une colonne pour les numéros de contrat ainsi qu'une colonne pour les numéros de sinistre. Afin de joindre les deux bases et obtenir une base de données complète, nous pouvons ainsi utiliser les colonnes de numéro de contrat. Cependant, nous avons constaté que cette jointure produisait principalement des valeurs manquantes, les numéros de contrat des deux bases ne correspondant pas. Cela est dû à des raisons de sécurité et de confidentialité des données. Des échanges ont encore lieu pour avoir une jointure avec des **données anonymisées**.

Dans le cadre de notre étude, nous ne pouvons donc pas utiliser cette jointure pour étudier en ligne à ligne les différents sinistres. Cette méthode aurait notamment pu être efficace pour connaître le montant exact des données d'exposition associé à chaque sinistre avec l'ensemble des informations détaillées associé au bien et cela peut permettre de vérifier si la correspondance des LoB est bien cohérente. Ne pouvant utiliser cette jointure, nous avons décidé de regrouper les sinistres par LoB, à savoir RÉSIDENTIEL, AFFAIRE, AGRICOLE et AUTOMOBILE par département. Nous avions également essayé de faire cette jointure au niveau des codes postaux et codes communes, mais en raison des nombreuses hypothèses lors de la jointure, nous nous retrouvions avec plusieurs codes postaux sans exposition en face et le cas inverse se retrouvait également.

En travaillant au niveau du département, nous perdons de nombreux détails concernant les sinistres mais nous faisons l'hypothèse que les sinistres auront un comportement

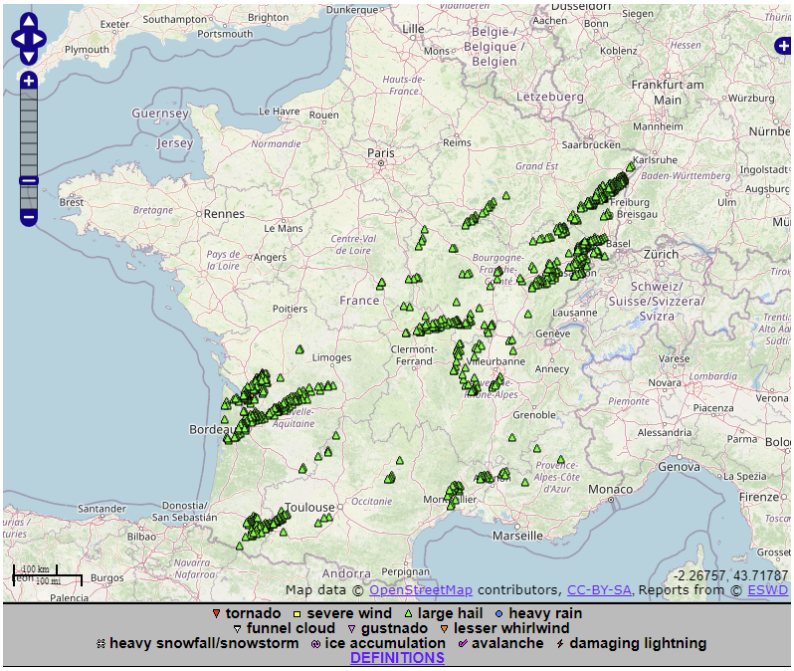
similaire au sein d'une même LoB. Concernant la ligne de métier RÉSIDENTIEL, nous avons tout de même une colonne «type de bâtiment»avec la mention de maison ou immeuble qui reste intéressante. En effet, face au risque de grêle, les maisons sont bien plus vulnérables que les immeubles et auront ainsi un comportement de sinistralité différent. Pour garder cette information malgré l'agrégation, nous décidons de calculer le pourcentage de maison par département au lieu d'abandonner les informations contenues dans cette colonne. Nous décidons également de le calculer pour la LoB AFFAIRE mais elle aura une moins grande utilité car la majorité des biens contenus dans cette LoB sont des immeubles. Enfin, ce niveau d'agrégation par département permet de réduire le bruit statistique induit par des erreurs opérationnelles de saisie par exemple et va permettre de capturer des effets d'accumulation de sinistres sur une zone plus large. Une amélioration à cette étude pourra toutefois être réalisée une fois un lien entre les numéros de contrat obtenu.

3.3 Données météorologiques

3.3.1 European Severe Weather Database (ESWD)

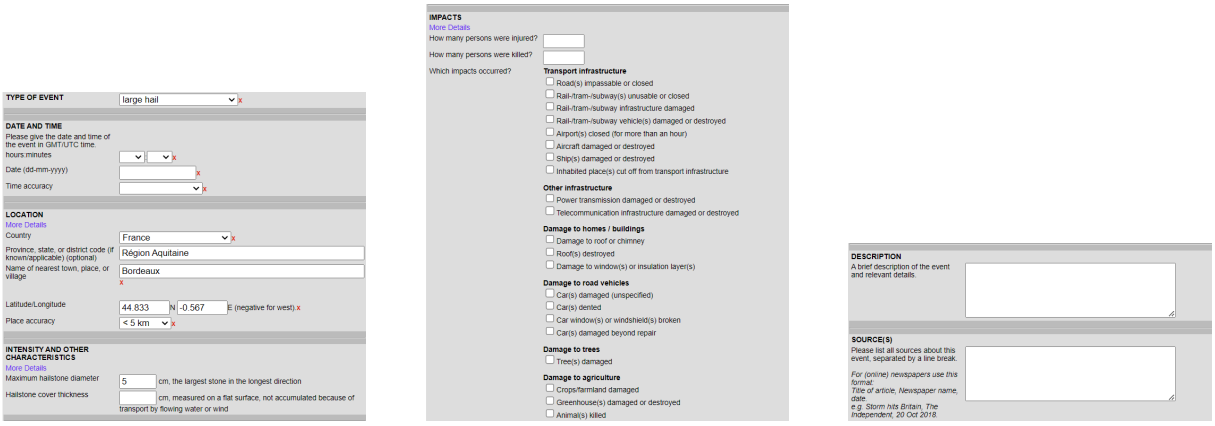
L'*European Severe Weather Database* (ESWD) [13], pouvant être traduit en français par base de données européenne sur les phénomènes météorologiques violents est une base de données élaborée par le *European Severe Storms Laboratory* (ESSL) qui est une organisation scientifique réalisant des recherches sur les orages, tornades, épisodes de précipitations intenses, avalanches et de grêles en Europe avec pour objectif d'améliorer la compréhension scientifique et la prévision des phénomènes météorologiques sévères en encourageant une coopération européenne sur le sujet.

Une visualisation est accessible à tous sans inscription et gratuitement via le lien suivant https://eswd.eu/cgi-bin/eswd.cgi et permet de voir en continu les éléments que l'on souhaite analyser, dont les évènements de grêle dans le cadre de notre étude. Un exemple est montré dans la visualisation 18.



Graphique 18: Visualisation de l'évènement QIARA de juin 2022 en France
Source : ESWD

Cette organisation collecte ainsi de nombreuses données météorologiques à partir de rapports sur les phénomènes observés pouvant être soumises par tout le monde à partir d'un formulaire à remplir sur le site de la base de données : https://www.eswd.eu/ . Cela encourage ainsi la participation du public.



Graphique 19: Formulaire d'envoi de rapport ESWD

Ensuite, chaque rapport recueilli fait l'objet d'un contrôle de qualité des données et reçoit un code de qualité selon le niveau de vérification de ces données.

[1]https://www.essl.org/cms/vision-mission-and-strategy-of-essl/

Niveau de qualité	Désignation
QC0	Rapports tels que reçus
QC0+	Rapports dont la plausibilité a été vérifiée,
QC1	Rapports confirmés par des sources fiables
QC2	Rapports ayant fait l'objet d'études de cas scientifiques

Table 3: Niveaux de vérification des données d'ESWD
Source : ESWD

Ce système de vérification implémenté par l'organisme permet ainsi de garantir un certain niveau de **fiabilité des données** ainsi que des analyses issues de cette base de données. Il est cependant important de souligner un potentiel **biais** dans la méthodologie de collecte de donnée. En effet, étant donné que cette base de données s'appuie sur des rapports soumis par des observateurs et non des données météorologiques relevées par des radars par exemple, cela peut introduire un **biais de représentation des données**. Ainsi, une commune densément peuplée peut être sur-représentée dans la base de données car la probabilité qu'un phénomène météorologique soit observé et signalé y est plus élevée. À l'inverse, des communes peu peuplées risquent d'être sous-représentées à cause d'un manque d'observateurs sur place. Ainsi, une zone avec une absence de donnée ne signifie pas forcément une absence de grêle. Pour atténuer ce biais, nous croiserons les données issues de cette base avec d'autres sources d'information comme des radars météo.

Nous passons en revue différentes informations fournies par ESWD dans la table 4.

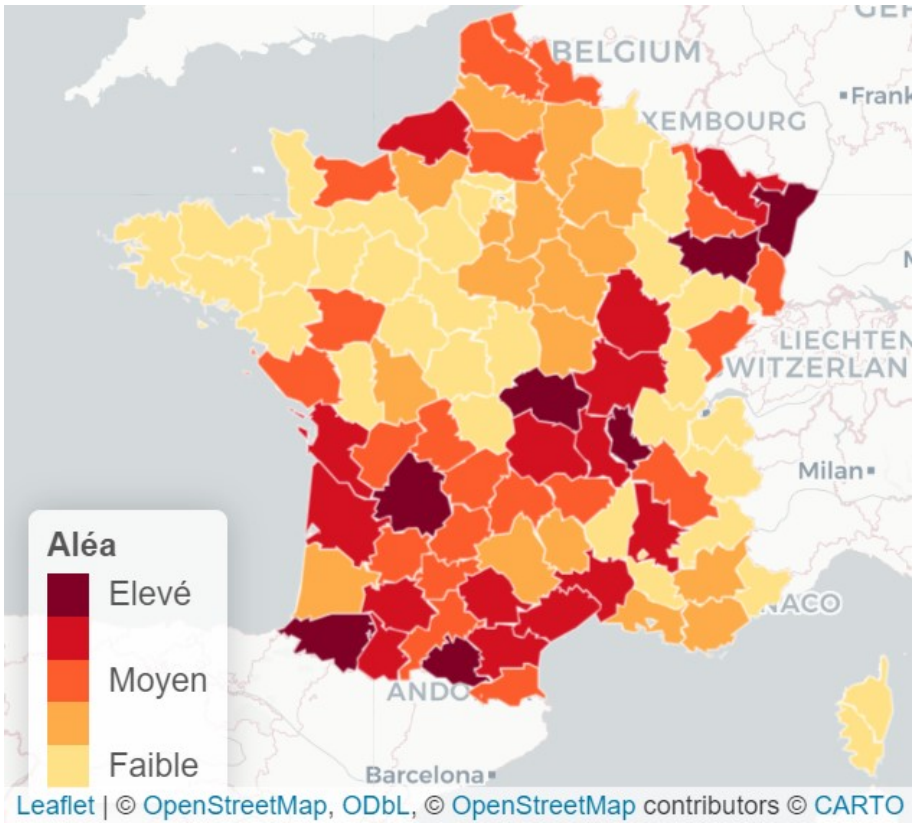
Variables	Détail
TIME_EVENT	Date et heure de l'évènement
COUNTRY	Pays
LATITUDE	Latitude de l'occurence de l'évènement
LONGITUDE	Longitude de l'occurence de l'évènement
TYPE_EVENT	Type de l'évènement : Hail (Grêle) pour notre cas
MAX_HAIL_DIAMETER	Diamètre maximal de grêlon observé
QC_LEVEL	Niveau de qualité des données

Table 4: Extrait des variables utilisées des données de ESWD
Source : ESWD, European Severe Weather Database (ewd.eu)

Étant donné qu'il s'agit d'une des rares bases ayant un historique relativement long, nous décidons d'utiliser ESWD pour établir une carte de la fréquence historique des grêles qui nous servira de carte d'aléa utilisée dans notre modélisation. Pour ce faire, nous allons garder les données entre 2006 et 2020 et nous allons analyser l'occurrence annuelle de la grêle par département. Ces départements sont identifiés grâce aux coordonnées de la base

de données. Puis nous diviserons ce nombre par l'étendue de la période d'observation, ici 15 années pour obtenir une fréquence. Enfin, pour une meilleure visualisation et pour simplifier les données pour notre modélisation, nous allons regrouper les fréquences en 5 catégories : 1 étant des départements peu touchés par la grêle tandis que 5 représente les départements régulièrement touchés par ces évènements.

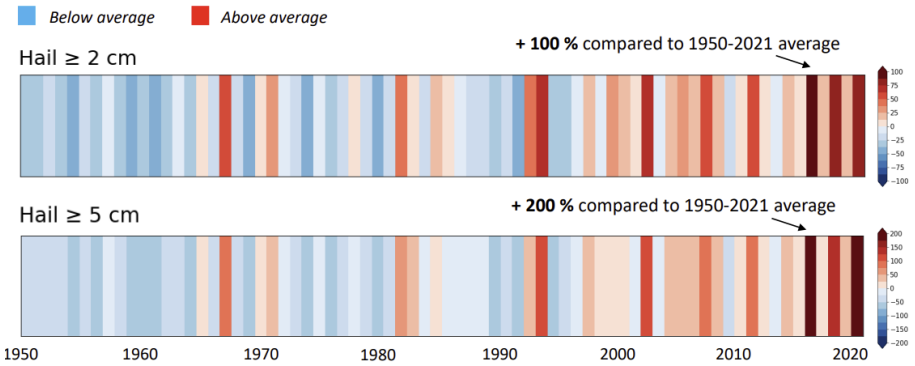
Pour notre modélisation, nous utiliserons ces données de ESWD pour identifier les différentes communes affectées par les épisodes de grêle, encore une fois grâce aux informations de la latitude et la longitude. Nous les agrégerons ensuite par département, créant ainsi la variable *NOMBRE_ COMMUNE*, qui représente le nombre de communes touchées par la grêle pour chaque évènement.



Graphique 20: Carte d'aléa de la grêle en France métropolitaine
Source : Carte construite à partir des données ESWD sur 2006-2020

Le laboratoire ESSL a par ailleurs présenté lors d'une conférence organisée par la *World Meteorological organization* en novembre 2022 à Genève [19], une analyse intéressante sur la fréquence des grêles en Italie du nord. Cette analyse montre une augmentation claire des épisodes de grêle depuis les années 2000 et plus particulièrement depuis 2015 symbolisé par les traits en rouge dans le graphique 21. Il faut cependant garder à l'esprit encore

une fois que le nombre de rapport soumis a grandement varié depuis 1950.



Graphique 21: Augmentation de la fréquence de grêle dans le nord de l'Italie
Source : ESSL (2022)

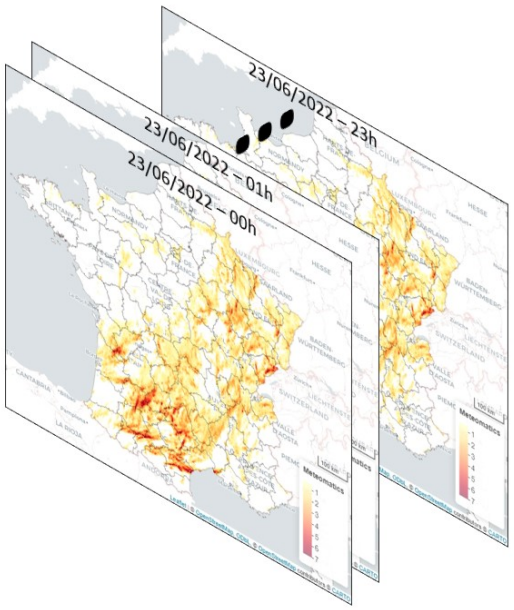
3.3.2 Meteomatics

Une autre source de données essentielle pour notre mémoire est **Meteomatics**. Il s'agit d'une société de technologie météorologique fondée en 2012 en Suisse qui se spécialise dans la collecte de données météorologiques à travers ses propres drones appelés Meteodrones, mais aussi de diverses bases de données comme les stations météologiques ou des données radars de différents instituts. Meteomatics fournit également des prévisions météorologiques à haute résolution à des fins commerciales via une *Application Programming Interface* (API), qui est un outil permettant aux utilisateurs d'accéder à des données précises sans avoir à toucher toutes les données sous-jacentes ayant permis de construire les données dérivées. Une comparaison peut être faite à un serveur au restaurant qui prendrait notre commande, à savoir la requête de données, et la donnerait en cuisine dont la personne responsable va interpréter la commande en choisissant soigneusement les ingrédients à sa disposition en cuisine et en les transformant avant de les remettre au serveur qui va nous apporter les données désirées. L'API serait donc le serveur de ce restaurant. L'API de Meteomatics est ainsi particulièrement utile dans le cadre de notre mémoire car les données seront plus facilement requêtables sous R et disponibles quotidiennement selon les limites définis par le contrat. Dans notre cas, nous allons utiliser l'API concernant la grêle.

Meteomatics s'appuie sur de très nombreuses sources de données [23], aussi bien des modèles de prévisions météorologiques comme le modèle prévisionnel AROME de Météo-France. La société utilise également des données de radars de plusieurs pays dont la France en plus des données satellitaires et des stations météorologiques publiques comme privées. Enfin, elle utilise également le modèle de réanalyse global ERA5 réalisés par

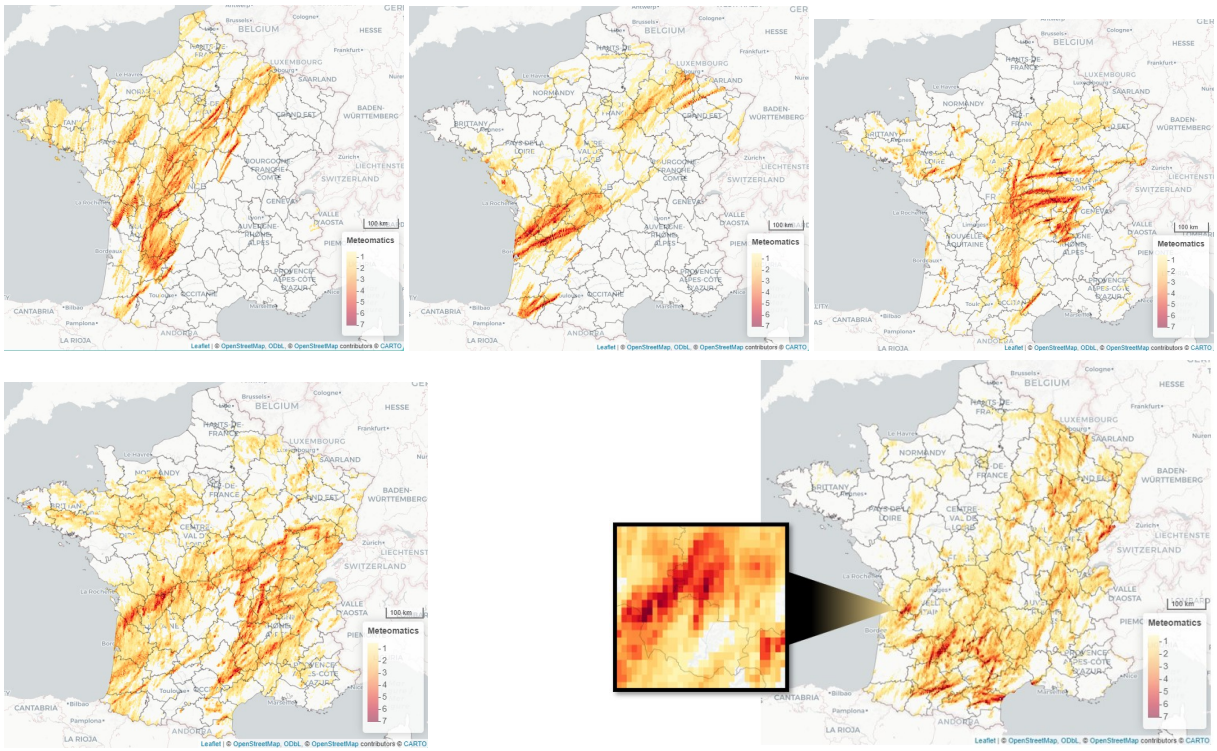
ECMWF (European Centre for Medium-Range Weather Forecasts) ou Centre européen pour les prévisions météorologiques à moyen terme en français. Ces données de réanalyse combinent des observations historiques avec des modèles météorologiques pour créer une représentation cohérente du passé. Parmi les variables principales utilisées [11] qui sont pertinentes dans notre analyse, nous relevons l'énergie potentielle de convection (CAPE en anglais). La CAPE représente l'énergie potentielle par unité de masse qu'une parcelle d'air plus chaude que son environnement peut libérer lorsqu'elle s'élève induite par l'énergie d'une parcelle d'air plus chaude que son environnement. Une masse d'air est plus chaude et moins dense que l'air environnant, elle tend à monter, ce qui peut entraîner des mouvements convectifs : une valeur élevée indique que l'atmosphère est instable et peut créer les conditions favorables à la naissance d'orages incluant ceux associés à la grêle.

L'API de Meteomatics offre un accès à des données de modélisation des tailles de grêlons avec une résolution spatiale de 2 km par 2 km et une fréquence horaire. Pour des raisons d'optimisation des requêtes, nous limiterons notre analyse à la France métropolitaine hors Corse. Cette haute résolution nous permet une analyse détaillée des grêles à l'échelle locale ce qui est important pour étudier l'impact de ces évènements. Étant donné que les grêles ne sont pas des phénomènes fréquents, la majorité des données sur la carte seront des zéros. Il est essentiel de prendre en compte cette caractéristique dans l'analyse pour mieux appréhender le phénomène. La carte 22 nous montre un exemple de visualisation des données de Meteomatics sur un jour à travers plusieurs heures.



Graphique 22: Données Meteomatics

Ces données Meteomatics nous seront ainsi précieuses pour caractériser au mieux les évènements de grêle ayant entraîné des montants de sinistres importants au sein du portefeuille d'AXA. Les cartes de la figure 23 nous montrent ainsi l'empreinte de grêle de la tempête de grêle QIARA du 19 au 23 juin 2022.



Graphique 23: Evènement de grêle QIARA du 19 au 23 juin 2023
Source : Meteomatics

3.3.3 Météo-France

Météo-France est le service météorologique et climatologique français qui met à disposition une base de données météorologiques. Météo-France est responsable de l'observation, de la prévision, et de l'étude des phénomènes météorologiques, ainsi que de l'émission d'alertes météorologiques pour l'ensemble des territoires français, aussi bien en métropole qu'en outre-mer. Depuis le 1er janvier 2024 [1], Météo-France offre un accès gratuit à ses données qui prennent en compte aussi bien des observations météorologiques en temps réel que des données climatologiques historiques ou des informations issus de radars. Cette démarche s'est réalisée dans l'optique de d'aider au développement d'innovations ainsi que pour favoriser la transition écologique et énergétique [22].

Dans le cadre de cette étude, nous allons principalement nous intéresser aux données

[1] https://www.info.gouv.fr/actualite/meteo-france-la-reutilisation-des-donnees-publiques-devient-gratuite

issues des stations météorologiques. Météo-France dispose ainsi d'un réseau assez dense de stations météorologiques réparties sur tout le territoire français, nous pouvons en compter près de 2 000 en France. Ces stations disposent de radars qui permettent de repérer différentes formes de précipitations comme la pluie, la neige ou encore la grêle dans le cadre de notre mémoire ainsi que de relever d'autres paramètres comme le taux d'humidité, la vitesse du vent par exemple. Ces radars possèdent une portée de 100 km environ.

Nous allons nous appuyer sur les données climatologiques de base avec un relevé quotidien afin de réaliser des analyses sur des données assez fines bien que davantage volumineuses que des données mensuelles et peuvent être téléchargées librement ici : https://meteo.data.gouv.fr/datasets/. Nous extrayons ces données par fichier *csv*. Ces données sont mises à jour selon les informations contenues : jusqu'en 2022, les données sont mises à jour mensuellement tandis que les données comprises entre 2023 et 2024 sont mises à jour quotidiennement. Ces bases étant réalisées par département, nous avons réalisé un algorithme qui nous permet de requêter les données entre 2020 et 2024 sur les différents départements afin de ne constituer qu'une seule et unique base de données regroupant les données des stations météorologiques des différents départements.

Nous avons deux types de fichier : un type appelé «RR-T Vent»et un autre appelé «Autres paramètres».

Les fichiers avec «Autres paramètres»font référence à des données comme le taux d'humidité, de pression, et l'occurrence ou non d'orage, de brume, de neige, de verglas, de précipitation ou encore de grêle et de grésil. Nous allons ainsi récupérer à partir de ces fichiers les occurrences de grêles et de grésil ainsi que leur date de mesure sur les différentes stations météorologiques.

Les fichiers avec «RR-T Vent»font référence aux données concernant la précipitation, la température ainsi que le vent. Nous présentons quelques-unes des variables principales que nous allons utiliser :

Nom paramètre	Détail
NUM_POSTE	Numéro Météo-France du poste sur 8 chiffres
LAT	Latitude
LON	Longitude
AAAAMMJJ	Date de la mesure (année mois jour)
RR	Quantité de précipitation tombée en 24 heures
TN	Température minimale sous abri
HTN	Heure de la température minimale (TN)
TX	Température maximale sous abri
FXY	Maximum quotidien de la force maximale horaire du vent moyenné sur 10 minutes, à 10 m
DXY	Direction de FXY (direction du vent)
XY	Heure de FXY (heure de la force maximale du vent)

Table 5: Paramètres des données météorologiques du fichier RR-T Vent de Météo-France

Il nous semble intéressant de prendre en compte des données sur la température car nous avons vu que les écarts de température entre la base et le sommet du nuage était à l'origine de la formation des orages pouvant potentiellement provoquer de la grêle. Ainsi, des écarts brusques de température pourraient induire la formation d'orage et ainsi expliquer la grêle. C'est également dans cette optique que nous gardons des mesures sur les précipitations pour détecter des orages. Enfin, nous avons décidé de prendre en compte des paramètres sur le vent comme son heure ainsi que sa direction.

En effet, plus la vitesse du vent est élevée, plus les grêlons sont projetés avec une énergie accrue, ce qui peut augmenter la gravité des dégâts. Nous pensons également que la direction du vent est un autre facteur qui peut amplifier les dommages causés par les grêles. L'orientation du vent par rapport aux surfaces exposées peut modifier la manière dont les grêlons frappent ces dernières. Lorsque les grêlons frappent les surfaces sous un angle incliné, cela pourrait occasionner davantage de dégâts sur les vitres par exemple. De plus, la direction du vent peut influencer la trajectoire des grêlons, affectant ainsi la distribution spatiale des dommages au sein d'un département par exemple. Les zones exposées directement au vent sont plus susceptibles de subir des dommages plus importants que celles protégées par des obstacles naturels comme de grands arbres ou des montagnes ou artificiels comme d'autres immeubles ou constructions diverses.

Cependant, nous avons décidé de ne pas retenir les données issues des fichiers de «RR-T Vent»en raison de l'absence de données pour la plupart des stations météorologiques pouvant provenir des installations possiblement hétérogènes selon les différentes stations. Pour mesurer le taux d'absence de données, nous avons croisé ces données avec les fichiers «Autres paramètres»pour relever les stations qui ont indiqué une occurrence de grêle afin de nous focaliser sur ces données et calculé le nombre de données manquantes. En dehors des données concernant la température, nous avions alors environ 70% des stations

météorologiques touchées par la grêle qui ne remontent pas de mesures sur la vitesse ou la direction du vent par exemple.

3.3.4 GSOD

Une autre base de données intéressante est le GSOD, *global Surface Summary of the Day* qui nous donne des informations météorologiques de manière quotidienne avec une couverture qui se veut mondiale contrairement à Météo-France dont les données se focalisent en France. Ces informations sont issues d'une autre base de données ISD, *Integrated Surface Dataset*, qui intègre des intègre des observations météorologiques de plus de 20 000 stations météorologiques[1]. Les données de ISD sont fournies par le NCEI *The National Centers for Environmental Information*, une agence gouvernementale américaine, qui gère de vastes archives de données atmosphériques, côtières, géophysiques et océaniques à l'échelle mondiale.[2]

GSOD compile ensuite ces différentes données et met à disposition des utilisateurs sans frais les données de plus de 9 000 stations météorologiques à travers le monde, bien que la couverture spatiale soit particulièrement dense en Amérique du Nord, en Europe, en Australie et dans certaines parties de l'Asie. En France spécifiquement, il y a un peu moins de 200 stations incluses dans GSOD, ce qui est bien moins que le nombre de stations météorologiques gérées par Météo-France. Cependant, ces données sont facilement requêtables sous R grâce à un paquet *GSODR* permettant d'obtenir les données des stations pour un pays et une année donnée facilement. Les données historiques remontent à 1929, avec une couverture plus complète à partir de 1973. Nous nous intéresserons particulièrement aux données à partir de 2020.

Chaque station rapporte différentes données météorologiques quotidiennes, incluant la température moyenne, la pression au niveau de la mer, la vitesse et la direction du vent, ainsi que des indicateurs d'occurrence de la grêle, de la pluie, de la neige, et des orages. Cela va ainsi permettre de comparer ces données à celles de Météo-France pour en vérifier la pertinence. Nous présentons dans la table 6 quelques-unes des variables principales que nous allons utiliser.

[1] https://www.ncei.noaa.gov/products/land-based-station/integrated-surface-database
[2] https://en.wikipedia.org/wiki/National-Centers-for-Environmental-Information

Nom paramètre	Détail
STN	Numéro Météo-France du poste sur 8 chiffres
LATITUDE	Latitude
LONGITUDE	Longitude
YEARMODA	Date de la mesure (année-mois-jour)
TEMP	Température moyenne sur la journée
MAX	Température maximale sur la journée
MIN	Température minimale sur la journée
WDSP	Vitesse Moyenne du vent
MXSPD	Vitesse du vent soutenue maximale enregistrée pendant la journée
GUST	Rafale de vent maximale enregistrée pendant la journée
PRCP	Maximum quotidien de la force maximale horaire du vent moyenné sur 10 minutes, à 10 m
R	Occurence de la pluie
H	Occurence de la grêle
T	Occurence d'orage

Table 6: Paramètres des données météorologiques de GSOD

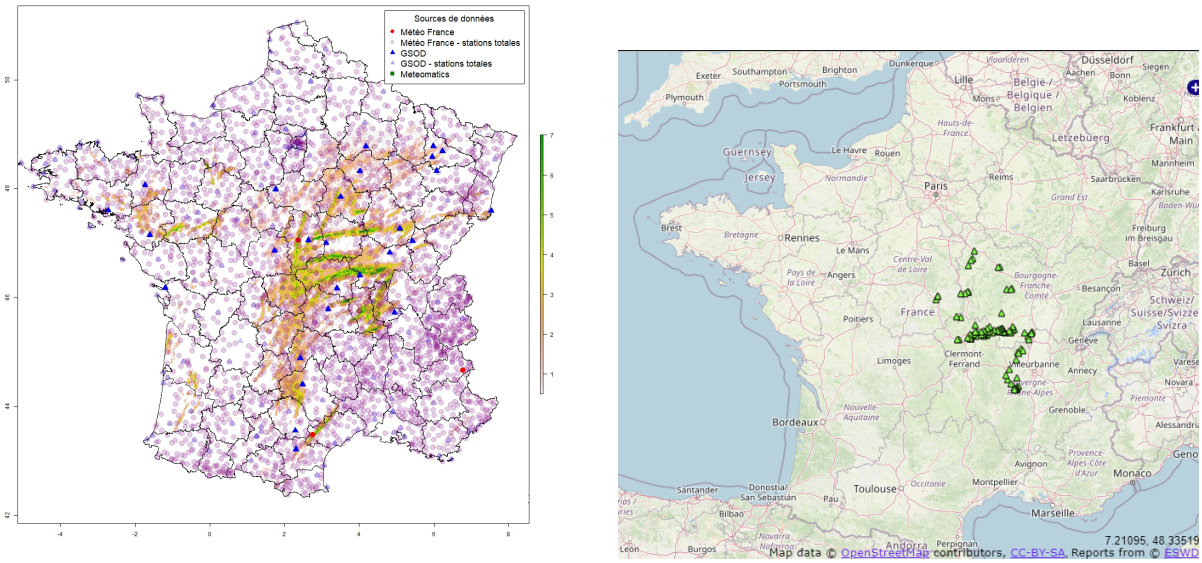
3.3.5 Comparaison des données

Maintenant que nous avons ces différentes bases de données en notre possession, nous allons pouvoir les comparer pour étudier leur pertinence dans le cadre de notre mémoire. Nous cherchons plus particulièrement à identifier les données qui indiquent clairement l'occurrence de grêle pour des évènements que nous savons s'être produits.

Chacune des bases de données ont leurs défauts. Par exemple, GSOD a certes une couverture mondiale, mais ne possède moins de 200 stations en France. Ainsi, nous pourrions nous attendre à une détection plus rare des évènements de grêle. Pour Météo-France, le nombre de stations météorologiques est certes 10 fois plus élevée, mais le caractère très localisé de la grêle pourrait rendre plus complexes la détection de ces évènements par les radars. De plus GSOD tout comme Météo-France n'indiquent que l'occurrence ou non des évènements. Ainsi, si les stations météorologiques détectent 2 évènements de grêle la même journée, la variable de grêle n'indiquera que «1 »pour l'occurrence de grêle.

Concernant les données d'ESSL, les données provenant d'observations de terrain, cela dépend fortement de la présence d'observateur sur les lieux et au moment de la grêle ce qui peut induire une sous-estimation des grêles sur certains endroits. Enfin, les données Meteomatics sont issues d'une modélisation réalisée par ses équipes. Ainsi, il est possible d'avoir des évènements plus ou moins sous-estimés.

Intéressons-nous à la tempête de grêle QIARA de juin 2022 pour comparer les données. Nous allons de plus nous concentrer sur le 21 juin 2022 pour étudier la pertinence des données quotidienne. Nous affichons le comparatif des cartes 25.



Graphique 24: Données Météo-France, GSOD et Météomatics (gauche) et ESSL (droite)

Cette comparaison nous révèle quelques analyses intéressantes et surprenantes. Nous observons d'après ces cartes que le département de l'Allier, en région Auvergne-Rhône-Alpes a été touché par cet évènement de grêle de manière importante avec une empreinte de grêle sous la forme de «griffes»appelées *hail streak* en anglais.

Sur la carte à gauche, nous avons l'ensemble des stations Météo-France affichées selon des points transparents et violets. Lorsque ces dernières indiquent une occurrence de grêle, les points sont alors affichés en rouge. Nous avons le même système avec GSOD mais cette fois-ci avec des triangles bleus. Nous n'observons que 3 points rouges, le premier vers le centre de la France dans le Cher, le second dans les Hautes-Alpes et le troisième dans le sud dans l'Hérault pour Météo-France. L'occurrence de grêle dans le Cher semble être corroboré avec les informations de GSOD, Meteomatics et ESSL. Cependant, le deuxième point semble particulièrement surprenant car les autres bases n'indiquent aucune activité de grêle, tandis que seul Meteomatics semble indiquer la présence de grêle en plus de Météo-France pour l'Hérault, bien que les stations de GSOD situées vers la frontière de l'Aude et du Tarn indiquent également l'occurrence de grêle.

Nous constatons ainsi la difficulté de l'observation des évènements de grêle. Il est également étrange que les stations de GSOD relèvent davantage de grêles que Météo-France. Nous pensons que les instruments de grêle ayant un coût élevé, toutes les stations météorologiques n'en sont pas forcément équipées.

Les données issues de GSOD, ESSL et Meteomatics nous semblent être plus perti-

nentes. Cependant, nous estimons que les stations de GSOD ne sont pas assez nombreuses pour assez bien couvrir le territoire français, nous risquerions de sous-estimer l'impact de la grêle. Nous allons donc poursuivre l'étude avec les données de Meteomatics et d'ESSL

3.4 Aléa, vulnérabilité et risque

Il est important de distinguer les termes d'aléa, de vulnérabilité et de risque qui couvrent des notions distinctes. La publication géographique Géoconfluences explique ainsi qu'un « risque est la possibilité qu'un aléa se produise et touche une population vulnérable à cet aléa » [18]. Dans le magazine n°53 de l'Actuariel de juin 2024, Myriam MERAD, directrice de la recherche au CNRS, prend l'exemple simple d'un pot de fleurs d'un immeuble qui tombe sous la force du vent. La chute du pot de fleurs représente un aléa. Cependant, cet aléa ne devient un risque que si une personne se trouve à l'endroit de la chute du pot de fleur à ce moment-là et qu'elle ne s'était pas préparée à la chute. Si la personne est assez agile pour éviter le pot après l'avoir vu tomber, il n'y a pas de risque. Elle continue l'explication en explicitant que le risque ici, est le risque d'invalidité ou de mort de la personne vulnérable exposée à l'aléa de la chute du pot de fleurs [1].

Revenons en détail sur ces différents termes.

L'**aléa** fait référence à un phénomène qui va être plus ou moins probable sur une zone et pour une période donnée. Cet aléa peut être caractérisé par différents paramètres comme sa nature, sa fréquence, son intensité, sa durée et sa localisation. Pour la grêle, l'aléa peut être ainsi quantifié par des paramètres tels que la taille des grêlons, la durée de l'épisode de grêle, la fréquence de ces derniers ainsi que les communes ou départements touchés. Les bases de données citées dans cette partie à savoir Meteomatics, Météo-France, GSOD et ESSL sont ainsi des données permettant de caractériser l'aléa de la grêle.

La **vulnérabilité** représente la sensibilité des enjeux exposés en cas d'occurrence de l'aléa. Dans le cadre de la grêle, cette vulnérabilité peut provenir d'une zone avec une densité de population particulièrement élevée qui peut accroître le nombre potentiel de victimes, ou les caractéristiques des biens exposés comme les bâtiments qui ont des toits qui offrent une résistance différente à la grêle selon les matériaux employés, les véhicules et leur type de carrosseries par exemple. Ces éléments peuvent ainsi venir augmenter ou diminuer la vulnérabilité face à l'aléa de grêle.

Enfin, le **risque**, ici naturel comme nous étudions la grêle, résulte du croisement entre

l'aléa et la vulnérabilité :

$$\text{Risque} = \text{Aléa} \times \text{Vulnérabilité} \tag{1}$$

Ainsi, il sera important de croiser les données d'aléa à d'autres sources de données permettant d'obtenir la vulnérabilité d'AXA face à l'aléa de grêle afin d'obtenir finalement une mesure des sinistres liés au risque de grêle.

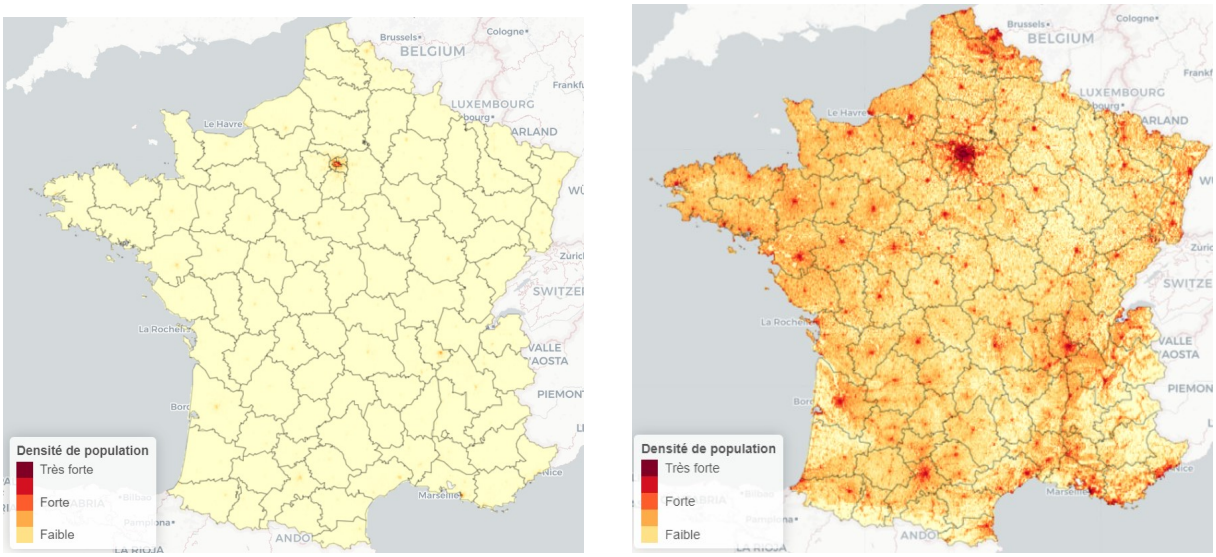
3.5 Données de la population

Afin de rajouter des données sur la vulnérabilité, nous avons décidé d'utiliser des données de la densité de la population fournie par le Global Human Settlement Layer (GHSL) qui est un projet initié par le Centre Commun de Recherche (JRC) de la Commission Européenne. Le GHSL vise à cartographier et à suivre l'évolution des zones habitées et construites. Cette base de données combine les informations des recensements avec celles sur la surface bâtie. Cette approche permet de prendre en compte la diversité d'occupation des territoires étudiés, en associant les données démographiques à l'emprise au sol des constructions. Cela résulte en un maillage précis du nombre d'habitants par maille d'une résolution assez fine de 1 km de côté. [1] [2]

Dans ce mémoire, nous allons nous appuyer sur les données de la population de 2020. La France présente des variations significatives en termes de densité de population. Par exemple les villes de Paris, Marseille ou Lyon sont beaucoup plus densément peuplées que d'autres villes. Cette hétérogénéité peut rendre difficile l'interprétation et la modélisation des données, car les grandes valeurs de densité dans les zones urbaines dominantes peuvent masquer les tendances et les risques dans les zones moins denses. Pour faire face à ce sujet, nous avons décidé de transformer les données pour réduire l'effet des grandes valeurs par l'application du logarithme. Nous affichons ainsi les deux cartes, avec et sans transformation dans la figure 25.

[1] https://data.jrc.ec.europa.eu/collection/ghsl

[2] https://artificialisation.developpement-durable.gouv.fr/bases-donnees/global-human-settlement-layer-ghsl



Graphique 25: Carte de la densité de la population de la France en 2020 (gauche), au logarithme de la densité (droite)
Source : global Human Settlement Layer

Ainsi, en combinant les données d'aléa avec les données de la population, nous allons pouvoir identifier les zones où la population est la plus exposée aux risques de grêle et mieux estimer les dommages potentiels.

4 Cadre théorique

4.1 Mesures de risques

4.1.1 Les quantiles

Dans le cadre de l'analyse des données de grêle, les quantiles sont des outils statistiques essentiels pour capturer les aspects de la distribution des données. Le quantile est un seuil qui ne sera dépassé que dans une proportion fixée de cas. Ainsi, la fonction faisant correspondre ce seuil à la proportion fixée est la fonction quantile. Par la suite, nous allons nous appuyer sur les définitions élaborées par M. Denuit et A. Charpentier (2004) [12].

Le quantile d'ordre α de la variable aléatoire X noté q_α est défini de la manière suivante :

$$q_\alpha = F_X^{-1}(\alpha) = \inf\{x \in \mathbb{R} \mid F_X(x) \geq \alpha\}, \quad \alpha \in [0,1]$$

Avec la fonction F_X^{-1}, la fonction quantile qui est associée à la fonction de répartition F_X. Certains quantiles ont également des noms spécifiques dans le langage courant. Ainsi, lorsque $\alpha = \frac{1}{2}$, nous parlons de médiane, et lorsque $\alpha = \frac{1}{4}$ ou $\alpha = \frac{3}{4}$, nous parlons

respectivement de premier quartile ou troisième quartile.

Dans le cadre de mesure de risque, la notion de quantile est utilisée sous l'appellation de *Value-At-Risk*. Pour un risque X et un niveau de probabilité α, où $\alpha \in [0,1]$, la VaR correspondante est notée $VaR[X,\alpha]$ et représente donc le quantile d'ordre α de X. Nous notons :

$$VaR[X,\alpha] = q_\alpha = F_X^{-1}(\alpha), \quad \alpha \in [0,1]$$

Dans le contexte de la grêle, ces quantiles vont ainsi nous permettre de mieux comprendre la distribution des diamètres de grêle et déterminer la proportion de grêlons qui va dépasser un certain seuil de manière plus robuste que les statistiques classiques comme la moyenne qui dans la majorité des cas nous sera inexploitable en raison de la prépondérance des valeurs nulles dans nos données.

4.1.2 Tail Value-at-Risk (TVaR)

Le quantile d'ordre α indique la taille des grêlons qui ne sera dépassée que dans au plus $1-\alpha$ % des cas. Cependant, il ne fournit aucune information sur les grêlons qui se produisent lorsque cette taille est dépassée. Or cette notion pourrait être primordiale pour comprendre les grêlons susceptibles d'occasionner des dommages importants.

C'est dans ce cadre-là que la Tail Value-at-Risk (TVaR) fait son apparition. La TVaR est une mesure de risque qui se concentre sur les tailles de grêlons attendues dans les cas les plus défavorables, c'est-à-dire au-delà d'un certain seuil fixé. Mathématiquement, la TVaR au seuil α (exprimé en pourcentage, par exemple 75%, 95% ou 99%) est définie comme la moyenne des pertes qui excèdent le quantile ou encore la Value-at-Risk au même seuil α:

$$\mathrm{TVaR}(X;\alpha) = \frac{1}{1-\alpha}\int_\alpha^1 \mathrm{VaR}(X;p)\,dp$$

La TVaR est souvent associée avec une autre mesure, appelée la *Conditionnal Tail Expectation*. Cette mesure au niveau de probabilité α et notée $CTE[X,\alpha]$ représente dans notre contexte la taille de grêlon attendue sachant que la VaR au seuil α est dépassée. Cela donne mathématiquement :

$$\mathrm{CTE}(X;\alpha) = E[X|X > \mathrm{VaR}(X;\alpha)]$$

Les auteurs parlent ainsi du concept de «perte moyenne dans les pires $1-\alpha$% des cas»ou encore du diamètre des grêlons dans les pires $1-\alpha$% des cas dans notre contexte d'estimation du comportement de la grêle. De plus, a CTE et la TVaR sont équivalentes

pour des risques dont la fonction de répartition est continue.

$$\text{CTE}(X;\alpha) = \text{TVaR}(X;\alpha), \quad \alpha \in [0,1]$$

Ainsi, ces métriques se lisent comme étant la moyenne des VaR d'un seuil supérieur à α. Par la suite, nous n'utiliserons que les termes de quantiles et de TVaR.

Cependant, dans le cadre de nos données de diamètres de grêle modélisés par Meteomatics, nous ne souhaitons pas avoir des métriques qui soient trop sensible aux valeurs extrêmes en cas d'erreur dans la modélisation. En effet, si la majorité des diamètres de grêles d'une liste de données sont comprises entre 1 et 3 avec une seule valeur à 5 cm de diamètre par exemple, nous risquerions sur surestimer les impacts de la grêle si ce cas est isolé. Ainsi, nous décidons de développer en plus de la TVaR, une nouvelle métrique appelée TVaR_med, qui s'appuie cette fois-ci sur la médiane des VaR d'un seuil supérieur à α. Il sera alors intéressant de comparer les différences de performances des indicateurs réalisés à partir de la moyenne (TVaR_mean par la suite) et de la médiane (TVaR_med).

Mathématiquement, si X est une variable aléatoire représentant les diamètres de grêle, et $\text{VaR}(X;\alpha)$ est la VaR au niveau α, la médiane conditionnelle $\text{TVaR_med}(X;\alpha)$ est donnée par:

$$\text{TVaR_med}(X;\alpha) = \text{Median}(X|X > \text{VaR}(X;\alpha))$$

Cette nouvelle TVaR_med est alors la valeur x telle que :

$$F_{X|X>\text{VaR}(X;\alpha)}(x) = 0,5$$

Cela signifie que 50% des pertes qui excèdent la VaR au niveau α seront inférieures ou égales à $\text{TVaR_med}(X;\alpha)$.

4.2 Modèles linéaires généralisés

Le modèle linéaire généralisé ou encore Generalized Linear Model (abrégé GLM par la suite) est un modèle statistique qui provient d'une généralisation du modèle linéaire et qui permet d'analyser et de quantifier la relation entre une variable réponse et plusieurs variables explicatives notamment utilisé en assurance non-vie. Dans cette partie, nous allons nous appuyer sur le cours d'assurance dommage de N. Baradel (2024)[8]. Pour

comprendre l'intérêt d'utiliser des GLM, repartons du modèle linéaire.

$$Y = \sum_{i=1}^{n} \beta_i X_i + \epsilon \tag{2}$$

Avec :

- Y étant la variable réponse du modèle
- X_i représentant les variables explicatives ou encore appelés prédicteurs
- β_i étant les paramètres du modèles qui ne sont pas observés et qui sont donc à modéliser
- ϵ étant le terme d'erreur

Le modèle linéaire repose sur la distribution normale du terme d'erreur $\epsilon \sim \mathcal{N}(0, \sigma^2)$ qui est donc l'écart entre les valeurs de Y observées et les valeurs $\hat{Y}$ prédites par le modèle

$$\epsilon = Y - \hat{Y} = Y - \sum_{i=1}^{p} \beta_i X_i \tag{3}$$

De plus, les modèles linéaires reposent sur l'hypothèse d'une relation linéaire entre l'espérance de la variable à expliquer et les variables explicatives, ce qui peut être très contraignant, notamment dans le cas de l'étude de l'impact des variables climatiques sur la sinistralité. Passer aux modèles linéaires généralisés permet à la variable réponse Y d'avoir une distribution d'erreur bien plus large, provenant de la famille de distribution exponentielle que nous expliciterons par la suite. Cette famille inclut par exemple la distribution normale, de Poisson, binomiale ou encore gamma.

Les modèles linéaires généralisés intègrent également une fonction de lien qui relie l'espérance de la variable réponse Y à une combinaison linéaire des variables explicatives X_i.

$$g(\mathbb{E}[Y]) = \beta_0 + \sum_{i=1}^{p} \beta_i \cdot X_i \tag{4}$$

Nous avons ainsi 3 composants dans un GLM :

- Y qui est la variable réponse et qui est une composante aléatoire suivant une loi de probabilité issue de la famille exponentielle.
- $(X_i)_{1 \leq i \leq p}$, qui sont les variables explicatives et qui sont des composantes déterministes.

- $g(.)$, la fonction de lien représentant la relation entre l'espérance de Y et les variables explicatives $(X_i)_{1\leq i\leq p}$.

Composantes du GLM

Variable réponse

Soit Y, la variable aléatoire et $(y_i)_{1\leq i\leq n}$ les observations de Y. La loi de Y doit appartenir à la famille exponentielle. La loi de Y est de la famille exponentielle si cette famille est dominée et que sa densité peut s'écrire sous la forme

$$f(y,\theta,\phi) = \exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right] \tag{5}$$

avec α, b et c qui sont des fonctions mesurables, θ est appelé paramètre naturel et ϕ est le paramètre de dispersion.

Pour Y, variable aléatoire, nous avons les résultats suivants:

$$\mathbb{E}[Y] = b'(\theta) \tag{6}$$
$$\mathbb{V}ar(Y) = b''(\theta)a(\phi) \tag{7}$$

Regardons maintenant un exemple dans le cas de la loi normale qui fait partie de la famille exponentielle. Elle s'écrit :

$$f(y;\theta,\phi) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(y-\mu)^2}{2\sigma^2}} = \exp\left(\frac{\mu y - \frac{\theta^2}{2}}{\sigma^2} - \frac{y^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\right). \tag{8}$$

On pose la fonction de lien canonique $\theta = \mu$, la densité se réécrit alors de la manière suivante :

$$f(y;\theta,\phi) = \exp\left(\frac{\theta y - \frac{\theta^2}{2}}{\sigma^2} - \frac{y^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\right). \tag{9}$$

On a alors : $b(\theta) = \frac{\theta^2}{2}$, $a(\phi) = \sigma^2$ *et* $c(y,\phi) = -\frac{y^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)$. On pose alors $\phi = \sigma^2$. La table 7 reprend ainsi les paramètres des lois usuelles.

Distribution	$\boldsymbol{\theta}$	$\boldsymbol{b(\theta)}$	$\boldsymbol{a(\phi)}$	$\boldsymbol{\mu}$
Bernoulli $\mathcal{B}(1,p)$	$\ln\left(\frac{p}{1-p}\right)$	$\ln(1+\exp(\theta))$	1	$p=\frac{\exp(\theta)}{1+\exp(\theta)}$
Poisson $\mathcal{P}(\lambda)$	$\ln(\lambda)$	$\exp(\theta)$	1	$\exp(\theta)=\lambda$
Normale $\mathcal{N}(\mu,\sigma^2)$	μ	$\frac{\theta^2}{2}$	σ^2	$\theta=\mu$
Gamma $\mathcal{G}(\nu,\mu)$	$-\frac{1}{\mu}$	$-\ln(-\theta)$	$\frac{1}{\nu}$	$-\frac{1}{\theta}=\mu$

Table 7: Paramètres des lois usuelles

Variable explicatives

Nous avons X, les observations des variables explicatives à valeurs dans $\mathbb{R}^d$. Avec $\beta_0 \in \mathbb{R}$ et $\beta \in \mathbb{R}$, le prédicteur linéaire qui est la composante déterministe du modèle est donnée par

$$\eta = \beta_0 + X'\beta \tag{10}$$

Fonction de lien

De plus, nous supposons qu'il existe une fonction g bijective telle que :

$$g(\mathbb{E}(Y|X)) = \eta = \beta_0 + X'\beta \tag{11}$$
$$\mathbb{E}(Y|X) = g^{-1}(\eta) = g^{-1}(\beta_0 + X'\beta) \tag{12}$$

Cette fonction g est appelée fonction de lien. La difficulté principale des modèles linéaires généralisés est ainsi de correctement estimer β_0 et β.

Calcul des paramètres β_0 et β par la méthode du maximum de vraisemblance

Nous estimons les paramètres β_0 et β par la méthode du maximum de vraisemblance. Nous considérons $(y_i)_{1\leq i\leq n}$, les observations des variables aléatoires indépendantes des $(Y_i)_{1\leq i\leq n}$ ainsi que les $(x_i)_{1\leq i\leq n}$, les observations des variables aléatoires indépendantes des $(X_i)_{1\leq i\leq n}$. Nous supposons également que $Y|X$ suit une fonction de densité issue de la famille exponentielle. Dans le cas de la fonction de lien canonique, la vraisemblance s'écrit de la manière suivante :

$$\mathcal{L}(y|x;\beta_0,\beta;\phi) = \prod_{i=1}^{n} f(y_i;\beta_0,\beta,\phi) = \prod_{i=1}^{n} f(y_i;\beta_0+x_i'\beta,\phi) \tag{13}$$

La log-vraisemblance s'obtient ainsi de la manière suivante :

$$\ell(y|x;\beta_0,\beta;\phi) = \sum_{i=1}^{n} \log(f(y_i;\beta_0 + x_i'\beta,\phi)) \tag{14}$$

$$= \sum_{i=1}^{n} \left[\frac{y_i(\beta_0 + x_i'\beta) - b(\beta_0 + x_i'\beta)}{a(\phi)} + c(y_i,\phi) \right] \tag{15}$$

Les équations du premier ordre du problème sont les suivantes :

$$(\hat{\beta}_0, \hat{\beta}) = \arg \max_{(\beta_0,\beta)\in\mathbb{R}^{d+1}} \ell((y_i \mid x_i)_{1\leq i\leq n}; \beta_0, \beta) \tag{16}$$

Elles ne nous permettent pas d'obtenir une solution explicite dans la plupart des cas. Il est donc nécessaire d'utiliser des méthodes numériques afin d'estimer le maximum de vraisemblance. Nous allons utiliser la fonction la fonction *glm* de R pour la suite de ce mémoire.

4.3 Méthodes de Machine Learning

Le *Machine Learning* en anglais, apprentissage automatique en français, est une branche de l'intelligence artificielle qui s'appuie sur des approches statistiques et mathématiques pour apprendre et s'améliorer à partir de l'expérience sans être explicitement programmé. Nous distinguons 4 types d'algorithme d'apprentissage : l'apprentissage supervisé, l'apprentissage non-supervisé, l'apprentissage semi-supervisé et l'apprentissage par renforcement. Dans le cadre de ce mémoire, nous n'aborderons que l'apprentissage supervisé qui nous sera utile par la suite. L'apprentissage supervisé implique l'entraînement d'un modèle sur ensemble de données comprenant des variables explicatives, usuellement notées $X_1, X_2, ...X_p$ pour p variables explicatives ainsi que d'une variable cible, souvent notée Y. L'objectif étant de prédire Y à partir des différentes variables explicatives.

Nous allons désormais passer en revue quelques outils qui nous permettront de comprendre la modélisation des données avec l'algorithme *XG Boost*.

4.3.1 Les arbres de décision

L'algorithme CART (*Classification And Regression Trees*) de BREIMAN (1984)[9] est une technique importante pour construire des arbres de décision. Ces arbres de décision permettent de partitionner de manière récursive l'espace des données en sous ensembles qui seront plus petits et semblables en fonction des caractéristiques des données. Un arbre de décision consiste ainsi en une suite de partitions de plus en plus fines de tous les individus que nous observons.

Un arbre de décision peut traiter des problèmes de régression mais aussi des problèmes de classification. Les problèmes de régression visent à prédire une variable cible de nature quantitative, un montant de sinistre par exemple tandis qu'en classification, nous allons entraîner un modèle qui va attribuer une étiquette de classe à la variable cible de manière à prédire cette étiquette pour de nouvelles variables explicatives. Cela peut être utile si nous souhaitons modéliser par exemple la survenance ou non de la grêle ou encore une classification des départements sinistrés que nous allons voir par la suite.

A chaque étape correspondant à un nœud, l'algorithme choisit une variable explicative X_i ainsi qu'une règle selon un seuil s et sépare les données en deux partitions selon les valeurs de X par rapport à s. Ainsi, nous aurons après la première séparation, deux groupes : le groupe de gauche si $X_i \leq s$ et le groupe de droite si $X_i > s$.

Le choix de la variable à utiliser X_i et le seuil s se réalise par un algorithme dit «glouton»ou *greedy* en anglais qui va parcourir l'ensemble des variables explicatives X_i ainsi que tous les seuils de manière à ce que la déviance calculée à chaque nouvelle partition soit minimale. Lorsqu'il n'est plus possible de séparer les observations avec une nouvelle règle, ce nœud devient alors le nœud terminal et est appelé feuille. Nous associons une valeur $\hat{y}$ à chacune des feuilles de l'arbre qui consiste en la moyenne des y de la feuille dans le cadre d'un problème de régression ou alors la classe majoritaire de la feuille dans le cas d'un problème de classification. Ainsi, pour obtenir une prédiction $\hat{y}$ de nouvelles variables explicatives, il nous suffira de suivre les nœuds de l'arbre en respectant les différentes règles et de relever la valeur issue de la feuille sur laquelle nous parvenons.

Différentes mesures de la déviance sont possibles pour mesurer l'erreur. Dans le cadre d'une régression, nous pouvons chercher à minimiser la somme des écarts entre la valeur observée y et la valeur moyenne $\hat{y}$ des nœuds obtenus après application de la règle, encore appelée minimisation de l'erreur des moindres carrés.

$$Deviance = \sum_{y \in \text{groupe gauche}} (y - \hat{y}_{\text{gauche}})^2 + \sum_{y \in \text{groupe droite}} (y - \hat{y}_{\text{droite}})^2 \tag{17}$$

avec $\hat{y_{droite}}$, la moyenne des observations du groupe gauche et $\hat{y_{droite}}$, la moyenne des observations du groupe droite.

Dans le cas d'une classification à K classes, de nombreuses métriques comme le critère

de Gini qui mesure l'impureté d'un nœud, ou encore appelé le degré d'inhomogénéité.

$$\text{Critère de Gini} = n_{gauche} \sum_{k=1}^{K} p_{k,Gauche}(1 - p_{k,Gauche}) + n_{droite} \sum_{k=1}^{K} p_{k,Droite}(1 - p_{k,Droite}) \tag{18}$$

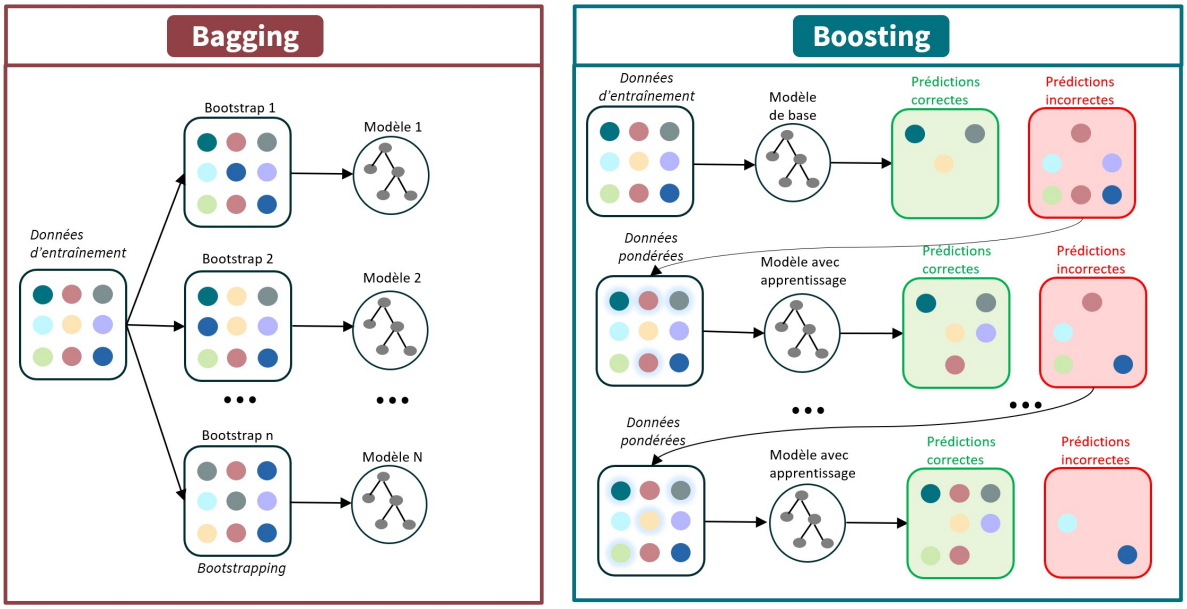
avec n_{gauche}, le nombre d'individus dans le groupe de gauche, n_{droite}, le nombre d'individus dans le groupe de droite, $p_{k,Gauche}$, la proportion d'individus de classe k dans le groupe de gauche et $p_{k,Droite}$, la proportion d'individus de classe k dans le groupe de droite. En cas de proportion parfaitement prédite, le critère de Gini sera alors égal à 0 et nous minimisons alors la déviance.

Ensuite, il est possible de fixer des paramètres afin d'éviter un sur-apprentissage ou augmenter la vitesse d'exécution de l'algorithme. Par exemple, nous pouvons fixer le nombre d'individus minimal à l'intérieur de chaque feuille ou encore limiter la profondeur de l'arbre correspondant au nombre maximal de nœud possible.

4.3.2 Gradient Boosting

Le *Gradient Boosting* est une approche adaptative, construisant un modèle prédictif en combinant séquence par séquence des modèles avec des pouvoirs prédictifs plus faibles, appelés en anglais *weak learners*, afin d'obtenir d'obtenir un modèle final avec un pouvoir prédictif plus fort. Cette approche se distingue des approches dites ensemblistes, comme dans les algorithmes de *bagging* ou des forêts aléatoires, *random forest*. Ces derniers s'appuient sur différents modèles prédictifs qui vont fournir différents résultats qui seront agrégés par la suite par une moyenne simple, une moyenne pondérée ou avec la majorité des prédictions par exemple.

L'approche séquentielle par le *Gradient Boosting* permet de construire à chaque étape un nouveau modèle qui corrige les erreurs précédentes.



Graphique 26: Méthode du bagging et du boosting
Inspiré de A. CHAWLA (2023)[1]

Dans le cadre de notre mémoire, nous allons nous appuyer sur la méthode du *Gradient Boosting Machine* présenté par FRIEDMAN (1999)[17]. Nous considérons ainsi un vecteur de variables explicatives $x = x_1, x_2, ..., x_p$ ainsi que de la variable cible y. L'objectif sera d'obtenir une estimation de la fonction optimale $F(\hat{x})$ qui relie x et y de manière à minimiser l'espérance d'une fonction de perte spécifique notée $\mathcal{L}(y, F(x))$.

$$\hat{F} = \arg\min_{F} \mathbb{E}_{y,x} \left[\mathcal{L}(y, F(x))\right] \tag{19}$$

avec $\mathcal{L}$, différentiable par rapport à F. Dans l'article, plusieurs fonctions de perte sont citées dont l'erreur quadratique pour un problème de régression ou encore le LogitBoost. Quelques fonctions classiques sont indiquées dans la table 8.

Type de problème	**Type**	**Fonction de perte $\mathcal{L}(y, F(x))$**
Classification	Perte 0-1	$\mathbb{I}(Y \neq f(X))$
	LogitBoost	$\log\left(1 + \exp(-2Y \cdot f(X))\right)$
Régression	$\mathbb{L}_2$ ou quadratique	$\frac{1}{2}(Y - f(X))^2$
	$\mathbb{L}_1$	$\lvert Y - f(X)\rvert$

Table 8: Exemples de fonctions de perte

[1]https://blog.dailydoseofds.com/p/an-animated-guide-to-bagging-and

L'algorithme commence ensuite par un modèle initial simple, puis ajoute séquentiellement de nouveaux modèles pour prédire les résidus du modèle, aussi appelés erreurs, combiné à chaque étape. Mathématiquement, pour un problème de régression, le modèle à l'itération m peut être exprimé comme :

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x)$$

où $F_{m-1}(x)$ est le modèle de l'itération précédente, $h_m(x)$ est le nouveau modèle faible ou *weak learner*, et γ_m est le pas d'apprentissage qui peut être fixé en paramètre. L'objectif est de minimiser une fonction de perte $\mathcal{L}(y, F(x))$ sur l'ensemble d'entraînement. À chaque itération, le *Gradient Boosting* ajuste $h_m(x)$ pour approximer le gradient négatif de la fonction de perte :

$$h_m(x) \approx -\left[\frac{\partial L(y, F(x))}{\partial F(x)}\right]_{F(x)=F_{m-1}(x)}$$

Maintenant que nous avons étudié le *Gradient Boosting*, voyons voir plus en détail le *XG Boost*, *eXtreme Gradient Boosting*. Le *XG Boost* est une implémentation du Gradient Boosting. Cet algorithme est reconnu pour sa vitesse et son efficacité du fait de quelques fonctionnalités particulières. Nous notons tout d'abord le fait que leur implémentation sous R ou sur Python utilise un calcul parallèle et utilise intelligemment la mémoire disponible sur l'ordinateur, ce qui améliore grandement la vitesse d'exécution de l'algorithme. De plus, il intègre une régularisation dans le processus d'apprentissage qui va nous permettre d'éviter un sur-apprentissage. Sous R, le package *XG Boost*, comprend la régularisation L1, aussi appelée régression de LASSO (*Least Absolute Shrinkage and Selection Operator*) utilisée par défaut et la régularisation L2, aussi appelée *Ridge regression*.

Dans le cadre du mémoire, nous allons avoir recours à l'algorithme *XG Boost* afin de modéliser différentes classes de sinistres pour d, un département, pour t, un évènement donné et et i un individu de l'échantillon : $C^i_{d,t} \in \{0, 1, 2, 3, 4, 5\}$. Nous volons minimiser les erreurs de classifications en prenant en compte la distance entre les classes observées et les classes modélisées. Pour cela, nous utilisons la fonction de perte appelée perte logarithmique qui va en effet pénaliser les mauvaises classifications. Elle est aussi appelée entropie croisée ou *log loss* en anglais. Le terme d'entropie est utilisé car elle permet de caractériser le niveau de désorganisation d'un système. Dans le cadre de notre modélisation multi-classes, elle nous permettra de comparer nos prédictions sont proches des observations. Plus les prédictions seront bonnes, plus le niveau de désordre sera faible impliquant une fonction de perte moindre.

Dans le cadre d'une classification à K classes avec ainsi une variable cible avec $y \in \{0, 1, ..., K\}$, nous avons la fonction de perte suivante :

$$\mathcal{L}(y_i, \hat{y}_i) = -\sum_{i=1}^{m}\sum_{j=1}^{K} \mathbb{I}_{y_i=j} \log(\hat{P}(y_i = j|x_i)) \tag{20}$$

Où

- y_i est la classe observée réelle de l'individu i,
- $\hat{y}_i$ est la classe prédite par le modèle pour l'individu i,
- m est le nombre d'individus dans l'échantillon,
- K est le nombre de classes,
- $\mathbb{I}_{y_i=j}$ est la fonction indicatrice qui vaut 1 si $y_i = j$ et 0 sinon,
- $\hat{P}(y_i = j|x_i)$ est la probabilité prédite que l'individu i appartienne à la classe j à partir des observations des variables explicatives x_i.

Ainsi, si nous essayons d'estimer $y_1 = 5$ et que la probabilité induite par la modélisation nous donne 100%, la fonction de perte vaut $-\log(1) = 0$ nous avons donc une perte minimale associée à cette prédiction, tandis que si la probabilité associée n'était que de 5% par exemple, dans ce cas nous rajouterions $-\log(0,05) = -(-1,3) = 1,3$ à la fonction de perte ce qui va pénaliser ce modèle. De la même manière, la fonction de perte nous donnera 2 pour une probabilité estimée à 1%. Plus la probabilité d'obtenir la valeur observée est improbable d'après le modèle, plus ce modèle sera fortement pénalisé par cette fonction de perte. Il s'agit ainsi d'une fonction de perte pertinente pour minimiser les erreurs de classification.

4.3.3 Mesures d'erreurs

Nous allons présenter dans cette partie quelques métrique qui vont nous permettre d'évaluer les différents modèles de ce mémoire. Nous allons nous focaliser sur le R^2, le RMSE et le MAE.

Coefficient de détermination R^2

Le coefficient de détermination R^2, encore appelé *R-Squared* en anglais permet de mesure la proportion de la variance de la variable cible Y expliquée par les variables explicatives X_i d'un modèle.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{21}$$

où y_i est la valeur observée de la variable cible, $\hat{y}_i$ est la prédiction obtenue par le modèle, $\bar{y}$ est la moyenne des valeurs réelles, et n est le nombre d'observations.

- Le R^2 varie généralement entre 0 et 1, où 1 indique que le modèle explique parfaitement la variance de la variable cible Y tandis qu'un R^2 proche de 0 signifie que le modèle ne capture pas bien la variance
- Un R^2 négatif indique que les prédictions sont moins bonnes que si l'on prédisait systématiquement la moyenne de la variable cible Y.
- Le R^2 permet ainsi de faciliter la comparaison entre différents modèles en normalisant l'erreur par rapport à la variance de la variable cible Y.
- Il est intéressant de comparer le R^2 sur l'échantillon d'entraînement d'un modèle à celui de l'échantillon de test. Un R^2 sur la base d'entraînement bien supérieur à celui de la base de test peut indiquer un cas de sur-apprentissage et un cas de sous-apprentissage le cas échéant

Erreur Quadratique moyenne - RMSE

Le RMSE (*Root Square Mean Error*) mesure l'écart moyen entre les prédictions et les valeurs observées, en termes de racine carrée de l'erreur quadratique moyenne. Cela permet ainsi de mesurer la dispersion des résidus et donc la concentration ou non des données modélisées des valeurs observées.Nous chercherons ainsi à utiliser les modèles qui minimisent cette métrique et indiquant ainsi un meilleur ajustement par rapport aux données observées.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{22}$$

Nous chercherons ainsi à utiliser les modèles qui minimisent cette métrique et indiquant ainsi un meilleur ajustement par rapport aux données observées.

Erreur Absolue moyenne - MAE

La MAE (*Mean Absolute Error*) est une métrique qui mesure l'écart moyen entre les valeurs observées et les valeurs prédites par le modèle en valeur absolue.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{23}$$

Comme dans le cas de la RMSE, nous chercherons à minimiser cette métrique.

5 Modélisation et résultats

Dans cette nouvelle partie, nous allons nous pencher au cœur de notre étude sur la modélisation des dommages causés par les évènements de grêle sur le portefeuille réassuré chez AXA à partir des bases de données présentées dans la section 3 et des modèles introduits dans la section 4. Notre approche se présente en cinq étapes clés :

1. L'agrégation des données Meteomatics
2. Modèle 1 : Classification des sinistres
3. Modèle 2 : Modélisation du taux de destruction
4. Modèle combiné : Croisement des modèles 1 et 2
5. Identification du type d'évènement

5.1 Agrégation des données Meteomatics

Les données fournies par Meteomatics se présentent sous la forme d'un raster. Un raster est un format de données géographiques représentant l'espace sous la forme d'une matrice de cellules organisées en lignes et en colonnes, représentant respectivement les latitudes et longitudes. Chaque cellule de cette matrice contient une valeur numérique qui dans notre étude est le diamètre de grêle prédit par Meteomatics. Ce format a un intérêt particulier pour les données météorologiques car il permet de représenter les surfaces avec des espacements réguliers ce qui s'avère très pratique dans le cadre d'analyse spatiales et de calculs sur de grandes zones géographiques tout en optimisant l'espace de stockage requis. Plus la taille de ces cellules est petite, plus la résolution est grande. Cette précision est visible dans le nombre de décimales présent dans les latitudes et longitudes. Pour la suite de notre étude, nous travaillons à une résolution de cellules 2 kilomètres par 2 kilomètres.

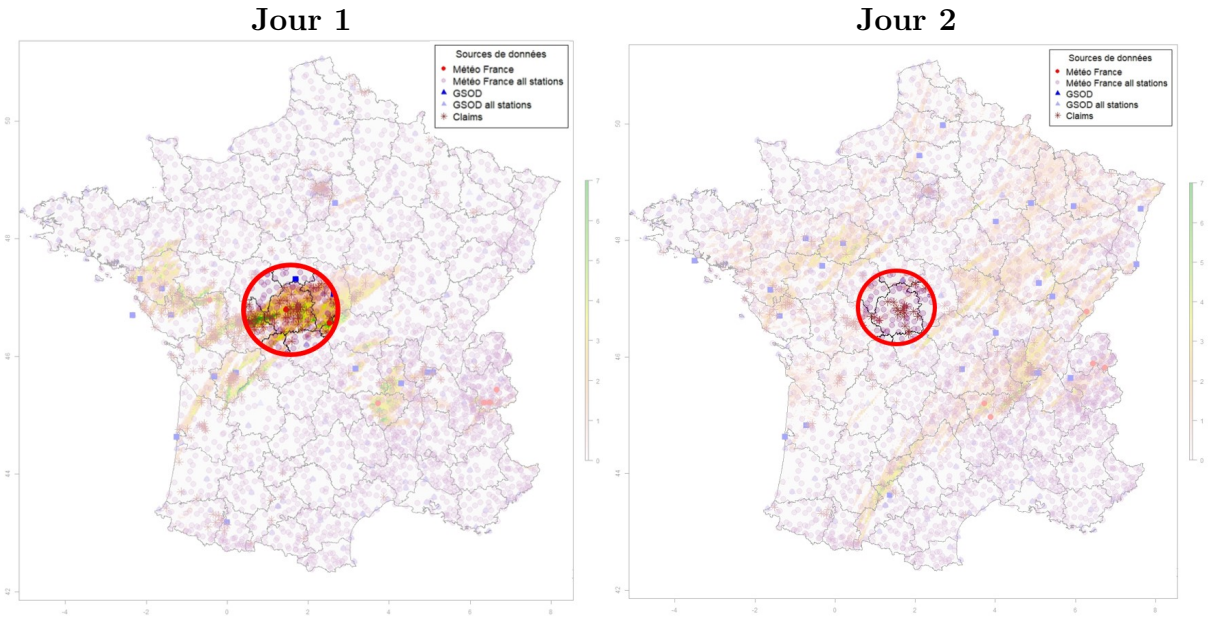
Une autre particularité de notre jeu de données réside dans sa dimension temporelle : pour un évènement de grêle s'étendant sur cinq jours par exemple, nous disposons de 120 rasters distincts, soit un par heure, permettant une observation quasi-continue de l'évolution du phénomène. Au total, pour une même cellule de 2 kilomètres par 2 kilomètres, nous avons alors 120 valeurs potentiellement différentes selon l'influence ou non de la grêle sur cette zone.

Ensuite, dans notre base de données des sinistres, nous possédons pour chaque sinistre, sa date d'occurrence ainsi que l'évènement de grêle auquel il est rattaché. Cette structure

de données nous offre deux axes d'analyse distincts pour les informations météorologiques fournies par Meteomatics :

- Une **analyse journalière**, permettant une étude fine de la sinistralité au jour le jour.
- Une **analyse par évènement**, offrant une vue plus globale de l'impact de chaque épisode de grêle.

L'approche journalière comporte cependant un souci lorsque l'on croise les données avec les données de la sinistralité. En effet, les déclarations ne sont pas totalement exactes et il peut y avoir des écarts de jours entre l'évènement de grêle ayant conduit à la sinistralité et la déclaration de ce sinistre de la part de l'assuré, en plus des erreurs possibles lors de la remontée des données au sein des bases. Ce cas de figure se voit lorsque nous visualisons les données de sinistralité avec les données d'occurrence de grêle modélisé par Meteomatics ainsi que quelques données issues des bases de données de Météo-France et GSOD. Certains jours alors que les occurrences de grêle sont avérées, nous n'avons pas de déclaration de sinistres, et ils arrivent soient avec un jour de décalage ou alors avec plusieurs jours. Dans ce cadre, faire une analyse journalière pourrait fausser les conclusions.



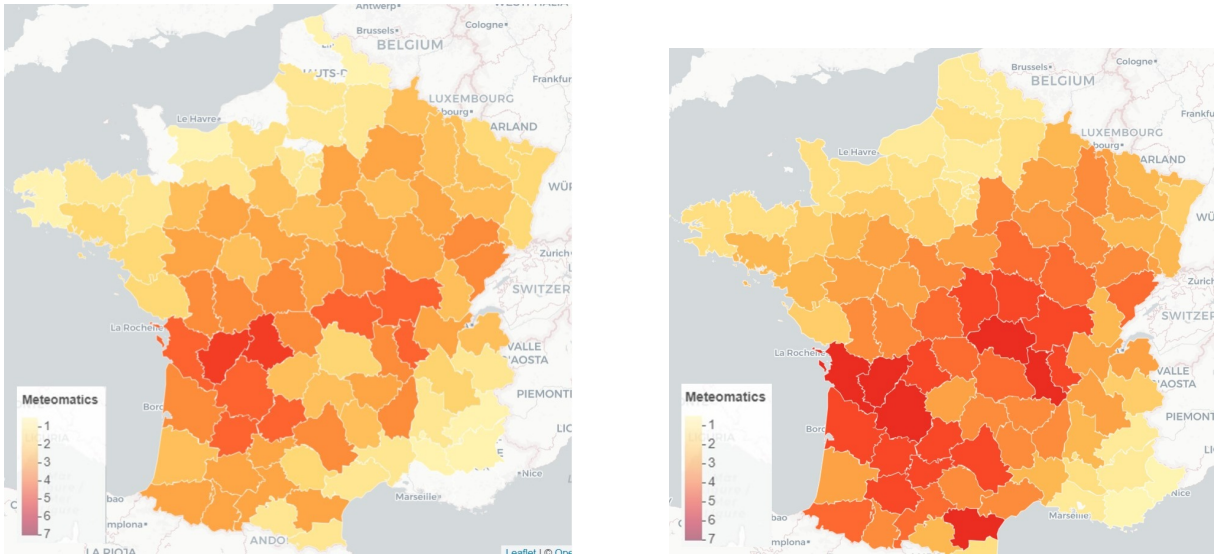
Graphique 27: Décalage entre déclaration et occurrence des évènements
Source : AXA, Meteomatics, Météo-France et GSOD

Nous constatons en effet sur ces images que des données de sinistres identifiées par des étoiles marrons ont été déclarées sur 2 jours relevés de l'évènement. Cependant, les indicateurs météorologiques fournis par Meteomatics qui sont sur une échelle de jaune à vert, vert pour les données les plus intenses ici, Météo-France ainsi que GSOD nous indique un évènement davantage survenu le premier jour. Ces derniers sont indiqués par des ronds rouges et des triangles bleu foncé. Lorsque les couleurs sont transparentes, cela indique la présence d'une station mais qu'elle n'a pas relevé d'évènements de grêle. Ces écarts peuvent également être dus à une grêle effectivement survenus sur les 2 jours mais pas assez intense pour être relevée par les radars le dernier jour. Ainsi, par prudence et pour prendre en compte les erreurs potentielles, nous décidons de poursuivre l'analyse des données au niveau de l'évènement et non au niveau journalier.

Ainsi, il est nécessaire de réorganiser les différents rasters associés à chaque évènement afin de les regrouper et de produire des données météorologiques agrégées pour chaque évènement et pour chaque département également pour être conforme aux données de notre base d'exposition et de sinistralité. Nous prendrons par la suite un exemple avec un évènement défini sur 5 jours, qui aurait ainsi besoin de 120 rasters pour chaque évènement.

5.1.1 Sens de l'agrégation

Maintenant que nous avons décidé sur quel pas de temps nous souhaitons réaliser l'analyse ainsi que la modélisation, deux choix s'offrent à nous. Nous pouvons dans un premier temps effectuer une agrégation spatiale, c'est-à-dire d'obtenir à partir des données Meteomatics d'une résolution de 2 km par 2 km, une donnée au niveau de chaque département, puis effectuer une agrégation au niveau des 120 fichiers horaires pour finalement obtenir une valeur par département et par évènement. Nous pouvons également faire l'inverse, à savoir d'abord agréger temporellement et ensuite agréger spatialement. Le sens d'agrégation a en effet son importance : en dehors de quelques agrégations simples comme le maximum, minimum ou la somme, les valeurs finales obtenues peuvent varier et refléter différentes caractéristiques de l'évènement.



Agrégation spatiale par le maximum puis agrégation temporelle par le quantile 90%

Agrégation spatiale par le quantile 90% puis agrégation temporelle par le maximum

Graphique 28: Importance du sens de l'agrégation des données Meteomatics
Source : Meteomatics

Nous observons en effet quelques différences notables entre les deux cartes de la France à partir des mêmes données Meteomatics dans le graphique 28. A gauche, nous avons d'abord pris le maximum de la valeur du diamètre de grêle au niveau de chaque département, puis nous prenons le quantile à l'ordre 90% sur les 120 fichiers horaires et à droite nous faisons l'inverse. Nous avons globalement des valeurs plus importantes à gauche du fait que nous prenons en premier lieu la valeur maximale au niveau du département. Cela signifie que si Meteomatics nous indique une zone de 2 km par 2 km avec une grêle de 3 cm de diamètre isolée, il s'agit tout de même cette valeur qui va être prise en compte alors que ce n'est pas forcément un bon indicateur de l'intensité de l'évènement. Tandis qu'en prenant tout d'abord le quantile à l'ordre 90%, nous nous assurons d'avoir plusieurs valeurs avec un tel diamètre de grêle pour éviter ces cas isolés. Cela se remarque particulièrement dans le Nord-Ouest ainsi que dans le Sud-Est de la France. Choisir la première carte de la figure 28 pourrait par exemple entraîner un montant plus important de sinistres modélisé car nous avons des signaux de grêle dans tous les départements. Après avoir exploré plusieurs combinaisons et exploré en détails les données, nous avons décidé d'élaborer en priorité des agrégations au niveau du département, et ensuite au niveau de l'évènement.

En effet, un épisode de grêle se manifeste sur des durées relativement courtes, de l'ordre de quelques minutes en général, voir plus longtemps dans certains cas d'évènements plus intenses ainsi que sur une échelle spatiale réduite, la grêle étant un phénomène davantage

local. Ainsi, une agrégation temporelle en premier lieu au niveau du raster pourrait produire des résultats peu représentatifs car les valeurs significatives seraient diluées par un grand nombre de cellules avec une valeur nulle signifiant une absence d'activité de grêle. Seules quelques analyses au niveau de quantiles de valeurs élevées pourraient être pertinentes dans ce cadre-là. En revanche, en agrégeant d'abord spatialement au niveau du département, il nous sera possible de capturer davantage les conséquences de la grêle au niveau d'un département, en prenant en compte par exemple la simultanéité des impacts de grêle dans différentes zones du département par exemple. Plusieurs communes peuvent en effet être affectées au même moment sur un département par un ou plusieurs orages de grêle. Le fait d'avoir plusieurs signaux de grêles au sein d'un département pourrait dans ce cas être révélateur d'un évènement de grêle d'une intensité conséquente et ainsi, être à l'origine d'un plus grand montant de sinistralité.

Ainsi, en réalisant une agrégation temporelle après avoir fait l'agrégation spatiale pourra alors fournir des informations plus précises sur les caractéristiques des évènements de grêle et mieux refléter la concentration des données dans certaines zones et améliorer la pertinences des indicateurs calculés. Concernant l'agrégation temporelle, nous allons privilégier des agrégations comme le maximum, la moyenne des valeurs les plus élevées par exemple.

Cependant, dans le cas où un ou plusieurs épisodes de grêles affecte une même zone, par exemple une ou plusieurs communes mais à des moments différents de la journée ou alors sur plusieurs heures consécutives, nous n'allons pas pouvoir retranscrire cette accumulation de grêle au niveau temporelle avec la méthode précédemment décrites. Bien que ces cas semblent moins fréquents, nous voulons nous laisser la possibilité de caractériser également ce cas de figures, qui peuvent être certes moins fréquents, mais certainement plus dévastateurs en termes de montant de sinistre.

Nous décidons ainsi d'élaborer en priorité plusieurs indicateurs qui agrègent d'abord spatialement, puis temporellement les modélisations de diamètres de grêle fournis par Meteomatics, tout en gardant quelques indicateurs qui vont effectuer l'agrégation dans le sens inverse.

5.1.2 Caractérisation des évènements de grêle

L'analyse précédente a permis de mettre le doigt sur l'importance de la compréhension des évènements de grêle afin d'élaborer des indicateurs pertinents. Maintenant, que nous

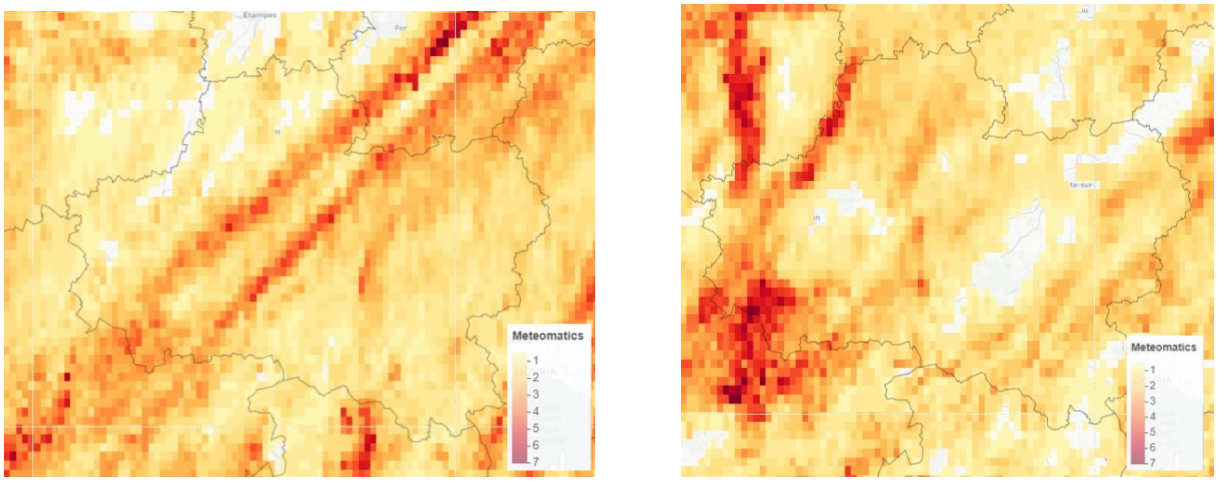
avons pris en compte le sens d'agrégation des données en fonction de certaines caractéristiques de la grêle, nous allons voir d'autres éléments qui méritent notre attention.

Nous pensons que des indicateurs comme la durée des épisodes de grêle, leur nombre ou encore la durée cumulée des différents épisodes de grêle peuvent être des indicateurs pertinents dans le cadre de notre modélisation. Nous décidons également de nous pencher davantage au travers d'indicateurs sur le nombre de cellules indiquant un diamètre de grêle compris entre 2 et 4 cm de diamètre et ceux d'un diamètre supérieur à 4 cm de diamètre pour observer des comportements qui peuvent être différents.

De plus, les données sur lesquels nous nous appuyons majoritairement dans notre analyse proviennent des modélisations de Meteomatics. Ces données nous semblent plutôt proches de la réalité en comparant avec les données de sinistres observées et en croisant les données avec Météo-France par exemple, mais elles peuvent tout de même introduire quelques erreurs de prédictions, avec potentiellement des valeurs isolées qui pourraient être élevées pouvant influencer nos analyses. Afin de garantir une plus grande robustesse des résultats, il sera important de s'intéresser aux quantiles qui offrent une représentation plus stable et moins sensible à ces valeurs extrêmes tout en prenant en compte le fait que la majorité des données est composée de valeurs nulles étant donné le caractère plus ou moins local des évènements.

Revenons quelques instants sur l'importance de cette localité. Les données Meteomatics nous fournissent des informations précieuses sur l'aléa à travers des informations sur la localisation de la grêle, sa fréquence et son intensité avec le diamètre de grêle modélisé. Afin de comprendre le risque de grêle pour AXA, il nous faut des informations également sur sa vulnérabilité face à la grêle, à savoir sa sensibilité face aux aléas de la grêle qui peut être influencée par de nombreux facteurs. Pour mieux comprendre cet enjeu, nous allons prendre le cas d'un département sur deux évènements différents.

Nous avons dans la figure 29 des représentations différentes pour un même département, celui du Loiret (45) sur deux évènements. Les données n'ont pas été agrégées au niveau du département car nous souhaitons garder la résolution initiale de 2 km par 2 km pour comprendre l'importance de la localité des évènements de grêle. Pour ce faire, nous avons pris la valeur maximale du diamètre de grêle sur le pas de temps horaire.



Graphique 29: Importance des caractéristiques de l'agrégation : Cas du Loiret du 19 au 23 juin 2022 (gauche) et du 02 au 06 juin 2022 (droite)

Nous observons alors une empreinte météorologique qui est totalement différente. En ce qui concerne l'intensité de l'évènement, nous remarquons dans les deux cas des diamètres de grêles qui atteignent des tailles conséquentes indiquées par la couleur rouge foncé. Nous pouvons même dire que la carte de droite semble plus intense que celle de gauche car ses couleurs dont davantage foncées dans le Sud-Ouest du département.

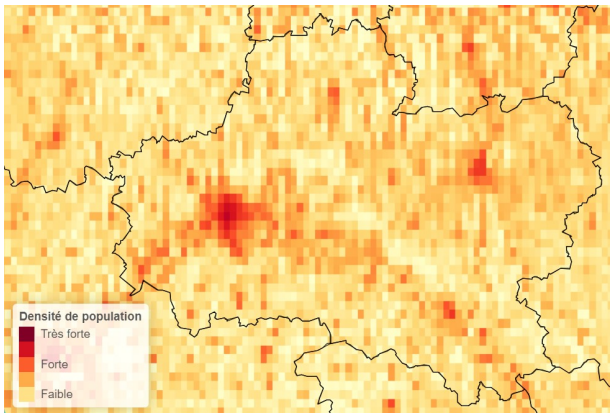
Ainsi, si nous prenons des quantiles élevés, l'agrégation au niveau du département pourrait nous donner une valeur plus élevée pour l'évènement de droite que celui de gauche. C'est ce que nous confirme le tableau 9 où nous observons des données similaires en dehors du quantile à l'ordre 99% où nous avons une plus grande valeur pour le second évènement.

Agg.Horaire	Agg.DPT	2022-06-02 au 2022-06-06	2022-06-19 au 2022-06-23
Max	Quantile.50%	1,5	1,5
Max	Quantile.75%	2	2
Max	Quantile.90%	3	3
Max	Quantile.99%	4,5	5,5

Table 9: Tableau comparatif des agrégations au Loiret
Source : Agrégations issues des données Meteomatics

Cependant, l'évènement de gauche a coûté plusieurs millions d'euros tandis que celui de droite n'a engendré que quelques dizaines de milliers d'euros. Comment expliquer un tel écart ? Il s'agit ici d'une question de la vulnérabilité d'AXA face à l'aléa de la grêle.

À gauche, nous observons que l'évènement touche la grande majorité du département dont la ville principale, Orléans avec près de 116 000 habitants en 2021[1] et visible par sa forte densité de population dans la carte 30, tandis que l'autre évènement n'est majoritairement touché qu'aux extrémités en épargnant Orléans notamment. Ainsi, l'évènement du 19 au 23 juin 2022 touche davantage de zone où AXA peut être sensible à cet aléa de grêle décrit par Meteomatics que lors de l'évènement du 02 au 06 juin 2022 à cause d'une présence accrue de la population dans les zones touchées par la grêle, et donc implicitement, un plus grand nombre de biens exposés à cet aléa.



Graphique 30: Densité de la population du Loiret

Ainsi, il est pertinent de créer de nouveaux indicateurs qui prennent en compte la carte de la population française pour intégrer non seulement l'aléa de la grêle mais aussi pour avoir une notion de vulnérabilité. En croisant de cette manière aléa et vulnérabilité nous avons ainsi une meilleure mesure du risque de grêle qui va nous être utile pour la suite de ce mémoire.

Cependant, dans le cadre actuel de notre analyse, nous décidons de ne pas intégrer les données de localisation des expositions d'AXA dans les indicateurs Meteomatics que nous élaborons. Cette décision est motivée d'une part par la précision relativement des données du portefeuille qui nous semble insuffisante comparée à celle de la densité de la population ou celle des données Meteomatics et d'autre part, nous cherchons distinguer et analyser séparément deux aspects clés du risque : l'aléa à travers les indicateurs issus des agrégations des données de Meteomatics et la vulnérabilité issue des données d'exposition d'AXA. En séparant ainsi ces deux composantes du risque, nous comprendrons mieux leurs contributions respectives au risque global de grêle au sein de notre modèle.

[1] https://www.insee.fr/fr/statistiques/2011101?geo=COM-45234

5.1.3 Fonctions retenues

Dans cette partie, nous allons présenter les différentes fonctions et méthodes retenues pour effectuer les agrégations afin de mieux appréhender les sinistres liés à la grêle.

Étant donné le caractère très localisé de la grêle, nous prenons tout d'abord en compte les quantiles, nous décidons de nous intéresser notamment aux quantiles d'ordre 50%,75%, 80%, 85%, 90%, 95% et 99%. Nous avons décidé de ne pas relever des quantiles d'un ordre inférieur à 50% car nous estimons que la majorité des données issus des agrégations donneraient des valeurs nulles ce qui n'aurait pas un grand intérêt dans notre étude. Ensuite, nous essayons d'avoir un pas relativement régulier pour essayer de capter au mieux différentes caractéristiques des grêles. Par exemple, si le quantile à l'ordre de 80% est de 2 cm de diamètre. Cela pourrait paraître faible mais cela veut dire que pour un phénomène local comme la grêle, 20% du département est touché par des grêlons d'un diamètre supérieur à 2 cm, ce qui est conséquent. Les quantiles à l'ordre 95% ou 99% vont davantage capter des phénomènes de grandes valeurs en queue de distribution des diamètres de grêle. Un quantile d'ordre 95% à 5 cm de diamètre peut être révélateur d'une grande sévérité sur certaines zones. Cependant, ces mesures vont être trop sensibles aux valeurs extrêmes qui risquent d'influencer fortement les données en cas d'erreurs

Pour mieux comprendre les comportements dans les queues de distribution qui risquent d'affecter davantage les expositions d'AXA, nous décidons également d'étudier la TVaR aux quantiles d'ordre 75%, 80%, 85% et 90% pour les mêmes raisons. La TVaR va mieux capter ces comportements en estimant les diamètres de grêles dans cette queue de distribution. Nous avons une mesure classique de cette TVaR à travers l'estimation de la moyenne des diamètres de grêle dépassant le quantile d'ordre fixé, mais nous élaborons également cette métrique à l'aide d'une médiane pour être encore moins sensible aux phénomènes trop extrêmes. Nous ne connaissons pas à ce stade quelle est la meilleure agrégation possible. Le but à travers ces multitudes de variables est d'essayer de repérer ainsi celle qui vont se révéler les plus pertinentes dans nos modèles. Mathématiquement, les fonctions sont obtenues en suivant les formules mathématiques présentées dans la section 4.1.

Nous avons également des indicateurs simples comme le maximum, ou des sommes classiques. Concernant les sommes, nous élaborons deux indicateurs selon des seuils de diamètres de grêle : un indicateur qui va prendre en compte toutes les données Meteomatics, nommé NWS_0 et un autre qui va simplement faire la somme des valeurs supérieures à un certain seuil S, ici fixé à 2 cm, NWS_2 car nous estimons qu'il s'agit de ces grêlons

là qui commencent à occasionner des dommages sur la toiture des bâtiments par exemple. En étudiant des indicateurs calculés à partir de la somme des diamètres de grêle, cela va nous permettre d'avoir une certaine mesure de l'importance des grêlons tombés sur une zone donnée. Dans toute cette partie, nous notons X, le diamètre de grêle modélisé par Meteomatics que nous cherchons à agréger, S, le seuil au delà duquel nous considérons les diamètres de grêle et P, le seuil de normalisation.

$$\text{NWS(X,S)} = \text{non_weighted_sum}(X, S) = \sum_{x \in X, x \geq S} x, \quad S \in \{0; 2\} \tag{24}$$

En plus de ces indicateurs de sommes, nous pensons qu'il sera également intéressant de nous pencher sur des sommes des diamètres de grêle que l'on pondère par leur propre valeur, ce qui revient à faire la somme des diamètres de grêle au carré WS ou encore pondéré par leur valeur au carré, ce qui revient dans ce cas à en prendre le cube, WSS pour amplifier l'effet des diamètres des grêles d'une taille plus importante. En effet, l'effet des tailles de grêlons peuvent avoir des effets non linéaires sur les montants de sinistres finaux. Nous prenons cette fois-ci toutes les valeurs, à savoir un seuil S de 0 pour les diamètres de grêlons.

$$WS(X, S) = weighted_sum(X, S) = \sum_{x \in X, x \geq S} x^2 \quad , \quad S = 0 \tag{25}$$

$$WSS(X, S) = weighted_squared_sum(X, S) = \sum_{x \in X, x \geq S} x^3 \quad , \quad S = 0 \tag{26}$$

Nous avons également pensé à des sommes pondérés et normalisés tout en gardant cette volonté d'amplifier les impacts des diamètres de grêlons les plus importants. Nous définissons une somme pondérée par l'exponentiel de la valeur divisée par un seuil, $WNES$. Nous allons cette fois-ci utiliser 2 et 4 comme seuil de normalisation P et 0 pour le seuil S, c'est à dire que nous prenons l'ensemble des diamètres de grêle. En divisant les valeurs par 2 ou 4 selon l'indicateur, nous allons donner plus de poids aux diamètres de grêlons tombés supérieurs à ces valeurs et au contraire amoindrir l'importance des grêlons de tailles plus faibles qui pourraient être de faux signaux de grêles. En prenant l'exponentiel nous retrouvons l'idée d'amplifier l'effet des plus gros grêlons. Les valeurs de P ont été choisies selon l'étude présentée dans la section 2.4.3.

$$WNES(X, P, S) = weighted_normalized_exp_sum(X, P, S) \sum_{x \in X, x \geq S} x \exp(x/P), \quad P \in \{2; 4\} \tag{27}$$

Ensuite, tout comme la somme, nous définissons des indicateurs autour de la moyenne. Nous ne réalisons pas d'indicateurs autour de la moyenne simple car la majorité des valeurs étant nulles, nous allons nous retrouver avec un indicateur peu pertinent. Nous créons donc des indicateurs autour de la moyenne du carré WM, et du cube des diamètres de grêlons WSM ou encore la moyenne simple des 5 plus grandes valeurs de diamètre de grêle, $NWMT_5$.

Concernant le seuil S pour ces indicateurs, nous avons longuement hésité à indiquer également des valeurs comme 2 ou 4 par exemple. Cependant, nous avons décidé de ne pas prendre en compte de seuil pour ces indicateurs cités. S sera donc égal à 0. En effet, pour ces derniers, le fait de prendre le carré ou le cube des diamètres de grêle vont nous donner une information sur l'importance des grêlons tombés. Cependant le fait d'avoir des valeurs nulles va pénaliser en quelques sortes ces valeurs et nous indiquer la localité nous non de ces évènements sans pour autant que les valeurs soient proches de 0 comme cela aurait été le cas avec une moyenne simple. Nous élaborons également la moyenne pondérée par l'exponentiel des diamètres de grêlons divisés par le seuil de normalisation P de 2 ou 4, $WNEM$ selon les indicateurs pour garder l'idée d'amplifier l'impact des diamètres de grêlons les plus importants.

$$WM(X) = weighted_mean(X) = \frac{\sum_{x \in X} x^2}{\sum_{x \in X} x} \tag{28}$$

$$WSM(X, S) = weighted_squared_mean(X) = \frac{\sum_{x \in X} x^3}{\sum_{x \in X, x \geq S} x^2} \tag{29}$$

$$WNEM(X, P) = weighted_norm_exp_mean(X, P) = \frac{\sum_{x \in X} x \exp(x/P)}{\sum_{x \in X} \exp(x/P)}, \quad P \in \{0; 2\} \tag{30}$$

$$NWMT_(X, T) = non_weighted_mean_top(X, T) = \frac{1}{T} \sum_{i=1}^{T} x_{(i)}, \quad T = 5 \tag{31}$$

Dans la dernière formule, $NWMT_5$, pour $T = 5$, les $x_1, x_2, ..., x_5$ représentent les 5 plus grandes valeurs, triées par ordre décroissant.

En dehors de ces valeurs, nous définissons également les indicateurs nb_2_4, nb_4, $Variance$. Commençons par les plus simples, à savoir le nombre de grêlons tombés entre 2 et 4, puis ceux au-delà de 4 cm de diamètre. Les grêlons entre 2 et 4 cm vont être le plus communs, et une accumulation de ces grêlons peut indiquer l'intensité de l'évènement. Tandis que les grêlons de plus de 4 cm de diamètre vont relever davantage de la sévérité de l'évènement, ces grêles-là étant capable de blesser des personnes.

$$nb_2_4 = nombre_grele_entre_2_4_cm = \sum_{i=1}^{N} \mathbb{I}_{\{2 \leq x_i \leq 4\}} \tag{32}$$

$$nb_4 = nombre_grele_superieur_4_cm = \sum_{i=1}^{N} \mathbb{I}_{\{x_i > 4\}} \tag{33}$$

Ensuite pour la variance, cela va nous permettre de quantifier la dispersion des tailles de grêlons autour de la moyenne. Une variance élevée indique que les tailles des grêlons sont très variées, tandis qu'une variance faible suggère que les tailles sont plus homogènes.

$$Variance = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})^2, \text{où } \bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i \tag{34}$$

Maintenant que nous avons passé en revue les différentes agrégations envisagées, il est essentiel de déterminer le sens de l'agrégation, comme nous l'avons vu dans la section 5.1.1.

Agrégation temporelle et spatiale

Les données de Meteomatics sont sous la forme de raster, chaque fichier horaire comporte en ligne les Latitudes et en colonne les longitudes. Le restant des cellules de la matrice étant rempli par les valeurs des diamètres de grêle modélisé. Ce format étant peu facilement manipulable, nous décidons de les transformer en table de données classiques ayant pour première colonne les latitudes, en deuxième colonne les longitudes. Les colonnes suivantes représenteront chaque heure des différents fichiers raster. Pour chaque coordonnée de point donnée par sa latitude et sa longitude, les valeurs des diamètres de grêle pour chaque heure sont insérées dans les colonnes correspondantes. Ainsi du format de la table 10, nous arrivons au format de la table 11.

Latitude / Longitude	**10.0**	**10.5**	**11.0**	**11.5**
45.0	0	0	0	0
45.5	0	1	3	2
46.0	0	1	2	1
46.5	0	2	3	1

Table 10: Diamètres de grêle sous format raster pour 1 heure, 00:00 par exemple

A partir de cette table, nous pouvons soit réaliser une agrégation par ligne, auquel cas comme il s'agit d'une agrégation par coordonnée, il s'agira d'une agrégation spa-

Latitude	Longitude	00:00	01:00	02:00
45.0	10.0	0	0	0
45.0	10.5	0	0	0
45.0	11.0	0	0	0
45.0	11.5	0	0	0
45.5	10.0	0	1	3
45.5	10.5	1	3	2
45.5	11.0	3	2	1
45.5	11.5	2	1	1
46.0	10.0	0	1	2
46.0	10.5	1	2	1
46.0	11.0	2	1	1

Table 11: Diamètres de grêle sous format table de donnée

tiale, soit une agrégation par colonne, auquel cas ce sera une agrégation temporelle. Concernant l'agrégation temporelle, elle sera plutôt directe car les différentes heures sont clairement affichées, cependant, il va nous falloir une étape supplémentaire pour obtenir les départements ou les communes associées aux coordonnées. Pour cela, il est courant d'utiliser des fichiers *shapefiles* qui contiennent les coordonnées géographiques des contours des communes ou des départements par exemple. Ces fichiers permettent de réaliser des analyses spatiales précises. Nous utiliserons notamment les packages *raster* et *sf* de R pour identifier les communes ou départements associées à ces coordonnées ainsi que les *shapefiles* fournis par le jeu de données publiques françaises : https://www.data.gouv.fr/fr/datasets/ pour la France et ceux de ce *github* pour l'Italie : https://github.com/openpolis/geojson-italy. Une fois les codes de département rattachés aux coordonnées, il sera ainsi plus simple d'agréger les données à partir de nos fonctions.

Les épisodes de grêle étant relativement de courte durée, pour les variables agrégées temporellement, puis spatialement, nous décidons de ne prendre en compte que le maximum de la valeur des diamètres de grêle pour chaque coordonnée sur le pas de temps horaire. Nous pensons en effet que pour une coordonnée donnée, nous n'allons pas avoir un nombre significatif de valeurs différentes sur la durée de l'évènement. Ainsi, pour chaque coordonnée, nous aurons une unique valeur. Ensuite, pour agréger spatialement, nous pouvons utiliser les codes de département. Pour cela, nous décidons de ne prendre en compte que les quantiles d'ordre 50%, 75%, 80%, 85%, 90%, 95% et 99%. De plus, pour identifier plus facilement les variables ayant d'abord été agrégée de manière temporelle, nous décidons d'ajouter un préfixe $T_$. Ainsi, la variable nommée $T_Max_Quantile_80$ signifie que les données Meteomatics ont d'abord été agrégée au niveau temporel, c'est à dire sur toute la durée de l'évènement en prenant le maximum des diamètres de grêle par

coordonnée, puis nous avons pris le quantile à l'ordre 80% par département.

Nous pouvons désormais nous concentrer sur les variables agrégées spatialement, puis temporellement. nous allons ici utiliser toutes les fonctions que nous avons décrites précédemment, à savoir les agrégations autour des sommes, des moyennes, de la variance, des quantiles ainsi que des TVaR et du maximum. Ces fonctions, combinées aux différents seuils des valeurs de diamètres de grêle pris en compte (souvent 0 cm ou 2 cm) et aux différentes normalisations de certaines fonctions comme $WNES$ ou $WNEM$ nous amène à 29 agrégations différentes.

Pour éviter d'avoir trop de variables à analyser, nous décidons dans la deuxième étape de l'agrégation de nous en tenir à 3 agrégations possibles, à savoir la somme sum, le maximum max ainsi que la moyenne des 5 plus grandes valeurs $NWMT_5$. Ainsi, pour une variable agrégée spatialement par le quantile à l'ordre 95% par exemple, nous aurons 3 variables finales une fois agrégées temporellement, à savoir $Quantile_95_sum$, $Quantile_95_max$, $Quantile_95_NWMT_5$. Cela multiplie ainsi grandement les variables à notre disposition. Il sera alors indispensable de définir des méthodes pour savoir quelles sont les variables les plus pertinentes.

Agrégations diverses

Pour étudier les caractéristiques des épisodes de grêle, nous avons décidé d'élaborer 3 nouvelles agrégations, à savoir $duree_max_grele$, $duree_grele$, nb_grele, nous avons dû faire quelques hypothèses pour leur réalisation. La première vise à obtenir la durée de l'épisode de grêle la plus longue au sein du département, la deuxième cumule la durée des différents épisodes sur le département et enfin le dernier compte le nombre de ces épisodes de grêle. Pour chaque évènement relevé dans notre base de données de sinistre, nous allons extraire les dates correspondantes des données de Meteomatics, nous savons donc que ces données possèdent le ou les épisodes de grêle ayant causé différents niveaux de perte. Cependant, réussir à caractériser ces différents épisodes n'est pas évident. En effet, il ne suffit pas de prendre des valeurs strictement positives car la modélisation indique parfois des diamètres de grêlons de 0,5 cm ou 1 cm de diamètre qui vont occasionner peu de dégâts et simplement ajouter du bruit à nos données. Ensuite, ces valeurs sont davantage susceptibles d'être de faux signaux de grêle.

Pour mieux coller à la volonté de modéliser nos sinistres observés dans notre portefeuille, notamment le résidentiel, nous avons décidé de ne prendre en compte pour élaborer

ce nouvel algorithme que les tailles de grêlons au-dessus de 2 cm de diamètre. Ensuite, nous agrégeons les données de diamètres de grêle avec le quantile à l'ordre 99% par département. Nous ne prenons pas le max pour éviter les potentielles valeurs extrêmes issus d'erreurs de modélisation. Pour chaque heure nous allons avoir au niveau du département différentes valeurs de diamètre de grêle. A partir de là, nous allons étudier les valeurs consécutives supérieures à 2 cm de diamètres pour essayer de caractériser les évènements de grêles survenus au sein de chaque département.

Cependant, il peut arriver que pour une certaine heure, le diamètre relevé puisse passer de 2 cm de diamètre à 1,5 cm, puis remonter à 2 cm pour l'heure suivante. Dans ce cas nous aurions une discontinuité alors qu'il est fort probable qu'il s'agisse du même épisode de grêle. Ainsi, nous ajoutons une condition de tolérance de discontinuité. Si la valeur descend une seule fois en dessous de 2 cm, c'est à dire pendant une heure, alors nous considérons qu'il s'agit du même évènement.

Les trois indicateurs vont être crées à partir de ce mécanisme : le nb_grele va relever le nombre d'épisodes continus, $duree_grele$, va sommer le nombre d'heures de chaque épisode et $duree_max_grele$ va relever le nombre d'heure maximal des différents épisodes de grêle. L'idée ici est de pouvoir identifier les différents épisodes de grêles.

Algorithm 1 Algorithme de détection des épisodes de grêle

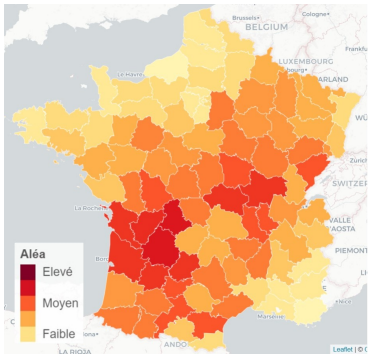
1: **Entrée :** valeur de diamètres de grêle par département et sur les différentes heures de l'évènement
2: **Agréger** les diamètres de grêle par le quantile à l'ordre 99% par département
3: **Identifier** les valeurs consécutives supérieures à 2 cm avec une tolérance de discontinuité d'une heure
4: **Sortie duree_grele :** Sommer la longueur des plages de données horaires consécutives
5: **Sortie duree_max_grele :** Prendre le maximum des longueurs des plages de données horaires consécutives
6: **Sortie nb_grele :** Compter le nombre de plages des données horaires consécutives

Données de la densité de population

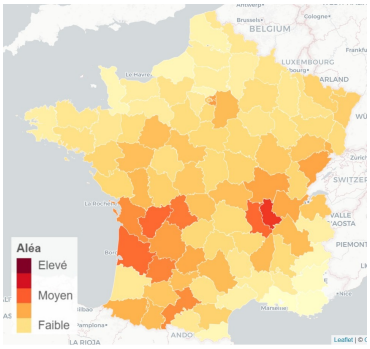
Maintenant que nous avons ces différentes agrégations, nous créons à nouveau chacune de ces agrégations, mais en prenant en compte cette fois-ci les données de la densité de la population avant toute application d'agrégation. Ainsi, à chaque fichier raster Meteomatics pour une heure donnée, nous la multiplions par la carte de la densité de la population présentée dans la section 3.5. Nous reproduisons les différentes agrégations présentées

dans cette partie. Désormais, nous aurons non seulement la variable *Quantile_95_max*, mais nous aurons également son pendant avec la prise en compte de la donnée de la densité. Pour les différencier, nous ajoutons le suffixe *_POP*. Ainsi, nous aurons la variable *Quantile_95_max_POP*. Cela nous donne une diversité conséquente de variables climatiques, près de 200, à tester pour comprendre la sinistralité liée à la grêle, ce qui est à la fois une richesse, mais également un facteur de complexité pour la modélisation.

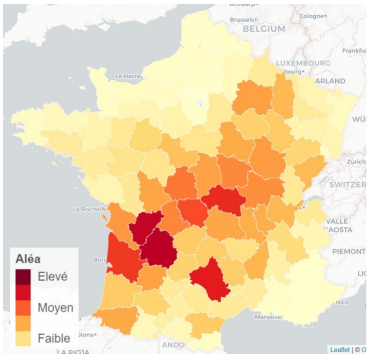
Ces différentes méthodes nous permettent d'obtenir plusieurs agrégations possibles, nous en affichons 6 à titre d'exemple obtenus à partir des données météorologiques de l'évènement QIARA de juin 2022.



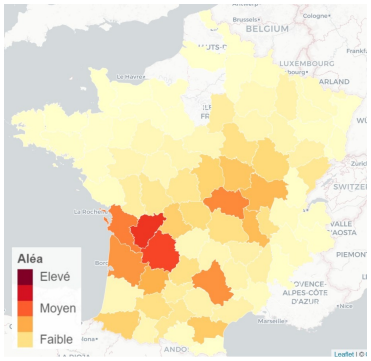
Graphique 31: Quantile_95_max



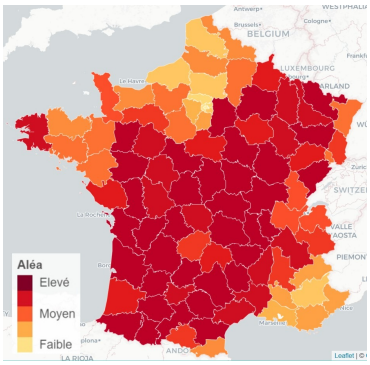
Graphique 32: Quantile_95_max_POP



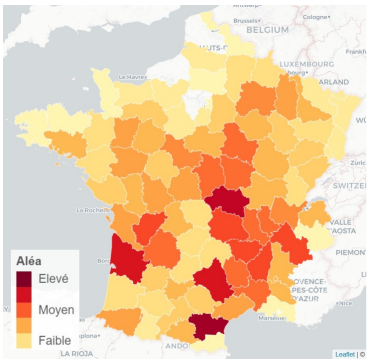
Graphique 33: nb_2_4_sum



Graphique 34: nb_4_sum



Graphique 35: Max_Max

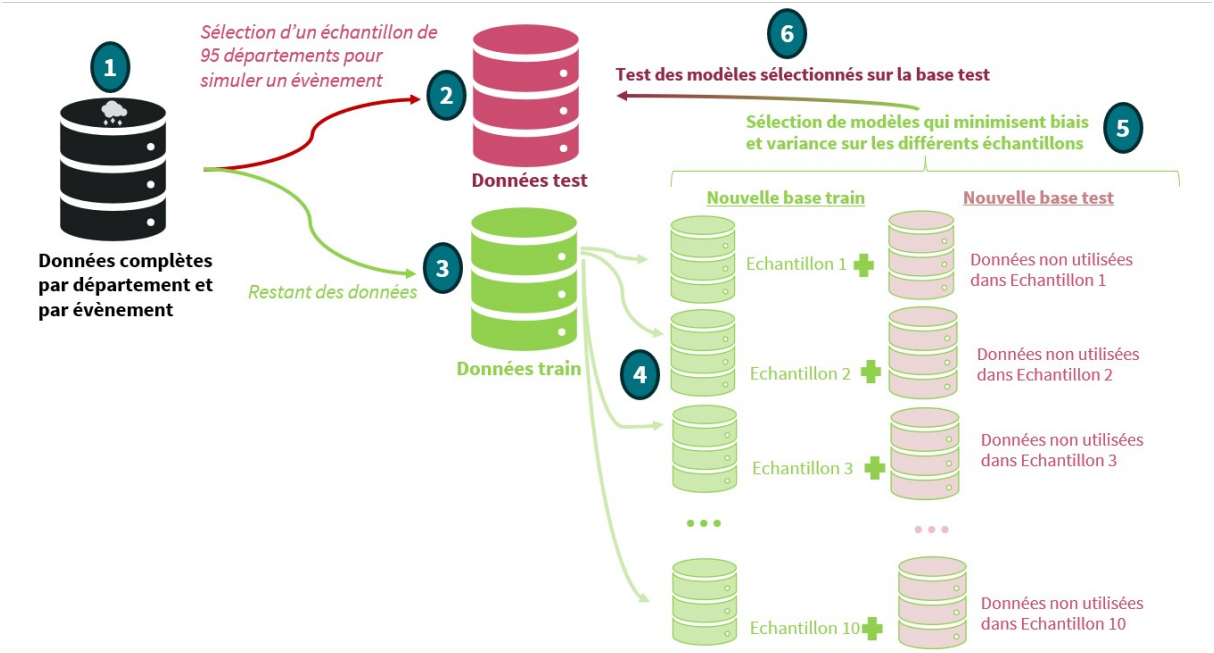


Graphique 36: Duree_Max_grele

5.2 Démarche pour la modélisation

Dans le cadre de notre mémoire, nous souhaitons modéliser la charge de sinistralité des évènements de grêle au niveau du portefeuille d'AXA. Nous pensons qu'une étape importante de cette démarche est la classification des sinistres observés au niveau du portefeuille en fonction des montants de sinistres qui sont agrégés par département et par évènement de grêle. Nous allons nous limiter à 6 classes allant de l'absence de sinistre par départe-

ment que nous allons appeler Classe 0, aux sinistres plus conséquents que nous allons catégoriser en Classe 5. Cette méthode repose sur plusieurs hypothèses mais également sur les caractéristiques des données à notre disposition. Ensuite, nous allons modéliser ces classes à partir de différentes méthodes de classification et sélectionner les variables les plus pertinentes. Enfin, à partir de ces classes prédites, nous allons utiliser un second modèle qui nous permettra d'obtenir un taux de destruction qui nous donnera directement le montant de sinistres prédits en appliquant l'exposition du portefeuille. Pour prendre en compte le manque de données disponible, nous allons d'abord entraîner selon le schéma 37, les différents modèles sur un sous-échantillon de la base de données initiale qui nous servira de base d'entraînement, sur laquelle nous allons générer plusieurs évènements grâce à la méthode développée dans la sous-section 5.3.2 et déterminer les modèles à faibles biais et variance avant de les tester sur l'échantillon test afin de s'assurer de la robustesse des modèles.



Graphique 37: Méthodologie d'entraînement et de test des données

5.3 Classification des sinistres

5.3.1 Identification des classes

Nous avons décidé d'adopter dans un premier temps une catégorisation des sinistres agrégés au niveau départemental qui vise à classer les départements uniquement en fonction de l'ampleur des dommages causée par les grêles.

Nous pensons qu'il existe des effets de seuil significatifs entre les différentes catégories de sinistres qui traduisent l'étendu des dégâts au niveau des biens exposés à la grêle. Par exemple, un sinistre répertorié en Classe 4 ou Classe 5, correspondant à des sinistres agrégés sur le département allant de plusieurs millions à des dizaines de millions d'euros, peut résulter du croisement entre un évènement de grêle de très forte intensité et d'une forte vulnérabilité du département à cet évènement. Cela va ainsi à la fois toucher la structure des bâtiments assurés par AXA à travers leur toiture, leurs murs ou leurs fenêtres par exemple pour les maisons et les immeubles, en plus de toucher violemment la carrosserie des expositions automobiles. En parallèle, un département sinistré répertorié en Classe 1 ou 2, représentant un sinistre agrégé allant de quelques milliers d'euros à des dizaines de milliers d'euros à l'échelle du département va davantage être révélateur d'un évènement peu intense et très localisé au sein du département. Cela va affecter des éléments de moins grande envergure comme des parties isolées des maisons ou des immeubles, ou alors un petit parc automobile. Ainsi, lorsque nous allons modéliser le taux de destruction des sinistres dans un second temps, ces effets de seuil peuvent avoir des effets non linéaires et non négligeables à prendre en compte.

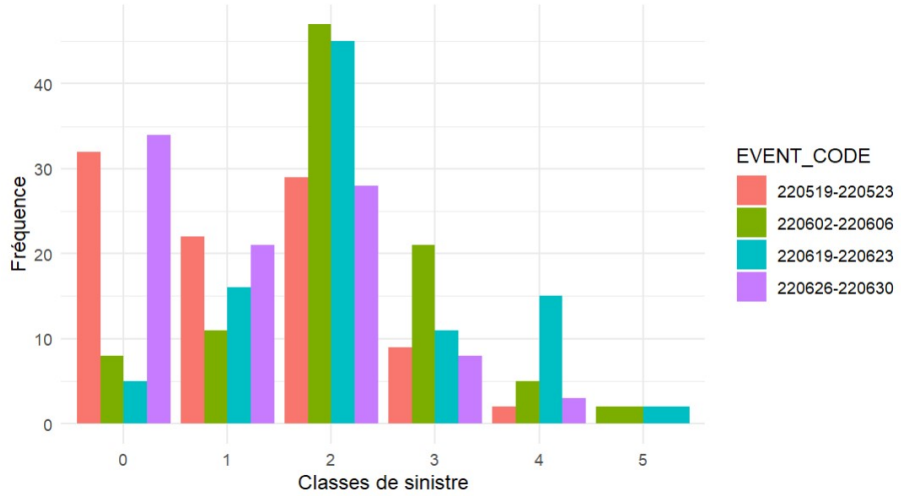
Cette distinction de classe de sinistres nous est essentielle, non seulement pour appréhender ces effets de seuil, mais également pour compenser le fait que nous n'avons pas pu établir une correspondance directe entre notre base d'exposition et notre base de sinistralité au niveau de chaque police d'assurance. En effectuant cette agrégation de sinistres par catégorie, nous essayons ainsi de mieux comprendre le comportement et la vulnérabilité de notre portefeuille face aux différentes intensités des évènements de grêle modélisées par les agrégations Meteomatics vues précédemment malgré l'absence de données individuelles dans notre base de données finale.

En effet, nous pouvons estimer raisonnablement que des départements sinistrés de classe 4 ou 5 sont liés à des fortes grêles intenses, avec une durée plus longue et touchant un nombre conséquent de biens assurés par AXA au sein du département comparés aux sinistres de classe 1 ou 2 qui vont quant à eux être associés à des évènements plus modérés en termes d'intensité et de vulnérabilité, nous touchons alors davantage à un risque de fréquence que de sévérité. Par la suite, nous allons simplifier l'expression de classes de départements sinistrés par «classes de sinistre ». Ainsi, lorsque nous ferons référence à des sinistres, nous parlerons toujours de sinistres agrégés par département et pour un évènement de grêle.

Classe	Description	Montant	Impact sur le département
0	Non touché par la grêle	Aucun sinistre	Aucune perte financière
1	Sinistres mineurs	Milliers d'euros	Impact superficiel sur les biens
2	Sinistres modérés	Dizaine de milliers d'euros	Dommage plus significatifs avec potentiellement plusieurs bien touchés
3	Sinistres importants	Centaine de milliers d'euros	Dommages considérables, impact sur la structure notamment
4	Sinistres majeurs	Millions d'euros	Dommages graves sur la structure et sur le contenu des biens, grands montants de réparation attendus
5	Sinistres exceptionnels	Dizaine de millions d'euros	Très fort impact économique sur le portefeuille

Table 12: Descriptif des classes de sinistres

En adoptant cette approche de classification des sinistres, nous visons à capturer avec ce modèle des nuances importantes dans la relation complexe entre l'intensité des évènements de grêle et leur impact financier sur le portefeuille tout en tenant compte des limitations liées aux données disponibles. Ainsi, cette méthode va nous permettre de modéliser de manière plus fine le taux de destruction et ainsi la charge de sinistre sur le portefeuille d'AXA à partir des données météorologiques.



Graphique 38: Distribution des classes des départements sinistrés selon les évènements

Nous affichons la distribution des classes de départements sinistrés par évènement dans le graphique 38. Si nous prenons l'évènement de grêle du 02 juin 2022 au 06 juin 2022, affiché en vert sur l'histogramme, nous voyons que durant cet évènement, plus d'une quarantaine de départements ont été sinistré à hauteur de quelques dizaines de milliers

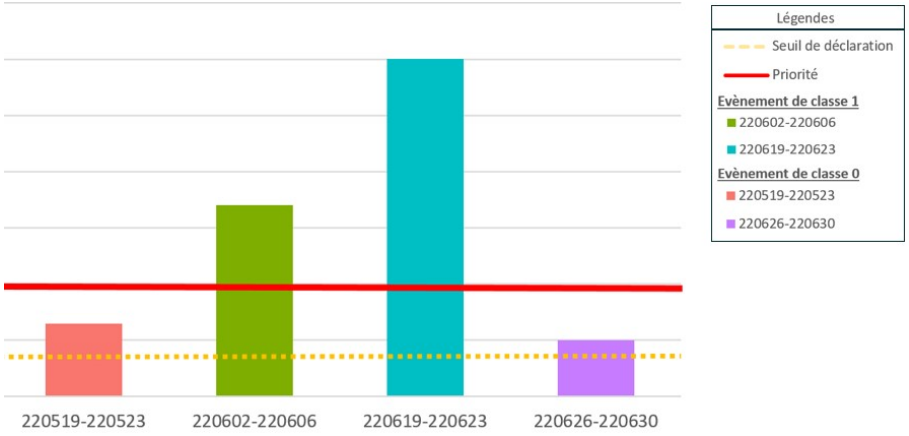
d'euros (sinistres de Classe 2) avec un petit nombre de départements ayant été touché à plus de dix millions d'euros (Classe 5). Notre objectif ici est de trouver des similarités en termes de distribution de sinistres entre les différents évènements.

Nous constatons des comportements similaires entre les évènements en rouge et en violet (respectivement évènement du 19 mai 2022 au 23 mai 2022 et 26 juin 2022 au 30 juin 2022) qui présentent tous les deux un grand nombre de départements qui n'ont pas été affectés par la grêle, d'où le fort nombre de Classe 0 et un nombre réduit de classes supérieures à 3 tandis que les deux autres évènements en vert et en bleu (respectivement les évènements du 02 juin 2022 au 06 juin 2022 et du 19 juin 2022 au 23 juin 2022) ont un nombre très faible de départements épargnés par la grêle (Classe 0), un nombre plus importants de départements de Classe 2 et une poignée de départements de Classe 4 et 5.

Cette distinction de distribution de départements sinistrés entre ces deux types d'évènements nous semble cohérente. En effet, nous avions vu dans la section 3.2.1 que nous avions à disposition les montants de sinistre dans deux cas de figure dans un contexte de réassurance des entités d'AXA :

- Cas 1 : Les montants de sinistres lorsque la charge de sinistre totale au niveau de l'évènement dépasse le seuil de déclaration mais restent inférieurs à la priorité du traité de réassurance.
- Cas 2 : Les montants de sinistres lorsque la charge de sinistre totale au niveau de l'évènement dépasse la priorité du traité de réassurance.

Après analyse de la sinistralité, nous avons identifié que le Cas 1 correspondait aux évènements en rouge et violet et que le Cas 2 correspondait à ceux en bleu et vert. Étant donné leurs comportements distincts, il nous semble également important que notre modélisation intègre cette différenciation. Ainsi, les événements du Cas 1 seront classés comme évènement de «Classe 0», tandis que ceux du Cas 2 seront affectés comme évènement de «Classe 1». Nous résumons ces analyses dans le graphique 39. Cette approche permettra une meilleure prise en compte des spécificités de chaque type d'événement dans notre modèle. Nous avons ainsi deux classifications : une classification pour les départements sinistrés par la grêle allant de 0 à 5, et une classification pour les évènements de grêle avec 1 ou 0.



Graphique 39: Distinction des classes d'évènements en fonction du montant total de sinistre des différents évènements

Nous gardons cependant en tête que nous élaborons ces profils d'évènements à partir d'une année exceptionnelle de sinistralité de grêle qu'est 2022. Il faudra ainsi vérifier si ces profils restent vrais pour les sinistres à venir.

5.3.2 Échantillonnage des évènements

Nous devons donc modéliser des classes de sinistres dans un premier temps. Pour ce faire et pour avoir des résultats robustes, nous devons séparer notre base de données en deux échantillons : une base d'entraînement sur laquelle nous allons entraîner les données et en tirer de premières conclusions et une base de test pour vérifier que notre modèle est cohérent avec les données et qu'il n'y a pas eu de sur-apprentissage. Nous nous confrontons à la problématique des évènements disponibles à notre étude. En effet, étant donné que nous avons dû restreindre notre étude au cas de la France ainsi que des évènements remontés en 2022, celui réduit les données sur lesquels entraîner le modèle et nous risquons alors d'élaborer des modèles avec du sur-apprentissage qui vont alors bien s'ajuster aux données d'entraînement mais en revanche, qui ne pourront pas correctement prédire les données de test. Pour tenter de résoudre ce problème, nous avons décidé de créer de nouveaux échantillons d'évènements à partir de notre base de données historiques avec des contraintes telles que ces nouveaux évènements créés soient cohérents avec ce qui a déjà pu être observé.

Pour cela nous avons développé un algorithme qui nous permet de générer de nouveaux échantillons d'évènements de grêle par département, en nous appuyant sur notre base de données existante. Cet algorithme respecte également une certaine structure et ainsi que les différentes contraintes issues des données initiales. Chaque échantillon généré maintient une taille constante de 95 départements, correspondant aux 95 départements

français métropolitains hors Corse, avec chaque département n'apparaissant qu'une seule fois par évènement pour générer des évènements réalistes. Concernant les contraintes, l'algorithme prend en entrée des contraintes spécifiques sur la distribution des classes de sinistres (de 0 à 5) au sein de chaque échantillon, permettant de reproduire des schémas de distribution observés selon l'intensité de l'évènement. Cette approche permet ainsi d'avoir une certaine la flexibilité dans la génération d'évènements avec une répartition de classes de sinistres prédéfinie pour une classe d'évènement donnée.

Cette méthode nous permet ainsi d'obtenir de nouveaux échantillons sur lesquels entraîner les données en évitant le sur-apprentissage tout en ajoutant de l'aléa dans ces données au niveau du montant total de sinistre par évènement. Cette dernière caractéristique est particulièrement intéressante pour nous car elle permet d'entraîner le modèle sur des scénarios de sinistres totaux qui sont cohérents avec les observations historiques mais qui ne se sont pas encore produits, renforçant ainsi la robustesse et la capacité prédictive du modèle face à des évènements futurs potentiellement plus extrêmes.

Algorithm 2 Algorithme de création de nouveaux échantillons

Entrée : Liste des départements, liste des classes de sinistres (0 à 5), contraintes pour chaque classe
Répertorier les classes de sinistres disponibles par département
Pour chaque classe i de 5 à 0 **Faire** une revue aléatoirement des départements
    Pour chacun de ces départements **Faire**
        Si le département n'a pas encore de classe **Alors**
            Lui affecter la classe i si elle est disponible en respectant les contraintes
        Fin Si
    Fin Pour
Fin Pour
Cas des départements restants : Attribuer une classe aléatoire parmi celles restantes
Sortie : Liste des départements avec les classes de sinistres attribuées

Nous avons précédemment établi des classes de sinistres. A partir des seuils définis, nous avons pu élaborer un profil des évènements en nous appuyons sur la fréquence de ces sinistres. Nous observons ainsi que pour les évènements de catégorie 1 nous retrouvons cette répartition:

Classe 0	Classe 1	Classe 2	Classe 3	Classe 4	Classe 5
10	15	20	18	10	2

Table 13: Exemple de distribution des classes de sinistres pour les évènements de catégorie 1

Lecture : un évènement de catégorie 1 sera simulé à partir de 2 départements de classe de sinistre 5, 10 départements de classe de sinistre 4 etc.

Par un raisonnement similaire pour la classe 0, nous obtenons la répartition suivante :

Classe 0	Classe 1	Classe 2	Classe 3	Classe 4	Classe 5
30	20	18	15	4	0

Table 14: Exemple de distribution des classes de sinistres pour les évènements de catégorie 0

Lecture : un évènement de catégorie 0 sera simulé à partir de 4 départements de classe de sinistre 4, 15 départements de classe de sinistre 3 etc.

Nous supposons que des évènements d'une même catégorie (0 ou 1) vont avoir en général la même répartition. Nous pouvons alors à partir de ces hypothèses créer des échantillons d'évènements à partir de notre algorithme cohérents avec ce qui s'est réellement produit.

Département	Echantillon 1	Echantillon 2	...	Echantillon n
01	2	1	...	2
02	4	5	...	4
...	...	...	...	...
95	1	1	...	2

Table 15: Exemple d'échantillon généré par l'algorithme 2

Cette étape est cruciale car nous voulons comprendre la relation entre classes de sinistres et agrégations issues des données Meteomatics mais nous ne voulons pas ajouter du bruit en associant des données incohérentes. Désormais, nous allons pouvoir nous pencher à nouveau sur l'élaboration d'échantillon d'entraînement et de test : nous générons tout d'abord pour les évènements de catégorie 1 et 0 un nouvel échantillon. Il est à noter que nous avons pu faire cela pour les évènements de catégorie 0, cependant pour les évènements de catégorie 1, à savoir les évènements les plus intenses, nous n'avions pas assez de départements sinistrés de classe 5, nous n'en avons que 4 au total dans l'historique. Ainsi, si nous utilisons 2 départements de classe 5 dans la base test, nous allons nous retrouver qu'avec 2 départements sinistrés de classe 5 dans l'échantillon d'entraînement et nous pensons que cela n'est pas assez pour entraîner les modèles. Nous décidons alors de ne

garder qu'un département de classe 5 dans la base de test. Cela reste cohérent avec le fait que l'année 2022 a connu une sinistralité exceptionnelle comparée aux autres années. Prendre qu'un seul département de classe 5 pour simuler un futur évènement n'est donc pas absurde. Toutes les données qui n'ont pas été utilisées pour constituer cette base de test se retrouvent dans la base d'entraînement.

Ensuite, nous aurions pu utiliser directement cette base d'entraînement pour nos modèles et voir sur la base test si les modèles sont concluants. Cependant, cette méthodologie ne nous permet pas de vérifier la variance de nos modèles sur différents échantillons. En effet, il peut arriver qu'un modèle prédise correctement les données dans certains cas mais qu'en présence d'un nouvel échantillon, cela ne soit pas le cas auquel cas nous ferions de nouveau face à un risque de sur-apprentissage. A cet égard, nous avons décidé d'utiliser à nouveau notre algorithme 2 pour créer une dizaine de sous-échantillons de notre base d'entraînement afin de faire une validation croisée de nos modèles et nous assurer de leur robustesse. Un bon modèle sera celui qui réduira au mieux le biais et la variance de ces différents échantillons.

Maintenant que nous avons traité le problème des données disponibles, l'objectif est désormais de prédire la classe de sinistralité à partir des variables issues de notre base de données. Nous avons tout d'abord les **variables historiques**, peu corrélées entre elles, qui sont l'exposition, le niveau d'aléa extrait de notre carte 20 ainsi que de la proportion de maison composant notre portefeuille résidentiel et le nombre de communes touchées au niveau du département des données issues de ESSL. Ensuite, nous avons nos **données climatiques** qui ont été agrégées de différentes manières, au total, nous avons environ 200 agrégations possibles. De plus, ces agrégations venant des mêmes données météorologiques de Meteomatics, nous avons de fortes corrélations. Dans le cas de modèles de régression, cela peut poser ainsi poser soucis.

Nous allons donc, à travers une boucle sur nos différentes variables climatiques agrégées, extraire celles qui permettent de modéliser au mieux les classes de sinistres déterminées précédemment. Nous préférons réaliser une boucle sur les différentes variables au lieu de faire un modèle avec plusieurs variables climatiques pour plusieurs raisons. Nous pensons que chaque agrégation apporte une compréhension différente du phénomène et aura une importance selon différents cas. Par exemple $nb_2_4_sum$, correspondant à la variable qui somme sur la durée de l'évènement le nombre de grêlons d'un diamètre compris entre 2 et 4 cm de diamètre tombés par département sera particulièrement utile pour étudier les nombreux grêlons qui peuvent tomber sur une zone tandis que nb_4_sum, correspondant

à la variable qui somme sur la durée de l'évènement le nombre de grêlons d'un diamètre supérieur à 4 cm de diamètre tombés par département, va davantage relever les sinistres qui relèvent des grêlons d'un diamètre conséquent.

Leur utilité dépend donc des évènements de grêles et garder une analyse distincte en fonction de ces variables nous permettra dans notre modèle final de comparer les différentes valeurs obtenues pour un même évènement à prédire afin de garder différentes projections possibles au sein desquelles se situerait le réel montant de sinistralité. De plus, dans notre cas de classification de sinistres, selon les classes prédites à partir des différents indicateurs nous allons pouvoir repérer quel type de grêle affecte le département. Ainsi, une prédiction d'une classe 5 avec nb_4_sum mais d'une classe 3 par $nb_2_4_sum$ indique un évènement qui comporte notamment de grêlons avec un diamètre élevé.

Ensuite, nous allons devoir repérer les meilleures variables d'agrégation. Nous avons tout d'abord pensé aux mesures obtenues à partir de la matrice de confusion qui synthétise la comparaison entre les données prédites et observées. Voici une matrice de confusion dans un cas de variable binaire avec vrai/faux :

	Prédictions	
Observations	Vrai	Faux
Vrai	vrai positif	faux négatif
Faux	faux positif	vrai négatif

Table 16: Matrice de confusion

Nous observons que 4 cas différents sont possibles :

1. **Les vrais positifs (VP)** : Ce sont les cas où le modèle a correctement prédit la classe positive observée.
2. **Les faux négatifs (FN)** : Il s'agit des cas où le modèle a prédit à tort une classe négative alors que l'observation était en réalité positive.
3. **Les faux positifs (FP)** : il s'agit ici des cas où le modèle a prédit à tort une classe positive alors que l'observation était en réalité négative.
4. **Les vrais négatifs (VN)** : ce sont les cas où le modèle a correctement identifié la classe négative.

Nous cherchons ainsi à obtenir les modèles qui maximisent les individus dans les cas 1 et 4 en minimisant les individus dans les cas 2 et 3.

A partir de cela, plusieurs mesures existent pour mesurer. Les principales sont les suivantes :

La justesse (aussi appelée *accuracy*) qui mesure la proportion de prédictions correctes parmi toutes les prédictions et qui est donnée par

$$\text{Justesse} = \frac{\text{VP} + \text{VN}}{\text{VP} + \text{VN} + \text{FP} + \text{FN}}$$

Il s'agit d'une mesure intéressante mais loin d'être parfaite. En effet, il faudra faire attention aux cas observés et leur présence dans l'échantillon. Dans le cas où nous souhaitons étudier l'occurrence de la grêle sur les départements en France par exemple, il s'agit d'un évènement peu fréquent. La proportion de ces évènements sera faible. Ainsi, prenons le cas où dans un échantillon, nous avons seulement 2% des départements touchés par la grêle. Un modèle qui prédirait qu'il n'y a jamais de grêle aurait une justesse de 98%. Or ce n'est absolument pas le modèle recherché alors que nous observerions tout de même une bonne performance.

La précision (*Precision*) qui mesure la proportion de vrais positifs parmi tous les cas prédits comme positifs.

$$\text{Précision} = \frac{\text{VP}}{\text{VP} + \text{FP}}$$

Il s'agit d'un indicateur intéressant qui va permettre de contrebalancer le cas précédent du modèle qui va prédire que des individus comme étant de classe positive.

Sensibilité (*Recall*) : Il mesure la proportion de vrais positifs parmi tous les cas réellement positifs.

$$\text{Rappel} = \frac{\text{VP}}{\text{VP} + \text{FN}}$$

Un équilibre doit être trouvé entre la précision et la sensibilité. Augmenter l'un tend en effet à diminuer l'autre indicateur. Il faudra ainsi déterminer si l'on souhaite en priorité minimiser les faux négatifs ou les faux positifs. Minimiser les faux négatifs dans le cas de l'occurrence de la grêle peut être intéressant pour éviter de sous-estimer le nombre d'évènements en prédisant moins d'évènements de grêle que ce qui est observé. Dans ce cas nous maximiserons la sensibilité. Dans le cas échéant, minimiser les faux positifs permettra d'éviter de surestimer le nombre d'évènements prédit et donc de surestimer la charge de sinistralité. La prudence viserait ici à préférer minimiser les faux négatifs et donc préférer maximiser la sensibilité.

Avec ces différents indicateurs en tête, nous allons pouvoir nous attacher à étudier les

différentes classes de sinistres que nous possédons, allant de 0 à 5.

Classe	0	1	2	3	4	5
0	20	2	1	0	0	0
1	12	15	5	2	0	0
2	4	2	21	10	0	0
3	0	0	2	7	4	0
4	0	0	0	1	3	1
5	0	0	0	1	0	1

Table 17: Exemple de matrice de confusion avec plusieurs classes de sinistres

Dans notre modélisation, nous ne faisons pas face à une variable binaire mais à une variable qui peut prendre 6 classes différentes. Cela amène plusieurs changements par rapport aux métriques précédemment citées. En effet, les différentes classes n'ont pas la même importance et la distance entre les classes est cruciale à prendre en compte. En effet, mal prédire 3 individus de classe 0 et 1 individu de classe 5 n'auront pas les mêmes conséquences financières.

Notion d'importance de classe

D'après le tableau 12 descriptif des classes, nous avons conscience de l'impact des différents sinistres sur le portefeuille d'AXA, les individus de classe 0 indiquant une absence de sinistralité jusqu'aux individus de classe 4 avec plusieurs millions d'euros et les individus de classe 5 montrant des sinistres de l'ordre de la dizaine de millions d'euros. Une erreur de prédiction de la classe 0 va entraîner une surestimation pour cet individu mais dont l'impact sera beaucoup plus léger qu'une erreur de prédiction pour la classe 4 ou 5. Un individu de classe 5 étant prédit en classe 4 par exemple peut entraîner plusieurs millions d'euros de sous-estimation qui entraînera des conséquences importantes dans la précision du modèle. Ainsi, nous voulons établir une nouvelle mesure d'erreur qui va tenir compte de l'importance de chaque classe et qui va leur affecter une pondération particulière pour en tenir compte.

Notion de distance entre les classes

En parallèle de la notion d'importance, la distance de classe entre les observations et les prédictions est primordiale. En effet, un individu de classe 5 prédit en classe 4 aura des conséquences limitées par rapport au cas où il serait modélisé en classe 2 par exemple, cas pour lequel les sinistres pourraient être sous-estimés de plusieurs dizaines de millions d'euros. Il est également important de mêler la notion de distance à la notion

des classes. Par exemple, pour un individu de classe 1, avec un sinistre à 3 000€ par exemple qui va être prédit en classe 3 à 150 000€, cela fera 2 classes d'erreurs pour une surestimation de 147 000€ qui reste acceptable. Cependant, si nous prenons une même distance d'erreur de 2 classes, appliqué pour un individu de classe 5, les conséquences sont loin d'être les mêmes. Admettons un sinistre ayant coûté 15 000 000€ au département lors de l'évènement. S'il est prédit en classe 3, à 150 000€ par exemple, cela entraîne une sous-estimation de 14 850 000€ ce qui est loin d'être négligeable. Nous allons ainsi ajouter une pénalité particulièrement forte pour les erreurs de prédictions concernant les classes 4 et 5.

Nous en sommes arrivés à la formule suivante que nous appliquerons à la matrice de confusion obtenue par les valeurs prédites et les valeurs observées :

$$\text{Weighted Error (WE)} = \sum_{i=0}^{5}\sum_{j=0}^{5} \mathbb{I}_{\{i\neq j\}} \times |j-i| \times x_{i,j} \times (i+1) \times \gamma_i \tag{35}$$

où:

- $\mathbb{I}_{\{i\neq j\}}$: Est la fonction indicatrice pour appliquer une pénalité supplémentaire uniquement sur les valeurs incorrectement prédites quand $i \neq j$.
- $|j-i|$: Représente la distance séparant la classe observée i de la classe prédite j.
- $x_{i,j}$: représente le nombre d'éléments incorrectement classés.
- $(i+1) \times \gamma_i$: est un facteur de pénalité qui sera d'autant plus fort que i sera grand. Ce facteur est paramétrable selon le degré de pénalité que nous souhaitons appliquer aux classes. Par exemple, nous pouvons affecter 10 à γ_5 pour la classe 5, 5 à γ_4 pour la classe 4 et 1 pour le reste.

5.3.3 Modélisation par Extreme Gradient Boosting (XG Boost)

Nous sommes ici dans une logique de classifications des sinistres de grêles au niveau du département par évènement. Un large panel d'outils s'offre à nous pour modéliser ces différentes classes. Dans ce cadre, nous avons opté pour une modélisation avec le *XG Boost* qui présente de nombreux avantages dans notre cas de figure. En effet, il s'agit d'un outil qui permet de combiner différentes sources de données de nature différentes, avec ici des données historiques et climatiques qui ne sont sur les mêmes échelles de données. Nous n'avons ainsi pas besoin de normaliser ou de standardiser les données pour les exploiter. Cela est particulièrement intéressant dans le cadre de données climatiques où l'échelle des

variables peuvent présenter des écarts très importants. Par exemple, si l'on prend des données agrégées créées à partir de quantiles, nous allons rester sur des valeurs comprises en général entre 0 et 7 sauf valeurs extrêmes.

Cependant, selon l'ampleur de l'évènement, les données s'appuyant sur la somme des diamètres de grêle ou d'autres agrégations peuvent être bien plus importantes que 7. Nous pouvons dire la même chose de l'exposition qui peut varier du million au milliard par exemple. Ainsi, utiliser *XG Boost* permet de s'affranchir d'une standardisation des données et permet également d'accélérer la durée des traitements des données. De plus, cela permet à la fois de faire de la régression mais également de faire de la classification que nous allons employer dans notre modélisation en plus de méthodes qui permettent de réduire les erreurs à travers le *boosting* tout en limitant le sur-apprentissage avec des méthodes de régularisation qui aide à prévenir le surajustement à l'aide de pénalités ajoutées aux modèles complexes. Un autre facteur déterminant qui a orienté notre choix vers cet algorithme d'apprentissage est sa capacité à modéliser des relations complexes et non linéaires, ce qui est particulièrement adapté à notre problématique de modélisation des sinistres de grêle.

Nous avons la **variable cible** $C^i_{d,t}$ qui est notre classe de sinistre pour le département d et l'évènement t où $C^i_{d,t} \in \{0, 1, 2, 3, 4, 5\}$ et $i \in \{0, 1\}$ correspondant à la classe de l'évènement, 0 si l'évènement est d'intensité faible, auquel cas nous resterons en dessous du seuil de rétention de la réassurance, 1 si nous sommes au-dessus.

Ensuite, nous avons les variables historiques qui sont l'exposition, le pourcentage de maison dans le département, le niveau d'aléa du département défini par 20 ainsi que le nombre de communes touchées par la grêle par département et les variables météorologiques sont présentées dans la section 5.1.3.
Ainsi nous notons $\mathbf{X}_{d,t} = \{\text{variables historiques}, \text{variable météorologique}\}$. Le modèle construit est donc le suivant :

$$\widehat{C}^i_{d,t} = \arg\max_c P(C^i_{d,t} = c \mid X_{d,t}).$$

À partir des variables explicatives et pour un type d'intensité d'évènement, nous déterminons la probabilité d'appartenir à chacune des classes, puis nous prenons la classe associée à la plus grande probabilité pour obtenir la classe prédite. Ensuite, nous appliquons ce modèle en boucle à chaque variable climatique et nous allons retenir les variables climatiques qui minimisent la métrique d'erreur élaborée précédemment. Enfin, pour avoir

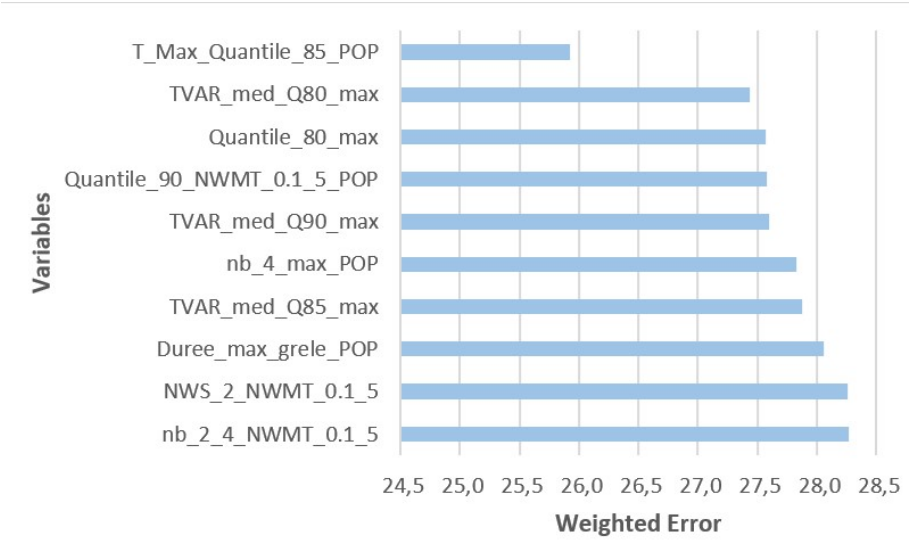
un résultat robuste, nous allons appliquer une méthode de validation croisée à partir des échantillons d'entraînement et de test élaborés grâce à l'algorithme 2. 10 échantillons au total sont générés, chaque échantillon servant tour à tour de test, le restant des données servent alors de données d'entraînement. Les données sont finalement agrégées à travers différentes métriques comme la moyenne, la médiane ou encore l'écart type pour identifier les variables qui présentent un faible biais et une faible variance.

Variables climatiques	RS	RMSE	MAE	**WE**	0_mean	1_mean	2_mean	3_mean	4_mean	4_median	4_sd	5_mean	5_median	5_sd
T_Max_Quantile_85_POP	0,38	0,85	0,56	**25,92**	0,87	0,24	0,60	0,70	0,23	0,12	0,20	0,60	1,00	0,52
TVAR_med_Q80_max	0,23	0,96	0,65	**27,43**	0,93	0,35	0,48	0,67	0,15	0,08	0,18	1,00	1,00	-
Quantile_80_max	0,26	0,94	0,64	**27,57**	0,93	0,37	0,49	0,66	0,14	0,04	0,18	1,00	1,00	-
Quantile_90_NWMT_5_POP	0,26	0,94	0,63	**27,58**	0,90	0,32	0,54	0,60	0,15	0,04	0,19	0,90	1,00	0,32
TVAR_med_Q90_max	0,24	0,95	0,64	**27,60**	0,96	0,28	0,53	0,62	0,15	0,08	0,20	1,00	1,00	-
nb_4_max_POP	0,30	0,91	0,59	**27,83**	0,96	0,31	0,58	0,65	0,15	0,08	0,11	1,00	1,00	-
TVAR_med_Q85_max	0,25	0,94	0,64	**27,88**	0,87	0,35	0,50	0,65	0,16	0,08	0,17	1,00	1,00	-
Duree_max_grele_POP	0,16	1,00	0,70	**28,06**	0,96	0,40	0,40	0,63	0,14	0,08	0,17	1,00	1,00	-
NWS_2_NWMT_5	0,27	0,93	0,60	**28,26**	0,96	0,28	0,60	0,58	0,23	0,23	0,13	1,00	1,00	-
nb_2_4_NWMT_5	0,28	0,92	0,60	**28,27**	0,93	0,28	0,58	0,60	0,15	0,12	0,14	1,00	1,00	-
TVAR_med_Q85_NWMT_5_POP	0,23	0,95	0,67	**28,29**	0,83	0,17	0,50	0,64	0,14	0,08	0,18	0,90	1,00	0,32
Quantile_95_max	0,21	0,97	0,66	**28,40**	0,96	0,33	0,50	0,62	0,16	0,08	0,21	1,00	1,00	-
TVAR_med_Q90_NWMT_5_POP	0,31	0,91	0,59	**28,45**	0,90	0,31	0,58	0,68	0,15	0,08	0,19	0,90	1,00	0,32
Duree_grele	0,30	0,91	0,60	**28,52**	0,94	0,20	0,58	0,65	0,12	0,08	0,13	0,90	1,00	0,32
TVAR_med_Q80_NWMT_5	0,15	1,01	0,69	**28,60**	0,87	0,31	0,46	0,65	0,18	0,08	0,22	0,80	1,00	0,42
WSS_0_sum_POP	0,29	0,92	0,62	**28,77**	0,91	0,19	0,53	0,67	0,22	0,15	0,16	0,50	0,50	0,53
TVAR_med_Q85_max_POP	0,23	0,95	0,66	**28,81**	0,83	0,29	0,50	0,63	0,14	0,04	0,20	0,90	1,00	0,32
TVAR_mean_Q85_max_POP	0,23	0,95	0,64	**28,81**	0,90	0,29	0,54	0,62	0,15	0,08	0,21	0,90	1,00	0,32
NWS_2_sum_POP	0,31	0,90	0,56	**28,84**	0,96	0,37	0,62	0,61	0,22	0,19	0,12	0,50	0,50	0,53
WNES_0_2_max_POP	0,21	0,97	0,67	**28,89**	0,80	0,30	0,48	0,66	0,21	0,23	0,11	0,50	0,50	0,53
TVAR_mean_Q90_max_POP	0,26	0,94	0,64	**28,89**	0,90	0,24	0,53	0,63	0,15	0,08	0,17	0,90	1,00	0,32
TVAR_med_Q90_NWMT_5	0,21	0,97	0,66	**29,04**	0,94	0,28	0,50	0,62	0,19	0,08	0,23	0,80	1,00	0,42
T_Max_Quantile_95_POP	0,32	0,90	0,59	**29,08**	0,90	0,27	0,58	0,69	0,12	0,08	0,15	0,50	0,50	0,53
WNEM_0_4_NWMT_5_POP	0,28	0,93	0,64	**29,20**	0,93	0,15	0,54	0,60	0,18	0,12	0,15	0,50	0,50	0,53
WNES_0_4_NWMT_5_POP	0,26	0,94	0,63	**29,26**	0,82	0,22	0,55	0,68	0,18	0,19	0,14	0,50	0,50	0,53

Table 18: Extrait de la modélisation de la classe de sinistralité pour les évènements de classe 1

Les données obtenues dans la table 18 sont triées par ordre croissant de la *Weighted Error* (WE) qui est notre métrique de mesure de l'erreur. Ensuite, les données *0_mean* à *5_mean* représentent les proportions d'individus correctement identifiés dans chaque classe. Si nous prenons la troisième ligne, nous lisons que pour le *Quantile_80_max*, nous avons une mesure de l'erreur WE estimée à 27,57 avec en moyenne sur les différents échantillons, 93% des départements de classe 0 correctement identifiés, 35% pour la classe 1 ou encore 100% des individus en classe 5 avec cependant un écart type pour la classe 5 de 52% indiqué par la colonne 5_*sd*. Nous observons que pour les variables avec une *Weighted Error* faible nous avons en moyenne un très bon taux d'individus de classe 5 identifiés, que ce soit selon la moyenne ou selon la médiane. Nous faisons en sorte de ne sélectionner que des éléments avec un écart type inférieur à 60%. A côté, nous continuons d'indiquer des métriques d'erreurs classiques comme le R^2 (Coefficient de détermination), le RMSE (*Root Mean Square Error*), ou encore le MAE (*Mean Absolute Error*). Nous notons un R^2 peu satisfaisant égal 0.38 pour le meilleur indicateur. Cependant, comme nous l'avons vu précédemment, le R^2 n'est pas un bon indicateur dans notre cas car nous accordons plus d'importances aux éléments de classe 4 et 5, d'où un excellent score

Weighted Error, qui est notre métrique de mesure d'erreur personnalisée, malgré un R^2 faible. La MAE nous indique tout de même une information intéressante, globalement, nous nous trempons en moyenne de moins d'une classe. De cette table, nous décidons de ne garder que les 10 premières variables.



Graphique 40: Sélection de variables par XG Boost - évènement de classe 1

Étudions quelques variables issues de notre sélection. $Max_Quantile_85_POP$ est une variable ayant d'abord été agrégée selon le maximum de la taille des grêlons par département normalisé par la densité de la population, et ensuite nous prenons le quantile à l'ordre 85% sur les différentes heures constituant l'évènement. Cela peut en effet être un bon indicateur d'un évènement qui a touché un département à différents moments de la journée. Le fait d'inclure la population permet de prendre en compte les grêlons tombés dans des zones densément peuplées. En effet, nous pouvons raisonnablement penser qu'un évènement de grêle touchant une commune peu densément peuplée aura une plus faible exposition et ainsi une charge réduite.

La TVAR nous permet de nous concentrer sur ce les comportements de la grêle dans les quantiles les plus élevés de la distribution. Cependant, contrairement à la TVAR classique, nous prenons la médiane et non la moyenne des valeurs au-delà du quantile désiré estimant que la médiane est moins sensible aux valeurs extrêmes que la moyenne.

Ensuite la variable $Duree_max_grele_POP$ permet de récupérer par département l'évènement de grêle le plus long repéré avec des tailles de grêlons au-delà de 2 cm de diamètre. Nous pouvons ainsi nous dire qu'un évènement ayant duré longtemps va occa-

sionner davantage de dommage.

Enfin, nous avons le $nb_2_4_NWMT_5$. Cela signifie que nous avons d'abord relevé le nombre de grêle compris entre 2 et 4 cm de diamètre par département, puis pour le premier, nous avons pris la moyenne non pondérée (*Non Weighted Mean Top* - NWMT) des 5 plus grandes valeurs. Cela peut révéler une information précieuse sur la grêle car ce sont souvent ces grêlons qui sont les plus fréquents et qui peuvent occasionner des dégâts au-delà de 2 cm de diamètre.

Maintenant que nous avons sélectionné les variables qui nous semblent les plus pertinentes afin de modéliser cette classe de sinistres, l'objectif final est de sélectionner et d'enregistrer un modèle *XG Boost* optimal pour chaque variable climatique, en retenant celui qui minimise la métrique d'erreur (*WE*) parmi les 10 échantillons obtenus grâce à la méthode expliquée dans le schéma 37. Le modèle *XG Boost* ainsi enregistré capture la structure de l'arbre de décision et les poids associés à chaque point de séparation (aussi appelé *split*), mais aussi les différents hyperparamètres et l'importance des différentes variables explicatives dans le modèle.

variable	Échantillon	RS	RMSE	MAE	WE	Classe 0	Classe 1	Classe 2	Classe 3	Classe 4	Classe 5
Duree_max_grele_POP	1	0,14	1,01	0,72	25,31	1,00	0,38	0,38	0,42	0,23	1,00
Duree_max_grele_POP	2	0,04	1,07	0,76	34,15	1,00	0,46	0,38	0,58	0,08	1,00
Duree_max_grele_POP	3	0,19	0,98	0,71	30,33	0,86	0,38	0,44	0,50	-	1,00
Duree_max_grele_POP	4	0,06	1,06	0,73	26,19	1,00	0,31	0,38	0,67	0,38	1,00
Duree_max_grele_POP	5	0,22	0,96	0,67	28,01	1,00	0,38	0,40	0,83	0,08	1,00
Duree_max_grele_POP	6	0,14	1,01	0,72	33,26	0,86	0,46	0,44	0,50	-	1,00
Duree_max_grele_POP	7	0,20	0,98	0,66	19,08	1,00	0,38	0,40	0,67	0,46	1,00
Duree_max_grele_POP	8	0,33	0,89	0,63	27,39	1,00	0,46	0,40	0,83	-	1,00
Duree_max_grele_POP	9	0,03	1,08	0,76	29,22	1,00	0,31	0,44	0,50	0,15	1,00
Duree_max_grele_POP	10	0,28	0,93	0,67	27,69	0,86	0,46	0,35	0,83	-	1,00

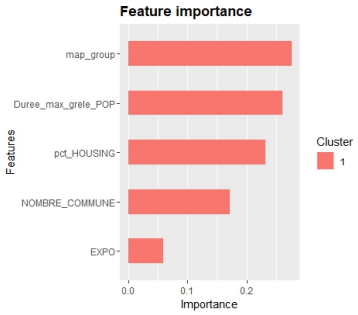
Table 19: Résultats de la variable Duree_max_grele_POP sur les différents échantillons

Ainsi, en nous focalisons sur $Duree_max_grele_POP$ par exemple sur les différents échantillons, nous remarquons que l'échantillon sur lequel le *XG Boost* a minimisé notre métrique d'erreur est le 7ème échantillon.

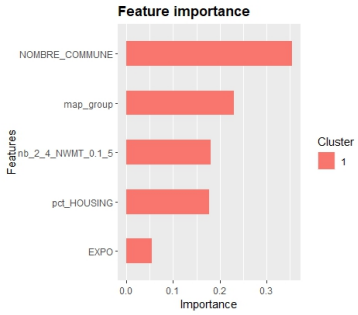
variable	Échantillon	RS	RMSE	MAE	WE	Classe 0	Classe 1	Classe 2	Classe 3	Classe 4	Classe 5
Duree_max_grele_POP	1	0,25	0,95	0,62	16,75	1,00	0,38	0,44	0,67	0,54	1,00
Duree_max_grele_POP	2	0,29	0,92	0,60	18,05	1,00	0,46	0,44	0,75	0,46	1,00
Duree_max_grele_POP	3	0,25	0,95	0,62	19,09	1,00	0,46	0,46	0,58	0,46	1,00
Duree_max_grele_POP	4	0,27	0,93	0,60	18,08	1,00	0,54	0,44	0,75	0,46	1,00
Duree_max_grele_POP	5	0,32	0,90	0,57	17,59	1,00	0,46	0,46	0,75	0,46	1,00
Duree_max_grele_POP	6	0,35	0,88	0,54	17,42	1,00	0,62	0,48	0,75	0,46	1,00
Duree_max_grele_POP	7	0,20	0,98	0,66	19,08	1,00	0,38	0,40	0,67	0,46	1,00
Duree_max_grele_POP	8	0,33	0,89	0,56	15,79	1,00	0,46	0,46	0,75	0,54	1,00
Duree_max_grele_POP	9	0,29	0,92	0,59	18,30	1,00	0,54	0,46	0,67	0,46	1,00
Duree_max_grele_POP	10	0,39	0,85	0,53	14,08	1,00	0,46	0,42	1,00	0,54	1,00

Table 20: Résultats de cette variable après sélection

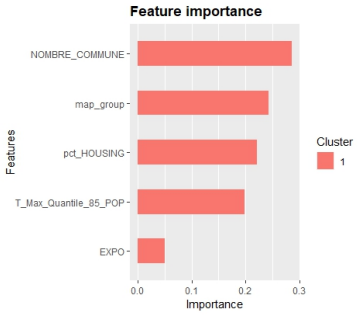
Nous décidons donc d'enregistrer le modèle associé à cet échantillon. Nous appliquons le même procédé pour les 9 autres variables climatiques.



Graphique 41: Importance des variables pour le modèle XGB avec duree_max_grele_pop



Graphique 42: Importance des variables pour le modèle XGB avec nb_2_4_NWMT_0.1_5



Graphique 43: Importance des variables pour le modèle XGB avec T_Max_Quantile_85_POP

Dans les graphiques précédents, nous montrons les visualisations obtenues avec la fonction *plot_importance()* de *XG Boost*. Ces graphiques illustrent l'impact relatif de chaque variable sur les prédictions du modèle. Cette notion d'importance prend en compte le nombre de fois où une variable est utilisée pour diviser les données dans l'ensemble des arbres du modèle ainsi que l'amélioration de la performance du modèle apportée par chacune des variables indiquées. Cela nous permettra ainsi de comprendre les résultats modélisés par la suite. Nous remarquons qu'en dehors des variables issues des agrégations climatiques étudiées précédemment, 3 variables reviennent régulièrement dans les variables les plus importantes:

- *map_group* : il s'agit du niveau d'aléa affecté à chaque département à partir de la base de données ESWD.
- *NOMBRE_COMMUNE* : cette variable représente le nombre de commune touchées par l'évènement de grêle selon ESWD.
- *pct_HOUSING* : il s'agit du pourcentage de maison de chaque département d'après le portefeuille d'AXA

Il nous semble ainsi cohérent que ces variables puissent justifier d'un département avec une grande classe de sinistre si un département est touchés sur plusieurs communes, avec un pourcentage élevé de maison plus vulnérables à la grêle en plus d'être sur un département avec un fort aléa de grêle.

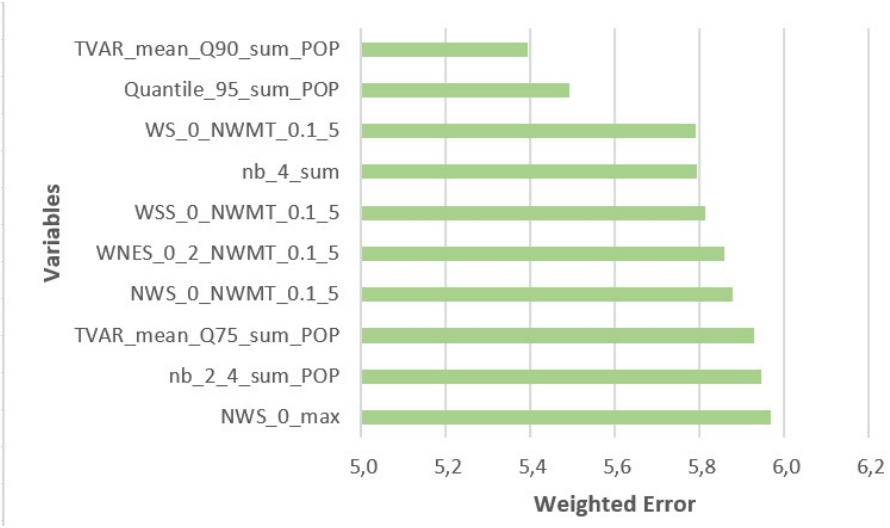
Résultats avec les évènements de classe 0:

De la même manière, nous retrouvons, les résultats pour les évènements de classe 0 dans la table 21.

Variables	RS	RMSE	MAE	WE	0_mean	1_mean	2_mean	3_mean	4_mean	4_median	4_sd	5_mean	5_median	5_sd
TVAR_mean_Q90_sum_POP	0,04	0,96	0,63	5,39	0,48	0,23	0,66	0,41	0,85	1,00	0,24	-1,00	-1,00	-
Quantile_95_sum_POP	0,05	0,95	0,63	5,49	0,45	0,16	0,71	0,44	0,75	0,75	0,26	-1,00	-1,00	-
WS_0_NWMT_5	0,03	0,96	0,64	5,79	0,53	0,14	0,65	0,50	0,75	1,00	0,35	-1,00	-1,00	-
nb_4_sum	-0,10	1,02	0,71	5,79	0,45	0,15	0,60	0,50	0,80	1,00	0,35	-1,00	-1,00	-
WSS_0_NWMT_5	0,00	0,98	0,66	5,81	0,51	0,18	0,63	0,44	0,75	0,75	0,26	-1,00	-1,00	-
WNES_0_2_NWMT_5	-0,01	0,98	0,66	5,86	0,51	0,12	0,65	0,50	0,75	0,75	0,26	-1,00	-1,00	-
NWS_0_NWMT_5	0,01	0,97	0,65	5,88	0,51	0,16	0,66	0,46	0,70	0,75	0,35	-1,00	-1,00	-
TVAR_mean_Q75_sum_POP	0,06	0,95	0,63	5,93	0,46	0,15	0,72	0,44	0,65	0,75	0,41	-1,00	-1,00	-
nb_2_4_sum_POP	-0,08	1,01	0,68	5,95	0,40	0,18	0,68	0,47	0,70	0,75	0,35	-1,00	-1,00	-
NWS_0_max	-0,03	0,99	0,67	5,97	0,46	0,19	0,64	0,49	0,65	0,50	0,34	-1,00	-1,00	-
Quantile_95_NWMT_5_POP	0,07	0,94	0,62	6,02	0,46	0,17	0,72	0,41	0,65	0,50	0,34	-1,00	-1,00	-
WNES_0_4_NWMT_5	-0,03	0,99	0,66	6,03	0,50	0,14	0,65	0,51	0,70	0,75	0,35	-1,00	-1,00	-
T_Max_Quantile_90_POP	0,13	0,91	0,58	6,03	0,47	0,18	0,78	0,34	0,60	0,50	0,39	-1,00	-1,00	-
Quantile_95_sum	-0,04	1,00	0,67	6,04	0,47	0,22	0,65	0,47	0,65	0,50	0,34	-1,00	-1,00	-
NWS_2_sum	0,02	0,97	0,65	6,09	0,45	0,13	0,68	0,48	0,60	0,75	0,46	-1,00	-1,00	-
nb_2_4_max_POP	0,01	0,97	0,64	6,10	0,44	0,17	0,71	0,40	0,65	0,50	0,34	-1,00	-1,00	-
Quantile_99_sum_POP	0,15	0,90	0,56	6,10	0,58	0,22	0,73	0,46	0,65	0,50	0,24	-1,00	-1,00	-
T_Max_Quantile_99_POP	0,01	0,97	0,64	6,11	0,48	0,21	0,69	0,39	0,65	0,50	0,24	-1,00	-1,00	-
WNES_0_4_sum	-0,01	0,98	0,64	6,11	0,58	0,14	0,68	0,46	0,65	0,75	0,41	-1,00	-1,00	-
nb_2_4_NWMT_5_POP	-0,05	1,00	0,68	6,13	0,39	0,17	0,69	0,38	0,70	0,75	0,35	-1,00	-1,00	-
T_Max_Quantile_99	-0,10	1,02	0,69	6,15	0,43	0,21	0,65	0,48	0,70	0,75	0,35	-1,00	-1,00	-
NWS_2_NWMT_5	0,02	0,97	0,65	6,18	0,52	0,18	0,62	0,48	0,65	0,50	0,34	-1,00	-1,00	-
NWS_2_max_POP	0,05	0,95	0,63	6,18	0,47	0,22	0,68	0,41	0,60	0,50	0,32	-1,00	-1,00	-
Quantile_95_NWMT_5	-0,02	0,99	0,67	6,19	0,46	0,15	0,64	0,50	0,60	0,50	0,32	-1,00	-1,00	-
Var_NWMT_5_POP	0,05	0,96	0,63	6,19	0,49	0,14	0,72	0,39	0,60	0,50	0,39	-1,00	-1,00	-

Table 21: Extrait de la modélisation de la classe de sinistralité pour les évènements de classe 0

À première vue, des nombres négatifs dans la colonne de la classe 5 peut interroger. Il s'agit simplement d'une méthode dans l'algorithme qui permet de détecter l'absence d'une classe dans les données. En effet, les évènements de classe 0 n'ont pas de départements touchés à plusieurs millions d'euros, ce qui peut en partie expliquer le fait que la rétention ne soit pas touchée. Ces éléments sont donc indiqués avec un -1 par défaut. Comme précédemment, nous observons que le R^2 est très faible mais cela nous importe peu. Les métriques d'erreurs sont faibles, mais cela est lié à l'absence de classe 5 ce qui diminue grandement l'ampleur des pénalités d'erreurs. Maintenant, les plus grandes pénalités sont celles associées à la classe 4. Nous retenons comme précédemment les 10 variables climatiques ayant minimisé notre métrique d'erreur.

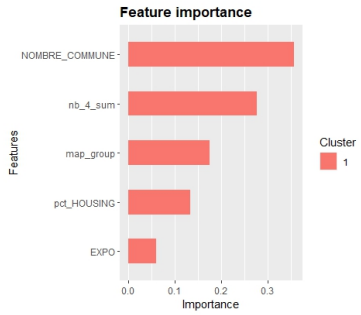


Graphique 44: Sélection de variables par XG Boost - évènement de classe 0

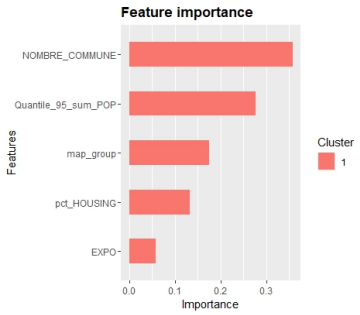
Nous retrouvons ici quelques variables que nous avions également dans le cas des évènements de classe 1. Nous allons donc nous focaliser sur les nouvelles agrégations comme le $WS_0_NWMT_5$ qui est la somme au carrée des diamètres de grêlons au niveau du département, puis nous en prenons la moyenne sur le pas de temps horaire sur l'évènement. Il peut ainsi s'agir d'un bon indicateur de l'ampleur de la grêle.

Nous avons également $WNES_0_2_NWMT_5$ qui représente la somme pondérée de l'exponentiel des diamètres de grêlons divisés par 2 mettant davantage en exergue les grêlons de plus de 2 cm de diamètre qui commencent à faire des dégâts sur les différents biens.

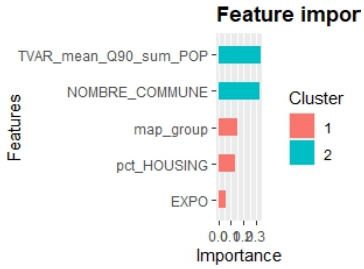
Enfin, de la même manière que pour les évènements de classe 1, nous regardons la métrique d'erreur associées aux variables climatiques sur les différents échantillons et nous gardons le modèle associé à la plus petite métrique d'erreur. Puis nous affichons le graphique de l'importance des variables de chacun de ces modèles et nous remarquons que celles qui sont les plus importantes sont celles relatives encore une fois aux agrégations réalisées, du nombre de communes touchées et du niveau d'aléa du département.



Graphique 45: Importance des variables pour le modèle XGB avec nb_4_sum



Graphique 46: Importance des variables pour le modèle XGB avec Quantile_95_sum_POP



Graphique 47: Importance des variables pour le modèle XGB avec TVAR_mean_Q90_sum_POP

5.4 Modélisation des taux de destruction

À partir de nos nouveaux évènements ainsi que de la modélisation des classes de sinistres, nous avons une approche plus précise de l'ordre de grandeur des sinistres en tenant compte des effets de seuil. Maintenant, nous allons associer ces données à nouveau à des données météorologiques et autres données du portefeuille afin de déterminer un taux de destruction, puis appliquer l'exposition correspondant à chaque département pour obtenir la charge de sinistralité de chaque département.

Le taux de destruction est défini comme suit :

$$TD_{d,t} = \frac{S_{d,t}}{E_{d,t}},$$

où $S_{d,t}$ représente le montant du sinistre agrégé dans le département d sur l'évènement t et $E_{d,t}$ représente l'exposition du portefeuille dans le département d sur l'évènement t.

Un problème s'est alors posé à nous au début de la modélisation lié à la variable du taux de destruction. Nous avons observé que les taux de destruction initiaux étaient extrêmement faibles, généralement de l'ordre de 10^{-5} ou 10^{-6} par exemple en raison des grandes expositions du portefeuille. Ces valeurs présentaient également une distribution fortement asymétrique étalée sur une grande plage entre 0 et 1, ce qui compliquait leur modélisation. En effet, un simple écart de l'ordre de 10^{-3} pourrait entraîner par la suite plusieurs millions d'euros d'écart une fois l'exposition appliquée.

Pour remédier à cette situation, nous avons décidé de réaliser une transformation logarithmique des données, permettant ainsi de gérer la large plage de valeurs et de réduire

l'influence des valeurs extrêmes. Cette transformation a conduit à une nouvelle variable, TD_LOG, dont les valeurs se situent désormais entre 0 et 4, avec une plus grande concentration autour de certaines valeurs par rapport aux valeurs initiales qui étaient très dispersées. Nous avons veillé à ce que la distribution de la variable transformée conserve les caractéristiques principales de la distribution originale de TD afin de maintenir une cohérence dans nos analyses. Cette transformation logarithmique ne sera cependant pas détaillée dans le cadre de ce mémoire pour des raisons de confidentialité des données.

Nous avons ainsi :

$$TD_LOG_{d,t} = f(TD_{d,t}).$$

où $f(.)$ est un transformation logarithmique appliquée aux taux de destruction initiaux.

Nous cherchons donc désormais à modéliser cette nouvelle variable à savoir TD_LOG à partir de nos données historiques ainsi que nos agrégations de données climatiques. Nous allons encore une fois séparer les évènements de classe 1 ayant dépassé le seuil de rétention et les évènements de classe 0 étant en dessous de ce seuil. Le modèle s'écrit alors de la manière suivante : Avec :

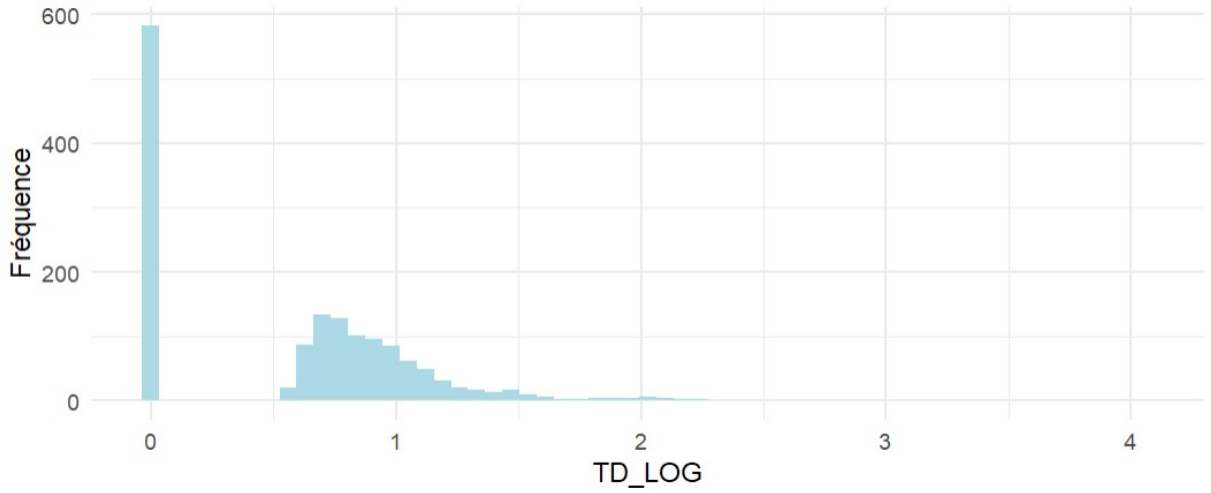
$$\mathbf{Z}^i_{d,t} = \{\text{variables historiques}, \text{variables météorologiques}, \mathbf{C}^i_{d,t}\}.$$

où $C^i_{d,t}$ représente les classes de sinistre pour le département d sur l'évènement t et pour l'évènement de classe i. Nous allons dans un premier temps utiliser les classes de sinistres observés dans l'historique afin de calibrer au mieux notre modèle en partant du principe que le modèle précédent a permis de capter parfaitement les classes de sinistres. Puis dans un second temps nous utiliserons $\widehat{C}^i_{d,t}$, à savoir les classes de sinistres prédites obtenus à partir de notre modèle.

Nous cherchons à obtenir :

$$TD_LOG^i_{d,t} = g(\mathbf{Z}^i_{d,t}; \boldsymbol{\beta}),$$

où $g(\mathbf{Z}^i_{d,t}; \boldsymbol{\beta})$ = est combinaison linéaire des variables explicatives.



Graphique 48: Distribution de la variable TD_LOG

L'histogramme du graphique 48 nous indique ainsi la distribution de la variable TD_LOG à partir des données recueillies à l'échelle des départements de la France métropolitaine sur les différents évènements. Nous observons une certaine concentration des données autour de 0.6 environ avec une décroissance à gauche et à droite de cette valeur avec une décroissance plus rapide à gauche. De plus, toutes les valeurs nulles représentent les départements qui n'ont pas été touchés par la grêle qui sont ainsi logiquement plus nombreux, la grêle étant un phénomène local. En effet, la modélisation précédente des classes de sinistralité par département nous permet d'identifier l'absence de sinistres. Cette situation correspond précisément aux cas où la classe de sinistre prédite est égale à 0. En étudiant les différentes familles de distribution et de modélisation, au vu des éléments précédemment cités, nous nous tournons vers une modélisation avec un modèle linéaire généralisé avec une distribution gaussienne.

Pour ce faire, nous allons adopter une démarche similaire que pour la classification des sinistres. Nous allons tout d'abord séparer notre base de données en deux bases, une qui nous servira d'échantillon test, et l'autre d'échantillon d'entraînement. Pour les générer, nous allons tout d'abord utiliser l'algorithme de création d'évènement pour créer l'évènement test, tous les éléments non utilisés serviront alors dans l'échantillon d'entraînement. Nous répétons ce mécanisme 2 fois : une fois pour les évènements que nous avons classés de classe 1, à savoir les évènements touchant la rétention, et ceux de classe 0 pour ceux qui ne l'ont pas atteint. Ensuite, parmi l'échantillon d'entraînement, nous créons à nouveau 10 échantillons d'évènements afin de contrôler le biais et la variance et éviter un potentiel sur-apprentissage. Les variables explicatives sont dans notre cas le pourcentage de maison au sein du département *pct_housing*, le niveau d'aléa du département *map_group* ainsi que les classes de sinistres allant de 0 à 5.

Afin d'utiliser cette dernière variable catégorielle dans notre modélisation, nous allons recourir à la création de variables indicatrices, aussi appelées variables «*dummy*». Cette approche est également connue en tant que codage *one-hot*. Cela consiste à transformer la variable catégorielle en plusieurs variables binaires prenant ainsi 0 ou 1 comme valeurs, chacune représentant une catégorie spécifique. Ainsi, pour notre variable à 6 niveaux (0 à 5), nous allons créer 5 variables *dummy*, la catégorie de référence, ici la catégorie de sinistre 0 étant implicitement représentée par l'absence de 1 dans toutes les autres variables. Le fait d'exclure cette catégorie permet d'éviter une multicolinéarité parfaite. En effet, la catégorie de sinistre 0 peut être obtenue à partir des des autres catégories de sinistre :

$$INC0 = 1 - (INC1 + INC2 + INC3 + INC4 + INC5)$$

où $INCi$ représente la variable indicatrice ou *dummy* pour la catégorie i du sinistre.

En plus de ces variables, nous avons les agrégations des données Meteomatics. Ces variables climatiques étant fortement corrélées, nous allons les tester une par une avec nos autres variables. Nous allons donc réaliser une boucle en partant du principe que nous connaissons parfaitement les classes de sinistres. L'objectif du modèle 2 ici, est de trouver le meilleur modèle qui à partir des classes de sinistres parfaitement prédites, nous permet d'approcher au mieux du TD_LOG observé. Ainsi dans le premier temps nous modélisons la catégorie de sinistre, et ici nous modélisons le taux de destruction sachant une connaissance parfaite de cette catégorie. Ce sont deux modèles calibrés indépendamment que nous allons combiner dans un troisième temps.

5.4.1 Modélisation appliquée aux évènements de classe 1

Nous obtenons le tableau 22. Il montre pour chaque ligne la l'agrégation climatique utilisée lors de la boucle ainsi que sa performance sur les 10 échantillons. Nous observons que les R^2 sur la base d'entraînement ($R^2 train.td.log$) et sur la base test ($R^2 test.td.log$) sont très bons pour les taux de destruction transformés, de l'ordre de 99% pour les entraînements et entre 93% et 99% pour les bases de test en moyenne avec un très faible écart type. Nous décidons de dé-transformer TD_LOG pour obtenir le taux de destruction $\widehat{TD}$. Cette opération s'effectue en appliquant la fonction inverse f^{-1} comme suit :

$$\widehat{TD^1_{d,t}} = f^{-1}(\widehat{TD_LOG^1_{d,t}})$$

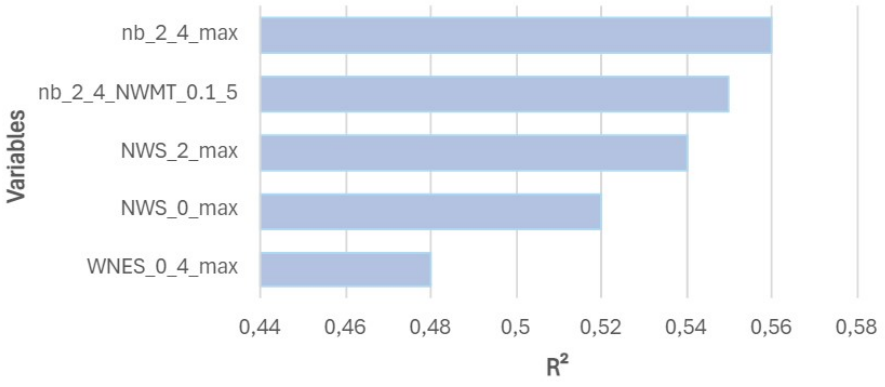
Une fois le taux de destruction $\widehat{TD^1_{d,t}}$ obtenu, nous l'appliquons à l'exposition correspondante du département $E_{d,t}$ pour calculer la sinistralité $\widehat{S^1_{d,t}}$:

$$\widehat{S^1_{d,t}} = \widehat{TD^1_{d,t}} \times E_{d,t}$$

Nous pouvons ainsi observer les R^2 des sinistres ($R^2 test.I$) ainsi modélisés pour voir à quel point nous nous éloignons ou non des valeurs observées. Nous oscillons dans la base d'entraînement autour de 0,95 mais cela chute à environ 0,54 environ pour les meilleures valeurs de la table avec un écart type conséquent. Cela vient notamment du fait qu'une faible erreur de prédiction en termes de taux de destruction peut entraîner de forts écarts avec l'exposition. De plus, ces écarts proviennent notamment des plus gros sinistres qui baissent le résultat global de l'ensemble des prédictions.

Variables	R^2train.td.log	R^2train.td.log.sd	R^2train.I	R^2train.I.sd	écart.train	R^2test.td.log	R^2test.td.log.sd	R^2test.I	R^2test.I.sd	ecart.test
nb_2_4_max	0,99	0,00	0,96	0,02	1%	0,94	0,01	0,56	0,48	-32%
nb_2_4_NWMT_5	0,99	0,00	0,96	0,02	-2%	0,93	0,01	0,55	0,48	-33%
NWS_2_max	0,99	0,00	0,99	0,01	-1%	0,93	0,01	0,54	0,52	-41%
NWS_0_max	0,99	0,00	0,97	0,02	-1%	0,93	0,01	0,52	0,52	-35%
WNES_0_4_max	0,99	0,00	0,98	0,01	-3%	0,93	0,02	0,48	0,56	-43%
TVAR_med_Q80_sum	0,99	0,00	0,94	5	-0,03%	0,93	0,01	0,47	0,47	-33%
WS_0_max	0,99	0,00	0,98	0,01	-3%	0,93	0,02	0,47	0,57	-45%
Quantile_80_sum	0,99	0,00	0,94	4	-0,03%	0,93	0,01	0,46	0,50	-33%
nb_4_max_POP	0,99	0,00	0,97	2	-0,01%	0,93	0,01	0,46	0,59	-40%
Quantile_75_sum	0,99	0,00	0,94	0,05	-3%	0,93	0,01	0,45	0,48	-32%
nb_2_4_NWMT_5_POP	0,99	0,00	0,92	6	-4%	0,92	0,01	0,45	0,50	-29%
Nb_grele	0,99	0,00	0,92	0,07	-4%	0,92	0,01	0,44	0,45	-27%
nb_4_NWMT_5_POP	0,99	0,00	0,96	0,03	-3%	0,92	0,02	0,44	0,59	-41%
WNES_0_4_NWMT_5	0,99	0,00	0,96	0,03	-2%	0,93	0,02	0,44	0,61	-40%
nb_2_4_sum	0,99	0,00	0,93	0,05	-3%	0,93	0,01	0,44	0,56	-29%
TVAR_mean_Q80_sum	0,99	0,00	0,93	0,05	-3%	0,93	0,01	0,43	0,53	-32%
Quantile_85_NWMT_5	0,99	0,00	0,93	0,05	-4%	0,93	0,01	0,43	0,48	-40%
WS_0_NWMT_5	0,99	0,00	0,96	0,02	-2%	0,93	0,02	0,42	0,63	-42%
Quantile_50_max	0,99	0,00	0,94	0,04	-1%	0,93	0,01	0,42	0,55	-37%
Quantile_50_max_POP	0,99	0,00	0,94	0,04	-2%	0,92	0,01	0,42	0,54	-40%
TVAR_med_Q80_sum_POP	0,99	0,00	0,93	0,05	-4%	0,93	0,01	0,42	0,51	-38%
TVAR_mean_Q80_NWMT_5	0,99	0,00	0,93	0,05	-4%	0,93	0,01	0,42	0,53	-38%
TVAR_med_Q75_sum	0,99	0,00	0,93	0,05	-3%	0,93	0,01	0,41	0,54	-34%
Quantile_85_sum	0,99	0,00	0,93	0,05	-3%	0,93	0,01	0,41	0,55	-35%
Quantile_75_sum_POP	0,99	0,00	0,93	0,05	-3%	0,92	0,01	0,40	0,53	-38%

Table 22: Résultats de la modélisation du taux de destruction pour les évènements de classe 1



Graphique 49: Sélection de variables par GLM - évènement de classe 1

Parmi l'extrait affiché dans la figure 22, nous retenons les 5 premières variables dans la figure 49 selon la performance en R^2 au niveau des sinistres prédits.

Nous observons ainsi que d'après notre méthode, 5 agrégations climatiques semblent bien modéliser notre cas. Essayons de comprendre ces variables. Tout d'abord nous avons le $nb_2_4_max$ et le $nb_2_4_NWMT_5$ cela signifie que nous avons d'abord relevé le nombre de grêle compris entre 2 et 4 cm de diamètre par département, puis pour le premier, nous avons pris la valeur maximale de ce nombre sur le pas de temps horaire de l'évènement dans le premier cas, et la moyenne non pondérée (*Non Weighted Mean Top*) des 5 plus grandes valeurs dans le second cas. Ces variables peuvent révéler une information précieuse sur la grêle car ce sont souvent ces grêlons qui sont les plus fréquents et qui peuvent occasionner des dégâts au-delà de 2 cm de diamètre.

Ensuite nous avons NWS_0_max et NWS_2_max, qui représente à l'échelle du département la somme non pondérée des diamètres de grêlons dans le premier cas, et la somme non pondérée des grêlons dont le diamètre est supérieur à 2 cm, et dans un second temps nous prenons la valeur maximale sur l'évènement. Cette simple somme selon le montant peut être un indicateur de l'ampleur des grêlons tombés au sol. La valeur avec un seuil de 2 cm est encore une fois révélatrice des grêlons qui occasionnent le plus de dommages.

Enfin nous avons $WNES_0_4_max$ qui est moins facilement interprétable. Il s'agit de la somme pondérée par l'exponentielle du diamètre de grêle tombé divisé par un seuil qui est établi à 4 cm de diamètre ici. Cette normalisation par 4 permet de mettre davantage en exergue les grêlons dont la taille de diamètre est supérieure à 4 cm et révélant un évènement qui va davantage toucher une question de sévérité. En effet, les grandes tailles de grêles sont moins fréquentes. Mettre ce nombre à l'exponentiel permet de mettre davantage en valeur ces cas-là et cet indice peut ainsi révéler un évènement où des grêlons de grandes tailles sont davantage tombés.

Maintenant, notre objectif est de retenir pour chacune de ces variables des coefficients qui permettent de modéliser le taux de destruction transformé. En raison d'une base de données qui n'est pas riche en évènement, nous décidons de regarder les résultats de chaque modèle sur les différents échantillons que nous pouvons voir dans le tableau 23, d'en récupérer les coefficients afin d'en prendre la moyenne pour éviter d'utiliser des coefficients obtenus sur des échantillons où il y a eu potentiellement du sur-apprentissage sur quelques spécificités des données. Nous vérifions à titre d'information que le modèle

final obtenu reste cohérent avec ce que nous cherchons à modéliser, visible dans la table 26. Nous allons nous focaliser sur le cas de la modélisation avec la variable climatique agrégée $nb_2_4_NWMT_5$.

Variable	Échantillon	R^2train.td.log	R^2train.I	écart_train	R^2test.td.log	R^2test.I	écart_test
nb_2_4_NWMT_5	1	0,99	0,95	-9%	0,92	-0,09	-94%
nb_2_4_NWMT_5	2	0,99	0,98	1%	0,94	0,75	-1%
nb_2_4_NWMT_5	3	0,99	0,98	1%	0,94	0,76	3%
nb_2_4_NWMT_5	4	0,99	0,93	-1%	0,92	0,88	-41%
nb_2_4_NWMT_5	5	0,99	0,98	0%	0,95	0,76	-0%
nb_2_4_NWMT_5	6	0,99	0,98	1%	0,93	0,78	1%
nb_2_4_NWMT_5	7	0,99	0,93	-4%	0,92	0,86	-47%
nb_2_4_NWMT_5	8	0,99	0,97	-6%	0,90	-0,56	-121%
nb_2_4_NWMT_5	9	0,99	0,98	1%	0,93	0,74	3%
nb_2_4_NWMT_5	10	0,99	0,98	-6%	0,93	0,57	-29%

Table 23: Entraînement des modèles sur les échantillons

Nous observons que les résultats du R^2 au niveau des taux de destructions transformés sont toujours bons, nous pensons ainsi avoir un bon modèle mais lorsque nous appliquons l'exposition, nous arrivons à obtenir des R^2 intéressants dans la plupart des cas avec 7 échantillons sur 10 avec un R^2 au niveau du sinistre estimé supérieur à 74% mais les 3 autres cas sont plutôt mauvais. Cela peut provenir d'un ou deux sinistres extrêmes mal modélisés ici ou encore de notre variable d'exposition qui ne tient pas assez compte de l'exposition au risque de grêle. Il faut ainsi garder en tête que la majorité des écarts observés viennent d'une poignée de département, notamment les sinistres de classe 5 qui coûtent plusieurs dizaines de millions d'euros, donc un écart de 10 millions d'euros sur cette ligne peut entraîner un écart conséquent au global mais ce n'est pas forcément un indicateur d'un mauvais modèle.

Désormais, nous allons nous atteler à obtenir les coefficients finaux à utiliser dans notre modèle, nous avons décidé de prendre la moyenne des coefficients sur chaque échantillon dont nous pouvons voir quelques extraits dans les tables 24 et 25. Nous vérifions que les coefficients associés aux variables aient bien une p-value inférieure à 0,05. La *p-value* représente la probabilité, sous l'hypothèse nulle, d'obtenir un résultat au moins aussi extrême que celui observé dans l'échantillon. Ainsi, dans le cas où cette valeur est inférieure au seuil usuellement utilisé de 0,05, la probabilité d'observer un effet au moins aussi extrême que celui constaté dans notre échantillon est peu probable et nous rejetons cette hypothèse nulle. Nous faisons toutefois attention à vérifier la cohérence des coefficients retenus dans le cadre de notre modèle. L'intervalle de confiance au seuil de 95% est également indiqué à titre d'information. Toute variable dont l'intervalle de confiance de son coefficient inclue la valeur zéro peut être considérée comme potentiellement non

significative pour le modèle.

Nous observons quelques résultats intéressants. En effet, les variables les plus significatives sont les classes de sinistre. Plus la classe représente une sinistralité élevée, plus le coefficient est important. Il existe donc une relation croissante entre taux de destruction et classe de sinistre. Deux résultats peuvent cependant paraître étonnants comme le coefficient négatif devant *map_group* qui représente le niveau d'aléa par département avec une p value proche de 0,05. Cela peut être révélateur d'une carte d'aléa historique qui ne représente plus assez bien l'exposition au risque d'aujourd'hui. Cela peut par exemple venir du fait que des départements issus des zones où l'aléa ont établi des stratégies visant à renforcer les toitures par exemple et ainsi limiter les dommages liés à la grêle. Nous décidons tout de même de poursuivre l'analyse en gardant en tête qu'une mise à jour de cette carte d'aléa pourra être utilise avec la prise en compte de nouveaux facteurs. Ensuite, le coefficient devant l'agrégation climatique paraît faible. Cela peut venir du fait que les informations climatiques ont majoritairement été préalablement pris en compte au sein des classes de sinistres et que dans le cadre de la modélisation du taux de destruction, cela va davantage venir affiner la prédiction.

Variables	Coefficients	Std_Error	Borne_inf_IC_0.95	Borne_sup_IC_0.95	P_value
(Intercept)	-0,046112451	1,26E-02	-0,070879532	-0,021345371	3,45E-04
pct_housing	0,094403926	1,86E-02	0,057852552	0,130955301	1,02E-06
map_group	-0,005190288	2,61E-03	-0,010303801	-7,67748E-05	4,82E-02
INC1	0,614413635	8,94E-03	0,596899999	0,631927272	1,22E-130
INC2	0,715965389	9,24E-03	0,697846463	0,734084315	1,37E-139
INC3	0,918449257	1,54E-02	0,888183075	0,948715438	7,75E-120
INC4	1,149352324	3,73E-02	1,076167211	1,222537438	2,06E-73
INC5	1,341329738	4,00E-02	1,262922272	1,419737204	4,39E-79
nb_2_4_NWMT_5	0,000287431	7,29E-05	0,000144553	0,000430309	1,15E-04

Table 24: Résultats GLM sur échantillon 2

Variables	Coefficients	Std_Error	Borne_inf_IC_0.95	Borne_sup_IC_0.95	P_value
(Intercept)	-0,050389594	0,012974499	-0,075819145	-0,024960043	1,45E-04
pct_housing	0,104601615	0,019113733	0,067139387	0,142063843	1,48E-07
map_group	-0,005699879	0,002664418	-0,010922043	-0,000477715	3,38E-02
INC1	0,615379786	0,00920052	0,597347098	0,633412475	1,43E-128
INC2	0,718244532	0,009486375	0,699651579	0,736837485	7,00E-138
INC3	0,922342004	0,015119696	0,892707945	0,951976063	1,03E-121
INC4	1,148513479	0,03725325	1,075498451	1,221528507	1,62E-73
INC5	1,343695874	0,039381731	1,266509101	1,420882648	2,94E-80
nb_2_4_NWMT_5	0,000281986	6,29312E-05	0,000158643	0,000405329	1,32E-05

Table 25: Résultats GLM sur échantillon 5

Ainsi, en prenant la moyenne des coefficients sur les 10 échantillons nous obtenons

pour nb_2_4_NWMT_5 :

$$TD_LOG = -0.05699 + 0.10688 \cdot \text{pct_housing} - 0.00386 \cdot \text{map_group} + 0.61413 \cdot \text{INC1} + 0.71294 \cdot \text{INC2} + 0.90235 \cdot \text{INC3} + 1.1668 \cdot \text{INC4} + 1.35461 \cdot \text{INC5} + 0.00032 \cdot \text{nb_2_4_NWMT_5} \quad (36)$$

Ensuite, nous avons voulu tester ces coefficients sur les différents échantillons qui étaient à notre disposition à titre d'information seulement. En effet, étant donné que nous allons appliquer des coefficients obtenus sur les mêmes échantillons qui nous ont permis d'obtenir ces coefficients, nous ne pouvons pas être sûrs d'une absence de surentraînement mais dans un premier temps, cela va nous permettre de voir si les résultats restent cohérents avec le tableau initial 23.

Variable	Echantillon	R^2_train	R^2_incurred_train	ecart_train	R^2_test	R^2_incurred_test	ecart_test
nb_2_4_NWMT_5.2	1	0,99	0,97	11%	0,94	0,49	-53%
nb_2_4_NWMT_5.2	2	0,99	0,98	-11%	0,94	0,78	-17%
nb_2_4_NWMT_5.2	3	0,99	0,98	-12%	0,93	0,78	-16%
nb_2_4_NWMT_5.2	4	0,99	0,91	5%	0,93	0,92	-27%
nb_2_4_NWMT_5.2	5	0,99	0,98	-11	0,95	0,78	-17%
nb_2_4_NWMT_5.2	6	0,99	0,98	-12%	0,93	0,78	-16%
nb_2_4_NWMT_5.2	7	0,99	0,92	4%	0,93	0,92	-28%
nb_2_4_NWMT_5.2	8	0,99	0,97	11%	0,93	0,49	-53%
nb_2_4_NWMT_5.2	9	0,99	0,98	-13%	0,93	0,78	-15%
nb_2_4_NWMT_5.2	10	0,99	0,97	6%	0,93	0,53	-44%

Table 26: Résultats du modèle retenu sur les échantillons

Nous notons une amélioration globale des résultats en termes de R^2. Concernant les échantillons pour lesquels nous avions un R^2 négatif pour les sinistres modéliser, désormais, ils se situent aux alentours de 0,51. Les écarts en termes de pourcentage entre les sinistres observés et prédits restent satisfaisants en étant inférieurs à 30% la plupart du temps.

Nous répétons la démarche sur les 4 autres variables et nous obtenons les valeurs suivantes que nous allons pouvoir utiliser par la suite.

Pour nb_2_4_max, nous obtenons :

$$TD_LOG = -0.05704 + 0.10688 \cdot \text{pct_housing} - 0.00387 \cdot \text{map_group} + 0.61411 \cdot \text{INC1} + 0.7116 \cdot \text{INC2} + 0.89774 \cdot \text{INC3} + 1.16779 \cdot \text{INC4} + 1.34478 \cdot \text{INC5} + 0.00019 \cdot \text{nb_2_4_max} \quad (37)$$

Pour WS_0_max, nous obtenons :

$$TD_LOG = -0.05511 + 0.10483 \cdot \text{pct_housing} - 0.00417 \cdot \text{map_group} + 0.61364 \cdot \text{INC1} + 0.71264 \cdot \text{INC2} + 0.88861 \cdot \text{INC3} + 1.16752 \cdot \text{INC4} + 1,34946 \cdot \text{INC5} + 2 \times 10^{-5} \cdot \text{WS_0_max} \quad (38)$$

Pour WNES_0_4_max, nous obtenons :

$$TD_LOG = -0.05568 + 0.10297 \cdot \text{pct_housing} - 0.00404 \cdot \text{map_group} + 0.6138 \cdot \text{INC1} + 0.71181 \cdot \text{INC2} + 0.89003 \cdot \text{INC3} + 1.16638 \cdot \text{INC4} + 1,35212 \cdot \text{INC5} + 2 \times 10^{-5} \cdot \text{WNES_0_4_max} \quad (39)$$

Pour NWS_2_max, nous obtenons :

$$TD_LOG = -0.05522 + 0.10666 \cdot \text{pct_housing} - 0.00426 \cdot \text{map_group} + 0.61346 \cdot \text{INC1} + 0.71209 \cdot \text{INC2} + 0.88673 \cdot \text{INC3} + 1.16359 \cdot \text{INC4} + 1,33849 \cdot \text{INC5} + 6 \times 10^{-5} \cdot \text{NWS_2_max} \quad (40)$$

5.4.2 Modélisation appliquée aux évènements de classe 0

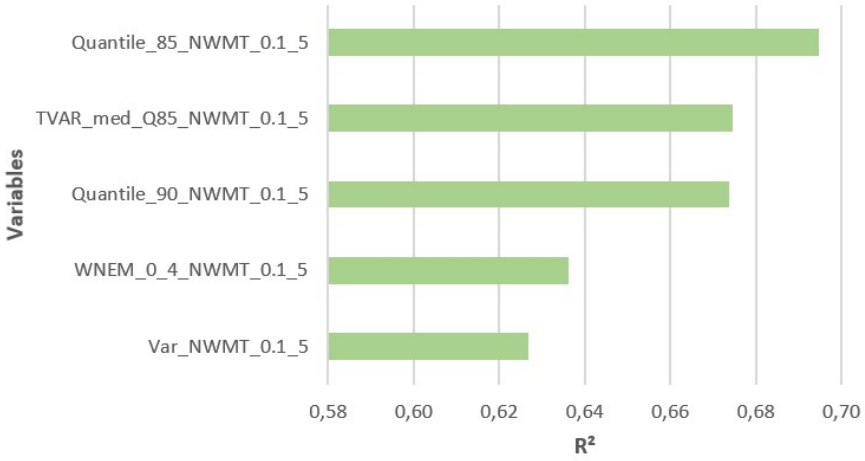
Pour les évènements de classe 0, nous réitérons exactement le même procédé. Dans cette partie nous ne présenterons ainsi que les résultats majeurs ainsi que les variables qui seront retenues.
Ainsi nous affichons dans la table 27 les agrégations climatiques suite à l'étude du R^2 moyens au niveau de la base d'entraînement et au niveau de la base de test, aussi bien en termes de taux de destruction transformé qu'en termes de montant de sinistres prédits en veillant à sélectionner les variables qui possèdent un écart type faible.

Variables	R²train.td.log	R²train.td.log.sd	R²train.I	R²train.I.sd	écart.train	R²test.td.log	R²test.td.log.sd	R²test.I	R²test.I.sd	ecart.test
Quantile_90_NWMT_5	0,98	0,00	0,81	0,01	-8%	0,98	0,00	0,67	0,21	6%
TVAR_med_Q85_NWMT_5	0,98	0,00	0,78	0,00	-9%	0,98	0,00	0,67	0,21	7%
Quantile_85_NWMT_5	0,98	0,00	0,83	0,01	-7%	0,98	0,00	0,69	0,22	8%
Var_NWMT_5	0,98	0,00	0,83	0,01	-7%	0,98	0,01	0,63	0,24	7%
WNEM_0_4_NWMT_5	0,98	0,00	0,84	0,01	-7%	0,98	0,00	0,64	0,24	7%

Table 27: Résultats de la modélisation du taux de destruction pour les évènements de classe 0

Les variables retenues ici ont toutes la même agrégation au niveau temporel, à savoir une moyenne des 5 plus grandes valeurs de l'agrégation au niveau du département. Tout d'abord, nous avons $WNEM_0_4_NWMT_5$ qui représente la moyenne pondérée de

l'exponentiel des diamètres de grêlons divisés par 4 mettant, comme dans le cas des évènements de classe 1, davantage en exergue les grêlons d'un grand diamètre occasionnant beaucoup de dégâts.



Graphique 50: Sélection de variables par GLM - évènement de classe 0

Ensuite nous avons la variance au niveau du département. Cette variance mesure la dispersion des diamètres de grêle autour de la moyenne, ce qui permet de quantifier l'hétérogénéité des chutes de grêle dans le département. Ainsi, une variance élevée peut indiquer une grande variabilité dans les tailles de grêlons, suggérant des orages de grêle plus intenses ou un phénomène très localisé avec potentiellement des grêlons de très gros diamètre sur un périmètre restreint et au contraire, une faible variance signifie des tailles de grêlons plus uniformes, ce qui peut indiquer des orages moins violents ou plus étendus sur le département.

Nous avons ensuite les quantiles à l'ordre 85% et 90%. Plus la valeur de ces quantiles est élevée au niveau du département, plus cela indique un évènement de forte intensité. Les quantiles à 85% ou 90% permettent de capturer les tailles de grêlons les plus importantes, qui sont souvent responsables des dommages les plus significatifs. De plus, contrairement à la variance, les quantiles sont moins sensibles aux valeurs extrêmes isolées, et cela offre donc une mesure plus stable des évènements de grêle intenses. De plus, un évènement avec un quantile 85% de 1,5 cm de diamètre indique que seulement 15% des observations sur le territoire dépassent cette valeur, donc l'évènement est peu intense alors que si ce diamètre est de 4 cm, cela indique un évènement intense.

Pour nous concentrer sur ce qu'il se passe dans les quantiles les plus élevés nous avons la TVAR qui va se concentrer sur ces 15% restant dans la distribution. Cependant,

contrairement à la TVAR classique, nous prenons la médiane et non la moyenne des valeurs au-delà du quantile 85% estimant que la médiane est moins sensible aux valeurs extrêmes.

5.5 Modélisation croisée Classe de sinistre et Taux de destruction

Nous avons précédemment établi un premier modèle visant à modéliser la classe de sinistre auquel va appartenir le département à partir de variables historiques et de variables climatiques, et ensuite indépendamment, nous avons établi un second modèle capable à partir de l'information de la classe de sinistralité, de modéliser un taux de destruction transformé, qui une fois une dé-transformation appliqué et multipliée par l'exposition réelle observée dans le portefeuille nous donne la sinistralité prédite.

L'objectif désormais est de combiner ces deux approches, à savoir utiliser dans le second modèle les prédictions réalisées par le premier et d'estimer la pertinence de cette méthode en étudiant l'écart entre les valeurs prédites et les valeurs observées. Nous allons encore une fois étudier séparément les évènements de classe 1 et 0. Nous avons pour chaque type de classe d'évènement une batterie de 10 modèles *XG Boost* pour modéliser la classe de sinistre et 5 modèles obtenus par GLM. Ce qui nous fait un jeu de 50 combinaisons possible. Nous allons maintenant étudier quelles sont les combinaisons qui s'écartent le moins possible des données observées en minimisant les écarts.

Nous allons encore une fois créer 10 échantillons sur lesquels entraîner nos données afin de voir en moyenne sur les différents cas, comment se comportent chacune des combinaisons en privilégiant celles qui ont un faible écart et une faible variance de résultats.

Modélisation appliquée aux évènements de classe 1

Nous trions les données par ordre décroissants du R^2 du montant de sinistre modélisé dans la table 28 et observons également le pourcentage d'écart entre les valeurs observées et les valeurs modélisées. Un pourcentage **d'écart positif** signifiera ainsi que nous **sous-estimons** la charge, et inversement. Nous observons que deux variables *XG Boost* sont particulièrement bonnes sur les différents échantillons, à savoir $Duree_max_grele_POP$ et $nb_2_4_NWMT_5$. Nous observons également que la variable $nb_2_4_NWMT_5$ a tendance à sous-estimer globalement les sinistres tandis que $Duree_max_grele_POP$ a au contraire tendance à surestimer. Il est intéressant de noter que du côté du GLM, les valeurs restent globalement dans les mêmes ordres de grandeur. Pour la suite de la modélisation, nous décidons de nous concentrer sur ces deux variables de la modélisation

XG Boost qui semblent les plus prometteuses.

Variable.XGB	Variable.GLM	R^2	R^2_sd	écart	écart_sd
nb_2_4_NWMT_5	nb_2_4_max	0,76	0,17	4%	0,12
nb_2_4_NWMT_5	nb_2_4_NWMT_5	0,75	0,15	4%	0,14
nb_2_4_NWMT_5	NWS_2_max	0,74	0,17	-1%	0,10
nb_2_4_NWMT_5	WNES_0_4_max	0,74	0,16	-1%	0,12
nb_2_4_NWMT_5	WS_0_max	0,73	0,17	-2%	0,11
Duree_max_grele_POP	nb_2_4_max	0,68	0,21	-14%	0,14
Duree_max_grele_POP	nb_2_4_NWMT_5	0,66	0,20	-16%	0,15
Duree_max_grele_POP	NWS_2_max	0,63	0,24	-21%	0,11
Duree_max_grele_POP	WNES_0_4_max	0,60	0,25	-23%	0,13
Duree_max_grele_POP	WS_0_max	0,58	0,26	-24%	0,13
NWS_2_NWMT_5	nb_2_4_NWMT_5	0,21	0,51	-13%	0,16
NWS_2_NWMT_5	nb_2_4_max	0,20	0,54	-12%	0,15
nb_4_max_POP	nb_2_4_NWMT_5	0,20	0,51	-16%	0,16
nb_4_max_POP	nb_2_4_max	0,19	0,55	-15%	0,15
TVAR_med_Q80_max	nb_2_4_max	0,17	0,55	-27%	0,15
TVAR_med_Q90_max	nb_2_4_max	0,17	0,55	-31%	0,16
TVAR_med_Q80_max	nb_2_4_NWMT_5	0,17	0,52	-29%	0,16
TVAR_med_Q90_max	nb_2_4_NWMT_5	0,17	0,52	-32%	0,17
TVAR_med_Q85_max	nb_2_4_max	0,13	0,57	-55%	0,16
TVAR_med_Q85_max	nb_2_4_NWMT_5	0,11	0,54	-56%	0,15

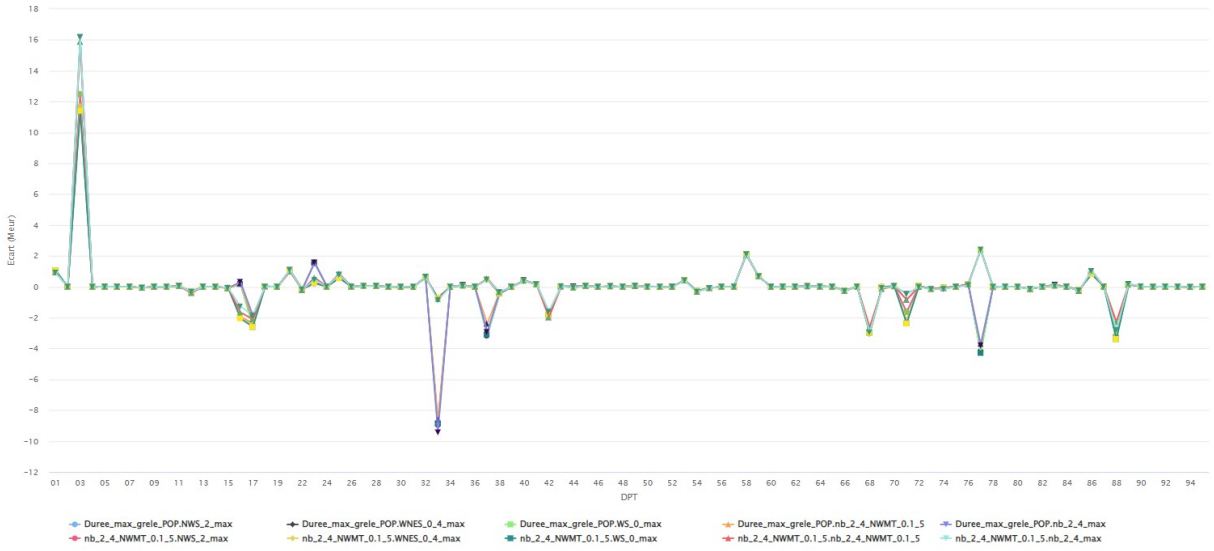
Table 28: Extrait des résultats de la modélisation croisée XGB et GLM
Évènements de classe 1

Nous procédons alors à la validation de la combinaison de ces modèles sur la base test isolée initialement dans la table 29.

Variable.XGB	Variable.GLM	R^2	écart
Duree_max_grele_POP	nb_2_4_max	0,42	-0,02
Duree_max_grele_POP	NWS_2_max	0,57	-0,13
Duree_max_grele_POP	WS_0_max	0,60	-0,17
Duree_max_grele_POP	nb_2_4_NWMT_5	0,46	-0,01
Duree_max_grele_POP	WNES_0_4_max	0,59	-0,16
nb_2_4_NWMT_5	nb_2_4_max	0,56	0,25
nb_2_4_NWMT_5	NWS_2_max	0,70	0,13
nb_2_4_NWMT_5	WS_0_max	0,73	0,09
nb_2_4_NWMT_5	nb_2_4_NWMT_5	0,57	0,24
nb_2_4_NWMT_5	WNES_0_4_max	0,72	0,11

Table 29: Résultat de la validation du modèle croisé

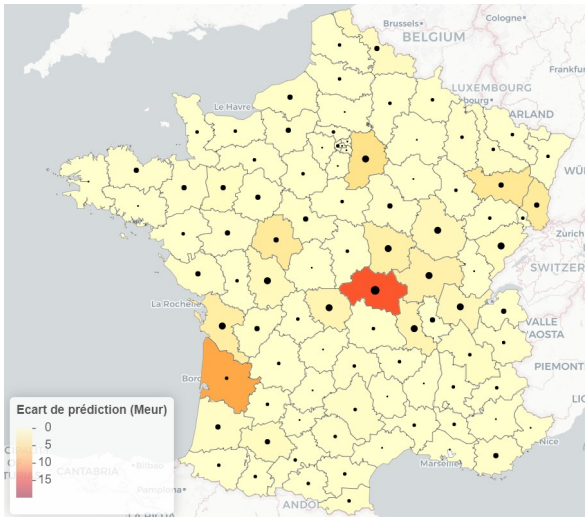
Nous observons que la majorité des combinaisons de modèles sélectionnés ont un écart inférieur à 17% du montant de sinistre observé en dehors de deux combinaisons où nous atteignons environ 25% d'écart. En faisant la moyenne des différentes combinaisons de ce tableau, nous obtenons un écart de 3% seulement avec les valeurs observées ce qui est très satisfaisant au niveau global de l'évènement sachant qu'il y a dans cette base de test un évènement qui coûte plusieurs dizaines de millions d'euros (sinistre de classe 5). Si nous sommes satisfaits de la prédiction au niveau de l'ensemble de l'échantillon, ce qui nous intéresse réellement est la pertinence de la modélisation à l'échelle des départements pour éviter d'attribuer des montants incohérents à ces derniers. En effet, nous ne souhaitons pas d'un faible écart venant d'une compensation non contrôlée des montants de sinistres prédits entre les départements.



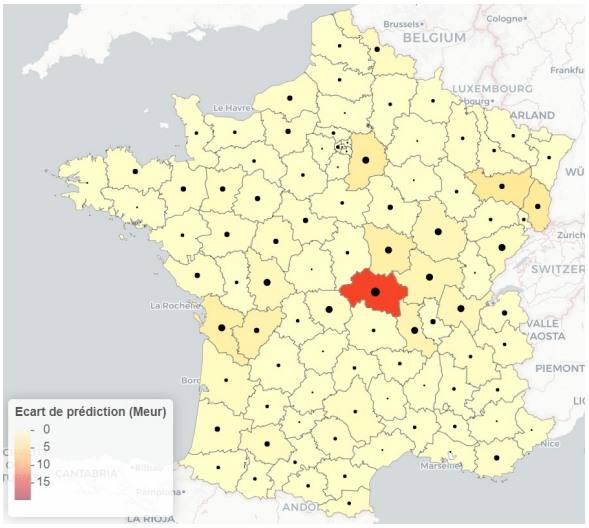
Graphique 51: Écart de prédiction par département selon les différentes combinaisons

Sur le graphique 51, nous affichons les écarts du montants prédits par département pour les combinaisons de variables ayant obtenu les meilleurs scores. Il est à noter que les données ont été modifiées et ne correspond pas aux réelles pertes observées dans le portefeuille mais restent cohérents dans l'ordre de grandeur. Nous observons que la plupart des écarts restent inférieurs à 3 millions d'euros ce qui est concluant. Nous notons tout de même des écarts importants par exemple dans le cas de l'Allier (département 3) et dans la Gironde (département 33). Dans les autres cas, les estimations sont cohérentes et par exemple dans le cas de la Seine-et-Marne (département 77), les écarts issus des modèles avec $nb_2_4_NWMT_5$ en variable pour XGB compensent les écarts avec ceux obtenus par $Duree_max_grele_POP$. Ainsi, cela montre qu'il est pertinent de garder ces différents indicateurs.

Pour mieux comprendre les écarts, nous décidons d'afficher deux combinaisons à titre d'exemple sur une carte : $Duree_max_grele_POP$ en XGB et WS_0_max en GLM dans la carte 52 ainsi que $nb_2_4_NWMT_5$ en XGB et NWS_2_max en GLM sur la carte 53.



Graphique 52: Carte des écarts de prédictions pour la combinaison Duree_max_grele_POP en XG Boost et WS_0_max en GLM

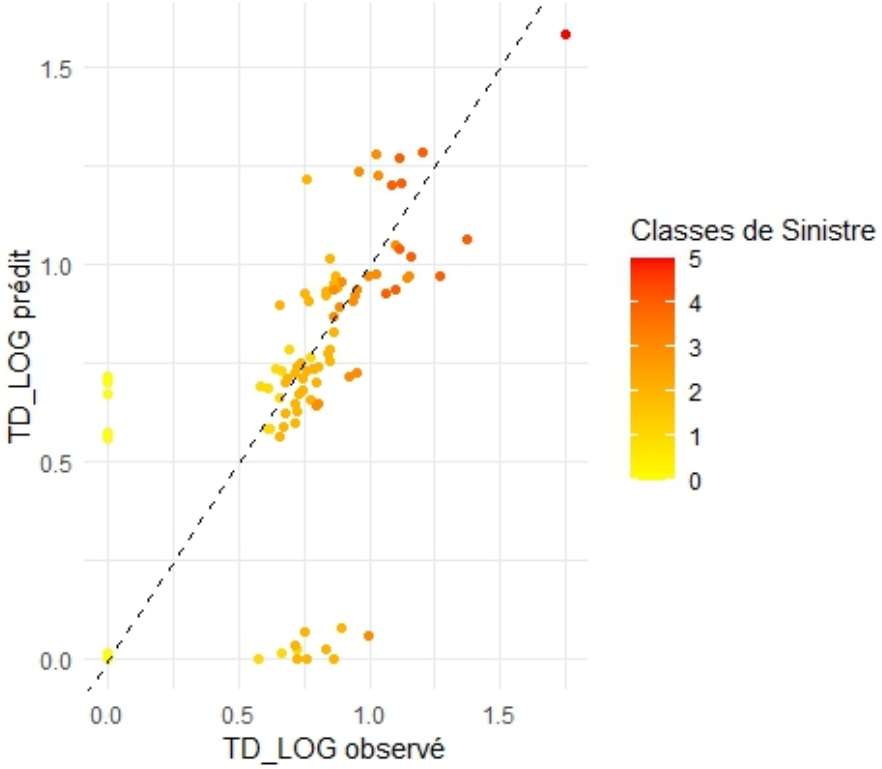


Graphique 53: Carte des écarts de prédictions pour la combinaison nb_2_4_NWMT_5 en XG Boost et NWS_2_max en GLM

Les écarts affichés dans un dégradé de couleurs du jaune (peu d'écart) au rouge (écarts importants) sont les écarts en valeur absolue. Afin de mettre en perspective ces écarts par rapport aux montants réels des évènements, nous avons indiqué par le rayon des points noirs la classe de sinistre auquel appartient le département. Les plus petits rayons indiquant la classe de sinistre 0, à savoir l'absence de sinistre et les plus importants indiquent la classe de sinistre 5 à plusieurs dizaines de millions d'euros. Nous notons que la plupart des prédictions ont un écart inférieur à 2 millions d'euros ce qui est un résultat très satisfaisant. Seule la Gironde présente un écart d'environ 10 Millions d'euros par rapport à un sinistre qui n'a coûté que quelques milliers d'euros.

Les écarts les plus importants concernent pour la plupart des sinistres de classe 4 ou 5 ce qui est compréhensible comme il s'agit des sinistres les plus élevés avec une forte volatilité. De plus, nous observons un écart de près de 15 millions d'euros dans le département 3, à savoir l'Allier. Cela peut paraître excessif à prime abord mais lorsque nous gardons en tête qu'il s'agit d'un sinistre à plusieurs dizaines de millions d'euros, cela semble raisonnable.

Nous allons désormais essayer d'expliquer l'origine de ces écarts à partir des résultats de la modélisation. Nous avons une sous-estimation dans le cas de l'Allier et une surestimation pour la Gironde. Nous pouvons commencer par l'analyse de l'Allier. Nous affichons dans le graphique 54, la comparaison entre les valeurs observées (en abscisse) et les valeurs prédites (en ordonnée) du taux de destruction transformé (TD_LOG) avec la bissectrice. La proximité des points à cette bissectrice indique la précision de notre modèle : plus les points sont proches de cette ligne, plus les valeurs prédites sont similaires aux valeurs observées. Pour avoir l'importance de chacun des points affichés, nous représentons également sur une échelle de couleur les classes de sinistres. Ainsi, une erreur de modélisation pour des classes de sinistre 0 ou 1 (jaune) aura une moindre importance comparée à une erreur de modélisation d'une classe de sinistre 5 (rouge).



Graphique 54: Comparaison des TD_LOG prédits et observés (évènements de classe 1)
Dans ce graphique, chaque point correspond au taux de destruction transformé lié à la sinistralité par département observé (en abscisse) et prédit (en ordonnée) pour l'évènement test généré par l'algorithme 2.

Nous observons que la majorité des valeurs se trouve autour de la bissectrice indiquée en trait pointillé. Certains points se trouvent néanmoins sur les axes, indiquant les cas où de la sinistralité a eu lieu sur le département ($TD_LOG > 0$) mais cela n'a pas été capturé par le modèle ($TD_LOG_prédit = 0$) et inversement. Cependant, ces points

sont pour la grande majorité des classes de sinistre inférieures à 2, ce qui limite l'impact des écarts observés. Le graphique révèle une observation importante concernant le département de l'Allier, qui présente le sinistre le plus important, de classe 5. Ce point est situé dans la partie supérieure droite du graphique et se trouve relativement proche de la bissectrice. Cela signifie que l'écart observé provient du léger décalage observé entre la valeur prédite et la valeur observée du taux de destruction transformé, écart décuplé par la suite par l'application de l'exposition du portefeuille AXA sur ce département afin d'obtenir le montant du sinistre. Afin de réduire cet écart, il sera intéressant d'ajuster la variable d'exposition pour qu'elle reflète au mieux l'exposition au risque de grêle du portefeuille résidentiel d'AXA.

Concernant le cas de la Gironde, l'écart provient d'une erreur de prédiction de la bonne classe de sinistralité, en effet, la modélisation par le *XG Boost* avec la variable agrégée *duree_max_grele_POP* indique un sinistre de classe 4 au lieu d'un sinistre de classe 2. Le graphique de l'importance des variables 41 nous montre la pertinence du niveau d'aléa du département (*map_group*), qui est associée au niveau 4 (sur un niveau maximal de 5), ce qui un niveau fort d'aléa en plus d'un pourcentage de maison relativement élevé. Pour donner un ordre de grandeur de ces valeurs, nous affichons les valeurs normalisées par le maximum et le minimum de ces variables X au sein de notre base de données dans la table 30. Cela nous indique que la grande valeur de la classe de sinistre attribué provient notamment par les valeurs de *map_group* ($\frac{4}{5} = 0,75$) ainsi qu'un pourcentage de maison au sein du portefeuille en Gironde relativement élevé, 59% du pourcentage maximum observé dans la base de données.

$$x_{\text{normalisé}} = \frac{x - \min(X)}{\max(X) - \min(X)} \tag{41}$$

	map_group	**Duree_max_grele_POP**	**pct_HOUSING**	NOMBRE_COMMUNE
Valeurs normalisées	0,75	0,29	0,59	0,25

Table 30: Détail des valeurs pour les variables du modèle XGB avec Duree_max_grele_POP

Avec ces différentes considérations, nous comprenons l'origine de ces quelques écarts. Nous notons que ces écarts sont relativement localisés sur quelques départements et qu'il existe des pistes d'amélioration, notamment une meilleure variable d'exposition au risque de grêle et l'intégration d'un plus grand volume de données qui pourraient affiner le modèle. Néanmoins, la modélisation reste globalement satisfaisante. Il fournit une bonne estimation des classes de sinistralité, particulièrement pour les classes les plus importantes, ce

qui nous permettra dans la finalité de ce modèle de prévenir les équipes adéquates d'une grosse sinistralité dans ce département par exemple tout en faisant preuve de prudence.

Nous observons les mêmes caractéristiques pour la seconde combinaison de variables avec une prédiction correcte cette fois ci du montant de sinistre en Gironde et un écart un peu plus important dans l'Allier causé par les arguments que nous avons étudiés précédemment.

Modélisation appliquée aux évènements de classe 0

Nous appliquons la même méthodologie pour les sinistres issus des évènements de classe 0 et nous obtenons le tableau des résultats des combinaisons des modèles de classe de sinistre ainsi que du taux de destruction avec les données triées par ordre décroissant du R^2 du montant de sinistre estimé.

Variable.XGB	Variable.GLM	R^2	R^2_sd	écart	écart_sd
Quantile_95_sum_POP	Quantile_85_NWMT_5	0,70	0,18	4%	0,12
TVAR_mean_Q90_sum_POP	Quantile_85_NWMT_5	0,70	0,19	1%	0,12
nb_4_sum	Quantile_85_NWMT_5	0,68	0,18	-10%	0,16
Quantile_95_sum_POP	TVAR_med_Q85_NWMT_5	0,68	0,19	3%	0,12
TVAR_mean_Q90_sum_POP	TVAR_med_Q85_NWMT_5	0,68	0,19	0%	0,13
Quantile_95_sum_POP	Quantile_90_NWMT_5	0,68	0,17	1%	0,12
TVAR_mean_Q90_sum_POP	Quantile_90_NWMT_5	0,67	0,18	-2%	0,13
nb_4_sum	TVAR_med_Q85_NWMT_5	0,67	0,19	-10%	0,16
nb_4_sum	Quantile_90_NWMT_5	0,65	0,17	-14%	0,16
Quantile_95_sum_POP	WNEM_0_4_NWMT_5	0,65	0,20	2%	0,13
TVAR_mean_Q90_sum_POP	WNEM_0_4_NWMT_5	0,65	0,20	-1%	0,13
Quantile_95_sum_POP	Var_NWMT_5	0,63	0,20	2%	0,12
TVAR_mean_Q90_sum_POP	Var_NWMT_5	0,63	0,20	-1%	0,13
nb_4_sum	WNEM_0_4_NWMT_5	0,62	0,20	-13%	0,17
nb_4_sum	Var_NWMT_5	0,61	0,20	-14%	0,16

Table 31: Extrait des résultats de la modélisation croisée XGB x GLM
Évènements de classe 0

En affichant les premières lignes, nous notons qu'il n'y a pas de combinaisons de modèles qui se démarquent particulièrement pour la classe de sinistres. Nous décidons de nous focaliser sur $Quantile_95_sum_POP$ et nb_4_sum pour la suite. Concernant les modèles de taux de destruction, $Quantile_85_NWMT_5$, $TVAR_med_Q85_NWMT_5$ et $Quantile_90_NWMT_5$ reviennent régulièrement. Les écarts en moyenne sont très satisfaisants de l'ordre de 3 à 13% en moyenne avec une tendance à surestimer les données (écart en pourcentage négatifs). L'écart type reste également raisonnable.

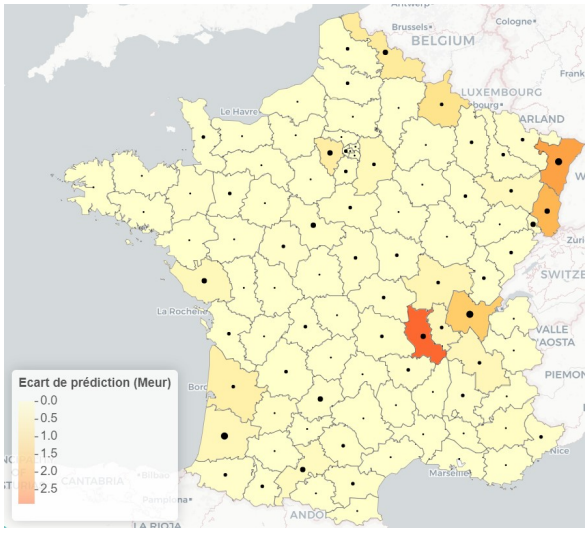
Nous procédons alors à la validation de la combinaison de ces modèles sur la base test isolée initialement dans la table 32.

Variable.XGB	**Variable.GLM**	**R²**	**écart**
Quantile_95_sum_POP	Quantile_90_NWMT_5	-1,80	-1,06%
Quantile_95_sum_POP	Var_NWMT_5	-1,89	-109%
Quantile_95_sum_POP	TVAR_med_Q85_NWMT_5	-1,76	-110%
Quantile_95_sum_POP	Quantile_85_NWMT_5	-1,85	-113%
Quantile_95_sum_POP	WNEM_0_4_NWMT_5	-2,55	-122%
nb_4_sum	Quantile_90_NWMT_5	-2,17	-138%
nb_4_sum	TVAR_med_Q85_NWMT_5	-1,96	-144%
nb_4_sum	Quantile_85_NWMT_5	-2,05	-144%
nb_4_sum	Var_NWMT_5	-2,41	-147%
nb_4_sum	WNEM_0_4_NWMT_5	-2,94	-158%

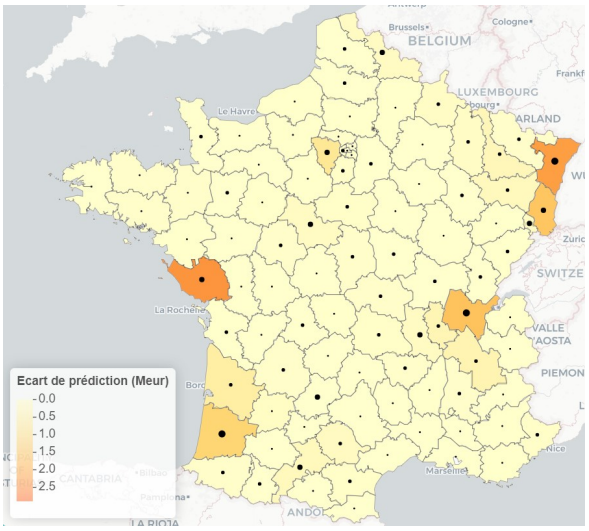
Table 32: Résultat de la validation du modèle croisé

Nous notons cette fois-ci des résultats au niveau du R^2 et de l'écart en pourcentage qui ne semblent pas satisfaisants. Nous avons en effet des R^2 fortement négatifs ainsi des montants estimés à plus du double de la valeur observée. Dans le cas des évènements de classe 0, nous n'avons pas d'évènements de classe 5, ce qui fait que l'ampleur des évènements de grêle est limitée. Ainsi, il suffit d'une mauvaise estimation d'un sinistre en classe 5 au lieu de 3 ou 4 par exemple pour entraîner une surestimation significative. Une autre hypothèse pour expliquer ces mauvais résultats serait le fait que les prédictions surestiment dans tous les cas les sinistres. Ainsi, même si les écarts sont limités à quelques millions d'euros entre les valeurs prédites et observées, au global l'accumulation des surestimations va entraîner de grands écarts et cela se verra notamment en termes de pourcentage car un sinistre à 10 000€ estimé à 100 000€ aura un écart estimé à ecart_pct $= \frac{10000-100000}{10000} = 900\%$. Pour comprendre ces écarts, nous allons examiner les écarts par département et selon les catégories de sinistres observées.

Les cartes 55 et 56 nous font davantage pencher vers l'hypothèse où les différents modèles surestiment le montant des sinistres sur les différents départements. Nous observons que dans la majorité des cas, nous avons des écarts inférieurs à 1 million d'euros. Cependant, les écarts de l'ordre de 3 millions d'euros vers l'est de la France dans le Haut Rhin et le bas Rhin dans les deux cas, et dans le département de la Loire dans la carte 55 et en Vendée dans la carte 56 creusent les écarts.

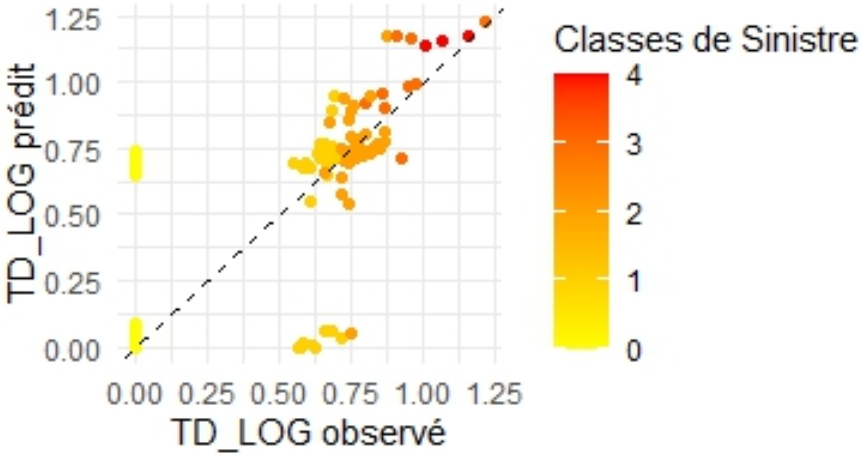


Graphique 55: Carte des écarts de prédictions pour la combinaison nb_4_sum en XG Boost et Quantile_90_NWMT_5 en GLM



Graphique 56: Carte des écarts de prédictions pour la combinaison Quantile_95_sum_POP en XG Boost et Quantile_85_NWMT_5 en GLM

Le graphique 57 nous confirme cette surestimation à travers une surestimation également des taux de destructions sur la plupart des départements, notamment pour les sinistres de classe 2 et 3. Étant donné qu'il s'agit d'une tendance globale, ces écarts proviennent d'une erreur de classification du premier modèle avec *XG Boost*.



Graphique 57: Comparaison des TD_LOG prédits et observés (évènements de classe 0)
Dans ce graphique, chaque point correspond au taux de destruction transformé lié à la sinistralité par département observé (en abscisse) et prédit (en ordonnée) pour l'évènement test généré par l'algorithme 2.

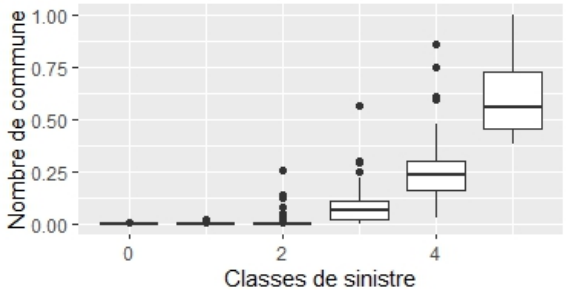
Intéressons-nous aux 3 écarts de prédictions majeurs pour comprendre leur origine, nous indiquons dans la table 33 les valeurs normalisées des variables utilisées dans le cas

de la modélisation *XG Boost* de la classe de sinistre.

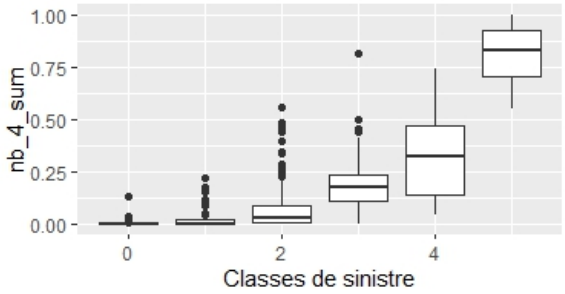
Département	Nombre_Communes_touchées	nb_4_sum
42	0,08	**0,41**
68	**0.30**	0,02
90	**0.29**	0,11

Table 33: Détail des valeurs pour quelques variables du modèle XGB avec Nb_4_sum

Nous remarquons pour les départements 68 et 90 que les prédictions ont notamment été influencées par les valeurs élevées du nombre de commune touchées, revenant à environ 30% du maximum de commune touchées par la grêle observée au sein de notre base de données. D'après le box-plot 58 (aussi appelé «boîte à moustache»), 0,30 est en effet associée d'après l'intervalle interquartile (entre le premier quartile et le troisième) à un sinistre de classe 4. Pour le département 42, nous remarquons qu'il s'agit plus particulièrement de la variable climatique agrégée par le nombre de grêle d'un diamètre supérieur à 4 cm de diamètre qui a poussé à une plus grande classe de sinistre. La valeur 0,41, d'après la distribution observée dans le box-plot 59 se situe dans l'intervalle interquartile associée à la classe de sinistre 4.



Graphique 58: Distribution des valeurs de Nombre de communes touchées en fonction de la classe de sinistre



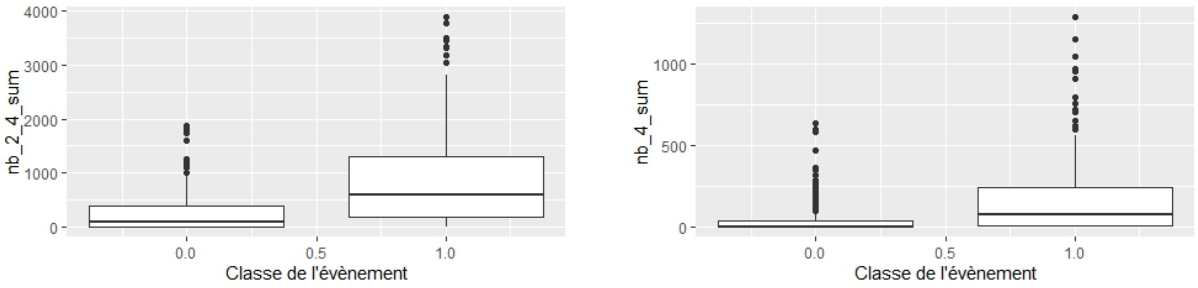
Graphique 59: Distribution des valeurs de nb_4_sum en fonction de la classe de sinistre

Ainsi, ces écarts proviennent notamment d'un nombre élevé de communes touchées par la grêle qui n'ont cependant pas eu des dommages aussi conséquents que ce qui a été observé par le passé. Cela peut être dû à la localisation des biens touchés au sein des différents département où le portefeuille n'était pas surexposé ou alors des mesures de prévention qui ont été mises en place par les communes ou les départements pour mieux faire face à la grêle. La modélisation montre quelques faiblesses dans le cas des évènements de classe 0, des améliorations sont à faire mais les résultats peuvent être utilisés avec prudence.

5.6 Modélisation des classes 1 et 0 pour l'évènement

Désormais, nous allons déterminer des indicateurs météorologiques qui nous permettraient de distinguer les évènements de classe 1 ou 0 pour savoir si ces derniers vont toucher ou non la rétention afin de savoir quel type de modèle appliquer afin d'éviter de sous-estimer ou surestimer de manière conséquente en nous trompant de modèle.

Pour effectuer cette classification, nous avons attribué des valeurs binaires (0 ou 1) à une nouvelle colonne INT_EVT (intensité de l'évènement) en fonction du dépassement de la rétention de réassurance. Nous attribuons 1 dans le cas où nous dépassons la rétention et 0 le cas échéant. Puis, nous avons passé en revue les différentes agrégations issues des données Meteomatics afin de voir si certaines d'entre elles étaient particulièrement sensibles à cette nouvelle variable et ainsi, évaluer leur sensibilité à l'ampleur globale des évènements.



Graphique 60: box-plot pour les variables nb_2_4_sum et nb_4_sum en fonction de la classe de l'évènement
Source : Meteomatics

Deux variables climatiques se sont avérées particulièrement pertinentes dans cette analyse comme nous pouvons le voir dans les box-plots 60 où nous distinguons des distributions distinctes de ces variables en fonction de la classe de l'évènement, avec une distribution concentrée en dessous de 200 pour $nb_2_4_sum$ et autour de 0 pour nb_4_sum :

- $nb_2_4_sum$: représente le nombre total de grêlons de diamètre compris entre 2 et 4 cm tombés sur l'ensemble du département durant l'évènement. Il s'agit d'une variable intéressante que ces types de grêlons, bien que moins dangereux que ceux de plus de 4 cm, peuvent entraîner des dégâts importants s'ils sont accumulés en grande quantité.

- nb_4_sum : correspond au nombre total de grêlons d'un diamètre supérieur à 4 cm tombés sur l'ensemble du département durant l'évènement. Cette variable est aussi intéressante car ces grêlons sont certes, moins fréquents que ceux de moins de 4 cm de diamètre mais peuvent être révélateurs d'un évènement de sévérité conséquente.

Nous avons ensuite calculés divers quantiles associés à ces deux variables selon la classe de l'évènement sur notre base de données que nous pouvons retrouver dans les tables 34 et 35 pour trouver un indicateur au-delà duquel nous pouvons affirmer qu'un évènement soit un évènement de type 1 ou 0. Pour cela, nous avons trouvé le quantile à l'ordre 75% pertinent pour distinguer ces cas-là et nous avons choisi le seuil de 1 000 grêlons entre 2 et 4 cm de diamètre pour la première variable et 200 grêlons de plus de 4 cm de diamètre sur le département sur tout l'évènement pour distinguer ces classes d'évènements. En effet, les évènements associés à savoir 220602-220606 et 220619-220623 sont les évènements ayant engendré la plus grande sinistralité et nous pensons donc que ces seuils sont cohérents. Nous pouvons ainsi déterminer à partir des données de Meteomatics l'ampleur de l'évènement de grêle et appliquer le modèle approprié en conséquence. Nous notons que cette classification peut être améliorée en prenant en compte d'autres variables d'aléa comme la vitesse du vent par exemple, qui a également une influence sur les dommages en cas de grêle.

Evènements.Grêle	Somme	Médiane	Quantile_60%	Quantile_75%	Quantile_90%	Quantile_95%	max
220519-220523	59 040	201	276	469	789	1 128	1 871
220602-220606	146 742	560	764	1 072	1 651	2 226	3 497
220619-220623	178 540	646	991	1 558	2 187	2 963	3 887
220626-220630	47 976	60	74	93	149	480	1 875

Table 34: Analyses pour la variable nb_2_4_sum

Evènements.Grêle	Somme	Médiane	Quantile_60%	Quantile_75%	Quantile_90%	Quantile_95%	max
220519-220523	8 788	9	17	55	172	230	360
220602-220606	34 930	84	145	258	426	747	1 156
220619-220623	36 300	63	107	236	486	714	1 046
220626-220630	10 248	20	31	52	97	297	874

Table 35: Analyses pour la variable nb_4_sum

Nous avons ainsi élaboré une méthode permettant de modéliser le risque de grêle. Dès la connaissance d'un évènement de grêle en France, que ce soit à partir de bulletins météorologiques ou des rapports des entités d'AXA, nous pouvons extraire en temps réel les données météorologiques fournies par Meteomatics et y intégrer les données de relevés de terrain issus des données ESWD ainsi que les données de la densité de population. Cette combinaison de sources nous permet d'obtenir une vision plus complète et détaillée de chaque événement de grêle. Ensuite, en appliquant les différents modèles présentés dans ce mémoire au portefeuille d'AXA, nous sommes capables d'obtenir une estimation de la charge de sinistre de cet évènement, aussi bien au niveau global qu'au niveau de chaque département potentiellement affecté par la grêle. Cependant, il est important de noter que des précautions restent nécessaires lors de l'interprétation des résultats.

6 Limites et perspectives d'évolution du modèle

6.1 Problématique de la donnée disponible

La principale limitation de notre étude réside dans la disponibilité limitée des données des sinistres de grêle. Dans le contexte de la réassurance, nous ne disposons que des informations transmises par les entités dans le cadre des traités de réassurance, et ce uniquement lorsque la sinistralité dépasse un certain seuil. Par conséquent, notre jeu de données se concentre principalement sur l'année 2022, année exceptionnelle intense pour la grêle, ce qui limite la profondeur temporelle de notre analyse. Malgré cette contrainte, nous avons identifié des caractéristiques qui nous semblent intéressantes et robustes. Nous avons également fait attention au biais et à la variance lors de notre étude afin d'éviter le sur-apprentissage du modèle.

Les résultats obtenus grâce au modèle présenté dans le cadre de mémoire prometteurs, montrant un potentiel intéressant pour l'analyse et la prédiction des risques liés à la grêle dans le contexte de la réassurance. Le modèle est amené à évoluer et à s'améliorer au fil du temps, notamment grâce à l'intégration de nouveaux évènements qui ne manqueront pas de survenir, notamment en raison de l'augmentation de la fréquence des épisodes de grêle liée au changement climatique.

Toujours dans une optique d'améliorer la qualité de notre modèle, nous envisageons également de travailler la granularité des données. L'analyse est réalisée à l'échelle du département mais nous pensons qu'il peut être intéressant de travailler à une maille plus fine comme le code postal ou le code commune. Cette démarche permettra d'obtenir des données plus précises, ce qui est particulièrement pertinent compte tenu du caractère local des phénomènes de grêle.

6.2 Prise en compte changement climatique

Une autre limite de notre modèle peut être l'absence de prise en compte des projections climatiques à long terme. Dans ce cadre, les rapports du Groupe d'experts intergouvernemental sur l'évolution du climat (GIEC ou IPCC en anglais) peuvent être intéressant à prendre en compte. Il s'agit groupe d'experts en charge de l'élaboration d'un rapport sur l'évolution du changement climatique dans le monde. Le GIEC a publié des scenarii appelés SSP (*Shared Socio-economic Pathways*) de projection climatique afin de mieux prendre en compte les effets potentiels du réchauffement climatique. Ces projections permettent en effet d'étudier l'évolution future des aléas climatiques qui peuvent signi-

ficativement impacter les dommages liés à la grêle dans le cadre de notre mémoire.

Parmi ces scénarios, le SSP2-4.5 est considéré comme un scénario où le rythme des émissions des gaz à effet de serre ne connaît pas de variations brutales, le "*business as usual*", tandis que le SSP5-8.5 représente un scénario plus pessimiste avec des émissions de gaz à effet de serre (GES) élevées, impliquant un échec des politiques liées à la réduction de ces émissions.

Le 6ème rapport du GIEC publié en mars 2023 atteste ainsi d'une hausse en fréquence et en coût moyen des évènements climatiques dont la grêle qu'il serait intéressant et important de prendre en compte. Autrement, cela pourrait impliquer une sous-estimation des risques liés aux impacts du changement climatique.

Dans les futures améliorations de notre modèle, nous envisageons ainsi d'incorporer ces scénarios climatiques pour avoir une vision plus complète et prospective des risques associés à la grêle dans un contexte de changement climatique.

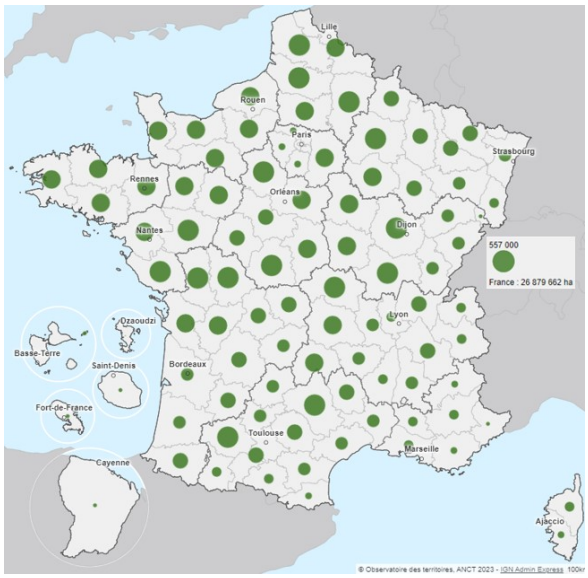
6.3 Prise en compte d'autres sources de données

Une autre piste d'amélioration de notre modèle consiste à intégrer d'autres sources de données. Nous pouvons par exemple penser à la composition des matériaux de toiture. En effet, la résistance à la grêle varie considérablement selon les matériaux utilisés, influençant ainsi la vulnérabilité des bâtiments. La Fondation de prévention des établissements cantonaux d'assurance établie en Suisse a publié en 2007 un rapport sur la protection contre la grêle et y réalise un classement de la résistance des différents matériaux de toiture face à la grêle [15]. Il sera dès lors intéressant d'associer ces données à la base de données nationale des bâtiments (BDNB) qui donne une cartographie du parc de bâtiments existants en France et notamment des données sur la toiture en France à l'échelle du département. Prendre en compte ces données permettrait d'affiner notre évaluation des risques et d'améliorer la précision de nos estimations de dommages.

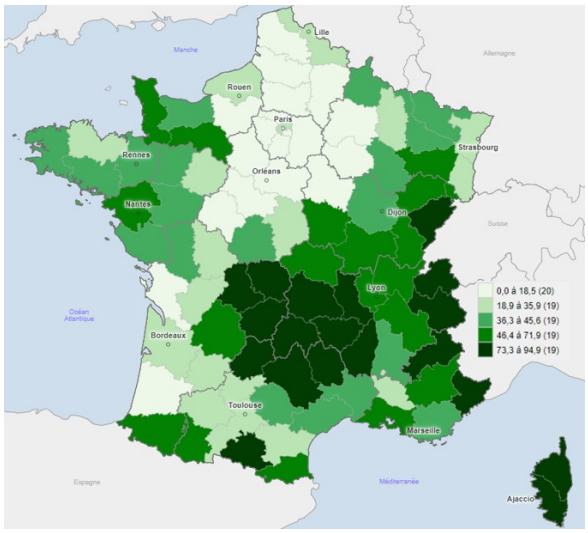
De plus, la prise en compte du vent peut être pertinente, notamment à travers sa vitesse et son orientation. L'orientation du vent peut en effet influencer l'angle d'impact des grêlons, tandis que sa vitesse peut amplifier les dégâts causés. Il est important de noter que les données Meteomatics modélisent le diamètre des grêlons sur une zone en fonction des conditions climatiques. Cependant, ces grêlons modélisés peuvent être déplacés de leur endroit de formation initiale sur une certaine distance selon la force et l'orientation du vent. Ainsi, une zone voisine à la zone de création des grêlons pourrait être affectée.

Par ailleurs, il serait pertinent d'étendre notre modélisation à d'autres LoB, telles que les lignes automobiles, les lignes commerciales et lignes agricoles. Chacune de ces lignes présente des caractéristiques et des vulnérabilités spécifiques face aux épisodes de grêle. Pour les lignes agricoles, par exemple, l'intégration de données sur les superficies agricoles et les types de cultures serait particulièrement utile. L'étude de Pùčik et al (2019) nous montrait en effet dans le graphique 17 l'impact des grêlons à partir d'un diamètre plutôt faible comparés aux maisons par exemple.

Dans ce cadre, l'AGRESTE [2], un organisme public français d'études et de statistiques sur l'agriculture, la forêt, les industries agroalimentaires, l'occupation du territoire, les équipements et l'environnement en zone rurale et l'ANCT [3], l'Agence nationale de la cohésion des territoires mettent à disposition de nombreuses analyses, études et données dont les cartes ci-dessous. Elles indiquent la surface agricole utile (SAU) qui un indicateur statistique destiné à évaluer le territoire consacré à la production agricole carte de gauche) et le pourcentage de prairie par rapport à la SAU (carte de droite). Nous pensons que ces données peuvent en effet être utiles dans le cadre de la modélisation des sinistres de grêles sur des zones agricoles.



Graphique 61: Carte des SAU en France en 2020 (hectares)
Source : ANCT (2020)



Graphique 62: Part des prairies dans la SAU en 2020 (%)
Source : AGRESTE (2020)

Enfin, pour les lignes commerciales et automobiles, l'intégration de données, les types de bâtiments commerciaux et la distribution du portefeuille automobiles permettrait d'affiner notre modèle.

Conclusions

Ce mémoire a ainsi pour objectif principal de concevoir un modèle permettant d'estimer les sinistres engendrés par la grêle sur le portefeuille d'AXA. Nous avons concentré nos travaux sur le portefeuille résidentiel à l'échelle départementale. Pour atteindre cet objectif, nous avons intégré diverses sources de données météorologiques qui ont été agrégées de manière à caractériser au mieux les épisodes de grêle, tant en termes de fréquence que d'intensité.

Cette base de variables permettant de représenter l'aléa climatique a été enrichie par des données démographiques et des données d'exposition du portefeuille d'AXA afin d'intégrer la notion de vulnérabilité. En croisant ces informations à travers une modélisation adaptée, nous avons pu estimer le risque associé à la grêle.

Notre approche se distingue par son caractère innovant, en combinant deux modèles avec un algorithme d'échantillonnage pour simuler des évènements, une méthode qui a été pensée pour pallier le manque de profondeur de l'historique de sinistralité et s'avère particulièrement pertinente pour les besoins de la réassurance. En effet, seules les montants issus évènements de grêle dépassant un certain seuil de déclaration étant remontés par les entités dans le cadre de traité de réassurance, cela limite fortement les données disponibles.

La modélisation s'est déroulée en deux étapes : tout d'abord, nous avons décidé de modéliser la classe de d'intensité sinistre, ce qui nous a permis d'estimer l'ampleur des dommages de la grêle par département. Ensuite, un second modèle s'appuyant sur les données météorologiques et les résultats du premier modèle nous permet d'estimer un taux de destruction, permettant ainsi d'obtenir l'estimation de la charge associée à chaque département. Cette approche a été appliquée à la fois aux sinistres susceptibles d'atteindre la rétention de réassurance et à ceux qui ne l'atteignent pas.

Les résultats montrent que notre modélisation produit des estimations plutôt satisfaisantes, avec la majorité des écarts limités au niveau départemental. Seuls un ou deux départements présentent des écarts plus conséquents. Ces écarts s'expliquent par le fait que nous utilisons deux modèles, avec des erreurs d'estimations possible pour chacun d'eux, mais aussi par la rareté et la complexité des sinistres les plus extrêmes, qui restent difficiles à estimer au million d'euros près. Nous sommes ainsi globalement satisfaits des résultats obtenus. Concernant les évènements de plus faible intensité, une tendance à la surestimation des dégâts a été observée, tout en restant globalement cohérente. Des

pistes d'amélioration sont possibles comme l'intégration de nouvelles données, telles que les caractéristiques des toitures des bâtiments, ou encore l'extension de la modélisation à d'autres portefeuilles, notamment pour les lignes automobiles, commerciales et agricoles.

Ce modèle permet ainsi d'obtenir une estimation rapide de la charge associée au risque de grêle dès la connaissance d'un évènement de grêle, et de déterminer si cet évènement va toucher la réassurance et à quelle hauteur. Cette capacité d'anticipation facilitera l'élaboration de stratégies adaptées par les équipes de réassurance d'AXA.

Liste des graphiques

1 Cartes de l'évènement QIARA de Juin 2022 en France selon différentes agrégations temporelles et géographiques des données Meteomatics 10
2 Méthodologie d'entraînement et de test des données 10
3 Carte des écarts de prédictions pour un évènement de classe 1 (gauche) et de classe 0 (droite) 12
4 Maps of the QIARA event in France (June 2022) based on different aggregations of Meteomatics data 14
5 Methodology for Training and Testing Data 14
6 Map of prediction discrepancies for a class 1 event (left) and a class 0 event (right) 16
7 Schéma de la formation de la grêle Source : ANELFA 21
8 Historique de la fréquence et du coût moyen grêle depuis 1990 Source : France Assureurs (2022) 23
9 Carte des grêlimètres en France Source : ANELFA 25
10 Grêlimètre Source : ANELFA 25
11 Plaque en polystyrène Source : ANELFA 25
12 Géographie du risque-agent grêle en France Source : F. VINET (1998) . . . 26
13 Saisonnalité de la grêle en France Source : ESWD (2006-2020) 27
14 Saisonnalité de la grêle en Italie Source : ESWD (2006-2020) 28
15 Cycle annuel de la grêle en Europe en distinguant les régions situées au nord (en bleu) et au sud du 46e parallèle nord (en rouge) Source : Pùčik et al (2019) 29
16 Fréquence horaire de la grêle en France Source : ESWD (2006-2020) 30
17 Fréquence relative du type d'impact en fonction du diamètre maximal de la grêle, Source : Pùčik et al. (2019) 30
18 Visualisation de l'évènement QIARA de juin 2022 en France Source : ESWD 41
19 Formulaire d'envoi de rapport ESWD 41
20 Carte d'aléa de la grêle en France métropolitaine Source : Carte construite à partir des données ESWD sur 2006-2020 43
21 Augmentation de la fréquence de grêle dans le nord de l'Italie Source : ESSL (2022) 44
22 Données Meteomatics 45
23 Evènement de grêle QIARA du 19 au 23 juin 2023 Source : Meteomatics . 46
24 Données Météo-France, GSOD et Météomatics (gauche) et ESSL (droite) . 51

25 Carte de la densité de la population de la France en 2020 (gauche), au logarithme de la densité (droite) Source : global Human Settlement Layer . 54
26 Méthode du bagging et du boosting 63
27 Décalage entre déclaration et occurrence des évènements Source : AXA, Meteomatics, Météo-France et GSOD 69
28 Importance du sens de l'agrégation des données Meteomatics Source : Meteomatics 71
29 Importance des caractéristiques de l'agrégation : Cas du Loiret du 19 au 23 juin 2022 (gauche) et du 02 au 06 juin 2022 (droite) 74
30 Densité de la population du Loiret 75
31 Quantile_95_max 83
32 Quantile_95_max_POP 83
33 nb_2_4_sum 83
34 nb_4_sum 83
35 Max_Max 83
36 Duree_Max_grele 83
37 Méthodologie d'entraînement et de test des données 84
38 Distribution des classes des départements sinistrés selon les évènements . . 86
39 Distinction des classes d'évènements en fonction du montant total de sinistre des différents évènements 88
40 Sélection de variables par XG Boost - évènement de classe 1 98
41 Importance des variables pour le modèle XGB avec duree_max_grele_pop 100
42 Importance des variables pour le modèle XGB avec nb_2_4_NWMT_0.1_5100
43 Importance des variables pour le modèle XGB avec T_Max_Quantile_85_POP100
44 Sélection de variables par XG Boost - évènement de classe 0 102
45 Importance des variables pour le modèle XGB avec nb_4_sum 103
46 Importance des variables pour le modèle XGB avec Quantile_95_sum_POP103
47 Importance des variables pour le modèle XGB avec TVAR_mean_Q90_sum_POP103
48 Distribution de la variable TD_LOG 105
49 Sélection de variables par GLM - évènement de classe 1 107
50 Sélection de variables par GLM - évènement de classe 0 113
51 Écart de prédiction par département selon les différentes combinaisons . . . 116
52 Carte des écarts de prédictions pour la combinaison Duree_max_grele_POP en XG Boost et WS_0_max en GLM 117
53 Carte des écarts de prédictions pour la combinaison nb_2_4_NWMT_5 en XG Boost et NWS_2_max en GLM 117
54 Comparaison des TD_LOG prédits et observés (évènements de classe 1) . 118

55 Carte des écarts de prédictions pour la combinaison nb_4_sum en XG Boost et Quantile_90_NWMT_5 en GLM 122
56 Carte des écarts de prédictions pour la combinaison Quantile_95_sum_POP en XG Boost et Quantile_85_NWMT_5 en GLM 122
57 Comparaison des TD_LOG prédits et observés (évènements de classe 0) . 122
58 Distribution des valeurs de Nombre de communes touchées en fonction de la classe de sinistre 123
59 Distribution des valeurs de nb_4_sum en fonction de la classe de sinistre . 123
60 box-plot pour les variables nb_2_4_sum et nb_4_sum en fonction de la classe de l'évènement Source : Meteomatics 124
61 Carte des SAU en France en 2020 (hectares) Source : ANCT (2020) 128
62 Part des prairies dans la SAU en 2020 (%) Source : AGRESTE (2020) . . 128

Liste des tables de données

1 Exemple de données d'exposition en Millions d'euros . . . 34
2 Exemple de correspondance code postal - code commune en France 38
3 Niveaux de vérification des données d'ESWD Source : ESWD 42
4 Extrait des variables utilisées des données de ESWD Source : ESWD, European Severe Weather Database (ewd.eu) 42
5 Paramètres des données météorologiques du fichier RR-T Vent de Météo-France 48
6 Paramètres des données météorologiques de GSOD 50
7 Paramètres des lois usuelles 59
8 Exemples de fonctions de perte 63
9 Tableau comparatif des agrégations au Loiret Source : Agrégations issues des données Meteomatics 74
10 Diamètres de grêle sous format raster pour 1 heure, 00:00 par exemple . . . 79
11 Diamètres de grêle sous format table de donnée 80
12 Descriptif des classes de sinistres 86
13 Exemple de distribution des classes de sinistres pour les évènements de catégorie 1 90
14 Exemple de distribution des classes de sinistres pour les évènements de catégorie 0 90
15 Exemple d'échantillon généré par l'algorithme 2 90
16 Matrice de confusion 92
17 Exemple de matrice de confusion avec plusieurs classes de sinistres 94
18 Extrait de la modélisation de la classe de sinistralité pour les évènements de classe 1 97
19 Résultats de la variable Duree_max_grele_POP sur les différents échantillons 99
20 Résultats de cette variable après sélection 99
21 Extrait de la modélisation de la classe de sinistralité pour les évènements de classe 0 101
22 Résultats de la modélisation du taux de destruction pour les évènements de classe 1 107
23 Entraînement des modèles sur les échantillons 109
24 Résultats GLM sur échantillon 2 110
25 Résultats GLM sur échantillon 5 110
26 Résultats du modèle retenu sur les échantillons 111

27 Résultats de la modélisation du taux de destruction pour les évènements de classe 0 112
28 Extrait des résultats de la modélisation croisée XGB et GLM Évènements de classe 1 115
29 Résultat de la validation du modèle croisé 115
30 Détail des valeurs pour les variables du modèle XGB avec Duree_max_grele_POP 119
31 Extrait des résultats de la modélisation croisée XGB x GLM Évènements de classe 0 120
32 Résultat de la validation du modèle croisé 121
33 Détail des valeurs pour quelques variables du modèle XGB avec Nb_4_sum 123
34 Analyses pour la variable nb_2_4_sum 125
35 Analyses pour la variable nb_4_sum 125
36 Récapitulatif des abréviations issues des agrégations des données Meteomatics 140

Bibliographie

[1] L'ACTUARIEL n°53. (2024, 23 juin). *Myriam Merad, Directrice de recherche au CNRS (aide à la décision en matière de risque) - L'Actuariel.* https://www.lactuariel.fr/metier/myriam-merad-directrice-de-recherche-au-cnrs-aide-a-la-decision-en-matiere-de-risque/

[2] AGRESTE (2020). Service statistique ministériel de l'agriculture. Base de données utilisée dans le mémoire: Part des prairies dans la SAU en 2020. https://stats.agriculture.gouv.fr/cartostat/partprairie20

[3] ANCT (2020). Agence nationale de la cohésion des territoires (2020). Base de données utilisée dans le mémoire: : Carte des SAU en France en 2020. https://www.observatoire-des-territoires.gouv.fr/outils/cartographie-interactive/agri.sau2020

[4] ANELFA. *Association Nationale d'Etude et de Lutte contre les Fléaux Atmosphériques.* https://www.anelfa.asso.fr/Association-Nationale-d-Etude-et-de-Lutte-contre-les-Fleaux-Atmospheriques.html

[5] AXA GLOBAL RE. https://www.linkedin.com/company/axa-global-re/about/

[6] AXA (2022). *Finalisation de la transformation d'AXA S.A. en réassureur interne du Groupe à l'issue de sa fusion avec AXA Global Re.* https://www.axa.com/fr/presse/communiques-de-presse/finalisation_de_la_transformation_d_axa_sa_en_reassureur_interne_du_groupe

[7] AXA (2023). *Rapport annuel 2023.* https://www-axa-com.cdn.axa-contento-118412.eu/www-axa-com/5d84b947-baeb-438a-8429-227512a1bd0a_axa_urd2023_accessibleb_vf.pdf

[8] BARADEL, N. (2024). *Cours d'assurance dommage* enseigné en 2024. https://www.nicolasbaradel.fr/enseignement/ressources/cours_assurance_dommage.pdf

[9] BREIMAN, L., FRIEDMAN, J., & STONE, C. J. (1984). *Classification and regression trees.* https://doi.org/10.1201/9781315139470

[10] CODE DES ASSURANCE. *Article L125-1 - Code des assurances - Légifrance.* https://www.legifrance.gouv.fr/codes/article_lc/LEGIARTI000047109084

[11] COPERNICUS. Sources des données de Copernicus : https://cds.climate.copernicus.eu/datasets/reanalysis-era5-single-levels?tab=overview

[12] DENUIT, M., CHARPENTIER, A., & BÉBÉAR, C. (2004). *Mathématiques de l'Assurance Non-Vie. Tome I : Principes Fondamentaux de Théorie du Risque.*

[13] ESWD. *European Severe Storms Laboratory.* https://eswd.eu/cgi-bin/eswd.cgi

[14] FÉDÉRATION FRANÇAISE DU BÂTIMENT (2024). *Indice FFB* : https://www.outils.ffbatiment.fr/federation-francaise-du-batiment/le-batiment-et-vous/en_chiffres/indices-index/Chiffres_Index_FFB_Construction.html

[15] FONDATION DE PRÉVENTION DES ÉTABLISSEMENTS CANTONAUX D'ASSURANCE (2007). *Répertoire de la protection contre la grêle.* https://cms.vkg.ch/media/ya3k4oen/hagel_f.pdf

[16] FRANCE ASSUREURS (2022). *L'assurance des évènements naturels en 2022, Assurances de biens et de responsabilité.*https://www.franceassureurs.fr/wp-content/uploads/lassurance-des-evenements-naturels-en-2022.pdf

[17] FRIEDMAN, J. H. (1999). *Greedy Function Approximation : A Gradient Boosting Machine. https://www.jstor.org/stable/2699986.* https://www.jstor.org/stable/2699986

[18] GÉOCONFLUENCE (2005). *Risque.* https://geoconfluences.ens-lyon.fr/glossaire/risque-s/@@download_pdf?id=risque-s&uid=91c90b6a9e5f8f6af5c40d61c3fc58b5

[19] HOLZER, A. M. , ESSL Director of Operations. (2022). *ESSLs' European Severe Weather Database (ESWD) – its manifold uses as an example of citizen science and NGO participation.* https://wmo.int/sites/default/files/4._Alois_HOLZER_20221102_ESSL_ESWD_at_WMO_region_VI_meeting.pdf

[20] De l'IA, intelligence artificielle générative en particulier https://chatgpt.com/ a été utilisée afin de reformuler mes propres idées

[21] KERAUNOS. *La grêle en France : climatologie de la grêle en France, statistiques grêle, records du plus gros grêlon, fréquence de la grêle, climat.* https://www.keraunos.org/climatologie/les-orages-en-france/grele-climatologie-france-records-saisonnalite-risque-degats

[22] MÉTÉO-FRANCE (2023).

- *La réutilisation des données publiques devient gratuite.* https://www.info.gouv.fr/actualite/meteo-france-la-reutilisation-des-donnees-publiques-devient-gratuite.
- *Orages : pourquoi grêle-t-il ?* https://meteofrance.com/actualites-et-dossiers/questions/orages-pourquoi-grele-t-il

[23] METEOMATICS. Sources des données de Meteomatics : https://www.meteomatics.com/en/weather-api/about

[24] PÚČIK, T., CASTELLANO, C., GROENEMEIJER, P., KÜHNE, T., RÄDLER, A. T., ANTONESCU, B., & FAUST, E. (2019). *Large Hail Incidence and Its Economic and Societal Impacts across Europe*, https://doi.org/10.1175/MWR-D-19-0204.1

Annexe

Récapitulatif des abréviations

Abréviations	Nom complet	Description
duree_max_grele	Durée Maximale Grêle	Durée maximale des différents épisodes de grêle en heure
duree_grele	Durée Grêle	Durée cumulée des épisodes de grêle en heure
Max	Maximum	Maximum des diamètres de grêlons
nb_2_4	Nombre Grêlons entre 2 et 4 cm de diamètre	Nombre de grêlons entre 2 et 4 centimètres de diamètre
nb_4	Nombre Grêlons au-dessus de 4 cm de diamètre	Nombre de grêlons d'un diamètre supérieur à 4 centimètres de diamètre
nb_grele	Nombre Grêle	Nombre d'épisode de grêle
NWMT	Non Weighted Mean Top (T)	Moyenne simple des T plus grandes valeurs des diamètres de grêlons
NWS	Non Weighted Sum (S)	Somme des diamètres de grêlons qui dépassent le seuil S
POP	Population	Prise en compte du logarithme de la densité de population de GHSL dans les calculs
Quantile	Quantile (α)	Quantile d'ordre α
Sum	Somme	Somme simple des diamètres de grêlons
TVaR_mean	Tail Value at Risk Mean (α)	Moyenne du diamètre des grêlons attendue sachant que le quantile d'ordre α est dépassé
TVaR_med	Tail Value at Risk Median (α)	Médiane du diamètre des grêlons attendue sachant que le quantile d'ordre α est dépassé
Variance	Variance	Variance des diamètres de grêlons
WM	Weighted Mean	Moyenne pondérée des diamètres de grêlons
WNEM	Weighted Normalised Exponential Mean (P)	Moyenne pondérée par l'exponentiel des diamètres de grêlons divisés par le seuil de normalisation P
WNES	Weighted Normalised Exponential Sum (P)	Somme pondérée par l'exponentiel des diamètres de grêlons divisés par le seuil de normalisation P
WS	Weighted Sum (S)	Somme pondérée des diamètres de grêlons au carré qui dépassent le seuil S
WSM	Weighted Square Mean (S)	Moyenne pondérée par le carré diamètres de grêlons qui dépassent le seuil S
WSS	Weighted Sum Square (S)	Somme pondérée par le carré des diamètres de grêlons au cube qui dépassent le seuil S.
WSM	Weighted Square Mean (S)	Moyenne pondérée par le carré diamètres de grêlons qui dépassent le seuil S
WSS	Weighted Sum Square (S)	Somme pondérée par le carré des diamètres de grêlons au cube qui dépassent le seuil S.

Table 36: Récapitulatif des abréviations issues des agrégations des données Meteomatics