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

le 9 Septembre 2022

Par : Mouhamed Moustapha NDOUR
Titre : Estimation de la charge à l'ultime pour les Évènements climatiques de Grande Ampleur

Confidentialité : Non

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

Membre présent du jury de l'EURIA :
Philippe LENCA
Signature :

Entreprise :
AXA France

Membres présents du jury de l'Institut des Actuaires :
Pierre CORREGE
Romain NOBIS
Signatures :

Directeur de mémoire en entreprise :
Clémentine VIE
Signature :

Invité :

Signature :

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

Signature du responsable entreprise :

Signature du candidat :

EURIA
EURo Institut d'Actuariat

6, avenue le Gorgeu
CS 93837
29238 Brest Cedex 3

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

Résumé

Les changements climatiques et l'atypisme lié à ceux-ci représentent des enjeux de plus en plus importants aujourd'hui. D'après le rapport de France Assureurs sur l'impact du changement climatique sur l'assurance à horizon 2050, les indemnisations cumulées par les assureurs d'ici 2050 avoisineraient les 143 Mds d'euros contre 74,1 Mds d'euros en 2019. Ainsi, sur les 30 prochaines années, on pourrait voir le montant des sinistres climatiques quasiment doubler. Compte tenu des enjeux financiers que cela représente, l'actuaire doit être à même d'estimer correctement ces évènements.

L'étude réalisée dans le cadre de ce mémoire tourne autour de la prédiction de la charge à l'ultime spécifiquement sur les Évènements climatiques de Grande Ampleur (EGA). Ainsi, dans un premier temps, il nous sera important de bien définir la notion d'Évènements de grande ampleur. Il a fallu dans un premier temps constituer une base consolidée et plus exhaustive d'évènements en utilisant des algorithmes de détection d'anomalies. Nous avons construit cette nouvelle base à partir de la base historique de données d'AXA qui est à une maille sinistres. On a pu ainsi créer une nouvelle base de détection d'anomalies constituée de zones ayant été touchées par un EGA.

Afin de prédire au mieux la charge à l'ultime, plusieurs méthodes de Machine Learning et de MLG sont testées. La méthode actuelle utilisée au sein d'AXA se base sur une détermination de l'évènement "le plus proche" de celui à estimer duquel on déduit le volume de sinistres et le coût moyen à l'ultime. Les modèles que nous avons testés et qui nous permettent d'obtenir les meilleurs résultats sont le gradient boosting pour le calcul du volume et le Modèle Linéaire Généralisé avec la loi log-normale pour le calcul du coût moyen. Nous confrontons ces résultats au modèle d'AXA en appliquant sur l'exemple de la tempête Eunice.

Mots clefs: Charge à l'ultime, Évènements de Grande Ampleur, Zonier, Machine Learning, MLG, Modèle de seuil, KNN, Gradient Boosting, Prédiction

Abstract

Climate change and the atypicality linked to it represent increasingly important issues today. According to the France Assureurs report on the impact of climate change on insurance by 2050, the cumulative indemnities paid by insurers by 2050 will be close to 143 billion euros, compared to 74.1 billion euros in 2019. Hence, over the next 30 years, we could see the amount of climate-related claims almost double. Given the financial stakes involved, the actuary must be able to correctly estimate these events.

The study carried out within the framework of this dissertation revolves around the prediction of the ultimate cost specifically on Large-Scale climatic Events (LSE). Thus, at first, it will be important to define the concept of Large Scale Events. Until now, all large-scale events were grouped together on two separate bases. It was therefore necessary to create a consolidated and more exhaustive database of events by using anomaly detection algorithms. We built this new database from the historical database that we have at the level of claims within AXA. We were thus able to create a new anomaly detection database made up of areas that had been affected by an EGA.

In order to best predict the ultimate cost, several Machine Learning and GLM methods are tested. The current method used at AXA is based on the determination of the "closest" event to the one to be estimated, from which we deduce the volume of claims and the average final cost. The models that we have tested and that allow us to obtain the best results are the gradient boosting for the calculation of the volume and the Generalized Linear Model (MLG) with log-normal law for the calculation of the average cost. We will compare these results with AXA's model by applying them to the example of the Eunice storm.

Keywords: Ultimate cost, Large-scale events, Zoner, Machine Learning, GLM, Threshold model, KNN, Gradient Boosting, Prediction

Note de synthèse

Mise en contexte

Ces dernières années, les sujets et travaux autour des risques climatiques se sont multipliés. Le dernier sommet sur le changement climatique (COP 26) nous rappelle à quel point il est crucial pour tous de bien comprendre et appréhender les différents enjeux climatiques. Ces enjeux sont d'autant plus importants pour les assureurs que l'évolution climatique a des conséquences directes sur la gestion des différents sinistres couverts ainsi que les coûts engendrés par ces derniers. Les Évènements climatiques de Grande Ampleur (EGA) sont au coeur des risques climatiques et représentent les évènements dont la charge (règlements + provisions) pour Axa France est supérieure à 1M€.

Au sein d'AXA France, il y a ainsi des enjeux multiples en ce qui concerne les estimations à chaud de la charge à l'ultime ; en plus des enjeux de provisionnement, elles permettent de déterminer ce qui est cessible à la réassurance mais également au-delà de ça de calibrer le dispositif de crise. En effet, il est important d'avoir un bon ordre de grandeur de cette charge pour avoir une idée du nombre de personnes à mobiliser opérationnellement sur le sujet afin d'être au bon niveau d'assistance pour les assurés. On effectue donc un suivi de l'évolution de la charge pour tous les évènements climatiques et pour ceux ayant des impacts financiers très importants (charge>1M€), on calcule une charge à l'ultime pour l'évènement en question.

Pour l'identification des évènements, AXA a signé un partenariat avec WikiPredict, filiale de France Meteo, qui envoie des alertes avec un indice de gravité par commune lorsqu'il y a des Évènements de Grande Ampleur. Le modèle actuel utilisé au sein d'AXA se base sur une bibliothèque d'évènements au sein de laquelle s'effectue une recherche de l'évènement le plus proche en terme de cadences d'ouvertures des sinistres pour estimer la charge à l'ultime suivant une règle de trois, sans tenir compte de la zone géographique. Il s'agit donc pour nous d'explorer de nouvelles méthodes d'automatisation du calcul de la charge à l'ultime. Pour ce faire, nous nous sommes tout d'abord concentrés sur un processus de construction et d'enrichissement de la base de données d'Évènements de Grande Ampleur, allant de la création d'un zonier à la détection d'anomalies. L'estimation de la charge à l'ultime sera ensuite effectuée à l'aide de cette nouvelle base de

données.

Modèle d'AXA France

La formule utilisée pour le calcul de la charge à l'ultime est la suivante :

$$Charge\ ultime = Volume\ ultime * Cout\ moyen$$

Pour déterminer la charge à l'ultime, il faut ainsi passer par l'estimation du coût moyen et du volume à l'ultime. Pour calculer le volume à l'ultime, il faut au préalable construire les cadences d'ouvertures cumulées des sinistres, c'est-à-dire l'évolution cumulée du nombre de sinistres ouverts journalièrement après la fin de l'évènement. C'est à partir des premières cadences d'ouvertures que l'on essaie de prédire le volume final de sinistres.

Chez AXA France, l'estimation du volume à l'ultime se fait en recensant dans un premier temps l'ensemble des évènements climatiques sur un historique d'au moins 10 ans puis en recherchant l'évènement le plus proche en terme de nombre d'ouvertures de sinistres. Le calcul pour déterminer l'évènement le plus proche se fait suivant la formule suivante :

$$Distance = \sum_{\substack{1 \leq i \leq n \\ 1 \leq j \leq n}} \left(x_{i,j}^2 - y_{i,j}^2\right)$$

avec *y la matrice de cadencement du nombre d'ouvertures pour l'évènement que l'on cherche à estimer et x la matrice de cadencement du nombre d'ouvertures d'un évènement de même type.*

On récupère ensuite le pourcentage du volume ultime atteint à la même vision de l'évènement le plus proche que l'on applique à l'évènement qu'on cherche à estimer :

$$Volume\ ultime = \frac{Volume\ sinistres}{\%\ du\ volume\ ultime\ de\ l'evenement\ le\ plus\ proche\ a\ la\ meme\ periode}$$

Pour le calcul du coût moyen, on utilise le coût moyen calculé pour des évènements passés de même nature.

L'exemple d'application à la tempête Eunice fait dans le corps du mémoire permet d'illustrer les étapes précédentes. L'exemple permet de déterminer un nombre d'ouvertures à l'ultime autour de 15000 sinistres au bout de 3 jours pour un coût moyen autour de 2500 euros avec une marge d'erreur de plus ou moins 20%. Cette marge d'erreur est

liée à la forte volatilité sur les cadences d'ouvertures et les coûts moyens d'une année à une autre.

Construction de la base de données

Notre objectif est donc d'enrichir notre catalogue d'évènements climatiques en utilisant plus d'informations sur les évènements détectés avec comme ambition d'améliorer la prédiction de la charge à l'ultime des nouveaux évènements.

Compte tenu des données à notre disposition, de leur forme et de leur contenu, on a besoin de redéfinir ce qu'est un EGA. On va, dans un premier temps, se concentrer sur la **détection d'anomalies** au sein de nos données : lorsque pour un lieu et une date/période de temps donnés, on observe un volume total de garanties sinistrées plus important/aberrant que d'habitude/la normale, alors il s'agira d'une anomalie.

La base de données initiale concerne un échantillon de près de 3 millions d'observations (2 914 000) pour des aléas climatiques tempête, grêle, inondation et séisme. Il s'agit d'une base de données avec arrêté des comptes à Juin 2022 et avec des sinistres survenus et clos de 1989 à 2021.

Le nombre de lignes pour lesquelles l'information relative au lieu du sinistre est manquante est de 1 526 130 lignes, ce qui correspond à 52.4% de nos données. La première partie de nettoyage et de traitement des données va nous permettre de récupérer plusieurs informations qui étaient manquantes pour le lieu du sinistre. Au final, on se retrouve à peu près avec 30% de lignes pour lesquelles on ne dispose d'aucune information possible à retraiter.Une bonne partie de ces lignes correspondent à des sinistres ayant eu lieu entre 1989 et 1999. Pour ces raisons, nous avons décidé de filtrer nos données sur l'horizon 2000-2021. On se retrouve désormais avec un peu plus 2 240 000 lignes allant de 2000 à 2021 et qui contiennent une information exploitable relative au lieu.

De plus, un examen préliminaire nous a montré qu'il existe une trop grande disparité entre les communes au cours de ces années d'historique en termes de volume ainsi qu'en termes de fréquence journalière : pour beaucoup de communes il y a beaucoup de jours au cours desquels il ne s'est rien passé. Afin "d'harmoniser" les volumes d'UP associé à chaque individu, nous avons procédé à l'élaboration d'un zonier. En effet, la France va être découpée de manière géométrique et équitable indépendamment des communes, de leurs coordonnées géographiques et de leurs frontières. La création du zonier nous a permis de prendre en compte l'appartenance de chaque individu à une zone plus étendue et d'atténuer les disparités.

A partir de ce zonier, on a testé des modèles de détection d'anomalies : les threshold

models et l'Isolation Forest.

Parmi les threshold models, on distingue le fixed threshold models et l'adaptive threshold models. Le fixed threshold model est basé sur une variable de seuil fixe et figée. Le seuil de 3000 a été déterminé par l'approche GPD de la théorie des valeurs extrêmes. L'adaptive threshold models est lui basé sur une variable de seuil adaptative : le principe reste le même que le premier modèle, c'est à dire qu'on compare chaque évènement à une valeur de référence mais celle-ci n'est plus unique et figée. Concrêtement, pour chaque individu/zone, on veut que X% des évènements qui ont eu lieu ressortent en anomalie. Les cibles testées sont 1%, 3% et 5%.

En ce qui concerne l'algorithme d'Isolation Forest, il détecte les anomalies au sein d'une série temporelle en utilisant le principe d'isolation, c'est-à-dire en s'intéressant à la distance entre l'observation analysée et le reste des données ; au lieu d'essayer de créer un modèle d'instances « normales », elle isole et modélise explicitement les points anormaux dans l'ensemble de données. Son avantage est qu'il fonctionne bien sur les gros volumes de données comme le nôtre.

Pour l'évaluation des différents modèles de détection d'anomalie, nous sommes dans le cadre d'un apprentissage non-supervisé. De ce fait, notre jeu de données ne contient pas d'exemples labellisés. Dans cette configuration, il a fallu faire le travail à la main et comparer du mieux possible les individus-dates classifiés en anomalies avec les données déjà existantes d'AXA. On considère notre modèle performant à partir du moment où au moins 90% des anomalies identifiées semblent coïncider avec des EGA ayant eu lieu.

Modèle	% d'anomalies coincidant avec des EGA
Fixed Threshold	68%
Adaptative threshold 1%	81%
Adaptative threshold 3%	78%
Adaptative threshold 5%	75%
Isolation Forest	93%

Table 1 – Résultats des modèles

Notre meilleur modèle est l'Isolation Forest. Sur les près de 2 240 000 sinistres recensés, l'algorithme d'Isolation Forest détecte près de 960 000 anomalies pour 2 544 zones agrégées affectées. Nous avons donc gardé ce modèle.

Calcul de la charge à l'ultime : Modélisation par Machine Learning et MLG

Dans l'optique de palier au manque d'informations relatives aux Évènements de Grande Ampleur, nous avons procédé précédemment à la création d'un nouveau catalogue, d'une nouvelle base consolidée et fiable prenant en compte le critère géographique afin de l'exploiter et de prédire au mieux la charge à l'ultime. Cette base est donc constituée de 960 000 lignes initialement correspondant à un peu plus de 32 000 communes (représentées par leurs codes communes) où on a détecté une anomalie, donc un probable EGA entre début 2000 et fin 2021 (possibilité donc de retrouver sur plusieurs lignes différentes une même commune). Pour chaque code commune, on a également la variable relative aux zones agrégées auxquelles chaque code commune est rattaché. On s'est vite rendu compte qu'à la maille commune, on manquait fortement d'exposition en terme de sinistres. Nous sommes donc passés à la maille zone agrégée que nous avons définie grâce à notre zonier. Nous incluons dans notre base toute variable jugée pertinente par rapport à la variable cible, qui dans notre cas correspond à la charge ultime.

Le calcul de la charge à l'ultime dans le cadre des EGA repose sur 2 estimations fondamentales :

- **l'estimation des volumes de sinistres à l'ultime :** nous allons pour cela utiliser une approche par **Machine Learning** avec les algorithmes de KNN, Random Forest et Gradient Boosting

  On commence à construire la base avec projection des cadences d'ouvertures de sinistres. De plus, on souhaite être à mesure d'estimer le volume final des sinistres en début d'évènement. Donc pour le modèle, on se limitera à la construction de cadences jusqu'à 3 jours après ouverture des sinistres.

DTSURV	JourSurv	month	Year	dureeEvenement	part_auto	part_transport	part_RC	part_dommages	part_constructions	J1	J2	J3	JFinal
2009-01-23	Friday	1	2009	2 days	0.188679	0.000000	0.0	0.995631	0.000000	201.0	332.0	433.0	1605.0
2009-01-24	Saturday	1	2009	2 days	0.039474	0.000000	0.0	0.999460	0.000000	3119.0	6931.0	10521.0	79286.0
2009-01-25	Sunday	1	2009	2 days	0.121951	0.000000	0.0	0.998309	0.000000	115.0	162.0	180.0	494.0

TABLE 2 – Aperçu de la base finale de construction des cadences d'ouverture

On décompose le jeu de données en 2 groupes : les données pour l'apprentissage (70%) et les données pour les tests (30%). Puis, on a calibré nos modèles par validation croisée. Une fois le calibrage réalisé et nos paramètres optimisés grâce à la validation croisée, on peut enfin procéder à l'entraînement des modèles sur notre échantillon d'apprentissage. On compare l'exactitude des prévisions grâce au

RMSE et au MAPE qui sont des métriques d'erreur souvent utilisées pour évaluer des modèles.

	KNN	Random Forest	Gradient Boosting
RMSE	6287.7403	5827.0283	5681.2081
MAPE	0.6150	0.2503	0.2253

TABLE 3 – Evaluation de la performance des différents modèles selon le MAPE et le RMSE

Le modèle qui a le RMSE le plus faible et le MAPE le plus faible est le gradient boosting. Celui-ci permet de donner une prédiction des volumes à l'ultime avec une marge d'erreur de plus ou moins 22%. Cette marge d'erreur peut paraître énorme, néanmoins cela reste un résultat plutôt satisfaisant quand on connaît la forte volatilité autour des EGA. De plus, l'analyse de l'importance des variables nous permet de constater l'importance de la variable de zone géographique qu'on a rajoutée.

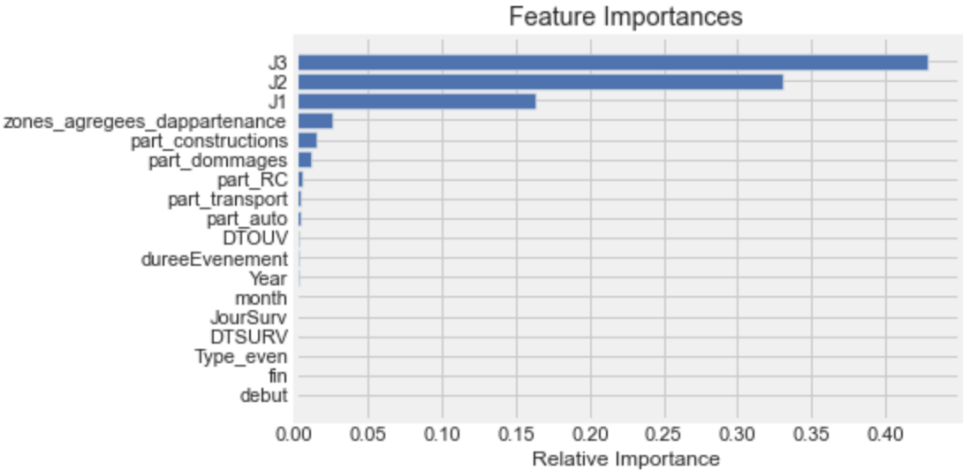


FIGURE 1 – Importance des variables avec la modélisation par gradient boosting

- **l'estimation du coût moyen :** nous allons pour cela utiliser une approche par **MLG** avec les lois Gamma et Log-Normales.

Exemple de la tempête

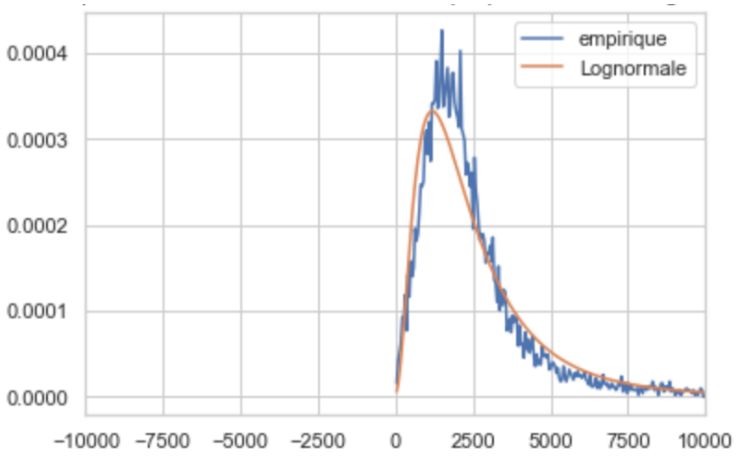


FIGURE 2 – Comparaison de la fonction de densité empirique avec une loi log-normale pour la tempête

Pour les 4 risques tempête, séisme, inondation et grêle, les meilleurs résultats ont tous été donnés avec le modèle log-normal. Ainsi, le MAPE calculé en utilisant ce modèle nous dit que notre coût moyen à l'ultime est assez bien prédit avec un intervalle de confiance autour de +/- 10% pour les 4 risques.

L'estimation de la charge à l'ultime pour un évènement donné se fera donc en réalisant le produit entre l'estimation du volume à l'ultime faite en section 4.2 et l'estimation du coût moyen qui vient d'être faite.

L'application des modèles retenus sur l'exemple de la tempête Eunice nous donne un coût moyen à l'ultime d'environ 2916 euros +/- 10% pour un volume à l'ultime autour de 10000 sinistres +/- 22%.

L'intervalle de confiance du volume estimé n'est pas forcément plus fin que celui du modèle d'AXA mais nous permet déjà d'avoir en début d'évènement (à JO+3) un bon ordre de grandeur contrairement au modèle d'AXA où l'estimation a été revue à la baisse 20 jours ouvrés après l'ouverture des sinistres liés à cet évènement.

Synthesis note

Context

In recent years, the number of topics and works around climate risks has multiplied. The latest summit on climate change (COP 26) reminds us how crucial it is for everyone to understand and grasp the various climate issues. These issues are all the more important for insurers as climate change has direct consequences on the management of the various claims covered as well as the costs generated by them. Large scale climatic events (LSE) are at the heart of climate risks and represent the events for which the cost (settlements + provisions) for Axa France is higher than 1M€.

At AXA France, there are multiple issues at stake when it comes to hot estimates of the ultimate burden ; in addition to provisioning issues, they make it possible to determine what is transferable to reinsurance, but also beyond that to calibrate the crisis mechanism. Indeed, it is important to have a good order of magnitude of this cost in order to have an idea of the number of people to be mobilized operationally on the subject in order to be at the right level of assistance for the insured. We therefore monitor the evolution of the cost for all climatic events and for those with very important financial impacts (cost>1M€), we calculate an ultimate cost for the event in question.

For the identification of events, AXA has signed a partnership with WikiPredict, a subsidiary of France Meteo, which sends alerts with a severity index by municipality when there are large-scale events. The current model used by AXA is based on a library of events from which the closest event in terms of opening rates of claims is searched to estimate the ultimate cost according to a rule of three, without taking into account the geographical area. It is thus a question for us of exploring new methods of automation of the calculation of the ultimate cost. To do so, we first focused on a process of construction and enrichment of the database of large-scale events, from the creation of a zonier to the detection of anomalies. The ultimate cost estimation will then be performed using this

new database.

AXA France model

The formula used to calculate the ultimate cost is as follows :

$$Ultimate\ cost = Ultimate\ volume * Average\ cost$$

In order to determine the ultimate cost, we must therefore estimate the average cost and the ultimate volume. To calculate the ultimate volume, we must first construct the cumulative opening rates of the claims, i.e. the cumulative evolution of the number of claims opened daily after the end of the event. We use the first opening rates to try to predict the final volume of claims.

At AXA France, the ultimate volume is estimated by first listing all the climatic events over a period of at least 10 years and then looking for the closest event in terms of the number of claims opened. The calculation to determine the closest event is done according to the following formula :

$$Distance = \sum_{\substack{1 \leq i \leq n \\ 1 \leq j \leq n}} \left(x_{i,j}^2 - y_{i,j}^2\right)$$

with *y the timing matrix of the number of claims for the event we want to estimate and x the timing matrix of the number of claims of an event of the same type.*

We then recover the percentage of the ultimate volume reached at the same vision of the nearest event that we apply to the event we want to estimate :

$$Ultimate\,volume = \frac{Claims\ volume}{\%\ of\ the\ ultimate\ volume\ of\ the\ closest\ event\ at\ the\ same\ period\ of\ time}$$

For the calculation of the average cost, we use the average cost calculated for past events of the same nature.

The example of application to the storm Eunice made in the body of the thesis allows to illustrate the previous steps. The example allows to determine a number of openings to the ultimate around 15000 claims after 3 days for an average cost around 2500 euros with a margin of error of more or less 20%. This margin of error is linked to the high

volatility of opening rates and average costs from one year to the next.

Construction of the database

Our goal is therefore to enrich our catalog of climatic events by using more information on the detected events with the ambition to improve the prediction of the ultimate cost of new events.

Given the data at our disposal, their form and content, we need to redefine what an EGA is. We will initially focus on the detection of anomalies within our data : when for a given location and date/period of time, we observe a larger/aberrant total volume of claims than usual/normal, then this will be an anomaly.

The initial database covers a sample of nearly 3 million observations (2 914 000) related to storm, hail, flood and earthquake risks. It is a database with a closing date of June 2022 and with claims occurring and closed from 1989 to 2021.

The number of rows with missing location information is 1 526 130 rows, which corresponds to 52.4 corresponds to 52.4% of our data. The first part of the cleaning and processing of the data will allow us to recover several pieces of information that were missing for the place of the disaster. In the end, we end up with about 30% of lines for which we have no possible information to process. A good part of these lines correspond to claims that took place between 1989 and 1999. For these reasons, we have decided to filter our data over the 2000-2021 horizon. We now have just over 2,240,000 rows from 2000 to 2021 that contain usable location information.

In addition, a preliminary examination has shown us that there is too much disparity between the communes during these historical years in terms of volume as well as in terms of daily frequency in terms of daily frequency : for many communes there are many days in which nothing happened. during which nothing happened. In order to "harmonize" the volumes of UP associated to each individual, we have proceeded to the elaboration of a zonier. Indeed, France is going to be cut in a geometrical and equitable way independently of the communes, of their geographical coordinates and of their borders. The creation of the zonier allowed us to take into account the belonging of each individual to a larger zone and to attenuate the disparities.

From this zonier, we tested anomaly detection models : threshold models and the Isolation Forest.

Among the threshold models, we distinguish the fixed threshold model and the adaptive threshold model. The fixed threshold model is based on a fixed threshold variable. The threshold of 3000 was determined by the GPD approach of the extreme value theory.

The adaptive threshold models is based on an adaptive threshold variable : the principle remains the same as the first model, i.e. each event is compared to a reference value, but this one is not unique and fixed anymore. Concretely, for each individual/zone, we want X% of the events that took place to be anomalous. The targets tested are 1%, 3% and 5%.

As for the Isolation Forest algorithm, it detects anomalies within a time series by using the isolation principle, i.e. by looking at the distance between the analyzed observation and the rest of the data ; instead of trying to create a model of "normal" instances, it explicitly isolates and models the anomalous points in the data set. Its advantage is that it works well on large volumes of data like ours.

For the evaluation of the different anomaly detection models, we are in an unsupervised learning framework. Therefore, our dataset does not contain labeled examples. In this configuration, we had to do the work by hand and compare as best as possible the individuals-dates classified as anomalies with the already existing AXA data. Our model is considered to be efficient when at least 90% of the identified anomalies seem to coincide with EGAs that have taken place.

Model	% Percentage of anomalies coinciding with EGA
Fixed Threshold	68%
Adaptative threshold 1%	81%
Adaptative threshold 3%	78%
Adaptative threshold 5%	75%
Isolation Forest	93%

TABLE 4 – Model results

Our best model is Isolation Forest. Of the nearly 2 240 000 claims, the Isolation Forest algorithm detects nearly 960 000 anomalies for 2 544 aggregated affected areas. We have therefore kept this model.

Calculation of the ultimate cost : Modeling by Machine Learning and GLM

In order to compensate for the lack of information on large-scale events, we have previously created a new catalog, a new consolidated and reliable database taking into account the geographical criterion in order to exploit it and to predict the ultimate cost. This database is therefore made up of 960,000 lines initially corresponding to a little over 32,000 communes (represented by their commune codes) where an anomaly

was detected, i.e. a probable EGA between the beginning of 2000 and the end of 2021 (possibility of finding the same commune on several different lines). For each commune code, we have also the variable relating to the aggregated zones to which each commune code is attached. We quickly realized that at the commune level, there was a strong lack of exposure in terms of claims. We therefore switched to the aggregated zone grid that we defined thanks to our zoner. We include in our base any variable deemed relevant to the target variable, which in our case corresponds to the ultimate cost.

The calculation of ultimate cost under the EGA is based on 2 fundamental estimates :

- **Estimation of ultimate claim volumes :** we will use an approach by **Machine Learning** KNN, Random Forest and Gradient Boosting algorithms.

  We start to build the database with a projection of the rate at which claims are opened. In addition, we want to be able to estimate the final volume of claims at the beginning of the event. Therefore, for the model, we will limit ourselves to the construction of rates up to 3 days after the opening of the claims.

DTSURV	JourSurv	month	Year	dureeEvenement	part_auto	part_transport	part_RC	part_dommages	part_constructions	J1	J2	J3	JFinal
2009-01-23	Friday	1	2009	2 days	0.188679	0.000000	0.0	0.995631	0.000000	201.0	332.0	433.0	1605.0
2009-01-24	Saturday	1	2009	2 days	0.039474	0.000000	0.0	0.999460	0.000000	3119.0	6931.0	10521.0	79286.0
2009-01-25	Sunday	1	2009	2 days	0.121951	0.000000	0.0	0.998309	0.000000	115.0	162.0	180.0	494.0

TABLE 5 – Overview of the final basis for building the claims rates

We decompose the dataset into 2 groups : the data for training (70%) and the data for testing (30%). Then, we calibrated our models by cross-validation. Once the calibration is done and our parameters are optimized thanks to the cross-validation, we can finally proceed to the training of the models on our learning sample. We compare the accuracy of the forecasts thanks to the RMSE and the MAPE which are error metrics often used to evaluate models.

	KNN	Random Forest	Gradient Boosting
RMSE	6287.7403	5827.0283	5681.2081
MAPE	0.6150	0.2503	0.2253

TABLE 6 – Evaluation of the performance of the different models according to the MAPE and the RMSE

The model with the lowest RMSE and the lowest MAPE is gradient boosting. This one gives a prediction of the ultimate volumes with a margin of error of more or less 22%. This margin of error may seem huge, but it is still a rather satisfactory result

when we know the high volatility around the EGA. Moreover, the analysis of the importance of the variables allows us to note the importance of the geographical zone variable that we added.

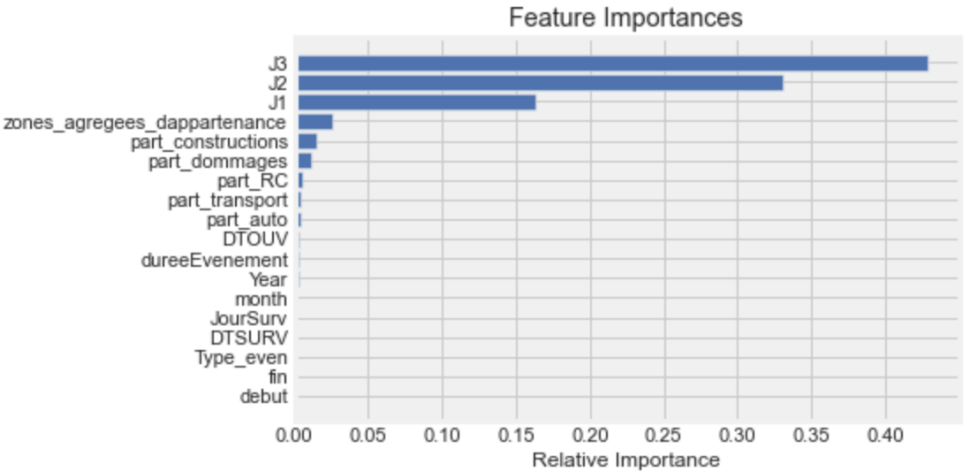


FIGURE 3 – Importance of variables with gradient boosting modeling

- **the average cost estimate :** we will use a **GLM** approach with Gamma and Log-Normal laws.

  Example of the storm

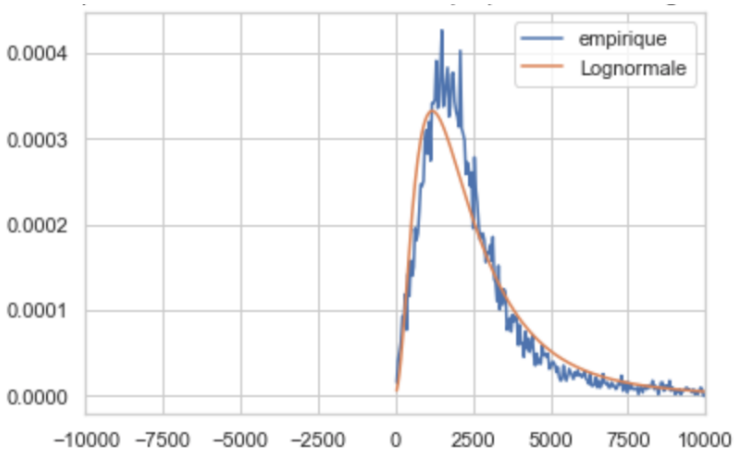


FIGURE 4 – Comparison of the empirical density function with a lognormal distribution for the storm

For the 4 risks storm, earthquake, flood and hail, the best results were all given with the log-normal model. Thus, the MAPE calculated using this model tells us that our average ultimate cost is fairly well predicted with a confidence interval around +/- 10% for the 4 risks.

The estimate of the ultimate cost for a given event will therefore be made by taking the product of the ultimate volume estimate made in section 4.2 and the average cost estimate just made.

The application of the models retained on the example of the Eunice storm gives us an average ultimate cost of about 2916 euros +/- 10% for an ultimate volume of about 10000 claims +/- 22%.

The confidence interval of the estimated volume is not necessarily finer than that of AXA's model, but it allows us to have a good order of magnitude at the beginning of the event (at JO+3), contrary to AXA's model where the estimate was revised downwards 20 working days after the opening of the claims related to this event.

Remerciements

Il m'est important de remercier les personnes qui m'ont aidé et soutenu pour le bon déroulé de ce mémoire. C'est dans ce que cadre qu'il me paraît indispensable de remercier ma tutrice en entreprise, Mme Clémentine VIE, manager de l'équipe Claims Analytics pour sa confiance, sa disponibilité et son soutien tout au long de la réalisation de ce mémoire.

Je remercie ma famille et mes amis de près ou de loin pour tout le soutien émotionnel qu'ils m'ont apporté en s'assurant que j'allais bien et en me motivant continuellement tout au long de cette année. Je remercie en particulier mon père d'être ma plus grande source d'inspiration et de motivation.

Je remercie bien entendu Mr Laurent IMBERT, responsable de la majeure Actuariat de l'École Supérieure d'ingénieurs Léonard de Vinci ainsi que Mr Franck VERMET, directeur de l'Euro Institut d'Actuariat qui m'ont permis d'acquérir et de développer de solides compétences indispensables pour évoluer dans le monde fascinant de l'actuariat.

Je remercie mes tuteurs d'alternance et de mémoire Mr Pierre AILLIOTT et Mme Laurence ABIVEN pour leur relecture et leurs précieux conseils.

Enfin, je remercie Wissam QURESHI, Ratiba MIKOU, Coline MERLIN ainsi que tous les membres de l'équipe Claims Analytics d'AXA France pour m'avoir bien accueilli et pour leur aide précieuse tout au long de ces mois de travail.

Table des matières

Introduction

Ces dernières années, la fréquence des évènements climatiques a augmenté de manière drastique dans le monde en général et en France en particulier. Face à cette vague croissante d'évènements climatiques qui pèse sur les comptes de résultat des assureurs, il est impératif de bien appréhender les différents enjeux économiques autour de l'estimation à chaud du coût de ces évènements pour les assureurs. En effet, les enjeux sont multiples. D'un point de vue opérationnel, c'est de réussir à calibrer le bon nombre de collaborateurs sur le terrain. D'un point de vue comptable, c'est de fiabiliser le calcul de la charge afin de mieux provisionner ces évènements climatiques.

Ainsi, avoir une bonne estimation de la charge à l'ultime qui représente le coût final prévisible des sinistres s'avère essentiel aujourd'hui pour tout assureur. Cela est d'autant plus important que les évènements climatiques coûtent de plus en plus cher aux assureurs.

Cependant, face à l'atypisme et la sévérité changeante de ces évènements, il est difficile aujourd'hui pour les assureurs de prédire de façon assez précise cette charge à l'ultime. En effet, les méthodes agrégées classiques telle que les méthodes de Chain Ladder et de Mack usuellement utilisées par les assureurs dans le calcul des provisions ne sont pas applicables car les hypothèses fondamentales ne sont pas respectées. L'hypothèse principale de la méthode de Chain Ladder repose sur le fait que les modèles d'évolution des pertes historiques doivent permettre de déterminer des modèles d'évolution des pertes futures. Ainsi, les facteurs de développement ne doivent pas dépendre de la survenance mais uniquement de la date de vision. Or, d'une année à une autre, les coûts relatifs aux Évènements de Grande Ampleur peuvent s'avérer très variables. Ce qui contrarie notre hypothèse.

Il est crucial de trouver d'autres moyens d'estimer au mieux cette charge. Une solution pour estimer le coût d'un évènement climatique est de projeter les volumes de sinistres attendus sur un évènement à partir des cadences d'ouverture du début de l'évènement. Cela requiert d'avoir des évènements de même nature comparables. Pour estimer le coût de l'évènement, il faudra associer un coût moyen à chaque sinistre.

Face à la montée en puissance de l'intelligence artificielle notamment dans le cadre de la gestion, de la classification et de la prédiction des données, les assureurs font d'avantage d'investissements dans la récupération et l'utilisation des données. Les méthodes d'apprentissage statistique sont encore peu utilisées dans la pratique, de par leur faible

maturité ainsi que leur faible consistance en données, mais représentent désormais des solutions de plus en plus explorées par les compagnies d'assurance pour l'estimation de la charge à l'ultime.

De précédents mémoires et travaux de recherche ont porté sur la modélisation de ces évènements qui reste fondamentale aujourd'hui pour mieux les appréhender. Cependant, notre travail ne se limitera pas à ça, il sera davantage axé sur l'estimation de la charge à l'ultime pour les Évènements de Grande Ampleur. Pour cela, nous allons challenger les méthodes dites agrégées utilisées aujourd'hui et qui se basent uniquement sur l'observation du nombre de sinistres ouverts ainsi que la typologie de l'évènement pour déterminer la charge. En utilisant uniquement ces caractéristiques, on fait face à un risque accru de sous/sur-estimation de la charge à l'ultime. Nous allons ainsi rajouter de nouvelles variables telle que la zone géographique afin d'enrichir notre base de données d'évènements.

Ainsi, dans un premier temps, nous commencerons par introduire le contexte dans lequel s'effectue ces travaux, en commençant par l'assurance non-vie en général et ses spécificités, avant d'entrer plus dans le détail des différentes branches assurantielles ainsi que les risques climatiques sur lesquels portent cette étude ; pour finir sur les notions de provisionnement et de réassurance. Dans un second temps, nous nous intéresserons à la définition des Évènements climatiques de Grande Ampleur et à la méthode d'estimation de la charge à l'ultime pour ces évènements. Les chapitres suivants consisteront à challenger cette méthode en commençant par enrichir notre base d'évènements avant de présenter notre approche d'estimation de la charge basée sur des méthodes de Machine Learning et de Modèles Linéaires Généralisés. Enfin, nous conclurons par une comparaison des 2 méthodes.

Chapitre 1

Éléments de contexte

Commençons par définir les notions clés permettant de cadrer le sujet. Ainsi, ce mémoire entre dans un corps de métier dynamique et en constante évolution qui est celui de l'assurance et plus spécifiquement de l'assurance non-vie.

1.1 Le monde de l'assurance

Le secteur de l'assurance occupe une place très importante dans le milieu de l'économie française. Par définition, l'assurance est une opération par laquelle une partie appelée « assureur » s'engage à réaliser une prestation au profit de la partie « assuré » lors de la survenance d'un risque aléatoire et moyennant le paiement d'une somme d'argent appelée « prime » ou « cotisation ». L'assurance est donc caractérisée par une inversion du cycle de production. Autrement dit, l'assuré paie un service avant même que l'entreprise ne le délivre ou ne crée une quelconque valeur.

1.1.1 L'assurance non-vie

Définition

Les assurances non-vie communément appelées assurances IARD (Incendie Accident Risques Divers) regroupent les assurances de choses et de responsabilité relatives aux dommages ainsi que les assurances santé liées à la personne, c'est-à-dire tout ce qui est en rapport avec les frais de santé sur la personne. Cette distinction dommages/personne est basée sur le principe d'indemnisation des sinistres :

- le principe indemnitaire : l'assuré se fait rembourser en fonction du préjudice subi, il n'y a aucune possibilité d'enrichissement de l'assuré
- le principe forfaitaire : ne dépend pas du préjudice subi, les montants en cas de préjudice sont contractuels et connus à l'avance

Quelques chiffres

L'assurance non-vie ne représente que 29,8% des cotisations en France (fig1.1) mais

est pourtant la plus développée dans le monde. Les charges des prestations réalisées par les assureurs en France en 2020 s'élèvent quant à elles à 182 Md€, soit 23,7% des charges totales (fig1.2).

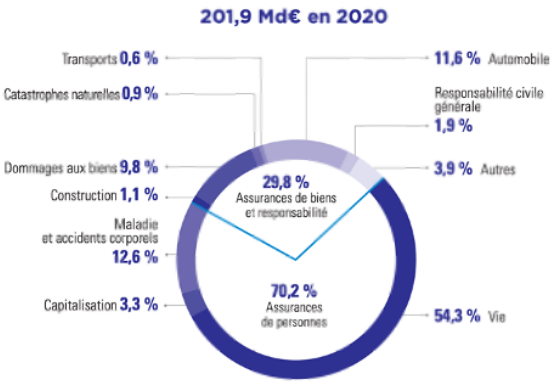


FIGURE 1.1 – Cotisations de l'assurance française en 2020 [4]

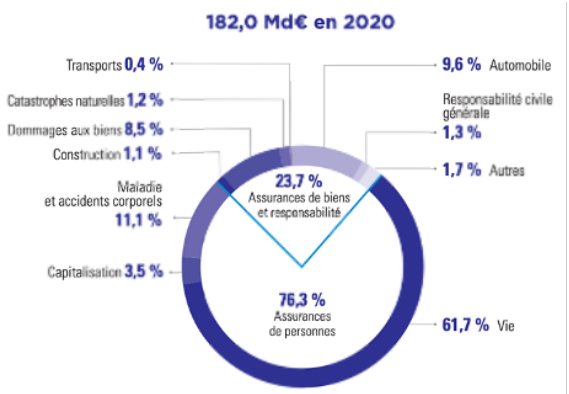


FIGURE 1.2 – Charges de l'assurance française en 2020 [4]

En assurance non-vie, on distingue **l'assurance des particuliers** qui est constituée majoritairement de la branche automobile qui constitue à elle seule près de 55,6% des cotisations (5) et **l'assurance des professionnels** où l'on peut retrouver principalement du dommage aux biens (37,1%) et la Responsabilité Civile (16,5%). On peut également avoir de l'assurance collective ou individuelle.

1.1.2 Zoom sur quelques branches de l'étude

Branche auto

L'assurance auto représente l'une des branches les plus importantes de l'assurance non-vie. En France, la branche auto, c'est 23,5 Md€ de primes collectées pour 17,4 Md€ de charges de prestations en 2020. Il s'agit d'un secteur en pleine croissance. Entre 2019 et 2020, elle a connu une hausse des primes de 3% pour une baisse des charges de 6,6% du fait notamment de la pandémie liée à la Covid-19. (5) Elle fait partie des assurances dites obligatoires au titre de la responsabilité civile. En effet, tout véhicule à moteur circulant en France est tenu d'être assuré « au tiers », c'est-à-dire d'être assuré face aux dommages que pourrait causer le véhicule aux autres.

Branche MRH

L'assurance MRH ou Multi Risques Habitations fait également partie des assurances obligatoires. Elle est obligatoire pour tous les propriétaires pour qui le logement est situé dans une copropriété. Elle est incluse dans la garantie Dommages aux biens qui s'applique

aussi bien aux particuliers qu'aux professionnels et qui permet d'être indemnisé en cas de dégradation ou perte de biens dans le cadre d'un contrat d'assurance habitation. En France, l'assurance MRH, c'est 11,7 Md€ de primes collectées pour 7,4 Md€ de charges de prestations en 2020.

Branche MRI

L'assurance MRI est relativement similaire à l'assurance multirisque habitation classique, à la seule différence qu'elle est spécifique aux immeubles. Au delà de la Responsabilité Civile, l'assurance Multirisque Immeuble prend également en compte une garantie de dommages aux biens.

Branche Risques Industriels

L'assurance des risques industriels cible spécifiquement les entreprises du secteur industriel. Depuis 2010, on fait face à des hausses de plus de 10% sur les risques industriels. Cette assurance permet ainsi aux entreprises du secteur industriel de se protéger contre les incendies, inondations, vols ou tentatives de vols, catastrophes naturelles, mais en partie les dommages liés à la production. Cette assurance est donc très importante notamment dans le contexte actuel où l'on fait face à une hausse des aléas climatiques et des catastrophes naturelles.

Branche Multirisque Professionnelle

L'Assurance Multirisque Professionnelle (MRP) est un contrat global destiné à couvrir l'ensemble des risques patrimoniaux et professionnels d'une entreprise. Par exemple, l'assurance Multirisque Professionnelle des commerçants va couvrir les dommages causés par un incendie, une tentative de vol, des actes de vandalisme ou encore les dommages aux tiers.

1.2 Les risques climatiques

Les risques climatiques représentent des risques émergeants majeurs pour toutes les compagnies d'assurance. Ceux-ci peuvent se décomposer en 3 principaux risques :

- Les risques physiques qu'on a tendance à appeler communément risques climatiques et qui sont rattachés directement aux phénomènes météorologiques
- Les risques de transition qui représentent les risques pour une société de ne pas pouvoir s'adapter à un environnement en pleine transition climatique
- Les risques de responsabilité qui sont quant à eux des risques qu'une société subissent des procès pour non-respect d'engagements vis-à-vis de l'environnement

Pour la suite, on ne se focalisera que sur les risques physiques climatiques qu'on appellera simplement risques climatiques.

Les risques climatiques représentent donc l'intégralité des risques liés aux aléas climatiques, allant de la tempête à l'inondation en passant par la sécheresse. Tous ces aléas ont des enjeux économiques importants pour la bonne gestion des assurés d'une compagnie d'assurance. En effet, les risques climatiques peuvent engendrer plusieurs autres risques tels que des risques financiers à tel point que la règlementation telle que la SFDR "Sustainable Finance Disclosure Regulation" oblige les investisseurs à évaluer à chaque fois les risques financiers dont l'origine serait due à des aléas climatiques.

L'étude Climat de France Assureurs datant d'Octobre 2021 nous montre la répartition des montants indemnisés par les assureurs sur l'horizon 1989-2019 pour les aléas tempêtes, inondations et sécheresse (fig 1.3). On remarque que près de la moitié des indemnisations portent sur la tempête. En regardant à une maille plus fine ces chiffres, avec la répartition du montant et du nombre de sinistres par catégorie d'assurés (particuliers et professionnels) (Table 1.1), on remarque un nombre de sinistres indemnisés assez élevé au niveau des tempêtes, ce qui explique leur charge importante. Mais en moyenne entre 1989 et 2019, le coût moyen global des sinistres tempêtes s'élève à 2 170 euros, contre 10 230 euros pour les inondations et 16 340 euros pour la sécheresse.

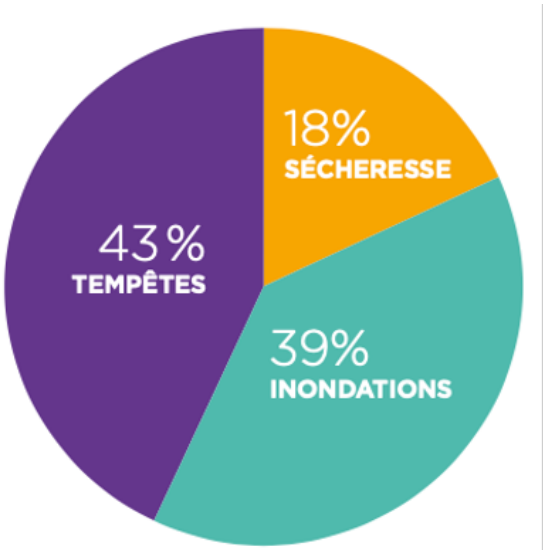


FIGURE 1.3 – Répartition du cumul des indemnisations versées par les assureurs au cours des 31 dernières années (1989 – 2019) (Etude Climat France Assureurs Octobre 2021)

De 1989 à 2019	Nombre de sinistres indemnisés	Charge (Md€ constants 2020)
INONDATIONS	**1 961 000**	**28,8**
Particuliers	1 480 000	15,1
Professionnels	481 000	13,6
TEMPÊTES	**10 105 000**	**31,6**
Particuliers	8 251 000	17,9
Professionnels	1 854 000	13,7
SÉCHERESSE	**843 000**	**13,8**
Ensemble des périls	**12 909 000**	**74,1**

TABLE 1.1 – Montant et nombre de sinistres par catégorie d'assurés (Etude Climat France Assureurs Octobre 2021)

Une partie des raisons de l'inflation des coûts des évènements climatiques est également l'augmentation de la couverture d'assurance de ces évènements, mais également l'augmentation du niveau de vie en général. Cette étude prospective de France Assureurs prévoit dans les 25 prochaines années un quasi-doublement des sinistres liés au climat.

Nous allons donc nous intéresser à différents aléas climatiques dans le cadre de ce mémoire. Il s'agit de la tempête, des inondations, de la grêle et des séismes.

1.2.1 Tempête

La tempête est un phénomène naturel qui se caractérise par des perturbations marquées par des vents violents pouvant dépasser un seuil de 90 km/h par rafales. Elle peut s'accompagner de violentes précipitations et d'orages.

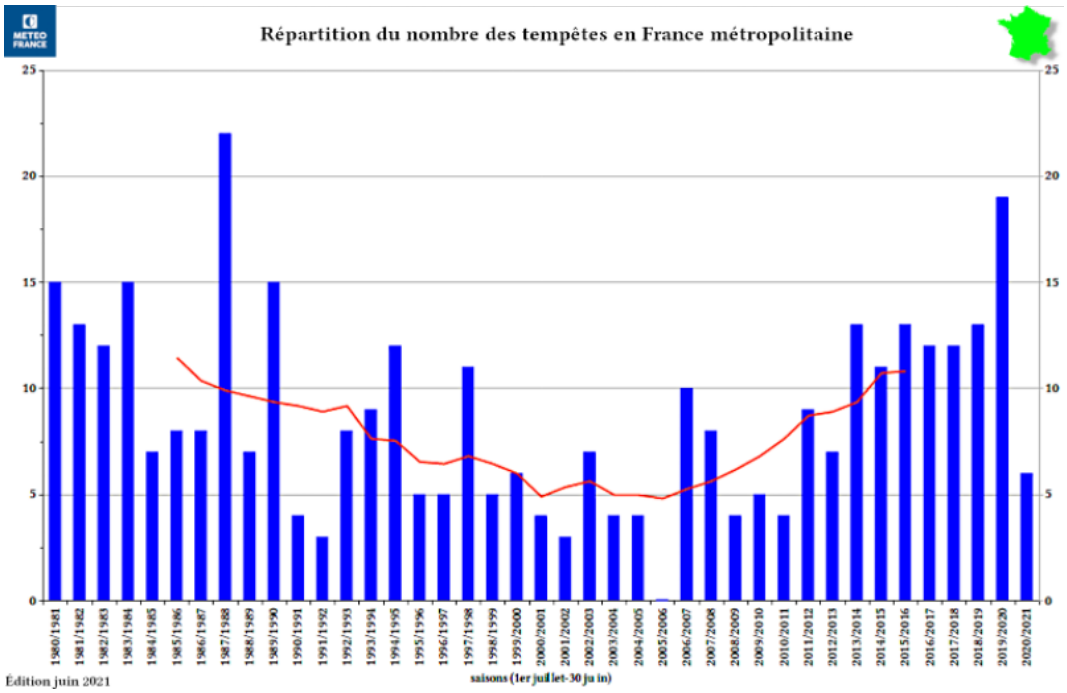


FIGURE 1.4 – Répartition du nombre de tempêtes en France métropolitaine entre 1981 et 2021 (METEO France)

Comme le montre la figure 1.3, les tempêtes représentent l'aléa climatique qui a engendré le plus de coûts sur l'horizon 1989-2019, notamment en raison des tempêtes Lothar et Martin en 1999 qui ont causé 92 décès et plus de 15 milliards d'euros de dommages. Pourtant, la figure ci-dessus (fig 1.5) issue de Meteo France sur la répartition du nombre de tempêtes en France métropolitaine montre qu'il y a eu peu de tempêtes recensées en 1999. Ainsi, lorsqu'on fait face à des évènements extrêmes tels que Lothar et Martin, il est difficile d'établir une corrélation entre le nombre de sinistres survenus et le coût total des sinistres sur une année.

1.2.2 Inondations

Le risque inondation est un des risques majeurs en assurance dommages. Ce dernier représente le 1er pôle d'indemnisation des catastrophes naturelles, avec un coût moyen

par année passé de 650 M€ à un peu plus d'1 Md€ entre 1990 et 2010 (voir fig 1.5). Cette hausse de la sinistralité est notamment due à quelques évènements majeurs dont les inondations du Gard en 2002 et du Rhône en 2003 ou encore les inondations liées à la tempête Xynthia en 2010.

En France, il existe 2 grands types d'inondations causées par des évènements météorologiques différents : les crues lentes de plaine, représentées par les inondations régionales, après le débordement des cours d'eau ordinaires, généralement après une période de précipitations continues, et les crues éclairs causés par de fortes et courtes précipitations.

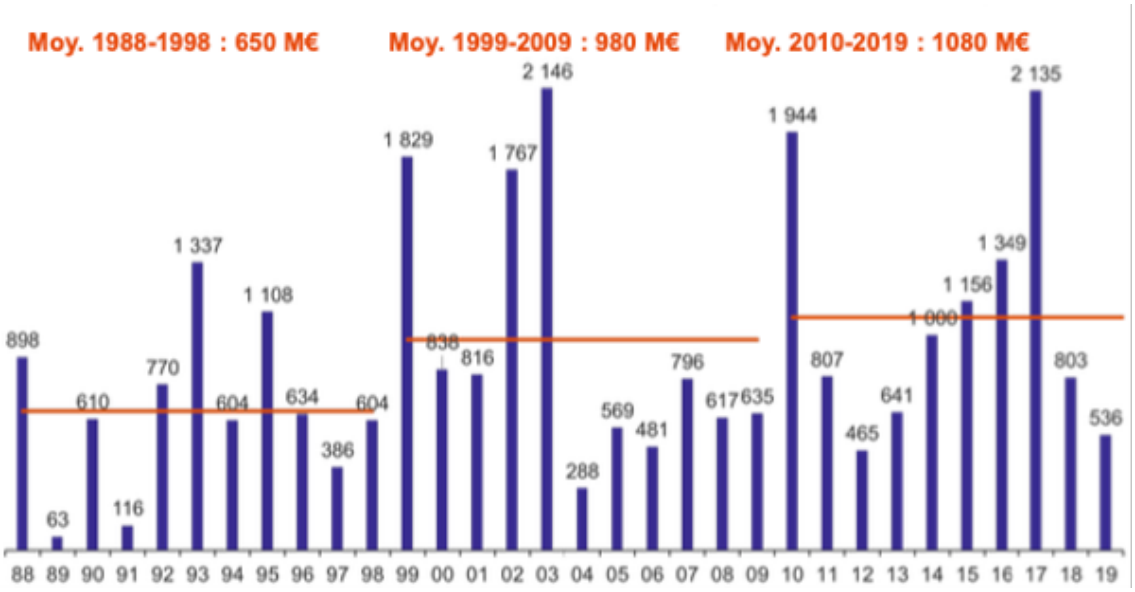


FIGURE 1.5 – Coûts moyens des sinistres inondation en 2020
[3]

1.2.3 Grêle

La grêle est un phénomène climatique complexe qui est de plus en plus observé et suivi à partir de divers paramètres atmosphériques. La grêle se produit généralement lors d'orages violents, dans ce qu'on appelle des cumulonimbus : il s'agit de nuages épais qui apparaissent lorsque des orages se produisent. On commence à parler de grêle dès lors que le diamètre des particules de glace est supérieur à 5 mm.

Les travaux autour de l'impact du changement climatique sur l'aléa grêle ont mis du temps à se lancer en partie à cause des méthodes d'identification des conditions de grêle qui sont difficiles à implémenter et à appliquer. Il s'agit d'un fléau qui touche énormément de secteurs d'activités, en particulier les agriculteurs en France. Ces derniers se couvrent généralement face à ce risque en souscrivant à des assurances classiques de dommages

aux biens. Concernant les habitations, tous les contrats habitation depuis 1992 incluent une Garantie Tempête/Grêle/Neige qui assure les dommages causés par la grêle.

La hausse de la sinistralité est notamment due à des épisodes orageux intenses de plus en plus fréquents. Les évènements orages/grêle du 20 au 23 mai 2022 ont causé près de 93 000 sinistres pour un coût avoisinant les 323 millions d'euros. Au seul cas de AXA, près de 4 000 sinistres ont été recensés sur cette période pour une charge autour de 41 M€.

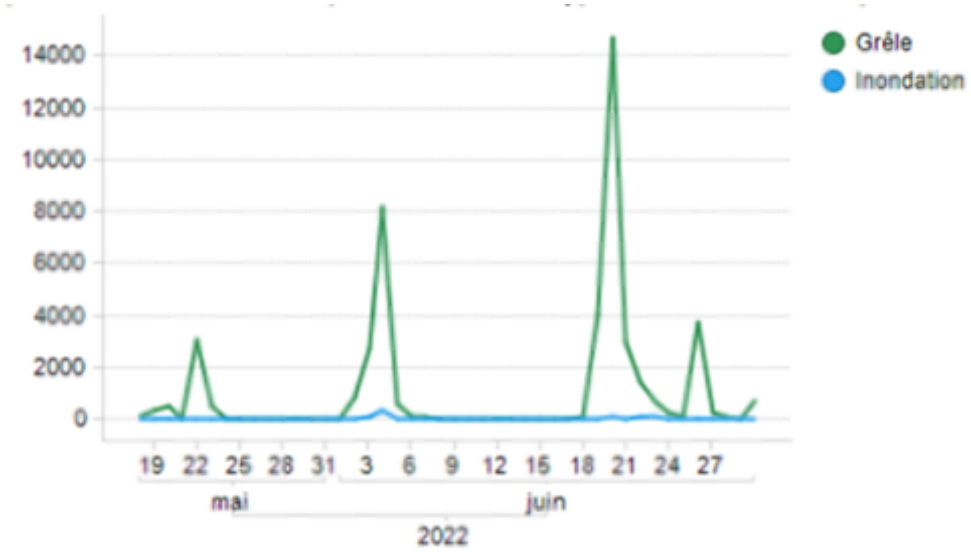


FIGURE 1.6 – Volume par date de survenance pour sinistre de type Cat-Nat et Climatique AXA France

1.2.4 Séisme

Les risques sismiques sont des risques aux conséquences dévastatrices non seulement pour la population mais également pour les compagnies d'assurance. Encore appelés tremblements de terre, les séismes se reflètent sur la surface par des vibrations dans le sol. Cela est causé par la fracturation de roches profondes due à l'accumulation massive d'énergie libérée par la création de failles au-delà d'un certain seuil.

En cas de tremblement de terre majeur, il est important que nous puissions estimer les dommages assurés qu'il peut occasionner à partir des propriétés physiques de celui-ci.

Le Japon est un pays entre-autres qui a été fortement touché par des séismes au cours des dernières décenies. En 2011, le pays avait notamment été frappé par un séisme qui a entraîné un tsunami et causé plus de 21 000 morts et des pertes assurées avoisinant les 35 Mds de dollars à lui seul. Cette tragédie a notamment entrainé au Japon une vague

d'émissions de nombreuses polices d'assurance couvrant le risque sismique.

Le risque de séisme représente donc un risque très important pour l'assureur mais également pour le réassureur qui porte souvent une grande partie du risque.

1.3 Coût d'un sinistre

Après avoir fait un court rappel sur l'activité de l'assurance non-vie en général ainsi que les différents risques climatiques sur lesquels portent cette étude, il est essentiel de bien définir et comprendre les différents enjeux et la problématique autour du calcul du coût d'un sinistre. Les leviers de maîtrise du coût d'un sinistre sur les évènements climatiques sont multiples. Lorsqu'on fait face à un Évènement climatique de Grande Ampleur ou médiatique, tel que l'épisode Cévenol du mois de septembre 2021, il est nécessaire pour toute compagnie d'assurance d'être à même d'estimer assez rapidement le nombre de sinistres qui seront ouverts ainsi que le coût que cela va engendrer. Ainsi, les projections de charge à l'ultime doivent être les plus précises possible afin de bien provisionner les évènements dans les comptes de l'assureur mais aussi pour maintenir une relation de confiance avec les réassureurs. Elles sont également utiles pour bien dimensionner la gestion opérationnelle des évènements climatiques.

Dans les sections qui suivent, nous commencerons par définir le cycle de vie d'un sinistre avant de nous pencher sur les enjeux autour de l'estimation de la charge à l'ultime, enjeux liés au provisionnement mais également à la réassurance.

1.3.1 Cycle de vie d'un sinistre

Comprendre les différentes étapes de la vie d'un sinistre est essentiel pour bien comprendre les enjeux autour du provisionnement et plus précisément de l'estimation de la charge à l'ultime.

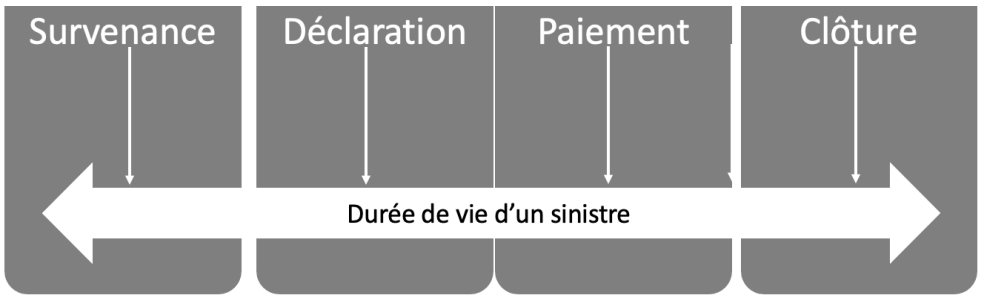


FIGURE 1.7 – Durée de vie d'un sinistre

La figure 1.7 nous donne un aperçu des différentes étapes du cycle de vie d'un sinistre. En effet, tout commence par la survenance du sinistre. Lorsque c'est le cas, les

gestionnaires déterminent un coût estimatif des sinistres survenus.

Cependant, la seconde étape qui consiste en la déclaration des sinistres n'est pas toujours réalisée dans les temps, notamment sur des branches longues comme en RC auto.

Et la 3ème étape consiste au paiement de tous les sinistres. Ainsi, la réglementation exige un provisionnement pour tous les sinistres survenus, même ceux qui n'ont pas encore été déclarés afin d'éviter des risques de sous-provisionnement à la clôture.

Un dernier élément à mentionner est la possibilité de rouvrir des sinistres précédemment fermés. Néanmoins cela arrive très rarement dans le cadre des risques climatiques car généralement l'assureur prend le temps nécessaire avant de clôturer un sinistre.

Les différentes étapes de déroulement d'un sinistre nous montrent donc à quel point bien provisionner est essentiel pour la bonne gestion du sinistre. On va donc parler plus en détail de quelques provisions.

1.3.2 Provisionnement

La charge à l'ultime représente le coût final rattaché à des sinistres. Cette charge à l'ultime est utilisée notamment à des fins de provisionnement. Le provisionnement est nécessaire dans toute compagnie d'assurance afin d'avoir une bonne vision des engagements de l'assureur vis-à vis de ses assurés à la clôture. Les provisions correspondent à la part des primes mise de côté par l'assureur pour faire face aux sinistres survenus ou à venir. Celles-ci se retrouvent donc au passif du bilan d'une compagnie d'assurance et se conforment aux normes comptables et prudentielles françaises, qui accordent une véritable importance à la transparence et l'amélioration des pratiques de provisionnement. Au-delà des raisons qui viennent d'être citées, le provisionnement est également très important pour des raisons fiscales car il a un impact sur le résultat, et donc un impact sur les impôts. Lorsque l'assureur augmente ses provisions, on parle de dotation et lorsqu'il les diminue, on parle de reprise. Il existe plusieurs types de provisions dont on peut citer :

- les provisions de primes : il s'agit principalement des Provisions pour Primes Non Acquises (PPNA) en droit français
- les provisions pour sinistres : il s'agit principalement des Provisions pour Sinistres à Payer (PSAP)
- les provisions mathématiques

Les définitions plus détaillées des provisions qui suivent se basent sur les définitions initiales tirées de l'article R331-6 du code des assurances.

Provisions pour Primes non Acquises PPNA

La Provision pour Primes Non Acquises représente une provision destinée à constater, pour l'ensemble des contrats en cours, la part des primes émises et celles restant à émettre se rapportant aux prochains exercices. Elle est calculée par défaut en utilisant la méthode prorata temporis, c'est-à-dire en se basant sur l'hypothèse homogène sur l'année. Entre 2 années consécutives N-1 et N, la variation de PPNA se calcule comme suit :

$$\delta PPNA = PPNA_N - PPNA_{N-1} = Prime\ emise_N - Prime\ acquise_{N-1}$$

Provisions pour Risques en Cours

La Provision pour Risques en Cours représente une provision destinée à constater, pour l'ensemble des contrats en cours, la charge des sinistres et des frais afférents aux contrats et qui n'est couverte ni par les primes, ni par la PPNA. Ainsi, lorsqu'un produit est sous-tarifé, par exemple si le ratio de sinistralité $\frac{S}{P}$ ultime des générations précédentes est supérieure à 1 depuis 2 ans, l'assureur est obligé de rajouter une provision qui constitue la PREC.

$$PREC_N = PPNA_N \cdot (max\frac{S}{P} - 1)$$

Provisions pour Sinistres à Payer

La Provision pour Sinistres à Payer représente la provision pour des sinistres survenus avant la clôture mais qui n'ont pas encore été réglés. L'article R331-6 définit exactement la PSAP comme étant « la valeur estimative des dépenses en principal et en frais, tant internes qu'externes, nécessaires au règlement de tous les sinistres survenus et non payés, y compris les capitaux constitutifs des rentes non encore mises à la charge de l'entreprise ». Elle est très importante car elle représente une grande partie des provisions techniques en assurance non-vie et est toujours calculée à la fin de chaque exercice. Celle-ci est constituée de plusieurs éléments.

Provisions Dossier/Dossier

Lorsque des sinistres se produisent, ils sont directement pris en charge par des gestionnaires qui font une première déclaration de provision (généralement sur une base forfaitaire pour les risques de masse). Cette première déclaration constitue la provision dossier/dossier et son calcul se base sur plusieurs critères tels que les coûts précédents des sinistres considérés, l'évolution de l'inflation, le niveau de responsabilité de l'assuré etc. Au cours de la vie du sinistre, les gestionnaires peuvent être amenés à réévaluer cette provision à la hausse ou à la baisse selon la nature et les caractéristiques propres au dossier. Elle décroît en même temps que les sinistres sont payés, jusqu'à la clôture où

elle s'annule.

Provisions IBNER

Les provisions IBNER ou Incurred But Not Enough Reported sont des provisions destinées à couvrir le manque potentiel de provisions en cas de sinistres survenus et déclarés à la date de clôture des états financiers. Ainsi, ces provisions servent à palier la possibilité que le montant des sinistres soit plus élevé que prévu à la clôture.

Provisions IBNYR

Les provisions IBNYR ou Incurred But Not Yet Reported sont des provisions destinées à couvrir le coût des sinistres survenus mais non encore déclarés à la date de clôture des états financiers. Ces provisions sont généralement calculées en s'appuyant sur des méthodes de fréquence/sévérité.

Les provisions IBNER et les provisions IBNYR constituent à elles 2 les provisions IBNR ou **tardifs**.

1.3.3 Décomposition de la charge à l'ultime

La charge ultime représente la charge finale prévisible que va coûter un sinistre et qui est évaluée régulièrement à l'aide de méthodes statistiques. On peut décomposer cette charge à partir des différents éléments qui constituent la PSAP ainsi que des sinistres déjà payés.

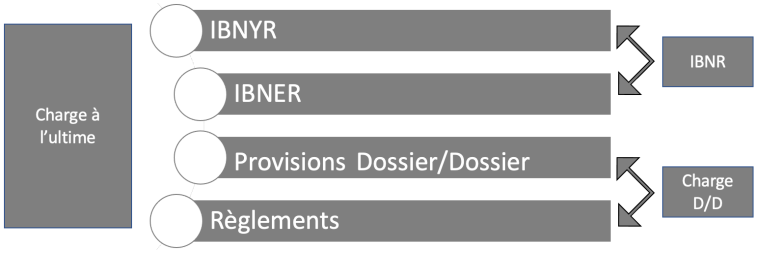


FIGURE 1.8 – Décomposition de la charge à l'ultime

En cas de sinistre, la charge à l'ultime, représente donc le coût final estimé du sinistre. Il est calculé en sommant le montant qui a déjà été réglé pour ce sinistre, les provisions dossier/dossier ainsi que les tardifs ou IBNR qui sont estimés par les actuaires. A la

clôture, la charge à l'ultime équivaut au montant total payé par l'assureur.

Pour l'estimation de cette charge à l'ultime, il existe 2 grandes méthodes :

- les méthodes déterministes (principalement Chain Ladder)
- les méthodes stochastiques

Nous nous limiterons dans le cadre de ce mémoire à présenter la méthode de Chain Ladder.

Méthode de Chain Ladder

La méthode de Chain Ladder fait partie des méthodes agrégées basées sur des triangles de liquidation ou **run-off** . Elle est la méthode la plus connue et la plus répandue. Commençons par une brève description du triangle.

Triangles de liquidation

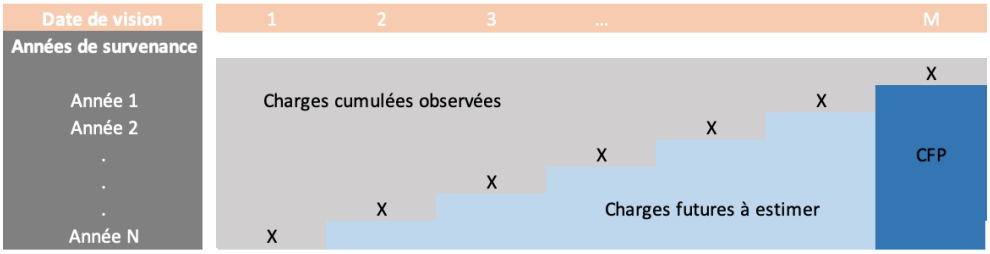


FIGURE 1.9 – Triangle de développement

La partie supérieure du triangle représente ces paiements cumulés. On notera que très souvent, initialement ces paiements sont en décumulés. Ces triangles peuvent se faire sur une année, un mois ... Au delà des paiements, on peut également faire des triangles sur le nombre de sinistres.

Les lignes représentent les exercices de survenance des sinistres tandis que les colonnes représentent les périodes de développement. La diagonale caractérisée par les X représente les paiements pour les sinistres survenus en année N et réglés la même année.

La méthode de Chain Ladder s'applique donc à des triangles de paiements cumulés. Soit le triangle de paiements cumulés $C_{i,j}$ ci-dessous :

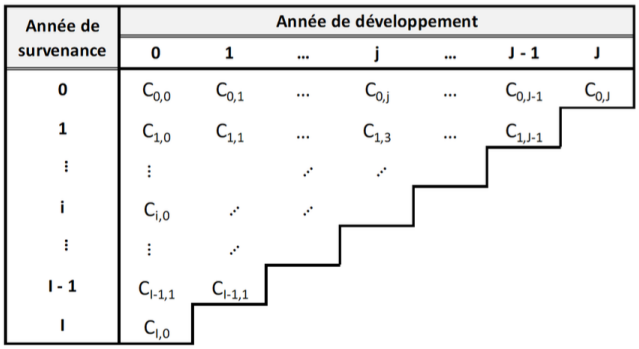


FIGURE 1.10 – Triangle de paiements cumulés

Hypothèses de Chain Ladder

L'hypothèse principale repose sur l'indépendance des ratios $f_{ij} = \frac{C_{i,j+1}}{C_{i,j}}$ de l'origine i, c'est-à-dire

$$\forall j = 0, ..., J-1 : \frac{C_{0,j+1}}{C_{0,j}} = \frac{C_{1,j+1}}{C_{1,j}} = ... = \frac{C_{Ij1,j+1}}{C_{Ij1,j}}$$

Dès lors, on définit les facteurs de développement de la manière suivante :

$$\forall j = 0, ..., J-1 : \hat{f}_j = \frac{\sum_{i=0}^{I-j-1} C_{i,j+1}}{\sum_{i=0}^{I-j-1} C_{i,j}}$$

Ces derniers permettent de compléter la partie inférieure du triangle et de calculer montant des réserves :

$$\hat{C}_{i,j+1} = \hat{f}_j * \hat{C}_{i,j}$$

et

$$\hat{R} = \sum_{i=0}^{I} \hat{R}_i = \sum_{i=0}^{I} \hat{C}_{i,J} - C_{i,J-i}$$

où R_i représente la provision au titre de l'année de survenance i et R la provision totale.

1.3.4 Calcul de la charge Dossier/Dossier

Une provision est allouée à l'ouverture de chaque sinistre. Cette provision est forfaitaire et dépend de 2 critères :

- du type de contrat (MRH, MRP, Auto 4 Roues...)
- des garanties ouvertes (Tempête, Inondation, CATNAT...)

Ce forfait est déterminé par rapport au forfait de l'année précédente sur la base d'un coût moyen forfaitaire prévisionnel (CMFP) auquel on applique un taux d'inflation en fonction de l'inflation qu'on observe sur le marché selon la branche.

Pour des sinistres ne dépassant pas un certain seuil, on applique directement cette formule :

$$CMFP_{N+1} = CMFP_N.Inflation$$

Au delà d'un certain seuil, on applique un coefficient de vieillissement :

$$Coef\_vieillissement : Coef(N-1/N) = \frac{CM(N-1/N)}{CM(N-1/N-1)}$$

$$CMFP_{N+1} = MAX(CMFP_N, CM_{S=n}^{vieilli}.Inflation)$$

avec *CM (N / N) : le coût moyen de l'exercice N vu en année N* .

On calcule ainsi le maximum entre le CMFP de l'année précédente et le coût moyen vieilli.

Une bonne prise en compte des coefficients d'inflation en fonction des branches est donc importante. Par exemple, si c'est de l'auto, on aura tendance à regarder l'inflation des réparations, des pièces etc, si c'est du non auto cela pourrait être le prix du bâtiment, des matières premières. On appliquera ensuite ces inflations selon les critères définis ci-dessus.

HISTORIQUE DES INDEXATIONS		2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	Proposition
Automobile	Corporel	1,07	1,08	1,08	1,08	1,08	1,08	1,08	1,08	1,08	1,02	1,02	1,02	1,02	1,02	1,020
	RC Matériels PP	1,02	1,02	1,02	1,02	1,06	1,04	1,03	1,04	1,05	1,06	1,04	1,05	1,07	1,086	1,075
	EN	1,02	1,02	1,02	1,02	1,06	1,04	1,03	1,04	1,05	1,06	1,04	1,05	1,07	1,097	1,075

TABLE 1.2 – Exemple d'inflation appliquée sur la branche auto, données fictives

Pour vérifier la pertinence de cette méthode et s'assurer qu'elle ne soit pas obsolète au fil des années, on regarde les sinistres clos des dernières années et combien ils ont évolué sur une durée de 3 ans. On applique donc du Chain Ladder uniquement pour vérifier que notre hypothèse est toujours valable et garder une stabilité dans le temps.

FAMUV	10	INOND		10	NATUR		10	TEMP		26	TEMP	
Nombre de dossiers												
	N	N+1	N+2	N	N+1	N+2	N	N+1	N+2	N	N+1	N+2
2 019	620	2 019	2 032	4 902	8 953	9 013	11 473	22 774	22 927	1 579	3 495	3 567
2 020	1 416	1 669		5 971	7 139		17 351	20 776		2 989	3 769	
2 021	2 576			6 146	-		10 845			1 913	-	
Taux de cloture par année de survenance												
	N	N+1	N+2	N	N+1	N+2	N	N+1	N+2	N	N+1	N+2
2019	37%	82%	94%	65%	90%	96%	57%	85%	95%	47%	81%	93%
2020	52%	88%		67%	92%		62%	90%		49%	85%	
2021	50%			48%			43%			33%		
Taux de sans suite par année de survenance												
	N	N+1	N+2	N	N+1	N+2	N	N+1	N+2	N	N+1	N+2
2019	19%	44%	47%	31%	45%	45%	15%	22%	22%	18%	29%	30%
2020	28%	38%		40%	45%		17%	21%		18%	30%	
2021	31%			25%			20%			20%		
CM sur les clos												
	N	N+1	N+2	N	N+1	N+2	N	N+1	N+2	N	N+1	N+2
2019	1 543 €	2 165 €	2 723 €	1 216 €	1 757 €	2 060 €	1 628 €	2 144 €	2 456 €	1 510 €	2 089 €	2 500 €
2020	2 024 €	2 897 €	3 644 €	1 012 €	1 617 €	1 896 €	1 693 €	2 249 €	2 576 €	1 691 €	2 164 €	2 590 €
2021	1 325 €	1 897 €	2 387 €	1 355 €	2 166 €	2 540 €	1 397 €	1 856 €	2 125 €	1 315 €	1 683 €	2 014 €
CMFP 2020	5 857 €			3 075 €			4 485 €			3 320 €		
CMFP 2021	5 974 €	2,50		3 260 €	1,28		4 754 €	2,24		3 519 €	1,75	

TABLE 1.3 – Exemple d'application de Chain Ladder pour vérification des hypothèses, données fictives

La figure ci-dessus représente des triangles de liquidation sur des familles de garanties climatiques (famuv) pour des nombres de dossiers, des taux de clôture et des coûts moyens sur les sinistres clos. En regardant le triangle des coûts moyens sur la colonne inond par exemple, des sinistres en 2019 étaient ouverts à 1286 euros, en 2021 ils ouvrent à 1325€, on applique la méthode de Chain Ladder avec les triangles de développement qui nous permet d'obtenir un montant prévisible de 2387€ et on vérifie que ce montant est toujours cohérent avec le forfait qu'on ouvre pour ces sinistres.

C'est donc cette charge Dossier/Dossier qui est renseignée par le gestionnaire de sinistre dès qu'un sinistre attritionnel est ouvert.

Dès qu'un évènement se produit, on a ainsi une information directe sur le nombre de sinistres ouverts ainsi que la charge Dossier/Dossier associée.

1.3.5 Réassurance

La réassurance, communément appelée assurance de l'assureur, se caractérise par un contrat suivant lequel moyennant une prime, l'assureur transfère une partie du risque dont il a la charge au réassureur. Contrairement à la coassurance, elle permet un partage du risque sans avoir à diviser le contrat de l'assuré.

Au-delà du transfert du risque et de la protection contre de potentielles pertes financières, l'assureur peut faire appel au réassureur pour de l'aide à la tarification, pour augmenter sa capacité de souscription ou encore pour avoir des résultats techniques plus stables dans le temps. L'assureur également appelé la cédante cède le risque au réassureur ou cessionnaire qui le prend suivant ce qu'on appelle un traité.

Les risques cédés sont par exemple des risques évènementiels tels que les séismes ou tempêtes, les risques de surfréquence en assurance santé par exemple...

La réassurance sert donc de levier financier et technique à l'assurance.

Il existe 2 types de réassurance : la réassurance proportionnelle et la réassurance non-proportionnelle.

Réassurance proportionnelle

Comme son nom l'indique, on parle de réassurance proportionnelle lorsque le réassureur gère une proportion du risque de l'assureur. On distingue dans ce cas 2 types de traités : le traité en quote-part et le traité en excédent de plein.

Traité en quote-part

Le traité en quote-part se caractérise par un partage du sort entre le réassureur et l'assureur. L'opération s'effectue de sorte à ce que le cessionnaire rembourse à la cédante une partie du risque correspondant au prorata de la prime qu'il a touchée sur la police. Il s'opère donc un échange de commissions entre l'assureur qui paie les commissions et le réassureur qui paie des rétro-commissions afin de couvrir les charges de gestion des sinistres.

Traité en excédent de plein

Contrairement au traité en quote-part, il n'y a pas de partage du sort intégral dans le traité en excédent de plein.

Réassurance non proportionnelle

On parle de réassurance non proportionnelle lorsque le réassureur accepte en contrepartie d'une portion de prime de prendre en charge un montant ne dépendant pas uniquement de la survenance ou non d'un sinistre mais également de deux autres paramètres que sont la priorité et la portée. La réassurance non proportionnelle peut s'appliquer sur la globalité du portefeuille ou par sinistre (exemple avec les catastrophes naturelles). On distingue dans ce cas 2 types de traités : le traité en excédent de sinistres et le stop loss.

Traité en excédent de sinistres (XS)

Dans un traité en excédent de sinistres, le réassureur s'engage à couvrir le risque au delà d'un certain seuil ou niveau de coût de sinistre.

Stop Loss

Dans un traité en Stop Loss, le réassureur s'engage à prendre en charge qu'une partie du rapport « sinistres à prime de l'année » dépassant un certain seuil. En d'autres termes, le réassureur n'intervient dans ce traité que lorsque l'assureur est en perte. Ce type de traité est très utilisé dans le cadre des catastrophes naturelles en raison de leur caractère

cyclique ou périodique.

Réassurance facultative

Dans la réassurance facultative, on y retrouve aussi bien de la réassurance proportionnelle que de la réassurance non proportionnelle. La particularité ici est que le réassureur après analyse a la liberté de l'accepter ou au contraire de refuser la couverture d'un risque que lui propose l'assureur.

Territorialité	Périls couverts	Couverture	Structure (en M€)	Capacité (en M€)
Monde entier hors France métropolitaine		par évènement	1000 XS 20	1020
	Grêle, orages, tornades	par évènement	980 XS 40	1020
France métropolitaine	Tempêtes "hivernales"	par évènement	2100 XS 200	2300
	Autres évènements naturels	par évènement	700 XS 50	750
Monde entier	Evènements non naturels hors attentat/terrorrisme	par évènement	600 XS 20	620

TABLE 1.4 – Exemple de traités de réassurance appliqués pour des évènements naturels et non naturels, données fictives

Réassurance et CAT NAT

En France, il existe une assurance pour les catastrophes naturelles dans laquelle intervient fortement la réassurance. En effet, lorsqu'on dénombre un volume de sinistres élevé à la suite d'un évènement climatique, les représentants des municipalités peuvent déclarer l'état de catastrophes naturelles pour leurs communes : on parle d'arrêtés CAT NAT. C'est dans ce cadre qu'intervient la Caisse Centrale de Réassurance qui réassure une grande partie de ces risques. Cela permet notamment une indemnisation plus rapide des sinistrés par les assureurs. Une base de données publique répertoriant tous les arrêtés CAT NAT en France peut alors être consultée.

La charge à l'ultime calculée dans le cadre ce mémoire est brute de réassurance.

1.4 Enjeux et objectifs du mémoire

Notre première partie nous a permis de mettre en relief notre environnement d'étude à savoir les risques climatiques dans le cadre de l'assurance non-vie et de ses sous branches. Elle a surtout permis d'introduire la notion de charge à l'ultime dont le calcul représente un enjeu majeur dans toutes les compagnies d'assurance aussi bien pour des raisons de provisionnement, réassurance et opérationnelles. Ainsi, avoir une bonne vision de la charge à l'ultime est nécessaire afin de mobiliser sur le terrain le bon nombre d'acteurs en fonction de la sévérité des évènements climatiques. Ces prévisions sont également très

attendues par France Assureurs qui représente les intérêts du secteur des assurances de l'autorité publique, et auprès des réassureurs telle que la Caisse Centrale de Réassurance (CCR) qui assure la réassurance et fournit les garanties de l'État contre les catastrophes naturelles.

Face à cet enjeu majeur, nous allons challenger la méthode d'estimation de la charge à l'ultime pour les Évènements climatiques de Grande Ampleur actuellement utilisée au sein d'AXA France. En effet, cette méthode se base uniquement sur la cadence d'ouvertures des sinistres par type d'évènement climatique sans prise en compte du critère géographique. Avec une base de données d'évènements assez restreinte, il sera donc question dans un premier temps d'identifier un modèle d'apprentissage automatique capable de détecter de nouveaux évènements. Dans un second temps, en rajoutant les variables géographiques et avec une base de données d'évènements enrichie, il sera question d'identifier un nouveau modèle permettant de tirer parti de la quantité et de la finesse des données dont il dispose pour estimer plus précisément la charge à l'ultime.

La première partie de la suite de ce mémoire consistera donc en une définition de ce qu'on appelle Évènement Climatique de Grande Ampleur (EGA) ainsi que la méthode d'estimation de la charge à l'ultime actuellement utilisée notamment au sein d'AXA France pour ces évènements.

La seconde partie sera plus centrée sur le processus de construction et d'enrichissement de la base de données d'Évènements de Grande Ampleur (EGA).

Il s'agira ensuite de tester des modèles de machine learning permettant de prédire au mieux cette charge à l'ultime.

Nous terminerons en comparant notre modèle avec le modèle actuellement utilisé au sein d'AXA.

Chapitre 2

Identification d'un Évènement Climatique de Grande Ampleur

2.1 Définition d'un EGA

Un Évènement climatique de Grande Ampleur (EGA) peut se caractériser par un ou plusieurs sinistres extrêmes.

Même parmi les scientifiques, le concept d'évènements météorologiques extrêmes reste difficile à définir. Les statisticiens, les physiciens et les spécialistes des sciences sociales ont leurs propres définitions des évènements météorologiques extrêmes. Bien que ces trois définitions soient complémentaires, elles ont toutes leur propre dimension.

Pour les statisticiens, une valeur extrême est ainsi nommée si la valeur mesurée (température, vitesse du vent) dépasse ce que l'on rencontre normalement. Les nombres déterminent si les évènements sont extrêmes.

Les physiciens en donnent une deuxième définition : les extrêmes correspondent à une classe d'évènements (cyclones tropicaux, tempêtes extratropicales, canicules, sécheresses, etc.), selon la région et sa description phénoménologique.

Enfin, les spécialistes des sciences sociales définissent les évènements par les dommages causés. En ce sens, lorsqu'un évènement affecte la société, on peut dire qu'il est extrême.

On retiendra tout simplement qu'un EGA est caractérisé par une probabilité de survenance relativement faible et un coût élevé.

Au sein d'AXA France, les EGA sont définis en utilisant uniquement la typologie de

l'évènement et les dates de survenance. Ils correspondent aux évènements dont **le coût ou la charge pour Axa (règlements + provisions) est supérieur à un certain seuil actuellement égal à 1M€** (3M€ avant 2021). Tous ces EGA sont recensés dans une base qui contient actuellement 176 évènements allant de 2009 à 2022.

debut	fin	TYPE_EVT	evt_new
2009-01-23	2009-01-25	TEMP.	1
2009-02-09	2009-02-11	TEMP.	2
2009-05-11	2009-05-13	GRELE	3
2009-05-20	2009-05-22	GRELE	4

TABLE 2.1 – Aperçu de la base d'EGA au sein d'AXA France

On parlera de sinistres graves dès lors que la charge d'un sinistre dépasse 150.000€. Les sinistres dont le coût pour Axa est inférieur à 150.000€ sont appelés sinistres attritionnels.

Ces dernières années, plusieurs compagnies d'assurance se sont vu nouer des partenariats avec des structures météorologiques tel que Météo France qui envoie des alertes avec un indice de gravité par commune lorsqu'un évènement pourrait se produire, ce qui permet aux compagnies de détecter et d'anticiper des évènements climatiques. Néanmoins, une alerte ne signifie pas forcément qu'on va avoir beaucoup de dommages et il peut arriver que des Évènements de Grande Ampleur aient lieu sans qu'on ait eu d'alertes au préalable (*Ex : Orages en Corse en Août 2022*).

2.2 Gestion opérationnelle des évènements climatiques

Lorsqu'un Évènement climatique de Grande Ampleur se produit, il est essentiel de mettre en place un bon dispositif de gestion de crise, c'est-à-dire avoir une idée du nombre de personnes à déployer sur le terrain. En effet, en fonction de l'ampleur d'un évènement, des cellules de crise plus ou moins importantes peuvent être créées. Un système de communication adéquat doit être mis en place et cela passe par des grilles de décision.

Grille de décision

Exemples de questions :
Est-ce qu'il y a une alerte rouge météo pour un département ?
Est-ce qu'il y a eu des évènements récemment dans la / les régions visées ?

Est-ce une région avec un contexte environnemental à risque ?

En fonction des réponses, on peut songer à plusieurs scénarios visant à déployer des mesures non seulement auprès du client mais également en interne. Il existe également un système d'alerte qui se met en place lorsqu'il y a un évènement.

Mise en place d'un système d'alertes

Lorsqu'un évènement climatique survient, les compagnies d'assurance reçoivent des alertes soit via des partenariats réalisés avec des sociétés météorologiques, soit via des logiciels en interne. Suite à ces alertes, elles reçoivent des informations sur les régions les plus touchées et les caractéristiques propres à l'évènement.

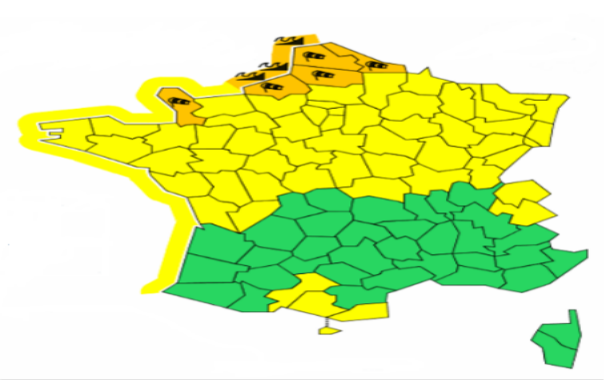


FIGURE 2.1 – Exemple d'informations reçues : zones les plus impactées par la tempête EUNICE (Source WikiPredict/AXA France)

Ainsi les compagnies peuvent dimensionner proprement et efficacement le niveau de leur intervention eu égard aux caractéristiques de l'évènement. Il est donc essentiel d'avoir un bon dispositif de gestion de crise encore plus dans le cadre d'Évènements de Grande Ampleur.

2.3 Calcul de la charge à l'ultime

Après avoir défini ce qu'était un Évènement Climatique de Grande Ampleur ainsi que le dispositif de gestion de crise mis en place lorsqu'un évènement se produit, on peut maintenant se pencher sur la méthode d'estimation de la charge à l'ultime.

2.3.1 Construction des cadences d'ouverture

La première étape du calcul de la charge à l'ultime consiste en la construction des

cadences d'ouverture de sinistres. En effet, une approche pour estimer le volume final de sinistres qui va être engendré par un évènement climatique est d'observer l'évolution du nombre de sinistres ouverts quelques jours après l'évènement. En pratique, on parvient à avoir une bonne idée des volumes sinistres que vont générer un évènement climatique trois jours ouvrés après la fin de l'évènement pour les tempêtes ou lors de chutes de grêle. Cela requiert d'avoir des évènements passés de même nature pour pouvoir projeter des volumes de sinistres. Pour estimer le coût de l'évènement, il suffira alors d'associer un coût moyen à chaque sinistre.

Prenons l'exemple de la tempête Eunice survenue du 17 février au 19 février 2022 :

		Nombre d'ouvertures	Retrait WE + JF	Cadences (cumulées)
vendredi	18/02/2022	390	390	390
samedi	19/02/2022	1 395		
dimanche	20/02/2022	1 045		
lundi	21/02/2022	925	3 365	3 754
mardi	22/02/2022	729	729	4 483

TABLE 2.2 – Exemple d'application de cadences sur une UP tempête

On voit qu'à l'ouverture on a eu 390 sinistres ouverts le vendredi. Il n'y a pas d'ouvertures de sinistres le week-end ni les jours fériés et il faut donc comptabiliser les ouvertures de sinistres ayant eu lieu à ces moments selon une méthode à définir : soit en les comptabilisant le vendredi précédant ou le lundi suivant le week-end ou en les séparant entre le vendredi et le lundi. Dans notre exemple, ils ont été comptabilisé le lundi suivant le week-end. L'étape suivante consiste enfin à cumuler les nombres d'ouverture afin de déterminer des cadences d'ouverture en cumulé. On passe ainsi de 390 sinistres enregistrés au 18/02/2022 à 4483 sinistres au 22/02/2022.

2.3.2 Choix de l'évènement le plus proche

Le but de cette étape est de déterminer l'évènement qui se rapproche le plus de celui que l'on cherche à estimer. Pour ce faire, lorsqu'un évènement survient, on commence par regarder l'historique d'évènements du même type (tempête, grêle...). Sur cet historique, on calcule la somme de la différence des carrés des matrices de cadencement du nombre d'ouvertures.

Cela revient à calculer la distance entre le cadencement de chaque évènement et celui à estimer :

$$Distance = \sum_{\substack{1 \leq i \leq n \\ 1 \leq j \leq n}} \left(x_{i,j}^2 - y_{i,j}^2\right)$$

avec *y la matrice de cadencement du nombre d'ouvertures pour l'évènement que l'on*

cherche à estimer et x la matrice de cadencement du nombre d'ouvertures d'un évènement de même type.

On procède ainsi à une sélection de l'évènement qui minimise la distance : il s'agira donc pour nous de l'évènement le «plus proche» en cadencement d'ouverture.

2.3.3 Estimation de la charge à l'ultime

De manière générale, la méthode la plus utilisée et la plus répandue pour l'estimation de la charge à l'ultime est la méthode de Chain Ladder.

Cependant, dans le cadre de ce mémoire et de l'estimation de la charge pour les Évènements climatiques de Grande Ampleur, cette méthode n'est pas utilisée car le raisonnement par année de survenance n'a aucun sens en raison de la très grande volatilité des évènements d'une année à une autre. Ainsi, ça ne satisfait pas l'hypothèse d'indépendance de Chain Ladder énoncée précédemment. D'autres méthodes de calculs sont alors envisagées.

Pour rappel, la charge à l'ultime représente le coût final prévisible des sinistres. Dès qu'on a un évènement, on doit être capable de donner une estimation à chaud de la charge à l'ultime. La première étape consiste donc à déterminer l'évènement le plus proche en terme de cadence d'ouverture.

Une fois qu'on a déterminé l'évènement le plus proche, on récupère le pourcentage du volume ultime atteint à la même vision de l'évènement le plus proche. Il s'agit du rapport entre le montant de volume à la même période et le volume ultime.

On applique ce pourcentage à l'évènement que l'on cherche à estimer :

$$Volume\ ultime = \frac{Volume\ sinistres}{\%\ du\ volume\ ultime\ de\ l'evenement\ le\ plus\ proche\ a\ la\ meme\ periode}$$

Enfin, une fois qu'on a estimé les volumes, on utilise le coût moyen qu'on retrouve sur une fourchette d'évènements passés plus récents et de même nature pour déduire la charge à l'ultime selon la formule suivante :

$$Charge\ ultime = Volume\ ultime * Cout\ moyen$$

Mise en "as if" des coûts moyens

L'utilisation de coûts moyens pour le calcul de la charge à l'ultime suppose qu'on ait des évènements réellement comparables. En effet, un évènement peut être réglé sur une période longue pouvant aller jusqu'à plusieurs années. Utiliser des coûts moyens passés

pour estimer la charge suppose alors que ces coûts moyens au cours des années calendaires soient comparables entre eux. Cependant, ce n'est pas toujours le cas en raison de facteurs impactants telles que l'inflation ou l'évolution de l'indice des prix. En guise d'alternative, il faut donc élaborer ce qu'on appelle une statistique "as-if", autrement dit comparable à une année de référence.

L'idée autour de cette statistique est qu'un évènement qui a coûté 2M€ en 2000 n'est pas comparable à un évènement qui a coûté 2M€ en 2020. Plusieurs facteurs tels qu'énoncés précédemment influent sur ces coûts là. Une solution est donc d'indexer les évènements par rapport à un indice de référence le taux d'inflation pour une branche donnée :

$CM_k^n = CM_k * \frac{I_n}{I_k}$

avec

- CM_k^n : coût moyen "as if" pour un évènement passé survenu en année k et vu en année n
- CM_k : coût moyen de l'évènement en année k
- I_n : Indice l'année de référence
- I_k : Indice l'année de survenance k

2.3.4 Exemple d'application à la tempête EUNICE

La tempête Eunice survenue du 17/02 au 19/02/22 génère de violentes rafales de vent sur l'extrême nord du pays, de 120 à 130 km/h dans les terres du Pas-de-Calais et zones limitrophes, 100 à 110 km/h en allant vers le département de la Manche et le sud du Nord ainsi qu'en Seine-Maritime, 90/100 km/h ailleurs. Sur la côte, des pointes à 140/150 km/h sont envisagées, notamment sur les caps exposées.

Projection des volumes de sinistres à l'ultime

L'étape suivante et par ailleurs la plus importante correspond à la projection à l'ultime des chiffres de sinistralité précédemment obtenus.

En effet, à partir de ces chiffres, on compare les cadences d'ouverture des sinistres précédents de même type et on garde celles qui se rapprochent le plus de la tempête Eunice (cf. Section 3.3). On limite notre seuil de distance à 400 ouvertures de sinistres et on détermine les tempêtes "les plus proches" de la tempête Eunice. Il s'agit ici des tempêtes Egon, Ciara et Aurore.

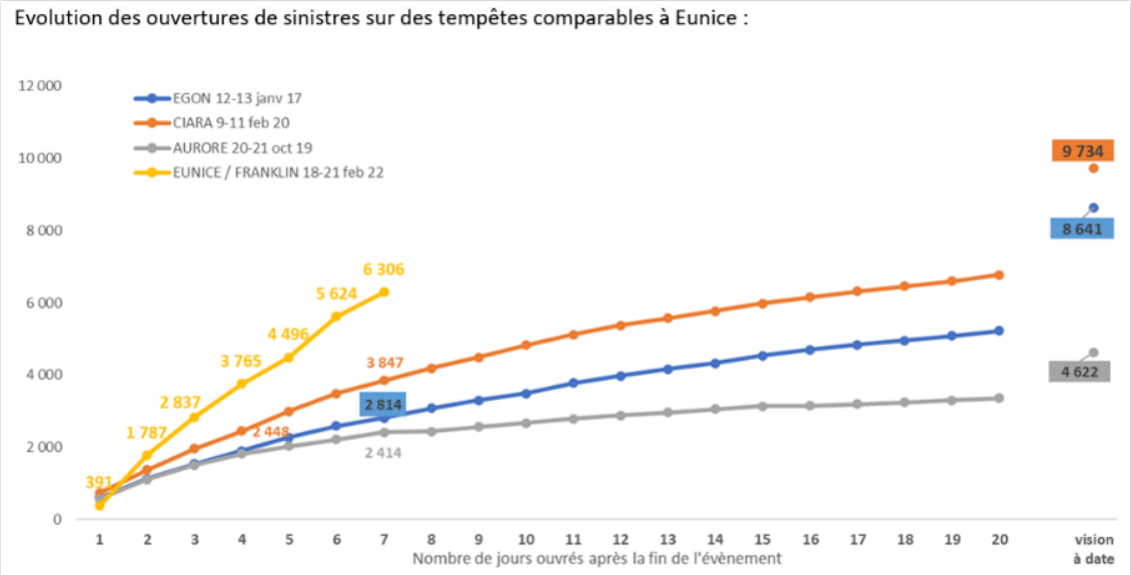


FIGURE 2.2 – Comparaison cadences d'ouverture tempête Eunice avec d'autres tempêtes similaires

L'évènement Eunice a généré un fort volume d'ouvertures durant les premiers jours contrairement aux autres tempêtes Egon, Ciara et Aurore.

La tempête Ciara a le cadencement du nombre d'ouvertures le plus proche de celle que l'on cherche à estimer. Viennent ensuite les tempêtes Aurore et Egon.

Une semaine après ouverture en ne comptant que les jours ouvrés, la cadence d'ouvertures de la tempête Eunice était de 6306 ouvertures.

A la même période, c'est-à-dire 7 jours après ouverture, le cadencement de la tempête Ciara était autour de 40% du nombre d'ouvertures final ($\frac{3847}{9734}$). On applique donc le cadencement du nombre d'ouvertures de la tempête Ciara pour en déduire la cadence d'ouverture finale prévisible de 15765 sinistres pour la tempête Eunice (soit $\frac{6306}{40\%}$). cf fig 2.5

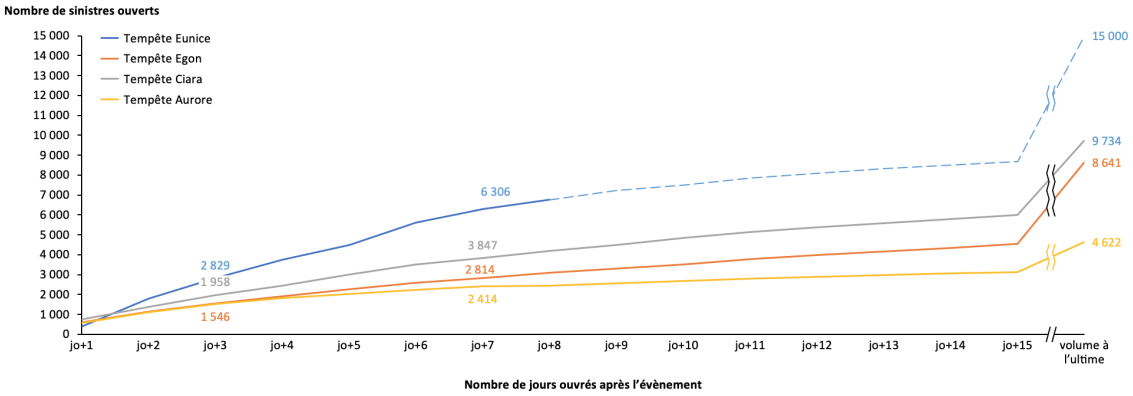


FIGURE 2.3 – Projection des cadences d'ouverture tempête Eunice

D'un évènement à un autre, il y a une forte volatilité sur les cadences d'ouvertures et les coûts moyens. De ce fait, lorsqu'on réalise nos différentes estimations, on se fixe une marge d'erreur plus ou moins variable selon la branche d'activité (Auto, Dommages Aux Biens etc). Après 3 jours ouvrés, on estime le volume à l'ultime entre 13000 et 18000 sinistres.

Plus on avance dans le temps, plus on est à mesure d'apporter une estimation précise de notre charge à l'ultime. Pour une vision au 17 mars 2022, soit 20 jours ouvrés après ouverture des sinistres, on obtient entre 12000 et 13500 sinistres ouverts estimés.

Estimation de la charge à l'ultime

Pour l'estimation de la charge à l'ultime, on regarde le coût moyen hors sinistres graves d'un évènement récent de même nature. Pour la tempête CIARA par exemple, le coût moyen était autour de 2500 euros. Celui-ci va nous permettre de déterminer la charge à l'ultime pour les sinistres relatifs à la tempête Eunice.

On obtient alors la projection du volume et de la charge à l'ultime suivants pour une vision au 17 mars 2022 :

		projection à l'ultime
	Nombre de sinistres	**Charge à l'ultime**
TOTAL	**12.000 - 13500**	**31-35 M€**
DAB particuliers		21 - 24 M€
Automobile		1,6 M€
DAB pro		5,7 - 6,7 M€
Dab agricoles		2,7 M€

TABLE 2.3 – Projection à l'ultime des sinistres pour la tempête Eunice

Concernant la répartition sur les différentes branches, on la détermine en appliquant des hypothèses basées sur la répartition moyenne des sinistres par type de contrat sur des évènements passés de même type (tempête ici en l'occurence).

Vérification des hypothèses de Chain Ladder pour les EGA

Nous aurions pu essayer de traiter ce sujet en utilisant les méthodes classiques de Chain Ladder. La méthode d'ailleurs utilisée précédemment fait penser à une forme déguisée de Chain Ladder. Prenons l'exemple de la tempête EUNICE. On a déjà détecté précédemment les évènements ressemblants. A partir de ces évènements ressemblants, construisons un triangle de développement des cadences d'ouverture.

Evènements "ressemblants"	JO+1	JO+2	JO+3	Jfinal
Aurore	573	1 109	1 513	4 622
Egon	612	1 141	1 546	
Ciara	736	1 381		
Eunice	390			

FIGURE 2.4 – Triangle de développement tempête Eunice avec des tempêtes ressemblantes

Pour l'application de la méthode de Chain Ladder, il faut au préalable que l'hypothèse sur les ratios des facteurs adjacents soit respectée, autrement dit qu'ils soient indépendants de la survenance.

Or les facteurs de développement dans le cadre de la tempête Eunice ne sont pas constants d'une journée d'ouverture à une autre comme le montre la figure 2.5.

Facteurs de développement	1,93455497	1,36490753	3,0548579
	1,86590352	1,3549518	
	1,8763587		

FIGURE 2.5 – Vérification hypothèse de Chain Ladder

L'hypothèse de Chain Ladder n'est pas bien respectée, on le voit notamment au jour JO+1 où les coefficients sont légèrement différents d'une ligne à une autre. C'est pour cette raison qu'elle n'est pas utilisée directement dans le cadre des Évènements de Grande Ampleur.

Néanmoins, la méthode d'estimation de la charge à l'ultime actuellement utilisée reste assez approximative avec des résultats qui se basent sur un historique de données assez limité et des coûts moyens pouvant être très volatiles d'une année à une autre. De plus, un facteur essentiel n'est pas pris en compte lors de cette estimation : le facteur géographique. Ainsi, nous allons par la suite construire une nouvelle base de données plus exhaustive et riche tenant compte du critère géographique. L'estimation de la charge à l'ultime sera donc réalisée à partir de cette nouvelle base.

Chapitre 3

Construction de la base de données

3.1 Objectifs

La règle d'identification des Évènements de Grande Ampleur semble évoluer au cours du temps et se base pour certaines compagnies d'assurance comme AXA principalement sur la charge en excluant d'autres paramètres.

Notre objectif est donc d'enrichir notre catalogue d'évènements climatiques en utilisant plus d'informations sur les évènements détectés avec comme ambition d'améliorer la prédiction de la charge à l'ultime des nouveaux évènements.

Compte tenu des données à notre disposition, de leur forme et de leur contenu, on a besoin de redéfinir ce qu'est un EGA. On va, dans un premier temps, se concentrer sur la **détection d'anomalies** au sein de nos données : lorsque pour un lieu et une date/période de temps donnés, on observe un volume total de garanties sinistrées plus important/aberrant que d'habitude/la normale, alors il s'agira d'une anomalie. Un EGA sera donc une anomalie ou un « regroupement » d'anomalies.

3.2 Récupération des sinistres

3.2.1 Périmètre

Comme énoncé précédemment, on cherche à détecter des anomalies. Pour cela on va utiliser une base de données sinistres propre à AXA contenant les sinistres sur lesquels a été ouverte au moins une UP (Unité de Prestation) climatique. On entend par UP la garantie activée pour l'indemnisation du sinistre. On a donc dans notre cas de figure une UP tempête, une UP inondation, une UP grêle et une UP séisme.

La base de données concerne un échantillon de près de 3 millions d'observations (2 914 000) pour des UP climatiques tempête, grêle, inondation et séisme. Il s'agit d'une base de données avec arrêté des comptes à Juin 2022 et avec des sinistres survenus et

clos de 1989 à 2021.

3.2.2 Variables

Les données proviennent de diverses sources que sont le dataware qui est la base de données cloud utilisée par AXA depuis 2017, l'infocentre qui était la base utilisée avant 2017 ainsi que les informations tirées des données de contrat (geoiris).

Variables	Type	Description
DTSURV	DateTime	Date de survenance du sinistre
DTM_Lieu_Du_Sinistre	Object	Lieu du sinistre
DTM_PostalCode	Integer	Code Postal
DTM_City	Object	Ville du sinistre
UP	Object	Garantie activée pour l'indemnisation du sinistre
infocentre_CDPOSURV	Float	Code Postal issu de l'infocentre
geoiris_CP	Float	Code Postal issu des données de contrats
geoiris_COMMUNE	Object	Nom de commune issu des données de contrats

TABLE 3.1 – Descriptif des variables géographiques et temporelles

Il y a donc 3 axes principaux dans notre étude : la dimension de lieu, de temps ainsi que la variable cible qui est le volume d'UP. En effet, les variables de géolocalisation et de temps sont des paramètres sur lesquels on peut jouer et qui vont être déterminants dans nos futures analyses. On les a ainsi a beaucoup retravaillées. La variable de lieu, et plus précisément son unité/sa maille, a un rôle central dans la pertinence et la cohérence des données et résultats. Dans nos données brutes, on est à la maille des communes (communes 'principales' mais également communes déléguées/rattachées qui se rapprochent des lieux dits sans pour autant en être).

L'identification d'un EGA va donc passer par 4 étapes :

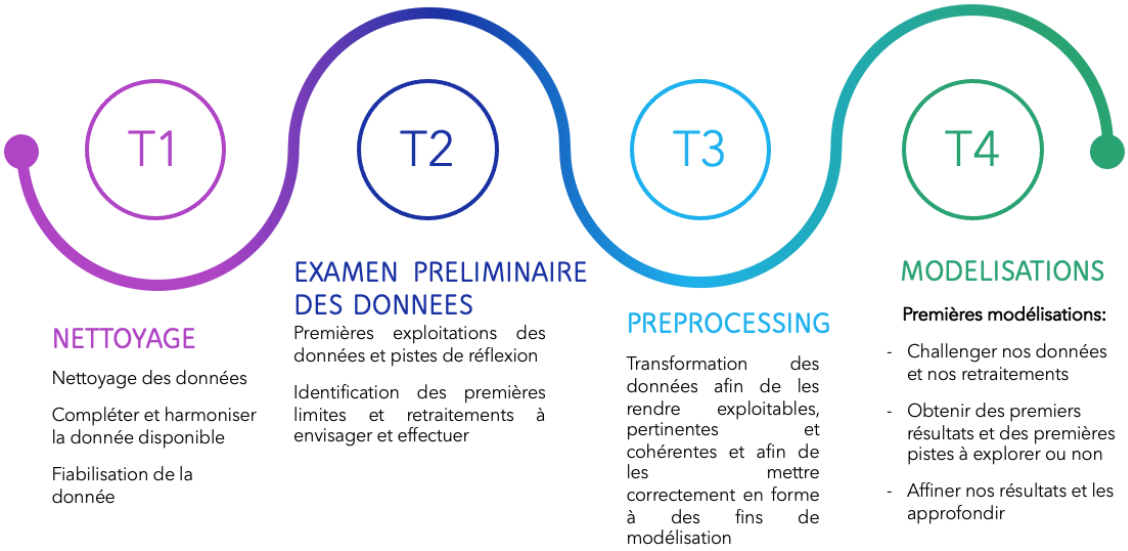


FIGURE 3.1 – Etapes de détection d'un EGA

3.3 Nettoyage des données

3.3.1 Premiers retraitements

Notre base de données initiale contient plusieurs sinistres pour des garanties climatiques tempête, inondation, grêle et séisme. Le nombre de lignes pour lesquelles l'information relative au lieu du sinistre (DTM) est manquante est de 1 526 130 lignes, ce qui correspond à 52.4% de nos données.

Certaines lignes peuvent être existantes mais l'information renseignée est fausse ou non exploitable.

Ex : DTM_Lieu_Du_Sinistre == 'RES.'

Cette variable lorsqu'elle est renseignée est, dans la très grande majorité des cas, présente sous la forme suivante : CP XXXX

Ex : DTM_Lieu_Du_Sinistre == '62000 ARRAS'

Une première correction consistera à dissocier le code postal (si présent) du lieu de sinistre contenu dans la variable. On procédera ensuite à une harmonisation et une correction des lieux de sinistre précédemment extraits grâce à des fonctions de nettoyage et traitements de texte.

	DTM_Lieu_Du_Sinistre	DTM_PostalCode	lieu_du_sinistre	code_postal
1	92190 MEUDON	92190	MEUDON	92190
2	75016 PARIS	75016	PARIS	75016
3	83700 ST RAPHAEL	83700	SAINT RAPHAEL	83700

TABLE 3.2 – Exemple dissociation Code postal du lieu du sinistre

Il est possible voire normal qu'après retraitement on ait davantage de valeurs manquantes. Ces nouvelles valeurs manquantes correspondent :

- aux anciennes valeurs aberrantes ou fausses
- aux valeurs mal renseignées
- aux valeurs non présentes en France
- aux valeurs contenant seulement un code postal

Ex : DTM_Lieu_Du_Sinistre == '28200' donne lieu_du_sinistre == Null mais code_postal == "28200"
Ex : DTM_Lieu_Du_Sinistre == 'RES.' donne lieu_du_sinistre == Null mais code_postal == Null

3.3.2 Seconds retraitements

Malgré l'étape précédente, on a toujours des lignes pour lesquelles le lieu du sinistre est manquant. Par conséquent, on a besoin d'aller plus loin dans le nettoyage et la recherche d'information en se concentrant sur les cas où le lieu du sinistre n'est toujours pas connu mais le code postal l'est dans les données brutes.

On utilise alors le code postal précédemment extrait afin de récupérer l'information relative au lieu du sinistre.

Lorsque l'on tombe sur des cas où la structure de la variable DTM_Lieu_Du_Sinistre ne se conforme pas à la nomenclature habituelle, on parcourt les autres variables contenant l'information code postal (DTM_PostalCode, infocentre_CDPOSURV...) ou on utilise des données externes (INSEE) afin de trouver des correspondances entre un code postal donné et le nom de la commune qui lui est associé.

Ex : DTM_Du_Lieu_Sinistre == '28200', le code_postal == "28200" donne lieu_du_sinistre == "PLOUHINEC"

Après emploi du code postal, il reste de cas pour lesquels le lieu et le code postal sont inconnus ou non renseignés dans nos données(absence de données ou absence de correspondances avec les données externes et officielles de l'INSEE). Mais il est possible de traiter les cas où le code commune est connu mais pas le lieu du sinistre ni le code postal.

- Rassemblement des codes communes déjà existants
- Correction de ces derniers

Ex : le code '00000' est transformé en Null, le code '1400' est transformé en "01400" (problème de format lors de l'importation)

	DTM_Lieu_Du_Sinistre	lieu_du_sinistre	code_postal	code_commune
1	null	MAREIL SUR MAULDRE	null	78368
2	null	CELSOY	null	52090
3	null	SAINT MAUR DES FOSSES	null	94068
4	null	DIJON	null	21231
5	null	CHAMPFROMIER	null	01081

TABLE 3.3 – Illustration de l'exemple dans la base

On effectue ensuite un mappage entre les codes et les communes grâce aux données

externes.

Ex : si on dispose du code commune 59350, on va récupérer comme lieu du sinistre LILLE.

Après application des premières corrections, on se retrouve avec 1 171 497 de lignes pour lesquelles l'information 'lieu du sinistre' est manquante, ce qui correspond à 40.23% de nos données (contre 52.4% de NaN dans les données brutes).

3.3.3 Dernier retraitement

On va ensuite tenter d'exploiter les données CONTRAT afin d'enrichir et d'harmoniser au maximum nos données. On exploite les données contrat en dernier recours car les données précédentes sont plus fiables car plus récentes et réellement relatives à l'endroit précis du sinistre. On a ainsi suivi une certaine hiérarchie au sein des données en fonction de leur qualité et de leur fiabilité.

Pour l'exploitation des données CONTRAT, on se concentre donc ici sur les lignes pour lesquelles notre nouvelle variable *"lieu_du_sinistre"* est nulle et pour lesquelles on ne dispose pas d'information exploitable au sein des autres variables (c'est-à-dire les données anciennes, données pour lesquelles *DTM_Lieu_Du_Sinistre* n'existe pas et n'était pas encore renseignée).

Le lieu du sinistre devient la commune contrat (*après correction et harmonisation*) si elle est renseignée.

Si ce n'est pas le cas :

- Le code postal contrat est renseigné et est valide : recherche de correspondance entre celui-ci et nos données externes
- Le code postal contrat n'est pas renseigné ou n'est pas valide : on ne peut rien faire et le lieu du sinistre demeure inconnu

On notera qu'au cas où certains codes dont on disposerait ne seraient pas des codes postaux mais des codes communes, on réalise également le mappage entre ces derniers et les données externes codes communes (on n'écrase pas les résultats déjà établis mais cela peut servir à les enrichir). On a également actualisé quand c'était possible les codes commune et postal qui n'existent plus (ancienne commune, commune rattachée à une autre etc).

En conclusion, il y a à peu près 30% de lignes pour lesquelles on ne dispose d'aucune information possible à retraiter telles que : lieu du sinistre, ville, code postal, code commune (toute source de donnée confondue) soit par ce que ces informations n'existent pas dans la base originelle soit parce qu'après revue des données, celles-ci se sont avérées

erronées/non utilisables. Une bonne partie de ces lignes correspondent à des sinistres ayant eu lieu entre 1989 et 1999 comme le montre la figure 3.5.

ANNEESURV	non_renseigne_donnees_initiales	code_commune_non_renseigne_donnees_corrigees	lieu_sinistre_non_renseigne_donnees_corrigees
1989	87.51	89.02	87.82
1990	93.98	98.34	93.99
1991	92.54	97.05	92.56
1992	93.39	98.11	93.39
1993	92.23	97.4	92.25
1994	90.65	97.18	90.66
1995	89.63	96.17	89.64
1996	84.95	92.43	84.99
1997	80.83	87.38	80.88
1998	63.59	70.04	63.85
1999	1.41	1.48	1.83
2000	1.64	1.88	2.29
2001	0.86	1.1	1.33
2002	1.11	1.47	1.69
2003	0.91	1.2	1.32
2004	0.99	1.27	1.31
2005	1.46	1.9	1.95
2006	1.18	1.58	1.63
2007	1.35	1.66	1.6
2008	1.45	1.91	1.91
2009	1.17	1.47	1.4
2010	1.85	2.55	2.08
2011	2.6	3.88	3.07
2012	1.4	2.46	1.7
2013	2.14	3.97	2.53
2014	2.84	5.65	3.66
2015	0.33	3.82	0.92
2016	0%	3.69	0.58
2017	0%	6.02	0.22
2018	0%	7.04	0.41
2019	0%	10.1	0.73
2020	0%	11.62	0.23
2021	0%	14.41	0.46

TABLE 3.4 – Répartition valeurs manquantes par année

Parmi les données pour lesquelles le lieu du sinistre est manquant, on compte 7 489 lignes pour lesquelles on dispose d'une information. Cependant, celle-ci n'a pas permis d'établir de manière fiable le lieu du sinistre.

De nombreux cas particuliers existent, et il est donc difficile de tous les traiter avec des fonctions génériques et/ou sans altérer les autres données déjà existantes et fiables.

Parfois, les données ne nous semblent pas erronées ou "bizarres" mais celles-ci sont malheureusement introuvables dans nos données externes (un code postal qui suit la nomenclature française et officielle mais qu'on ne retrouve pas dans les données externes

anciennes et actuelles).

Pour ces raisons, nous avons décidé de filtrer nos données sur l'horizon 2000-2021 en ne gardant que les données de ce scope qui contiennent une information géographique exploitable. On se retrouve désormais avec un peu plus 2 240 000 lignes allant de 2000 à 2021 et qui contiennent une information exploitable relative au lieu.

3.4 Examen préliminaire des données

3.4.1 Première tranformation et regroupement des données

On souhaite passer de la notion d'UP à la notion d'évènement journalier (observation de sinistres jour après jour)

On procède à la création d'un premier dataset simple avec les paramètres suivants :

- Lieu : lieu du sinistre représenté par les variables géographiques code_commune_actualise (après nettoyage et actualisation c'est-à-dire correction des codes communes anciens qui n'existent plus), et code_postal_actualise (après nettoyage correction et remplacement des codes anciens par les nouveaux quand c'est possible)
- Temps : date de survenance
- Variable d'intérêt : le nombre d'UP associé au lieu et au temps t
- Prise en compte et regroupement par UP (répartition volume et charge)

	code_commune_actualise	code_postal_actualise	UP	DTSURV	occurrence_up	lieux	merlin_chg
0	01014	01100	GRELA	2008-05-29	29	ARBENT	57.58212
1	01024	01340	TEMP	2020-03-01	1	ATTIGNAT	0.37678
2	01033	01200	TEMP	2001-11-14	1	VALSERHONE	0.29300

TABLE 3.5 – Aperçu du dataset créé, Données réelles

On a créé ce dataset car on souhaite passer de la notion d'UP à la notion d'évènement journalier (observation de sinistres jour après jour).

Plus précisément, on a souhaité passer de 1 ligne = 1 UP d'un sinistre au lieu L et au temps T à 1 ligne = 1 lieu observé au temps T et pour lequel on a observé X fois l'UP grêle/tempête etc et Y sinistres. Ainsi, les individus observés et manipulés sont des lieux

dont les caractéristiques connues et exploitées sont les volumes d'UP et la charge.

Quelques chiffres

Le nombre de communes dont le nombre d'UP cumulé entre 2000 et fin 2021 est inférieur ou égal à 100 : 43 279

Le nombre de communes dont le nombre d'UP cumulé entre 2000 et 2021 est inférieur ou égal à 10 : 36 717

La variance du nombre cumulé d'UP par commune : 1114.78

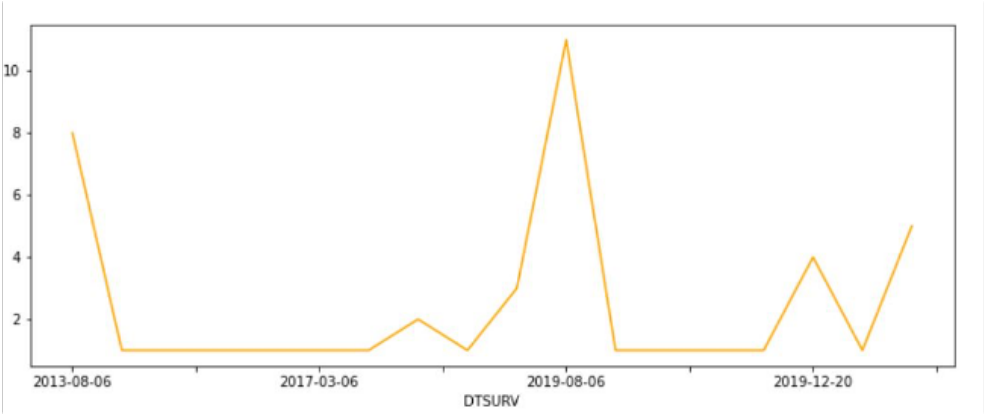


FIGURE 3.2 – Volume total d'UP observé au cours du temps au sein du lieu de sinistre «Champdieu»

Notre examen préliminaire nous a montré qu'il existe une trop grande disparité entre les communes au cours de ces années d'historique en termes de volume ainsi qu'en termes de fréquence journalière : pour beaucoup de communes il y a beaucoup de jours au cours desquels il ne s'est rien passé.

- On dispose de séries temporelles trop partielles pour certains individus
- Appliquer un modèle de Machine Learning sur ce genre de séries temporelles partielles et hétérogènes n'est pas recommandé (absence de résultats cohérents et pertinents)

On pourrait songer à enrichir nos données afin de disposer de données journalières entre 2000 et fin 2021 mais cela ne ferait pas totalement disparaître ce problème : un algorithme de détection d'anomalies de Machine Learning' aurait tendance à identifier chaque évènement où le nombre d'UP est différent de 0 comme une anomalie ou une observation au comportement anormal par rapport au reste de la série (même si celui-ci serait très faible comme dans le cas de Champdieu, car un volume subitement égal à 5

serait une anomalie si pendant des jours/années celui-ci était égal à 0-1)

On ne peut pas exclure d'analyser les données jour après jour car il est possible qu'un arrêté CATNAT ne couvre qu'une seule journée (environ 30% des arrêtés).

Les données à la maille géographique 'communes' sont difficilement comparables entre elles au cours du temps. On ne risque de détecter des anomalies uniquement lorsque celles-ci ont lieu dans les grandes villes et donc de les sous-estimer, ou au contraire, on risque de sur-estimer le nombre d'EGA en voulant être trop conservateurs (et ne pas exclure les communes plus "modestes", mais par conséquent on identifiera l'intégralité ou presque des évènements/dates au sein des métropoles comme relevant d'anomalies).

Notre objectif principal va donc être de créer de nouveaux individus, c'est-à-dire de nouvelles unités géographiques afin "d'harmoniser" les volumes d'UP associés à chaque individu. En effet, on veut rendre les individus davantage comparables les uns avec les autres mais aussi obtenir des individus et observations pouvant être utilisés par un algorithme de Machine Learning.

3.4.2 Méthodologie

Afin "d'harmoniser" les volumes d'UP associé à chaque individu, nous avons procédé à l'élaboration d'un nouveau référentiel géographique : nous avons créé une nouvelle unité géographique afin de découper le territoire français dont nous disposons dans nos données initiales ; la France ne sera plus une union de communes mais une union de zones géométriques plus ou moins grandes. Cette déconstruction va nous permettre de créer de nouveaux individus.

Voici une description rapide de la méthodologie employée à travers ces principales étapes :

- On va créer un nouveau zonier qui n'est plus basé sur les communes.
- La France est découpée de manière géométrique et équitable. En effet, le pays a été découpé indépendamment des communes, de leurs coordonnées géographiques et de leurs frontières.
- La taille de ces zones est paramétrable ; il ne faut pas créer des zones trop petites sous peine de reproduire notre problème, ni des zones trop grandes afin de conserver une granularité et une maille géographique assez précise et tenant compte des singularités et différences entre les territoires français.

Le paramètre est appelé la «résolution» : plus la résolution est grande, plus les zones vont être petites et donc plus nombreuses au sein du polygone que l'on cherche à diviser ; à l'inverse, plus la résolution sera petite, et plus les zones seront grandes. Pour des

raisons techniques, on ne peut pas dépasser une résolution égale à 8. Nous nous sommes vite rendus compte qu'une résolution inférieure à 4 produisait des zones beaucoup trop larges. Après plusieurs essais, nous avons déterminé qu'une résolution de 5 était un bon compromis et permettait de répondre de manière satisfaisante à notre problématique.

On va notamment avoir besoin, pour certains individus, c'est-à-dire zones, d'effectuer des regroupements.

» Certaines vont être potentiellement plus ou aussi petites que les communes de départ ou bien encore toujours trop peu significatives et représentatives en termes de nombre d'évènements observés. Dans ce cas là, notre problème de départ, c'est-à-dire le manque d'évènements et d'UP observés ne serait pas résolu.

» On va devoir procéder à une agrégation des zones géométriques créées en utilisant une méthode de seuil. En effet, un individu I doit être agrégé et faire partie d'un plus grand ensemble si son volume global d'UP est jugé insuffisant, c'est-à-dire si celui-ci est inférieur à une valeur de seuil donné.

3.5 Preprocessing : Création d'un zonier

3.5.1 Premier découpage spatial et réarrangement des données

Les communes sont maintenant englobées et pour certaines à cheval sur plusieurs zones géométriques hexagonale.

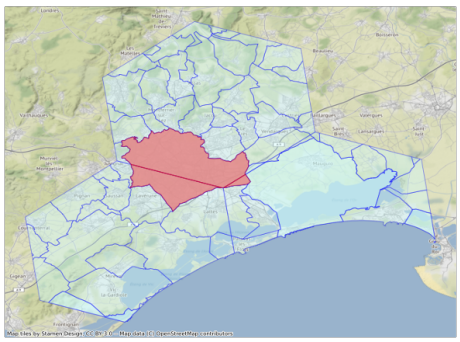


FIGURE 3.3 – Montpellier après 1er découpage

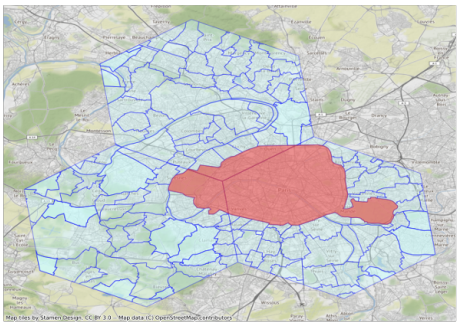


FIGURE 3.4 – Paris après 1er découpage

Sur les deux exemples de Montpellier et Paris, les deux villes apparaissent en rouge. Elles sont à cheval et font partie de trois zones hexagonales (en bleu).

individus_communes	individus_hex
Montpellier	zone_hex1
Montpellier	zone_hex2
Montpellier	zone_hex3
Paris	zone_hex4
Paris	zone_hex5
Paris	zone_hex6

TABLE 3.6 – Nouvelle structure de données

L'unicité des individus est ici bien assurée par la variable "individus_hex", et non plus par les communes. Les communes deviennent une composante des zones hexagonales.

3.5.2 Deuxième retraitement spatial des données – consolidation des nouveaux individus grâce à des regroupements

Comme expliqué précédemment, nous n'avons aucune garantie que les nouvelles unités géographiques précédemment créées constituent des individus davantage «homogènes» entre eux et dont les séries temporelles sont satisfaisantes d'un point de vue statistique.

On va donc procéder à un regroupement de zones afin d'observer un volume d'UP plus élevé et "pertinent" au sein de celles-ci et ainsi créer des séries temporelles exploitables et pouvant être utilisées au sein d'un modèle de Machine Learning.

Puis on va calculer, pour chaque zone, le nombre d'UP total recensé (données plus fiables et suffisantes en quantité).

	lieu_du_sinistre	code_commune	code_postal	UP	DTSURV	occurrence_up	merlin_chg	outre_mer
1	ABREST	03001	03200	TEMP	2016-06-01	1	1.757	0
2	AIGUEFONDE	81002	81200	TEMP	2014-02-12	1	1.172	0
3	ALENYA	66002	66200	TEMP	2020-01-21	1	0.56578	0
4	AMBARES ET LAGRAVE	33003	33440	GEL	2020-03-01	1	0	0
5	ANCHE	37072	37500	TEMP	2018-05-26	1	1.51592	0

TABLE 3.7 – Aperçu du dataset créé, données fictives

3.5.3 Identification des zones à traiter et qui ne se suffisent pas à elle-même

On utilise pour cela une variable de seuil que l'on a au préalable identifiée. Les zones n'atteignant pas ce seuil vont donc être agrégées avec d'autres afin de satisfaire notre

contrainte

Calcul et identification de la valeur de seuil

Nous avons utilisé pour cela une commune de référence, c'est-à-dire une commune pour laquelle :

- on observe un volume global d'UP entre 2000 et fin 2021 suffisant
- la fréquence d'évènements est suffisante
- on observe des valeurs extrêmes (maximum locaux)

C'est en se basant sur ces critères que nous avons sélectionné la ville de LILLE . Note SEUIL = volume global d'UP observé à LILLE entre 2000 et fin 2021

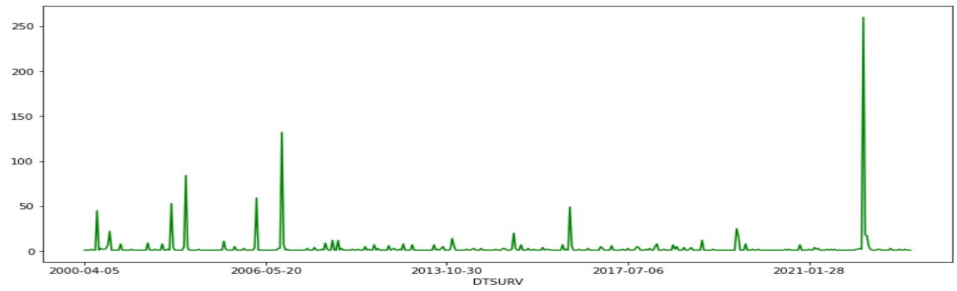


FIGURE 3.5 – Volume global d'UP observé à LILLE entre début 2000 et fin 2021

Remarques, limites et critiques

La méthode est purement déterministe et dépend entièrement du seuil choisi. Il se peut que nous n'ayons pas sélectionné le plus pertinent. L'optimisation de cet hyperparamètre est cependant difficile : comment juger directement de la pertinence de l'output ?

Il faut pour cela avoir suffisamment de recul sur la donnée et disposer de données plus concrètes et chiffrées (volume d'UP, nombre d'anomalies identifiées et donc les résultats découlant des futurs algorithmes)

Application d'une marge d'erreur

Au lieu d'appliquer un seuil strict, on pourrait prendre en compte et appliquer une marge d'erreur : si pour la zone z, son volume d'UP est compris entre [0 ; seuil – erreur], alors celle-ci doit être agrégée (au lieu de l'intervalle plus restrictif [0, seuil]).

Ainsi, on a fixé une marge d'erreur à 30% qui peut sembler énorme mais cela permettait d'éviter d'agréger un voisin pour quelques évènements manquants, c'est-à-dire

de passer d'une zone à quelques évènements du seuil à une zone à une zone comptant un nombre d'évènements dépassant par exemple 3, 4, 5 fois la valeur de seuil.

Application

Pour chaque individus/zone z, on calcule le volume total d'UP qui lui est associé entre 2000 et fin 2021. Pour chaque zone n'atteignant pas le seuil S, on examine ses 25 plus proches voisins (en termes géographiques : longitude/latitude). Pour ce faire, on utilise la méthode des k plus proches voisins et le module 'spatial' du package Python 'scipy' et le module 'KDTree'.

Algorithme k-dimensional tree

Cet algorithme est notamment utilisé pour trouver rapidement les voisins les plus proches d'un point et est privilégié à l'algorithme des k plus proches voisins (KNN) pour la gestion de données spaciales.

L'idée générale derrière le kd-tree est qu'on a un arbre binaire où chaque nœud représente un hyperrectangle aligné sur l'axe. Chaque nœud spécifie un axe et divise l'ensemble de points selon que leurs coordonnées le long de cet axe sont supérieures ou inférieures à une certaine valeur.

Lors de la construction, la sélection de l'axe et du point de division adopte la règle du "point médian glissant", qui garantit que les cellules ne deviendront pas toutes plus longues et plus fines.

Les r voisins les plus proches d'un point donné peuvent être déterminés à partir de l'arbre (éventuellement, il ne peut renvoyer que ceux situés à la distance maximale de ce point).

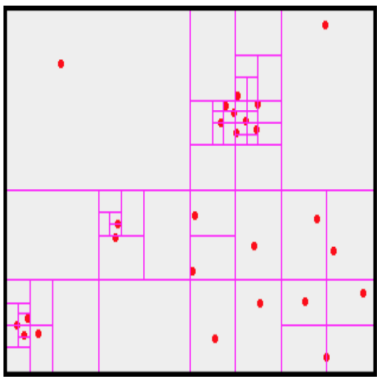


FIGURE 3.6 – Principe du KDTree [34]

Donc pour chaque zone n'atteignant pas le seuil S qu'on s'est fixé préalablement, on examine ses 25 plus proches voisins et on lui agrège autant de voisins que nécessaire afin d'atteindre ce seuil.

Néanmoins, il se peut que regrouper 25 + 1 zones (le '+1' correspondant à la zone intiale pour laquelle on calcule et identifie ses plus proches voisins) ne soit toujours pas suffisant. Dans tous les cas, nous avons choisi de nous limiter à 25 voisins/zones à agréger.

Nous sommes très (trop) conservateurs car dans les faits, en moyenne, seulement entre 5 et 6 voisins sont nécessaires à chaque individu. Regardons un exemple avec la commune de REMIREMONT.

La commune de REMIREMONT est localisée au sein de la zone rouge ; la zone agrégée créée autour d'elle et pour laquelle elle appartient apparaît en bleue.

7 voisins ont été nécessaires afin d'atteindre la valeur de seuil S. On obtient maintenant 2544 individus/zones. Celles-ci peuvent partager des points communs et se superposer, cela est normal et voulu et se justifie d'un point de vue méthodologique et statistique.

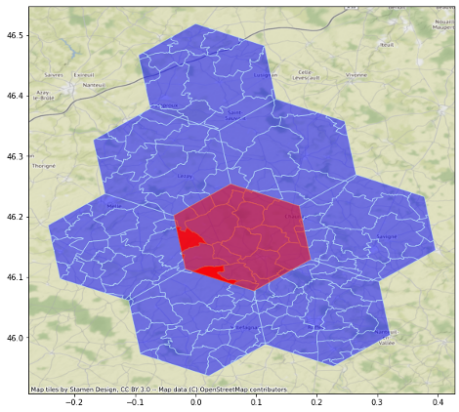


FIGURE 3.7 – Exemple avec la commune de REMIREMONT

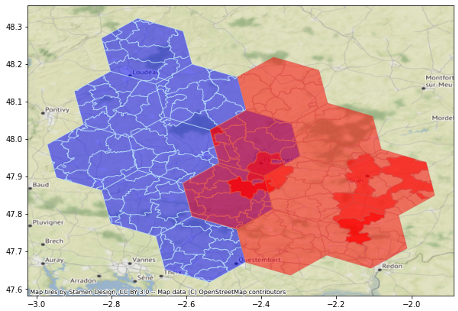


FIGURE 3.8 – Exemple avec la commune de REMIREMONT

3.5.4 Transformation des données originelles relatives aux sinistres et UP

Cette étape est nécessaire afin d'implémenter par la suite des modèles de classification et de détection d'anomalies. En effet, nous avons besoin de modifier nos données originelles et plus précisément les individus observés : nous devons exprimer nos communes sous le nouveau référentiel géographique afin de pouvoir exploiter dans un modèle de

Machine Learning les nouveaux individus créés.

Décomposition des communes en sous-zones

On veut obtenir un dataset comme suit afin de développer par la suite notre modèle de Machine Learning de détection d'anomalies : 1 ligne = une commune exprimée sous son code commune ainsi que les zones hexagonales la composant.

code_commune	hex_id_unique
01001	841f91dfffffffff
01002	841f911fffffffff, 841f912fffffffff

TABLE 3.8 – Aperçu du dataset créé, données fictives

lieu_du_sinistre	code_commune	UP	occurrence_up	DTSURV	hex_id_unique_commune
WIRWIGNES	62896	NATUR	1	2009-12-24	85194823ffffff, 8519483bffffff
SAINTE SIGOLENE	43224	NATUR	1	2013-11-20	851f906bffffff, 851f92a7ffffff
SAINT MARTIN URIAGE	38422	GEL	1	2014-02-01	851f9ecbffffff
GILLY SUR ISERE	73124	NATUR	2	2015-02-25	851f9e83ffffff, 851f9e93ffffff
BASTELICACCIA	20032	GRELE	1	2016-10-09	851eb2a3ffffff, 851eb2abffffff

TABLE 3.9 – Aperçu du dataset créé, données réelles

Application du nouveau référentiel – Eclatement des communes en sous-zones

On veut obtenir un dataset comme suit : 1 ligne = une zone z hexagonale. On peut donc retrouver plusieurs fois une même commune si celle-ci est à cheval sur plusieurs zones géométriques hexagonales

hex_id_unique	code_commune
8418409fffffffff	56009
8418410fffffffff	56086
841f911fffffffff	01002
841f912fffffffff	01002

TABLE 3.10 – Aperçu du résultat après application du nouveau référentiel sur des données fictives

Afin de ne pas modifier nos données relatives aux volumes/charges d'UP, nous avons besoin de redistribuer les variables numériques d'intérêt entre les zones.

En effet, pour l'instant, si on éclate simplement nos données sans toucher aux autres variables, on se retrouve dans le cas suivant : si la commune C compte au temps t, N UP observées et est constituée des zones H1 et H2, après éclatement, la commune C compte au total 2N* UP.

La solution la plus simple a été de redistribuer de manière équitable les N UP entre les zones H1 et H2, c'est-à-dire d'attribuer 0.5N à H1 et 0.5N à H2 afin de ne pas compter en double, triple etc nos UP et de modifier profondément nos données et par extension les futurs résultats.

	DTSURV	UP	occurrence_up	merlin_chg	lieu_du_sinistre	code_postal_actualise	commune_ou_commune_de_rattachement	code_commune_actualise	hex_id_unique_commune
0	2020-10-06	NATUR	0.333333	0.136	LOCMARIA	56360	LOCMARIA	56114	85184083fffffff
1	2020-10-06	NATUR	0.333333	0.136	LOCMARIA	56360	LOCMARIA	56114	85184093fffffff
2	2020-10-06	NATUR	0.333333	0.136	LOCMARIA	56360	LOCMARIA	56114	85184097fffffff

TABLE 3.11 – Visualisation et illustration du principe de redistribution des variables numériques entre les communes appartenant à une même zone hex : aperçu sur le cas de LOCMARIA et de la zone 85184083ffff

3.5.5 Agrégation des variables numériques d'intérêt

Il faut maintenant effectuer la liaison entre chacune des zones z que l'on a fait apparaître dans nos données avec la zone étendue E à laquelle elles appartiennent.

On veut obtenir un dataset comme suit : 1 ligne = 1 sous-zone hexagonale z. Cependant, les valeurs numériques d'intérêt "volume" et "charge" sont celles observées au sein

de la zone étendue E à laquelle z appartient.

- Si z se trouve autour de zones identifiées comme potentiellement concernées par un EGA, on suppose que la proximité entre les deux ne doit pas être ignorée et que z doit également être identifié comme zone d'alerte/EGA potentiel.

- Notre problème était le suivant : on disposait d'individus/localisations trop précis, créant de grandes disparités comme évoqué précédemment.

- Prendre en compte l'appartenance de chaque individu à une zone plus étendue nous permet d'atténuer ces disparités et de prendre en compte le caractère «diffus» de certains évènements (dans une certaine mesure).

	DTSURV	hex_id	zone_etendue	total_volume_up	total_merlin_chg	lieux	code_postal_actualise	EGA_isolationForest
0	2000-01-02	8518444bfffffff	851846b3fffffff, 8518445bfffffff, 8518444fffff...	0.666667	0.203333	ERGUE GABERIC	56110, 29390, 29900, 29340, 29180, 29510, 2992...	0
1	2000-01-05	8518444bfffffff	851846b3fffffff, 8518445bfffffff, 8518444fffff...	0.666667	0.071333	ERGUE GABERIC	56110, 29390, 29900, 29340, 29180, 29510, 2992...	0
2	2000-04-14	8518444bfffffff	851846b3fffffff, 8518445bfffffff, 8518444fffff...	1.000000	0.000000	ERGUE GABERIC	56110, 29390, 29900, 29340, 29180, 29510, 2992...	0
3	2000-05-09	8518444bfffffff	851846b3fffffff, 8518445bfffffff, 8518444fffff...	0.666667	2.902000	ERGUE GABERIC	56110, 29390, 29900, 29340, 29180, 29510, 2992...	0

FIGURE 3.9 – Aperçu du nouveau dataset agrégé

3.6 Modélisation EGA et théorie des valeurs extrêmes

Pour détecter un EGA, on va utiliser des modèles de détection d'anomalie et faire appel à la théorie des valeurs extrêmes. Un EGA pourra donc être une anomalie ("positive", un maximum local) ou un ensemble d'anomalies au sein d'une série temporelle, c'est-à-dire un évènement inhabituel qui semble s'éloigner et ne pas se comporter comme les autres évènements.

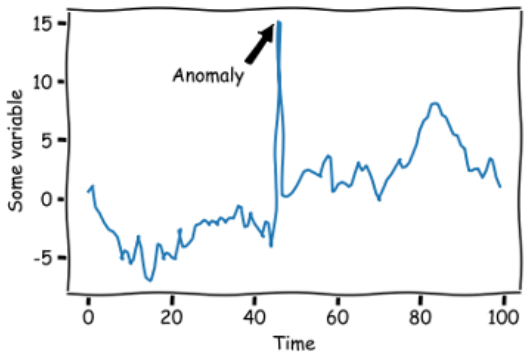


FIGURE 3.10 – Illustration d'une anomalie

Dans le cadre des évènements extrêmes, le théorême central limite n'est plus applicable : c'est là qu'intervient la théorie des valeurs extrêmes. La théorie des valeurs extrêmes fait partie de la large famille des statistiques et se distingue pour son étude spécifique des valeurs aberrantes ou extrêmes au niveau des queues de distribution. En ce sens, elle est très utilisée pour l'étude d'Évènements de Grande Ampleur qui se caractérisent par leur rareté et leur violence. L'approche la plus utilisée aujourd'hui se base sur l'étude des excès (POT : Peaks Over Threshold) et la distribution de Pareto généralisée qui reposent sur des seuils.

3.6.1 Threshold models : Approche POT

En statistique, un modèle à seuil représente tout modèle dans lequel une valeur seuil, ou un ensemble de valeurs seuils, est utilisé pour distinguer des plages de valeurs dans lesquelles le comportement prédit par le modèle varie de manière importante.

Il a été en premier introduit en 1980 par Tong et Lim. Tong constate l'existence d'un cycle limite, l'irréversibilité temporelle et la dépendance amplitude-fréquence qui ne peuvent pas être représentés par des modèles linéaires. Il propose ainsi une classe de modèles autorégressifs avec des seuils qui permettent l'incorporation de phénomènes non linéaires.

Soient $X_1, ..., X_n$ des variables aléatoires réelles indépendantes et identiquement distribuées de fonction de répartition F. Pour étudier les valeurs extrêmes, on regarde les excédents, c'est-à-dire les valeurs qui sont au dessus d'un certain seuil u.

On appelle les excès des variables $X_1, ..., X_n$ au dessus du seuil, l'ensemble des variables aléatoires $Y_1, ..., Y_n$ telles que $Y_i = X_i - u$, $\forall i \in \{1, ..., n\}$.

Dans les modèles à seuil, on cherche à déterminer la probabilité conditionnelle de F

par rapport au seuil u pour toutes valeurs au dessus de ce seuil :

$\forall 0 \leq y < x_F - u$, avec $x_F = sup\{x \in R : F(x) < 1\}$

$$P(X > u + y | X > u) = \frac{F(u+y)-F(u)}{1-F(u)}$$

Pour cela, nous allons passer par la distribution conditionnelle des excès au-delà d'un certain seuil.

$$F_u(y) = P(X - u \leq y | X > u)$$

Le but de cette méthode de dépassement de seuil est de trouver la loi de probabilité permettant d'obtenir une bonne estimation de cette distribution conditionnelle. C'est dans ce cadre que Pickands (1975), Balkema et de Haan (1974) ont proposé le théorème suivant.

3.6.2 Théorème de Pickands-Balkema-de Haan

Reprenons les variables aléatoires $X_1, ..., X_n$ iid de fonction de répartition F. On suppose que l'appoximation de la distribution du maximum par une distribution du GEV [17] est satisfaite :

$$P(M \leq x) = F_{\beta,\alpha,\xi}(x)$$

avec $M = max(X1, ..., X_n)$

et $F_{\beta,\alpha,\xi}$ la fonction de répartition pour la loi GEV de paramètres α, β, ξ.

Ainsi, on peut approcher la distribution des excédents pour u assez grand par la **loi de Pareto généralisée** suivante :

$$G_{\sigma\xi}(y) = \begin{cases} 1 - (1 - \xi\frac{y}{\tilde{\sigma}})^{-\frac{1}{\xi}}, \ \forall\ \xi \neq 0 \\ 1 - exp(-\frac{y}{\tilde{\sigma}}), \ si\ \xi = 0 \end{cases}$$

- Si $\xi < 0,\ y \in [0, Min(-\frac{\sigma}{\xi}, x_F - u)]$
- Si $\xi \geq 0,\ y \in [0, , x_F - u]$

où $G_{\sigma,\xi}$ la fonction de répartition de la loi de Pareto généralisée (GPD), $\xi \in R$ le paramètre de forme et $\hat{\sigma} > 0$ le paramètre d'échelle.

Ainsi, pour une distribution donnée, s'il y a convergence des maxima vers la GEV, il y aura convergence des excès vers la GPD.

3.6.3 Choix du seuil

Dès lors, le choix du seuil u à utiliser lors de l'ajustement du modèle GPD est très important.

Le choix du seuil peut être fait en visualisant graphiquement la durée résiduelle moyenne, c'est-à-dire l'espérance au delà du seuil. La courbe doit être linéaire au-delà du seuil auquel le modèle GPD devient valide.

Soit $X - u_0 | X > u_0 \sim GPD(\sigma, \xi)$, $\xi < 1$, alors $\forall u \geq u_0$

$$MRL(u) = \mathbb{E}(X - u | X > u) = \frac{\sigma_{u_0} + \xi_u}{1 - \xi}$$

avec u la valeur du seuil, ξ le paramètre de forme, σ le paramètre d'échelle.

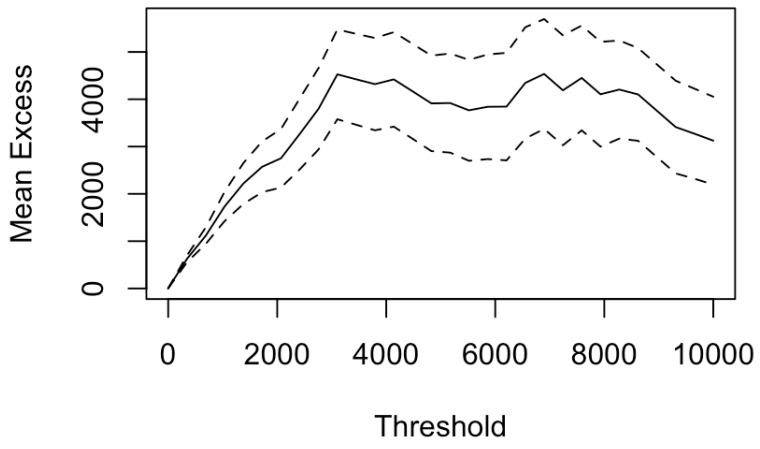


FIGURE 3.11 – Durée de vie résiduelle moyenne pour le volume d'UP

En analysant la courbe, on observe une linéarité entre 3000 et 4000. On part donc du principe que notre seuil se situe entre 3000 et 4000.

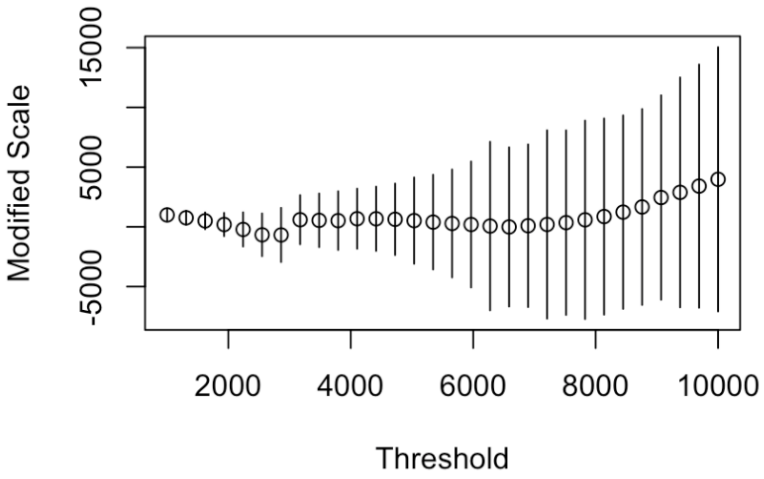


FIGURE 3.12 – Stabilité par seuil du paramètre d'échelle

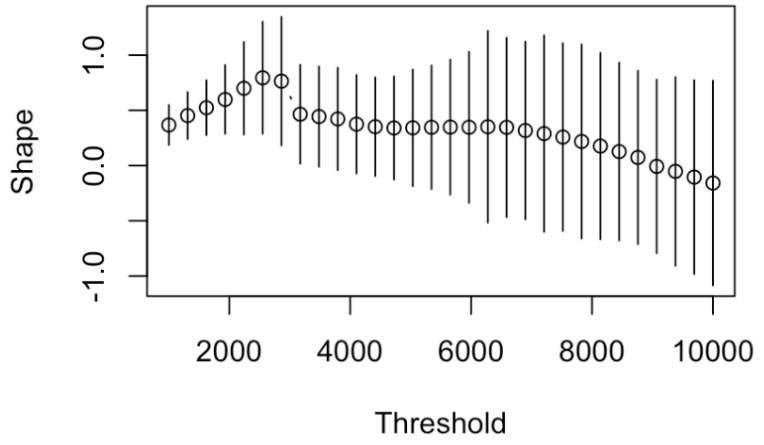


FIGURE 3.13 – Stabilité par seuil du paramètre de forme

Si notre approximation GPD est valable pour les dépassements d'un seuil donné u, dans ce cas l'espérance de vie résiduelle doit évoluer de manière linéaire au dessus de ce seuil et nos paramètres doivent être constants au dessus de ce seuil.

En pratique, on doit donc trouver un seuil suffisamment élevé pour que l'approximation asympotique de la distribution des excès par la GPD soit correcte mais cela entraîne une diminution du biais. Et en même temps, pour avoir une estimation précise des paramètres de la GPD, on doit avoir un seuil pas trop élevé, ce qui entraînerait une variance plus petite et donc potentiellement une mauvaise approximation de la loi asymptotique. Dans la pratique, une valeur de seuil trop basse va avoir pour effet d'inclure trop de zones

et d'identifier trop d'évènements comme étant des anomalies, mais aussi de "stigmatiser" les zones comme les grandes métropoles en identifiant presque chaque évènement/jour comme étant une anomalie potentielle. On parle du **dilemme biais-variance**. Si on prend un trop grand seuil, on va augmenter nos intervalles de confiance car il y aura une diminution du nombre d'observations. Une valeur de seuil trop haute va donc avoir pour effet d'exclure les zones plus "modestes" en termes d'UP et donc de sous-estimer le nombre d'anomalies.

Après analyse des graphiques précédents, un seuil autour de 3000 nous paraît cohérent et nous permet d'avoir un nombre d'observations suffisant avec $\sigma = 867,28$, $\xi = 0,16$.

Les graphiques suivants à savoir le "probability plot" et le "quantile plot" permettent de vérifier la qualité d'ajustement par la loi GPD et de confirmer le choix du seuil.

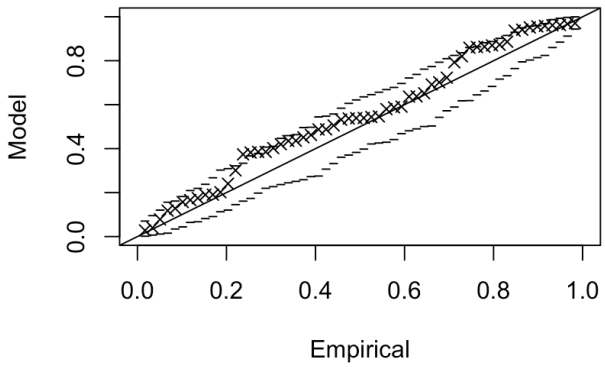


FIGURE 3.14 – Probability Plot

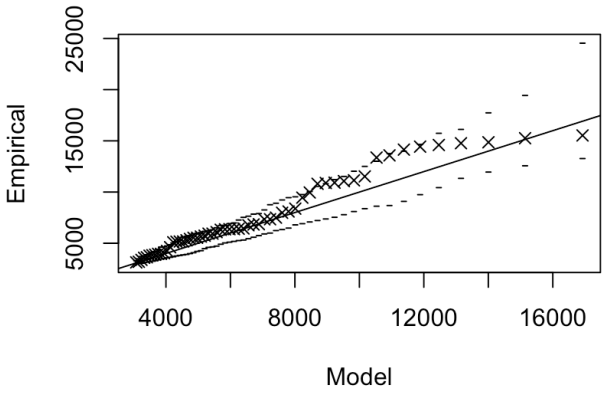


FIGURE 3.15 – Quantile Plot

Nous allons donc utiliser une valeur de seuil de 3000. Dans la pratique statistique, on distingue 2 catégories de modèles de seuil souvent utilisées :

- Fixed threshold Model
- Adaptive thresold Model

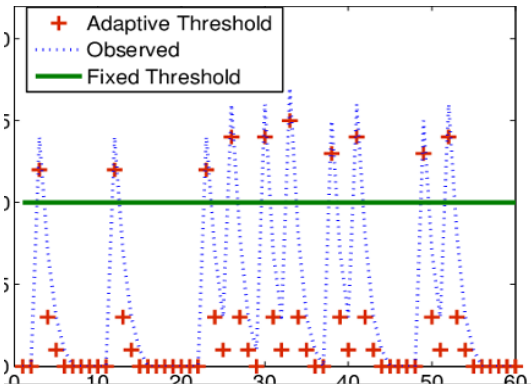


FIGURE 3.16 – Illustration de modèles de seuil

Fixed Threshold model

Il s'agit d'un modèle déterministe et simple basé sur une variable de seuil/de référence fixe et figée. Cette méthode consiste à définir un seuil u de sorte à ce qu'un pourcentage défini des données se trouve au dessus de ce seuil. Dans notre cas, on fixe donc notre valeur de seuil à 3000.

Application : on applique la comparaison au seuil avec les volumes d'UP observés au niveau agrégé ; le résultat sera à interpréter au niveau de la zone individuelle hex.

On examine chaque individu/zone l'un après l'autre. On compare chaque volume d'UP observé aux différentes dates au sein de la zone z avec la valeur de seuil.

- Si volume >= seuil : l'évènement est identifié comme anomalie
- Sinon : l'évènement ne semble pas anormal

Exemple :
On se concentre sur la zone Z1. La zone Z1 appartient à la plus grande zone agrégée ZA1. Dans nos données, on recense des UP associés à la zone Z1 pour les temps t1, t2 et t3. Les volumes d'UP indiqués sont en réalité ceux de la zone ZA1 au sein de laquelle on retrouve et est inclue Z1.

Si au temps t1, le volume total d'UP observé dépasse la variable de seuil, on dira bien

que c'est le couple (Z1, t1) qui est une potentielle anomalie, et non pas le couple (ZA1, t1).

Remarques et critiques

Ce modèle peut présenter des limites, car une valeur de seuil universelle raisonnable et pertinente pour l'intégralité de nos données est très difficile à trouver voire n'existe pas car le territoire que l'on couvre (et donc par extension le nombre d'individus présents dans nos bases) est vaste.

Cette méthode est très «basique» et ne prend pas particulièrement en compte la dimension métier ou encore les spécificités des évènements climatiques.

Résultats

- Nombre total d'anomalies identifiées par la méthode du seuil fixe : 5 408
- Proportion d'anomalies : 0.76%
- Valeur de seuil utilisée : 3000 - un nombre observé d'UP supérieur à 3000 est donc considéré comme une anomalie.

En comparant nos résultats, c'est-à-dire les anomalies détectées avec notre modèle de de seuil fixé à 3000, avec la base de données réelle d'EGA, on conclut que le Fixed Threshold Model n'est pas performant. Parmi les anomalies détectées par ce modèle, seulement 68% coïncide réellement avec un EGA ayant eu lieu.

Adaptative Threshold model

Il s'agit d'un modèle paramétrable basé sur une variable de seuil/de référence adaptative : le principe reste le même que le premier modèle, c'est à dire qu'on compare chaque évènement à une valeur de référence mais celle-ci n'est plus unique et figée. L'objectif est de prendre en compte les disparités entre les individus (atténuées par le redécoupage géographique mais elles n'ont pas pour autant pas totalement disparu).

Application : Pour rappel, on applique la comparaison au seuil avec les volumes d'UP observés au niveau agrégé ; le résultat sera à interpréter au niveau de la zone individuelle hex :

- On calcule ainsi percentile 1-X% pour chaque individu
- Chaque individu se voit donc associé une valeur de seuil personnalisée

- On compare, pour un individu donné, chacun de ses volumes d'UP observés au seuil précédemment établi.

Autrement dit : pour chaque individu/zone, on veut que X% des évènements qui ont eu lieu ressortent en anomalie.

Cibles testées : 1%, 3% et 5%.

Il est plus facile dans cette configuration de faire varier la valeur de seuil car :

- Il suffit de calculer le percentile X% ce qui est techniquement facile
- Intuitivement il est plus facile de tester des valeurs car on sait que les percentiles à considérer et à tester seront entre 90 et 99% car par définition, une anomalie est associée à la notion d'exception. Par conséquent, tester des valeurs de percentiles trop basses (considérer par exemple que 50% de nos dates et événéments doivent ressortir en anomalies) n'a pas de sens.

Paramètre/cible	Nombre total d'anomalies identifiés	Proportion
5%	34 212	5.25%
3%	20 890	3.205%
1%	7 772	1.19%

TABLE 3.12 – Nombre total d'anomalies identifiés selon la cible

Prenons l'exemple du Nord-Ouest de Paris :

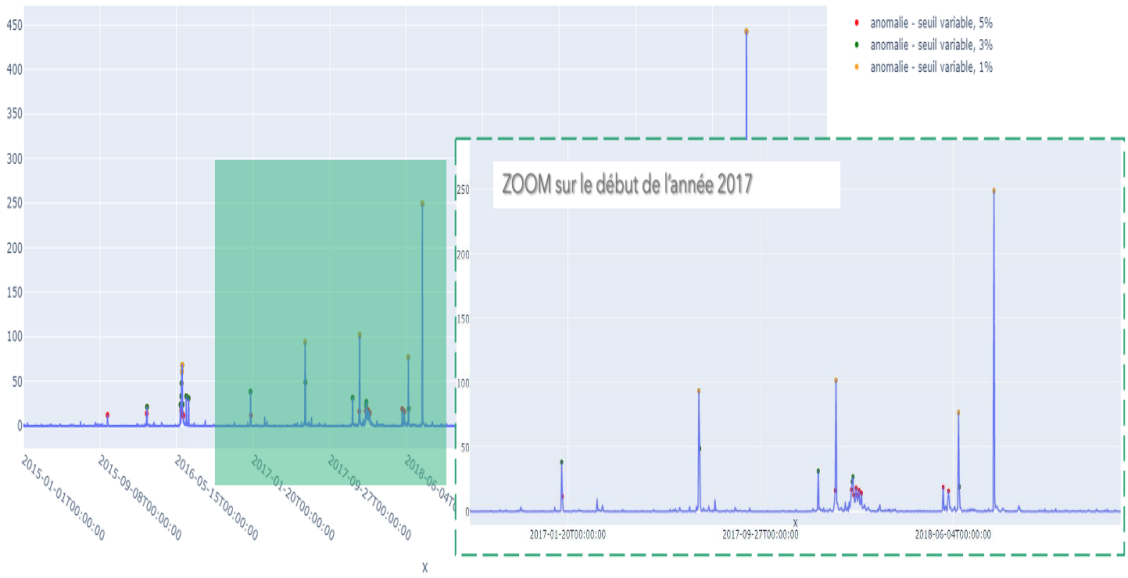


FIGURE 3.17 – Résultat de l'algorithme de détection d'anomalies (seuil variable, cible 5%) pour la zone autour de la ville de Paris entre 2015 et fin 2021

Evaluation des modèles

Nous sommes dans le cadre d'un apprentissage non-supervisé. De ce fait, notre jeu de données ne contient pas d'exemples labellisés. Dans cette configuration, il a fallu faire le travail à la main et comparer du mieux possible les individus-dates classifiés en anomalies avec les données déjà existantes d'AXA. On considère notre modèle performant à partir du moment où au moins 90% des anomalies identifiées semblent coïncider avec des EGA ayant eu lieu.

Dans notre base de données historique d'évènements recensés, un EGA est caractérisé par les variables de date de début, date de fin, le numéro d'évènement et le type d'évènement comme le montre la figure suivante :

evt_new	debut	fin	type_evt
1	2009-01-23	2009-01-23	TEMP.
2	2009-02-09	2009-02-11	TEMP.
3	2009-05-11	2009-05-13	GRELE

TABLE 3.13 – Aperçu de la base de données de recensement d'EGA

En comparant nos résultats, c'est-à-dire les anomalies détectées avec nos deux modèles de seuils, avec la base de données réelle d'EGA, on conclut que le Fixed Threshold Model n'est pas performant et que parmi les Adaptative Threshold Models, le plus performant semble être celui de paramètre 1%. Cela reste des méthodes assez déterministes.

Modèle	% d'anomalies coincidant avec des EGA
Fixed Threshold	68%
Adaptative threshold 1%	81%
Adaptative threshold 3%	78%
Adaptative threshold 5%	75%

TABLE 3.14 – Résultats modèles de seuil

Nous avons donc décidé de tester ensuite une solution davantage technique et qui ne dépend pas autant de choix d'hyperparamètres : CART (Classification And Regression Trees) et l'algorithme d'Isolation Forest.

3.6.4 Isolation Forest

Le modèle d'IsolationForest fait partie des algorithmes **CART** et est donc basé sur des **arbres de décisions**. Il s'agit d'un algorithme **non-supervisé**. [10]

Les techniques les plus couramment utilisées pour la détection d'anomalies sont basées sur la construction d'un profil de ce qui est «normal» : les anomalies sont dans ce cas signalées et identifiées comme des observations non conformes au profil normal. Cependant, il semble plus difficile d'isoler et identifier un point 'normal' plutôt qu'un point aberrant et anormal (cf l'image ci-dessous).

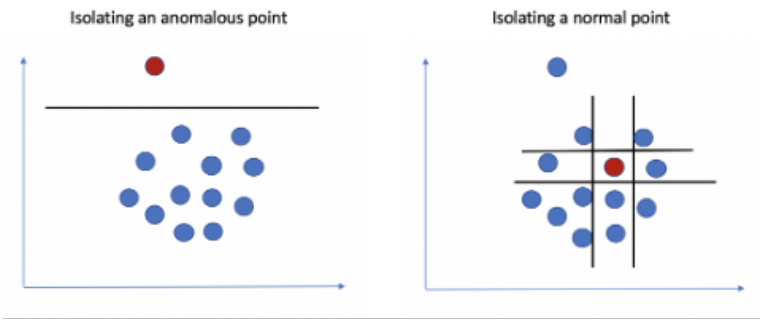


FIGURE 3.18 – Comparaison d'isolation entre un point normal et un point aberrant [27]

Algorithme

» **En entrée** : Soient les variables aléatoires $X_1, ..., X_n$ indépendantes, ϵ_Φ un échantillon aléatoire de X de taille Φ

» **En sortie** : On obtient un groupe de I arbres binaires.

- Si ϵ_Φ ne peut pas être divisé :
  - Alors retourner un noeud externe de taille Φ
- Sinon :
  - Sélectionner une variable aléatoirement
  - Découper cette variable aléatoirement selon un seuil de séparation (toute valeur dans la plage des valeurs minimum et maximum de la variable sélectionnée).
  - Répéter les deux étapes précédentes jusqu'à l'isolation d'une donnée.
  - Répéter les étapes précédentes de façon récursive. [21]

L'algorithme IsolationForest détecte les anomalies au sein d'une série temporelle en utilisant le principe d'isolation, c'est-à-dire en s'intéressant à la distance entre l'observation analysée et le reste des données ; au lieu d'essayer de créer un modèle d'instances « normales », elle isole et modélise explicitement les points anormaux dans l'ensemble de données.

Pour appliquer cet algorithme, on effectue un Traitement individuel c'est-à-dire que l'on applique l'algorithme tour à tour sur chacun de nos individus (zones) - on filtre nos données par individu et on les isole. On obtient donc la série temporelle de l'individu i et celle-ci constitue l'input de l'algorithme.

On calcule un score d'anomalie pour chaque observation de la série temporelle de l'individu i : il s'agit de la probabilité d'être une anomalie.

L'algorithme isole chaque date observée de manière récursive : il choisit une variable et fixe un seuil de coupure au hasard. Il évalue ensuite si cela permet d'isoler une date en particulier

Dans notre cas, nous utilisons pour l'instant une unique variable qui est le volume global d'UP. Une piste d'amélioration pourra être d'intégrer la charge dans le modèle.

L'algorithme effectue des découpes aléatoires dits des splits autour de chaque observation. Le nombre de splits nécessaires pour isoler une observation particulière donnée nous indique s'il s'agit ou non d'une anomalie. Plus le nombre de splits est faible et plus la probabilité est forte que notre observation, c'est-à-dire la date observée soit une anomalie : la probabilité d'être une anomalie (c'est-à-dire le score) est donc fonction du nombre de découpes aléatoires.

ET LES ARBRES DANS TOUT CA ? L'APPORT DE LA METHODE ENSEMBLISTE

Une des **limites** du processus de découpes aléatoires est la suivante : il est possible d'isoler à tort une des observations de la masse des données.

Pour palier au risque précédemment décrit, on peut exploiter **les arbres de décisions multiples**, c'est-à-dire les forêts aléatoires : on génère plusieurs arbres de décisions afin de bâtir un estimateur. Chaque arbre effecture une séquence aléatoire de découpes et calcule un score d'anomalies

On calcule ensuite la moyenne de l'ensemble des résultats, ce qui permet d' annuler les petites erreurs car la majorité des arbres/estimateurs va l'emporter : les méthodes ensemblistes garantissent davantage la robustesse des résultats.

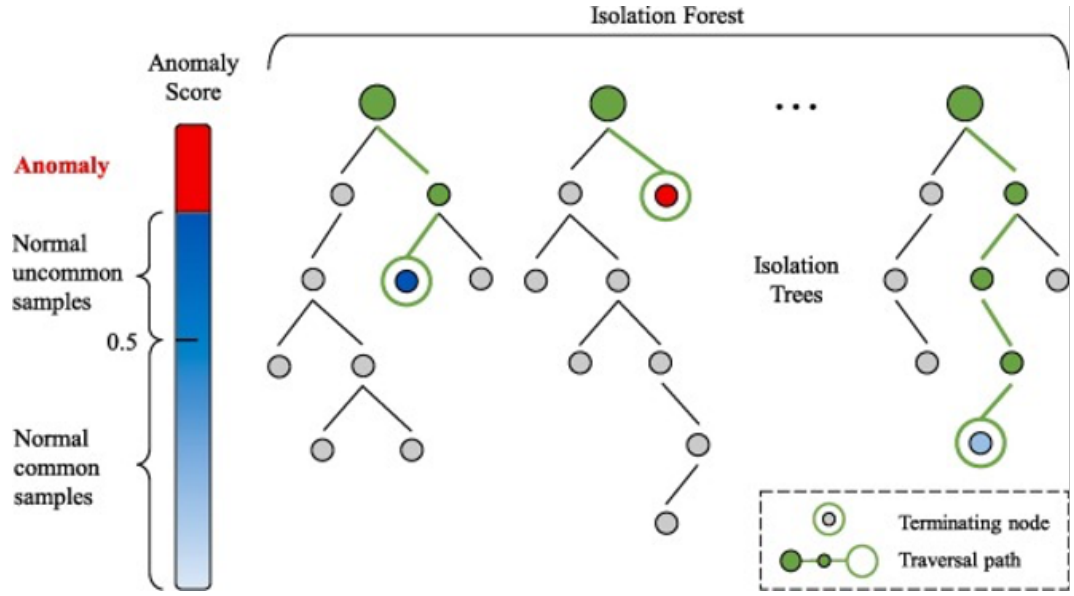


FIGURE 3.19 – Illustration de la méthode des forêts d'isolation [28]

Regardons par exemple ce que renvoie l'algorithme pour la région Ile de France.

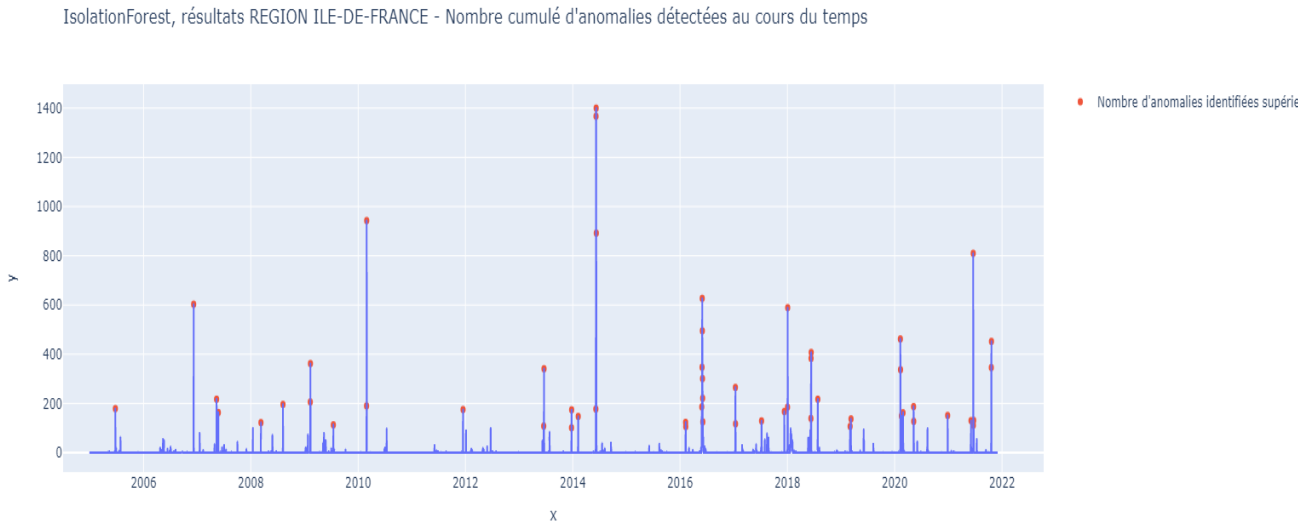


FIGURE 3.20 – Résultat Isolation Forest : Nombre cumulé d'anomalies détectées au cours du temps en Ile de France

Nous avons besoin de revenir à une forme d'output plus 'lisible' et intuitif afin d'exploiter ces résultats : on ne peut pas rester à la maille agrégée/nouveau référentiel car les données AXA initiales ainsi que les données d'apprentissage des EGA déjà existantes sont exprimées à une maille plus traditionnelle comme nous l'avons déjà vu, qui est le territoire français et son découpage en commune. Voici la forme de dataset produite :

DTSURV	EGA_isolationForest	code_commune_actualise	lieu_du_sinistre	code_postal_actualise	commune_ou_commune_de_rattachement	occurrence_up	merlin_chg
2017-07-10	NaN	71550	UCHIZY	71550	UCHIZY	8	27.04973
2020-10-20	NaN	43140	BLAVOZY	43140	LE MONTEIL	1	0.50010
2006-12-08	1.0	45137	ESCRENNES	45137	ESCRENNES	2	0.00000
2000-11-28	NaN	42095	FIRMINY	42095	FIRMINY	1	0.00000
2014-11-27	NaN	86066	CHATELLERAULT	86066	CHATELLERAULT	1	2.15700

TABLE 3.15 – Aperçu du nouveau catalogue créé incluant notre nouvelle donnée d'anomalies

Sur les près de 2 240 000 sinistres recensés, l'algorithme d'Isolation Forest détecte près de 960 000 anomalies pour 2 544 zones agrégées affectées. Plusieurs anomalies par zone détectées coïncident avec des évènements de notre base de données réelles d'évène-

ments ayant eu lieu. Cet algorithme nous permet d'obtenir une proportion d'anomalies coïncidant avec un EGA de notre base réelle d'évènements de 93%. Nous avons donc gardé ce modèle.

Chapitre 4

Méthode d'estimation de la charge à l'ultime : Modélisation par Machine Learning et MLG

Dans le chapitre 2, nous avons vu comment était calculée la charge à l'ultime pour les EGA à une maille agrégée, c'est-à-dire en ayant une vision par évènement (date de début, date de fin, type d'évènement et cadence d'ouverture). Cette méthode bien que simpliste et facile à reproduire manque terriblement de précision. En effet, prenons par exemple le cas d'un évènement qui dure 5 jours et dont le pic se trouve en début d'évènement, cela peut créer un décalage au niveau des cadences d'ouverture et donc fausser la règle de trois et in fine l'estimation de la charge à l'ultime. En plus de cela, l'atypisme des évènements fait que les règles de proportionnalité ne peuvent pas toujours être appliquées. C'est bien pour cela également que les méthodes classiques de provisionnement dans ce cas précis ne sont pas utilisées pour l'estimation de la charge à l'ultime.

De plus, dans un environnement affecté par le changement climatique, de nouvelles informations telle que la localisation géographique sont nécessaires, voire primordiales pour évaluer de façon plus précise les risques climatiques auxquels sont confrontés les entreprises de manière plus granulaire et sur le long terme. La figure 4.1 qui présente la carte de la France avec un indice de sinistralité par région nous montre que les évènements climatiques sont très localisés géographiquement et il est important de bien prendre en compte cette notion géographique si l'on veut avoir des estimations plus précises de la charge.

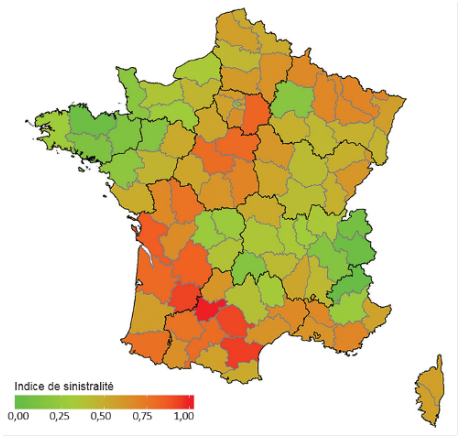


FIGURE 4.1 – Indice géographique de sinistralité en France métropolitaine hors DOM-TOM Source : Fondation pour l'innovation politique, données France Assureurs

Comme précisé dans le chapitre 2, le calcul de la charge à l'ultime dans le cadre des EGA repose sur 2 estimations fondamentales :

- **l'estimation des volumes de sinistres à l'ultime :** nous allons pour cela utiliser une approche par **Machine Learning**
- **l'estimation du coût moyen :** nous allons pour cela utiliser une approche par **MLG**

4.1 Analyse rapide de la base de données enrichie

La première partie du travail effectué notamment dans le chapitre 3 a consisté en la création d'un nouveau catalogue, d'une nouvelle base consolidée et fiable prenant en compte le critère géographique afin de l'exploiter et de prédire au mieux la charge à l'ultime.

Nous avons pu enrichir notre base de données d'évènements en prenant en compte les variables géographiques. On se retrouve désormais avec une nouvelle base de données contenant un peu plus de 960 000 lignes initialement correspondant à un peu plus de 32 000 communes (représentées par leurs codes communes) où on a détecté une anomalie, donc un probable EGA entre début 2000 et fin 2021 (possibilité donc de retrouver sur plusieurs lignes différentes une même commune). Pour chaque code commune, on a également la variable relative aux zones agrégées auxquelles chaque code commune est rattaché.

On s'est vite rendu compte qu'à la maille commune, on manquait fortement d'exposition en terme de sinistres. Nous sommes donc passés à la maille zone agrégée en utilisant le découpage par zone effectué dans le chapitre précédent afin d'avoir une meilleure exposition. On s'est donc retrouvé avec une base contenant 2544 zones.

code_commune_actualise	nbsin
01001	3
01004	133
01005	23
01007	11
01008	5
...	...
95676	16
95678	10
95680	34
95682	13
95690	2

32339 rows × 1 columns

TABLE 4.1 – Nombre de sinistres par commune

zones_agregees_dappartenance	nbsin
85184083ffffff, 85184093ffffff_8518409bffffff_85184097ffffff_8518442bffffff_8518442fffffff_85184477ffffff_8518443bffffff_85184423ffffff	24
85184093ffffff, 85184083ffffff_8518409bffffff_85184097ffffff_8518442bffffff_8518442fffffff_85184477ffffff_8518443bffffff_85184423ffffff	43
85184097ffffff, 85184093ffffff_85184083ffffff_8518442fffffff_8518409bffffff_8518442bffffff_85184423ffffff_8518443bffffff_85184477ffffff	10
8518409bffffff, 85184083ffffff_85184093ffffff_85184097ffffff_8518442bffffff_8518442fffffff_85184477ffffff_85184463ffffff_8518443bffffff	14
85184403ffffff, 85184413ffffff_8518440ffffff_85184407ffffff_8518441bffffff_85184417ffffff_8518440bffffff	319
...	...
853975abffffff, 853975bbffffff_853975a3ffffff_853975b3ffffff_853975b7ffffff_853975affffff_853975a7ffffff_85397587ffffff	13
853975affffff, 853975a7ffffff_853962dbffffff_853962d3ffffff_853975a3ffffff_8539666bffffff_853975b7ffffff_853975abffffff_853962c3ffffff_8539666fffffff	97
853975b3ffffff, 853975bbffffff_8539664fffffff_853975a3ffffff_8539664bffffff_85397587ffffff	312
853975b7ffffff, 853975a3ffffff_8539667bffffff_853975a7ffffff_8539664fffffff_8539666bffffff	602
853975bbffffff, 853975abffffff_853975b3ffffff_853975a3ffffff_85397587ffffff_8539664bffffff_853975b7ffffff	169

2544 rows x 1 columns

TABLE 4.2 – Nombre de sinistres par zone

Maintenant qu'on a cette base, nous allons réaliser une analyse descriptive et exploratoire rapide de celle-ci en rajoutant de nouvelles variables explicatives pour l'estimation

de la charge à l'ultime. Cela va notamment nous permettre de voir des éventuelles corrélations entre les variables.

4.1.1 Ajout de nouvelles variables

Pour rappel, nous avions une base de données sinistres contenant le code commune actualisé où a eu lieu le sinistre, la zone agrégée à laquelle il appartient, la date de survenance du sinistre. Nous avons ainsi rajouté des variables qui nous semblent pertinentes pour la suite de notre modélisation. Nous avons rajouté quelques variables telles que le type d'évènement qui n'était pas pris en compte dans la modélisation des anomalies, (Voir ensemble des variables dans l'annexe).

Variables	Type	Description
DTOUV	DateTime	Date d'ouverture du sinistre
DTCLOT	DateTime	Date de clôture du sinistre
cdentite	Object	Entreprise ou Pro/Particulier
niv2	Object	Branche d'activité
part_auto	Float	% de sinistres auto
part_dommages	Float	% de sinistres dommages
part_transport	Float	% de sinistres transport
part_constructions	Float	% de sinistres constructions

TABLE 4.3 – Descriptif de quelques variables rajoutées

4.1.2 Quelques visualisations

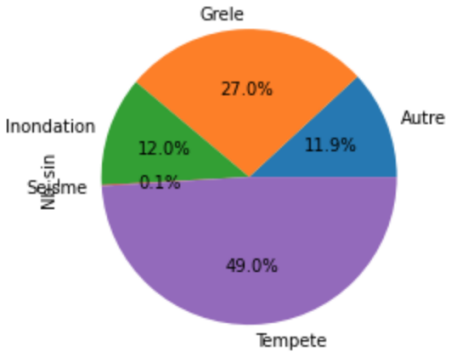


FIGURE 4.2 – Répartition du volume de sinistres par type climatique

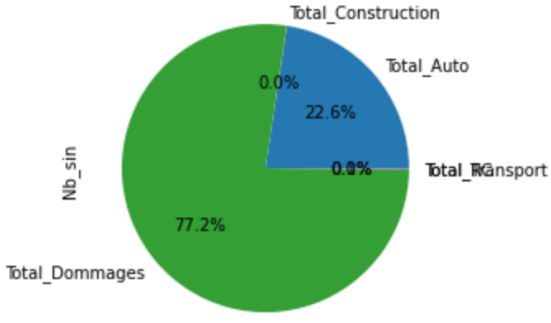


FIGURE 4.3 – Répartition du volume de sinistres par branche

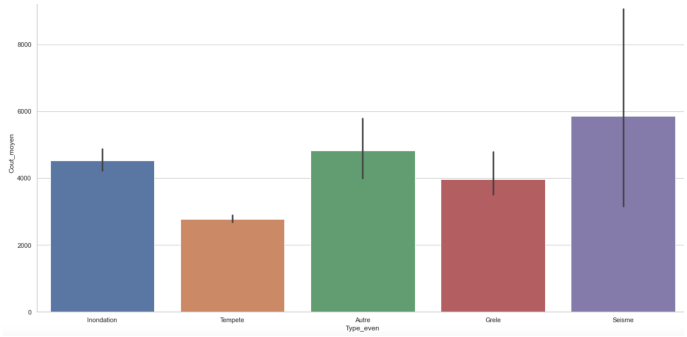


FIGURE 4.4 – Répartition du coût moyen par type climatique

Notre base est donc constituée de différentes variables dont les variables utilisées dans le chapitre 2. Elle est toujours constituée des UP grêle, tempête, inondation et séisme. Elle contient notamment les branches d'activité les plus impactées que sont les branches dommages avec près de 83% de la charge et l'auto avec un peu plus de 16% de la charge. Chaque ligne est caractérisée par une commune pour laquelle on a détecté une anomalie, les communes concernées ainsi que le nombre de sinistres comptabilisés.

4.2 Approche par Machine Learning pour le volume

Le Machine Learning ou apprentissage automatique se caractérise par un ensemble de méthodes basées sur des modèles mathématiques et statistiques qui donnent aux ordinateurs la capacité d'apprendre et de résoudre des tâches sans être spécifiquement programmés pour le faire. Contrairement aux algorithmes traditionnels qui appliquent des règles prédéterminées, le machine learning apprend et fabrique ses propres règles.

Au cours des dernières années, avec notamment l'essor du big data, le machine learning a pris une toute autre dimension, les entreprises ayant désormais accès à un grand nombre de variables différentes qui, lorsqu'elles sont corrélées, peuvent être des atouts décisionnels extrêmement puissants. Ces prédictions permettent de formuler une hypothèse avec une marge d'erreur de plus en plus étroite.

Les compagnies d'assurance se sont progressivement adonné au Machine Learning en l'appliquant à leurs modèles d'assurance. En effet, l'apprentissage automatique aura notamment permis de créer des corrélations entre de grandes quantités de données et d'évènements, ou encore de réaliser des analyses prédictives fiables pouvant challenger les méthodes déjà existantes.

En Machine Learning, on retrouve des algorithmes d'**apprentissage supervisé** et des algorithmes d'**apprentissage non supervisé**.

En apprentissage supervisé, la machine travaille à partir de données étiquetées. C'est-à-dire qu'elle a déjà connaissance des réponses des valeurs à prédire. On cherche donc à déterminer une fonction f qui permet de prédire la variable cible Y déjà connue à partir de variables explicatives X.

En apprentissage non supervisé par contre, les données ne sont pas étiquetées. La machine propose alors des réponses à partir d'analyses préalables et/ou de regroupements de données. Le but étant par exemple de créer des classes avec une certaine homogénéité.

Dans le cadre de l'estimation de la charge à l'ultime, nous nous intéressons à l'apprentissage supervisé. En effet, on cherche à prédire une variable cible Y qu'on connaît déjà à partir de variables explicatives X comme le montre la figure 4.5. Si la variable cible est qualitative, on parle de classification, si elle est quantitative, on parle de régression. La variable cible représente ici la charge à l'ultime brute de réassurance pour un évènement donné. Il s'agit donc d'une régression.

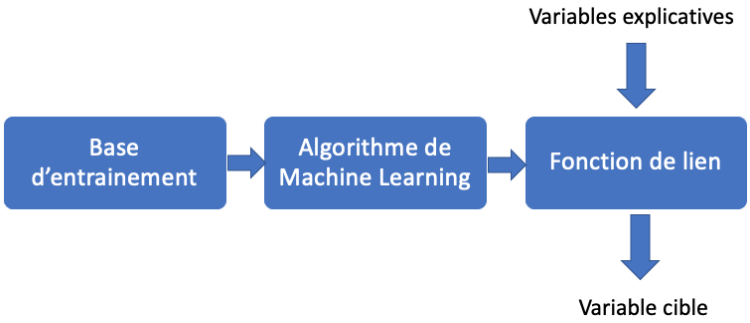


FIGURE 4.5 – Processus Apprentissage Supervisé

On cherche à estimer la fonction de lien qui permet de bien prédire la variable cible Y. L'utilisation et la justification d'un modèle de Machine Learning passe donc par plusieurs étapes essentielles.

4.2.1 Etapes d'évaluation d'un modèle de Machine Learning

La figure 4.3 ci-dessous montre les différentes étapes de l'évaluation d'un modèle de Machine Learning.

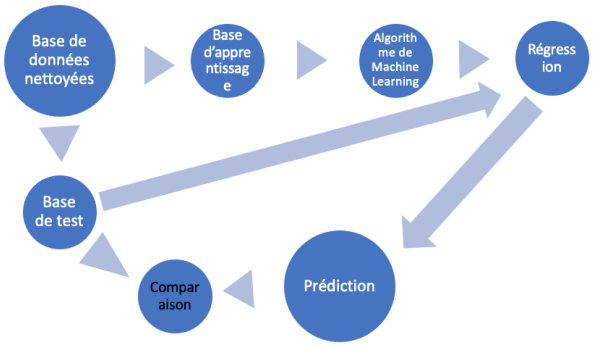


FIGURE 4.6 – Etapes de l'évaluation d'un modèle de Machine Learning

La première étape est donc d'avoir une base de données nettoyée. On a effectué une analyse rapide de notre base de données dans la section précédente.

On divise ensuite l'ensemble de données obtenu en deux parties :

- Base d'entrainement : ensemble de données utilisées par l'algorithme pour apprendre le modèle. Il se compose de la majorité des données pour espérer le meilleur résultat.
- Base de test : ensemble de données utilisées pour valider le modèle implémenté par l'algorithme. Il se compose d'une petite partie des données.

Le modèle est construit uniquement à l'aide de la base d'entraînement. A partir de ça, on fait une prédiction de la variable cible Y de la base de test que nous comparons aux données réelles connues au niveau de la base de test. Une fois la prédiction faite, il faut évaluer la performance du modèle. En fonction des résultats, certains paramètres peuvent devoir être modifiés pour rendre le modèle plus efficace. On peut alors faire une prédiction sur la base de nouvelles données.

Ces différentes étapes seront davantage détaillées dans la suite de cette étude.

Maintenant qu'on a établi les principales étapes d'évaluation d'un modèle de Machine Learning, nous allons parler des différents modèles de machine learning utilisés pour l'estimation du volume de sinistres à l'ultime.

4.2.2 KNN : K-Nearest Neighbors

La méthode actuelle d'estimation du volume de sinistres à l'ultime se base sur une estimation de l'évènement le plus proche de celui à estimer. L'algorithme que nous avons commencé à tester et qui au premier abord se rapproche de cette méthode de calcul est

celui des K plus proches voisins. L'algorithme des k plus proches voisins ou K-nearest neighbors (kNN) fait partie des algorithmes d'apprentissage supervisé, c'est-à-dire capables de prédire une variable Y à partir de variables d'entrée annotées X. [26] .

La première étape consiste en une sélection du nombre de voisins. Ce choix est essentiel pur le bon fonctionnement de l'algorithme. Le plus proche voisin est déterminé à partir d'une distance arbitraire d(.,.).

Soit L l'ensemble des données de l'échantillon d'apprentissage :

$$L = \{(y_i, x_i), i = 1, ..., n_L\} \qquad y_i \in \{1, ..., c\} \; et \; \mathrm{x}_i = (x_{i1}, \; ..., \; x_{ip})$$

avec y_i la variable cible pour l'individu i et le vecteur x_i les variables de prédiction de l'individu i. Soient deux individus caratérisés par p covariables, la distance euclidienne entre ces deux individus est définie par :

$$d((x_1, x_2, ..., x_p), (u_1, u_2, ..., u_p)) = \sqrt{(x_1 - u_1)^2 + (x_2 - u_2)^2 + ... + x_p - u_p)^2}$$

Il existe d'autres mesures de distances :

- la distance de Manhattan : d(x,y)= $\sum_{i=1}^{m} |x_i - y_i|$
- la distance de Minkowski : d(x,y)= $(\sum_{i=1}^{n} |x_i - y_i|)^{\frac{1}{p}}$

En fonction de la distance trouvée et du nombre de k voisins choisis initialement, on détermine les k plus proches voisins. Notre kNN étant utilisé pour de la régression, c'est la moyenne des variables des K plus proches observations qui sera utilisée pour la prédiction.

4.2.3 Forêts aléatoires (Random forest)

Nous avons ensuite testé l'algorithme des forêts aléatoires. Introduit en 2001 par Léo Breiman [7], l'algorithme des forêts aléatoires est un algorithme très performant d'apprentissage statistique de problèmes de classification. La définition donnée par Breiman est la suivante :

« *Soit { $\hat{h}$(.,θ_1), ..., $\hat{h}$(.,θ_q) } une collection de prédicateurs par arbre, où (θ_1, ..., θ_q) est une suite de variables aléatoires i.i.d., indépendantes de l'échantillon d'apprentissage*

L_n. Le prédicateur des forêts aléatoires est obtenu par agrégation de cette collection de prédicateurs. » [19]

L'algorithme des forêts aléatoires permet d'agréger les prédictions de chacun des classifieurs et prédire la classe d'une nouvelle observation grâce au Bagging. Le Bagging est une combinaison de Bootstrap et de Aggregating : **Bagging** = **B**ootstrap + **agg**regat**ing**. [14]

Le principe du bootstrap est similaire à la statistique classique. Nous avons un échantillon qui est composé de n variables aléatoires (X_1, ..., X_n) indépendantes et identiquement distribuées de fontion de répartition inconnue F. Nous devons estimer la loi de T_n= T(X_1, ..., X_n). T_n est un estimateur d'une grandeur θ telle que :

$$\theta := \mathrm{T(F)}$$

T_n permet de calculer la variance, le biais et les intervalles de confiance. Pour estimer ces différents indicateurs, il faut remplacer F par la fonction de répartition empirique F_n que l'on obtient à partir de l'échantillon [6] :

$$F_n(x) = \frac{1}{n}\sum_{k=1}^{n} \mathbf{1}_{\mathbf{X_k \leq x}} \qquad \mathbf{x} \in \mathbb{R}$$

En machine learning, pour maximiser les performances en prédiction, les méthodes d'ensemble utilisent plusieurs algorithmes d'apprentissage. L'Aggregating consiste à regrouper un ensemble de valeurs selon un ou plusieurs principes d'agrégation. L'agrégation est la méthode utilisée pour former des agrégats.

Ainsi, le Bagging est une méthode qui permet d'améliorer la stabilité et la précision des algorithmes. Le Bagging permet d'utiliser un même algorithme d'entraînement pour chaque prédicteur, mais en l'entraînant sur des sous-ensembles différents extraits aléatoirement du jeu de données. Le tirage s'effectue avec remise. Ensuite, on construit un arbre CART sur cet échantillon. CART signifie Classification And Regression Trees. Le principe de CART est de diviser de manière récursive une partition d'entrée en sous partition optimale pour la prédiction. La partition d'entrée correspond à la racine de l'arbre et les sous-partitions correspondent aux deux noeuds fils obtenus par la première découpe, et ainsi de suite.

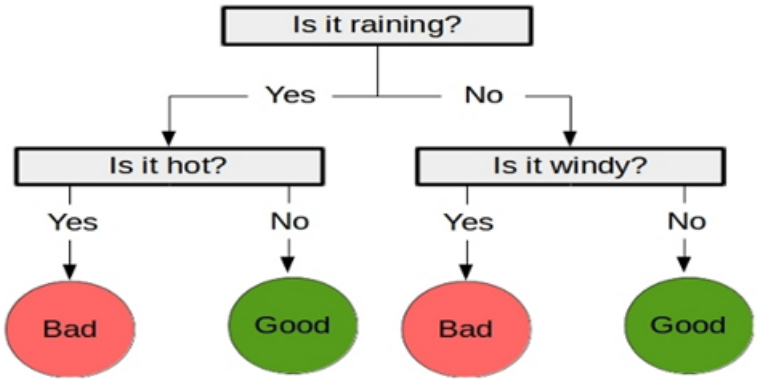


FIGURE 4.7 – Exemple modèle de CART binaire [2]

4.2.4 Gradient Boosting

L'algorithme Gradient Boosting est un cas particulier de Boosting. Le boosting est une technique basée sur des ensembles qui consiste en des classificateurs (modèles) agrégés développés séquentiellement sur des échantillons d'apprentissage, dont les poids individuels sont progressivement corrigés. Les classificateurs sont pondérés en fonction de leurs performances.

Cette technique de boosting est principalement utilisée pour les arbres de décision. L'idée principale est d'agréger à nouveau plusieurs classificateurs, mais de les créer de manière itérative. Ces "petits classificateurs" sont généralement de simples fonctions paramétrées, généralement des arbres de décision, où chaque paramètre est un critère de division de branche. Les super-classificateurs finaux sont les poids (via un vecteur w) de ces petits classificateurs.

L'objectif est trouver la valeur approximative f(x) avec une somme pondérée des "weak learners" c'est-à-dire les modèles avec des performances faibles : [12]

$$\text{f(x)} = \sum_{i=1}^{M} \beta_m h(x, \theta_m)$$

avec h(x,θ) qui correspond aux weak learners et θ_m qui représente l'ensemble des paramétres qui caractérisent la fonction h.

Nous venons de voir les différents algorithmes de Machine Learning utilisés. Nous allons maintenant passer à la modélisation.

4.2.5 Construction de la base avec cadences d'ouverture

On cherche à projeter des cadences d'ouvertures de sinistres, donc on doit à partir des informations contenues dans notre base de donnée initiale et construire ces cadences sur le même principe de construction que la méthode d'AXA, c'est-à-dire en prenant en compte les jours fériés et les week-ends. Ainsi, à partir de l'information sur le nombre de sinistres ainsi que la date d'ouverture des sinistres, on construit nos cadences cumulées. Comme énoncé en introduction de ce chapitre, se limiter à une vision par date de début et date de fin peut donner des résultats qui ne reflètent pas assez bien la réalité. Ainsi, regarder un évènement par jour de survenance pourrait être une première solution permettant de projeter au mieux les cadences d'ouverture et d'avoir une meilleure estimation de la charge à l'ultime.

De plus, on souhaite être à mesure d'estimer le volume final des sinistres en début d'évènement. Donc pour le modèle, on se limitera à la construction de cadences jusqu'à 3 jours après ouverture des sinistres. (cf. table. 4.4)

DTSURV	JourSurv	month	Year	dureeEvenement	part_auto	part_transport	part_RC	part_dommages	part_constructions	J1	J2	J3	JFinal
2009-01-23	Friday	1	2009	2 days	0.188679	0.000000	0.0	0.995631	0.000000	201.0	332.0	433.0	1605.0
2009-01-24	Saturday	1	2009	2 days	0.039474	0.000000	0.0	0.999460	0.000000	3119.0	6931.0	10521.0	79286.0
2009-01-25	Sunday	1	2009	2 days	0.121951	0.000000	0.0	0.998309	0.000000	115.0	162.0	180.0	494.0

TABLE 4.4 – Aperçu de la base finale de construction des cadences d'ouverture

4.2.6 Décomposition de la base de données en base d'apprentissage et base de test

On souhaite décomposer le jeu de données en 2 groupes : les données pour l'apprentissage et les données pour les tests. L'important ici est de définir une base d'apprentissage ou base d'entrainement qui regroupe le maximum d'informations et représente assez bien nos données afin que le modèle puisse bien apprendre. Par conséquent, il est nécessaire de prendre en compte toutes les particularités susceptibles d'altérer les performances de notre modèle tels que les effets de la saisonnalité. Ainsi, pour ce mémoire, on considère qu'une base d'apprentissage de 70% des données est assez représentative, et donc 30% des données correspondront aux données de test.

4.2.7 Calibrage par validation croisée ou Cross Validation

En Machine Learning, on cherche souvent à calibrer nos modèles avant de les entraîner afin d'avoir les paramètres optimaux avant l'entraînement sur l'ensemble de la base d'apprentissage. Nous utilisons pour cela la technique de validation croisée ou Cross Validation.

La validation croisée est une méthode statistique permettant d'évaluer et de comparer des algorithmes d'apprentissage. Elle consiste à diviser les données en deux parties : une pour l'apprentissage du modèle et une pour valider le modèle. Dans une validation croisée typique, les ensembles d'apprentissage et de validation doivent être croisés lors de cycles consécutifs afin que chaque point de données ait une chance d'être validé. La forme de base de la validation croisée est la validation croisée k-fold. D'autres formes de validation croisée sont des cas particuliers de validation croisée k-fold ou impliquent des cycles répétés de validation croisée k-fold.

Dans la validation croisée, les données sont d'abord divisées en k segments de taille égale. Ensuite, k itérations d'apprentissage et de validation sont effectuées de telle sorte qu'à chaque itération, les données de différents plis sont conservées pour validation et les k-1 plis restants sont utilisés pour l'apprentissage.

4.2.8 Entraînement des modèles sur la base d'apprentissage

Une fois le calibrage réalisé et nos paramètres optimisés grâce à la validation croisée, on peut enfin procéder à l'entraînement des modèles sur notre échantillon d'apprentissage.

4.2.9 Résultats des modèles et calculs des erreurs d'estimation

La dernière étape consiste ensuite à analyser les résultats des modèles en traçant par exemple le graphe des résidus. Pour l'évaluation de la performance des modèles et le calcul des erreurs, il existe plusieurs métriques d'erreurs que l'on peut utiliser. Nous allons utiliser 2 métriques d'erreur souvent utilisées pour évaluer l'efficience des modèles :

- le RMSE : Root Mean Squared Error
- le MAPE : Mean Absolute Percentage Error

RMSE : Root Mean Squared Error

Il s'agit de la racine carrée de l'erreur quadratique moyenne. Elle est souvent utilisée pour mesurer l'écart des résidus, c'est-à-dire l'écart entre la valeur observée et la valeur estimée.

$$RMSE = \sqrt{\frac{1}{n}\Sigma_{i=1}^{n}\Big(y_i - \hat{y_i}\Big)^2}$$

où n représente la taille de l'échantillon, y est la valeur réelle, $\hat{y_i}$ la valeur prédite par le modèle et n la taille de l'échantillon.

MAPE : Mean Absolute Percentage Error

Le MAPE mesure le pourcentage de l'erreur absolue moyenne. Elle permet donc de calculer en valeur absolue la moyenne arithmétique des écarts entre les valeurs prédites et les valeurs réelles.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y_i}|}{y_i}$$

où n représente la taille de l'échantillon, y_i la valeur observée, et $\hat{y_i}$ la valeur prédite.

Le MAPE se distingue du RMSE par son interprétabilité plus simple. En effet, le RMSE ne donne pas de grande indication sur ce que fait le modèle tel qu'un pourcentage d'erreur ou un intervalle de confiance.

4.2.10 Résultats

On compare l'exactitude des prévisions grâce au RMSE et au MAPE pour les risques tempête, inondation, grêle et séisme.

	KNN	Random Forest	Gradient Boosting
RMSE	6287.7403	5827.0283	5681.2081
MAPE	0.6150	0.2503	0.2253

TABLE 4.5 – Evaluation de la performance des différents modèles selon le MAPE et le RMSE

Le modèle qui a le RMSE le plus faible et le MAPE le plus faible est le gradient boosting. Celui-ci permet de donner une prédiction des volumes à l'ultime avec une marge d'erreur de plus ou moins 22%. Cette marge d'erreur peut paraître énorme, néanmoins cela reste un résultat plutôt satisfaisant quand on connaît la forte volatilité autour des EGA.

Analysons l'apport de la variable géographique dans notre meilleur modèle :

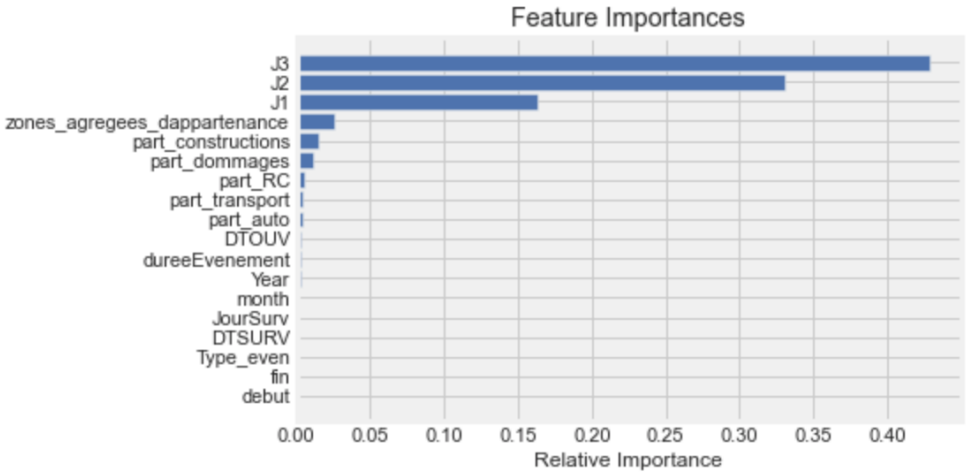


FIGURE 4.8 – Importance des variables avec la modélisation par gradient boosting

Comme on le voit sur la figure ci-dessus, la variable géographique représentée par la variable "zones_agregees_dappartenance" arrive en 4ème position après les variables "J1", "J2" et "J3". Le positionnement de ces 3 variables n'est pas surprenant vu qu'il s'agit des premières cadences d'ouverture qui sont donc très représentatives de la ca-

dence finale. Bien que beaucoup moins importante que ces 3 premières variables, la zone géographique a quand même eu son importance dans le modèle.

Passons maintenant à l'estimation du coût moyen.

4.3 Approche par MLG pour le coût moyen

Un des modèles les plus couramment utilisés pour l'estimation du coût moyen à l'ultime est le modèle linéaire généralisé.

Les MLG ou modèles linéaires généralisés sont des modèles très répandus et font partie de la grande famille exponentielle. Ils sont souvent utilisés pour l'estimation du coût moyen. Ils sont une continuité des modèles linéaires classiques dont on rappelle l'équation :

$$Y_i = \sum_{i=1}^{n} \alpha_i X_i + \epsilon_i$$

où Y_i est notre variable cible, α_i les paramètres de notre modèle, X_i les variables explicatives du modèle, ϵ_i les erreurs liées à l'écart entre les valeurs prédites et les valeurs observées.

Dans les modèles linéaires classiques, la variable à prédire est à valeur réelle et requiert une condition de normalité au niveau de ϵ_i. C'est ainsi qu'interviennent les modèles linéaires généralisés dont l'avantage est de passer outre cette condition.

Ainsi, un modèle MLG est un modèle permettant de prédire une variable cible Y est caractérisé par :

- une fonction de lien de la forme :

  $$g(E(y)) = \sum_{i=1}^{n} \alpha_i X_i + \epsilon_i, \text{ avec } E(y) \text{ l'espérance de } Y$$

- une fonction d'erreur qui est caractéristique d'une famille de distribution donnée (Log-normale, GAMMA, Poisson etc)

Pour notre modélisation, nous avons testé les 2 lois généralement utilisées pour le calcul du coût moyen : la loi log-normale et la loi Gamma.

4.3.1 La loi log-normale

Considérons X une variable aléatoire à valeurs dans $\mathcal{R}_*^+$ suivant la loi log-normale de moyenne μ et d'écart type σ : $Y \sim \mathcal{LN}(\mu, \sigma^2)$

Soit $X = \exp(Y)$, alors on a :

$E(X) = \exp(\mu + \frac{\sigma^2}{2})$

et

$V(X) = \exp(2\mu + \sigma^2)(\mathrm{e}^{\sigma^2} - 1)$

4.3.2 la loi de Gamma

La loi Gamma fait partie de la famille exponentielle. Sa fonction de densité est donnée par :

$$f(x) = \begin{cases} \frac{\lambda e^{-\lambda x}(\lambda x)^{\beta-1}}{\phi(\beta)} & \geq 0 \\ 0 & < 0 \end{cases}$$

avec

$\phi(x) = P(X \leq x) = \int_{-\infty}^{x} f(u)du$

et λ et β les paramètres de la loi GAMMA

4.3.3 Critères pour le choix de la loi

Déviance

La déviance est caractérisée par une variation de la log-vraisemblance. Pour rappel, la log-vraisemblance est définie par :

$logL(\Theta|y_1, ..., y_n) = \sum_{i=1}^{n} logL(\Theta|y_i)$

où $y_1, ..., y_n$ des variables aléatoires indépendantes et Θ le paramètre de la loi

Critère AIC

Le critère d'information d'Akaike ou AIC est un des critères utilisés pour comparer la qualité d'ajustement entre deux modèles. Il se définit comme suit :

$AIC = -2log(L) + 2p$

où $log(L)$ est la log-vraisemblance, et p le nombre de paramètres.

Critère BIC

Le critère d'information bayésien ou BIC est également un des critères utilisés pour comparer la qualité d'ajustement entre deux modèles. Il se définit comme suit :

$BIC = -2log(L) + p.log(n)$

où $log(L)$ est la log-vraisemblance, p le nombre de paramètres et n le nombre d'individus dans l'échantillon.

4.3.4 Base de données utilisée

Il s'agit de la même base de données que celle utilisée pour la construction des cadences d'ouvertures. Sauf qu'on s'intéresse cette fois-ci aux coûts moyens. On calcule donc un coût moyen par zone et par type d'évènement comme le montre le tableau ci-dessous :

zones_agregees_dappartenance	Type_even	Cout_moyen
['85184093fffffff, 4477fffffff__8518443bfffffff__85184423fffffff']	Inondation	4132.280000
	Tempete	2301.178571
['85184097fffffff, 423fffffff__8518443bfffffff__85184477fffffff'\n '85184093fffffff, 477fffffff__8518443bfffffff__85184423fffffff'\n '85184083fffffff	Inondation	250.925000
	Tempete	2666.936667

TABLE 4.6 – Aperçu de la base pour le coût moyen

4.3.5 Résultats

On calcule un coût moyen par type d'évènement.

Tempête

Pour le coût moyen des tempêtes, on obtient la distribution suivante qu'on compare avec la loi choisie :

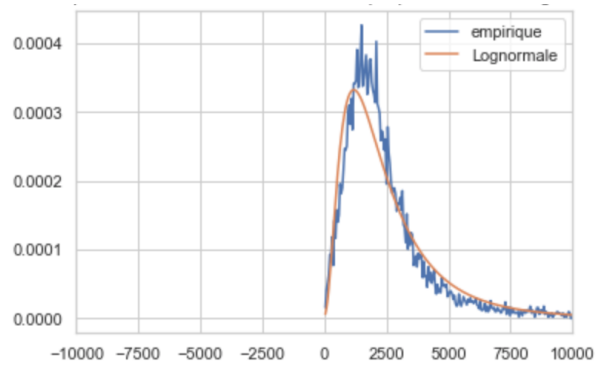


FIGURE 4.9 – Comparaison de la fonction de densité empirique avec une loi log-normale pour la tempête

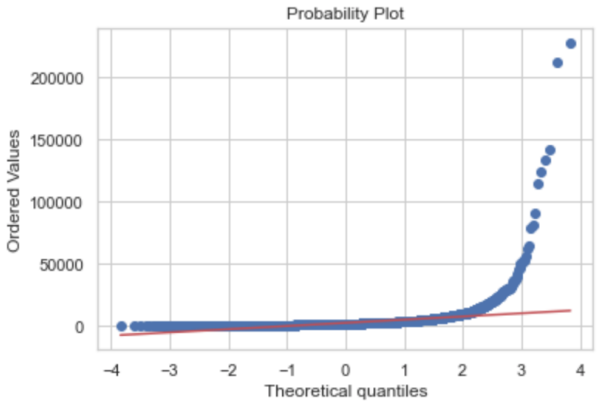


FIGURE 4.10 – Probability plot tempête

On a retenu la loi log-normale qui présente une bonne qualité d'ajustement pour les faibles montants bien que moins précise pour les valeurs se trouvant en queue de distribution. De plus, sur le plan statistique, le tableau ci-dessous nous indique que le modèle log-normal est mieux ajustable aux données de coûts moyens que le modèle GAMMA car ayant des valeurs de déviance et de critère AIC, BIC plus faibles.

	Déviance	Critère AIC	Critère BIC
Log-normale	56445.6696	57994.1779	58000.9267
Gamma	57451.0899	58873.3999	58880.1487

TABLE 4.7 – Tableau récapitulatif des critères de validation de modèles classiques pour la tempête

Grêle

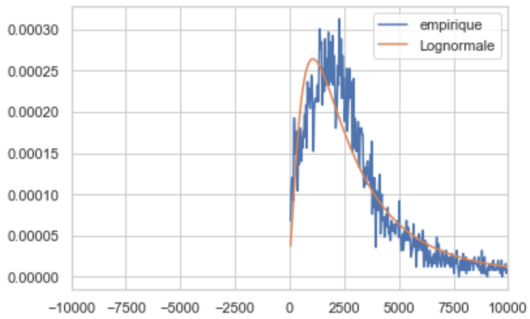


FIGURE 4.11 – Comparaison de la fonction de densité empirique avec une loi log-normale pour la grêle

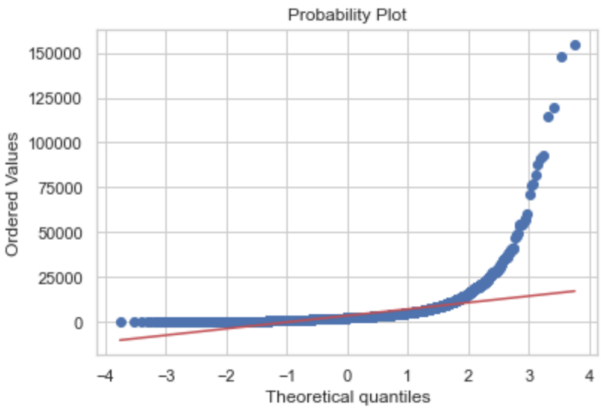


FIGURE 4.12 – Probability plot grêle

	Déviance	Critère AIC	Critère BIC
Log-normale	56301.2447	57868.8880	57875.6368
Gamma	57534.9552	58918.4484	58925.1972

TABLE 4.8 – Tableau récapitulatif des critères de validation de modèles classiques pour la grêle

De la même manière que pour la tempête, on retient la loi log-normale pour la grêle.

Inondation

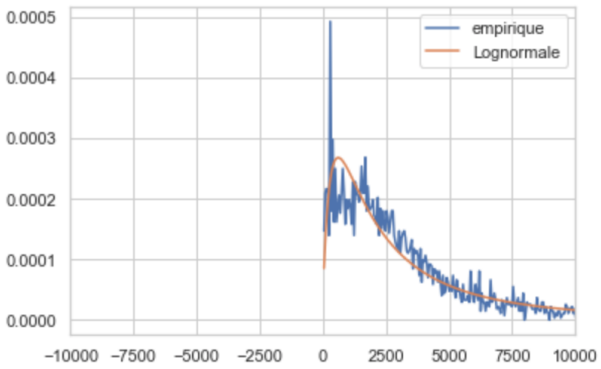


FIGURE 4.13 – Comparaison de la fonction de densité empirique avec une loi log-normale pour l'inondation

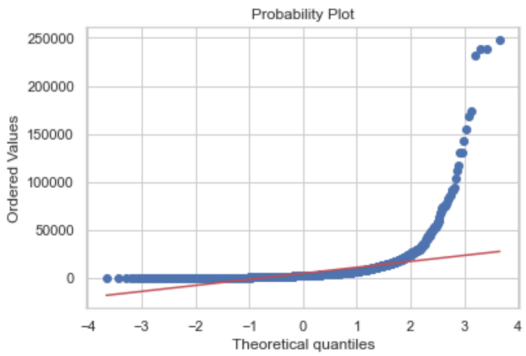


FIGURE 4.14 – Probability plot inondation

	Déviance	Critère AIC	Critère BIC
Log-normale	56099.2552	57699.2426	57705.9914
Gamma	57562.6624	58949.6333	58956.3821

TABLE 4.9 – Tableau récapitulatif des critères de validation de modèles classiques pour l'inondation

De la même manière que pour la tempête et la grêle, on retient la loi log-normale pour l'inondation.

Séisme

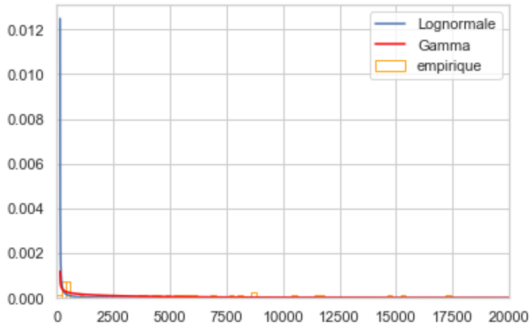


FIGURE 4.15 – Comparaison de la fonction de densité empirique avec une loi GAMMA pour les séismes

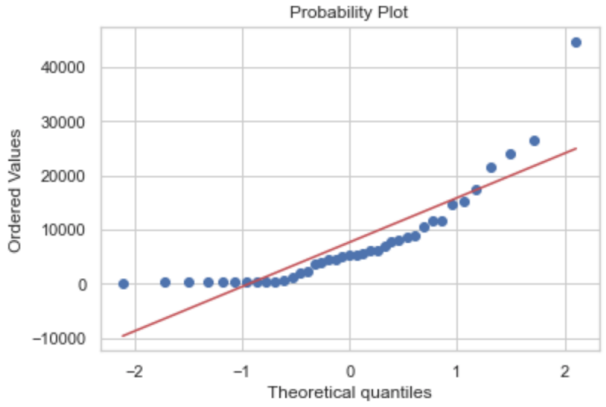


FIGURE 4.16 – Probability plot séisme

	Déviance	Critère AIC	Critère BIC
Log-normale	56113.2819	57648.5432	57655.1092
Gamma	57372.9053	58774.2175	58780.7738

TABLE 4.10 – Tableau récapitulatif des critères de validation de modèles classiques pour l'inondation

Pour le séisme également, on retient la loi log-normale.

4.3.6 Evaluation de la performance du modèle choisi

Comme expliqué précédemment, nous avons choisi la loi log-normale pour notre MLG. Tout comme pour l'estimation du volume, nous allons utiliser le MAPE comme métrique d'erreur pour mesurer la qualité de prédiction du modèle. Voici un tableau récapitulatif des résultats obtenus.

	Tempête	Grêle	Inondation	Séisme
MAPE	0.0957	0.0937	0.0965	0.0971

TABLE 4.11 – Evaluation de la performance du modèle selon le MAPE

Ainsi, le MAPE nous dit que notre coût moyen à l'ultime est assez bien prédit avec un intervalle de confiance autour de +/- 10% pour les 4 risques.

L'estimation de la charge à l'ultime pour un évènement donné se fera donc en réalisant le produit entre l'estimation du volume à l'ultime faite en section 4.2 et l'estimation du coût moyen qui vient d'être faite. On va calculer selon le MAPE :

$Intervalle\ charge\ ultime = [cout\ moyen\ min * volume\ sinistres\ min,\ cout\ moyen\ max * volume\ sinistres\ max]$

En appliquant les modèles retenus à la tempête Eunice, on trouve un coût moyen à l'ultime d'environ 2816 euros +/- 10% pour un volume à l'ultime autour de 10800 sinistres +/- 22%.

Au niveau du volume, l'intervalle de confiance lié à l'estimation faite par le modèle n'est pas forcément plus fin que celui du modèle d'AXA mais nous permet déjà d'avoir en début d'évènement (à JO+3) un bon ordre de grandeur contrairement au modèle d'AXA où l'estimation du volume de sinistres pour la tempête EUNICE a notamment été revue à la baisse 30 jours ouvrés après l'ouverture des sinistres liés à cet évènement. En effet, on est passé d'un intervalle entre 13500 et 18000 sinistres prédits à JO+3 à un intervalle entre 11500 et 13000 à JO+30. En regardant enfin au niveau de la charge, on voit que notre modèle est capable de prédire dès le début d'évènement la charge de manière plus

fiable que le modèle d'AXA. En effet, à JO+3, notre modèle prédit une charge autour de 31,5M€ alors que le modèle d'AXA prédisait 42M€ pour finalement avoir une estimation à JO+30 entre 31 et 35 M€ qui se rapproche plus de l'estimation faite par notre modèle à JO+3. Le tableau ci-dessous donne un récapitulatif final des résultats de nos modèles en comparant au modèle d'AXA France.

	Modèle AXA	Modèle ML/GLM
Base de données	Moins exhaustive	Plus exhaustive
Variables explicatives utilisées	Typologie d'évènement Dates de survenance Cadences d'ouverture	+ Zones géographiques, Types de client , Branches (Auto, DAB…)
Marge d'erreur moyenne pour l'estimation du volume des sinistres	15%	22%
Marge d'erreur moyenne pour l'estimation du coût moyen	15%	10%
Projection volume de sinistres Eunice	15700 (13 500 – 18 000) à JO+3 11 500 – 13000 à JO+30	10800 (8000 – 13 000) à JO +3
Projection de la charge finale prévisible Eunice	42 M€ (35 – 50 M€) à JO +3 31- 35 M€ à JO +30	31,5 M€ (23 M€ – 39 M€) à JO+3

Charge à date pour Eunice: 31 M€
Volume à date pour Eunice: 11 000

FIGURE 4.17 – Comparaison des modèles

Ainsi, nous avons pu créer un modèle prenant en compte plus de variables pertinentes dont principalement la variable géographique. Ce qui nous a permis d'avoir une estimation plus précise du coût moyen et plus fiable du volume pour in fine avoir une estimation globalement plus fiable de la charge.

4.4 Limites

Au regard des résultats trouvés, nous avons pu démontrer que les modèles retenus, à savoir le gradient boosting pour le volume de sinistres et le MLG log-normal pour le coût moyen ouvrent des perspectives prometteuses menant vers une estimation de la charge à l'ultime plus précise dans le cadre des Évènements climatiques de Grande Ampleur. Toutefois, il demeure certains points clés qu'il serait pertinent d'améliorer.

Premièrement, en raison des contraintes liées au manque d'informations géographiques, le scope de données allant initialement de 1989 à 2021 a été restreint à 2000-2021. Les Évènements de Grande Ampleur étant justement marqués par leur rareté, un échantillon plus large permettrait un meilleur apprentissage des modèles.

De plus, on ne prend pas en compte l'évolution du portefeuille dans notre étude. Ce n'est pas la même chose d'avoir 100 sinistres en 2010 dans une zone et 100 sinistres en 2020. En effet, si entre temps il y a eu beaucoup plus ou moins de souscriptions dans la zone, cela impacte directement le portefeuille et donc la sinistralité.

Par ailleurs, pour des raisons de gain de temps, nous avons utilisé les mêmes variables explicatives pour les différents évènements climatiques de notre étude. Des variables spécifiques à certains aléas climatiques tel que le taux de pluviométrie pour la tempête peuvent être des variables très pertinentes et qu'il serait bien de considérer. En effet, certaines données propres à un type d'évènement climatique peuvent être récupérées en open data ou encore chez des partenaires météorologiques et pourraient permettre d'améliorer la prédiction de nos modèles.

Conclusion

Les récents évènements orageux en Corse nous rappellent à quel point il est essentiel d'avoir un modèle adéquat d'évaluation de la charge à l'ultime surtout lorsqu'il peut y avoir des défaillances au niveau des structures d'alertes météorologiques comme cela a pu être vu.

Se situant dans le cadre des Évènements climatiques de Grande Ampleur, ce mémoire vise à challenger le modèle d'estimation de la charge à l'ultime utilisé actuellement au sein d'AXA France et basé uniquement sur la typologie d'un évènement et les cadences d'ouverture. L'objectif était donc de proposer une nouvelle méthode d'estimation de cette charge à l'ultime grâce à des algorithmes d'apprentissage automatique et en prenant en compte notamment le critère géographique. En effet, une des limites frappantes de la méthode actuelle réside dans la non-utilisation de variables géographiques pour l'estimation de la charge à l'ultime.

Lorsqu'un évènement climatique a lieu et que des sinistres sont ouverts, suivant la méthode actuelle, on détermine l'évènement de même nature le plus proche (en terme de cadences d'ouvertures et de coût moyen) duquel on prédit la charge à l'ultime suivant une règle de trois. Celle-ci est alors estimée sans tenir compte de la zone ou des zones affectées par le dit évènement. Et cette estimation est souvent faite avec des marges d'erreur pouvant des fois être très élevées.

Ainsi, dans un premier temps, nous nous sommes attelés à construire une nouvelle base de données plus riche afin de l'exploiter pour la prédiction de la charge à l'ultime. Pour la construction de cette base, il a fallu au préalable effectuer un véritable nettoyage et traitement des données afin de récupérer le plus de données géographiques possible. Un zonier a d'ailleurs été créé afin d'harmoniser les volumes d'UP (c'est-à-dire les volumes de garanties liées aux sinistres) associés à chaque individu. Ce zonier a notamment été utilisé pour l'étape suivante qui consistait en la détection d'anomalies ou de valeurs extrêmes. Les modèles de seuil avec la théorie des valeurs extrêmes ont été testés en premier lieu avant d'essayer la méthode d'Isolation Forest qui nous a renvoyé le meilleur résultat avec 93% d'anomalies distinguées coïncidant avec un EGA.

Nous avons ensuite utilisé cette nouvelle base pour l'estimation de notre charge à

l'ultime suivant des méthodes de Machine Learning et de MLG. En effet, le calcul de la charge à l'ultime au sein d'AXA France se faisant en deux temps, avec une première estimation du volume de sinistres à l'ultime et une seconde estimation du coût moyen desquels on déduit la charge à l'ultime en faisant un produit des deux. L'estimation du volume a été réalisée en testant trois modèles de Machine Learning dont le meilleur résultat a été réalisé avec le gradient boosting pour une prédiction du coût moyen avec une marge d'erreur autour de 22%. Celle-ci peut sembler élevée comparée à la marge d'erreur autour de 15% pour le modèle d'AXA. Cependant, elle est plus fiable d'autant plus que l'estimation faite par le modèle d'AXA est souvent revue à la hausse ou à la baisse en fonction des évènements comme on a pu le voir précédemment avec l'exemple d'application à la tempête Eunice. Concernant l'estimation du coût moyen, le MLG avec la loi log-normale nous a permis d'avoir une estimation avec une marge d'erreur plus faible (10%) pour chaque aléa climatique (tempête, inondation, séisme, grêle). In fine, on on obtient une estimation globalement plus fiable de la charge à l'ultime comme on a notamment pu le voir sur l'exemple de la tempête Eunice où on a pu déterminer dès JO+3 une bonne estimation plus précise de la charge que le modèle d'AXA à la même période.

Différents axes d'évolution sont envisageables après cette étude. Il serait intéressant de tester ce modèle sur de nouveaux évènements climatiques dont la sévérité est bien plus importante que celles de notre base de données afin de conforter les performances du modèle. Pour ce mémoire, nous nous sommes focalisés sur quatre aléas climatiques, il serait intéressant de regarder également ce que donnerait notre modèle sur d'autres aléas climatiques telle que la sécheresse.

Annexe A

Annexes

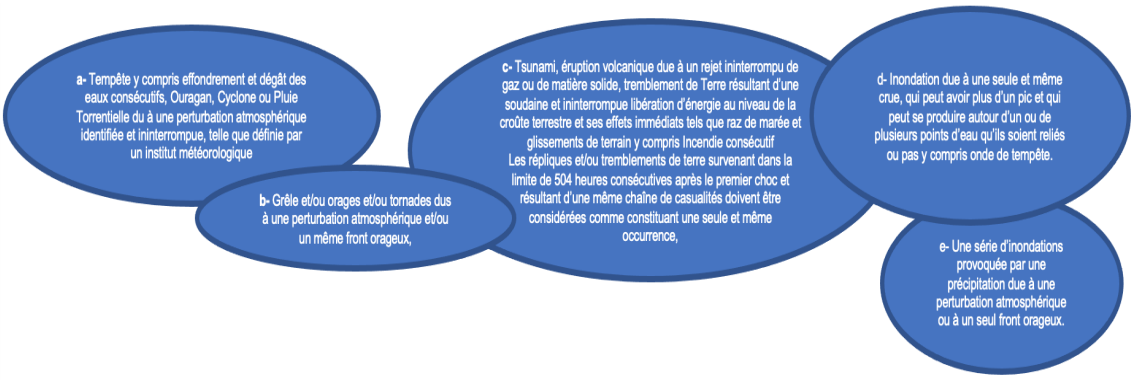
A.1 Etude France Assureurs : Projection de l'ensemble des périls climatiques à l'horizon 2050

A.2 Définition plus exhaustive d'un évènement côté réassurance AXA

A.2.1 Évènement naturel : 3 niveaux

Niveau 1

Dans les Traités de Réassurance, un Évènement est défini comme suit : Constituent un Évènement tous les sinistres assurés issus directement d'une même cause et survenant dans la même période et dans la même zone géographique. Une telle cause correspond au péril ayant provoqué directement les sinistres ou, lorsqu'il y a plusieurs périls dont la chaîne de causalité a occasionné les sinistres, au péril ayant déclenché la chaîne de causalité.



Niveau 2

Si le nombre de périls ne peut pas être déterminé en fonction des dispositions ci-dessus (niveau1), les parties acceptent de se référer à l'avis de Météo France pour les évènements autres que tremblements de terre, et à l'avis du Centre sismologique euro-méditerranéen (CSEM) pour les évènements tremblements de terre. Si les parties n'arrivent pas à trouver un accord dans un délai de 14 jours suivant l'avis de l'expert, les dispositions ci-dessous (niveau 3) devront s'appliquer.

Niveau 3

Si les dispositions du niveau 1 ci-dessus ne peuvent pas s'appliquer et si un accord ne peut pas être trouvé comme énoncé ci-dessus (niveau 2), un Évènement devra être

déterminé en fonction d'une période continue débutant à la date et à l'heure définies par la CÉDANTE mais en aucun cas plus tôt que le premier sinistre individuel enregistré par la CÉDANTE pour cette catastrophe, et dont la durée est limitée aux clauses horaires.

A.2.2 Évènement périls non naturels : Définition

Constitue un seul et même sinistre dans le cadre d'un évènement non naturel, l'ensemble des indemnités dues par la CÉDANTE au titre des dommages consécutifs à cet évènement, subis par deux ou plusieurs risques, et ce, quel que soit le nombre de polices d'assurances frappées.

A.3 Liste de toutes les variables

Variables	Type	Description
DTSURV	DateTime	Date de survenance du sinistre
Year	DateTime	Année de survenance relativement à DTSURV
DTM_Lieu_Du_Sinistre	Object	Lieu du sinistre
DTM_PostalCode	Integer	Code Postal
DTM_City	Object	Ville du sinistre
UP	Object	Garantie activée pour l'indemnisation du sinistre
infocentre_CDPOSURV	Float	Code Postal issu de l'infocentre
geoiris_CP	Float	Code Postal issu des données de contrats
geoiris_COMMUNE	Object	Nom de commune issu des données de contrats
debut	DateTime	Date de début d'évènement
fin	DateTime	Date de fin d'évènement
dureeEvenement	Integer	Durée d'évènement
zones_agregees_dappartenance	Object	Zones étendues obtenues après construction du zonier
nb_zones	Integer	Nombre de zones affectées par un évènement ou une anomalie
Type_even	Object	Type d'aléa climatique
DTOUV	DateTime	Date d'ouverture du sinistre
DTCLOT	DateTime	Date de clôture du sinistre
cdentite	Object	Entreprise ou Pro/Particulier
niv2	Object	Branche d'activité
part_auto	Float	% de sinistres auto
part_dommages	Float	% de sinistres dommages
part_transport	Float	% de sinistres transport
part_constructions	Float	% de sinistres constructions

Table A.1 – Liste de toutes les variables

A.4 Répartition du volume de sinistres par année et mois de survenance

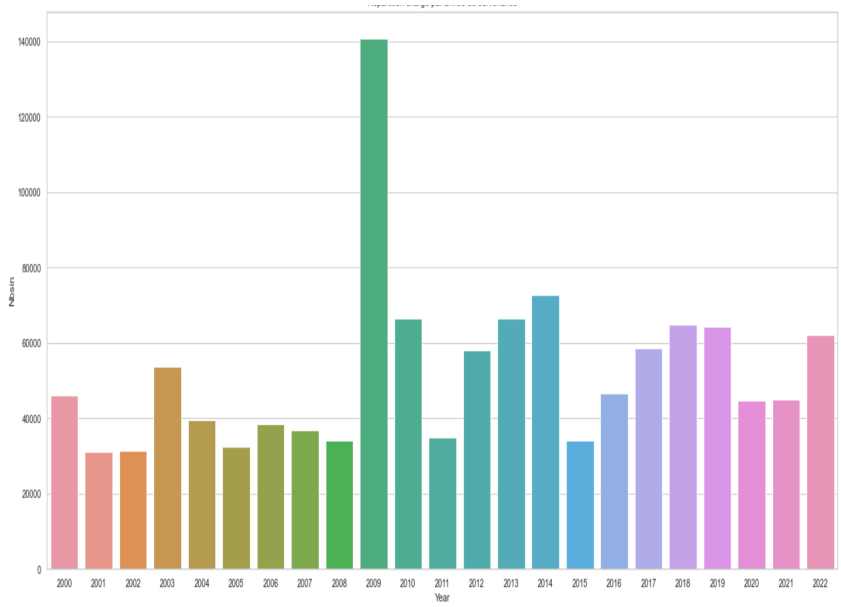


FIGURE A.1 – Répartition du volume de sinistres par année de survenance

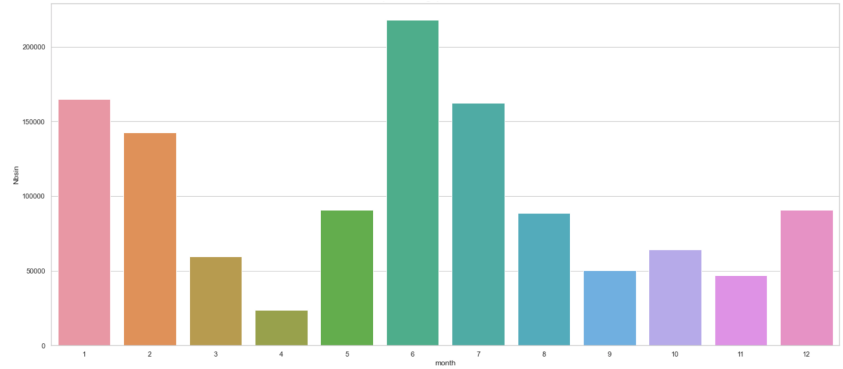


FIGURE A.2 – Répartition du volume de sinistres par mois de survenance

A.5 Répartition du coût moyen par branche

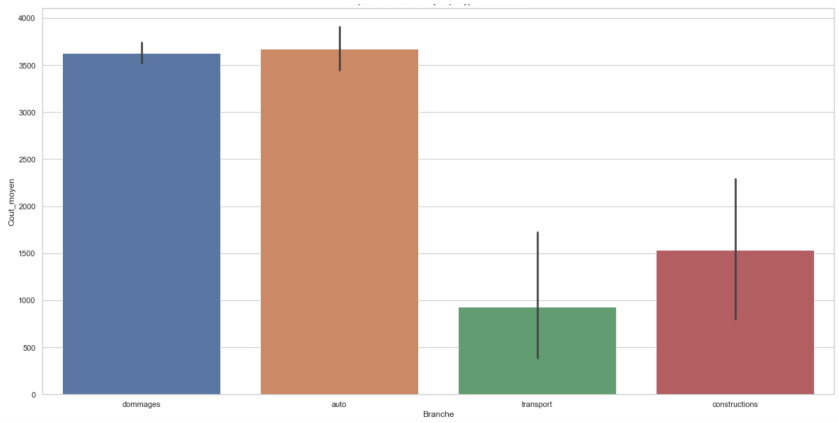


FIGURE A.3 – Répartition du coût moyen par branche

A.6 Matrices de correlation pour les variables catégorielles et numériques

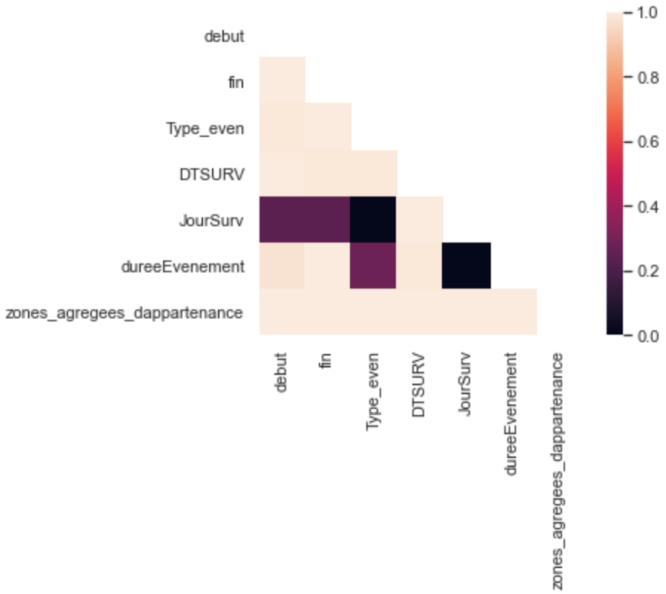


FIGURE A.4 – Matrice de correlation après test de Cramer pour les variables catégorielles

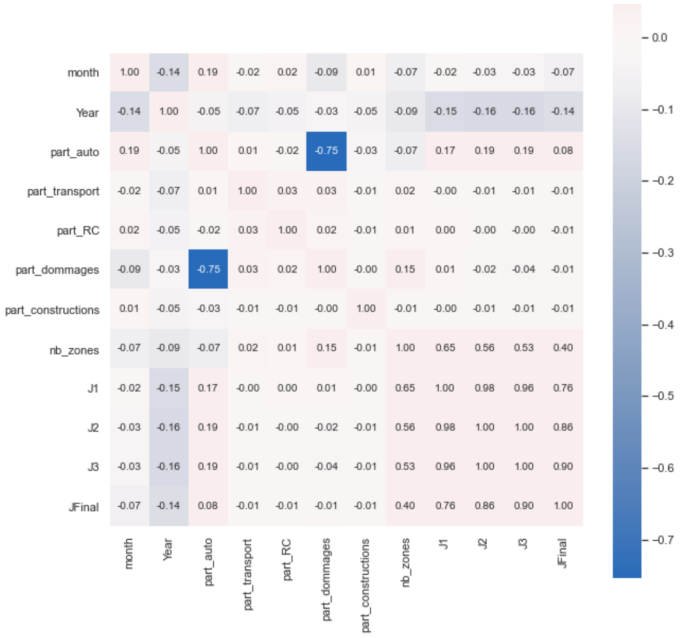


FIGURE A.5 – Matrice de correlation pour les variables numériques

Table des figures

Liste des tableaux

Bibliographie

[1] Sara Angeli Aguiton. La pluie, le rendement et l'assurance. les risques climatiques en agriculture au sénégal. *HAL*, 2020.

[2] Amir Ali. Decision tree with practical implementation. *Wavy AI Research Foundation*, 2018.

[3] France Assureurs. Impact du changement climatique sur l'assurance à l'horizon 2050. *France Assureurs*, 2021.

[4] France Assureurs. L'assurance française données clés 2020. *France Assureurs*, 2021.

[5] Nathalie BEDI. Modélisation du risque de tempête en france métropolitaine. *Institut des actuaires : Mémoire d'actuariat*, 2018.

[6] S Bottard. Application de la méthode du bootstrap pour l'estimation des valeurs extrêmes dans les distributions de l'intensité des séismes. *Revue de Statistique Appliquée, tome 44, no 4, p. 5-17*, 1996.

[7] Leo Breiman. Random forests. *Statistics Department University of California Berkeley*, 2001.

[8] Covéa. Changement climatique & assurance :quelles conséquences sur la sinistralité à horizon 2050 ? *Livre blanc Covéa*, 2022.

[9] DataScientest. Bagging en machine learning : de quoi s'agit-il ? *DataScientest*.

[10] DataScientest. Isolation forest. *DataScientest*. `https://datascientest.com/isolation-forest`.

[11] Arnaud Donguy. Contribution de l'information géographique aux m´etiers de l'assurance pour la gestion des evenements d'ampleur. *HAL*, 2012.

[12] Francis Duval. Gradient boosting techniques for individual loss reserving in non-life insurance. *Mémoire Université du Québec À Montréal*, 2019.

[13] Armelle Guillou et Alexandre You. Introduction à la théorie des valeurs extrêmes : Applications en actuariat. *Institut des actuaires*, 2011.

[14] A. C. Davison et D. V. Hinkley. Bootstrap methods and their application. *Cambridge University Press*, 1997.

[15] CCR et METEO FRANCE. Conséquences du changement climatique sur le coût des catastrophes naturelles en france à horizon 2050. *CCR, METEO FRANCE*, 2018.

[16] Arthur Charpentier et Michel Denuit. Mathématiques de l'assurance non-vie, tome ii : Tarification et provisionnement. *Economica*, 2010.

[17] Noureddine Benlagha et Michel Grun-Réhomme et Olga Vasechko. Les sinistres graves en assurance automobile : Une nouvelle approche par la théorie des valeurs extrêmes. *Revue MODULAD*, 2009.

[18] Cédric Fleury. Le kd-tree : une méthode de subdivision spatiale. *INSA de Rennes*, 2007.

[19] Robin Genuer. Forêts aléatoires : aspects théoriques, sélection de variables et applications. *HAL*, 2011.

[20] GIEC. Rapport 2022 du giec : une nouvelle alerte face au réchauffement climatique. *GIEC*, 2022.

[21] Fouad Jabiri. Applications de méthodes de classification non supervisées à la détection d'anomalies. *Université Laval, Mémoire de Statistiques P.18-24*, 2020.

[22] Will Koehrsen. Hyperparameter tuning the random forest in python. *Towards Data Science*, 2018.

[23] Martin Koppe. Introduction à la théorie des valeurs extrêmes : Applications en actuariat. *CNRS Le journal*, 2022.

[24] Iago Pereira Lemos, Antônio Marcos Gonçalves Lima, and Marcus Antônio Viana Duarte. Thresholdmodeling : A python package for modeling excesses over a threshold using the peak-over-threshold method and the generalized pareto distribution. *The Journal of Open Source Software*, 2011.

[25] Gwladys Mao. Estimation des coûts économiques des inondations par des approches de type physique sur exposition. *HAL*, 2020.

[26] Eve Mathieu-Dupas. Algorithme des k plus proches voisins pondérés et application en diagnostic. *HAL*, 2010.

[27] Amol Mavuduru. How to perform anomaly detection with the isolation forest algorithm. *Towards Data Science*, 2021.

[28] Dina Mohamed, Ayman El-Kilany, and Hoda M. O. Mokhtar. A hybrid model for documents representation. *International Journal of Advanced Computer Science and Applications, Vol. 12, No. 3*, 2021.

[29] ALICE Pauthier. L'assurance des risques climatiques. *OGéoD IRIS*, 2009.

[30] Julius Quiquet. Méthode d'estimation de la charge ultime en rc corporelle automobile basée sur des données individuelles. *Institut des actuaires : Mémoire d'actuariat*, 2007.

[31] Youcef Rahmani. Esg et valeur fondamentale des entreprises. *Droit et croissance*, 2015.

[32] Scikit-learn. Gradient boosting regression. *Scikit-learn*. `https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingRegressor.html`.

[33] Scikit-learn. Isolation forest. *Scikit-learn*. `https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.IsolationForest.html`.

[34] Scipy. Kdtree. *Docs.scipy*. `https://docs.scipy.org/doc/scipy/reference/generated/scipy.spatial.KDTree.html`.

[35] Gwladys Toulemonde. Estimation et tests en théorie des valeurs extrêmes. *HAL*, 2008.

[36] C. Tuleau-Malot. Préesentation de l'algorithme cart. *Laboratoire de mathématiques Université de Nice*.