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

Par : Raphaël Tumminello

Titre du mémoire :

Exploitation de modèles stochastiques de catastrophes naturelles pour une tarification technique

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

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

Membres présents du jury de l'Institut des Actuaires — signature

Membres présents du jury de la filière

Entreprise :

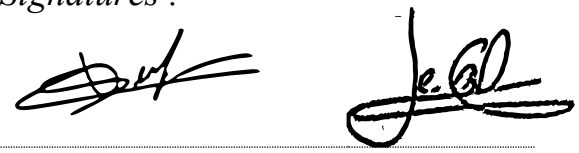
Nom : GIE Axa

Signature :

Directeurs de mémoire en entreprise :

Nom : Emmanuel Delafosse, Rozenn Le Calvez

Signatures :



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

Signature du responsable entreprise

Signature du candidat

Résumé

D'année en année, les dégats économiques et corporels liés aux catastrophes naturelles tendent à augmenter. Cette tendance est exacerbée par le réchauffement climatique qui provoque des phénomènes météorologiques violents de plus en plus fréquents et par l'augmentation de l'exposition (physique et humaine) accompagnant la croissance économique et l'urbanisation.

Pour assurer ces risques, il est primordial de les quantifier. Cependant, l'historique des catastrophes naturelles n'est en général pas assez fourni pour effectuer des méthodes statistiques classiques de tarification. C'est pourquoi les actuaires travaillent étroitement avec des experts en catastrophes naturelles capables, par des modèles physiques, de calibrer des modèles stochastiques générant des scénarios fictifs probables. Ces modèles sont appelés modèles catastrophes naturelles ou modèles CAT. Ce mémoire a pour objectif d'exploiter le modèle CAT pour la Suisse afin de tarifer l'assurance dommages du péril séisme.

Le modèle CAT est articulé en 4 modules (module exposition, aléa, vulnérabilité et financier) que nous allons détailler dans ce mémoire. Il permet d'obtenir une distribution de pertes site à site sur un portefeuille d'assurés. Afin d'être capable de reproduire le plus fidèlement possible les simulations de ce modèle CAT, nous avons calibré différents modèles. Déjà, nous avons mis en place un modèle GLM classique coût/fréquence. Ensuite nous avons modélisé directement le taux de destruction obtenu par des GLMs utilisant la famille Tweedie. Enfin, nous avons calibré des modèles innovants de machine learning. Tous ces modèles ont été comparés pour déterminer lesquels donnent les meilleurs résultats. Nous avons également développé une méthode de modélisation de l'impact des limites et franchises sur la perte brute pour compléter la tarification.

Mots clés

Catastrophes naturelles, Modèle Cat, Tarification, Séisme, Assurance dommages, GLM, Machine Learning, Taux de destruction, Conditions financières, AEP, OEP

Abstract

From year to year, the economic and physical damage caused by natural disasters tends to increase. This trend is exacerbated by global warming, which causes increasingly frequent violent meteorological phenomena, and by the increase in exposure (physical and human) accompanying economic growth and urbanization.

To insure these risks, it is essential to quantify them. However, the history of natural disasters is generally not sufficient to carry out classic statistical pricing methods. This is why actuaries work closely with experts in natural catastrophes who are able, using physical models, to calibrate stochastic models generating probable fictitious scenarios. These models are called natural catastrophe models or CAT models. The objective of this thesis is to use the CAT model for Switzerland in order to price the damage insurance of the earthquake hazard.

The CAT model is articulated in 4 modules (exposure, hazard, vulnerability and financial module) that we will detail in this paper. It allows us to obtain a site-by-site distribution of losses on a portfolio of insureds. In order to be able to reproduce the simulations of this CAT model as faithfully as possible, we have calibrated different models. First, we set up a classic GLM cost-frequency model. Then we directly modeled the destruction rate obtained by GLMs using the Tweedie family. Finally, we calibrated innovative machine learning models. All these models were compared to determine which ones give the best results. We have developed a method for modeling the impact of limits and deductibles on gross loss to complete the pricing.

Key words

Natural disasters, Cat Model, Pricing, Eartquake, Property insurance, GLM, Machine Learning, Destruction rate, Financial conditions, AEP, OEP

Synthèse

L'objectif de ce mémoire est d'exploiter le modèle catastrophe naturelle, modèle interne développé par les équipes de modélisation catastrophe naturelle du GIE AXA, afin de proposer une tarification technique pour le péril séisme. Ce mémoire sera accompagné d'exemples liés à la Suisse.

Dans de nombreuses zones géographiques, il est très difficile d'exploiter les données de pertes suite à des catastrophes naturelles. En effet ces évènements sont par essence rares voire très rares, et ne donnent pas lieu à des données assez fournies pour appliquer des méthodes statistiques classiques. Par conséquent, l'assureur a besoin d'exploiter un modèle catastrophe naturelle, dont on distingue 2 types :

- Les modèles vendeurs proposés par des compagnies externes (comme par exemple AIR, RMS, EQE), modèles pour lesquels on n'obtient que les valeurs financières sans avoir le modèle détaillé avec la vulnérabilité des biens et les évènements considérés.
- Les modèles internes développés par l'assureur lui-même, permettant de modéliser le phénomène physique sous-jacent à l'origine des catastrophes naturelles en plus de l'impact financier. La différence est que l'assureur a l'expertise du modèle physique en interne.

Le GIE AXA a développé son propre modèle catastrophe naturelle pour le péril séisme, nous sommes donc dans le deuxième cas dans le cadre de ce mémoire.

Structure du mémoire

Le mémoire débute par l'étude du risque assurantiel lié aux catastrophes naturelles ainsi que leur assurabilité, puis par un focus sur le péril séisme qui sera étudié dans ce mémoire.

Dans une deuxième partie, nous rentrons dans les détails du modèle catastrophe naturelle et expliquons chacune de ses étapes pour le péril séisme. La structure du modèle catastrophe naturelle est le suivant :

- **Module Aléa :** il construit un catalogue d'événements contenant une multitude de scénarios réalistes et probables. L'occurence des séismes est modélisée par un processus de Poisson, nécessitant le traitement de déclustering du catalogue de sismicité ; il s'agit de méthodes permettant d'enlever les répliques de séismes ou de les traiter comme un même événement et ainsi de respecter au maximum la notion d'indépendance des processus de Poisson.
- **Module Exposition :** il décrit les risques présents dans le portefeuille (localisation géographique des biens assurés, leur valeur assurée ainsi que leurs caractéristiques),
- **Module Vulnérabilité :** il associe à chaque bien assuré et touché par une catastrophe une perte économique en fonction des caractéristiques des biens ainsi

que de l'intensité de l'évènement. Ce module nécessite l'utilisation de courbes de fragilité qui sont définies selon les caractéristiques des bâtiments,

— **Module Financier :** il applique des conditions financières de l'assurance à chaque perte brute. Il s'agit donc de prendre en compte limites, franchises, réassurance, coassurance...

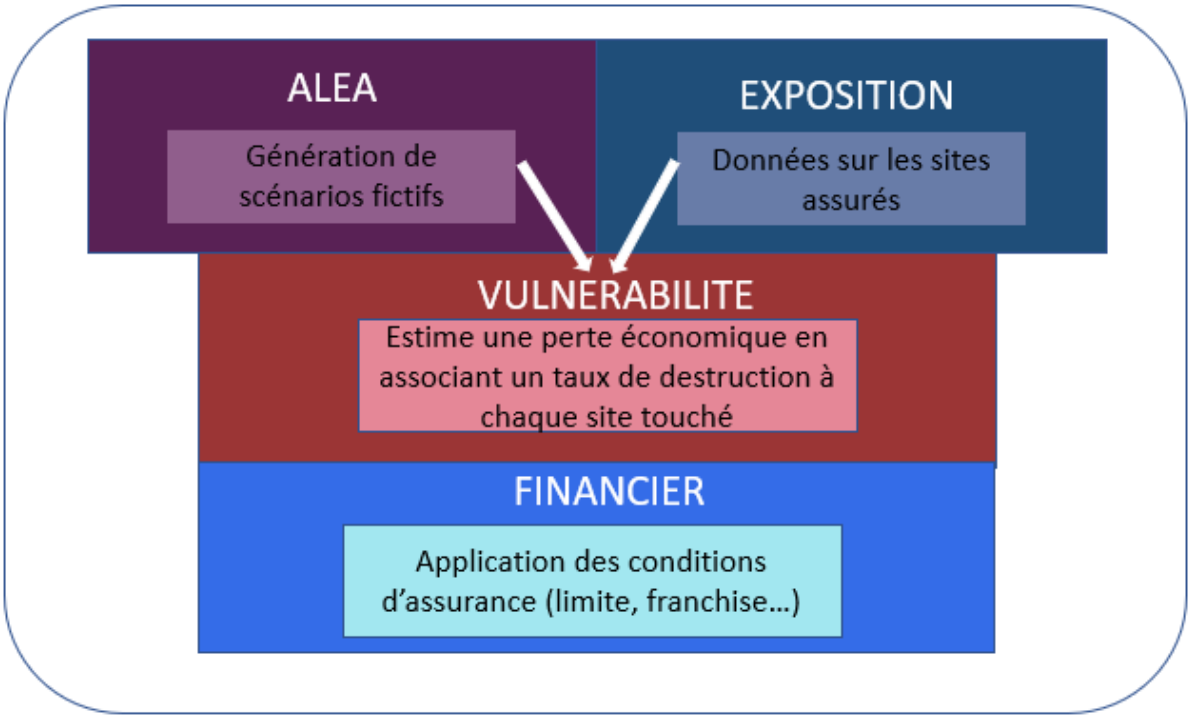


FIGURE 1 – Structure du modèle catastrophe naturelle

Enfin, dans une troisième et dernière partie, nous développerons les techniques mises en place afin d'exploiter le modèle stochastique de catastrophes naturelles pour une tarification technique du péril.

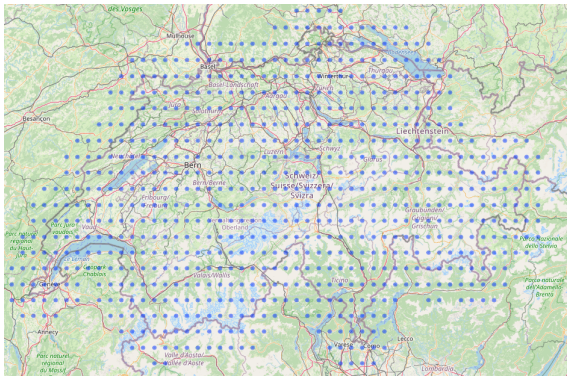
La base de données utilisée

Les données de perte que nous utilisons dans ce mémoire sont issues du modèle catastrophe naturelle. A partir d'un portefeuille d'assurés en entrée du modèles, nous obtenons une distribution de pertes financières. Le portefeuille d'assurés sur lequel nous faisons tourner le modèle est le résultat de la concaténation de 2 portefeuilles :

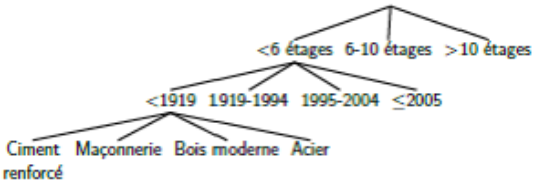
— Le portefeuille d'assurés réels de 2020, c'est à dire les bâtiments réellements assurés contre le risque sismique pour l'année 2020.

— Un portefeuille fictif, aussi appelé portefeuille représentatif, qui est un ensemble d'assurés fictifs. Nous avons la possibilité d'ajouter des assurés fictifs à l'étude car l'utilisation d'un modèle CAT pour obtenir des pertes ne nécessite pas forcément des assurés réels en entrée. L'intérêt d'ajouter des assurés fictifs est de pouvoir obtenir un portefeuille d'assurés couvrant l'entièreté du pays, et avec toutes les combinaisons de variables explicatives possibles. Ainsi, cela permet d'avoir une vision exhaustive du risque. La figure 2 résume la méthode pour construire le

portefeuille représentatif ; dans un premier temps on quadrille le pays pour avoir des bâtiments fictifs dans tout le pays, et ensuite, pour chacun des bâtiments on combine toutes les modalités de variables explicatives possibles en suivant l'arborescence des variables liées aux bâtiments.



(a) Quadrillage de la Suisse



(b) Arborescence des variables

FIGURE 2 – Création du portefeuille fictif

Enfin, l'objectif de cette partie est de modéliser le plus précisément les pertes simulées site à site d'un modèle CAT pour chaque couverture en calibrant des modèles statistiques usuels type GLM ou des modèles innovants tels que les réseaux de neurones et XGBoost. Les variables de risque à notre disposition pour effectuer la modélisation sont les suivantes :

— Des variables géographiques : Cresta (*Catastrophe Risk Evaluation and Standardising Target Accumulations*), le vs30 (qui représente la vitesse moyenne des ondes de cisaillement à 30 mètres de profondeur) ainsi que le *Peak Ground Acceleration* (PGA, accélération maximale du sol) qui est un très bon indicateur du risque sismique.

— Des variables liées aux bâtiments assurés : le type de structure, le nombre d'étages ainsi que l'année de construction.

Parmi ces variables, seul le PGA est une variable continue car les autres sont soit qualitatives par nature, soit découpées en classes par défaut pour faire tourner le modèle CAT. Ainsi, pour calibrer correctement les GLMs effectués, la variable PGA est discrétisée par arbre de régression, nous permettant d'obtenir 5 classes de risques liées au PGA.

Modélisation des pertes brutes

Nous avons tout d'abord comparé plusieurs GLMs, en affectant des poids plus élevés aux assurés réels qu'aux assurés fictifs :

1. Un modèle GLM fréquence/coût moyen ;
2. Un unique modèle GLM Tweedie, où la variable à expliquer est le taux de destruction (comprenant les zéros et valeurs positives) ;
3. Un GLM par zone de risque (zone risquée/non risquée) ;
4. Un GLM Tweedie avec des informations géographiques supplémentaires.

Ensuite, nous avons étudié des modèles de machine learning ; le XGBoost ainsi que le réseau de neurones, toujours dans un but de modélisation des pertes et afin de comparer

ces modèles avec des modèles GLMs plus classiques.
Afin de comparer tous les modèles calibrés, nous avons utilisé les indicateurs suivants :

— Le RMSE (*Root Mean Square Error)* : la fonction de perte quadratique classique. Le RMSE représente la moyenne du carré des erreurs site par site,

— Le RMAE (*Root Mean Absolute Error)* : la fonction de perte absolue. Le RMAE représente la moyenne de la valeur absolue des erreurs site par site,

— L'erreur cumulée : elle est définie comme la différence entre la somme des pertes réalisées et la somme des pertes du modèle de tarification,

— L'indice de Gini est un indicateur mesurant la capacité de segmentation du modèle et calculé à partir d'une courbe appelée courbe de Lorenz.

Le récapitulatif de ces indicateurs pour tous nos modèles est donc présenté dans les figures 3 et 4 :

	MODELES GLM							
	Modèle 1		Modèle 2		Modèle 3		Modèle 4	
Indicateurs	coût-fréquence		Unique Tweedie		Tweedie/Gamma par zone de risque		Tweedie avec plus d'informations géographiques	
Set	Train	Test	Train	Test	Train	Test	Train	Test
Erreur globale	-1,76%	-2,13%	1,52%	-2,25%	1,49%	1,84%	1%	1,55%
RMSE	0,603	0,649	0,380	0,421	0,366	0,413	0,379	0,418
RMAE	0,314	0,322	0,224	0,229	0,215	0,222	0,223	0,227
Gini	60,74%	60,44%	70,54%	70,80%	70,88%	71,07%	70,47%	70,57%

FIGURE 3 – Indicateurs pour les modèles GLMs

	MODELES Machine learning			
	Modèle 5		Modèle 6	
Indicateurs	Réseau de neurones		Xgboost	
Set	Train	Test	Train	Test
Erreur globale	-0,10%	0,26%	3,20%	3,14%
RMSE	0,364	0,405	0,270	0,271
RMAE	0,206	0,210	0,480	0,479
Gini	71,57%	71,79%	73,00%	67,00%

FIGURE 4 – Indicateurs pour les modèles de machine learning

Il est finalement intéressant de garder 2 modèles : le modèle 2 et le modèle 5. Le modèle 2 est le meilleur modèle GLM car il est le meilleur compromis entre complexité du modèle et indicateurs obtenus. L'avantage des modèles GLMs est qu'ils permettent d'expliquer facilement le tarif final car on peut en sortir une grille de coefficients à appliquer à la prime selon les modalités de variables explicatives. Au niveau des modèles de machine learning, on préférera le modèle 5 au modèle 6 car on trouve de meilleurs indicateurs sur la base de test. Les indicateurs sont également améliorés par rapport aux modèles

GLMs, mais le côté boite noire des réseaux de neurones est problématique pour une tarification en assurance. Le modèle de réseau de neurones peut au moins servir de modèle de référence car il donne de très bons indicateurs.

Modélisation des conditions d'assurance

Les modèles précédents permettent d'obtenir les pertes brutes attendues. A cette étape, il reste l'application des conditions d'assurance pour compléter la tarification, c'est à dire la franchise et la limite. A cet effet, nous mettons en place une méthode d'interpolation d'une courbe d'exposition empirique avec une courbe d'exposition d'une loi MBBEFD. Par exemple, l'interpolation de la courbe d'exposition dans le cas d'une franchise donne la figure suivante :

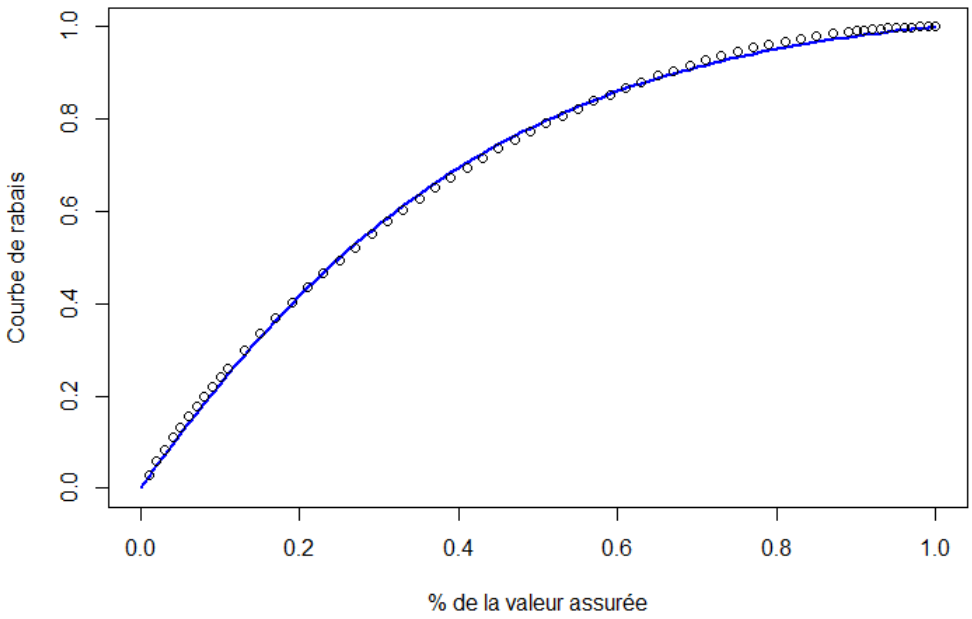


FIGURE 5 – Interpolation d'une courbe d'exposition d'une loi MBBEFD (cas d'une franchise)

En reliant le module modélisation de la perte brute avec l'application des conditions d'assurance, nous obtenons une tarification complète.

Synthesis

The aim of this paper is to use the natural disaster model, an internal model developed by the natural disaster modeling teams of GIE AXA, in order to propose technical pricing for the earthquake hazard. This paper will be accompanied by examples related to Switzerland.

In many geographical areas, it is very difficult to exploit the data of losses due to natural disasters. Indeed these events are in essence rare or very rare, and do not provide enough data to apply classical statistical methods. As a result, the insurer needs to operate a catastrophe model (or CAT model), of which we distinguish 2 types :

— Vendor models offered by external companies (such as AIR, RMS, EQE), models for which only financial values are obtained without having the detailed model with the vulnerability of the assets and the events considered.

— The internal models developed by the insurer itself, allowing to model the underlying physical phenomenon at the origin of natural disasters in addition to the financial impact. The difference is that the insurer has the expertise of the physical model internally.

The GIE AXA has developed its own natural disaster model for the peril seism, we are therefore in the second case in the context of this paper.

Thesis Structure

The dissertation begins with a study of the insurance risk associated with natural disasters as follows and their insurability, and then with a focus on the earthquake risk which will be studied in this paper.

In a second part, we go into the details of the catastrophe model and explain each of its steps for the earthquake hazard. The structure of the CAT model is as follows :

— **Hazard module** : it builds a catalog of events containing a multitude of realistic and probable scenarios. The occurrence of earthquakes is modeled by a Poisson process, requiring the declustering treatment of the seismicity catalog ; these are methods allowing to remove aftershocks of earthquakes or to treat them as the same event and thus to respect the notion of independence of Poisson processes.

— **Exposure Module :** it describes the risks that are in the portfolio (geographical location of the insured buildings, their insured value and their characteristics),

— **Vulnerability Module :** it associates an economic loss to each insured building affected by a disaster according to the characteristics of the building and the intensity of the event. This module requires the use of fragility curves which are defined according to the characteristics of the buildings,

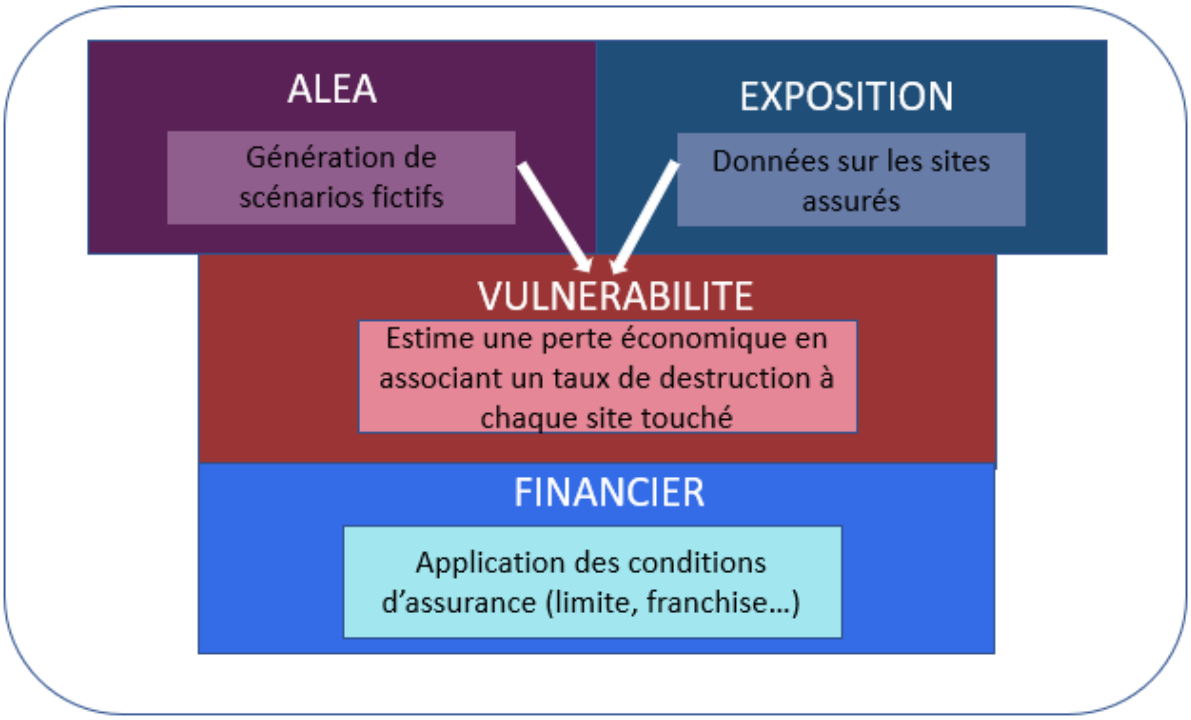


FIGURE 6 – CAT model structure

— **Financial Module** : it applies the financial terms of the insurance to each gross loss. Then we take into account limits, deductibles, reinsurance, co-insurance...

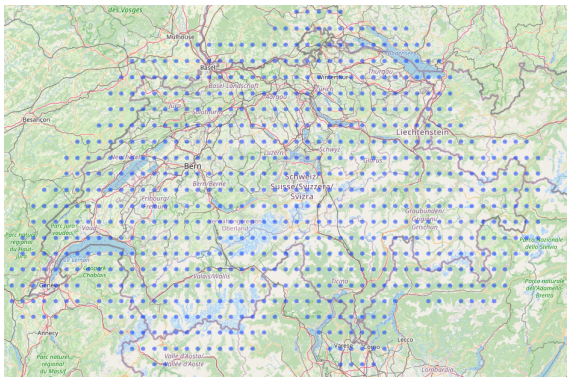
Finally, in a third and last part, we will develop the techniques in order to exploit the stochastic model of natural disasters for a technical pricing of the risk.

The database used

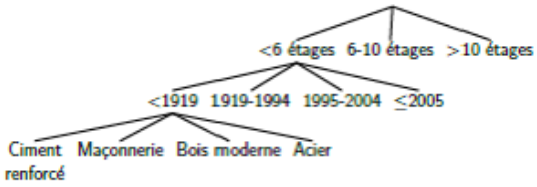
The losses data that we use in this paper is from the catastrophe model. From a portfolio of insureds at the input of the model, we obtain a distribution of financial losses. The portfolio of insureds on which we run the CAT model is the result of the concatenation of 2 portfolios :

— The real insured portfolio for 2020, i.e. the buildings actually insured against seismic risk for the year 2020.

— A fictitious portfolio, also known as a representative portfolio, which is a set of fictitious insureds. We have the ability to add fictitious insureds to the study because the use of a CAT model to obtain losses does not necessarily require real insureds as input. The interest of adding fictitious insureds is to be able to obtain a portfolio of insureds covering the entire country, and with all possible combinations of explanatory variables. Thus, it provides an exhaustive vision of the risk. The figure 7 summarizes the method for constructing the representative portfolio ; first of all, the country is squared in order to have fictitious buildings throughout the country, and then, for each of the buildings, all the possible explanatory variable modalities are combined by following the tree structure of the variables related to the buildings.



(a) grid of switzerland



(b) Variable tree structure

FIGURE 7 – Creation of the fictitious portfolio

Finally, the aim of this part is to model as accurately as possible the simulated site-to-site losses of a CAT model for each coverage by calibrating usual statistical models such as GLM or innovative models such as neural networks and XGBoost. The risk variables that at are our disposal to carry out the modeling are as follows :

— Geographical variables : Cresta (*Catastrophe Risk Evaluation and Standardising Target Accumulations*), the vs30 (which represents the average shear wave velocity at a depth of 30 meters) as well as the *Peak Ground Acceleration* (PGA) which is a very good seismic risk indicator.

— Variables related to the insured buildings : type of structure, number of floors and year of construction.

Among these variables, only the PGA is a continuous variable because the others are either qualitative in nature or cut into default classes to run the CAT model. Thus, to correctly calibrate the GLMs performed, the PGA variable is discretized by regression tree, allowing us to obtain 5 classes of risks related to the PGA.

Gross loss modeling

We first compared several GLMs, assigning higher weights to real insureds than to fictitious insureds :

1. A GLM model frequency/severity ;
2. A unique Tweedie GLM, where the response variable is the destruction rate (including zeros and positive values) ;
3. A GLM by rik zone (risky/unrisky area) ;
4. A Tweedie GLM with additional geographical information.

Then, we studied models of machine learning ; the XGBoost as well as the neural network, always with the aim of modeling losses and in order to compare these models with more classical GLM models.
In order to compare all the calibrated models, we used the following indicators :

— The RMSE (*Root Mean Square Error)* : the classical quadratic loss function. The RMSE represents the mean square of the errors site by site,

— The RMAE (*Root Mean Absolute Error)* : The RMAE represents the average of the absolute value of errors site by site.,

— The cumulated error : It is defined as the difference between the sum of realized losses and the sum of losses in the pricing model,

— The Gini index is an indicator measuring the segmentation capacity of the model and calculated from a curve called the Lorenz curve..

The summary of these indicators for all of our models is therefore presented in the figures 8 and 9 :

	MODELES GLM							
Indicateurs	Modèle 1 coût-fréquence		Modèle 2 Unique Tweedie		Modèle 3 Tweedie/Gamma par zone de risque		Modèle 4 Tweedie avec plus d'informations géographiques	
Set	Train	Test	Train	Test	Train	Test	Train	Test
Erreur globale	-1,76%	-2,13%	1,52%	-2,25%	1,49%	1,84%	1%	1,55%
RMSE	0,603	0,649	0,380	0,421	0,366	0,413	0,379	0,418
RMAE	0,314	0,322	0,224	0,229	0,215	0,222	0,223	0,227
Gini	60,74%	60,44%	70,54%	70,80%	70,88%	71,07%	70,47%	70,57%

FIGURE 8 – Indicators for GLM models

	MODELES Machine learning			
Indicateurs	Modèle 5 Réseau de neurones		Modèle 6 Xgboost	
Set	Train	Test	Train	Test
Erreur globale	-0,10%	0,26%	3,20%	3,14%
RMSE	0,364	0,405	0,270	0,271
RMAE	0,206	0,210	0,480	0,479
Gini	71,57%	71,79%	73,00%	67,00%

FIGURE 9 – Indicators for Machine learning models

It is finally interesting to keep 2 models : model 2 and model 5. Model 2 is the best GLM model because it is the best compromise between the complexity of the model and the indicators obtained. The advantage of GLM models is that they allow the final tariff to be easily explained because a grid of coefficients to be applied to the premium can be derived from them according to the explanatory variables. At the level of machine learning models, model 5 is preferred to model 6 because better indicators can be found on the test basis. The indicators are also improved compared to GLM models, but the black box side of neural networks is problematic for insurance pricing. The neural network model can at least serve as a reference model because it gives very good indicators.

Insurance conditions modeling

Previous models provide the expected gross losses. At this stage, there remains the application of the insurance conditions to complete the pricing, i.e. the deductible and the limit. For this purpose, we set up a method of interpolation of an empirical exposure

curve with an exposure curve of an MBBEFD law. For example, the interpolation of the exposure curve in the case of a deductible gives the following figure :

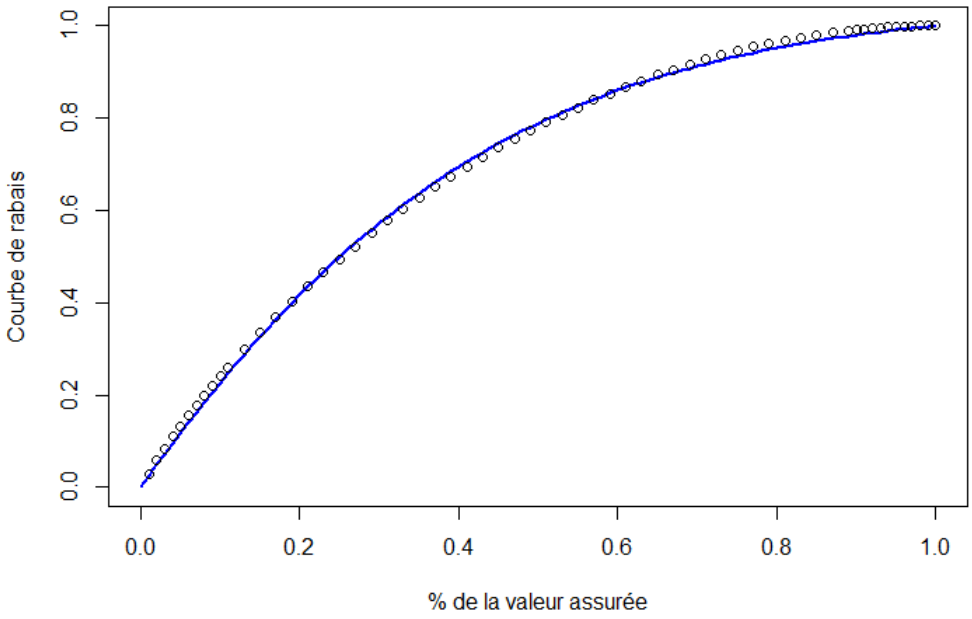


FIGURE 10 – Interpolation of an MBBEFD exposure curve (case of a deductible)

By linking the gross loss modeling module with the application of insurance conditions, we obtain a complete pricing.

Remerciements

Je tiens tout d'abord à remercier Rozenn Le Calvez et Emmanuel Delafosse de m'avoir guidé durant toute la rédaction de ce mémoire, ainsi que pour leurs multiples relectures. Je tiens également à les remercier de m'avoir donné l'opportunité d'intégrer l'équipe Data Analytics, Actuarial and Underwriting Solutions du GIE AXA, qui a été une expérience très enrichissante.

Un grand merci également à Florient Aubry et Jennifer Bounan pour leur aide et leurs conseils avisés. Je remercie tous ceux qui ont de près ou de loin contribué à l'écriture de ce mémoire en répondant à mes questions ou en me fournissant des documents.

Enfin, je souhaiterais remercier ma tutrice académique pour ce mémoire, Maud Thomas, pour ses conseils.

Pour finir, merci à mes parents pour leur soutien tout au long de mes études.

Table des matières

Table des figures

Liste des tableaux

Introduction

La dernière décennie (de 2010 à 2019) a été très marquée par les catastrophes naturelles. En effet, cette décennie a été record en termes de pertes économiques et nombre de décès causés par une catastrophe naturelle. En particulier, le péril tremblement de terre, qui est le péril étudié dans ce mémoire, a causé de lourds dégâts. Par conséquent, il est de plus en plus important pour un assureur de maîtriser le risque sismique et les dommages qu'ils peuvent engendrer.

Pour maîtriser le risque sismique, l'assureur doit être capable de quantifier le risque. Mais dans les régions longtemps épargnées par un séisme, il est très probable que le risque soit sous-estimé. En effet, une telle catastrophe étant tellement rare dans certaines régions qu'appliquer des méthodes statistiques classiques sur l'historique des sinistres est contre-productif par manque de données. Pour combler ce manque, l'assureur peut décider d'exploiter les sorties des outils de modélisation commerciaux, dont les plus connus sont RMS, EQE et AIR. Cependant, le côté boite noire de ces modèles est un frein ; en effet ces modèles ne donnent que des valeurs de pertes sur un portefeuille d'assurés mais ne dévoilent pas les évènements sous-jacents ainsi que les vulnérabilités des bâtiments. La directive solvabilité 2 exige aux compagnies d'assurance une bonne connaissance de leurs risques. Ainsi, certaines sociétés, dont Axa, développent leurs propres modèles catastrophes naturelles.

L'étude, qui a été faite pour la Suisse, se décompose en 3 chapitres :

Dans le premier chapitre, on étudie le risque assurantiel lié aux catastrophes naturelles dans leur ensemble et leur assurabilité. Un focus sur le péril séisme est également effectué.

Dans le deuxième chapitre, nous allons décortiquer le modèle catastrophe développé par les experts séismes d'Axa et rentrer dans les détails de chaque module : le module exposition, le module aléa, le module vulnérabilité et le module financier.

Enfin, dans un troisième et dernier chapitre, nous développerons des méthodes de type GLM et des méthodes innovantes de machine learning afin de modéliser le plus précisément possible les pertes site à site obtenues par le modèle catastrophe naturelle. Nous comparerons les méthodes calibrées pour déterminer le ou les meilleurs modèles à garder. Enfin, afin de compléter la tarification, nous développerons une méthode de modélisation de l'impact des franchises et limites sur la perte brute.

Pour des raisons de confidentialité, les valeurs numériques utilisées dans ce mémoire ont été normalisées ou modifiées.

Chapitre 1

Présentation générale

1.1 Le groupe AXA

Axa est un Groupe international d'assurance d'origine française créée en 1985, présent dans 64 pays et comptant près de 107 millions de clients ainsi que 126 000 employés. En 2019, son chiffre d'affaires a été de 105 milliards d'euros, avec un bénéfice net à 3.86 milliards d'euros.

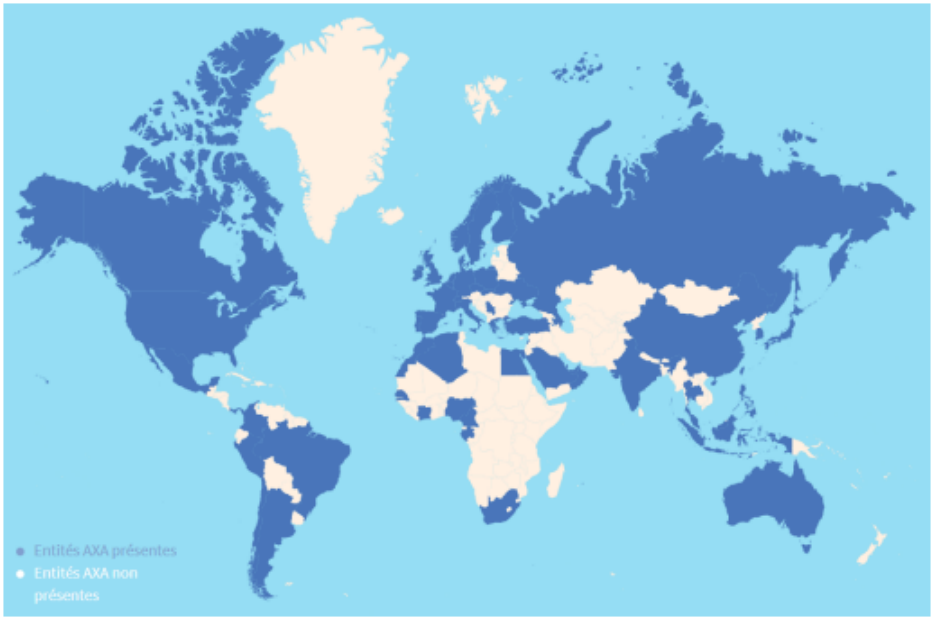


FIGURE 1.1 – Présence du groupe Axa dans 64 pays

L'entreprise est présente dans tous les secteurs de l'assurance, que ce soit l'assurance non-vie (biens, responsabilité civile, ...), de l'assurance vie (contrats de vie, de décès, d'épargne et de retraite, ...), de la santé prévoyance, ou encore la réassurance.

1.2 Le GIE Axa

Le siège mondial du Groupe AXA, le GIE AXA (Groupement d'Intérêt Economique), est situé au cœur du 8ème arrondissement de Paris. Il est l'élément central de la gouvernance du Groupe Axa et agit selon trois axes majeurs :

- De missions régaliennes : développement de la stratégie Groupe, établissement et publication des états financiers, gestion du bilan, gestion du portefeuille d'activités, protection et promotion de la marque, gestion des dirigeants et des valeurs du Groupe ainsi que représentation du Groupe vis-à-vis des parties prenantes externes,
- De missions de gestion du risque et de contrôle,
- De sa position et de son rôle d'actionnaire des entités opérationnelles du Groupe.

1.3 L'équipe Data Analytics, Actuarial and Underwriting Solutions

L'équipe Data Analytics, Actuarial and Underwriting Solutions est membre du GIE d'AXA. Son activité générale est relative aux activités d'assurance non-vie du Groupe, au support et transfert de bonnes pratiques aux entités. Pour cela, l'équipe se concentre sur des aspects métiers en proposant une plateforme Internet permettant de suivre les risques P&C (*Property & Casualty*), ainsi que la gestion des accumulations pour les couvertures cybernétiques ou liées aux catastrophes naturelles. La plateforme permet notamment un pilotage dynamique de l'activité d'assurance du Groupe et de réguler l'appétence et le transfert de risque des entités locales au travers d'indicateurs actuariels et de cartes de risques. Pour atteindre ces objectifs, l'équipe utilise son expertise en modélisation pour produire de fines évaluations.

1.4 Présentation du risque catastrophe naturelle

Dans ce mémoire, nous allons étudier en détail le péril séisme. Nous commençons d'abord par une présentation générale des catastrophes naturelles ainsi que leur assurabilité.

Une catastrophe naturelle peut être définie comme étant un évènement extrême, d'origine naturelle, entraînant des conséquences dramatiques sur les territoires qu'il affecte. On distingue plusieurs catégories d'évènements de catastrophes naturelles : les évènements géophysiques (comme les tremblements de terre), les évènements météorologiques (comme les tempêtes), les évènements hydrologiques (comme les inondations). La table 1.1 ci-dessous donne la classification des différents types de catastrophes naturelles.

Catégorie d'évènement	Type d'évènement
Géophysique	Tremblement de terre Mouvement de masse Activité volcanique
Météorologique	Tempête Température extrême
Hydrologique	Inondation Glissement de terrain Action des vagues
Climatologique	Sécheresse Vidange brutale d'un lac glaciaire Feu de forêts
Biologique	Epidémie Invasion d'insectes Accident animalier

TABLE 1.1 – Différents types de catastrophes naturelles

La décennie venant de s'achever (de 2010 à 2019) a été lourdement impactée par les catastrophes naturelles, générant une perte économique de 2 980 milliards de dollars, dont 845 milliards de pertes assurées, selon un rapport publié par le courtier Aon. Notamment frappées par le tsunami de Tohoku, qui a causé 210 milliards de dollars de pertes totales dont 40 milliards de pertes assurées et 15 880 morts, ces dix années ont été les plus coûteuses de l'histoire. Par ailleurs, la fréquence de ces évènements ne cesse de croître comme le résume la Figure 1.2.

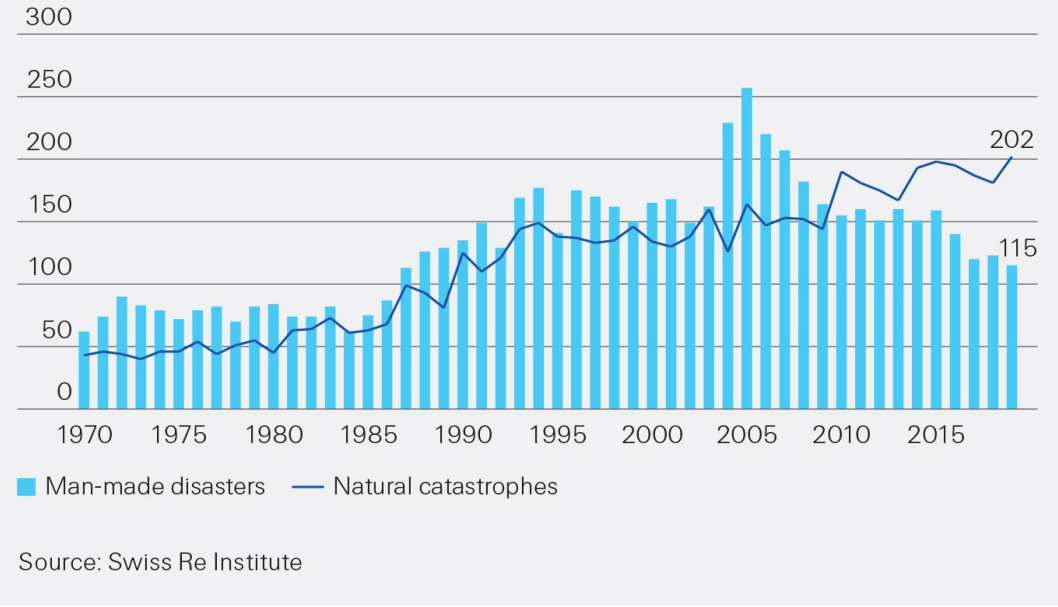


FIGURE 1.2 – Nombre d'évènements catastrophiques de 1970 à 2019 ([1])

La tendance est très largement haussière en ce qui concerne les catastrophes naturelles, qui ont même dépassé les catastrophes « man-made », c'est-à-dire causées par l'action de l'homme. Les évènements catastrophiques étant par définition des évènements engendrant de lourdes pertes économiques, les assurances ont un rôle très important

à jouer pour limiter ces pertes. On peut voir qu'une partie non négligeable des pertes économiques n'est pas assurée dans la figure 1.3 :

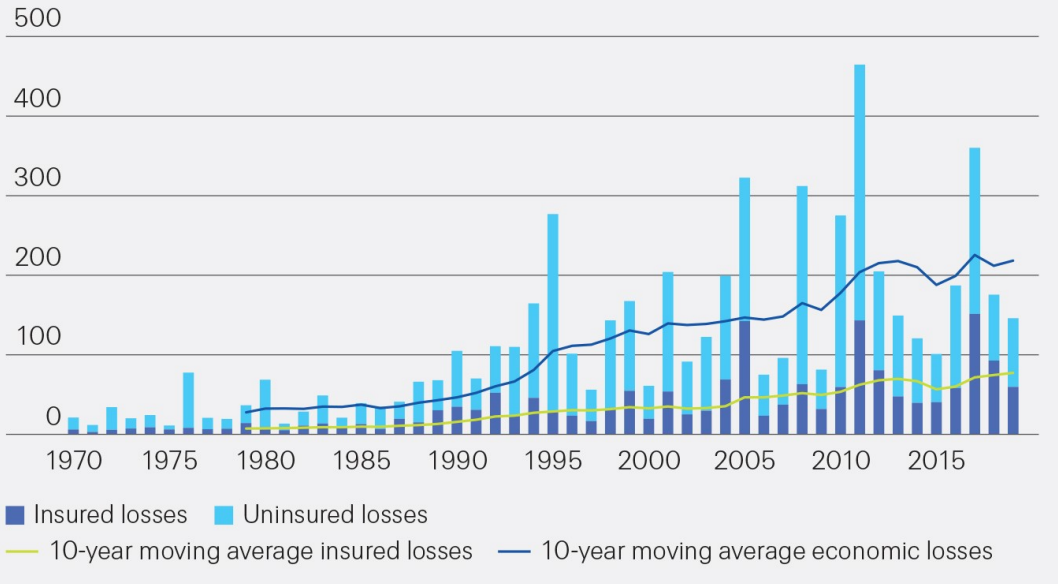


FIGURE 1.3 – Proportion des pertes assurées et non assurées ([1])

Comme le montre cette figure, la différence entre la perte économique et la perte assurée est élevée. Par exemple, en 2019, les pertes économiques totales résultant d'évènements catastrophiques s'élèvent à 140 milliards de dollars, pour 56 milliards de pertes assurées, ce qui représente un ratio d'assurabilité de seulement 40%, ce qui est peu, d'autant plus que les montants en jeu sont très élevés.
Jusqu'à présent, l'augmentation des pertes dues aux catastrophes naturelles peuvent être dues à l'accumulation croissante de l'exposition (humaine et physique) qui accompagne la croissance économique et l'urbanisation. Dans les décennies à venir, le changement climatique sera l'un des nombreux facteurs qui contribueront davantage à l'augmentation de ces pertes. En particulier, avec le réchauffement des températures mondiales, la fréquence des phénomènes météorologiques violents et les pertes qui en résultent augmenteront.

Dans ce mémoire, nous allons nous concentrer sur le péril séisme, qui est une des catastrophes ayant l'impact économique le plus important, comme le montre la distribution des pertes économiques par péril (figure 1.4) :

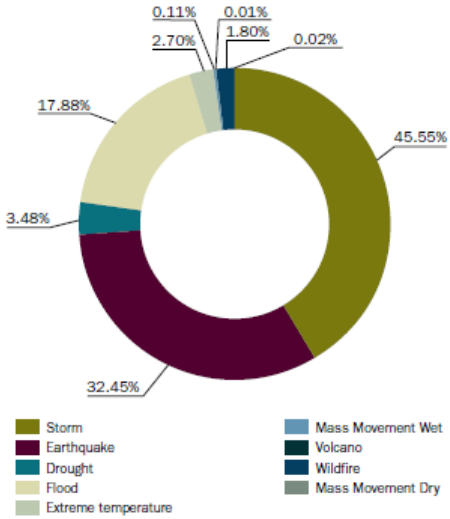


FIGURE 1.4 – Impact économique par catastrophe naturelle entre 2003 et 2013 ([2])

On peut voir que les tempêtes et les tremblements de terre sont les deux catastrophes naturelles avec l'impact économique le plus élevé (respectivement 45% et 32% de l'impact total).
Parmi les catastrophes naturelles, les tremblements de terre sont donc particulièrement dévastateurs, tant en termes de nombre de morts que de pertes économiques. Les séismes de basse fréquence, qui peuvent se produire toutes les quelques centaines ou milliers d'années et qui se caractérisent par une très haute libération d'énergie, peuvent engendrer d'énormes dégâts matériels et des pertes humaines dans une zone relativement restreinte, et également des pertes indirectes importantes liées à l'interruption d'activité potentielle. Par exemple, 2 séismes assez récents ont engendré une perte humaine dramatique.

— Sumatra-Andaman, de magnitude 9.3 en 2004, qui a causé 229 000 morts,

— Haïti, de magnitude 7 en 2010, qui a causé environ 250 000 morts.

Les séismes les plus destructeurs en termes de perte humaine se concentrent généralement dans les pays en développement, où la structure des bâtiments n'est pas adaptée pour subir de lourds séismes. En outre, des milliers d'évènements sismiques plus fréquents caractérisés par une libération d'énergie moyenne à faible se produisent chaque année, causant des pertes économiques et humaines cumulées importantes.

1.5 Focus sur le péril tremblement de terre

Un séisme ou tremblement de terre est une secousse du sol résultant de la libération brusque d'énergie accumulée par les contraintes exercées sur les roches. Il existe différents types de séismes, différenciés selon leur origine :

— les séismes tectoniques : ils sont de loin les plus fréquents et dévastateurs. Ils sont engendrés par la rupture d'une faille,

— les séismes d'origine volcanique : ils résultent de l'accumulation de magma dans la chambre magmatique d'un volcan,

— les séismes d'origine polaire : les glaciers et la couche de glace présentent une certaine élasticité, mais les avancées différenciées et périodiques (rythme saisonnier marqué) de coulées de glace provoquent des cassures dont les ondes élastiques génèrent des tremblements de terre,

— les séismes d'origine artificielle : de faible à moyenne magnitude, ils sont dus à certaines activités humaines telles que barrages, pompages profonds, extractions minières, explosions souterraines ou nucléaires, ou même bombardements.

La fréquence des séismes, appelée « sismicité », est plus élevée le long de bandes relativement étroites qui se situent le long des limites des plaques tectoniques. Cela est logique, puisque les séismes tectoniques, dont il est question dans ce mémoire, sont la conséquence des mouvements relatifs et des collisions de ces plaques. D'ailleurs, près de 90% des séismes du monde se produisent autour de l'océan Pacifique, près de la côte est de l'Asie et de la côte ouest de l'Amérique, dans une zone également appelée « ceinture de feu ». De plus, le Japon, la Californie, l'Amérique du Sud et l'Asie du Sud-Est sont les zones sismiques les plus actives du monde.

Les tremblements de terre qui se produisent dans des zones isolées, loin de la population et des activités humaines, causent rarement des morts ou des dégâts. Par exemple, en 1957, 1964 et 1965, trois grands tremblements de terre, respectivement de magnitude 8,6, 9,0 et 8,7 et de profondeur 33 km, 25 km et 36 km, se sont produits en Alaska. Bien qu'il s'agisse de trois des dix plus grands tremblements de terre observés lors du dernier siècle, ils ont causé un total d'environ 130 morts. À l'inverse, les tremblements de terre modérés qui touchent des zones très peuplées où les pratiques de construction sont médiocres peuvent faire de nombreuses victimes. Il n'est donc pas surprenant que, parmi les dix tremblements de terre les plus meurtriers depuis 1980, la plupart ont eu lieu au Moyen-Orient et en Asie du Sud-Est (Sumatra, Chine, Pakistan, Iran, Arménie, Inde et Turquie), régions caractérisées par une forte densité de population et dans des régions où il existe un manque relatif de structures antisismiques (par exemple, Haïti). Pour illustrer ce point, on peut comparer les impacts du tremblement de terre de 2003 dans le sud-est de l'Iran (Magnitude 6,6, profondeur 10km) avec ceux du tremblement de terre de Northridge en Californie en 1994 (Magnitude 6,7, profondeur 19km). Bien que ces deux évènements présentent des caractéristiques similaires, le nombre de victimes est significativement différent du fait des différences de vulnérabilité des bâtiments. Le tremblement de terre américain a fait 61 morts, tandis que celui qui a eu lieu en Iran a fait plus de 31 000 victimes. Paradoxalement, le séisme californien a entraîné une perte de 44 milliards de dollars US alors que l'évènement iranien n'a généré que 7 milliards de dollar US de pertes.

Dans les parties suivantes, nous allons présenter la théorie de base sur les séismes tectoniques.

1.5.1 Les bases de la tectonique des plaques

La terre est composée de trois couches principales : la croûte, le manteau et le noyau. La croûte et le manteau supérieur forment la lithosphère, qui est la couche rigide externe. Cette lithosphère, qui a une épaisseur d'environ 100 km, est divisée en plusieurs morceaux, appelés plaques tectoniques.
Les séismes tectoniques se produisent aux limites des plaques tectoniques. Les plaques

tectoniques se déplacent constamment et lentement, mais il arrive que la friction entre elles les bloque et les empêche de se déplacer. Le reste des plaques continue à se déplacer, ce qui entraîne une pression accrue sur la section bloquée. Finalement, cette section bloquée succombe à la pression, et les plaques se déplacent rapidement les unes devant les autres. Ce mouvement provoque un tremblement de terre tectonique. Les ondes d'énergie libérée se déplacent à travers la croûte terrestre et provoquent les secousses que nous ressentons à l'endroit du tremblement de terre.

1.5.1.1 Frontières des plaques tectoniques

Un séisme tectonique se produit à l'endroit où les plaques tectoniques se rencontrent, une zone appelée frontière. Trois types distincts de frontières, ou marges, ont été identifiées, dont les caractéristiques influencent la nature des tremblements de terre :

— **Frontière convergente** : lorsque deux plaques s'enfoncent l'une dans l'autre, elles forment une frontière convergente. Par exemple, la plaque océanique de Nazca, au large des côtes d'Amérique du Sud, le long de la Fosse du Pérou-Chili, s'enfonce dans la plaque sud-américaine et est subduite sous celle-ci. Ce mouvement soulève la plaque sud-américaine, créant ainsi les montagnes des Andes. La plaque de Nazca se brise en de plus petites parties qui sont bloquées pendant une longue période avant de se déplacer soudainement pour provoquer des tremblements de terre,

— Une **frontière divergente** se produit lorsque deux plaques s'éloignent l'une de l'autre, créant ainsi une nouvelle croûte, comme la dorsale médio-atlantique, qui s'étend de l'océan Arctique jusqu'au-delà de la pointe sud de l'Afrique. Sur des millions d'années, elle a provoqué des déplacements de plaques sur des milliers de kilomètres,

— Une **frontière transformante** se produit lorsque les plaques glissent horizontalement les unes sur les autres, sans détruire ni produire de croûte. Le mouvement des plaques crée des marges de plaques en zigzag et produit des séismes peu profonds. Le fond de l'océan abrite la plupart des failles transformantes, mais certaines, comme la zone de faille de San Andreas en Californie, se produisent sur terre.

1.5.1.2 Types de failles

Une faille est une surface tridimensionnelle où des blocs de roche se sont brisés. La roche située d'un côté de la faille passe devant la roche de l'autre côté. Une ligne de faille s'étend le long du sol à l'endroit où la faille coupe la surface de la Terre. Les failles sont de toutes tailles et on les trouve partout dans le monde. Lors d'un tremblement de terre, la roche située d'un côté de la faille glisse soudainement par rapport à l'autre côté, horizontalement, verticalement ou selon un angle quelconque entre les deux.

Les trois principaux types de failles sont :

— **Faille normal**e : une faille normale se forme lorsque le bloc situé au-dessus de la faille se déplace vers le bas par rapport au bloc situé en dessous,

— **Faille inverse** : une faille inverse (de poussée) se forme lorsque le bloc supérieur se déplace vers le haut et au-dessus du bloc inférieur,

— **Faille décrochante** : une faille décrochante se forme lorsque deux blocs glissent l'un sur l'autre dans une direction horizontale parallèle à la ligne de faille.

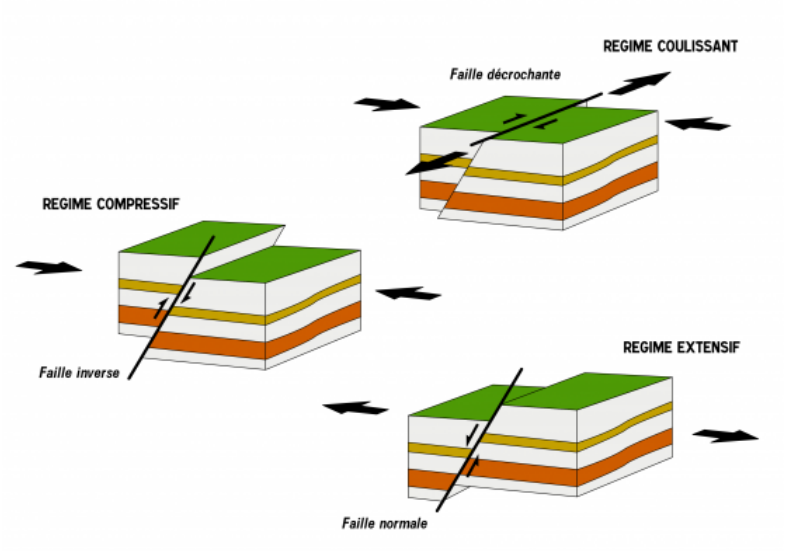


FIGURE 1.5 – Les trois types de failles ([3])

1.5.2 Mesure de l'intensité d'un séisme

La « force » ou intensité des secousses produites par un tremblement de terre sur un site donné peut être exprimée soit par des mesures qualitatives ou non instrumentales, soit par des mesures quantitatives ou instrumentales.
Avant le développement de la mesure instrumentale, une méthode descriptive était traditionnellement utilisée pour établir la taille des tremblements de terre. L'une des échelles les plus courantes était celle de Mercalli modifiée (MMI), qui exprime l'intensité des effets du tremblement de terre observés en un lieu donné sur la base d'une échelle discrète allant de I à XII, où I correspond à « généralement non ressenti par les personnes » et XII à « destruction totale ».
Les échelles instrumentales, d'autre part, décrivent quantitativement les niveaux de mouvement du sol sur la base d'enregistrements de mouvements forts réels. Les instruments utilisés pour mesurer les vibrations du sol produites par les tremblements de terre sont appelés sismomètres, ou sismographes.
Plusieurs échelles ont été définies pour estimer la taille d'un tremblement de terre à sa source, par la mesure instrumentale des secousses produites sur certains sites. L'une d'entre elles est l'échelle de Richter, qui est une échelle sans limite supérieure et qui exprime la relation logarithmique entre la taille du tremblement de terre et l'amplitude des plus grandes ondes enregistrées sur un sismogramme spécifique, en corrigeant la distance entre l'instrument d'enregistrement et l'épicentre du tremblement de terre. Sur la base de cette représentation, l'augmentation d'une unité de magnitude correspond approximativement à une multiplication par 30 de l'énergie libérée.
Bien que la magnitude de Richter soit l'échelle la plus couramment utilisée pour mesurer la taille d'un tremblement de terre, pour les séismes de grande intensité, une mesure plus précise aussi utilisée est la magnitude instantanée, M_w. La magnitude instantanée d'un tremblement de terre est une mesure de la quantité d'énergie estimée à partir des

relevés des sismomètres et elle est directement liée à l'étendue physique de la faille et aux propriétés mécaniques de la roche.

1.5.3 Conséquences des tremblements de terre

L'une des questions majeures dans la détermination des secousses du sol est la « réponse du site », qui désigne un ensemble de différents phénomènes physiques résultant de la propagation des ondes sismiques dans des formations géologiques proches de la surface ou dans des configurations géométriquement irrégulières à la surface de la terre. Du point de vue technique, cet effet est important car il peut entraîner une plus grande amplitude et une plus longue durée des secousses, ainsi que des modifications de leur fréquence. La gravité et l'étendue spatiale de ces effets dépendent de la profondeur de la couche de sol, de sa teneur en humidité et de la nature de la formation géologique.
En plus des bâtiments, d'autres structures telles que les routes, les chemins de fer, les ponts, les tunnels et les pipelines sont largement vulnérables aux dommages causés par les défauts de surface. Le moyen le plus efficace de limiter ces dommages est d'éviter les constructions à proximité immédiate des failles actives. Lorsque cela n'est pas possible, certaines mesures d'atténuation, telles que l'installation de pipelines au-dessus du sol ou l'utilisation de connexions flexibles, peuvent être envisagées.

À l'approche d'un site, les ondes sismiques sont capables de produire une large gamme d'effets, que l'on peut classer en effets « directs » et « indirects » :

Les effets directs

- **Tremblement (ou mouvement) du sol** : il s'agit de la vibration dans les trois directions du sol causée par le passage des ondes sismiques. C'est l'une des principales causes de l'effondrement partiel ou total des structures. Les accélérations, vitesses et déplacements induits par les tremblements de terre peuvent endommager ou détruire un bâtiment, à moins qu'il n'ait été conçu et construit ou renforcé pour résister aux séismes,
- **Défaut de surface** : Il s'agit du décalage ou la déchirure de la surface du sol par un mouvement différentiel le long d'une faille lors d'un tremblement de terre. Les déplacements varient de quelques millimètres à plusieurs mètres, et les dommages augmentent généralement avec un déplacement plus important. Le tassement, le soulèvement, le basculement et l'enfoncement des bâtiments ont été observés à la suite de plusieurs tremblements de terre dans le monde. Les dommages importants sont généralement limités à une zone étroite le long de la faille, bien que des ruptures subsidiaires puissent se produire à une distance donnée de la faille principale. La longueur des ruptures de surface peut aller jusqu'à plusieurs centaines de kilomètres. La proximité de la rupture de la faille est un facteur majeur dans le niveau de destruction, et donc la proximité d'une faille active est un facteur majeur dans l'estimation du risque sismique.

Les effets indirects

- Glissements de terrain,
- Chute de pierres,
- Liquéfaction des sols,

— Tsunamis et seiches,
— Inondations soudaines,
— Incendies,
— Contamination des eaux.

1.5.4 Gestion du risque sismique

Réduire les conséquences humaines, sociales et économiques des tremblements de terre et veiller à ce que les infrastructures et les installations restent pleinement opérationnelles en cas de séisme sont deux des plus grands défis auxquels la société est confrontée. En fait, la connaissance des pertes potentielles causées par les tremblements de terre est une étape essentielle pour sensibiliser le public, assurer une planification budgétaire adéquate, évaluer et allouer les ressources nécessaires à l'atténuation des risques et fixer les priorités en matière de modernisation.

Le risque sismique peut être défini comme la probabilité ou la possibilité de dépasser un niveau prédéfini de dommages et, par conséquent, de pertes dues à des tremblements de terre pour un élément à risque donné, sur une période de temps déterminée. Le risque sismique englobe plusieurs types de risques. En effet, il représente l'ensemble des pertes, y compris les décès, les blessures et les pertes économiques, générées par différents tremblements de terre, évaluées au cours d'une période donnée. Par conséquent, l'évaluation de la perte implique des efforts multidisciplinaires : sismologie, géophysique, géologie, ingénierie géotechnique, ingénierie structurelle, planification régionale et urbaine et gestion des risques. Bien entendu, dans le cadre des assurances, on s'intéressera particulièrement aux pertes économiques résultant de séismes.

Le risque sismique est la combinaison entre 3 éléments : l'aléa sismique en un point donné, la vulnérabilité des enjeux qui s'y trouvent exposés (personnes, bâtiments, infrastructures...) ainsi que leur exposition. Nous détaillerons plus tard ces 3 composantes (voir chapitre 2).

Il est impossible, par l'action de l'homme, de limiter l'occurrence et l'intensité des séismes (mis à part les séismes d'origine artificielle, mais ils sont peu fréquents et de faible intensité, donc sans graves conséquences). Mais il est possible d'augmenter la résistance des bâtiments exposés, et surtout protéger les vies humaines en cas de séisme : c'est l'objectif de la réglementation parasismique. Ces réglementations dépendent en général de deux paramètres :

— du **type de bâtiment** : distinction entre les bâtiments à « risque normal » pour lesquels « les conséquences d'un séisme demeurent circonscrites à leurs occupants et à leur voisinage immédiat » et les bâtiments à « risque spécial » pour lesquels « les effets sur les personnes, les biens et l'environnement de dommages même mineurs résultant d'un séisme peuvent ne pas être circonscrits au voisinage immédiat desdits bâtiments, équipements et installations »,
— de la **zone géographique** : certaines zones sont bien entendus plus risquées que d'autres.

La figure 1.6 représente la carte de risque sismique de l'Europe, avec un risque accru dans le sud :

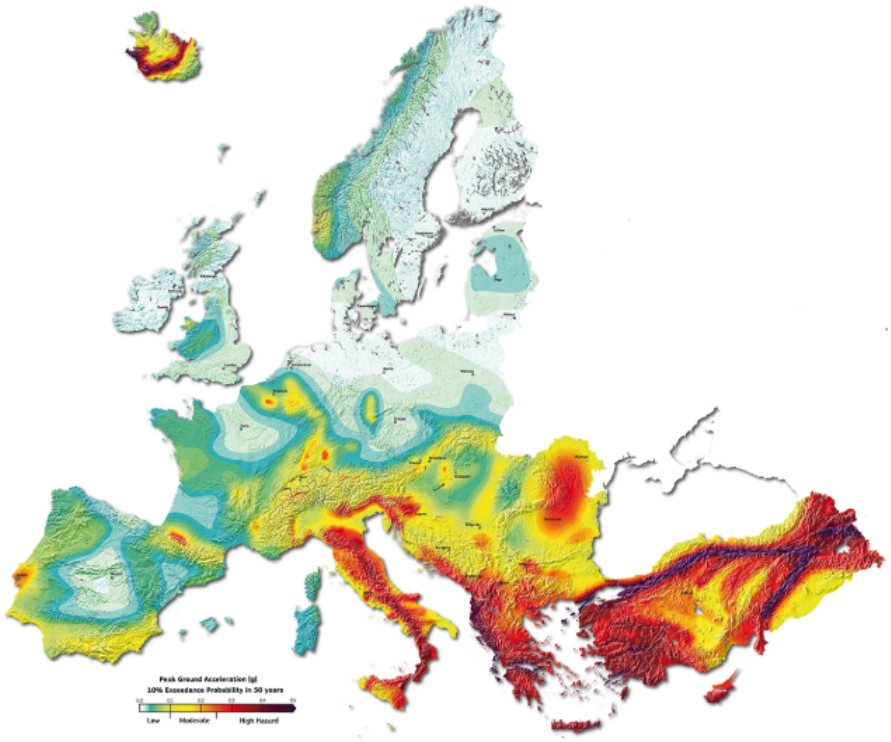


FIGURE 1.6 – Zonage sismique en Europe ([4])

Cette carte européenne des aléas représente l'accélération maximal au sol (PGA pour *Peak Ground acceleration* en anglais, en g) avec une probabilité de dépassement de 10% en 50 ans, soit une période de retour de 475 ans. Une présentation de cette grandeur sera fait dans la partie 2.2.6 de ce mémoire.

La souscription d'assurance contre le risque sismique dans les différents pays dépend de plusieurs facteurs, tels que la perception du risque, l'intervention de l'Etat, le niveau de risque sismique et le type de couverture. Par exemple, alors qu'en Nouvelle-Zélande, la réglementation de l'État rend l'assurance contre les tremblements de terre obligatoire, en Italie, le taux de pénétration (qui représente la part des assurés dans la population totale) de l'assurance est extrêmement faible (pour le tremblement de terre Émilie en 2012, seules 10% des pertes étaient assurées, alors que pour les tremblements de terre antérieurs touchant des zones résidentielles, seulement 2% des pertes l'étaient) parce que les incitations à l'achat privé d'une couverture d'assurance sont limitées.
Le niveau des pertes dépend généralement non seulement de la magnitude du tremblement de terre (évènements les plus importants), mais aussi de la vulnérabilité des bâtiments (principale cause du nombre de morts), et de la valeur des biens et activités exposés (directement liée aux pertes économiques les plus importantes). En effet, les tremblements de terre eux-mêmes ne blessent ni ne tuent les gens, et ils ne causent pas de dommages. Les bâtiments, les chutes d'objets et les effets induits (par exemple, les tsunamis, les glissements de terrain) le font.

Pour tarifer l'assurance dommages du péril tremblement de terre, il faut être capable de quantifier le risque. Pour cela, on utilise un modèle, appelé modèle catastrophe naturelle. Ce modèle, articulé en 4 modules, est présenté dans le chapitre suivant.

Chapitre 2

Modèle catastrophe naturelle

2.1 Introduction aux modèles catastrophe naturelle

Dans les régions rarement touchées par des évènements catastrophiques, il est très difficile de quantifier le risque pour un assureur. Ainsi, en utilisant uniquement les données de sinistres, il est très probable qu'un assureur sous-estime le risque lié à ces catastrophes, car les évènements forts engendrant des dommages lourds sont très rares, n'ont même parfois jamais frappé, mais sont probables. Par conséquent, il est nécessaire de compléter la vision historique du risque (venant des sinistres) par des scénarios fictifs probables afin de mieux appréhender le risque.
L'utilité des scénarios fictifs est bien résumé dans la figure 2.1. En effet, ces scénarios fictifs vont permettre d'obtenir la forme complète de la distribution des pertes liée au portefeuille.

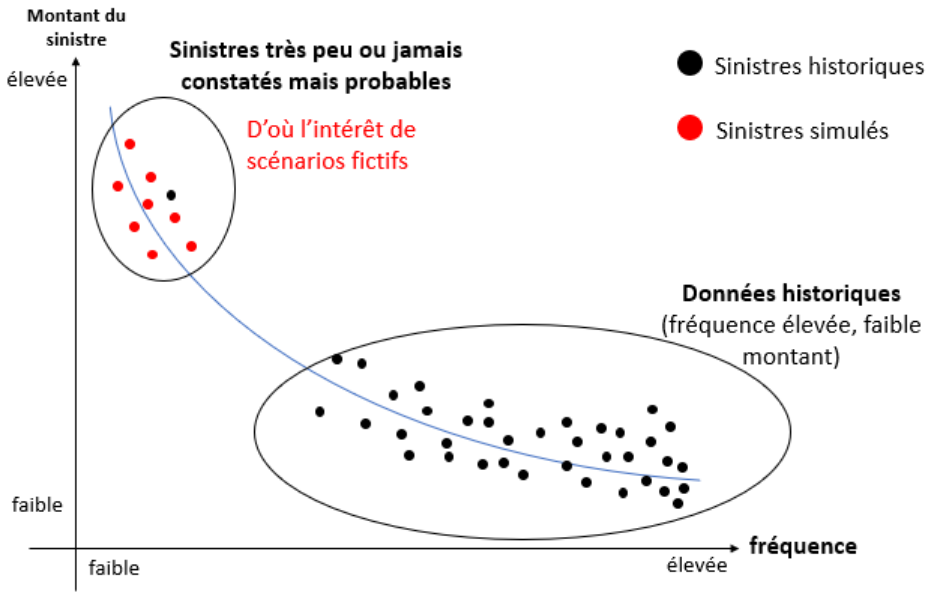


FIGURE 2.1 – Intérêt de l'utilisation de scénarios fictifs

Dans le cadre des catastrophes naturelles, l'utilisation d'un modèle catastrophe (ou modèle CAT) est donc plus que préconisée. Un modèle est une représentation simplifiée de la réalité. A partir de toutes les données dont on dispose, on va extraire un maximum d'informations permettant de représenter au mieux la répartition de la sinistralité liée au péril étudié. Ce modèle CAT est divisé en 4 modules, que nous allons expliciter dans les sous-parties suivantes :

- **Module Exposition :** il décrit les risques présents dans le portefeuille. On retrouve donc la localisation géographique des biens assurés, leur valeur assurée, ainsi que leurs caractéristiques,
- **Module Aléa :** il construit un catalogue d'événements contenant une multitude de scénarios réalistes et probables. Ce module nécessite des connaissances physiques et statistiques,
- **Module Vulnérabilité :** il associe à chaque bien assuré et touché par une catastrophe une perte économique en fonction des caractéristiques des biens ainsi que l'intensité de l'évènement. Ce module permet donc de mettre en commun les informations du module exposition et aléa afin d'estimer une perte économique par risque,
- **Module Financier :** il applique des conditions financières à chaque perte brute. Il s'agit donc de prendre en compte limites, franchises, réassurance, coassurance...

La figure 2.2 résume l'architecture du modèle CAT :

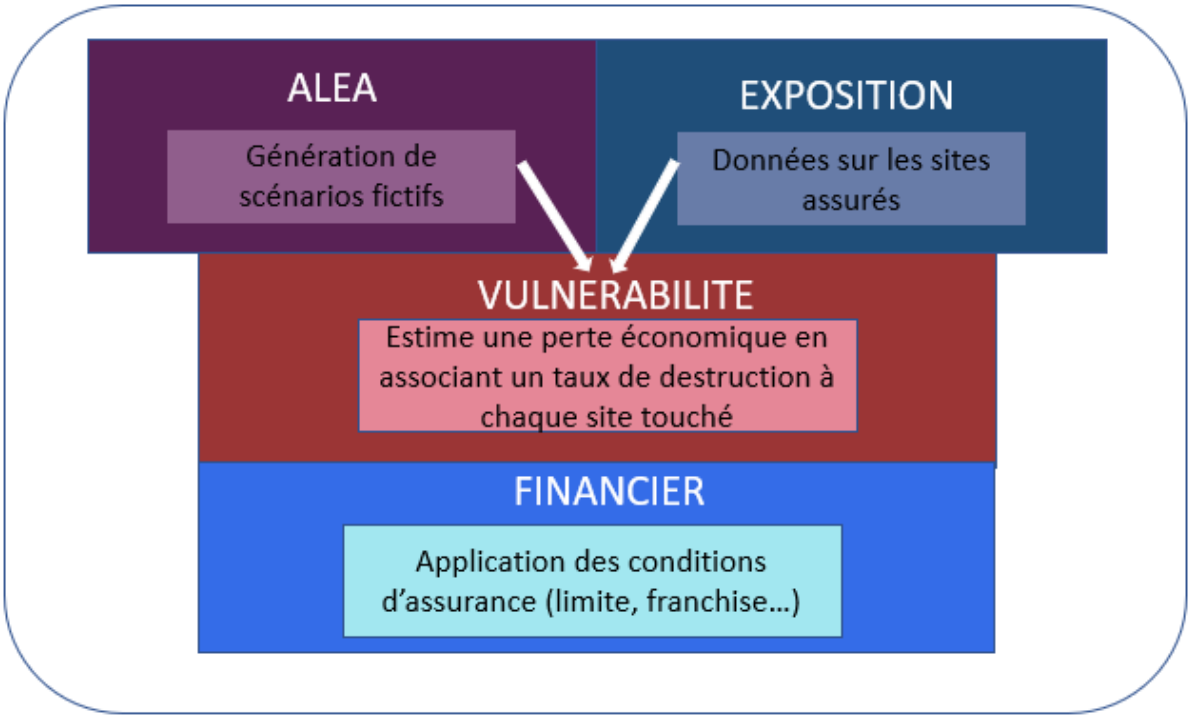


FIGURE 2.2 – Les modules d'un modèle CAT

Le modèle CAT permet donc de quantifier le risque réel d'un portefeuille en le décomposant en différentes composantes qui seront développées dans ce chapitre. En sortie de ce modèle, on retrouve différents résultats exploitables à différents niveaux :

— il permet d'obtenir une carte de risque basée sur le module aléa et présentant les zones les plus risquées vis à vis de la vulnérabilité, sans prise en compte du portefeuille d'exposition,

— il permet de simuler des scénarios d'évènements et les pertes qui s'en suivent afin d'en déduire un taux de destruction et ainsi pouvoir modéliser celui-ci par des méthodes telles que le GLM pour avoir un modèle multiplicatif,

— il fournit deux indicateurs importants :

1. l'AEP (*Aggregate exceedance probability*), défini par la perte annuelle cumulée pour une période de retour donnée. Cette grandeur permet de déterminer les quantiles des distributions de pertes, et en particulier celui à 99.5%, qui correspond précisément à la valeur du SCR requis par Solvabilité II pour le module Catastrophe en Assurance non-vie.

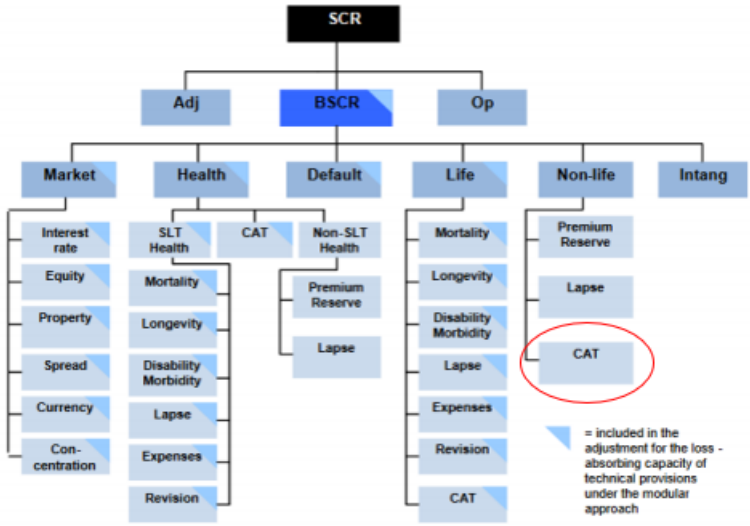


FIGURE 2.3 – SCR CAT (source : acpr.banque-france.fr)

2. l'OEP (*Occurrence exceedance probability*), défini par la perte annuelle maximale pour une période de retour donnée. L'étude de l'OEP est appropriée pour les contrats ne s'appliquant qu'aux risques rares et extrêmes. Les modalités et plus particulièrement la capacité d'un programme non proportionnel de réassurance sont choisies en fonction de l'OEP, qui renseigne sur les sinistres les plus graves.

— il permet de connaître et contrôler l'influence de l'ajout d'assurés au portefeuille. En effet, on considère que la limite d'un traité de réassurance (portée + priorité) est fixée au montant de la perte maximale survenant avec une probabilité de 0,5%. En d'autres termes il s'agit de l'OEP pour un temps de retour de 200 ans. Ainsi, si l'ajout d'un assuré influence l'OEP_{200}, il faudra étendre la couverture du traité ou alors couvrir la police avec de la réassurance facultative.

Dans les parties suivantes, nous allons rentrer dans le détail des modules du modèle CAT.

2.2 Module aléa

Comme précisé précédemment, l'objectif de ce module est de constituer un catalogue de séismes fictifs et probabilisés.

2.2.1 Catalogue sismique

Tout d'abord, il est nécessaire de récupérer l'historique des séismes qui se sont déroulés dans la zone étudiée pour constituer un catalogue sismique. Bien entendu, cet historique est limité aux données récoltées par l'homme. Par exemple, la base de données SHARE (Seismic Hazard harmonization in Europe) a recensé plus de 1500 séismes historiques autour de la Suisse (figure 2.4)

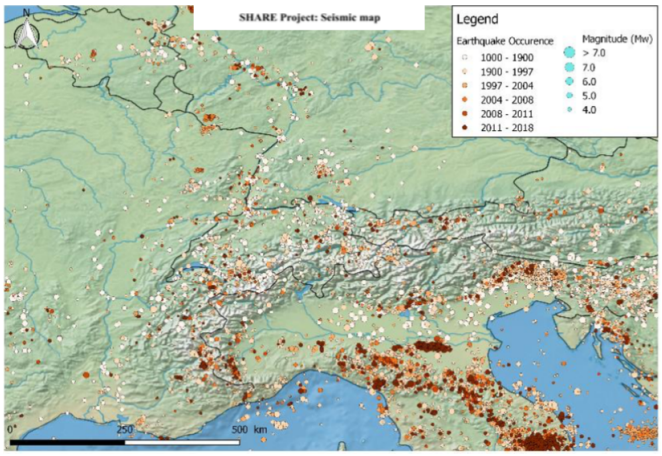


FIGURE 2.4 – Séismes historiques recensés depuis plus de 1000 ans

Les séismes enregistrés par un réseau de stations sismologiques composent la partie instrumentale du catalogue. La précision de leur localisation et de l'estimation de leur magnitude dépend fortement de la localisation des stations par rapport aux séismes. Les séismes ayant eu lieu avant la mise en place de tels instruments sont également pris en compte. Leur intensité est estimée par les dégâts occasionnés et reportés dans des comptes-rendus. La réalisation d'études d'aléa nécessite un catalogue de sismicité le plus complet et le plus homogène possible. Idéalement, il doit être constitué d'une seule et même mesure caractérisant les évènements sismiques. Dans les faits ceci est impossible et il est toujours nécessaire de compiler les informations de différents catalogues qui couvrent des périodes de temps différentes, des gammes de magnitudes différentes ou une région différente.

2.2.2 Définition des sources

Ensuite, on définit les sources que l'on trouve dans la zone étudiée. Dans le modèle, il y a 2 types de sources :

— Les zones sismotectoniques : il s'agit de zones géographiques dans lesquelles la probabilité d'occurrence d'un séisme de caractéristiques données peut être considérée homogène en tout point : ces zones s'articulent en général autour d'une même faille ou d'une même structure tectonique. Dans ces zones, la détermination du modèle

de récurrence se fait quasiment exclusivement à partir de la sismicité instrumentale et historique,

— Les failles : on retrouve 37 failles actives aux alentours de la Suisse (dans un rayon de 300 km), surtout dans les pays voisins.

La carte ci-dessous (figure 2.5) représente les caractéristiques et localisations des failles. Ces informations proviennent de la base de données INVG (*Instituto Natzionale di geofisca e Vulcanologia*).

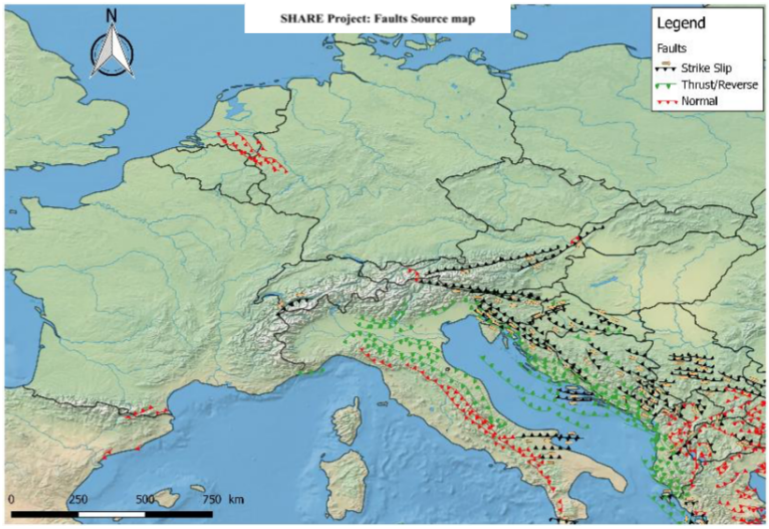


FIGURE 2.5 – Failles en Suisse et ses alentours

Pour chacune de ces zones sources considérées, un modèle de récurrence est défini.

2.2.3 Définition du modèle de récurrence

Historiquement, une première relation entre la magnitude (M) et le nombre de séisme ($N(M)$) sur une période de temps donnée a vu le jour en 1944, sur la base du modèle de Gutenberg-Richter (GR) :

$$N(M) = 10^{a-bM}$$

$N(M)$ est le nombre de séismes de magnitude supérieure ou égale à M par unité de temps, a et b sont des constantes (définies en général par région). 10^a représente le nombre total de tremblements de terre de magnitude supérieure ou égale à 0, et b est le taux de séismicité (avec une valeur proche de 1). Le paramètre b est d'autant plus élevé que la fréquence de séismes dans la zone étudiée est faible. Un des problèmes majeurs de cette première formule est qu'elle renvoie un résultat quelle que soit la valeur de M. Ainsi, une fréquence est associée aux valeurs très élevées de magnitude, même dans les zones où l'occurrence de sinistres à magnitude très élevée est impossible.

C'est pour cette raison qu'en 1969, une fonction prenant en compte, pour chaque source étudiée, une magnitude minimale (M_{min}) et maximale (M_{max}) est apparue la fonction suivante :

$$f_{GR}(x) = \frac{\exp\left[-\beta(x - M_{min})\right]}{1 - \exp\left[-\beta(M_{max} - M_{min})\right]} \mathbb{1}_{(M_{min} \leq x \leq M_{max})}$$

Il s'agit d'une fonction de densité (on vérifie en effet facilement qu'il s'agit d'une fonction positive d'intégrale égale à 1). Cette fonction s'appelle version tronquée de la loi de

Gutenberg Richter. Il s'agit actuellement du modèle de récurrence le plus couramment utilisé.
Ainsi le nombre de séismes s'exprime par la formule suivante :

$$N(m) = N(M_{min}) \int_{m}^{M_{max}} f_{GR}(x)dx$$

$$\iff N(m) = N(M_{min}) \frac{\exp\left[-\beta(m - M_{min})\right] - \exp\left[-\beta(M_{max} - M_{min})\right]}{1 - \exp\left[-\beta(M_{max} - M_{min})\right]} \mathbb{1}_{(M_{min} \leq m \leq M_{max})}$$

où $N(m) = 0$ pour $m < M_{min}$ et $m > M_{max}$.
Les magnitudes maximales et minimales sont déterminées en fonction de la sismicité et/ou des caractéristiques tectoniques localisées dans la zone.

La valeur de $N(M_{min})$ peut être plus ou moins facile à déterminer selon les sources. En effet, on distingue deux approches :

- Dans le cas de zones sismotectoniques, le nombre de données à disposition permet en général de déterminer statistiquement $N(M_{min})$,
- Dans le cas de failles, il y a moins de données à disposition ; il faut donc procéder autrement. Ce taux est estimé par le rapport entre le taux de moment annuel $\dot{M}o$ et le moment sismique maximal de la faille Mo_{max}, pondéré par une fonction correspondant au modèle sismique associé. Ces notions physiques ne seront pas étudiées en détail dans ce mémoire. Le lecteur qui veut en apprendre plus sur ces considérations pourra lire le papier de A. Deif and I. El-Hussain ([16])

2.2.4 Définition du modèle de probabilité

Un modèle de probabilité de la sismicité décrit la relation entre l'occurrence d'un ou plusieurs séismes d'une magnitude donnée en fonction du temps considéré. Les probabilités sont déterminées à partir des fréquences obtenues grâce par exemple aux formules citées précédemment. Nous allons lister différents modèles d'occurrences de tremblement de terre possibles. Ces modèles peuvent ensuite être injectés dans un calcul probabiliste de l'aléa sismique.

2.2.4.1 Estimation de l'occurrence par un processus de Poisson

Le modèle classique et largement utilisé considère que l'occurrence des séismes suit un processus de Poisson, dont voici la définition :

Définition (Processus de Poisson). Un processus de Poisson de densité $\lambda > 0$ est un processus de comptage $(N(t))_{t \geq 0}$ tel que :

1. Le processus est à accroissements indépendants : $\forall t_0 \leq t_1 < ... < t_k$, les varianles aléatoires $N_{t_k} - N_{t_{k-1}}, ..., N_{t_1} - N_{t_0}$ sont indépendantes. Autrement dit, les nombres d'occurrences dans des intervalles de temps disjoints sont indépendants les uns des autres.
2. Pour tout $(s, t) \in \mathbb{R}_+^2$, N(s+t)-N(s) suit la loi de Poisson de paramètre λt. Autrement dit, la probabilité d'occurrence dans un intervalle de temps est proportionnelle à la longueur de cet intervalle, le coefficient de proportionnalité étant λ.

On rappelle également la définition d'un processus de comptage :

Définition (Processus de comptage). Un processus de comptage est une suite de variables aléatoires réelles $(N(t))_{t\geq 0}$ telles que :

1. $N(0) = 0$;
2. $\forall$t≥ 0, $N(t) \in \mathbb{N}^*$;
3. $t \longmapsto N(t)$ est croissante.

Le paramètre λ peut être fixé, pour plusieurs intervalles de magnitude, à partir de la loi tronquée de Gutenberg-Richter présentée précédemment.

Les hypothèses du processus de Poisson impliquent que les séismes sont supposés se produire de manière indépendante et aléatoire dans le temps et l'espace. Dans la réalité les séismes ne sont pas indépendants les uns des autres. En effet, il peut exister par exemples des répliques de gros séismes qui ont lieu à la suite d'un séisme majeur et qui dépendent clairement de celui-ci. L'occurrence d'un séisme modifie la capacité de la faille dans les alentours à générer d'autres séismes. C'est pourquoi des méthodes de traitement des catalogues de sismicité (appelées méthodes de declustering) ont été développées dans le but de respecter au maximum cette notion d'indépendance, en retirant tous les évènements sismiques correspondant à des séismes précurseurs et des répliques.
Le modèle de Poisson est le modèle le plus simple à mettre en place et le plus couramment utilisé. Il s'agit d'un modèle sans mémoire, c'est à dire que la probabilité d'occurrence est indépendante de la date du dernier séisme. Ce modèle est particulièrement approprié lorsqu'aucune autre information que le taux de récurrence n'est disponible. De ce fait, il n'intègre pas la physique du cycle sismique, selon laquelle la probabilité d'occurrence d'un séisme majeur sur une faille ayant déjà produit un séisme majeur est très faible tant que le temps écoulé n'a pas permis une nouvelle accumulation de contraintes sur la faille. Cependant, seules quelques régions sont assez documentées pour pouvoir utiliser des modèles à mémoire, ce qui explique la large utilisation des modèles poissoniens. Le modèle de Poisson est celui qui a été utilisé pour le modèle interne CAT dans ce mémoire.

Dans la partie suivante, nous allons développer quelques méthodes de declustering, méthodes permettant de respecter au maximum la notion d'indépendance nécessaire pour les modèles de Poisson.

2.2.4.2 Quelques méthodes de declustering

Dans cette partie, nous allons expliciter des méthodes de declustering, qui sont notamment utilisées pour l'application de modèles de Poisson dans un but de respecter au maximum l'hypothèse d'indépendance du processus de Poisson. Ces méthodes permettent de décomposer un catalogue de séismes en 3 types de séismes :

— les séismes précurseurs : il s'agit des séismes qui précèdent les chocs principaux,

— Les chocs principaux, qui sont parfois précédés de séismes précurseurs et généralement suivis de répliques,

— Les répliques, qui ont lieu après les séismes majeurs ou chocs principaux.

Les séismes précurseurs et les répliques sont des séismes dépendants, tandis que les chocs principaux sont indépendants. Pour toutes les méthodes de declustering, il est nécessaire de définir une distance en termes de temps et d'espace, ainsi qu'un seuil afin de pouvoir décomposer le catalogue.

Méthode par fenêtrage

Les techniques de fenêtrage sont un moyen simple d'identifier les chocs principaux et les répliques. Pour chaque tremblement de terre du catalogue de magnitude M, les chocs ultérieurs sont identifiés comme des répliques s'ils se produisent dans un intervalle de temps déterminé T(M), et dans un intervalle de distance L(M). Dans le cas où un séisme B de magnitude plus élevée est capté dans le fenêtrage initial déterminé à partir de la magnitude d'un séisme A, le fenêtrage sera réinitialisé en prenant B comme référence et A sera considéré comme un précurseur.

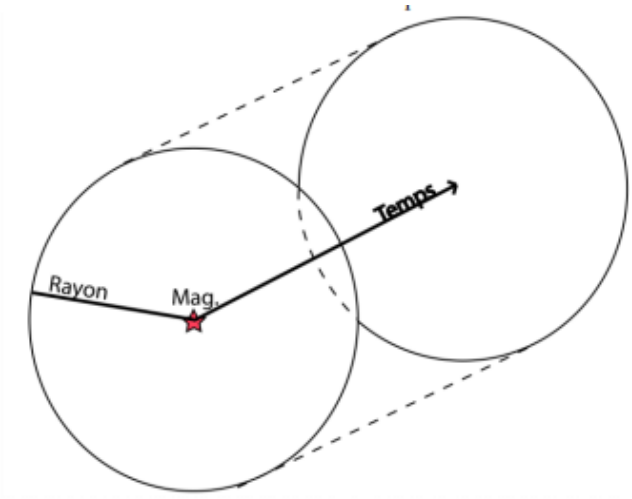


FIGURE 2.6 – Principe de la méthode par fenêtrage ([9])

Pour avoir une idée des ordres de grandeur utilisées pour le fenêtrage, on retrouve dans la table 2.1 des valeurs particulières suite aux travaux de Gardner et Knopoff en 1974 et permettant l'identification des séismes dépendants, c'est à dire les précurseurs et répliques :

M	L(km)	T(days)
2.5	19.5	6
3	22.5	11.5
3.5	26	22
4	30	42
4.5	35	83
5	40	155
5.5	47	290
6	54	510
6.5	61	790
7	70	915
7.5	81	960
8	94	985

TABLE 2.1 – Fenêtres d'identification des séismes dépendants (Gardner et Knopoff, 1974)

Cette méthode constitue l'une des formes les plus simples d'identification des séismes dépendants. Cependant, elle reste très simpliste. En effet, elle ignore les répliques secondaires et d'ordre supérieur (c'est-à-dire les répliques de répliques) : si un tremblement de terre C tombe dans les fenêtres de déclenchement des deux chocs principaux potentiels A et B, alors seul le plus grand choc A ou B est conservé comme le véritable choc principal de C, indépendamment de la possibilité que C puisse être sensiblement plus proche dans l'espace et le temps que l'autre choc. L'extension de la faille pour les chocs de plus grande ampleur n'est également pas pris en compte, étant donné qu'on considère des fenêtres spatiales circulaires.

Méthode par Cluster

Reasenberg (1985) a introduit une méthode d'identification des répliques en reliant les tremblements de terre à des clusters selon des zones d'interaction spatiales et temporelles. L'algorithme de Reasenberg permet de relier le déclenchement des répliques au sein d'un groupe de séismes : si A est le choc principal de B, et B le choc principal de C, alors tous les séismes A, B et C sont considérés appartiennent à un même groupe. Lors de la définition d'un groupe, seul le plus grand tremblement de terre est finalement conservé pour être le choc principal du cluster.

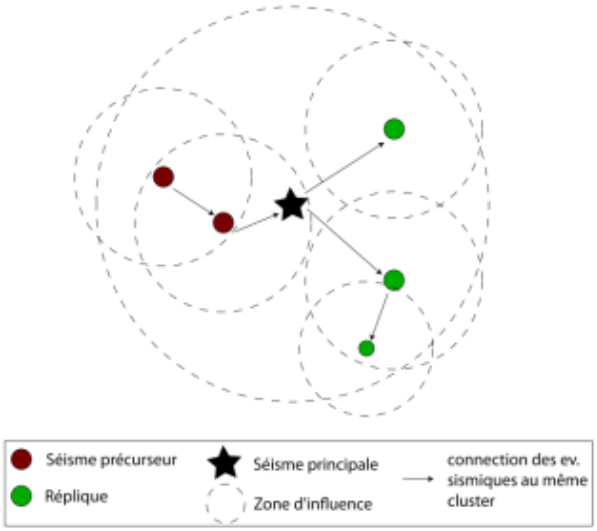


FIGURE 2.7 – Principe de la méthode par cluster ([9])

De la même manière que pour la méthode de fenêtrage, les zones d'interactions des séismes sont modélisées à partir de paramètres spatio-temporels. La relation d'interaction spatiale de Reasenberg est définie par le seuil défini par la fonction suivante : d(km) $= 0.4M_0 - 1.943 +$ k, où k vaut 1 pour la distance au séisme principal et 0 pour le dernier séisme du cluster. L'extension temporelle de la zone d'interaction est basée sur la loi d'Omori. Tous les séismes liés définissent un cluster, pour lequel le plus grand tremblement de terre est considéré comme le choc principal et les petits tremblements de terre sont divisés en répliques et précurseurs. Cet algorithme est très populaire dans la communauté sismologique. Le dernier algorithme disponible (CLUSTER 2000, disponible sur le site USGS) regroupe les paramètres écrits dans la table 2.2 :

Paramètre	Standard	Min	Max
T_{min}	1	0.5	2.5
T_{max}	10	3	15
p_1	0.95	0.9	0.99
x_k	0.5	0	1
x_{meff}	1.5	1.6	1.8
r_{fact}	10	5	20

TABLE 2.2 – Paramètres d'input pour la méthode de Reasenberg

avec

— $T_{min}(jours)$ est le délai minimal pour observer le prochain séisme dans un même cluster,

— $T_{max}(jours)$ est le délai maximal pour observer le prochain séisme dans un même cluster,

— p_1 est la probabilité de détecter un prochain événement dans les délais considérés,

— x_{meff} est la limite inférieure de magnitude pris en compte dans le cluster,

— x_k est un coefficient qui permet de calculer la limite inférieure de magnitude à prendre en compte dans un cluster : $x_{meff} = x_{meff} + x_k M$, où M est l'ampleur de l'événement le plus important du groupe,

— $r_{fact}(km)$ est un facteur applicable au rayon a de l'évènement le plus récent dans le but de considérer sa potentielle association avec l'évènement principal du cluster.

Declustering stochastique

Alternative aux méthodes déterministes de declustering citées précédemment, il existe également des méthodes de séparation probabilistes. Nous allons en développer une dans cette partie, basée sur le modèle ETAS. Ce modèle peut être représenté par une intensité conditionnelle, définie par la probabilité qu'un événement se produise à l'instant t en fonction de l'historique :

$$\lambda(t,x,y) = \mu(x,y) + \sum_{k,t_k<t} \kappa(m_k) g(t-t_k) f(x-x_k, y-y_k|m_k)$$

où $\mu(x,y)$ est une fonction d'intensité supposée être indépendante du temps, et les fonctions $g(t)$ et $f(x,y|m_k)$ sont respectivement les fonctions du moment de l'occurrence et du lieu. $\kappa(m_k)$ représente le nombre d'évènements attendus dépendant d'un premier évènement de magnitude m_k. Dans cette formule, on peut clairement voir que le risque à l'instant t dépend de l'historique. Chaque événement du passé (c'est-à-dire les $t_k < t$) contribue au risque à l'instant t, d'une valeur $\kappa(m_k)g(t-t_k)f(x-x_k, y-y_k|m_k)$. En règle générale, il est préférable de structurer la fonction g de sorte que les points du passé lointain ont moins d'influence que les points plus récents, et la fonction f de telle sorte que les points éloignés ont moins d'influence que les points proches.
Supposons que les événements soient numérotés dans l'ordre chronologique de 1 à N. Dans l'équation précédente, la probabilité qu'un événement j soit déclenché par le i$^{\text{ème}}$ événement est, pour $j > i$,

$$\rho_{ij} = \frac{\kappa(m_i) g(t_j - t_i) f(x_j - x_i, y_i - y_j|m_i)}{\lambda(t_j, x_j, y_j)}$$

Cette quantité représente la contribution relative du i$^{\text{ème}}$ événement au taux d'occurrence au moment et le lieu de l'événement j. De plus, la probabilité que l'événement j soit un événement déclenché (ou évènement dépendant) vaut donc :

$$\rho_j = \sum_{i=1}^{j-1} \rho_{ij}$$

Et la probabilité que l'évènement j soit un évènement indépendant vaut :

$$\phi_j = 1 - \rho_j = 1 - \sum_{i=1}^{j-1} \rho_{ij}$$

En utilisant les valeurs définies précédemment, il est alors possible de faire tourner un algorithme afin de séparer les évènements en évènements indépendant et dépendants. Cet algorithme, qui s'appelle classification stochastique de clusters de séisme, est précisé au début de la page suivante :

1. Calculer ϕ_j et ρ_{ij} pour $j = 1, 2, ..., N$ et $i = 1, 2, ..., j-1$.
2. Pour chaque évènement j, $j = 1, ..., N$, générer une variable aléatoire uniforme sur $[0, 1]$ U_j.
3. Pour chaque j, si $U_j < \phi_j$, sélectionner j comme un évènement initial. Sinon, I_j est un évènement parent de l'événement j où $I_j = Min(\text{k} : \phi_j + \sum_{i=1}^{k-1} \rho_{ij} \geq U_j, 1 \leq k < j)$.

On peut illuster cet algorithme par le schéma de la figure 2.8 :

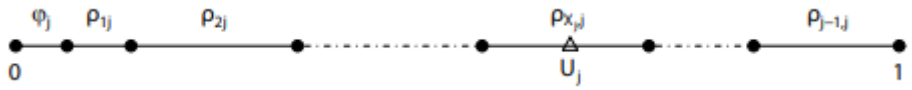


FIGURE 2.8 – Exemple d'une étape de l'algorithme

Dans cet exemple, x_j sortira comme l'évènement parent de j.

Cet algorithme nous permet de créer des "familles de séismes, il suffit ensuite de sélectionner dans chaque groupe le séisme le plus gros comme séisme indépendant pour construire le catalogue de fond.
Ces méthodes ne sont que 3 méthodes parmi une longue liste, mais ce sont les plus utilisées (surtout la méthode par cluster et la méthode par fenêtrage).

2.2.4.3 Estimation de l'occurrence par chaines de Markov

Bien que dans le modèle interne ainsi que dans la plupart des modèles CAT du péril séisme, le modèle de Poisson est utilisé comme modèle de probabilité, il peut être intéressant d'évoquer, dans les grandes lignes, une autre méthode basée sur les chaines de Markov afin de prédire l'occurrence des séismes. Cette méthode utilise les informations historiques existantes ainsi que sur les données géologiques avec un accent particulier sur la « mémoire » de la localisation spatiale. Elle comprend également l'évaluation de l'influence de la localisation d'un tremblement de terre sur la localisation du prochain tremblement de terre, c'est-à-dire la dépendance vis à vis de la localisation des séismes qui se sont déjà produits. Cette méthode doit être utilisée dans les zones avec un fort historique d'évènements, ce qui est très rare pour les séismes (cette méthode n'est par exemple par applicable en Suisse). Elle a tout de même été expérimentée par exemple pour la Turquie où l'activité sismique est très élevée.

Les notations que nous allons utiliser dans cette partie sont les suivantes :

— $X = (X_i)_{i \in \mathbb{N}}$ un processus stochastique,
— $i_0, i_1, ..., i_{n-1}, i, j$ les états du processus de respectivement $X_0, X_1, ..., X_{n-1}, X_n, X_{n+1}$ où $n \in \mathbb{N}$.

Une chaine de Markov est une séquence de variables aléatoires telles que pour tout n, le « prochain état » du processus X_{n+1} est indépendant des états « passés », c'est à dire de $X_0, X_1, ..., X_{n-1}$. Mathématiquement, cela donne :

Définition (Chaine de Markov). Le processus X est une chaine de Markov si

$$\mathbb{P}(X_{n+1} = j | X_0 = i_0, X_1 = i_1, ..., X_{n-1} = i_{n-1}, X_n = i) = \mathbb{P}(X_{n+1} = j | X_n = i) = P_{ij}$$

Compte tenu d'un catalogue sismique et d'un instant initial, au cours de chaque intervalle de temps Δt , l'état de chaque région de la zone étudiée peut être l'une des deux valeurs suivantes : 0 ou 1, correspondant, respectivement, à l'absence ou à la présence de tremblements de terre d'une magnitude supérieure ou égale à une valeur seuil M_0. Prenons l'exemple d'un pays divisé en quatre régions. On a donc $2^4 = 16$ états possibles. Pour un intervalle donné Δt, s'il n'y a pas de tremblements de terre dans toutes les régions, on écrit 0000 pour l'état 0, s'il y a un (ou plusieurs) tremblement(s) de terre uniquement dans la région 1, on écrit 1000 pour l'état 2, s'il y a un (ou plusieurs) tremblement(s) de terre uniquement dans la région 2, on écrit 0100 pour l'état 3,..., et s'il y a un (ou plusieurs) tremblement(s) de terre dans toutes les régions, nous écrivons 1111 pour l'Etat 15. Par conséquent, l'ensemble des états possibles est regroupé dans la table 2.3.

Etat	Région			
0	0	0	0	0
1	1	0	0	0
2	0	1	0	0
3	1	1	0	0
4	0	0	1	0
5	1	0	1	0
...				
14	0	1	1	1
15	1	1	1	1

TABLE 2.3 – Etats possibles pour les différentes régions

Un paramètre important à déterminer est l'intervalle de temps Δt qui est utilisé pour déterminer les états du système. Pour une trop petite valeur de Δt, l'état 0 (aucun tremblement de terre dans aucune région) sera le plus fréquent, et la transition de l'état 0 à l'état 0 sera dominante. De plus, les probabilités autres que P_{00} peuvent même être très faibles, de sorte qu'il soit très peu probable de passer dans un autre état que 0.
Inversement, pour une grande valeur de Δt, l'état 15 (tremblements de terre dans toutes les régions) sera plus fréquent que les autres, et la transition de l'état 15 à l'état 15 sera dominante.

Enfin, une dernière propriété utile est la suivante :

Propriété. La loi de la chaîne de Markov $X = (X_n)_{n\geq 0}$ est caractérisée par le couple constitué de sa matrice de transition P, et de sa loi initiale (la loi de X_0) : pour tout $n \geq 1$ la loi jointe de $(X_0, X_1, \ldots, X_n)$ est donnée par

$$\mathbb{P}\left(X_0 = i_0, X_1 = i_1, \ldots, X_{n-1} = i_{n-1}, X_n = i_n\right) = \mathbb{P}\left(X_0 = i_0\right) \, P_{i_0,i_1} \, P_{i_1,i_2} \, \ldots \, P_{i_{n-2},i_{n-1}} \, P_{i_{n-1},i_n}.$$

La matrice de transition P doit être déterminée selon l'historique des séismes dans la zone étudiée et selon l'intervalle de temps choisi. Nous n'entrerons pas plus dans les détail au sujet des chaines de Markov pour le modèle de probabilité des séismes, d'autant plus que dans le cas précis de la Suisse, ce modèle est inutilisable car il n'y a pas assez d'historique

de sinistres. Si l'on désire en savoir plus sur le sujet, un papier plus complet qui s'intitule a été écrit par Serpil Ünal et Salih Çelebioglu ([10])

2.2.5 Génération d'un catalogue stochastique

Le catalogue stochastique est généré en tenant compte des paramètres de chaque source du modèle (zone et faille). Il est composé des paramètres suivants :

- **Magnitude :** Un ensemble d'événements est généré pour chaque source pour une gamme de magnitude allant adaptée à la source en question,
- **Localisation :** La localisation des événements stochastiques dépend du type de source :
  - Dans le cas d'une faille, les événements sont générés sur le plan de projection de la surface de la faille,
  - Dans le cas d'une zone sismotectonique, les évènements sont localisés selon une grille de taille dépendante de la magnitude.
- **Fréquence :** Chaque événement est associé à des fréquences (valeurs moyennes et incertitudes), obtenues gràce à la formule décrite précédemment.

Le catalogue stochastique est aussi composé d'autres paramètres comme la profondeur et la géométrie de la rupture, mais nous ne détaillerons pas ces points dans ce mémoire.

2.2.6 Génération d'empreintes

Un catalogue stochastique est généré en tenant compte des caractéristiques de la source sismique (emplacement et profondeur de l'épicentre, géométrie de la rupture, magnitude). Cet ensemble d'événements est pleinement représentatif du niveau de connaissance de chaque source. Les empreintes sont calculées avec l'outil Openquake (application open source permettant de calculer les risques sismiques) pour chaque événement susceptible de générer des dommages dans la zone d'étude. Ce calcul prend en compte :

- Les équations GMPE (pour *Ground Motion Prediction Equation*, c'est à dire équations de prédiction du mouvement du sol). Les équations de prédiction des mouvements du sol, ou relations d'≪ atténuation ≫, permettent de prédire le niveau de secousse du sol et l'incertitude qui lui est associée sur un site ou un emplacement donné, en fonction de la magnitude du séisme, de la distance entre les sources et les sites, des conditions locales du sol, du mécanisme de faille, etc. Les GMPE sont utilisées efficacement pour estimer les mouvements du sol en vue d'une utilisation dans les analyses probabilistes des risques sismiques.
  Selon le contexte sismologique, différentes équations codées dans le moteur Openquake sont utilisées pour générer l'empreinte de chaque événement,
- Des informations sur le type de sol en termes de Vs30. Cette grandeur, exprimée en m/s, représente la vitesse moyenne des ondes de cisaillement à 30 mètres de profondeur. Le risque est augmenté avec une valeur faible de Vs30, correspondant à un sol mou. La carte (figure 2.9) ci-dessous représente le type de sol sur le territoire suisse.

FIGURE 2.9 – Carte vs30

Le résultat d'une analyse probabiliste des risques sismiques (PSHA) peut prendre différentes formes :

— Une courbe d'aléa sismique, qui représente la fréquence (ou le taux) annuelle de dépassement d'une mesure d'intensité des secousses du sol sur un site donné. Cette grandeur est ici l'accélération maximale du sol (*PGA*, pour Peak Ground Acceleration). Cette grandeur est grandement utilisée lorsqu'il s'agit de tremblements de terre et est un très bon indicateur du risque sismique. La figure 2.10 présente un exemple de courbe d'aléa sismique.

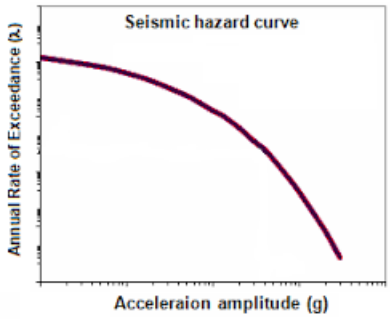


FIGURE 2.10 – Courbe d'aléa sismique

Elle représente la fréquence annuelle moyenne de dépassement d'une mesure de l'intensité du mouvement du sol (ici le PGA) pour un site donné. On peut observer que des fréquences plus élevées correspondent à une valeur de PGA plus faible, et inversement, ce qui est tout à fait logique,

— Une autre sortie (qui nous intéresse particulièrement) est une carte de risque (caractérisée par le PGA, c'est à dire l'accélération maximale du sol). Le niveau de probabilité indiqué sur les cartes des aléas est défini par les autorités locales en fonction du niveau de protection contre les tremblements de terre qu'elles décident d'assurer. Le niveau 10% en 50 ans (c'est à dire un PGA ayant 10% de chance de se produire en 50 ans) est un niveau de probabilité de dépassement typique adopté dans de nombreux pays à sismicité moyenne ou haute. Cela correspond au PGA pour une période de retour de 475 ans. La figure 2.11 représente la carte de risque de la Suisse. On remarque que le risque est relativement faible, avec cependant quelques zones au sud, à l'est et au nord avec un risque plus élevé.

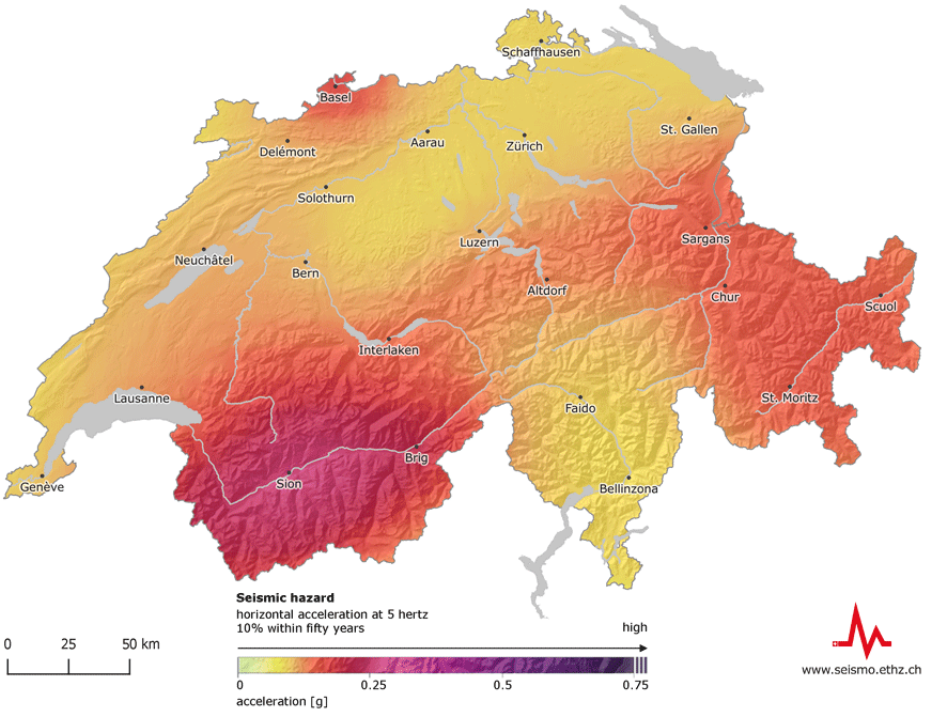


FIGURE 2.11 – PGA du modèle pour une période de retour de 475 ans (seismo.ethz.ch)

Accélération maximale du sol

Nous allons maintenant présenter plus en détail ce qu'est l'accélération maximale du sol (*Peak Ground Acceleration*), qui est une variable très importante dans la suite de l'étude car c'est avec cette grandeur qu'on définit les cartes de risques comme le montre la figure 2.11.
L'accélération maximale du sol (PGA) est égale à l'accélération maximale du sol qui s'est produite lors d'un tremblement de terre à un endroit donné. Parfois, le PGA est divisé en deux composantes, l'une horizontale et l'autre verticale (car les séismes se produisent généralement dans plusieurs directions). Les PGA horizontaux sont plus grands que les verticaux, mais ce n'est pas toujours vrai, surtout à proximité des grands séismes. Le PGA est un paramètre important (également connu sous le nom de mesure d'intensité) pour l'étude des tremblements de terre par des experts. Contrairement aux échelles de Richter et de magnitude instantanée, il ne s'agit pas d'une mesure de l'énergie totale d'un tremblement de terre, mais plutôt de la force avec laquelle la terre tremble en un point géographique donné.

Nous venons d'évoquer le fait qu'il était équivalent de parler de PGA pour une période de retour de 475 ans, et de PGA ayant une probabilité de 10% d'être atteint en 50 ans. Nous allons prouver cette équivalence. Si x est la valeur du PGA pour une période de retour de 475 ans, alors la probabilité que x ne soit jamais atteint en 50 ans vaut $(1-\frac{1}{475})^{50} = 90\%$. Ainsi, la probabilité qu'il soit atteint en 50 ans vaut 10%. On retrouve donc une équivalence entre les 2 formulations.

2.3 Module exposition

Contrairement aux autres modules, celui-ci ne repose pas sur des considérations techniques. En effet, l'objectif de ce module est de collecter et formater les données utiles à la modélisation. L'exposition est basée sur deux caractéristiques :

- **La localisation :** C'est une information capitale afin de mener à bien une étude du risque. Cette donnée ne demande pas une précision extrême contrairement à d'autres périls (comme par exemple l'inondation où quelques mètres peuvent suffire à modifier grandement le risque). La localisation reste cependant importante, mais une précision de l'ordre du kilomètre peut être suffisante.
- **La valeur assurée :** Il s'agit du montant indiqué au contrat et représentant l'engagement de l'assureur. Celle-ci est découpée en 3 garanties :
  - La garantie Bâtiment, qui couvre les dégâts physiques liés au bâtiment,
  - La garantie Contenu, qui couvre les objets assurés,
  - La garantie pertes d'exploitation, qui couvre pertes ou un manque à gagner pour une entreprise à la suite d'un évènement catastrophique. Il s'agit donc de pertes indirectes, qui ne sont pas physiques.

On retrouve aussi des informations liées aux caractéristiques des bâtiments assurés. Ces caractéristiques vont directement influer sur la vulnérabilité du bâtiment, et sont donc listées dans la partie suivante traitant du module vulnérabilité.

2.4 Module vulnérabilité

Le module de vulnérabilité a pour but de relier les modules aléa et exposition afin d'associer une perte économique aux biens assurés. Pour une même intensité, l'ampleur des dommages peut être très diverse. En effet, les dégâts sur les bâtiments peuvent varier selon les caractéristiques du bien sinistré. Par exemple, une maison construite il y a une cinquantaine d'années sera plus vulnérable qu'une maison récente, pour un évènement de même intensité. Dans la suite, on caractérisera la perte économique par le taux de destruction (DR pour *destruction rate*), qui représente la perte en pourcentage de valeur assurée. Parler de taux de destruction plutôt que de perte en valeur permet d'avoir des résultats qui ne dépendent pas de l'ordre de grandeur de la somme assurée. Le taux de destruction se définit comme suit :

$$DR = \frac{\text{Charge}}{TIV}$$

où TIV est la *total insured value*, c'est à dire la valeur assurée totale du bien, et Charge la charge totale résultante du sinistre. Pour déterminer la vulnérabilité de chaque bâtiment, il est donc crucial de prendre en compte :

- le géocodage,
- Le nombre d'étages du bâtiment,
- L'année de construction : l'année de construction représente une approximation de l'année de conception de la structure, qui détermine en fait le code sismique utilisé pour le dimensionnement des éléments structurels, du renforcement et des détails et, par conséquent, l'action sismique minimale à laquelle la structure devrait pouvoir résister,

— Le type de structure du bâtiment : comme les 2 propriétés de bâtiments précédentes, le type de structure est un paramètre de risque. En effet, un bâtiment en maçonnerie sera plus vulnérable qu'un bâtiment en acier par exemple,

— le type d'occupation du bâtiment.

Ce module va donc relier les caractéristiques citées ci-dessus et l'exposition d'un bien avec un taux de destruction à l'aide de fonctions. Ces fonctions proviennent de l'analyse de sinistres dans les cas où il y a suffisamment de données (cas des inondations en France par exemple), et/ou d'une étude d'experts et de chercheurs spécialistes dans le péril étudié (c'est généralement le cas des séismes pour lesquels il y a trop peu de sinistres à étudier).
La partie vulnérabilité du modèle est basée sur des fonctions de fragilité, qui décrivent la probabilité de dépasser un niveau spécifique de dommage correspondant à différents états limites. Les fonctions de fragilité donnent une vision plus précise des dommages potentiels d'un bâtiment dus à un tremblement de terre que les fonctions de vulnérabilité qui ne sont qu'une synthèse des fonctions de fragilité et qui établissent un lien direct entre une mesure d'intensité et une perte. On retrouve sur la figure 2.12 un exemple de courbes de fragilité :

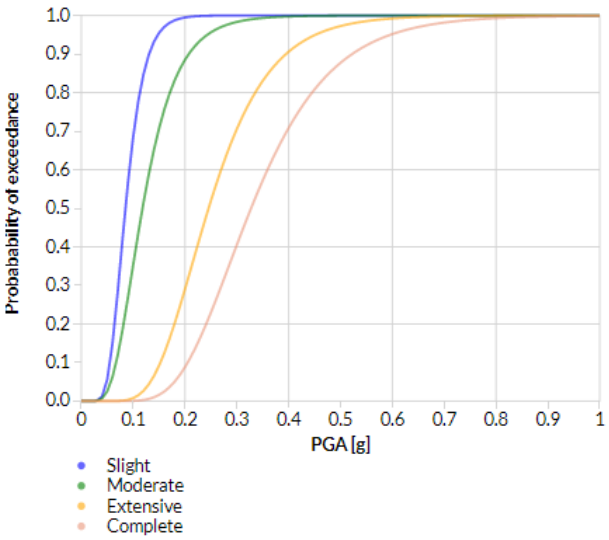


FIGURE 2.12 – Exemple de courbe de Fragilité

Chaque courbe de fragilité est associée à un seuil de dégât subi par le bâtiment (dans l'exemple léger, modéré, étendu, destruction totale), et représente donc la probabilité conditionnelle de dépasser chaque seuil de dégât ds_i sachant une valeur de PGA donnée. Mathématiquement, cela donne donc la formule suivante : $\mathbb{P}(D \geq ds_i | X = x)$ où D est la variable aléatoire représentant le dommage subi par un bâtiment, X la variable représentant le PGA et x une valeur particulière de X. Par exemple, la courbe orange décrit la fonction $\mathbb{P}(D \geq ds_2 | X = x)$ où ds_2 représente l'état de dommage correspondant à des dégâts étendus.

A partir de cette courbe de fragilité, il est possible d'en déduire une courbe de vulnérabilité. Une courbe de vulnérabilité associe un taux de destruction à un ensemble de valeur de PGA. Pour pouvoir construire la courbe, il est nécessaire d'associer une valeur de taux de destruction aux états de dommages ds_i. On peut pour cela utiliser par exemple la table

2.4. Il s'agit des valeurs par défaut utilisées par Hazus (*Hazard US* : c'est un outil d'analyse des risques naturels basé sur un système d'information géographique, développé et distribué gratuitement par la *Federal Emergency Management Agency* (FEMA)) :

Etat de dommage	Taux de destruction moyen
Léger	2%
Modéré	10%
Etendu	50%
Destruction totale	100%

TABLE 2.4 – Exemple de taux de destruction par état de dommage

Pour construire la courbe de vulnérabilité, on souhaite donc calculer la quantité suivante pour chaque valeur de PGA : $\mathbb{E}(L|X = x)$ où L représente le taux de destruction du bâtiment. Cette valeur peut se décomposer de la façon suivante :

$$\mathbb{E}(L|X = x) = \sum_{i=1}^{4} \mathbb{E}\left(L|D \in [ds_i, ds_{i+1}[\right) \times \mathbb{P}(D \in [ds_i, ds_{i+1}[\,|X = x)$$

Et l'on peut calculer $\mathbb{P}(D \in [ds_i, ds_{i+1}[\,|X = x) = \mathbb{P}(D \geq ds_i|X = x) - \mathbb{P}(D \geq ds_{i+1}|X = x)$, résultat obtenu grâce à la courbe de fragilité. Le premier terme de la somme précédente de la somme découle de la table 2.4.

La figure 2.13 est un exemple de courbe de vulnérabilité :

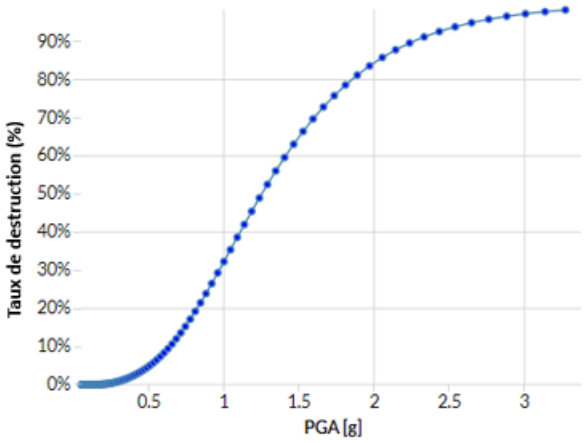


FIGURE 2.13 – Exemple de courbe de Vulnérabilité

Un ensemble complet de courbes de fragilité pour chaque combinaison de variables caractéristiques du bâtiment est pris en compte dans le modèle pour estimer la perte de chaque événement stochastique considéré. Ce calcul est réalisé avec l'outil Oasis (une plateforme open source de modélisation des catastrophes naturelles), qui gère ces entrées (empreinte du risque et fonctions de fragilité) avec les caractéristiques du portefeuille pour calculer la perte. Cet outil est calibré pour échantillonner le taux de destruction sur la base des probabilités obtenues pour une mesure d'intensité donnée. Pour chaque combinaison de variables caractéristiques du bâtiment, il existe des courbes de fragilité associées. Elles sont modélisées par des fonctions de répartition d'une loi log-normale.

Par conséquent, elles sont entièrement définies par 2 paramètres : la moyenne et l'écart type. On regroupe dans la table 2.14 un exemple de paramètres de ces courbes :

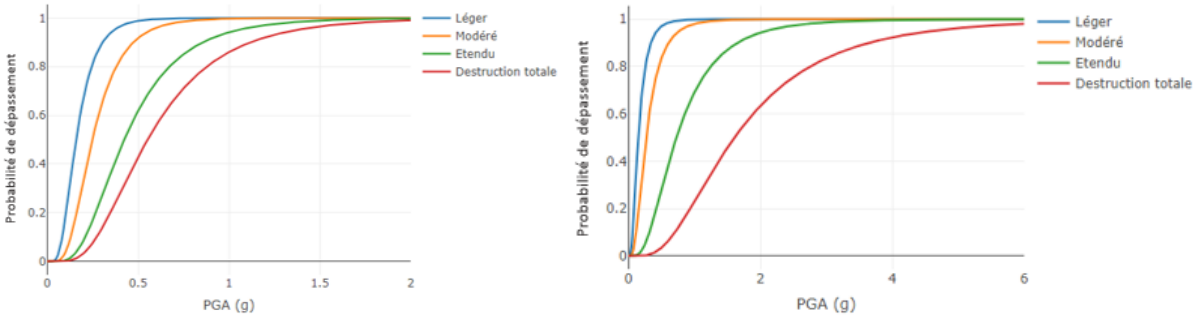
Reinforced Concrete										
class			DS1 (slight)		DS2 (moderate)		DS3 (extensive)		DS4 (complete)	
Code Level	Rise	Class	μ	σ	μ	σ	μ	σ	μ	σ
No Code	Low-Rise	NC-LR	-2,033	0,416	-1,387	0,426	-1,185	0,480	-0,932	0,524
	Mid-rise	NC-MR	-1,738	0,482	-1,204	0,499	-1,108	0,498	-1,021	0,497
	High-rise	NC-HR	-1,878	0,411	-1,183	0,441	-0,967	0,495	-0,711	0,533
Low Code	Low-Rise	LC-LR	-1,831	0,505	-1,390	0,504	-0,856	0,548	-0,592	0,550
	Mid-rise	LC-MR	-1,574	0,387	-1,190	0,394	-0,930	0,403	-0,694	0,427
	High-rise	LC-HR	-1,543	0,462	-0,957	0,449	-0,711	0,474	-0,431	0,520
Moderate Code	All	MC-LR	-1,523	0,458	-1,014	0,377	-0,693	0,398	-0,187	0,518
High code	All	HC-All	-1,897	0,640	-1,309	0,640	-0,315	0,640	0,476	0,640

FIGURE 2.14 – Paramètres définissant les courbes de fragilité

Dans cet exemple, on croise le type de structure « béton armé » avec toutes les modalités de l'année de construction (*Code Level*) et le nombre d'étages (*Rise*). Dans le tableau, on a des paramètres μ et σ pour chaque état (léger, modéré, étendu, destruction totale). Comme ces paramètres correspondent aux paramètres d'une fonction de répartition d'une loi log-normale, on peut tracer des exemples de courbes de fragilité sur la figure 2.15 :

— Celles liées au croisement béton armé/Low Code/Low Rise (première figure). Les valeurs de μ et σ correspondantes sont surlignées en jaune,

— Celles liées au croisement béton armé/High Code/High Rise (deuxième figure). Les valeurs de μ et σ correspondantes sont surlignées en bleu.



(a) béton armé/Low Code/Low Rise (b) béton armé/High Code/High Rise

FIGURE 2.15 – Courbes de fragilité

On a donc une multitude de courbes de fragilité (pour chaque croisement de variables), desquelles on déduit les courbes de vulnérabilité. Ces courbes de vulnérabilité servent à déterminer un taux de destruction pour la garantie bâtiment, c'est à dire les dégâts physiques subis par celui-ci. Afin de modéliser les pertes liées aux autres garanties (donc le contenu et l'interruption d'activité), il a été choisi de définir des taux de destruction basées sur le taux de destruction de la garantie bâtiment. La table 2.5 est un exemple de relativité que l'on peut trouver entre les garanties :

Etat du dommage	Taux de destruction Bâtiment	Taux de destruction Contenu	Taux de destruction BI
Destruction totale	100%	100%	100%
Etendu	50%	25%	25%
Modéré	10%	5%	5%
Léger	2%	1%	1%
Aucun dommage	0%	0%	0%

TABLE 2.5 – Relativité entre les garanties

Cette approche est justifiée car le contenu est généralement moins endommagé que le bâtiment en cas de tremblement de terre si le bâtiment n'est que partiellement endommagé.

2.5 Module financier

Ce module permet d'appliquer toutes les conditions d'assurance aux pertes brutes résultant d'un évènement catastrophique. Dans ce module, il n'y a pas de considération physique du péril étudié, il s'agit uniquement de déterminer la perte financière nette de l'assureur. Ces conditions financières s'appliquent d'abord au niveau de chaque site, puis dans certains cas on applique aussi des conditions au niveau des polices qui peuvent comporter plusieurs sites.
Les différentes caractéristiques à prendre en compte sont :

- La coassurance : c'est une opération par laquelle plusieurs sociétés d'assurances garantissent au moyen d'un seul contrat un même risque ou un même ensemble de risques. Ainsi, chaque assureur devra payer le coût du ou des sinistres en fonction du pourcentage correspondant à son niveau d'engagement dans la couverture du risque,
- La franchise et la limite : l'assureur paye la part du montant des sinistres compris entre la franchise et la limite, le reste étant à la charge de l'assuré,
- La réassurance : Lors de la survenance d'une catastrophe naturelle entrainant un très gros sinistre, l'assureur ne peut en général pas régler la totalité de celui-ci. La réassurance permet de pallier ce problème.

On présente sur la figure 2.16 un exemple très simple d'application de franchises et limites pour une police multi-sites (2 sites dans notre cas).

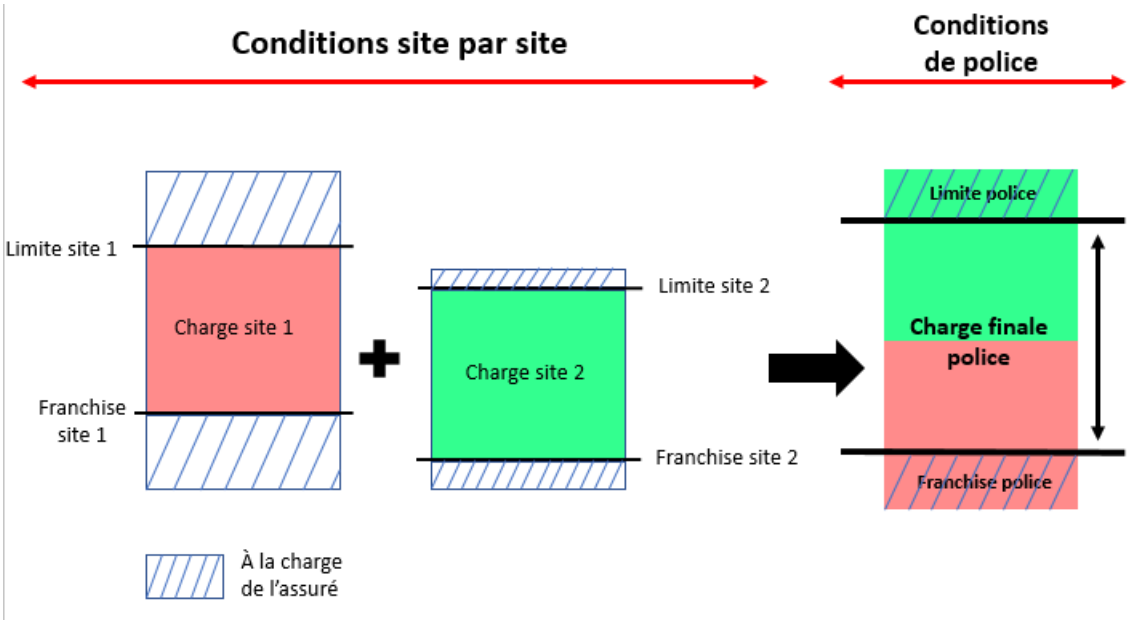


FIGURE 2.16 – Conditions d'assurance pour une police multi-sites

2.6 Sorties du modèle

Le modèle CAT donne en sortie une distribution de pertes. Pour la représenter, on distingue deux types de courbes, que l'on appelle les courbes EP (pour *exceedance probability*). Dans les formules suivantes on note N le nombre de sites assurés et X_i la perte associée au site i, pour i compris entre 1 et N.

- **La courbe AEP (aggregate exceedance probability) :** Il s'agit de la probabilité annuelle de dépasser un certain montant de pertes. Elle se définit par la formule suivante :

$$AEP(x) = \mathbb{P}\left(\sum_{i=1}^{N} X_i \geq x\right)$$

- **La courbe OEP (occurrence exceedance probability) :** Il s'agit de la probabilité annuelle que la perte maximale dépasse un certain seuil. L'OEP prend donc uniquement en compte l'évènement engendrant la plus grande perte.

$$OEP(x) = \mathbb{P}\left(\max(X_1, X_2, ..., X_n) \geq x\right)$$

Généralement, on représente ces grandeurs par des courbes représentant en abscisse une période de retour et en ordonnée une perte associée à cette période de retour. La période de retour est couramment utilisée lorsqu'il s'agit de sinistres rares. Elle est définie par l'inverse de la probabilité de survenance de la grandeur considérée.
On peut faire quelques remarques sur ces courbes.

— Par définition, on a toujours $AEP \geq OEP$ avec égalité uniquement lorsqu'il n'y a qu'un évènement dans l'année,

— Une valeur particulière est intéressante pour le risk management : il s'agit de la valeur de l'AEP pour une période de retour de 200 ans, c'est à dire une probabilité d'occurrence de 0.5%. En effet l'entrée en vigueur de la directive européenne

Solvabilité 2 contraint les entreprises à respecter une exigence quantitative de fonds propres détenus pour se protéger contre la perte la plus importante susceptible de se produire avec une probabilité d'une chance sur 200 dans l'année. Ainsi la valeur de l'AEP pour un temps de retour de 200 ans est un indicateur très important,

— L'OEP aide à optimiser les traités de réassurance, en quantifiant la distribution du coût maximal annuel d'un événement. En effet, l'OEP ne renseigne que sur les sinistres de sévérité haute, ce qui va aider à construire les contrats de réassurance.

Nous allons maintenant expliquer comment obtenir ces courbes sont calculées.

Le modèle CAT renvoie un *Event Loss Table* (souvent appelé ELT). Ce tableau regroupe chacun des scénarios stochastiques générés dans le modulé aléa, On retrouve dans ce tableau des informations pour chacun des événements comme la fréquence de survenance, les paramètres liés à la distribution de pertes ou encore l'intervalle de magnitude. Un exemple de ce type de tableau est présenté dans la table 2.6, avec des chiffres inventés :

Id Séisme	Fréquence	Coût	Ecart type	Valeur assurée	Magnitude
1	0.01	1 000 000	100 000	10 500 000	[7 ;8[
2	0.01	3 000 000	1 000 000	15 000 000	[6 ;7[
3	0.02	2 000 000	800 000	20 000 000	[5 ;6[
4	0.02	10 000 000	5 000 00	100 000 000	[5 ;6[
5	0.03	6 000 000	3 000 000	50 000 000	[3 ;4[
...					

TABLE 2.6 – Exemple d'Event loss table

A partir de l'ELT, le modèle effectue des simulations pour créer un autre tableau, le *Year Event Loss Table* (souvent appelé YELT). Les étapes pour créer l'YELT à partir de l'ELT sont les suivantes :

1. On décide sur combien d'années on fait la simulation : dans notre cas, on le fait sur 1 million d'années. On choisit un grand nombre d'années afin de tirer des évènements majeurs qui se déroulent avec une fréquence très basse ;
2. On détermine un nombre moyen d'événements qui se produisent dans une année pour chaque intervalle de magnitude selon les fréquences de l'ELT ;
3. Pour chaque année de l'YELT :
   — on simule une loi de Poisson de paramètre fixé à l'étape précédente pour déterminer le nombre d'événements se produisant au cours de l'année pour chaque intervalle de magnitude ;
   — On tire aléatoirement dans l'ELT le nombre d'évènements déterminé à l'étape précédente pour chaque intervalle de magnitude ;
   — On associe une perte selon les évènements tirés, l'information venant également de l'ELT.

Cela nous permet d'obtenir une *Year Event Loss Table*, souvent appelé YELT et représenté dans la table 2.7.

Année	Séisme Id	Perte	Location
1	2	3 500 000	a
2	1	1 000 000	b
2	3	1 200 000	c
3	4	13 000 000	d
...			
1 000 000	5	5 000 000	e

TABLE 2.7 – Exemple de Year Event loss table

Une réduction de dimension est ensuite appliquée à ces 1 million d'années, tout en essayant de garder les mêmes valeurs d'OEP et AEP que celles calculées pour le tableau complet. Cette réduction est nécessaire pour éviter de traiter de trop lourdes bases de données. Après cette réduction, on obtient alors un YELT de 10 000 lignes.

A partir de ce dernier tableau, il est alors possible de calculer les AEP et OEP du modèle. Il suffit, année par année de calculer le maximum des pertes (pour l'OEP) et la somme des pertes (pour l'AEP). Les courbes se construisent ensuite en calculant les quantiles associés à ces valeurs. Cela nous donne, pour le modèle suisse, la figure 2.17 :

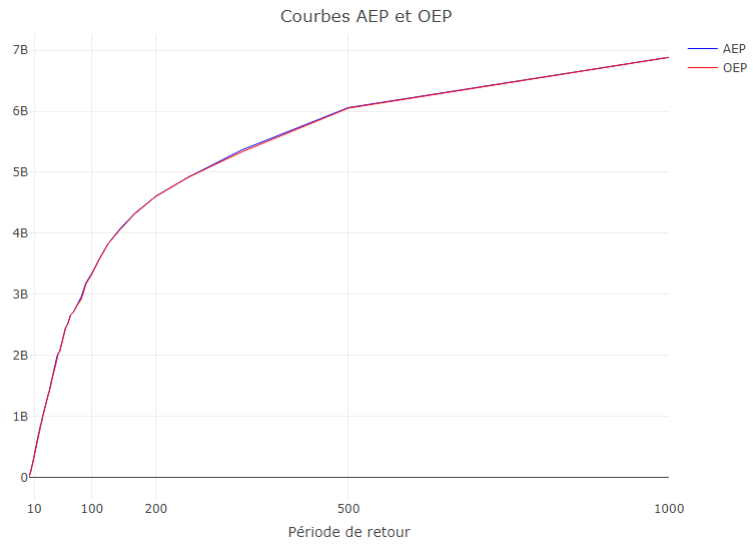


FIGURE 2.17 – Courbes EP du modèle suisse

Dans le cas particulier de la Suisse, les courbes AEP et OEP sont quasiment confondues. Cela s'explique par le fait que c'est un pays qui est assez peu risqué. On retrouve donc en sortie de modèle seulement 1 évènement par année en général (parfois même aucun). Dans les pays où la sinistralité est plus importante, on retrouve une différence significative entre ces deux courbes car plusieurs sinistres sont simulés par an. En général, pour un assureur, l'AEP et l'OEP sont très proches pour les événements de type tremblement de terre, tandis qu'ils peuvent être très différents pour les événements de type tempête, du fait de la présence d'un grand nombre d'événements de faible intensité.

Pour introduire le chapitre 3, nous allons également évoquer une autre sortie du modèle CAT. Il s'agit d'une tableau qui résume l'YELT et qui donne des pertes financières site

par site. Pour créer ce tableau, on associe à chaque site de notre portefeuille la **sinistralité moyenne** par année. On a donc, grâce au modèle CAT, la possibilité d'obtenir des pertes sur un portefeuille de sites assurés mis en entrée de ce modèle. A partir de ces pertes et dans un but de tarification du risque séisme dans le cadre des assurances, nous allons calibrer des modéles statistiques et de machine learning afin de modéliser les pertes à partir de différentes variables explicatives liées aux spécificités des bâtiments assurés ainsi que les variables géographiques utilisées par le modèle CAT.

Chapitre 3

Tarification

L'objectif de cette partie est de modéliser le plus précisément les pertes simulées site à site d'un modèle CAT pour chaque couverture en calibrant des modèles statistiques usuels type GLM ou des modèles innovants de machine learning. En fin de chapitre, nous verrons une méthode pour intégrer les conditions financières à la modélisation.

3.1 Intérêt des modèles statistiques sur les pertes simulées par le modèle CAT

La base de données sur laquelle nous nous appuyons pour calibrer des modèles statistiques est la base de pertes site à site issues du modèle CAT. Ainsi, nous effectuons des modèles sur des données déjà issues d'un modèle. Plusieurs raisons expliquent cela :

- Les entités ont besoin d'une grille tarifaire précise afin de pouvoir diffuser immédiatement un tarif aux potentiels assurés. Un grille tarifaire est facilement implémentable dans n'importe quel système IT. Devoir faire tourner un modèle CAT à chaque fois qu'un assuré souhaite effectuer un devis est très chronophage (cela peut prendre 1 ou plusieurs jours selon la puissance des machines utilisées). Par conséquent, il est important de pouvoir s'appuyer sur une grille tarifaire. Pour des petits risques (par exemple une faible valeur assurée et/ou une police mono-site), il semble pertinent d'utiliser cette grille prédéfinie grâce aux modèles statistiques type GLM. Cependant au cas par cas, pour de gros risques (correspondant à une valeur assurée élevée et/ou une police avec beaucoup de sites assurés), il pourrait être intéressant d'utiliser directement les pertes issues du modèle CAT en effectuant un run pour ces cas précis,
- On a la possibilité, lorsque l'on calibre des modèles statistiques, de choisir les variables explicatives à prendre en compte. On peut donc répondre à des contraintes opérationnelles des entités, qui, par exemple, pourraient vouloir limiter le nombre de variables explicatives discriminant la prime.

3.2 Notion de portefeuille fictif (ou représentatif)

Les méthodes que nous allons développer dans les parties suivantes seront accompagnées d'exemples liés à la Suisse, une étude sur ce pays ayant été réalisée pendant mon alternance au sein de l'équipe.
On a, dans le cas des catastrophes naturelles, la possibilité d'ajouter des assurés fictifs à notre étude car les pertes économiques résultant de tremblements de terre sont générées par un modèle CAT (qui a été développé dans le chapitre 2) ; il ne s'agit donc pas des véritables sinistres des assurés. En effet, les catastrophes de type séisme sont très peu fréquentes (mais peuvent être dévastateurs), il y a donc trop peu de sinistres en base de données pour n'utiliser que ceux-ci dans la modélisation. On présente ci-dessous les raisons qui nous poussent à utiliser un portefeuille fictif (aussi appelé portefeuille représentatif) ainsi que la démarche pour le construire.

Faisons tout d'abord la liste des caractéristiques utiles à la modélisation du risque sismique :
On retrouve déjà la localisation des sites assurés. Contrairement aux hypothèses que l'on peut avoir habituellement en assurance « plus classique », dans le cas des catastrophes naturelles, les sites assurés d'un portefeuille ne sont pas indépendants par rapport au risque. Par exemple, le canton du Valais a le plus grand aléa sismique de la Suisse, c'est à dire qu'il s'agit du canton qui a le plus de chance d'être touché par un séisme.
Des informations plus détaillées concernant les caractéristiques du bâtiment assuré sont également nécessaires. En effet, il faut tenir compte de certaines propriétés pour étudier l'impact d'un évènement :

- Le type de structure,
- Le nombre d'étages,
- L'année de construction.

On retrouve également le type d'occupation, qui est une variable liée au secteur d'activité. Nous étudierons plus en détail ces variables dans la partie 3.3 de ce mémoire, traitant de l'exposition au risque. Afin d'avoir une vision exhaustive du risque auquel fait face un pays, il faut être capable de modéliser celui-ci selon toutes les possibilités de variables explicatives. Si l'on prend par exemple le portefeuille existant de la Suisse (c'est à dire le portefeuille d'assurés réels), certaines régions sont très peu assurées par rapport à d'autres. Ceci est tout à fait normal car les assurés ont tendance à se couvrir plus souvent lorsqu'ils sont dans les zones les plus risquées. Mais pour être capable de construire un tarif technique pour tous types d'assurés, y compris dans les régions moins risquées, il faut avoir une base de sites couvrant tout le territoire et ayant toutes les caractéristiques possibles. C'est pour cette raison qu'il est nécessaire de construire un portefeuille fictif qui va couvrir l'entièreté du pays, possédant toutes les combinaisons de variables explicatives possibles. Ce portefeuille sera ensuite concaténé au portefeuille existant pour former une unique base de données exhaustive du point de vue des variables de risque.
Pour créer le portefeuille fictif, on procède en deux temps. Tout d'abord, on quadrille tout le pays pour le couvrir entièrement, puis pour chaque point créé, il s'agira de construire des bâtiments fictifs avec tous les croisements de variables possibles.

3.2.1 Quadrillage du pays

Tout d'abord, on génère des points couvrant la totalité du pays afin de quadriller celui-ci. Pour la Suisse, cela donne un total de 768 points répartis sur tout le territoire. Tous les points sont espacés de la même distance. Le logiciel QGIS, permettant de manipuler des données géographiques, nous a permis d'effectuer cette génération de points. Ces points sont représentés sur la figure 3.1

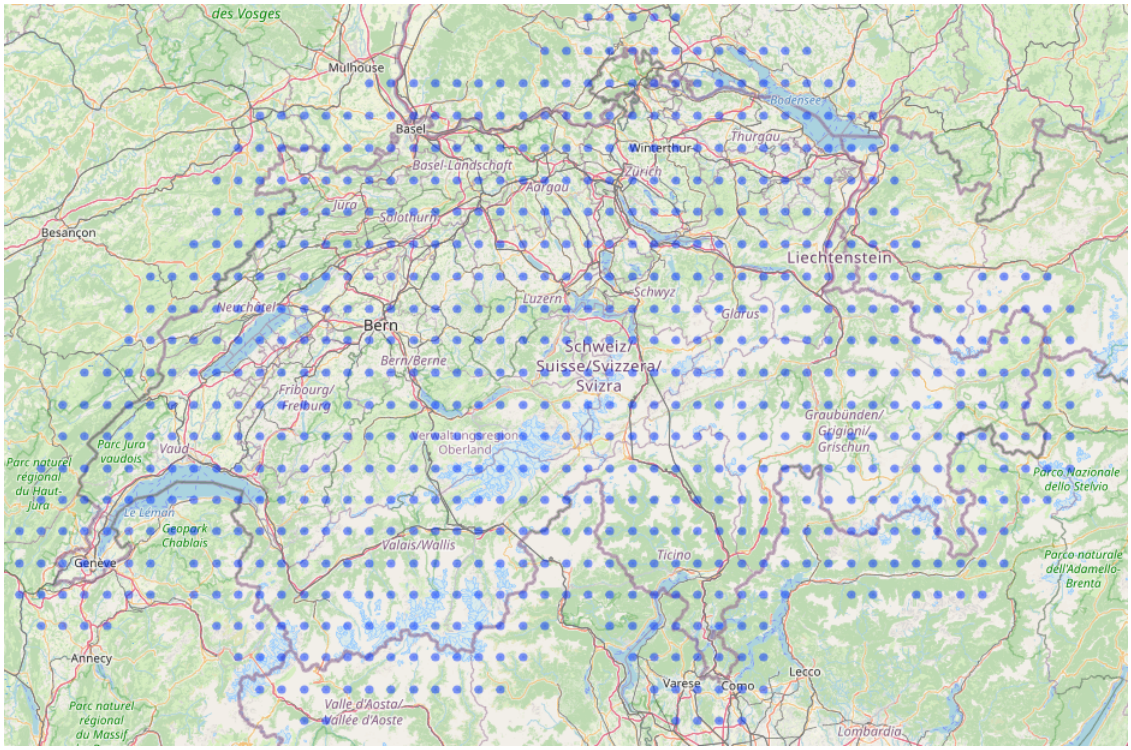


FIGURE 3.1 – Génération d'une grille de points en Suisse

Pour chacun de ces points, qui sont considérés comme des sites assurés, on va « construire » des batiments qui posséderont tous les croisements de variables explicatives possibles.

3.2.2 Génération de sites assurés fictifs

Après avoir quadrillé le pays de sites fictifs, nous allons parcourir chaque site, et y associer des variables explicatives, qui sont ici les caractéristiques propres au bâtiment. Afin d'être exhaustif, on parcourt chaque possibilité pour chaque variable afin de construire tous les croisements possibles. L'arborescence (figure 3.2) représente l'ensemble des combinaisons possibles avec, dans l'ordre, le nombre d'étages, l'année de construction puis le type de structure du bâtiment. Seules les branches de gauche ont été précisées par souci de clarté, mais il est très simple d'imaginer à quoi ressemblerait l'arbre complet.

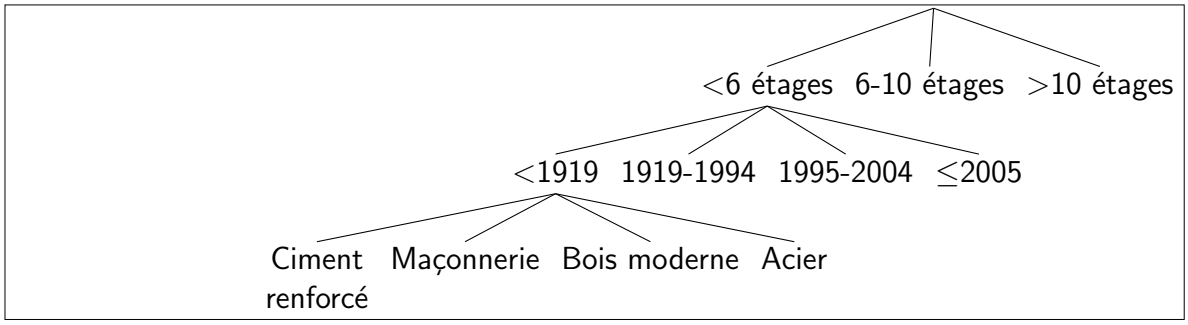


FIGURE 3.2 – Arborescence des variables liées au bâtiment

Ainsi, pour chaque point généré sur la carte, on y associe des bâtiments avec toutes les combinaisons de variables explicatives possibles, c'est à dire avec toutes les branches de la table ci-dessus. Finalement, on se retrouve avec une base de données couvrant la totalité du territoire, et exhaustive en ce qui concerne les variables explicatives. La valeur assurée qu'on associe à chaque bâtiment est peu importante car on travaillera pour la modélisation avec des taux de destruction, qui est indépendante de l'ordre de grandeur de la valeur assurée. On donne donc une unique valeur assurée pour tous les bâtiments. Finalement, on associe les 3 garanties possibles à chaque bâtiment créé pour avoir une base d'assurés fictifs exhaustive. Une fois que l'on dispose d'une base complète, l'obtention de pertes se fait par le biais d'un modèle CAT.

Il faut être très vigilant avec cette notion de portefeuille représentatif. En effet, certains bâtiments fictifs créés sont irréalistes et n'existeront jamais dans le portefeuille réel. Il faudra prendre en compte cette limite lors de la modélisation.
Dans la partie suivante, nous allons présenter en détail de quoi est constituée la base de données qui sera utilisée pour la calibration de nos modèles.

3.3 Base de données pour la modélisation

Une fois la base d'assurés constituée (qui on le rappelle est une concaténation entre assurés réels et assurés fictifs), on peut la mettre en entrée du modèle CAT et le faire tourner. On obtient alors, entre autres, une base de pertes site à site. Nous allons commencer par décrire cette base.

Les variables que l'on a à disposition ainsi que leurs modalités ont été précisées dans la table 3.1. On retrouve donc :

- Des informations sur l'exposition du site assuré,
- Des informations liées aux caractéristiques des bâtiments assurés, qui seront les principales variables explicatives du modèle de tarification,
- Des informations sur la variable réponse qui est le taux de destruction du bâtiment.

Nom de la variable	Description	Modalités
TIV	Valeur assurée totale du bâtiment	Variable quantitative
Année de construction	Année de construction du bâtiment	<1919 1919-1994 1995-2004 >=2005
Type de structure	Type de structure du bâtiment	Maçonnerie Béton armé Béton armé préfabriqué Acier
Nombre d'étages	Nombre d'étages du bâtiment	<6 étages 6-10 étages >10 étages
PGA	*Peak Ground Acceleration*	Variable quantitative
Cresta	Canton suisse	Valais, Zurich, Bâle, Genève...
vs30	Représente la vitesse moyenne des des ondes de cisaillement à 30 mètres de profondeur (m/s)	<750 >=750
Taux de destruction	Taux de destruction du bâtiment, défini par le rapport entre la charge du sinistre et la TIV.	Variable quantitative exprimée en %

TABLE 3.1 – Descriptif des variables en base de données

Cresta est l'acronyme du terme anglais *Catastrophe Risk Evaluation and Standardising Target Accumulations*. Ce terme a été fondé par des compagnies de réassurance dans le but d'établir un système uniforme au niveau mondial de contrôle des risques d'accumulation des risques naturels, en particulier les tremblements de terre, les tempêtes et les inondations.

Afin d'avoir une idée générale de l'influence des variables explicatives sur le taux de destruction, les quelques graphes univariés ci-dessous peuvent être intéressants.
Commençons par représenter par des boxplots le taux de destruction des sites sinistrés selon les caractéristiques du bâtiment, c'est à dire le type de structure, le nombre d'étages et l'année de construction :

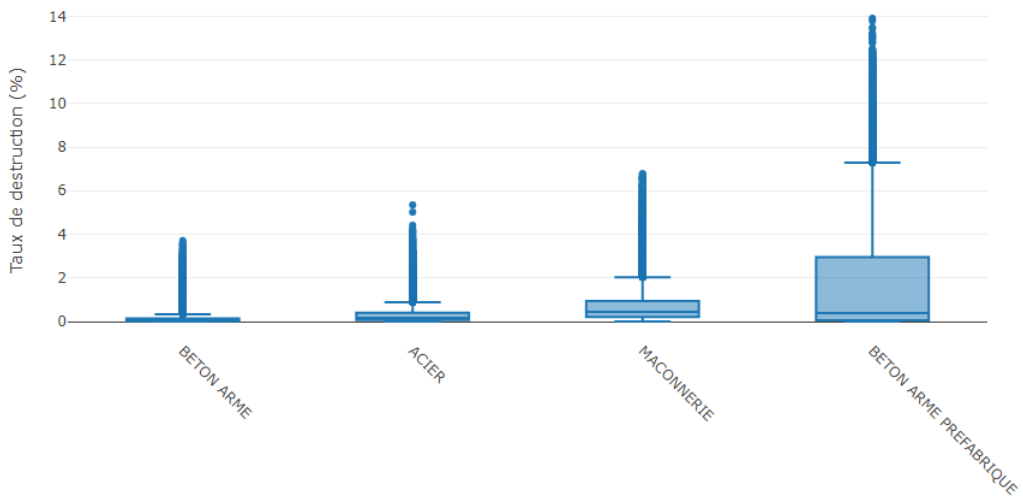
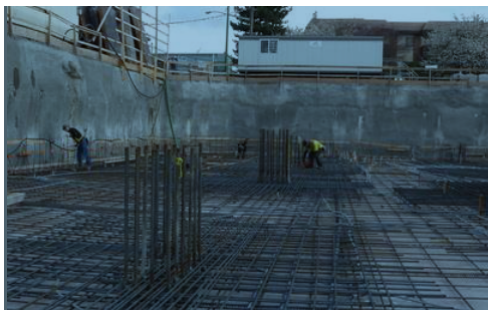


FIGURE 3.3 – Taux de destruction en fonction du type de structure

Voici 4 images de bâtiments en construction représentant les 4 types de structure présents dans notre base.



(a) Béton armé

(b) Acier

(c) Maçonnerie

(d) Béton armé préfabriqué

FIGURE 3.4 – Types de structure en image

Le boxplot suivant (figure 3.5) représente le taux de destruction selon l'année de construction :

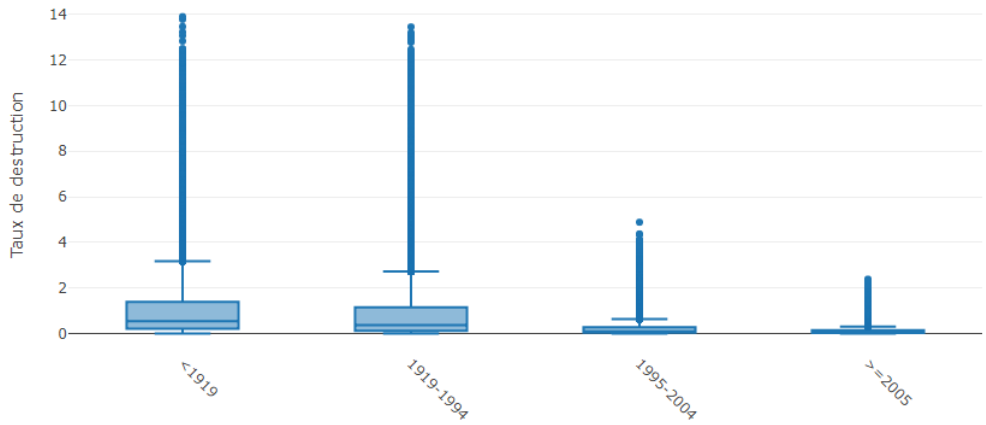


FIGURE 3.5 – Taux de destruction en fonction de l'année de construction

Enfin, la dernière variable à représenter est le nombre d'étages :

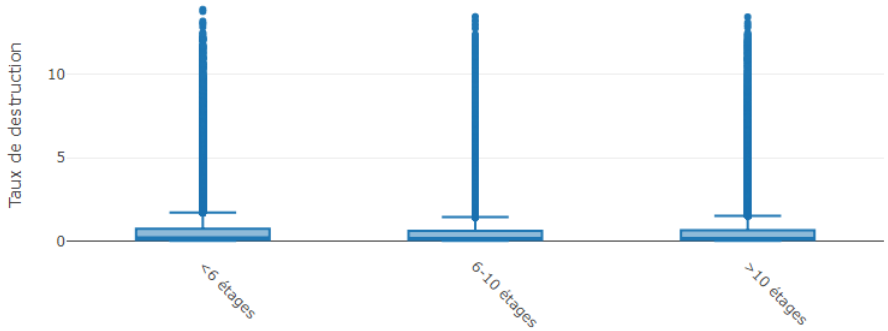


FIGURE 3.6 – Taux de destruction en fonction du nombre d'étages

A travers ces quelques graphes, on peut faire plusieurs remarques :

— ils ne représentent qu'une seule variable, sans prise en compte de l'influence des autres sur le taux de destruction. Ils permettent donc d'avoir une vision très générale. Une tendance se dégage tout de même pour certaines variables,

— les bâtiments en béton armé semblent les moins touchés en terme de sévérité, contrairement au béton armé préfabriqué. De plus les dégats, plus les bâtiments sont vieux et plus ils sont touchés sévèrement par un séisme. Cela est logique car, les constructions antisismiques ont beaucoup évolué avec le temps.

— la variable « Nombre d'étages » semble peu influente sur le taux de destruction. Cette tendance se vérifie lors de la modélisation effectuée un peu plus tard dans ce mémoire, où nous avons décidé de ne pas prendre en compte cette variable comme variable explicative.

On représente maintenant le taux de destruction selon les variables géographiques à notre disposition, qui sont des variables dépendantes de l'endroit exact où se trouve le

bien assuré. On retrouve donc le Cresta (pour *Catastrophe Risk Evaluation and Standardising Target Accumulations*), le vs30 (qui représente la vitesse moyenne des ondes de cisaillement à 30 mètres de profondeur) ainsi que le facteur de risque lié au PGA :

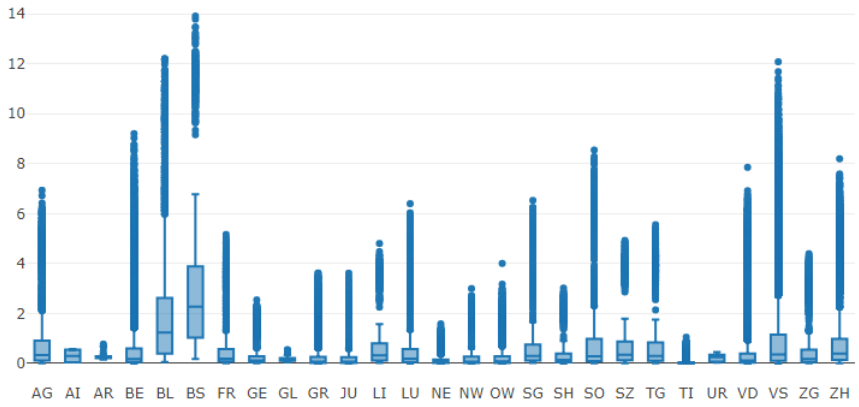


FIGURE 3.7 – Taux de destruction en fonction du Cresta

La carte des cantons de la Suisse, avec leur abréviation, est également donnée :

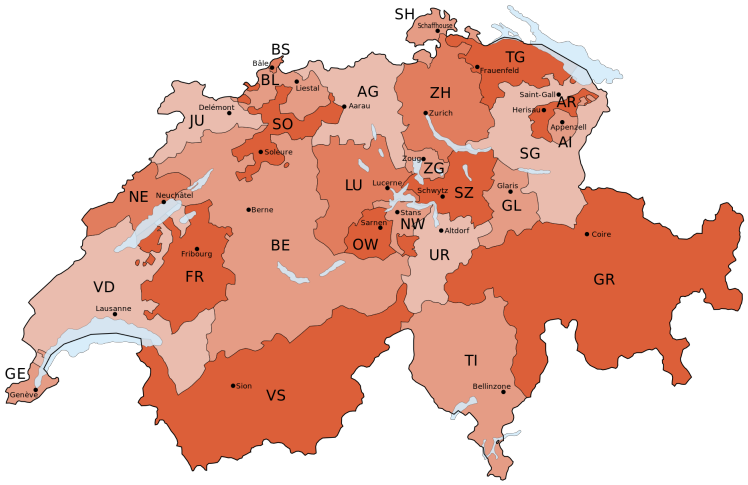


FIGURE 3.8 – Cantons de Suisse

On retrouve maintenant ci-dessous le taux de destruction selon le vs30 :

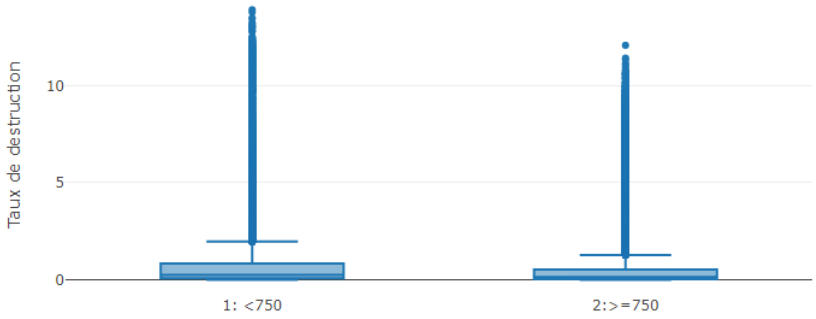


FIGURE 3.9 – Taux de destruction en fonction du vs30

Une variable géographique n'a pas encore été présentée, il s'agit du PGA. En sortie de modèle CAT, il s'agit d'une variable quantitative. Pour la modélisation (surtout pour les modèles GLMs), il est nécessaire de découper cette variable en classes. C'est ce que nous faisons dans la partie 3.3.1. A noter que les autres variables sont déjà découpées en classes par le modèle CAT.

3.3.1 Découpage du PGA en classes

Le *Peak Ground Acceleration*, une grandeur que l'on obtient pour chaque site assuré en sortie de modèle est une variable quantitative comprise dans notre cas précis entre $0.0766g$ et $0.2667g$. On rappelle que c'est une grandeur qui décrit l'accélération maximale du sol et qui caractérise le mouvement de sols soumis à des ondes sismiques ; elle est liée à la vitesse du sol se déplaçant lors d'un séisme. Ce paramètre dépend de l'intensité de la secousse, mais aussi de la nature géologique du sous-sol.

Un histogramme de cette grandeur est représenté en figure 3.10 :

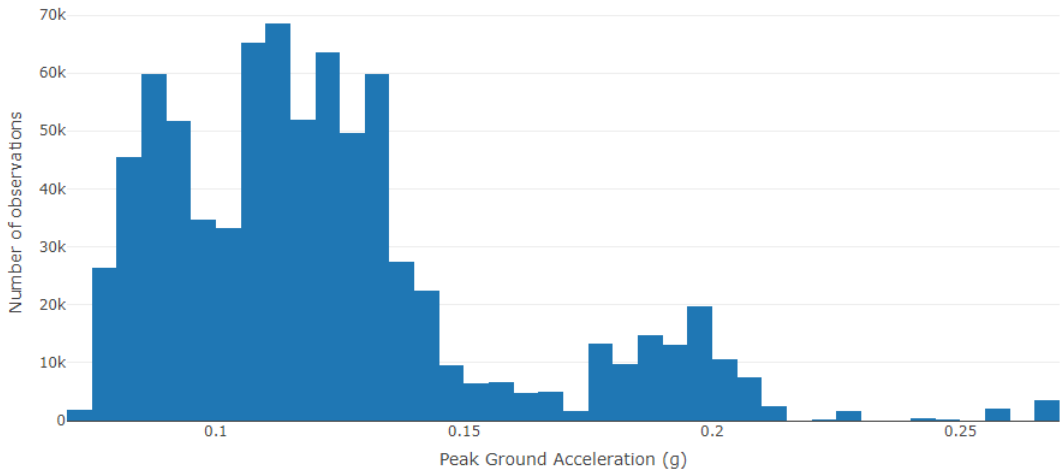


FIGURE 3.10 – Histogramme du Peak Ground Acceleration

L'objectif de cette partie est de découper le PGA en plusieurs classes de risque. En d'autres termes, on souhaite discrétiser cette variable. Nous allons ainsi effectuer une discrétisation supervisée qui consiste à découper la variable en tranches en fonction de la variable à expliquer, qui est ici le taux de destruction.
Naturellement, les premières méthodes de discrétisation qui viennent en tête sont :

— un découpage par des intervalles de taille égale,
— un découpage par quantile.

Ces méthodes permettent d'obtenir des classes homogènes en termes de population et sont très simples à mettre en place. Mais elles ne permettent pas d'avoir des classes homogènes en termes de sinistralité. Pour obtenir des classes de risque homogènes en termes de sinistralité, nous utiliserons des arbres de régression.

3.3.1.1 Arbres de régression

La discrétisation par des arbres de régression consiste à utiliser un arbre de régression pour identifier les points optimaux qui déterminent les bornes des intervalles des classes de risque. L'objectif est de construire des sous-groupes les plus homogènes du point de vue de la variable à prédire « taux de destruction ».

Les arbres de régression sont très efficaces pour des tailles d'échantillon importantes et faciles à implémenter comme ils ne requièrent pas d'hypothèses sur la distribution des variables. Le principe de fonctionnement est le suivant : pour expliquer une variable, le système recherche le critère le plus déterminant et découpe la population en sous populations possédant ce critère. Ces arbres ne sont pas uniquement utilisés pour construire des classes. En effet, ils sont en général utilisés pour de la modélisation classique, c'est à dire pour expliquer une variable réponse selon plusieurs variables explicatives. Dans ce cas, il faut mettre en entrée la totalité des variables explicatives ainsi que la variable réponse. Si l'on souhaite uniquement utiliser les arbres pour la construction de classes, il suffit de ne prendre en compte qu'une seule variable explicative, qui est la variable que l'on souhaite découper (ici le PGA), ainsi que la variable réponse qui est la variable cible (ici le taux de destruction) qui nous permettra de découper en classes le PGA.

Les principaux algorithmes d'arbres de décision sont :

— CART (*Classification And Regression Tree*), qui est adapté à tout type de variable,
— CHAID (*Chi-Square Automation Interaction Detection*), initialement réservé à l'étude des variables discrètes et qualitatives. Cet algorithme ne nous intéresse donc pas car on étudie uniquement des variables quantitatives.

Nous allons donc résumer le fonctionnement de l'algorithme CART, qui est celui utilisé par la fonction *rpart* de R du package portant le même nom. Pour plus de détails à propos de cet algorithme, de nombreux papiers existent et sont déjà trés complets (voir [14]).

Le principe général de CART est de partitionner récursivement l'espace d'entrée X de façon binaire, puis de déterminer une sous-partition optimale pour la prédiction. Bâtir un arbre CART se fait en deux étapes. Une première phase est la construction d'un arbre maximal, qui permet de définir la famille de modèles à l'intérieur de laquelle on cherchera à sélectionner le meilleur, et une seconde phase, dite d'élagage, qui construit une suite de sous-arbres optimaux élagués de l'arbre maximal (l'élagage d'un arbre signigie dépouiller un arbre de ses branches inutiles). Détaillons chacune de ces étapes ci-dessous. On se place dans un cas général d'un arbre de régression avec p variables explicatives.

1. **Construction de l'arbre maximal T_{max}.** A chaque pas du partitionnement, on découpe une partie de l'espace en deux sous-parties. On associe alors naturellement un arbre binaire à la partition construite. Les nœuds de l'arbre sont associés

aux éléments de la partition. Par exemple, la racine de l'arbre est associée à l'espace d'entrée tout entier. Ses deux nœuds fils sont associés aux deux sous-parties obtenues par la première découpe du partitionnement, et ainsi de suite. La figure 3.11 illustre la correspondance entre un arbre binaire et la partition associée, avec un exemple en 2 dimensions, c'est à dire 2 variables explicatives X^1 et X^2.

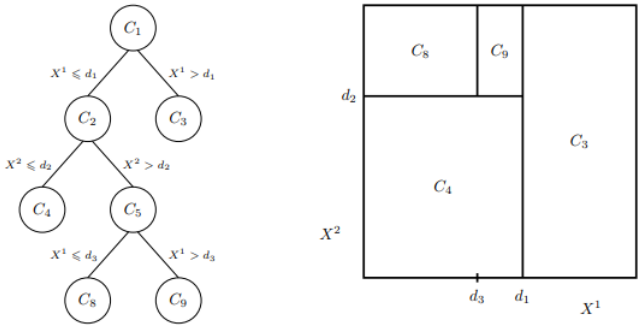


FIGURE 3.11 – Construction d'un arbre de régression et partition

Détaillons maintenant la règle de découpe. L'espace d'entrée est $\mathbb{R}^p$, où p est le nombre de variables. Partons de la racine de l'arbre (associée à $\mathbb{R}^p$ tout entier), qui contient toutes les observations de l'échantillon d'apprentissage. La première étape de CART consiste à découper au mieux cette racine en deux nœuds fils. Nous appelons coupure (ou découpe ou même split) un élément de la forme $\{X^j \leq d\} \cup \{X^j > d\}$. Découper suivant $\{X^j \leq d\} \cup \{X^j > d\}$ signifie que toutes les observations avec une valeur de la j-ième variable plus petite que d vont dans le nœud fils de gauche, et toutes celles avec une valeur plus grande que d vont dans le nœud fils de droite. La méthode sélectionne alors la meilleure découpe, c'est-à-dire le couple (j,d) qui minimise une certaine fonction de coût. Dans le cas d'arbres de régression, on cherche à minimiser la variance intra-groupes résultant de la découpe d'un nœud t en 2 nœuds fils t_L et t_R (L pour left et R pour right). La variance d'un nœud t étant définie par $V(t) = \frac{1}{\#t} \sum_{i,x_i \in t} (y_i - \bar{y}_t)^2$ où $\bar{y}_t$ est la moyenne des observations présentes dans le nœud t. L'idée est donc de minimiser :

$$\frac{1}{n} \sum_{(x_i,y_i) \in t_L} (y_i - \bar{y}_{t_L})^2 + \frac{1}{n} \sum_{(x_i,y_i) \in t_R} (y_i - \bar{y}_{t_R})^2 = \frac{\#t_L}{n} V(t_L) + \frac{\#t_R}{n} V(t_R)$$

Une fois la racine de l'arbre découpée, on se restreint à chacun des nœuds fils et on recherche alors, suivant le même procédé, la meilleure façon de les découper en deux nouveaux nœuds, et ainsi de suite. Les arbres sont ainsi développés, jusqu'à atteindre une condition d'arrêt. Une règle d'arrêt classique consiste à ne pas découper des nœuds qui contiennent moins d'un certain nombre d'observations. Les nœuds terminaux, qui ne sont plus découpés, sont appelés les feuilles de l'arbre.

2. **Elagage de T_{max}.** Le principe est de développer l'arbre au maximum, c'est à dire T_{max}, puis de le remonter en partant des feuilles et supprimer les nœuds dont la division n'améliore pas significativement l'arbre. Un arbre maximal, c'est à dire un arbre issu de la phase précédente, est généralement inexploitable pour plusieurs raisons :

— Le nombre de feuilles est trop important. Or si l'on assimile le nombre de feuilles à la complexité du modèle, on obtient un modèle de grande complexité,

— L'arbre ainsi construit est beaucoup trop fidèle aux données d'apprentissage. En effet, puisque l'erreur d'un arbre T se mesure par $R(T) = \sum_{t \in T} R(t)$ on constate que si l'on compare les erreurs des arbres emboîtés, l'erreur est d'autant plus faible que le nombre de feuilles augmente, et $R(T_{max}) \approx 0$. $R(t)$ est défini, dans notre cas, par $R(t) = \frac{1}{n} \sum_{X_i \in t} (y_i - \bar{y}_t)^2$ qui représente la moyenne des résidus, autrement dit la somme des écarts au carré entre les y_i du nœud t et leur moyenne.

A contrario, un arbre constitué uniquement de la racine a une très faible complexité mais un biais élevé. Le principe de cette phase d'élagage consiste à faire intervenir un critère pénalisé qui va permettre d'effectuer un compromis entre la fidélité aux données et la dimension de l'arbre et à déterminer ainsi un "meilleur" sous-arbre pour chaque dimension. Ce critère est le suivant. Le critère (à minimiser) utilisé est le suivant :

$$R_\alpha(T) = R(T) + \alpha|T|$$

où :

— $|T|$ est le nombre de feuilles de l'arbre T,

— α un réel positif pénalisant la complexité de l'arbre,

— $R(T)$ représente l'erreur quadratique et caractérise donc la qualité de l'ajustement.

La fonction *rpart* utilisée sous R a une limite de profondeur d'arbre, c'est à dire que les arbres construits ne peuvent pas avoir plus qu'un certain nombre d'étages, ce qui limite le nombre de nœuds et ainsi le temps de calcul.

Une fois l'arbre construit, la fonction *prp* du package *rpart.plot* permet d'avoir une représentation de l'arbre. Dans notre cas, on obtient l'arbre de la figure 3.12 :

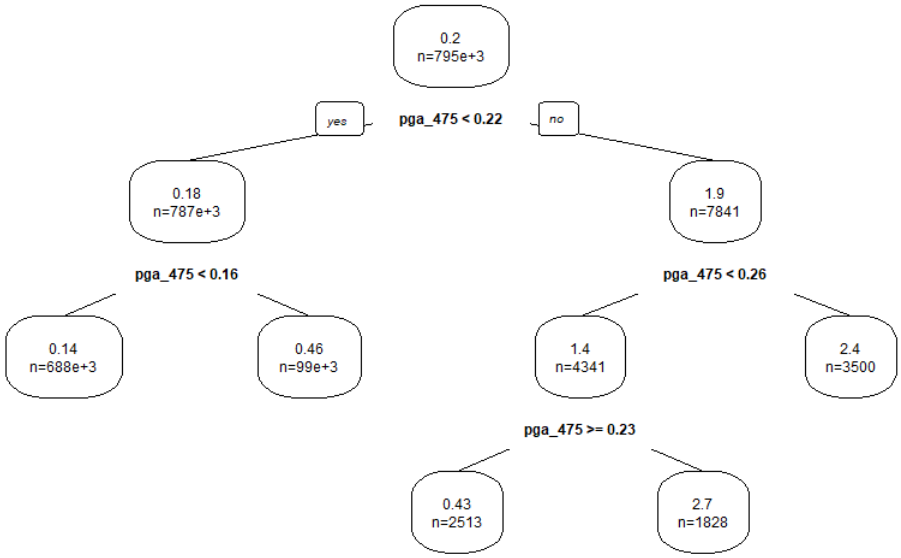


FIGURE 3.12 – Arbre résultat pour la discrétisation du PGA

Il semblerait qu'il soit optimal de considérer 5 classes de PGA :

— $[0.07, 0.16]$,
— $[0.16, 0.22]$,
— $[0.22, 0.23]$,
— $]0.23, 0.26]$,
— $[0.26, 0.267]$.

Le package *leaflet* de R, permettant de créer des cartes, nous permet de représenter ces zones de risque (figure 3.13) :

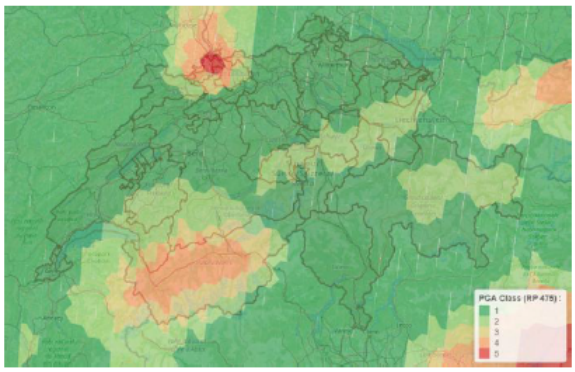


FIGURE 3.13 – Représentation géographique des 5 classes de risque

On retrouve bien un risque accru dans le nord et le sud de la Suisse.

A partir des pertes simulées par le modèle catastrophe naturelle, nous allons dans les parties suivantes utiliser des méthodes de modélisation statistique afin de modéliser la perte site à site. Bien entendu, le découpage en classes que nous venons d'effectuer nous sera très utile car pour certains modèles (comme le GLM) il est nécessaire d'avoir des variables explicatives découpées en classes au préalable. La première méthode que nous allons présenter dans la partie suivante est le modèle linéaire généralisé (ou GLM pour *Generalized linear model*).

3.4 Modèle linéraire généralisé

Définissons tout d'abord les notations utilisées dans cette partie.

— $Y = (Y_1, Y_2, ..., Y_n)'$ le vecteur colonne représentant les variables réponses. Il s'agit donc de ce que l'on souhaite modéliser,
— $X_i = (X_{i,1}, X_{i,2}, ..., X_{i,p})'$ le vecteur colonne représentant les variables explicatives pour la variable réponse Y_i,
— X la matrice de taille $n \times p$ dont les lignes sont les vecteurs lignes X_i',
— $\beta = (\beta_1, \beta_2, ..., \beta_p)'$ le vecteur colonne correspondant aux p paramètres du modèle.

Définition (famille exponentielle). Un modèle statistique $(\Omega, \mathcal{F}, (\mathbb{P}_{\theta,\phi})_{\theta\in\Theta,\phi>0}))$ est appelé famille exponentielle si les probabilités $(\mathbb{P}_{\theta,\phi})$ admettent une densité f par rapport à une mesure dominante avec

$$f_{\theta,\phi}(y) = c_\phi(y)\exp\left(\frac{y\theta - a(\theta)}{\phi}\right)$$

— θ s'appelle le paramètre canonique et ϕ le parametre de dispersion, souvent considéré comme un paramètre de nuisance (un paramètre de nuisance est un paramètre qui n'est pas d'un intérêt immédiat mais qui doit être pris en compte dans l'analyse des paramètres d'intérêt),

— $a(\theta)$ est de classe C^2 et convexe,

— $c_\phi(y)$ ne dépend pas de θ et sert à normaliser la fonction pour qu'il s'agisse bien d'une densité.

La table 3.2 est une liste non exhaustive de distributions usuelles appartenant à une famille exponentielle.

Loi	θ	$a(\theta)$	ϕ
Loi normale $\mathcal{N}(m,\sigma^2)$	m	$\frac{\theta^2}{2}$	σ^2
Loi gamma $\Gamma(k,\lambda)$	$\frac{-1}{\lambda}$	$-k\ln(-\theta)$	1
Loi de poisson $\mathcal{P}(\lambda)$	$\ln(\lambda)$	$\exp(\theta)$	1
Loi binomiale $\mathcal{B}(n,p)$	$\ln(\frac{p}{1-p})$	$n\ln(1+\exp(\theta))$	1
Loi binomiale négative NegBin(p,r)	$\ln(p)$	$-r\ln(1-\exp(\theta))$	1

TABLE 3.2 – Exemples de distributions usuelles appartenant à une famille exponentielle

De plus, pour les lois appartenant à une famille exponentielle, il est possible d'exprimer leur variance et espérance selon les paramètres de la définition :

Proposition (famille exponentielle). Si Y est distribuée selon une loi appartenant à une famille exponentielle, alors

$$\mathbb{E}[Y] = a'(\theta) \text{ et } Var[Y] = \phi a''(\theta)$$

Un modèle est un modèle linéaire généralisé (ou GLM) s'il vérifie les hypothèses suivantes :

1. $Y|X = x$ appartient à une famille exponentielle ;
2. $g\left(\mathbb{E}[Y|X]\right) = X\beta$ où g est une fonction bijective appelée fonction de lien.

On retrouve dans la table 3.3 une liste de fonctions de lien que l'on retrouve très régulièrement :

Nom de la fonction de lien	Formule
Lien identité	$g(x) = x$
Lien ln	$g(x) = \ln(x)$
Lien inverse	$g(x) = \frac{1}{x}$
Lien logit	$g(x) = \frac{x}{1-x}$

TABLE 3.3 – Fonctions de lien usuelles

Parmi ces fonctions, une est particulièrement intéressante car elle engendre un modèle multiplicatif ; il s'agit de la fonction de lien ln. En effet, pour une telle fonction, on a $\ln(\mathbb{E}[Y|X]) = X\beta$ et donc $\mathbb{E}[Y_i|X_i] = \exp(X_i'\beta) = \prod_{k=1}^{p} \exp(X_{i,k}\beta_k)$ pour tout i entre 1 et n.

3.4.1 Estimation des paramètres

Nous allons voir dans cette partie comment se fait l'estimation de β, qui est le vecteur utilisé ensuite pour prédire les valeurs de la variable réponse Y. Pour la suite, nous utiliserons les notations suivantes :

$$\begin{cases} \eta_i = X_i'\beta \\ \mu_i = \mathbb{E}[Y_i \mid X_i] = g^{-1}(X_i'\beta) = g^{-1}(\eta_i) \\ \theta_i = (a')^{-1}(\mu_i) = (a')_T^{-1}(g^{-1}(X_i'\beta)) = (a')^{-1}(g^{-1}(\eta_i)) \end{cases}$$

En reprenant les notations de la famille exponentielle, la log-vraisemblance s'écrit :

$$\ell(\beta) = \sum_{i=1}^{n} \ln f(Y_i;\beta,\phi) = \sum_{i=1}^{n} \underbrace{\left\{ \ln c_\phi(Y_i) + \frac{Y_i\theta_i - a(\theta_i)}{\phi} \right\}}_{:=\ell_i(\theta_i)}$$

Nous souhaitons donc trouver, pour chaque $j \in \{1, ..., n\}$, l'estimateur du maximum de vraisemblance de β_j. On écrit et calcule donc la dérivée de ℓ par rapport à β_j :

$$\frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial \ell_i(\theta_i)}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial \ell_i}{\partial \theta_i}\frac{\partial \theta_i}{\partial \beta_j}$$

La première dérivée de la somme se calcule comme suit :

$$\frac{\partial \ell_i}{\partial \theta_i} = \frac{Y_i - a'(\theta_i)}{\phi} = \frac{Y_i - \mu_i}{\phi}$$

Et la seconde :

$$\frac{\partial \theta_i}{\partial \beta_j} = \frac{\partial \theta_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j} = \frac{\partial \theta_i}{\partial \eta_i}x_{i,j} = \frac{x_{i,j}}{g'(\mu_i)a''(\theta_i)}$$

Le calcul de $\frac{\partial \theta_i}{\partial \eta_i}$ se fait en calculant $\left(\frac{\partial \theta_i}{\partial \eta_i}\right)^{-1}$.
Par conséquent, l'estimateur du maximum de vraisemblance de β est solution de

$$X'D\left(Y - g^{-1}(X\beta)\right) = 0.$$

où D est la matrice diagonale dont les coefficients sont égaux à $\frac{1}{g'(\mu_i)a''(\theta_i)}$.

Dans le cas général, on ne sait pas résoudre explicitement cette équation. L'estimateur du maximum de vraisemblance est calculé numériquement, souvent grâce à des algorithmes de descente de gradient.

3.4.2 La famille Tweedie

Dans ce chapitre, nous allons évoquer les GLM basés sur la famille Tweedie. Les distributions Tweedie généralisent de nombreuses distributions classiques appartenant à la famille exponentielle et souvent utilisées dans les GLM comme la loi Normale, de Poisson, de Gamma ou encore la Normale inverse.

3.4.2.1 Présentation générale

Définition (loi de Tweedie). Une loi X appartenant à une famille exponentielle et ayant une variance de la forme $\mathbb{V}(\mu) = \phi\mu^{\xi}$ est appelée loi de Tweedie, où ξ et $\phi \in \mathbb{R} \setminus]0, 1[$ et $\mu = \mathbb{E}(X)$.
Le paramètre ξ est appelé *paramètre de puissance* ou *paramètre d'index* de la famille Tweedie.
Par exemple, les lois citées précédemment vérifient bien cette égalité. En effet, on a :

- $\mathbb{V}(\mu) = \sigma^2\mu^0 = \sigma^2$ pour la loi normale $\mathcal{N}(\mu, \sigma^2)$: $\phi = \sigma^2$ et $\xi = 0$,
- $\mathbb{V}(\mu) = \mu^1$ pour la loi de Poisson $\mathcal{P}(\mu)$: $\phi = 1$ et $\xi = 1$,
- $\mathbb{V}(\mu) = \sqrt{k}\mu^2$ pour la loi Gamma $\Gamma(k, \theta)$: $\phi = \sqrt{k}$ et $\xi = 2$,
- $\mathbb{V}(\mu) = \mu^3$ pour la loi inverse Gaussienne $\mathcal{IG}(\mu, \lambda)$: $\phi = 1/\lambda$ et $\xi = 3$.

On regroupe dans la table 3.4 les différentes lois de Tweedie selon les valeurs de ξ :

Distribution	xi
Stable Extreme	$\xi < 0$
Normal	$\xi = 0$
Pas de loi appartenant à la famille exponentielle	$0 < \xi < 1$
Poisson	$\xi = 1$
Poisson-Gamma	$1 < \xi < 2$
Gamma	$\xi = 2$
Stable Positive	$\xi > 2$

TABLE 3.4 – Lois de Tweedie selon ξ

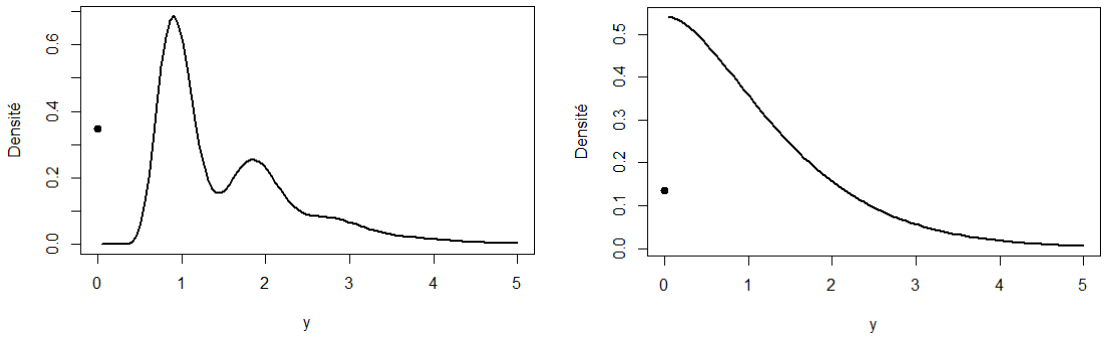
Cette famille permet de couvrir un grand nombre des lois pouvant être utilisées dans un GLM sur R simplement en faisant balayer ξ dans un intervalle adapté.
Les distributions de Tweedie peuvent être utilisées pour modéliser plusieurs types de données, car elles peuvent être de différents types. Selon les valeurs de ξ, les lois de Tweedie associées servent donc à modéliser des données différentes :

- pour $\xi \leq 0$, les lois sont adaptées pour modéliser des données continues négatives et positives,
- pour $\xi = 1$, les lois sont adaptées pour modéliser des données discrètes, donc à valeurs dans $\mathbb{N}$. On peut penser au cas classique en tarification pour modéliser la fréquence d'un sinistre,

— pour $1 < \xi < 2$, les lois sont adaptées pour modéliser des données continues positives, mais avec une masse en 0. Ce cas sera développé dans la partie suivante, car nous utiliserons ces lois pour modéliser le taux de destruction sur nos données sismiques,

— pour $\xi > 2$, les lois sont adaptées pour modéliser des données continues positives. Typiquement, on utiliser ces lois pour modéliser le coût d'un sinistre.

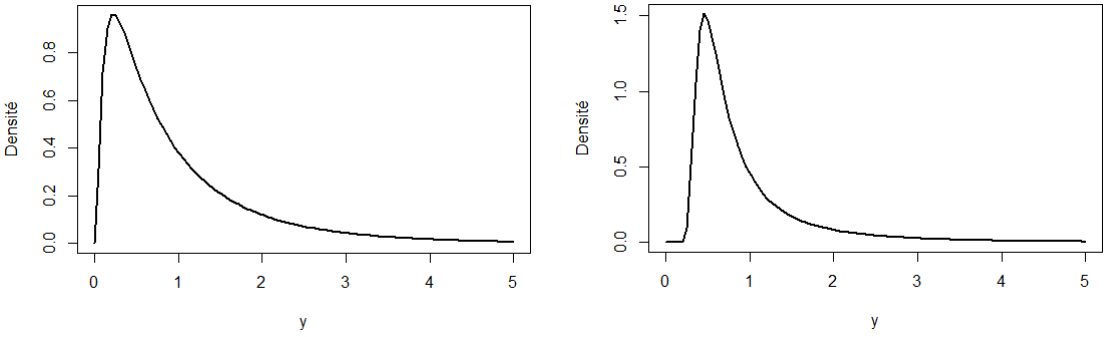
On retrouve ci-dessous des exemples de densités de lois de Tweedie pour plusieurs valeurs de ξ construites grâce au package *tweedie* du logiciel R :
Tout d'abord, on a 2 exemples pour $1 < \xi < 2$. On voit bien que dans ce cas, les lois sont continues pour $y > 0$ et ont une masse en 0 (représentée par le point en abscisse 0). La valeur de ϕ a été fixée à 1.



(a) Cas $\xi = 1.05$ (b) Cas $\xi = 1.5$

FIGURE 3.14 – Exemples de densités de lois de Tweedie $(1 < \xi < 2)$

Puis on retrouve 2 exemples dans le cas $\xi > 2$:



(a) Cas $\xi = 1.05$ (b) Cas $\xi = 1.5$

FIGURE 3.15 – Exemples de densités de lois de Tweedie $(\xi > 2)$

Dans notre cas, nous souhaitons modéliser le taux de destruction engendré par les séismes. Un séisme ayant une fréquence faible (donc beaucoup de sites ne sont pas touchés et ont un taux de destruction de 0%), on souhaite donc modéliser une variable continue pour les valeurs strictement positives, et avec des zéros. Le cas $1 < \xi < 2$ nous intéresse donc particulièrement. On appelle ce type de distribution une distribution **Poisson-Gamma**.

Il est également possible de déterminer la relation entre les paramètres d'une loi appartenant à une famille exponentielle et ceux d'une Tweedie $TW_\xi(\mu, \phi)$. En effet, en reprenant les différentes notations précédentes, on a les égalités suivantes :

$$\theta = \begin{cases} \frac{\mu^{1-\xi}}{1-\xi} & \text{for } \xi \neq 1 \\ \log(\mu) & \text{si } \xi = 1 \end{cases}$$

et

$$a(\theta) = \begin{cases} \frac{\mu^{2-\xi}}{2-\xi} & \text{for } \xi \neq 2 \\ \log(\mu) & \text{si } \xi = 2 \end{cases}$$

3.4.2.2 Estimation du paramètre ξ

Pour lancer un GLM, il est nécessaire de spécifier une valeur de ξ. Tout d'abord, il est simple de trouver un encadrement de cette valeur selon les données que l'on souhaite modéliser. En effet, on a vu dans la partie précédente que certaines valeurs de ξ sont adaptées à certains types de données. Par exemple, pour modéliser une grandeur positive continue avec une masse en zéro, on aura forcément $1 < \xi < 2$. Pour estimer le plus précisément possible une valeur de ξ, il y a différentes méthodes possibles.

Utilisation de la définition d'une Tweedie

La première méthode utilise le fait que pour une loi de Tweedie, $Var(Y) = \mathbb{V}(\mu) = \phi\mu^\xi$, ce qui donne, par un passage au logarithme, $\ln(Var(Y)) = \ln(\phi) + \xi \ln(\mu)$. Une méthode simple pour estimer ξ est de diviser les données en un petit nombre de groupes et de tracer le logarithme des variances du groupe par rapport au logarithme des moyennes du groupe, et effectuer une régression linéaire afin de déterminer le coefficient directeur, qui permet d'approcher ξ. Cependant, l'estimation de ξ peut dépendre de la manière dont les données sont réparties en groupes.

Méthode de la vraisemblance profilée

Une autre méthode, plus rigoureuse et qui ne dépend pas de la façon dont on groupe les données, est de calculer l'estimateur du maximum de la vraisemblance profilée de ξ. Un moyen pratique d'organiser les calculs est d'utiliser la vraisemblance profilée pour estimer ξ. La vraisemblance profilée se calcule de la manière suivante : différentes valeurs de ξ sont choisies afin d'être testées, puis le GLM de Tweedie est ajusté pour chacune de ces valeurs, en considérant que ξ est fixe. Une log-vraisemblance est calculée pour chaque valeur de ξ. La valeur de ξ donnant la plus grande log-vraisemblance est l'estimateur du maximum de la vraisemblance profilée. Un tracé du logarithme de la vraisemblance profilée contre diverses valeurs de ξ est souvent utile. L'une des difficultés de cette méthode est que la fonction de vraisemblance des lois de Tweedie doit être calculée. Cependant, il n'existe pas de formule fermée pour la densité des lois de Tweedie, sauf pour certains cas particuliers. Seulement quelques distributions particulières, telles que la Normale, correspondant à une distribution Tweedie avec $\xi = 0$, la Poisson, correspondant à une distribution Tweedie avec $\xi = 1$, la Gamma, correspondant à une distribution Tweedie avec $\xi = 2$ et l'Inverse Gaussienne, correspondant à une distribution Tweedie avec $\xi = 3$, peuvent être exprimées sous forme fermée.
Toutefois, des méthodes numériques permettant d'évaluer avec précision les densités de

Tweedie sont utilisées dans la fonction R tweedie.profile() (du package tweedie de R) pour calculer l'estimation du maximum de la vraisemblance profilée de ξ.

3.5 Sélection de modèles

De nombreux indicateurs peuvent être utilisés afin d'évaluer la pertinence d'un modèle, et comparer plusieurs modèles entre eux.
Afin d'évaluer la performance prédictive des différents modèles, nous utilisons différentes mesures, qui sont extrêmement utilisés dans toute analyse comparative :

- **le RMSE (Root Mean Square Error) :** la fonction de perte quadratique classique. Le RMSE représente la moyenne du carré des erreurs site par site. Les surestimations des observations ne sont pas compensées par les sous-estimations car on prend le carré des erreurs.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}$$

- **le RMAE (Root Mean Absolute Error) :** la fonction de perte absolue. Le RMAE représente la moyenne de la valeur absolue des erreurs site par site.

$$RMAE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}|Y_i - \hat{Y}_i|}$$

- **L'erreur cumulée :** Elle est définie comme la différence entre la somme des pertes réalisées et la somme des pertes du modèle de tarification. Elle donne une idée de l'équilibre du modèle.

$$EC = \sum_i Y_i - \sum_i \hat{Y}_i$$

- **La courbe de Lorenz et l'indice de Gini associé :** il s'agit d'un outil également très utilisé qui permet de mesurer la capacité de segmentation d'un modèle. Pour définir ces notions, nous allons expliquer l'utilité première de la courbe de Lorenz, et comment la transposer pour nos problématiques. La courbe de Lorenz est la représentation graphique de la fonction qui, à la part x des détenteurs d'une part d'une grandeur, associe la part y de la grandeur détenue. Elle a été développée par Max O. Lorenz en 1905 en vue d'une représentation graphique des inégalités de revenus. Utilisons alors les notations suivantes :
  - $(X_1, X_2, ..., X_n)$ des variables aléatoires représentant les revenus de n individus de la population et $(x_1, x_2, ..., x_n)$ une réalisation de cet échantillon,
  - $(x_{(1)}, x_{(2)}, ..., x_{(n)})$ la statistique d'ordre, c'est à dire les revenus triées par ordre croissant,
  - $L_p = \frac{\sum_{i=1}^{p} x_{(i)}}{\sum_{i=1}^{n} x_i}$ la proportion de revenus des p individus ayant les plus bas revenus par rapport au revenu total gagné par tous les individus.

  La courbe de Lorenz est la représentation graphique de la répartition des revenus au sein d'une population. Plus précisément, la courbe de Lorenz représente le nuage de points $(i/n, L_p)$ pour $p = \{1, ..., n\}$.

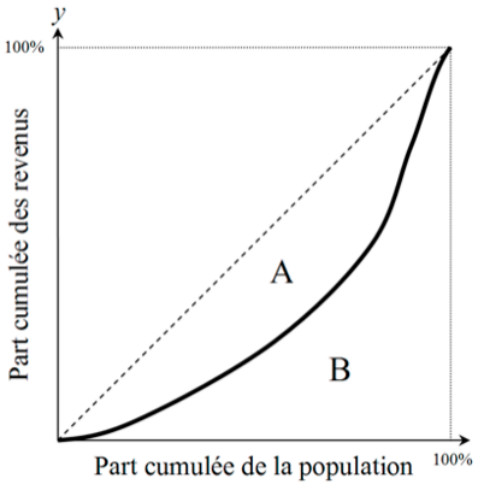


FIGURE 3.16 – Courbe de Lorenz

La droite en pointillés représente la situation dans laquelle la distribution des revenus disponibles serait parfaitement égalitaire. C'est-à-dire lorsque les 10% les plus modestes auraient 10% des revenus disponibles total, les 20% les plus modestes détiendraient 20% de l'ensemble des revenus disponibles, etc. La courbe de Lorenz part de l'origine (0% de la population perçoit 0% du revenu) et se termine au point (1 ; 1) ou (100% de la population perçoit 100% du revenu).

L'indice de Gini, noté G, se définit de la manière suivante :

$$G = \frac{A}{A+B}$$

Il est clair que $A + B = 0.5$, ce qui nous permet d'écrire l'indice de Gini $G = 2A$ (ou $G = 1 - 2B$). Il est donc égal à deux fois la surface comprise entre la courbe de Lorenz et la diagonale entre pointillés. Cette valeur est clairement comprise entre 0 et 1. Un indice de Gini proche de 0 signifierait que tous les revenus sont équitablement répartis sur la population. En revanche, un indice proche de 1 signifierait que les revenus sont très mal répartis.
Plus formellement, si l'on note $L(x)$ la fonction représentant la courbe de Lorenz, le calcul de G peut se faire de la manière suivante :

$$G = 2\int_0^1 L(x)dx - 1$$

Voila comment est défini historiquement l'indice de Gini. En sciences actuarielles, on trouve diverses définitions de l'indice de Gini selon le résultat que l'on souhaite regarder. Nous allons définir ci-dessous quelle définition nous utilisons dans le cadre de ce mémoire. Cet indice nous permettra de comparer certains modèles, en particulier leur segmentation du risque.

Définition de l'indice de Gini pour notre étude

Les données utilisées pour construire l'indice de Gini dans notre cas sont les valeurs prédites notées $(\hat{y_1}, \hat{y_2}, ..., \hat{y_n})$ et les valeurs observées que l'on note $(y_1, y_2, ..., y_n)$. Nous ordonnons alors les valeurs de y_i selon les valeurs ordonnées par ordre décroissant des $\hat{y_i}$. On note $(y_i^{\text{ord}})_i$ les y_i ordonnées de la sorte. Nous représentons alors la courbe avec :

— En abcisse : l'exposition cumulée, c'est à dire les $\frac{i}{n}$ pour i entre 1 et n.

— En ordonnée : La part du risque cumulé sur le risque totale en respectant l'odre défini précédemment, c'est à dire, pour tout i entre 1 et n, les $\sum_{k=1}^{i} y_i^{\text{ord}} / \sum_{k=1}^{n} y_i$

On note alors L la fonction ainsi définie. Cette courbe nous permet alors de voir si le modèle segmente bien le risque. Nous cherchons un modèle qui proposera un taux de destruction le plus segmenté possible c'est-à-dire le plus fort possible pour les sites particulièrement à risque et le plus faible pour les sites peu vulnérables. L'indice de Gini se calcule alors avec la formule $G = 2\int_0^1 L(x)dx - 1$. Plus l'indice de Gini est élevé et plus le modèle est bon pour différencier les grosses pertes des petites. Des exemples de courbes seront donnés lorsque l'on rentrera dans le détail des modèles.

3.6 Affectation de poids aux portefeuilles fictif et réel

On rappelle que la base de données totale d'assurés que l'on a en entrée du modèle est une fusion entre 2 bases différentes :

— Une base ne comprenant que les assurés réels, c'est à dire ceux qui ont réellement souscrit une assurance en cas de sinistre sismique. On compte 33 000 assurés réels,

— Une base fictive ne comprenant que des assurés fictifs et permettant d'avoir une meilleure vision du risque. On compte 760 000 assurés fictifs.

Les bâtiments compris dans la base fictive sont très utiles car ils enrichissent l'étude, mais il paraît plus approprié de donner plus de poids aux assurés réels pour plusieurs raisons :

— Etant donné le faible nombre d'assurés réels devant les fictifs, on souhaite augmenter le poids des assurés réels afin qu'ils aient plus d'importance dans la modélisation,

— Parmi les biens assurés fictifs, certains bâtiments ne sont pas réalistes et ne seront jamais assurés dans la zone étudiée. C'est pourquoi il faut également limiter l'influence de ce type de bâtiment, en diminuant le poids dans la modélisation des bâtiments fictifs.

Ainsi, nous allons utiliser l'argument *weight* de la fonction *glm* de R afin de donner plus d'importance aux biens assurés réels. Cet argument est bien entendu par défaut un vecteur de taille le nombre d'observations et uniquement composé de 1. Pour créer notre vecteur personnalisé de poids, on prend en compte plusieurs critères :

1. La présence ou non des variables explicatives liées au bâtiment du portefeuille fictif dans le portefeuille d'assurés réels. On rassemble toutes les combinaisons existantes de variables présentes dans la base d'assurés réels, et on compte le

nombre d'occurrences pour chacune de ces combinaisons. Ensuite on parcourt le portefeuille fictif, et pour chaque ligne :

- Si la combinaison de variables existe dans le portefeuille réel, on y associe le poids correspondant,
- Sinon, on y associe le poids correspondant au 1er quartile du vecteur de poids du portefeuille réel. On a choisi le premier quartile pour donner un faible poids aux bâtiments qui n'existent pas dans le portefeuille réel.

On appelle « poids_1 » les poids résultants de cette étape.

2. La proximité géographique des sites du portefeuille fictif avec les sites assurés réels. On procède, pour chaque ligne du portefeuille fictif, de la manière suivante :
   - On calcule, grâce à la longitude et latitude, la distance à chaque site du portefeuille réel grâce à la formule suivante (pour 2 sites s_1 et s_2) :

     $$D_{s_1,s_2} = 6371 \times \arccos\left[\sin(lat_{s_1})\sin(\text{lat}_{s_2}) + \cos(\text{lat}_{s_1})\cos(\text{lat}_{s_2})\cos(\text{long}_{s_2} - \text{long}_{s_1})\right]$$

   - On prend la distance minimale entre toutes les distances aux sites assurés réels et associe un poids en conséquence. La valeur choisie est $\frac{1}{D_{min}}$.

   On appelle « poids_2 » les poids résultants de cette étape

3. On choisit, pour chaque site du portefeuille fictif, un poids final valant $\log(\text{poids}_1 \times \text{poids}_2)$.

L'idée est donc d'associer des poids aux sites du portefeuille fictif, selon un critère de proximité géographique avec les sites du portefeuille réel et selon l'existence des types de bâtiments dans ce même portefeuille réel. Les valeurs que nous avons sélectionnées sont arbitraires, l'idée étant surtout d'instaurer une hiérarchie entre les sites en termes de poids dans la modélisation.

Finalement, nous associons aux sites assurés réels un poids de la forme $\alpha \times P$ où P est le max de tous les poids associés au portefeuille fictif. On choisit évidemment un α supérieur à 1.
La limite de cette approche est que ces poids ont été choisis de manière arbitraire sans pouvoir réellement tester la pertinence des poids choisis. Il n'existe pas de méthodes précises pour les déterminer, le but étant dans tous les cas de donner plus de poids aux assurés réels.

3.7 Résultats des modèles

Afin de déterminer la modélisation la plus efficace pour reproduire la perte économique annuelle moyenne de chaque couverture provenant du modèle interne, plusieurs techniques ont été testées parmi lesquelles on retrouve :

1. Un modèle GLM fréquence/coût moyen où la fréquence est la probabilité d'avoir une perte annuelle moyenne et le coût est le taux de destruction moyen en cas de perte annuelle moyenne positive.
2. Un unique modèle GLM Tweedie, où la variable à expliquer est le taux de destruction (comprenant les zéros et valeurs positives).
3. Un GLM par zone de risque (zone risquée/non risquée).

4. un GLM Tweedie avec des informations géographiques supplémentaires
5. 2 modèles de machine learning : le réseau de neurones ainsi que le XGBoost.

Le but de cette partie est donc de confronter tous ces modèles et choisir quel modèle nous souhaitons retenir. Pour le déterminer, nous allons :

— Utiliser les indicateurs définis dans la partie 3.5,

— Tout en prenant en compte les avantages et inconvénients de chaque modèle.

On comparera, après présentation de chaque modèle, les résultats obtenus. Pour les modèles GLM, il faut que toutes les variables quantitatives soient discrétisées. Ainsi, on utilisera les 5 classes de risque construites à partir du PGA dans la partie 3.3.1. Pour les méthodes de machine learning, il n'est pas nécessaire d'effectuer cette discrétisation.

3.7.1 Modèle fréquence/Coût moyen

Le premier modèle que nous mettons en place est un modèle de fréquence et coût moyen, qui est très largement utilisé en modélisation assurantielle.

Modèle de fréquence

Nous devons retraiter légèrement notre base de données afin d'avoir la variable cible voulue. Nous créons donc la variable suivante :

$$Y_i = \begin{cases} 1 \text{ si la perte annuelle totale est positive} \\ 0 \text{ sinon} \end{cases}$$

Cette variable sera notre variable à expliquer pour le modèle de fréquence.

Bien entendu, nous n'allons pas utiliser toutes les variables explicatives pour ce modèle, mais seulement les variables liées à la zone géographique, qui sont le **cresta**, le **PGA** ainsi que le **type de sol**. On effectue alors, pour la modélisation de la fréquence, un GLM Binomial avec la fonction de lien logit. L'erreur globale obtenue est la suivante, pour les bases de test et de train :

	Train	Test
Erreur globale	-30%	-19%

TABLE 3.5 – Erreur globale modèle de fréquence

Les résultats de fréquence sont plutôt mauvais, mais nous verrons juste après les indicateurs obtenus une fois la fréquence et sévérité combinées.

Modèle de sévérité

Pour le modèle de sévérité, on ne garde que les sites touchés et on modélise le coût des sinistres. Pour la modélisation de la sévérité, nous choisissons un GLM Gamma avec la fonction de lien log. Les indicateurs obtenus pour le modèle de sévérité sont les suivants :

	Train	Test
Erreur globale	-6.9%	2.25%
RMAE	0.298	0.229
RMSE	0.492	0.421
Gini	52%	70.80%

TABLE 3.6 – Indicateurs pour le modèle de sévérité

Les résultats de sévérité sont bons.

Combinaison Modèle de fréquence et coût

Regardons maintenant ce qui nous intéresse vraiment pour être capable de comparer le modèle coût/fréquence avec les autres modèles mis en place. On combine les modèles de coût et de fréquence ; en découlent alors les résultats suivants :

	Train	Test
Erreur globale	-1.76%	-2.13%
RMAE	0.314	0.322
RMSE	0.603	0.649
Gini	60.74%	60.44%

TABLE 3.7 – Indicateurs pour le modèle coût/fréquence

Malgré les résultats moyens du modèle de fréquence, la combinaison du modèle fréquence et sévérité donne un résultat convenable. Mais on peut facilement imaginer que les autres modèles testés donneront de meilleurs résultats étant donné la faiblesse du modèle de fréquence.

3.7.2 GLM Tweedie

On souhaite dans cette partie modéliser directement le taux de destruction par un unique GLM Tweedie, sans découpage coût et fréquence. Les variables explicatives prises en compte sont :

— Le Cresta, PGA et type de sol comme variable géographique,

— Ł'année de construction et le type de structure comme variable liée au bâtiment,

— **L**e type d'occupation et le type de couverture.

Pour déterminer la valeur du paramètre ξ à utiliser pour la loi de Tweedie, nous utilisons la méthode de la vraisemblance profilée décrite dans la partie 3.4.2.2, grâce à la fonction tweedie.profile du package *tweedie*. Cette fonction a besoin, en argument, de la formule utilisée dans le GLM et des données utilisées. Un argument facultatif *do.plot* fixé à *TRUE* permet d'avoir une représentation graphique du résultat. On précise également en argument les valeurs de ξ que l'on souhaite tester. Comme précisé précédemment, on souhaite une valeur de ξ comprise entre 1 et 2, car cela correspond à des lois Poisson-Gamma. Le graphe de la figure 3.17 représente la log-vraisemblance en fonction des différentes valeurs de ξ :

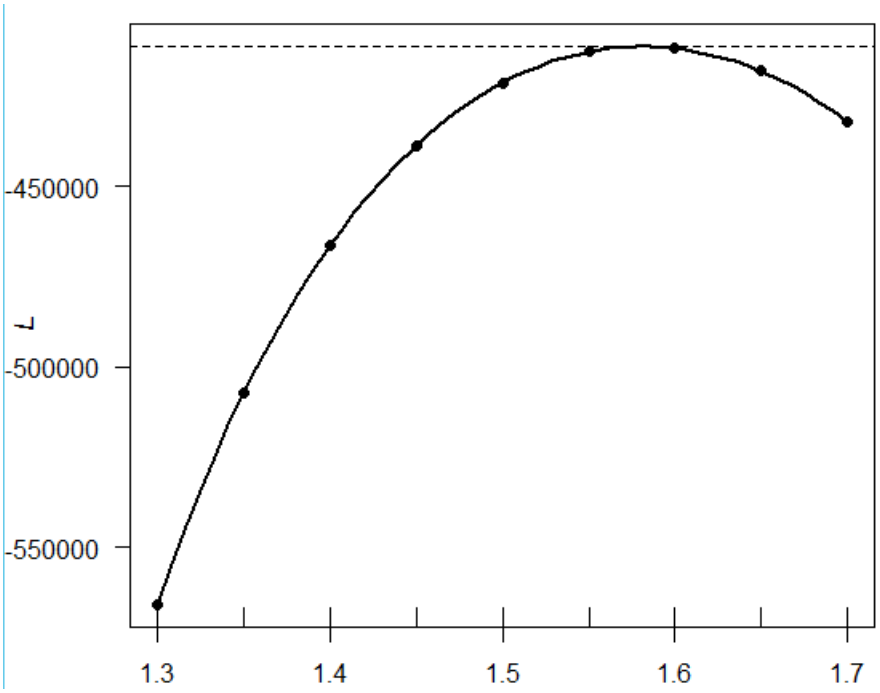
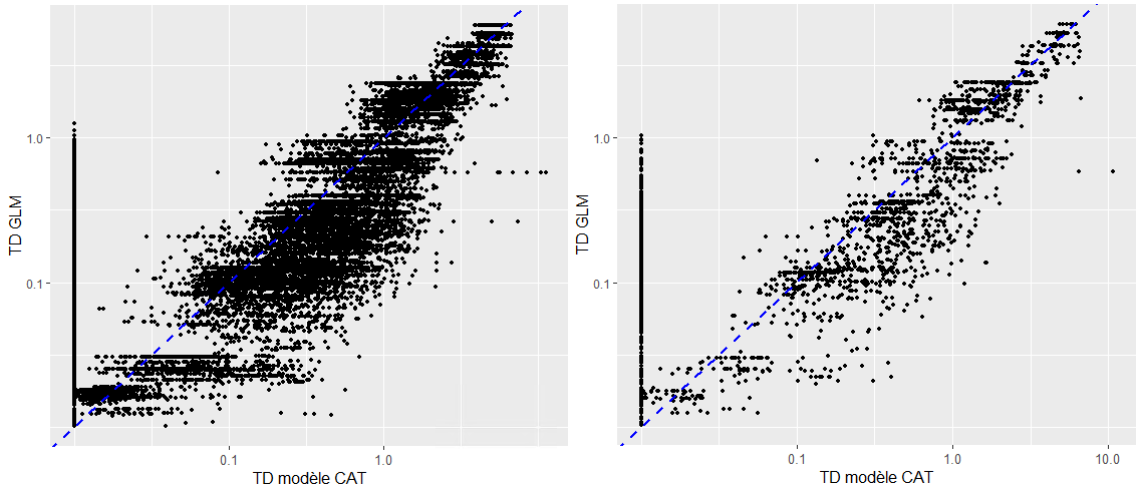


FIGURE 3.17 – Log-vraisemblance pour plusieurs valeurs de ξ

On remarque que la valeur maximale de log-vraisemblance est atteinte pour ξ valant 1.58. On utilise donc cette valeur de ξ pour le modèle.

Résultats du GLM

On représente sur la figure 3.18 le taux de destruction (TD) prédit par le GLM en fonction du taux de destruction à prédire (qui est celui provenant du modèle CAT).



(a) Base d'apprentissage

(b) Base de test

FIGURE 3.18 – Taux de destruction prédit vs taux de destruction modèle CAT

La base d'apprentissage a été réduite aux données du portefeuille réel, pour rendre le graphique plus clair. La tendance observée est plutôt bonne. En effet, les points sont relativement proches de la droite d'équation y=x tracée en bleu.

Les différents indicateurs obtenus sont regroupés dans la table suivante :

	Train	Test
Erreur globale	1.52%	2.25%
RMAE	0.224	0.229
RMSE	0.380	0.421
Gini	70.54%	70.80%

TABLE 3.8 – Indicateurs pour le modèle 2

Les indicateurs de ce modèle sont bons et nettement meilleurs que le modèle précédent.

Comme on pouvait s'y attendre, on remarque que les zones de risque (c'est à dire les classes de PGA) ont un poids très important sur la modélisation finale. En effet, on a regroupé dans la table 3.9 le facteur multiplicatif appliqué à la prime selon les zones de risque (en d'autres termes, les modulations liées aux zones de risque données par le GLM).

Zone de risque	Modulation
1	1
2	1.189
3	2.432
4	7.246
5	10.124

TABLE 3.9 – Modulations par classe de risque

Pour obtenir ce tableau, il suffit d'appliquer la fonction exponentielle à la modulation en sortie de GLM, car nous avons choisi la fonction de lien log.

Dans les indicateurs, on peut voir qu'on utilise l'indice de Gini qui a été défini dans la partie 3.5 portant sur la sélection de modèles. La courbe permettant d'obtenir l'indice de Gini est précisée à la figure 3.19 :

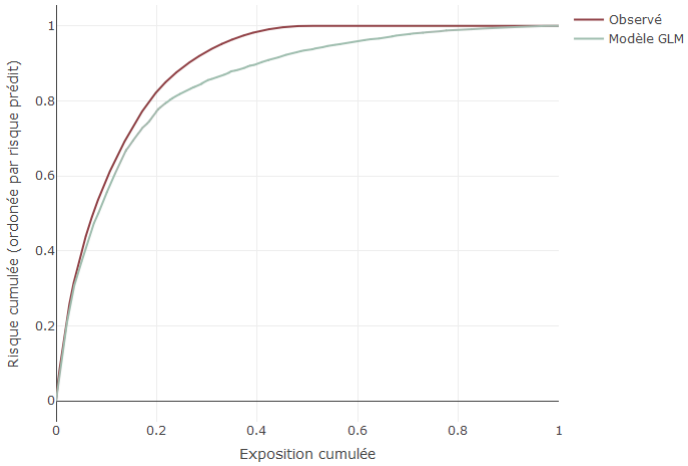


FIGURE 3.19 – Courbes permettant le calcul de l'indice de Gini

On remarque que les 2 courbes sont très proches. En effet, l'indice de Gini des taux de destruction observée (c'est à dire issus du modèle CAT) est de 78%, et on rappelle que celui du modèle GLM tweedie est de 70%. Dans tous les cas, l'objectif est d'avoir l'indice de Gini le plus élevé en comparant les modèles entre eux.

3.7.3 GLM par zone de risque

Dans cette partie, nous allons effectuer 2 GLMs : un pour les zones risquées et un pour les zones non risquées. Nous définirons en dessous ce que l'on entend par zone risquée. Expliquons tout d'abord ce qui nous a poussé à calibrer 2 GLMs différents. Faisons tout d'abord une représentation simple du taux de destruction par un histogramme :

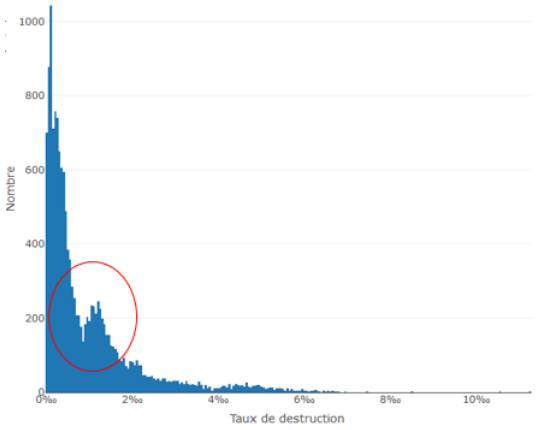


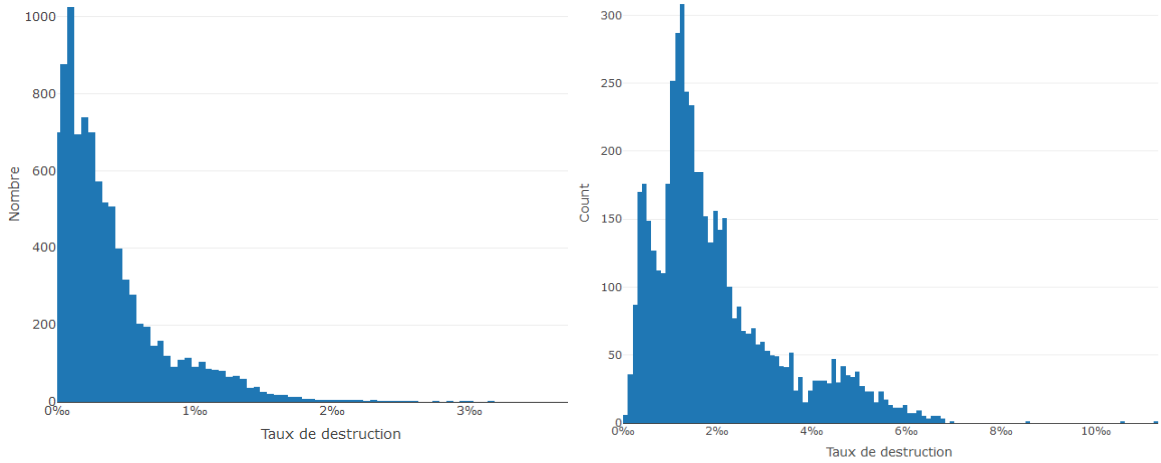
FIGURE 3.20 – Histogramme du taux de destruction

On remarque un creux entouré en rouge sur la figure ci-dessous, qui peut difficilement être capté par un unique GLM. C'est pourquoi nous souhaitons effectuer 2 GLMs différents, pour éviter ce type de creux dans les données à expliquer. Nous séparons alors les données en 2 de la manière suivante :

— Les bâtiments appartenant aux zones « risquées », c'est à dire des scores de risque liés au PGA valant 4 et 5,

— Les bâtiments appartenant aux zones « non risquées » ou « moins risquées », c'est à dire des scores de risque liés au PGA valant 1, 2 et 3.

Regardons les histogrammes du taux de destruction une fois nos deux bases séparées :



(a) Zone non risquée

(b) Zone risquée

FIGURE 3.21 – Histogrammes du taux de destruction selon les zones de risque

On remarque que pour les zones non risquées, ce creux disparaît. En revanche, pour les zones risquées, il est encore un peu présent, mais moins prononcé. Nous allons alors procéder à 2 GLMs distincts :

— Pour les zones non risquées, un GLM Tweedie de paramètre ξ valant 1.005,

— Pour les zones risquées où l'on remarque que le pic des valeurs ne se situe pas à 0 ou très proche de 0, on procède à un GLM Gamma.

Pour calculer le taux de destruction prédit, on effectue le calcul suivant :

$$TD^{\text{pred}} = TD_{\text{GLM Tweedie}} \mathbb{1}_{\text{Zone risque} \leq 3} + TD_{\text{GLM Gamma}} \mathbb{1}_{\text{Zone risque} > 3}$$

Les indicateurs résultants de la combinaison des deux modèles sont les suivants :

	Train	Test
Erreur globale	1.49%	1.84%
RMAE	0.215	0.222
RMSE	0.366	0.413
Gini	70.88%	71.07%

TABLE 3.10 – Indicateurs pour le modèle 3

Les indicateurs obtenus ne se sont pas sensiblement améliorés par rapport au modèle précédent (le modèle 2) qui est plus facile à mettre en oeuvre.

3.7.4 GLM Tweedie avec plus d'informations géographiques

Dans les modèles précédents, l'information géographique caractérisant le risque est le *Peak Ground Acceleration* pour un temps de retour de 475 ans, c'est à dire pour un évènement étant censé se dérouler tous les 475 ans. Une autre formulation de la période

de retour de 475 ans peut être la suivante : il s'agit également du PGA ayant 10% de chance de se produire en 50 ans. (l'équivalence entre les 2 formulations a été prouvée en fin de partie 2.2.6)

Une période de retour de 475 ans est communément utilisée pour caractériser le risque sismique. Dans cette partie, nous allons étudier les résultats avec des informations géographiques supplémentaires telles que :

- le PGA avec une période de retour de 200 ans,
- le PGA avec une période de retour de 1000 ans,
- le PGA avec une période de retour de 1000 ans,
- l'accélération spectrale pour une période naturelle de $0.1s$ et avec une période de retour de 475 ans,
- l'accélération spectrale pour une période naturelle de $0.3s$ avec une période de retour de 475 ans.

L'accélération spectrale tente de reproduire l'accélération maximale subie par un immeuble. L'immeuble est modélisé par un oscillateur vertical de masse nulle ayant la même période de vibration que le bâtiment. Elle s'exprime également en fonction de g et est caractérisée par une période naturelle (en s) et un coefficient d'amortissement propre à chaque bâtiment. Ainsi, c'est un peu différent du PGA qui caractérise l'accélération maximale du sol. L'accélération spectrale se focalise donc plus sur le bâtiment.

Toutes ces grandeurs sont corrélées positivement entre elles, comme le démontre la représentation de la matrice de corrélation (figure 3.22) :

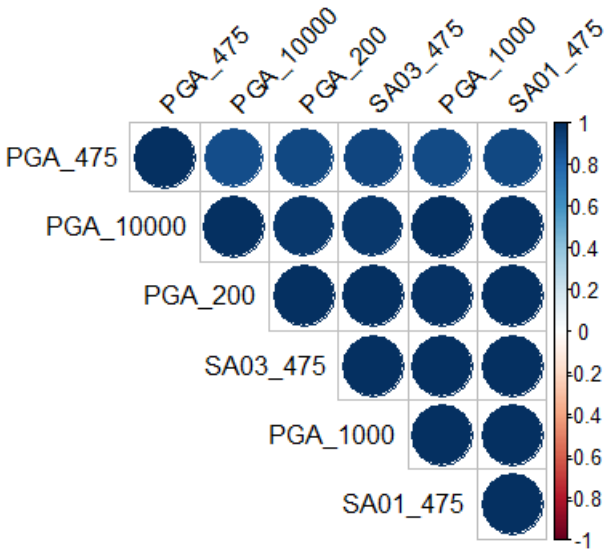


FIGURE 3.22 – Corrélation entre les variables géographiques

Etant donné les corrélations positives entre toutes les variables géographiques ajoutées, on peut imaginer qu'elles n'amélioreront pas le modèle, le PGA pour une période de retour de 475 ans semblant être suffisant. Nous effectuons malgré tout un GLM Tweedie après avoir ajouté ces variables géographiques.

Les indicateurs obtenus sont les suivants :

	Train	Test
Erreur globale	1%	1.55%
RMAE	0.223	0.227
RMSE	0.379	0.418
Gini	70.47%	70.57%

TABLE 3.11 – Indicateurs pour le modèle 4

L'ajout de toutes les variables, qui en plus sont corrélées positivement entre elles, n'améliore pas significativement le modèle 2, qui est un GLM effecuté avec moins de variables. (les indicateurs sont même très proches). Par conséquent, l'apport de toutes les variables géographiques n'est pas suffisant.

Nous avons donc calibrer plusieurs modèles GLM et avons obtenu des indicateurs déjà satisfaisants. Pour tenter d'améliorer ces indicateurs (et donc obtenir de meilleurs modèles), nous allons étudier, tester et calibrer 2 modèles de machine learning : le réseau de neurones et XGBoost.

3.7.5 Modèles de machine learning

3.7.5.1 Présentation du package R h2o

Les modèles de *machine learning* seront calibrés grâce au package *H2o* de R, dont voici une rapide présentation. « H2O » est une bibliothèque open source de *machine*

learning très populaire et est connue pour sa rapidité et son évolutivité. Elle est conçue pour produire des performances beaucoup plus rapides que les autres outils de *machine learning*, en particulier lorsque les données deviennent plus volumineuses. Elle prend en charge un ensemble d'algorithmes modernes *machine learning*, notamment le deep learning, les arbres d'ensemble tels que *XGBoost*, *RandomForest*, etc. « H2o » peut être utilisé via différentes interfaces, dont R et Python, afin que les utilisateurs de ces langages puissent facilement accéder à la puissance de H2O. Nous pouvons accéder aux fonctionnalités de « H2O » grâce au mécanisme des API (*application programming interface*). Ainsi, les fonctions du package *h2o* de R permettent faire des appels API depuis notre session R directement.
Nous allons maintenant présenter théoriquement les méthodes utilisées dans ce mémoire, en commençant par le réseau de neurones.

3.7.5.2 Réseau de neurone

Les réseaux de neurones se présentent comme des ensembles d'unités (aussi appelées neurones formels ou nœuds) connectés entre elles par des synapses. Chaque unité étant activée en recevant une donnée et en renvoyant une autre. Ces ensembles d'unités sont organisés en niveaux appelés couches, et chaque couche envoie ses données à la couche suivante. Ils tirent leur mécanisme du fonctionnement du cerveau humain. En effet, le cerveau humain est organisé en neurones biologiques et en synapses. Le neurone biologique est composé d'un corps, d'un axone et de dendrites. Les synapses servent de liens entre deux neurones biologiques. Le tout forme un réseau interconnecté de neurones qui communiquent. L'information se transmet d'un neurone à l'autre par l'intermédiaire des synapses, qui ajustent leur comportement afin d'assurer le transfert.

Les réseaux de neurones peuvent être utilisés à des fins différentes. On a d'une part l'utilisation de ceux-ci en classification avec les réseaux de Kohonen et d'autre part, on les retrouve en méthodes de prédictions (perceptrons, réseaux à fonctions radiales de base). Les réseaux descriptifs sont dits à apprentissage non supervisé tandis que les réseaux prédictifs sont dits à apprentissage supervisé.

Dans ce mémoire et dans un but de tarification, nous utiliserons des réseaux prédictifs.

3.7.5.3 Théorie

L'unité de base du modèle (illustré à la figure 3.23) est le neurone formel, un modèle inspiré du neurone l'être humain. Chez ce dernier, les signaux de sortie des neurones voyagent à des intensités variables le long des jonctions synaptiques et sont ensuite agrégés comme entrée pour l'activation d'un neurone connecté.

Dans le modèle, la valeur d'entrée du neurone formel est égale à la combinaison linéaire $\alpha = \sum_{i=1}^{n} w_i x_i + b$, où :

- les $(w_i)_{[\![i\in 1,n]\!]}$ représentent les poids de connexion associés aux entrées $(x_i)_{[\![i\in 1,n]\!]}$,
- b représente une valeur d'entrée seuil, appelée biais.

La sortie du neurone formelle vaut $f(\alpha)$. Cette fonction f représente la fonction d'activation non linéaire utilisée dans l'ensemble du réseau.

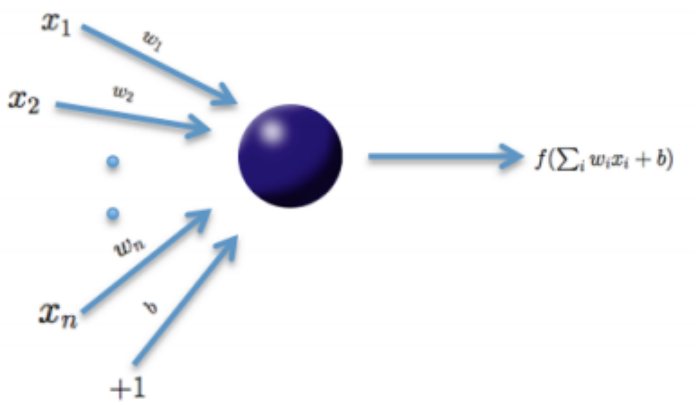


FIGURE 3.23 – Neurone formel

Nous avons donc expliqué le fonctionnement d'un neurone formel, qui est l'unité de calcul du réseau de neurones. Un réseau de neurones est une combinaison horizontale de plusieurs couches verticales de neurones formelles et partageant les mêmes entrées, comme l'illustre la figure 3.24 :

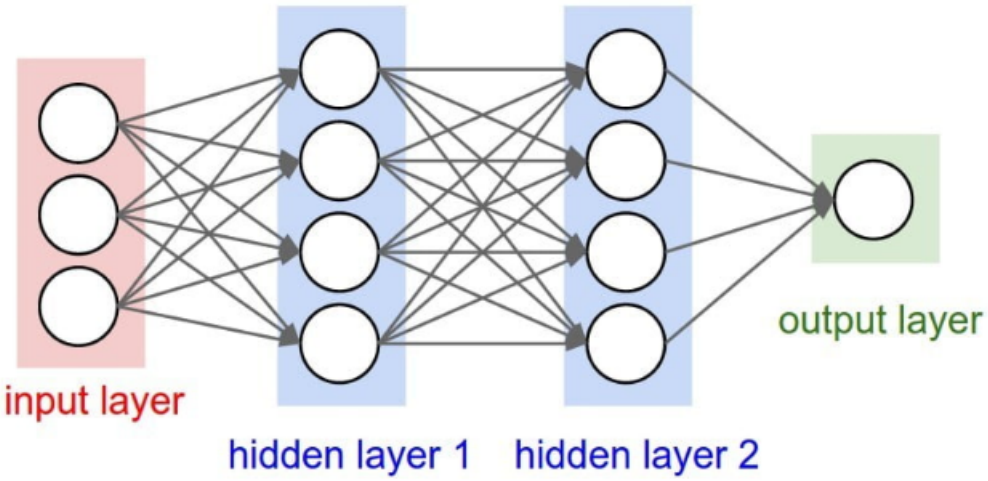


FIGURE 3.24 – Réseau de neurones

Les réseaux neuronaux multicouches sont donc constitués de plusieurs couches d'unités neuronales interconnectées.

La première couche (la plus à gauche sur la figure), aussi appelée couche d'entrée, reçoit en entrée l'ensemble des variables explicatives. Ensuite, ces entrées sont propagées à travers plusieurs couches successives, appelées couches cachées. Elles portent un tel nom car elles n'ont aucun contact avec l'extérieur (elles ne correspondent à aucune entrée ni sortie). Un unique et dernier neurone, connecté à la dernière couche (la plus à droite), retourne la sortie du modèle.

Les poids reliant les neurones et les biais avec d'autres neurones déterminent pleinement la sortie de l'ensemble du réseau, et ces poids sont adaptés pour minimiser l'erreur sur les données d'apprentissage. Plus précisément, pour chaque couche, l'objectif est de

minimiser une fonction de perte

$$L(W, B)$$

Ici W représente la suite $(W_i)_{i\in[\![1\,:N-1]\!]}$ où W_i désigne la matrice de poids reliant les couches i et $i+1$ pour un réseau de N couches ; de même, B représente la suite $(b_i)_{i\in[\![1\,:N-1]\!]}$ où b_i désigne la colonne vecteur de biais pour la couche $i+1$.

Précédemment, nous avons introduit la fonction d'activation non linéaire f. Notons ici que x_i et w_i désignent respectivement les valeurs d'entrée du neurone et leurs poids ; α désigne la combinaison linéaire $\alpha = \sum_{i=1}^{n} w_i x_i + b$. Les possibilités de fonction d'activation pour le package H2O sont regroupées dans la table 3.12 :

Fonction	Formule	Domaine
tanh	$f(\alpha) = \frac{e^{\alpha}-e^{-\alpha}}{e^{\alpha}+e^{-\alpha}}$	f(.) $\in [-1, 1]$
Linéaire rectifié	$f(\alpha) = \max(0, \alpha)$	f(.) $\in \mathbb{R}$
Maxout	$f(\alpha) = \max(w_i x_i + b)$, redim si f(.)$> 1$	f(.) $\in]-\inf, 1]$

TABLE 3.12 – Fonctions d'activation du package H2o de R

Il est difficile de déterminer la « meilleure » fonction d'activation à utiliser ; chacune d'entre elles peut être plus performante que les autres dans des scénarios distincts, mais les modèles de recherche par grille peuvent aider à comparer les fonctions d'activation et autres paramètres. La fonction d'activation par défaut est la « linéaire modifiée ».
Les différentes possibilités pour la fonction de perte L(W, B) sont résumées dans la table 3.13. La valeur par défaut du système applique la règle d'utilisation typique du tableau selon que l'on procède à une régression ou à une classification. Nous utilisons la notation y_j^{pred} pour décrire le résultat prédit par le modèle, et y_j pour le résultat réel :

Fonction	Formule
Erreur quadratique	L(W,B)$=\frac{1}{2}\sum_{i=1}^{p}(y_i - y_i^{pred})^2$
Cross entropy	L(W,B)$=-\sum_{i=1}^{p}\left(\ln(y_i)y_i^{pred} + \ln(1-y_i)(1-y_i^{pred})\right)$

TABLE 3.13 – Fonctions de perte du package H2o de R

Typique, la fonction de perte Erreur quadratique est utilisée pour la régression, tandis que la fonction Cross Entropy est utilisée généralement pour la classification. Ainsi, dans notre cas, nous utiliserons l'erreur quadratique car nous souhaitons effectuer une régression. L'objectif est donc de trouver les W et B qui minimisent cette fonction de perte. Les fonctions de pertes utilisées peuvent toutes s'écrire sous la forme $L(W, B) = \sum_{i=1}^{p} L_i(W, B)$. Par exemple, dans le cas de l'erreur quadratique, $L_i(W, B) = \frac{1}{2}(y_i - y_i^{pred})^2$. Cette notation (L_i) sera réutilisée ci-dessous.
La procédure visant à minimiser la fonction de perte $L(W, B)$ dans le package « H2O » est une version parallélisée de la descente de gradient stochastique. La descente de gradient stochastique standard a été résumée en début de page suivante (voir liste 3.1). Nous expliquons ensuite les points positif d'une descente de gradient stochastique par rapport à une descente de gradient classique. La constante α indique le taux d'apprentissage, qui contrôle la taille des pas pendant la descente du gradient.

1. Initialiser W et B.
2. On itère, tant que le critère de convergence n'est pas atteint :
   (a) On tire aléatoirement une observation dans l'échantillon d'entrainement (d'où le terme « stochastique »),
   (b) On met à jour les valeurs de w_{jk} et b_{jk} selon l'observation tiré précédemment :
   — $w_{jk} = w_{jk} - \alpha \frac{\partial L_i(W,B|j)}{\partial w_{jk}}$
   — $b_{jk} = b_{jk} - \alpha \frac{\partial L_i(W,B|j)}{\partial b_{yk}}$

Liste 3.1 – Algorithme de descente du gradient stochastique

Ici, le terme « stochastique » vient du fait que le gradient est basé sur une seule observation de l'échantillon d'apprentissage. En raison de sa nature stochastique, le chemin vers le minimum recherché n'est pas « direct » comme dans la Descente de gradient classique, mais peut aller en « zig-zag » si nous visualisons l'évolution du coût en 2 dimensions. Cependant, il a été démontré que la Descente de Gradient Stochastique converge presque sûrement vers le minimum de coût global si la fonction de coût est convexe (ou pseudo-convexe).

Ainsi, si le nombre d'observations dans l'échantillon d'entrainement est important, voire très important, l'utilisation de la descente en gradient classique peut prendre trop de temps car à chaque itération, lorsque l'on met à jour les valeurs des paramètres, on parcourt l'ensemble de l'échantillon. Ainsi, l'utilisation de la méthode de descente de gradient stochastique sera bien plus rapide. Notons que dans l'étape 2.(a), il est possible de sélectionner aléatoirement plusieurs observations dans l'échantillon d'apprentissage, plutôt qu'une seule.

La descente du gradient stochastique est rapide et efficace sur le plan de la mémoire, mais elle n'est pas facilement parallélisable sans devenir lente. Pour paralléliser cet algorithme, la méthode « HOGWILD ! » est utilisée. Cette méthode est très technique et ne sera pas développée dans ce mémoire. Le lecteur intéressé pourra se référer à l'article [18] de F.Niu.

Résultats du réseau de neurone

On obtient, après calibration du modèle, les indicateurs suivants :

	Train	Test
Erreur globale	-0.10%	0.26%
RMAE	0.206	0.210
RMSE	0.364	0.405
Gini	71.57%	71.79%

TABLE 3.14 – Indicateurs pour le modèle de réseau de neurones

Ce modèle améliore assez sensiblement les modèles GLMs précédemment calibrés. Cependant, on retrouve des points négatifs à utiliser un réseau de neurones pour la tarifi-

cation, notamment à cause du côté boite noire de la méthode ainsi que la complexité d'implémentation. Nous allons dans la partie suivante étudier et implémenter un dernier modèle, également assez populaire, le XGBoost.

3.7.6 eXtreme Gradient Boosting

L'algorithme XGBoost, connu par sa capacité prédictive et sa vitesse d'exécution, est un algorithme d'apprentissage supervisé dont le principe et de combiner les résultats d'un ensemble de modèles plus simple et plus faibles afin de fournir une meilleur prédiction. On parle d'ailleurs de méthode d'agrégation de modèles. L'idée est donc simple : au lieu d'utiliser un seul modèle, l'algorithme va en utiliser plusieurs qui serons ensuite combinés pour obtenir un seul résultat.

L'idée principale de l'algorithme est de construire séquentiellement des sous-arbres à partir d'un arbre original de telle sorte que chaque arbre suivant réduise les erreurs du précédent. De cette façon, les nouveaux sous-arbres mettront à jour les résidus précédents afin de réduire l'erreur de la fonction de coût. Afin de mesurer la robustesse de l'algorithme, une fonction appelée fonction objectif est utilisée. Cette fonction est la somme de deux fonctions : la fonction de perte et la fonction de régularisation. La fonction de perte mesure la capacité prédictive du modèle tandis que la fonction de régularisation contrôle la complexité du modèle et est utilisée afin d'éviter le problème de sur-apprentissage.

Si le modèle finale contient K arbres, alors on définit la valeur prédite de l'entrée i, $\hat{y}_i$, comme suit :

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i),\ f_k \in F$$

où f_k est un arbre indépendant de F, l'espace des arbres de régression, et $f_k(x_i)$ le score de l'arbre k pour les entrées x_i.
La fonction objectif du XGBoost, notée L et qui est la fonction à optimiser, est donnée comme suit :

$$\mathcal{L} = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

L'apprentissage des arbres du modèle se fait donc par minimisation de la fonction objectif $\mathcal{L}$. La fonction de perte $l(y_i, \hat{y}_i)$ évalue la différence entre la prédiction $\hat{y}_i$ et la valeur réelle y_i et est souvent définie comme étant la fonction de perte quadratique. Le terme utilisé ici pour éviter le problème du surajustement en pénalisant la complexité du modèle est :

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda ||\omega||^2$$

où γ et λ sont des paramètres de régularisation, T et ω sont respectivement le nombre de feuilles et les scores sur chaque feuille.

Pour l'apprentissage des arbres, une stratégie additive est adoptée. Nous écrivons ci-dessous la valeur de la prédiction à l'étape t sous la forme $\hat{y}_i^t$.

$$\begin{aligned}
\hat{y}_i^{(0)} &= 0 \\
\hat{y}_i^{(1)} &= f_1(x_i) = f_1(x_i) + \hat{y}_i^{(0)} \\
\hat{y}_i^{(2)} &= f_1(x_i) + f_2(x_i) = f_2(x_i) + \hat{y}_i^{(1)} \\
&\dots \\
\hat{y}_i^{(t)} &= \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i)
\end{aligned}$$

À chaque étape t, la fonction objectif peut s'écrire de la manière suivante :

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t)}\right) + \sum_{i=1}^{t} \Omega(f_i)$$

$$= \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t) + \text{ constant}$$

Ainsi, à chaque étape, on se ramène à un problème de minimisation de la fonction $\mathcal{L}^{(t)}$.

L'algorithme XGBoost a été implémenté sous R grâce au package *xgboost*. Les paramètres à fixer (liste non exhaustive) pour faire tourner la fonction sont les suivants :

- nrounds : contrôle le nombre maximum d'itérations,
- η : Il contrôle le taux d'apprentissage, c'est-à-dire la vitesse à laquelle notre modèle apprend des modèles dans les données. Après chaque cycle, il réduit les poids des caractéristiques pour atteindre le meilleur optimum,
- γ : Il contrôle la régularisation (ou empêche le surajustement). La valeur optimale du gamma dépend de l'ensemble des données et d'autres valeurs de paramètres,
- max depth : contrôle la profondeur maximale des arbres. Plus ce paramètre est élevé, plus le risque de sur-apprentissage est grand,
- nombre minimum d'observations par feuille,
- λ : contrôle la régularisation sur les poids.

3.7.6.1 Choix des paramètres

Afin de fixer les paramètres du modèle, nous allons utiliser la recherche par quadrillage. L'idée est plutôt simple ; en effet, après avoir choisi la liste des possibilités pour chacun des hyperparamètres, nous faisons tourner un modèle pour chaque valeur de cette liste. Finalement, on ne conserve évidemment qu'un seul modèle, celui qui donne les résultats les plus satisfaisants (nous verrons dans un exemple simple ce que l'on appelle satisfaisant).

C'est une technique intéressante et performante mais avec l'inconvénient majeur du temps de traitement car toutes les valeurs de la liste prédéfinie doivent être testées, ce qui représente un grand nombre de modèles.

Un exemple simple de détermination d'un seul paramètre a été faite dans ce mémoire pour bien comprendre le principe de la méthode. Pour la détermination de plusieurs paramètres paramètres, la technique est la même mais il faut croiser toutes les possibilités. Dans l'exemple, nous allons déterminer le paramètre η à utiliser dans le modèle. La liste des η prédéfinis sont 0.05, 0.1, 0.2, 0.5, 1. Nous fixons tous les autres paramètres pour que l'exemple soit simple. Nous nous intéressons alors à 2 résultats :

— La **vitesse de convergence**, en représentant le RMSE selon le nombre d'itérations pour voir à partir de quelle itération la valeur du RMSE se stabilise (figure 3.25) :

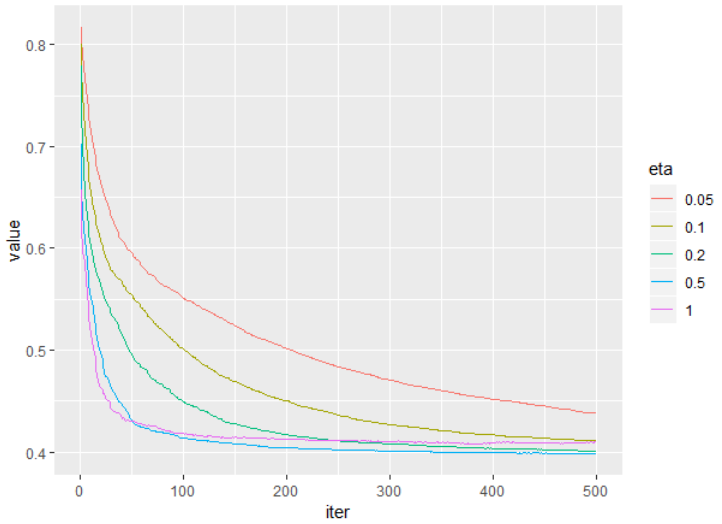


FIGURE 3.25 – Tableau de convergence

Ce graphique est en accord avec le fait que les grandes valeurs d'êta entrainent une convergence plus rapide du modèle,

— La table 3.15 regroupant les **valeurs du RMSE** pour les différentes valeurs de η :

η	0.05	0.1	0.2	0.5	1
RMSE	0.4375	0.4125	0.403	0.399	0.4125

TABLE 3.15 – RMSE selon les valeurs de η

L'objectif est d'avoir le meilleur compromis, c'est à dire un RMSE faible et une vitesse de convergence élevée. Dans cet exemple, les valeurs de RMSE obtenus après 500 itérations sont très proches. Nous choisissons η=0.5 car pour cette valeur, on a la $2^{\text{ème}}$ meilleure vitesse de convergence et le meilleur RMSE.
L'exemple développé est une simplification d'une recherche par quadrillage. En effet, nous n'avons testé qu'une seule variable, alors qu'il faut normalement en combiner plusieurs, ce qui rend les temps de calcul bien plus long et les résultats difficielemnt visuables sur un graphique.

Importance des variables

La fonction XGBoost nous donne la possibilité de visualiser les variables qu'elle a le

plus utilisé pour construire sa règle de décision finale. Un avantage de l'utilisation de l'algorithme XGBoost est qu'après la construction des arbres, il est relativement simple de récupérer des scores d'importance pour chaque variable. En général, l'importance fournit un score qui indique l'utilité de chaque caractéristique dans la construction des arbres de décision au sein du modèle. Plus un attribut est utilisé pour prendre des décisions clés avec les arbres de décision, plus son importance relative est élevée. Cette importance est calculée explicitement pour chaque attribut dans l'ensemble de données, ce qui permet de classer et de comparer les attributs entre eux. Dans notre cas, le graphique des 10 variables les plus importantes est représenté sur la figure 3.26 :

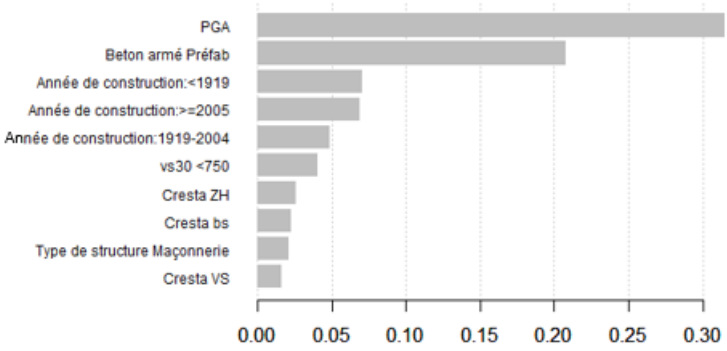


FIGURE 3.26 – Importance des variables par XGBoost

Sans surprise, le PGA est la variable étant la plus utile dans la construction des arbres de décision. En effet, il s'agit de la variable permettant de construire la carte de risque.

Résultats du XGBoost

Nous faisons ensuite tourner le XGBoost, pour modéliser les taux de destruction.
La table récapitulative des indicateurs principaux du modèle sont regroupés ci-dessous :

	Train	Test
Erreur globale	3.198%	3.139%
RMAE	0.2702	0.2727
RMSE	0.399	0.479
Gini	73%	67%

TABLE 3.16 – Indicateurs pour le modèle 4

Les différents indicateurs liés aux études de l'erreur ne sont pas meilleurs que les modèles précédemment implémentés. Seul l'indice de Gini pour la base d'apprentissage est amélioré. Parmi les 2 méthodes de machine learning testés, on préférera tout de même le réseau de neurones il donne de meilleurs résultats sur la base de test.

3.7.7 Comparaison des résultats et confrontation des modèles

On regroupe, dans la figure 3.27, la totalité des indicateurs pour chaque modèle GLM.

	MODELES GLM							
Indicateurs	Modèle 1 coût-fréquence		Modèle 2 Unique Tweedie		Modèle 3 Tweedie/Gamma par zone de risque		Modèle 4 Tweedie avec plus d'informations géographiques	
Set	Train	Test	Train	Test	Train	Test	Train	Test
Erreur globale	-1,76%	-2,13%	1,52%	-2,25%	1,49%	1,84%	1%	1,55%
RMSE	0,603	0,649	0,380	0,421	0,366	0,413	0,379	0,418
RMAE	0,314	0,322	0,224	0,229	0,215	0,222	0,223	0,227
Gini	60,74%	60,44%	70,54%	70,80%	70,88%	71,07%	70,47%	70,57%

FIGURE 3.27 – Indicateurs pour chaque modèle GLM

On regroupe ensuite dans la table 3.28 les indicateurs pour les modèles de machine learning :

	MODELES Machine learning			
Indicateurs	Modèle 5 Réseau de neurones		Modèle 6 Xgboost	
Set	Train	Test	Train	Test
Erreur globale	-0,10%	0,26%	3,20%	3,14%
RMSE	0,364	0,405	0,270	0,271
RMAE	0,206	0,210	0,480	0,479
Gini	71,57%	71,79%	73,00%	67,00%

FIGURE 3.28 – Indicateurs pour les modèles de machine learning

On peut faire les remarques suivantes :

Modèle 1

En divisant la modélisation en fréquence et coût, on voit que le plus difficile est de définir précisément la fréquence des séismes. En effet, le GLM Poisson pour le modèle de fréquence donne de mauvais résultats. La valeur élevée de l'erreur moyenne du modèle fréquence x coût provient donc sûrement du modèle de fréquence. En revanche, le modèle de sévérité est bon.

Modèle2

La modélisation du taux de destruction combinant fréquence et gravité dans un unique modèle permet d'améliorer les indicateurs statistiques par rapport au modèle 1.

Modèle3

Créer des modèles pour les zones peu risquées et les zones risquées permet d'améliorer très légèrement les indicateurs, mais pas de manière significative. On pourra donc, pour

des raisons de simplicité, préférer le modèle 2 qui est un unique modèle avec des indicateurs très proches du modèle 3.

Modèle 4

Dans ce modèle, nous avons plus d'informations géographiques mais une nouvelle fois, les indicateurs ne sont pas significativement améliorés pour le préférer au modèle 2. Ce n'est pas étonnant car les variables ajoutées sont très corrélées avec une variable déja présente dans les autres modèles, le PGA pour une période de retour de 475 ans.

Modèle 5

Utiliser un modèle plus complexe et innovant de réseau de neurones permet d'améliorer les indicateurs des GLMs. Cependant, il nécessite une lourde implémentation au niveau IT, et il ne permet pas d'expliquer facilement le taux de destruction obtenu comme pour c'est le cas d'un GLM multiplicatif à cause de son côté boite noire.

Modèle 6

Ce modèle donne de bons indicateurs, mais on préférera tout de même le modèle 5 parmi les modèles de machine learning calibrés car le modèle 5 a de meilleurs résultats sur la base de test.

Finalement, on peut conclure qu'il semble préférable de garder 2 modèles : le modèle 2 et le modèle 5. Le modèle 2 permettra d'expliquer facilement le tarif final car on peut facilement en sortir une grille de coefficients à appliquer à la prime selon les modalités de variables explicatives. Des coefficients de GLMs sont très facilement implémentable dans n'importe quel système IT. Le modèle 5 améliorera quelque peu la tarification, avec le point négatif du côté boite noire des réseaux de neurones. Comme il donne de bons résultats, il peut par exemple au moins être utilisé comme modèle de référence.

On représente finalement à la figure 3.29 la courbe permettant de calculer l'indice de Gini pour :

— Le modèle GLM,

— Le « meilleur » modèle de machine learning, c'est à dire le réseau de neurones,

— Le taux de destruction observé provenant du modèle CAT.

Les modèles développés permettent d'obtenir une perte brute attendue sur un portefeuille d'assurés, sans prise en compte des franchises et limites. La partie suivante porte sur l'intégration de ces conditions d'assurances.

3.8 Intégration des conditions d'assurance

L'objectif de cette partie est de modéliser les courbes de rabais concernant la franchise et la limite qui sont les deux conditions financières à appliquer à la prime brute pour

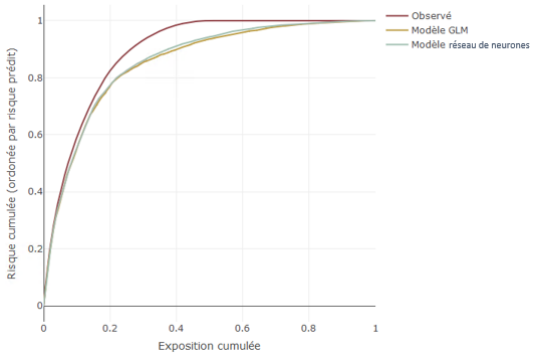


FIGURE 3.29 – Courbes permettant le calcul des indices de Gini

obtenir la prime nette. Cela nous permettra de connaitre l'influence de n'importe quelle valeur de limite ou franchise sur la perte totale.

3.8.1 Quelques notions théoriques

3.8.1.1 La prise en compte des conditions d'assurance

Notons X la variable aléatoire représentant la perte ground up, c'est à dire la perte sans prise en compte des conditions financières. Il s'agit bien entendu d'une variable aléatoire positive $\mathbb{E}(X)$ représente donc les pertes attendues.

Influence d'une limite

Supposons que les pertes sont limitées par une certaine limite L dans un contrat d'assurance. Alors, la perte attendue en prenant en compte cette limite vaut $\mathbb{E}(\min(X, L))$. Ce minimum peut s'exprimer de manière simple grâce aux quelques calculs ci-dessous :

$$\mathbb{E}[\min(X, L)] = \int_0^L x f(x)dx + \int_L^\infty L f(x)dx$$

Nous allons maintenant expliciter les 2 termes de cette somme. En rappelant que $F(t) = \int_0^t f(x)dx$ et $F(\infty) = 1$, on a alors :

$$\int_L^\infty L f(x)dx = L\left(\int_0^\infty f(x)dx - \int_0^L f(x)dx\right) = L - LF(L)$$

De plus, une intégration par partie simple (en utilisant le fait que $f(x) = F'(x)$) nous permet d'avoir :

$$\int_0^L x f(x)dx = LF(L) - \int_0^L F(x)dx$$

Par conséquent, le minimum peut s'écrire de la manière simple suivante :

$$\mathbb{E}[\min(X, L)] = L - \int_0^L F(x)dx = \int_0^L 1 - F(x)dx = \int_0^L S(x)dx$$

où S représente la fonction de survie. Un exemple simple (figure 3.30) est représenté ci-dessous, avec une limite de 100 et l'aire sous la courbe de S(x) entre 0 et L en bleu.

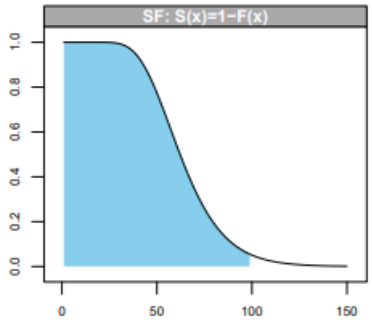


FIGURE 3.30 – Perte attendue avec prise en compte de la limite

Prise en compte d'une franchise

On suppose ici qu'en plus de la limite L, le contrat a une franchise d. La perte attendue après prise et compte de la limite et de la franchise est :

$$\mathbb{E}[\min(X, L)] - \mathbb{E}[\min(X, d)]$$

Un exemple simple est représenté à la figure 3.31, avec une limite de 100 et une franchise de 80.

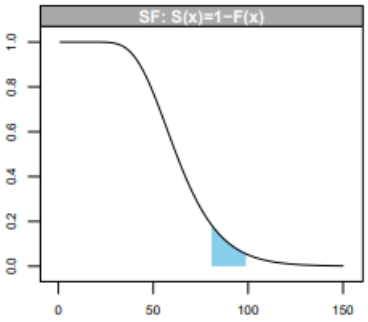


FIGURE 3.31 – Perte attendue avec prise en compte de la limite et la franchise

On réécrit la formule précédente de la manière suivante, en factorisant par $\mathbb{E}[X]$:

$$\mathbb{E}[\min(X, L)] - \mathbb{E}[\min(X, d)] = \mathbb{E}[X] \left(\frac{\mathbb{E}[\min(X, L)]}{\mathbb{E}[X]} - \frac{\mathbb{E}[\min(X, d)]}{\mathbb{E}[X]} \right) = \mathbb{E}[X] \left(G_L(X) - G_d(X) \right)$$

Nous préférons cette écriture car elle permet d'introduire la fonction G, appelée courbe d'exposition. Dans le cas d'une franchise par exemple, la courbe d'exposition représente le pourcentage de perte attendue retenu par l'assuré.

La figure 3.32 représente un exemple de fonction G, avec des valeurs prises par défaut dans le cas d'une franchise.

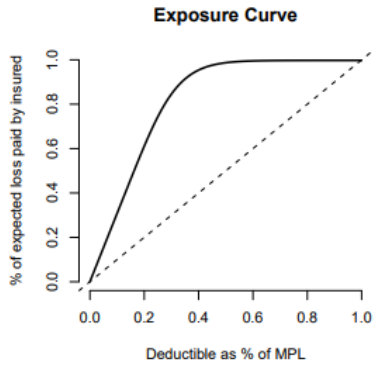


FIGURE 3.32 – Courbe d'exposition

Dans la courbe ci-dessous, la franchise est donnée en pourcentage du sinistre maximum probable (MPL signifie Maximum probable loss). Elle est également souvent donnée en pourcentage de la valeur assurée totale. La pente de la courbe est liée à la sévérité de la distribution des pertes. Plus la courbe est proche de la diagonale, plus la proportion de pertes totales est élevée (une perte totale est une perte valant la totalité de la valeur assurée). Si toutes les pertes étaient des pertes totales, la courbe d'exposition serait identique à la diagonale. Des périls et des expositions différents auront des courbes d'exposition différentes.

Nous allons étudier dans la partie suivante une distribution dont la fonction d'exposition est régulièrement utilisée pour modéliser ces courbes de rabais : il s'agit de la loi MBBEFD.

3.8.1.2 La distribution MBBEFD

Dans cette partie, nous allons donner quelques généralités sur la courbe d'exposition G_x de la loi MBBEFD ainsi que sur la loi en elle-même. La courbe d'exposition des lois MBBEFD (issue des travaux de Maxwell-Boltzmann, Bose-Einstein, Fermi-Dirac) est très utilisée pour modéliser les courbes de rabais. Cette courbe d'exposition dépend de 2 paramètres b et g et se définit de la manière suivante :

$$G_{b,g}(x) = \begin{cases} x & \text{si } g = 1 \text{ et } b = 0 \\ \frac{\ln(1+(g-1)x)}{\ln(g)} & \text{si } b = 1 \text{ et } g > 1 \\ \frac{1-b^x}{1-b} & \text{si } bg = 1 \text{ et } g > 1 \\ \frac{\ln(((g-1)b+(1-gb)b^x)/(1-b))}{\ln(gb)} & \text{si } b > 0, b \neq 1, bg \neq 1 \text{ et } g > 1 \end{cases}$$

Dans la suite, nous nous placerons dans le cas le plus général correspondant à la quatrième ligne de l'accolade.

Il existe un lien entre la fonction de répartition et la fonction d'exposition. En effet, on a le résultat ci-dessous qui est un cas général et non spécifique aux lois que nous étudions :

$$F(x) = \begin{cases} 1 \text{ si x} = 1 \\ 1 - \frac{G'(x)}{G'(0)} \text{ pour } 0 \leq x < 1 \end{cases}$$

En utilisant cette égalité et le fait que $F'(x) = f(x)$, on peut donner une formule de la densité d'une loi MBBEFD(b,g) :

$$f(x) = \frac{(b-1)(g-1)\ln(b)b^{1-x}}{((g-1)b^{1-x} + (1-gb))^2}$$

3.8.2 Modélisation des conditions financières

Nous allons dans cette partie décrire la méthodologie mise en place afin de modéliser les conditions financières, c'est à dire la limite et franchise.
La base de données utilisée est la *Year Event Site Loss Table*, c'est à dire une table représentant les pertes par Année x évènement x site.

Lorsque l'on a la main sur le modèle, ce qui est notre cas avec le modèle interne catastrophe naturelle d'AXA, nous avons la liberté de faire tourner ce modèle à notre guise, avec la possibilité de faire varier les limites et franchises comme on le souhaite, et de regarder l'influence sur la perte. Ainsi, on a la possibilité de mettre en place la méthodologie suivante :

1. Calcul de l'impact sur la perte d'une franchise et limite de x% de la valeur assurée (en faisant varier x).
2. Pour chaque site, on calcule l'effet minimum de ces limites et franchises pour chaque x découlant de l'étape 1. Cela nous permet d'obtenir, pour chaque x, une courbe d'exposition empirique (on a pris ici l'effet minimum, c'est à dire le quantile à 1%, car on souhaite être prudent quant à l'influence des limites/franchises sur la prime finale. Il est cependant tout à fait possible de prendre d'autres valeurs, comme par exemple l'effet moyen ou médian si l'on souhaite une influence plus importante des limites et franchises sur la prime).
3. On fait une interpolation des points empiriques avec une courbe d'exposition MBBEFD à 2 paramètres en cherchant à minimiser l'erreur quadratique moyenne.

La fonction *ecmbbefd* de R nous permet d'obtenir la courbe d'exposition d'une loi MBBEFD. Afin de minimiser l'erreur quadratique moyenne, nous utilisons l'algorithme BFGS présenté précédemment. Il s'agit d'un algorithme de type quasi-Newton. À partir d'une valeur initiale x_0 et une matrice Hessienne approchée B_0 (qui est souvent la matrice identité), on décrit ci dessous les étapes de cet algorithme :
tant que gradient $||\nabla f(x_k)|| < \epsilon$:

1. On trouve p_k en résolvant $B_k p_k = -\nabla f_{x_k}$ où B_k est une approximation de la matrice Hessienne à l'étape k et $\nabla f(x_k)$ est le gradient de f évalué en x_k ;
2. On met à jour la valeur suivante : $x_{k+1} = x_k + \alpha_k p_k$ où α_k est un pas ;
3. On calcule $y_k = \nabla f_{x_{k+1}} - \nabla f_{x_k}$;
4. $\mathrm{B}_{k+1} = \mathrm{B}_k + \frac{(\mathbf{y}_k \mathbf{y}_k^\top)}{(\mathbf{y}_k^\top \mathbf{s}_k)} - \frac{(\mathrm{B}_k \mathbf{s}_k \mathbf{s}_k^\top \mathrm{B}_k)}{(\mathbf{s}_k^\top \mathrm{B}_k \mathbf{s}_k)}$;
5. k=k+1.

Résultats

Les paramètres trouvés par minimisation de l'erreur quadratique sont les suivants :

$$g = 2.134 \text{ et } b = 0.002$$

On représente à la figure 3.33 en bleu la fonction G avec ces valeurs particulières de g et b, ainsi que les points à interpoler.

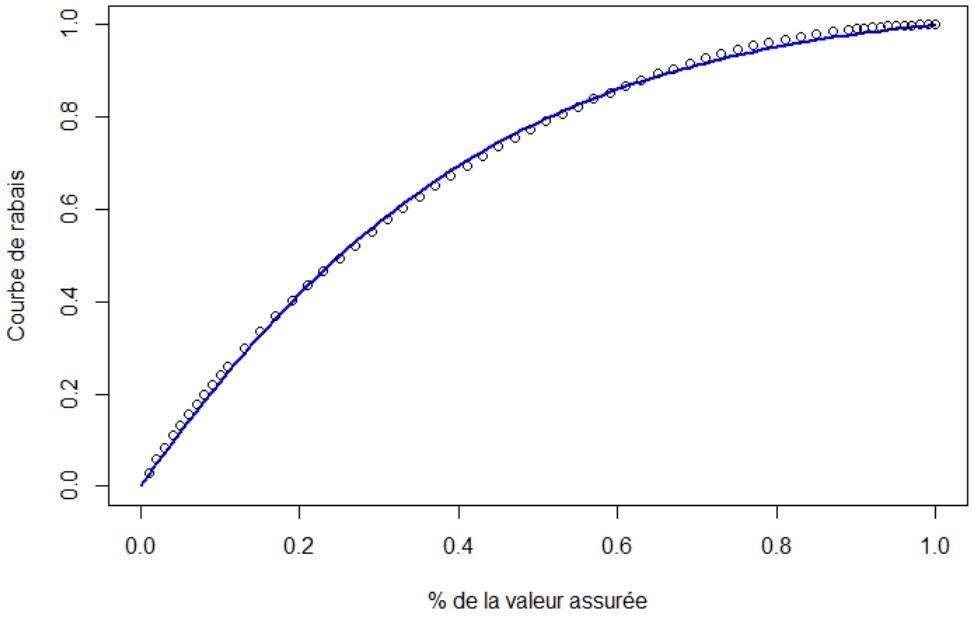


FIGURE 3.33 – Interolation d'une courbe d'exposition d'une loi MBBEFD (cas d'une franchise)

L'erreur quadratique résultant de cet exemple vaut 9.25×10^{-5}. On retrouve ici une courbe de la classe MBBEFD qui approche particulièrement bien la courbe d'exposition empirique. Pour de tels paramètres, on regroupe dans la table 3.17 quelques valeurs de G :

x	G(x)
1%	2%
2%	5%
3%	7%
4%	9%
5%	12%
10%	21%
15%	36%
20%	40%
25%	50%
30%	57%

TABLE 3.17 – Quelques valeurs de la courbe de rabais obtenue (exemple d'une franchise)

Une fois la méthode appliquée pour la franchise, on procède de la même manière pour obtenir l'influence de la limite sur la prime.

La méthode décrite précédemment nous permet donc d'appliquer des conditions d'assurance aux pertes brutes attendues obtenues par les modèles calibrés dans la partie 3.7. Ainsi, la tarification est complète et peut-être intégrée dans les processus de souscription. La figure 3.34 résume alors le cheminement pour obtenir une tarification complète :

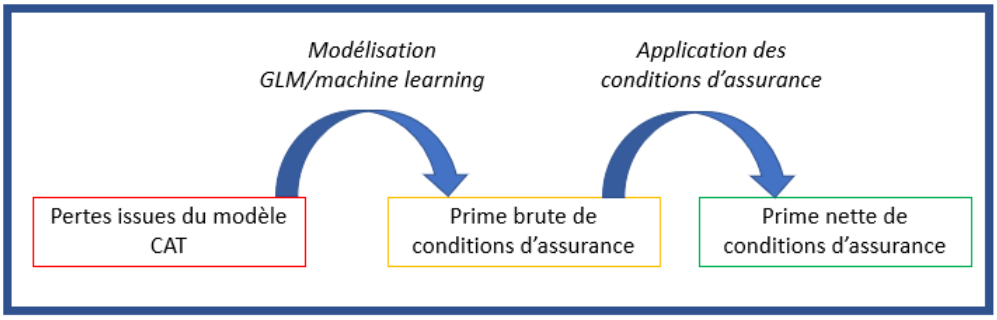


FIGURE 3.34 – Etapes de la tarification

Il convient de noter que cette méthode d'application de la courbe de rabais fonctionne bien pour les polices mono-site, où la franchise et la limite peuvent être facilement interprétées en proportion du montant total assuré. D'ailleurs, l'interpolation a été effectuée sur des résultats obtenus site par site.
Ainsi, pour les polices multisites et en cas de franchise/limite par police, cette méthodologie doit être adaptée et l'utilisabilité du modèle est compromise. Nous ne prendrons pas en compte ces cas particuliers dans ce mémoire, il serait cependant intéressant de dévélopper des méthodes afin d'étendre l'étude aux polices multi-sites.

Conclusion

L'objectif de ce mémoire a été d'exploiter le modèle catastrophe naturelle construit par les modélisateurs CAT d'AXA pour le péril séisme.

Après une introduction aux catastrophes naturelles dans le chapitre 1, nous avons détaillé dans le chapitre 2 les différents modules constitutifs d'un modèle CAT. Les pertes site à site issues de ce modèle servent de données d'entrée pour construire un modèle de tarification présenté dans le troisième et dernier chapitre.

Grâce au modèle CAT et à sa liberté d'utilisation, nous avons pu :

— Ajouter des bâtiments assurés fictifs (en plus des assurés réels) couvrant la totalité des combinaisons de variables explicatives dans un but de faire tourner le modèle CAT sur une base d'assurés exhaustive au niveau du risque.

— Tester de nombreuses valeurs de franchises et limites différentes afin de dégager une distribution de l'influence des ces valeurs sur la totalité des sites.

Pour reproduire le plus fidèlement possible les pertes site à site obtenues par le modèle CAT, nous avons calibré plusieurs modèles. Après avoir effectué un GLM classique coût/fréquence, nous avons modélisé directement les taux de destruction par des GLMs de la famille Tweedie. Ensuite, nous avons étudié des méthodes innovantes de machine learning : l'algorithme XGBoost et le réseau de neurones. Notons que la variable la plus discriminante des modèles, le PGA (*Peak Ground Acceleration*, l'accélération maximale du sol), a au préalable été découpé en classes par un arbre CART. Pour la modélisation, nous avons affecté des poids aux bâtiments assurés selon qu'ils proviennent de la base fictive ou de la base d'assurés réels, avec des poids plus élevés pour les assurés réels. Après calibration de tous les modèles cités, nous avons choisi de sélectionner 2 modèles :

— Un modèle GLM Tweedie unique expliquant directement les taux de destruction ;

— Le modèle de réseau de neurones.

Le modèle de réseau de neurones est celui qui nous permet d'obtenir les meilleurs indicateurs testés. Cependant, le côté boite noire de ce type de modèle rend la tarification difficilement utilisable pour la souscription contrairement aux GLMs qui sont des méthodes transparentes permettant de sortir directement les coefficients à appliquer à la prime pour chaque combinaison de variables explicatives, ce qui est facilement implémentable dans un système IT. Le modèle de réseau de neurones pourra cependant au moins servir de modèle de référence.

Enfin, le modèle CAT a également pu être exploité pour la modélisation de l'impact des conditions financières sur la perte brute. En effet le fait d'avoir pu tester diverses valeurs de limites et franchises nous a permis de déterminer l'effet de celles-ci sur la perte brute, pour ensuite effectuer une interpolation avec une courbe d'exposition d'une loi MBBEFD. La méthode développée est efficace pour les polices mono-sites mais nécessite d'être adaptée pour les polices multi-sites où les franchises et limites sont définies par police. L'étude pourrait donc être étendue à ces cas particuliers.

L'étude réalisée permet donc de fournir aux équipes de modélisation d'AXA un ensemble de codes potentiellement réutilisables pour une modélisation du risque sismique pour d'autres pays. Bien entendu, les conclusions dépendront du pays étudié mais la démarche est réplicable.

Bibliographie

[1] URL : https://www.swissre.com/institute/research/sigma-research/sigma-2020-02. *sigma 2/2020 : Natural catastrophes in times of economic accumulation and climate change* 2020.

[2] URL : https://www.emdat.be/.

[3] URL : https://www.geodiversite.net/.

[4] URL : http://www.seismo.ethz.ch/.

[5] Robin Genuer et Jean-Michel Poggi. *Arbres CART et Forêts aléatoires, Importance et sélection de variables* 201è.

[6] Maud Thomas. *Économétrie de l'assurance non-vie. Cours ISUP.* 2019.

[7] Hamza El Hassani. *Modélisation stochastique des innondations en France et applications en Réassurance.* 2017.

[8] Florient Aubry. *Exploitation actuarielle des modèles catastrophes naturelles pour éclairer et piloter dynamiquement la souscription de risques.* 2019.

[9] Aurélien Boiselet. *Cycle sismique et aléa sismique d'un réseau de failles actives : le cas du rift de Corinthe (Grèce).* 2014.

[10] Serpil Ünal et Salih Çelebioglu. *A Markov Chain Modelling Of The Earthquakes Occuring In Turkey.* 2014.

[11] Abdelhak Talbi, Kazuyoshi Nanjo, Kenji Satake, Jiancang Zhuang et Mohamed Hamdache. *Comparison of seismicity declustering methods using a probabilistic measure of clustering.* 2013.

[12] Julien Saunier. *Courbes d'exposition : Approximation par les distributions MBBEFD et Pareto via maximum de vraisemblance et intervalles de confiance. Analyse en fonction du capital assuré.* 2013.

[13] Christophe Dutang, Markus Gesmann et Giorgio Spedicato. *Exposure rating, destruction rate models and the mbbefd package.* 2019.

[14] Terry M. Therneau, Elizabeth J. Atkinson, Mayo Foundation. *An Introduction to Recursive Partitioning Using the RPART Routines.* 2019.

[15] Jiancang Zhuang, Yosihiko Ogata, and David Vere-Jones. *Stochastic Declustering of Space-Time Earthquake Occurrences.* 2019.

[16] A Deif and I El-Hussain. *Seismic moment rate and earthquake mean recurrence interval in the major tectonic boundaries around Oman.* 2012.

[17] Thomas van Stiphout, Jiancang Zhuang, David Marsan. *Models and Techniques for Analyzing Seismicity, Seismicity Declustering.* 2012.

[18] F.Niu. *HOGWILD ! : A Lock-Free Approach to Parallelizing Stochastic Gradient Descent.* 2011.