

Mémoire présenté le :

**pour l'obtention du Diplôme Universitaire d'actuariat de l'ISFA  
et l'admission à l'Institut des Actuaires**

Par : Léo Thiebault

Titre Analyse de sensibilité d'un modèle de risque sismique

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

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

*Membre présents du jury de l'Institut  
des Actuaires*

signature

*Entreprise : AXA GRM*

Nom :

Signature :

*Directeur de mémoire en entreprise :*

Nom : *Adrien Pothon*

Signature :

Invité :

Nom :

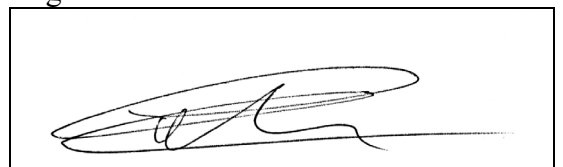
Signature :

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

Signature du responsable entreprise



Signature du candidat



# ANALYSE DE SENSIBILITÉ D'UN MODÈLE DE RISQUE SISMIQUE

Mémoire d'actuariat

I.S.F.A. Lyon

THIEBAULT Léo

Tuteur Professionnel :  
POTHON Adrien

Tuteur Académique :  
RIBEREAU Pierre

Mars 2023

---



*Pour ma mère.*

## Remerciements

Je tiens tout d'abord à remercier chaleureusement mon tuteur professionnel Adrien Pothon et mon tuteur pédagogique Pierre Ribereau pour leur pédagogie, leur patience et leurs conseils avisés.

Merci également à l'ensemble de l'équipe *R&D Modélisation Catastrophes Naturelles* d'AXA Group Risk Management : Madeleine-Sophie Deroche, Hugo Rakotoarimanga, Mathis Joffrain, Rémi Meynadier et Anyssa Diouf pour leur gentillesse et leur accueil. Enfin je remercie Quentin Ramadier, consultant Sia Partners au sein d'AXA GRM, pour sa disponibilité, sa bonne humeur et ses bons conseils.

Merci enfin aux chercheurs et professionnels m'ayant accordé de leurs temps pour discuter méthodes et résultats : Jérémy Rohmer, Ingénieur Recherche au Bureau de recherches géologiques et minières (BRGM), Bertrand Iooss, Ingénieur Recherche à EDF et Marouane Il Idrissi, Doctorant CIFRE à EDF.

# Sommaire

<b>Partie 1 : Assurance et Catastrophes naturelles</b>	<b>5</b>
<b>I Définir les catastrophes naturelles</b>	<b>6</b>
<b>II Le régime d'indemnisation des catastrophes naturelles</b>	<b>8</b>
II.1 Contexte et principes . . . . .	8
II.2 Réforme du régime . . . . .	10
<b>III Les enjeux pour l'assureur de la modélisation catastrophes naturelles</b>	<b>11</b>
III.1 Optimisation du <i>Solvency Capital Requirement</i> (SCR) . . . . .	12
III.2 Optimisation des traités de réassurance . . . . .	13
III.2.1 Réassurance : définition et intérêt . . . . .	13
III.2.2 Le développement du marché . . . . .	14
III.2.3 Rentabilité des réassureurs . . . . .	15
III.2.4 Les diverses formes de réassurance . . . . .	16
III.2.4.1 Quote-part ou QP ( <i>Quota share</i> ) . . . . .	17
III.2.4.2 Excédent de pleins ou XP ( <i>Surplus Share</i> ) . . . . .	18
III.2.4.3 Excédent de sinistre ou XS ( <i>excess of loss</i> ou XL) . . . . .	18
III.2.4.4 Excédent de perte ( <i>Stop Loss</i> ou SL) . . . . .	20
<b>Partie 2 : Modélisation des pertes dues aux tremblements de terre</b>	<b>21</b>
<b>I Brève histoire de la sismologie</b>	<b>22</b>
<b>II Les étapes de la modélisation</b>	<b>23</b>
<b>III Module de vulnérabilité : RISK UE LM1</b>	<b>24</b>

III.1	La méthode EMS-98 . . . . .	24
III.1.1	L'échelle d'intensité macrosismique . . . . .	24
III.1.2	L'échelle de dommage aux constructions . . . . .	24
III.1.3	Classification des différentes structures de bâtiment en classe de vulnérabilité . . . . .	25
III.1.4	Matrices de probabilité de dommage . . . . .	26
III.2	La méthode Risk-UE . . . . .	27
III.2.1	Formalisme et intérêt de la logique floue dans le cadre de Risk-UE LM1 . . . . .	27
III.2.2	Estimation du niveau de dommage moyen $\mu_D$ . . . . .	32
III.2.3	Estimation des distributions de dommage . . . . .	33
III.2.3.1	Modélisation par loi binomiale $Bin(5, \frac{\mu_D}{5})$ . . . . .	33
III.2.3.2	Modélisation par loi bêta à quatre paramètres $Beta(r, t - r, a, b)$ . . . . .	33
III.3	Une méthode pour déterminer un unique niveau de dommage . . . . .	35
 <b>Partie 3 : Les sources d'incertitudes associées à la modélisation</b>		<b>37</b>
 <b>I Les incertitudes de type « Modèle »</b>		<b>38</b>
I.1	Loss Ratios, <b>LR</b> . . . . .	38
I.2	Structure du bâtiment, <b>STR</b> . . . . .	39
I.3	Modèle de distribution de pertes, <b>MOD</b> . . . . .	42
 <b>II Les incertitudes de type « Paramètre »</b>		<b>44</b>
II.1	Intensité macrosismique, <b>MMI</b> . . . . .	44
II.2	Indice de vulnérabilité, <b>VI</b> . . . . .	47
 <b>Partie 4 : Analyse globale de sensibilité</b>		<b>50</b>

<b>I</b>	<b>Constitution de la base de données</b>	<b>51</b>
I.1	Représentation des sources d'incertitude . . . . .	51
I.2	Méthode d'échantillonnage . . . . .	52
I.2.1	L'échantillonnage par hypercube latin ( <i>Latin Hypercube Sampling, LHS</i> ) . . . . .	52
I.2.2	Formalisation . . . . .	53
I.2.3	En pratique . . . . .	54
I.2.3.1	L'échantillonnage de la variable <b>STR</b> . . . . .	54
I.2.3.2	L'échantillonnage des variables catégorielles <b>STR, MOD</b> et <b>LR</b> . . . . .	55
I.3	Statistiques descriptives des pertes observées . . . . .	55
<b>II</b>	<b>Décomposition fonctionnelle de la variance</b>	<b>57</b>
II.1	Décomposition ANOVA . . . . .	57
II.2	Décomposition de Hoeffding-Sobol . . . . .	58
II.3	Indices de Sobol . . . . .	59
II.4	Variables dépendantes : limitations de Sobol et intérêt de Shapley . . . . .	60
<b>III</b>	<b>Indices de Shapley</b>	<b>62</b>
III.1	Théorie des jeux coopératifs . . . . .	62
III.2	Application en analyse de sensibilité . . . . .	63
III.3	Estimation et interprétation des indices de Shapley . . . . .	64
<b>Partie 5 : Importance des variables et analyse locale de sensibilité</b>		<b>65</b>
<b>I</b>	<b>Importance d'une variable et modèle de substitution</b>	<b>66</b>
<b>II</b>	<b>Critères de sélection de modèle statistique</b>	<b>67</b>
II.1	La <i>Mean Squared Error</i> (MSE) et la <i>Root Mean Squared Error</i> (RMSE)	67

II.2	La <i>Mean Absolute Error</i> (MAE)	67
II.3	Le critère d'information d'Akaike	68
II.4	Le critère d'information bayésien	68
II.5	Choix des critères d'entraînement et de sélection des modèles	69
<b>III Modèles Paramétriques : du modèle linéaire multiple à la régression bêta</b>		<b>70</b>
III.1	Régression linéaire multiple	70
III.1.1	Présentation du modèle	70
III.1.2	Mise en place et critique du modèle	71
III.2	Régression Bêta	74
III.2.1	Introduction au modèle linéaire généralisé	74
III.2.2	Extension à la distribution bêta	75
III.2.3	Mise en place et critique du modèle	76
<b>IV Modèle semi-paramétrique : GAMLSS (<i>Generalized Additive Model for Location, Scale and Shape</i>)</b>		<b>78</b>
IV.1	Présentation du modèle	78
IV.2	Mise en place et critique du modèle	80
<b>V Modèles non-paramétriques : du modèle MARS aux arbres de régression</b>		<b>81</b>
V.1	MARS ( <i>Multivariate adaptive regression splines</i> )	81
V.1.1	Présentation du modèle	81
V.1.2	Mise en place et critique du modèle	82
V.2	CART ( <i>Classification And Regression Trees</i> )	82
V.2.1	Présentation du modèle	82
V.2.2	Limitations	83



V.3	Gradient Boosting . . . . .	84
V.3.1	Présentation du modèle . . . . .	84
V.3.2	Mise en place et critique du modèle . . . . .	85
V.4	Random Forest . . . . .	85
V.4.1	Mise en place et critique du modèle . . . . .	86
<b>VI</b>	<b>Méthodes additives d'attribution de caractéristique</b>	<b>88</b>
VI.1	Méthodes locales . . . . .	88
VI.2	Propriétés souhaitables . . . . .	89
VI.3	Valeurs SHAP ( <i>SHapley Additive exPlanation</i> ) . . . . .	90
VI.4	Indicateur SHAP d'importance de contribution . . . . .	91
<b>Partie 6</b>	<b>: Extension de l'étude à un portefeuille d'assurances</b>	<b>93</b>
<b>I</b>	<b>Contexte de l'étude</b>	<b>94</b>
I.1	Impacts socio-économiques du séisme . . . . .	94
I.2	Description du portefeuille d'assurances . . . . .	96
<b>II</b>	<b>Construction de la base de données du modèle étendu</b>	<b>98</b>
II.1	Association du portefeuille et des niveaux d'intensité macrosismique . . . . .	98
II.2	Représentation des incertitudes et échantillonnage . . . . .	99
II.3	Application de la méthode Risk-UE LM1 et obtention du <i>loss ratio</i> au niveau portefeuille . . . . .	100
<b>III</b>	<b>Résultats des analyses de sensibilité</b>	<b>102</b>
III.1	Indices de Shapley . . . . .	102
III.2	Indicateur d'importance de contribution . . . . .	103
<b>IV</b>	<b>Conclusion</b>	<b>105</b>

<b>Annexes</b>	<b>107</b>
Annexe A : Détails de la mise en place de la régression linéaire multiple. . . .	107
Annexe B : Détails de la mise en place de la régression bêta. . . . .	108
Annexe C : Détails de la mise en place du GAMLSS bêta inflaté en 0 et 1. . .	110
Annexe D : Détails de la mise en place du modèle MARS . . . . .	111
Annexe E : Optimisation des hyperparamètres du modèle XGBoost . . . . .	113
Annexe F : Quelques outils graphiques d'IA Explicable . . . . .	115
<b>Références</b>	<b>123</b>
<b>Liste des figures</b>	<b>130</b>
<b>Liste des tableaux</b>	<b>132</b>

## Résumé

Mots-clés : Analyse de sensibilité, Catastrophes naturelles, Réassurance, Indices de Shapley, IA explicable, SHAP

Effective depuis 1<sup>er</sup> janvier 2016, la Directive européenne Solvabilité II impose à chaque assureur et réassureur d'être en capacité de comprendre les risques inhérents à son activité afin de pouvoir allouer suffisamment de capital pour les couvrir. Dès lors, il est devenu essentiel de pouvoir déterminer les risques pouvant directement impacter la solvabilité des organismes d'assurance, notamment les événements naturels extrêmes et les pertes économiques qu'ils peuvent entraîner. Pour cela, les acteurs du marché ont recours à des modèles internes ou externes qui leur permettent également d'optimiser leurs traités de réassurance et besoins en capital réglementaire.

L'objectif de ce mémoire est de réaliser une analyse de sensibilité globale et locale d'un modèle de risque sismique. La première partie définit d'abord les catastrophes naturelles et les enjeux qu'elles représentent pour les assureurs et réassureurs. Ensuite, la deuxième partie est consacrée tout particulièrement au module de vulnérabilité et dommage Risk-UE Loss Model 1 dont les sources d'incertitude sont présentées en troisième partie. Dès lors, la quatrième partie a permis de réaliser une analyse globale de sensibilité à partir des valeurs de Shapley. La cinquième partie détermine quant à elle la meilleure réplique du modèle sur laquelle sera appliqué certains concepts d'IA explicable comme les valeurs SHAP pour effectuer une analyse locale de sensibilité et obtenir un indicateur d'importance de contribution. Enfin, la dernière partie est dédiée à un cas assurantiel concret : l'analyse de sensibilité d'un portefeuille de polices d'assurance lors du séisme du 23 novembre 1980 dans la région italienne d'Irpinia.

## Abstract

Key words : Sensitivity Analysis, Natural catastrophes, Reinsurance, Shapley indices, Explainable AI, SHAP

Effective since January 1, 2016, the European Solvency II Directive requires each insurer and reinsurer to be able to understand the risks inherent to its activity in order to allocate sufficient capital to cover them. Therefore, it has become essential to determine the risks that can directly impact the solvency of insurance organizations, in particular extreme natural events and the economic losses they can cause. To do so, market players use internal or external models that also allow them to optimize their reinsurance treaties and regulatory capital requirements.

The aim of this professional thesis is to perform a global and local sensitivity analysis of a seismic risk model. The first part defines natural catastrophes and the issues they represent for insurers and reinsurers. Then, the second part is dedicated to the vulnerability and damage module Risk-UE Loss Model 1 whose sources of uncertainty are presented in the third part. Then, the fourth part consisted of a global sensitivity analysis based on Shapley indices. The fifth part allowed to determine the best replica of the model on which some concepts of explainable AI will be used such as SHAP values to perform a local sensitivity analysis and obtain an indicator of contribution importance. Finally, the last part is dedicated to a concrete insurance case : the sensitivity analysis of a portfolio of insurance policies during the earthquake of November 23, 1980 in the Italian region of Irpinia.

## Introduction

Les catastrophes naturelles se caractérisent comme des événements de faible fréquence provoquant des niveaux de dommages extraordinaires. S'imposant comme des événements d'histoire, leurs récits historiques sont autant de témoignages de l'histoire humaine des peuples régionaux. D'après Bernard Bousquet [2], certains séismes antiques sont par exemple si détaillés que l'on peut évoquer la géographie dans laquelle se produit l'évènement et établir le zonage résultant de la décroissance de l'intensité sismique à l'intérieur de l'aire sismique.

Dans un contexte plus assurantiel, Solvabilité II a encore plus renforcé la nécessité d'étudier l'impact des catastrophes naturelles. La Directive européenne impose notamment à chaque assureur et réassureur d'être à même de comprendre les risques inhérents à son activité afin de pouvoir allouer suffisamment de capital pour les couvrir. Dès lors, pour estimer leurs expositions aux risques catastrophes naturelles et ainsi optimiser leurs traités de réassurance et leurs besoins en capital réglementaire, les acteurs du marché de l'assurance ont recours à des modèles internes ou externes.

En effet, contrairement à l'assurance IARD ou santé, où l'on peut appliquer un modèle fréquence - coût grâce à un historique de sinistres importants et une volatilité faible, l'assurance de catastrophes naturelles nécessite des modèles complexes spécifiques à chaque phénomène physiques. Elles sont en effet trop rares pour constituer un historique significatif et leurs volatilités est bien trop forte.

En particulier, la construction d'un modèle de pertes dues aux mouvements du sol à l'échelle d'une ville, d'une région ou d'un pays demande de choisir des paramètres pertinents tous associés à une incertitude plus ou moins importante. Dès lors, il convient de se poser deux questions :

1. Dans quelle mesure chaque paramètre est-il responsable de l'incertitude associée à la valeur de la perte estimée ?
2. Quelles sont les variables qui contribuent le plus à cette même valeur ?

Dans ce mémoire, nous répondrons à ces questions dans le cadre du modèle de vulnérabilité et dommage Risk-UE Loss Model 1 (LM1). Développé entre 2001 et 2004 dans le cadre du projet européen RISK-UE, ce modèle permet de déterminer un indice de vulnérabilité pour un bâtiment donné. Celui-ci permettra d'obtenir, en fonction de l'intensité macrosismique, une distribution de probabilités de dommages.

Après une première partie dédiée aux généralités sur les catastrophes naturelles et l'importance de leurs modélisations pour les assureurs, nous présenterons les différents paramètres du modèle à l'étude. Dès lors, pour répondre à la première question posée, nous réaliserons une analyse de sensibilité par calcul des indices Shapley.

Ensuite, pour répondre à la seconde question, nous déterminerons quel modèle de régression réplique au mieux le comportement du modèle initial. Ce modèle de substitution nous permettra alors d'appliquer certains concepts d'IA explicable comme les valeurs SHAP pour effectuer une analyse locale de sensibilité et obtenir un indicateur d'importance de contribution.

Ensuite, pour répondre à la seconde question, nous déterminerons quel modèle de régression réplique au mieux le comportement du modèle initial. Ce modèle de substitution nous permettra alors d'appliquer certains concepts d'IA explicable comme les valeurs SHAP pour effectuer une analyse locale de sensibilité et obtenir un indicateur d'importance de contribution.

Enfin, la dernière partie sera dédiée à un cas assurantiel concret : l'analyse de sensibilité globale et locale d'un portefeuille de polices d'assurance lors du séisme du 23 novembre 1980 dans la région italienne d'Irpinia.

Partie 1 :  
Assurance et Catastrophes naturelles

## I Définir les catastrophes naturelles

D'après l'Institut National de la Statistique et des Études Économiques (INSEE), une catastrophe naturelle est caractérisée par l'intensité anormale d'un agent naturel lorsque les mesures habituelles à prendre pour prévenir ces dommages n'ont pu empêcher leur survenance ou n'ont pu être prises. Elle provoque généralement des bouleversements importants et de grands dégâts matériels et humains.

Un événement naturel n'est donc pas à lui seul une catastrophe, surtout s'il se produit loin des zones habitées. En effet, la figure ci-dessous montre que, sur 1046 événements dommageables d'origine naturelle ayant eu lieu en 2021, 45 sont classifiés comme catastrophes, soit environ 4,30%.

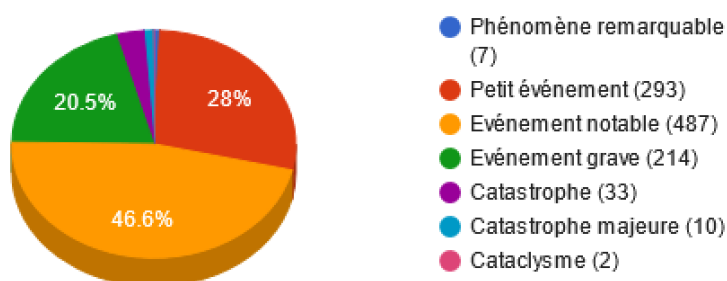


FIGURE 1 – Source : Répartition des événements dommageables d'origine naturelle par gravité survenus dans le monde en 2021. Source : *catnat.net*.

Parmi les événements les plus susceptibles d'être qualifiés comme catastrophe naturelle, on retrouve :

- Séisme ;
- Mouvement de terrain ;
- Ouragan ;
- Tornade ;
- Tsunami ;
- Inondation ;
- Vague de froid ;
- Sécheresse ;
- Feu de forêt ;
- Éruption volcanique.



Selon Munich Re, les dommages dus aux catastrophes naturelles en 2021 s'élèvent à 280 milliards de dollars. L'ouragan Ida (26 août 2021 – 4 septembre 2021) est la catastrophe naturelle la plus importante de l'année, ayant causé à lui seul au moins 65,25 milliards de dollars en dommages, dont 36 milliards de dollars de pertes assurées.

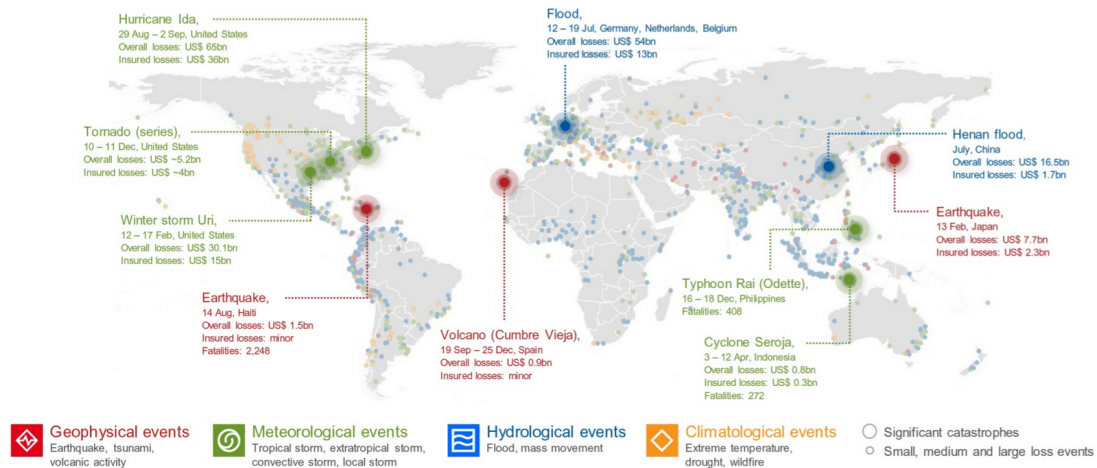


FIGURE 2 – Carte des événements et catastrophes naturels remarquables dans le monde en 2021. Source : Munich Re.

Toute la difficulté pour définir une catastrophe naturelle repose sur la notion d'intensité *anormale* d'un phénomène naturel. On retrouve d'ailleurs, dans le journal officiel du Sénat du 23 juin 1988, une question de M. Pierre Brantus suite à la loi du 13 juillet 1982 relative à l'indemnisation des victimes de catastrophes naturelles. Celui appelle l'attention du ministre de l'économie, des finances et du budget de l'époque sur les conséquences de l'absence de définition formelle de la notion de catastrophe naturelle. La réponse du ministère fut la suivante :

*« Cette notion d'intensité "anormale" est certes difficile à saisir dans une définition car elle supposerait la prise en compte de nombreux paramètres qui, aussi complets soient-ils, ne permettront jamais de répondre à tous les cas d'espèce. Le législateur, conscient de cette difficulté, n'a pas voulu donner une définition plus précise à la notion de catastrophe naturelle, laissant au Gouvernement le soin d'interpréter et de qualifier les faits permettant l'application de ce nouveau régime d'indemnisation, par la reconnaissance au cas par cas au moyen d'arrêtés interministériels ».*

## II Le régime d'indemnisation des catastrophes naturelles

### II.1 Contexte et principes

Après d'importantes inondations dans les vallées de la Saône, du Rhône et dans le Sud-Ouest de la France, le Parlement institue le 13 juillet 1982 avec la loi n° 82-600 un régime spécial d'indemnisation des catastrophes naturelles, dit « régime Cat-Nat ». Cette loi est fondée sur l'alinéa 12 du préambule de la Constitution du 27 octobre 1946, qui dispose que : « la Nation proclame la solidarité et l'égalité de tous les Français devant les charges qui résultent des calamités nationales ».

On peut lire à l'article premier de la loi n°82-600 : « les contrats d'assurance, souscrits par toute personne physique ou morale autre que l'État et garantissant les dommages d'incendie ou tous autres dommages à des biens situés en France, ainsi que les dommages aux corps de véhicules terrestres à moteur, ouvrent droit à la garantie de l'assuré contre les effets des catastrophes naturelles sur les biens faisant l'objet de tels contrats. En outre, si l'assuré est couvert contre les pertes d'exploitation, cette garantie est étendue aux effets des catastrophes naturelles, dans les conditions prévues au contrat correspondant ».

Depuis l'instauration de la loi, c'est la Caisse Centrale de Réassurance (CCR) qui est habilitée à délivrer aux sociétés d'assurance qui en font la demande, une couverture de réassurance illimitée, bénéficiant de la garantie de l'État, pour les risques de catastrophes naturelles en France.

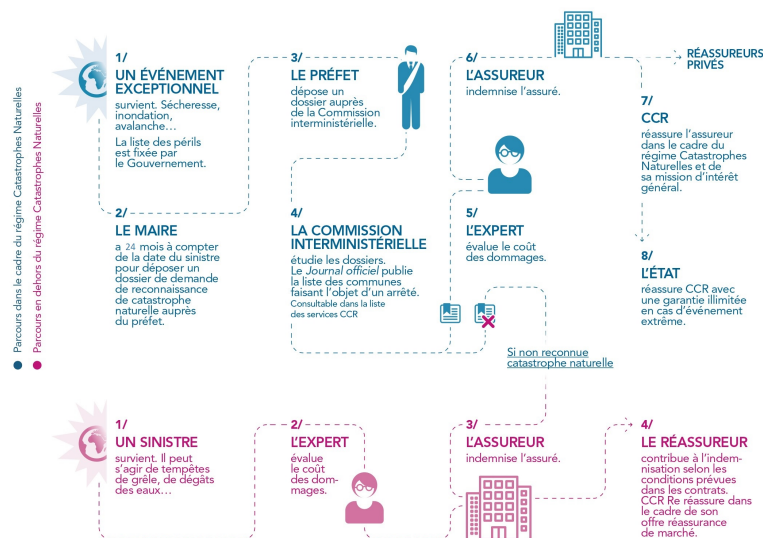


FIGURE 3 – Les étapes du régime Cat-Nat. Source : CCR.

Au titre de la garantie Cat-Nat, les assurés peuvent prétendre à l'indemnisation de leurs dommages matériels résultant directement d'une catastrophe et à une prise en charge des frais de démolition et de déblais des biens assurés ou encore des frais de pompage, de nettoyage et de désinfection des locaux sinistrés. À l'inverse, les dommages corporels ne sont pas couverts.

Il est ici important de préciser que les dommages résultant de tempêtes, de chute de grêle ou de neige ont été exclus du régime des catastrophes naturelles depuis 1990. Ces événements sont compris dans la garantie TGN (Tempête, Grêle, Neige). Par rapport à la garantie tempête, elle ne couvre que les dommages causés par le vent résultant d'un événement cyclonique, à condition que les vents maximaux enregistrés ou estimés sur la zone sinistrée ne dépassent pas une moyenne de 145 km/h sur dix minutes ou 215 km/h en rafales. Dans le cas contraire, c'est alors la garantie catastrophe naturelle s'applique.

L'assuré conserve à sa charge une partie de l'indemnité due par l'assureur, les garanties catastrophes naturelles prévoyant des franchises minimums légales définies à l'Annexe I de l'article A125-1 du code des assurances. Elles sont obligatoires, et non rachetables. Depuis le 1er janvier 2001, elles s'établissent comme suit :

TABLE 1 – Franchises minimales légales. Source : CCR.

Biens à usage d'habitation et autres bien à usage non professionnels	Dommages directs	380€	Sécheresse 1520€
Biens à usage professionnel	Dommage directs	10% minimum 1140€	Sécheresse 10% minimum 3050€
	Pertes d'exploitation	3 jours ouvré minimum 1140€	

Auparavant, une modulation des franchises était prévue pour les communes ne possédant pas de PPRNP (Plans de Prévention des Risques Naturels Prévisibles) ou si un PPRNP avait été prescrit pour le risque faisant l'objet de l'arrêté, mais que celui-ci n'avait pas été approuvé dans le délai de cinq ans suivant la date de la prescription. Toutefois, elles ont disparu le 1er janvier 2023 avec la réforme du régime des catastrophes naturelles.

## II.2 Réforme du régime

Parue au Journal Officiel du 29 décembre 2021, la réforme du régime des catastrophes naturelles est entrée entièrement en vigueur le 1er janvier 2023. Les objectifs de cette réforme sont de rendre moins opaque la procédure de reconnaissance des catastrophes naturelles, de redéfinir les délais de procédure et d'indemnisation et de renforcer la prise en charge des sinistrés.

<b>Simplification de la procédure pour la reconnaissance de l'état de catastrophe naturelle</b>	<b>Des délais plus avantageux pour l'assuré</b>	<b>Généralisation de l'indemnisation</b>
<ul style="list-style-type: none"><li>- Dépôt d'un dossier de reconnaissance de l'état de catastrophe naturelle par les communes : 24 mois après la survenance du sinistre (18 mois auparavant)</li><li>- Publication au Journal officiel de l'arrêté de reconnaissance de l'état de catastrophe naturelle : 2 mois à compter du dépôt des demandes des communes (3 mois auparavant)</li><li>- Les communes et/ou les sinistrés pourront demander la communication des documents ayant permis la prise de décision</li><li>- Recours gracieux facilités en cas de refus de reconnaissance</li><li>- Création au niveau départemental d'un délégué à la reconnaissance de l'état de CAT NAT</li><li>- Nomination d'un référent CAT NAT dans chaque préfecture pour assister les communes dans les démarches</li><li>- Création d'une commission nationale consultative des catastrophes naturelles</li></ul>	<ul style="list-style-type: none"><li>- L'assuré dispose de 30 jours à compter de l'arrêté pour déclarer le sinistre (10 auparavant)</li><li>- L'assureur a un mois, à compter de la réception de la déclaration de sinistre ou de la date de publication de l'arrêté si elle est postérieure, pour prendre position et informer l'assuré sur la mise en jeu de la garantie CAT NAT et l'éventuelle mission d'un expert</li><li>- L'assureur doit, dans le mois qui suit la réception de l'état estimatif transmis par l'assuré, ou du rapport d'expertise, proposer une indemnisation ou une réparation en nature</li><li>- L'assureur devra communiquer le rapport d'expertise à l'assuré</li><li>- À partir de l'accord de l'assuré sur la proposition d'indemnisation, l'indemnité devra être versée dans les 21 jours</li><li>- Prescription des sinistres sécheresse après 5 ans (2 ans auparavant)</li></ul>	<ul style="list-style-type: none"><li>- Suppression de la modulation de franchise pour les sinistrés qui résident dans des collectivités territoriales n'ayant pas encore adopté un plan de PPRNP</li><li>- Prise en charge par tous les assureurs des frais de relogement d'urgence</li><li>- Plus d'indemnisation des seuls dommages matériels directs mais indemnisation permettant de mettre un terme aux « désordres existants »</li><li>- Tout refus d'assurance (ou résiliation) en raison de l'importance du risque Cat Nat qui pèse sur le bien, pourra être contesté devant le bureau central de tarification qui pourra imposer le contrat à l'assureur</li></ul>

FIGURE 4 – Les principaux éléments de la réforme du régime Cat-Nat.

### III Les enjeux pour l'assureur de la modélisation catastrophes naturelles

Inspirée de la réforme Bâle II du secteur bancaire, la Directive Solvabilité II ayant pris effet le 1er janvier 2016 impose à chaque assureur et réassureur de comprendre les risques inhérents à son activité afin de pouvoir allouer suffisamment de capital pour les couvrir. La Directive repose sur trois grands piliers :

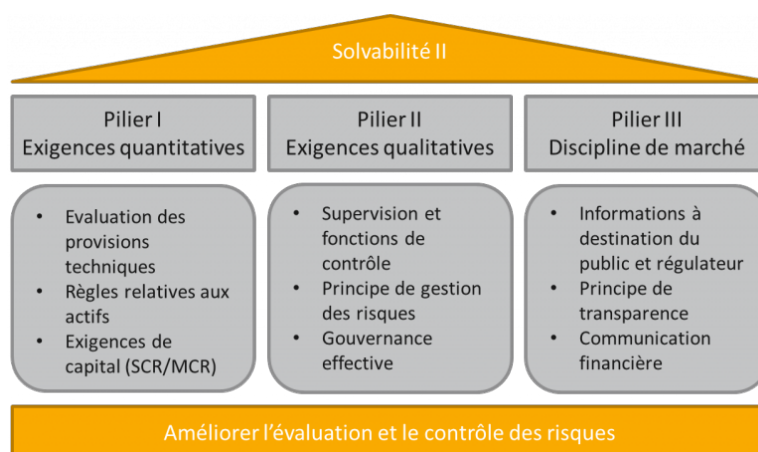


FIGURE 5 – Les grands piliers de Solvabilité II.

Dès lors, pour modéliser avec précision leurs expositions aux risques naturels, les acteurs du marché de l'assurance ont recours à des modèles complexes. Bien que ces modèles puissent être développés en interne, il est souvent difficile pour un assureur de taille modeste de réaliser lui-même cette modélisation. Il peut alors faire appel à des acteurs externes comme Moody's RMS, CoreLogic ou Verisk, trois acteurs majeurs de la modélisation catastrophes naturelles.

AXA est aujourd'hui l'un des seuls assureurs à internaliser cette modélisation sur l'ensemble des périls couverts, ce qui lui permet d'être au plus près du portefeuille d'assurés et de proposer des solutions uniques aux clients, comme Risk Scanning d'AXA XL ou les offres d'assurance paramétrique d'AXA Climate. De plus, un développement en interne permet de maîtriser au mieux les données et une connaissance approfondie des modèles. Cette connaissance permet notamment :

- De définir avec précision leurs besoins en capital réglementaire ;
- D'optimiser leurs traités de réassurance.

C'est dans ce cadre qu'une étude de sensibilité prend toute son importance. Pour un assureur, il est en effet primordial de comprendre dans quelle mesure chaque paramètre est responsable de l'incertitude associée à la valeur de la perte estimée.

### III.1 Optimisation du *Solvency Capital Requirement* (SCR)

Le SCR est défini dans Solvabilité 2 comme le montant de fonds propres à détenir pour limiter la probabilité de ruine à un an à 0,5%. La Directive indique qu'il peut se déterminer par deux approches différentes :

1. SCR formule standard : calculé selon une approche modulaire, il introduit lors de son calcul des corrélations entre les différents modules pour constater des bénéfices de diversification.

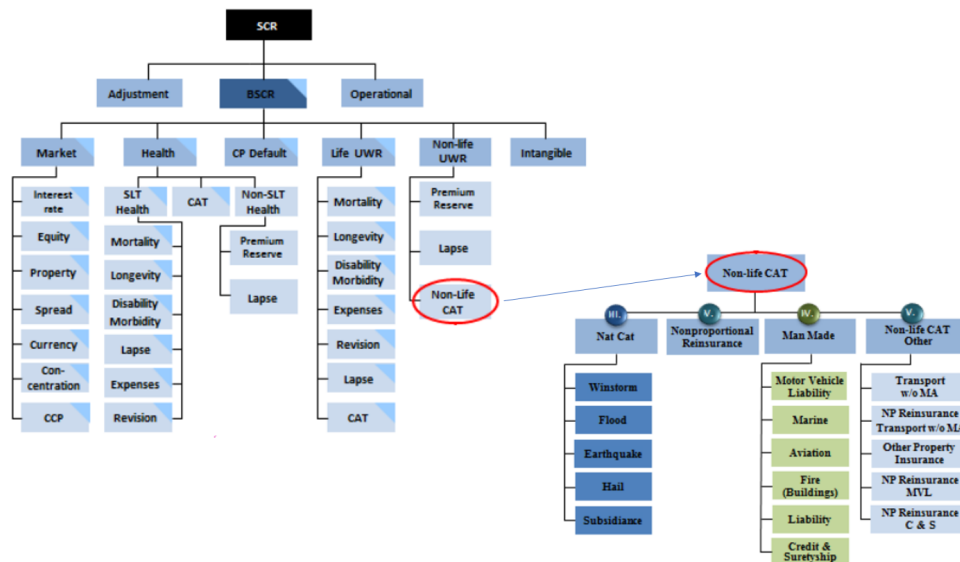


FIGURE 6 – Les modules du SCR formule standard et le sous-module risque catastrophe. Source : ACPR.

2. Modèle interne : bien qu'il possède un coût important de développement et de mise en place, il permet à l'assureur de développer un modèle sur son propre profil de risque permettant de traiter de manière spécifique chaque classe de risque liée à son bilan. Le modèle doit être soumis à une validation du superviseur.

## III.2 Optimisation des traités de réassurance

L'optimisation des traités de réassurance peut apporter plusieurs avantages aux compagnies d'assurance. En choisissant les couvertures les plus appropriées pour leurs besoins, les assureurs peuvent non seulement réduire les coûts liés aux primes de réassurance, mais également améliorer leur profil de risque. Cela peut se traduire par une réduction des pertes techniques et des réserves, ainsi qu'une amélioration de la qualité de leurs portefeuilles d'assurance.

### III.2.1 Réassurance : définition et intérêt

Maurice Picard et André Besson ont défini une opération de réassurance comme : « un contrat sur lequel un réassureur (dit cessionnaire) vis-à-vis d'un assureur professionnel (dit cédant) qui répond seul et intégralement vis-à-vis des assurés des risques par lui assurés, prend en charge moyennant rémunération tout ou partie des sommes dues ou versées aux assurés à titre de sinistres ». Plus trivialement, un réassureur peut se décrire comme l'assureur des sociétés d'assureur.

Selon Delacroix [5], la réassurance permet d'abord d'augmenter la capacité de souscription de la cédante et d'améliorer sa marge de solvabilité. Ensuite, elle protège le bilan de la cédante contre la survenance de sinistres extrêmes comme les catastrophes naturelles, caractérisées par une volatilité forte et des sinistres importants. La réassurance permet également un lissage dans le temps des résultats de la cédante impliquant entre autres une diminution du coup du capital. Ce phénomène de lissage s'observe d'ailleurs directement dans la figure ci-dessous :

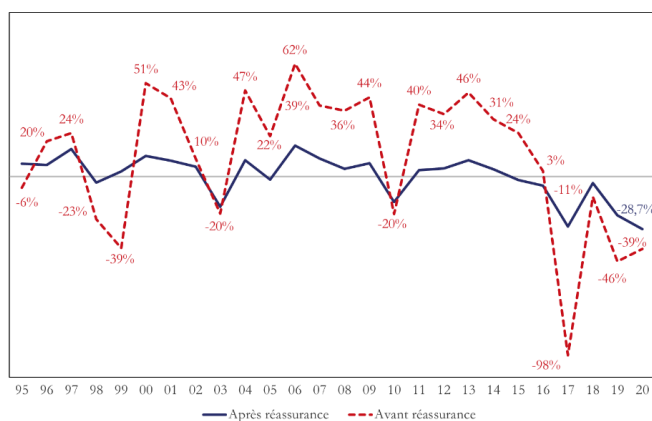


FIGURE 7 – Impact de la réassurance sur le résultat technique représenté en pourcentage des primes avant et après réassurance. Source : Fédération Française de l'Assurance (FFA).

### III.2.2 Le développement du marché

Le premier premier contrat de réassurance remonte à 1370, lorsque deux assureurs vénitiens qui assuraient la marchandise d'un bateau décidèrent de céder le risque à un troisième assureur pour l'étape la plus risquée du voyage. Par la suite, le marché s'est développé avec le raffinement des méthodes de réassurance, les besoins croissants de capacité des secteurs industriels en développement et la survenance des phénomènes naturels. Par exemple, *Cologne Ré*, la première compagnie de réassurance professionnelle indépendante, a été fondée en 1846 suite aux pertes causées par le grand incendie de Hambourg. De même, les besoins en capacité suites aux pertes de l'ouragan Andrew en 1992 ont entraîné une croissance importante des réassureurs bermudiens.

Aujourd'hui, le marché s'est entièrement développé et les sociétés de réassurance sont à majorité multibranches. Les acteurs du marché les plus importants sont les suivants :

TABLE 2 – Les 10 premiers réassureurs mondiaux selon les primes émises brutes en 2020. Source : A.M. Best.

Rang	Compagnie	Pays	Chiffre d'affaires 2020 (Md\$)			Fonds propres (Md\$)	Ratio en % (activité non vie seule)	
			Total	Non vie	Vie		Sinistres à primes	Combiné
1	Munich Re	Allemagne	45,846	30,237	15,609	36,845	74,7	105,6
2	Swiss Re	Suisse	36,579	21,512	15,067	27,258	78,7	109
3	Hannover Re	Allemagne	30,421	20,568	9,853	14,543	72,8	101,6
4	SCOR	France	20,106	8,795	11,311	7,588	70,2	100,2
5	Berkshire Hathaway	États-Unis	19,195	13,333	5,862	451,336	81	106,5
6	China Re	Chine	16,665	6,422	10,243	15,772	68	101,8
7	Lloyd's	Royaume Uni	16,511	16,511	-	45,010	73,1	110,3
8	Canada Life Re	Canada	14,552	-	14,552	21,137	-	-
9	Reinsurance Group of America	États-Unis	12,583	-	12,583	14,352	-	-
10	Korean Re	Corée du Sud	7,777	6,427	1,350	2,261	84,6	99,5

Dans le même temps, il est intéressant de remarquer le développement du marché alternatif qui représente près de 100 milliards de dollars d'actifs totaux et une contribution de plus de 15% du capital total de réassurance mondial. Ce marché comprend notamment des Insurance-Linked Securities qui sont des titres et des produits dérivés permettant de transférer les risques de catastrophe des compagnies d'assurances et de réassurances à des investisseurs des marchés financiers.



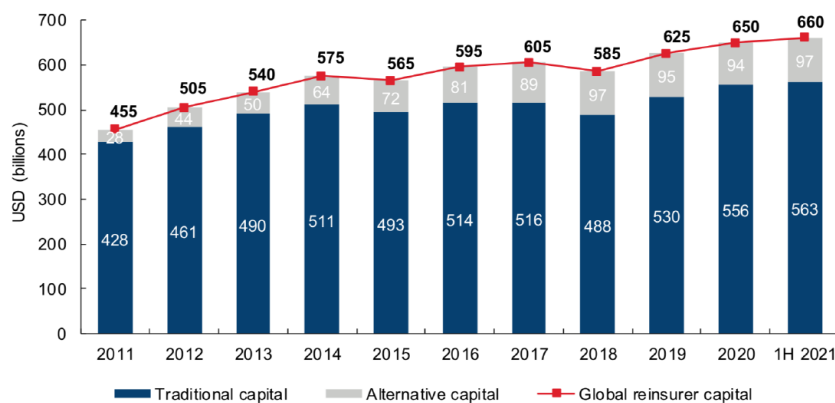


FIGURE 8 – Evolution du capital de la réassurance. Source : Aon.

### III.2.3 Rentabilité des réassureurs

Le ratio combiné (*combined ratio*) est fréquemment utilisé comme mesure de rentabilité pour évaluer leurs performances. Si le ratio combiné dépasse 100%, les dépenses sont supérieures aux recettes. L'assureur peut toutefois compenser ses pertes techniques par ses bénéfices financiers. Le ratio combiné est défini comme suit :

$$Combined\ Ratio = \frac{S + E}{P} \quad (1)$$

Avec :

- $S$  le montant réglé au titre des sinistres ;
- $E$  les dépenses encourues ;
- $P$  le montant des primes acquises.

Un ratio combiné supérieur à 100% est généralement dû à une sinistralité anormalement élevée. L'année 2011 a par exemple été marquée par une série de catastrophes naturelles sans précédent, notamment le séisme de Christchurch en Nouvelle-Zélande, le séisme et le tsunami de Tōhoku au Japon et les inondations en Thaïlande. De même, la saison cyclonique 2017 dans l'océan Atlantique nord a été particulièrement intense avec les ouragans Maria et Harvey. Enfin, le ratio combiné a de nouveau dépassé les 100% en 2020 en raison de l'épidémie de Covid-19. Toutefois, il semble que l'on assiste, depuis le début de l'année 2021, à un retour sous le seuil des 100%.

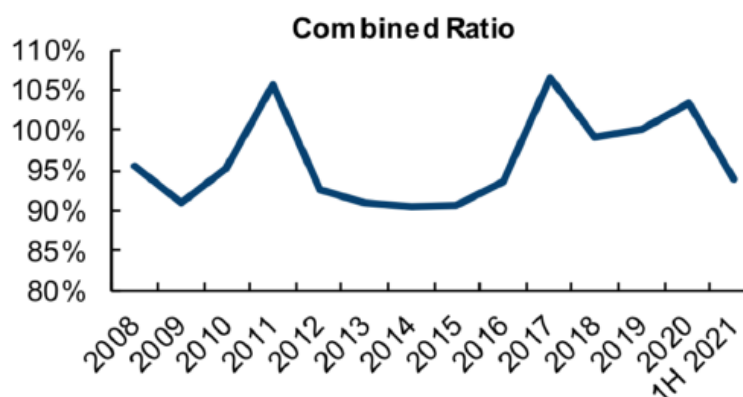


FIGURE 9 – Evolution du ratio combiné des réassureurs. Source : A. Delacroix.

On observe bien le côté cyclique de la réassurance à travers l'évolution des ratios combinés : lorsque les prix sont élevés, on dit que l'on est en *hard market* et, lorsque les prix sont bas, on parle de *soft market*. Delacroix [5] ajoute d'ailleurs que les cycles des différentes branches de la réassurance ne sont pas nécessairement corrélés.

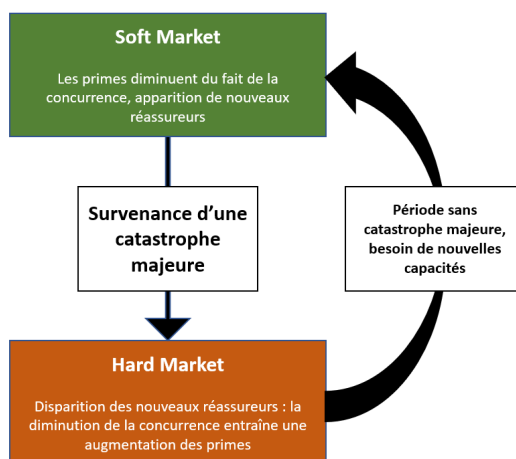


FIGURE 10 – Le cycle de la réassurance.

### III.2.4 Les diverses formes de réassurance

On distingue d'abord la réassurance obligatoire de la réassurance facultative. En réassurance obligatoire, la cédante est tenue de céder tous les risques qui répondent aux conditions spécifiées dans le contrat de réassurance, et le réassureur est tenu de les accepter sans effectuer de sélection. En revanche, dans le cas de la réassurance facultative, l'accord s'établit pour chaque risque, police par police. La cédante a la liberté de proposer les risques qu'elle souhaite au réassureur, qui a également la liberté de les accepter ou de les refuser.

Il est ensuite nécessaire de distinguer la réassurance proportionnelle, où le réassureur paie un montant proportionnel à celui versé par l'assureur et la réassurance non proportionnelle qui représente les autres types de traités.

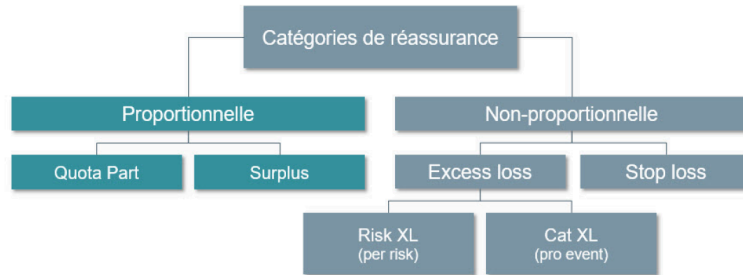


FIGURE 11 – Les différentes catégories de réassurance.

### III.2.4.1 Quote-part ou QP (*Quota share*)

Dans ce type de traité, le réassureur partage un pourcentage équivalent des primes et des sinistres du portefeuille de la cédante. Ce pourcentage, appelé taux de cession, est identique pour tous les risques en portefeuille, quelle que soit la somme assurée. Ainsi, le profil de portefeuille demeure le même à une transformation homothétique près.

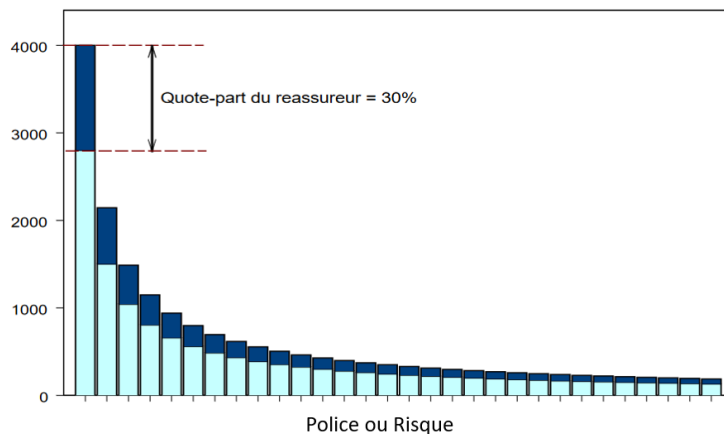


FIGURE 12 – Résultat d'une réassurance en quote-part sur un portefeuille de sinistres. *Lecture : dans cet exemple, le taux de cession est fixé à 30%.*

Bien que le quote-part permette une économie de fonds propres et une couverture illimitée, il présente toutefois un défaut majeur : la cédante peut aussi avoir à céder certains risques qu'elle aurait eu la capacité de conserver sans faire appel à la réassurance.

### III.2.4.2 Excédent de pleins ou XP (*Surplus Share*)

Dans le cas d'un traité en excédent de pleins, le réassureur prend en charge la portion de risque dépassant un seuil appelé plein de rétention jusqu'à un engagement maximal. Cela permet à la cédante d'homogénéiser la mutualisation et de conserver une part non négligeable des primes. D'après la CCR, un tel traité est cependant plus complexe à mettre en oeuvre et demande une gestion relativement lourde.

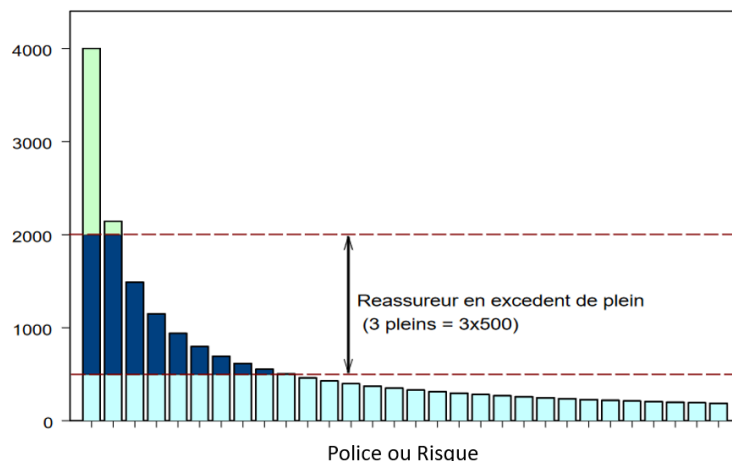


FIGURE 13 – Résultat d'une réassurance en excédent de plein sur un portefeuille de sinistres. *Lecture : dans cet exemple, la cédante paye un montant maximal par sinistre de 500€ et obtient une capacité supplémentaire auprès d'un réassureur correspondant à 3 pleins, soit 1500€.*

### III.2.4.3 Excédent de sinistre ou XS (*excess of loss* ou XL)

Ce contrat fonctionne comme un contrat d'assurance avec franchise. Bien qu'il pondère moins le calcul de la marge de solvabilité que la réassurance proportionnelle, il est adapté à la quasi totalité des branches et permet une couverture totale au-delà de la franchise, dans la limite du plafond de couverture. Il peut être assorti d'une garantie de reconstitution de garantie limitée ou illimitée. Celle-ci permet de reconstituer la couverture en cas de charge sinistre, selon les termes du contrat de réassurance, moyennant le paiement éventuel d'une prime supplémentaire. Si l'on note  $S_{XS}$  le montant à la charge du réassureur, alors :

$$S_{XS} = \min(\max(X - \text{Priorité}, 0), \text{Portée}) \quad (2)$$

Avec :

- $S_{XS}$  le montant à la charge du réassureur ;
- $X$  le montant d'un sinistre couvert par le traité ;
- *Priorité* le montant correspondant à la franchise déductible. le réassureur s'engage à payer pour tous les sinistres dépassant cette franchise et uniquement pour le montant de ce dépassement ;
- *Portée* l'engagement maximum du réassureur sur un sinistre.

En pratique, on le présente de la manière suivante : *portée XS priorité* et se décline généralement sous deux formes. Tout d'abord, le XS par risque qui couvre la cédante pour chaque risque individuel. Dans ce cas, la priorité et la portée sont appliquées aux sinistres individuels. Ensuite, le XS par événement qui couvre la cédante pour l'ensemble des polices sinistrées du fait d'une même cause. Ici, la priorité et la portée sont appliquées à l'événement et non pas individuellement à chaque sinistre.

Il peut également comprendre certaines clauses spéciales de type *aggregate*. Cela peut être une *Annual Aggregate Limit* (AAL) qui correspond au plafond annuel d'engagement du réassureur ou bien une *Annual Aggregate Deductible* (AAD) qui est à une franchise annuelle sur la somme agrégée des paiements dus par le réassureur à la cédante.

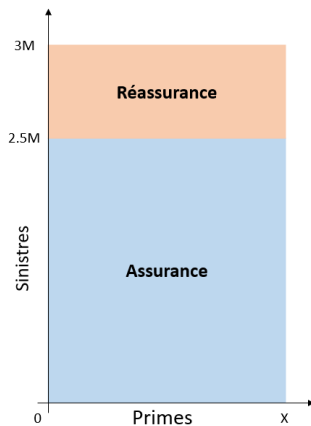


FIGURE 14 – Schéma d'un excédent de sinistre 500 000 € XS 2 500 000 € (*excess of loss* ou XL). Source : Wikipédia.

Un XS est souvent découpé en plusieurs tranches indépendantes qui sont toutes cotées séparément. On utilise généralement le *Rate on Line* (RoL) comme indicateur du prix d'une tranche. Souvent utilisé lors des renouvellements pour mesurer l'évolution du prix de la réassurance, il se définit comme :

$$RoL = \frac{\text{Prime de réassurance}}{\text{Portée}} \quad (3)$$

On retrouve également l'utilisation du *Payback* qui représente le nombre d'années de primes nécessaires pour payer un sinistre traversant totalement la tranche.

$$Payback = \frac{1}{RoL} = \frac{Portée}{Prime\ de\ réassurance} \quad (4)$$

#### III.2.4.4 Excédent de perte (*Stop Loss* ou SL)

Le principe du *Stop Loss* est le même que celui de l'XS, mis à part qu'il protège le ratio S/P de l'année. C'est notamment la couverture qu'utilise la CCR pour le régime Cat-Nat. Ici, la priorité et la portée sont en général exprimées en pourcentage de la prime directe.

$$S_{SL} = \min \left( \max \left( \frac{\sum X_i}{P} - Priorité, 0 \right), Portée \right) \times P \quad (5)$$

Avec :

- $S_{SL}$  le montant à la charge du réassureur ;
- $P$  le montant des primes acquises pendant l'exercice ;
- $X_i$  le montant du  $i$ -ème sinistre de la période de couverture couvert par le traité ;
- *Priorité* le montant correspondant à la franchise déductible. le réassureur s'engage à payer pour tous les sinistres dépassant cette franchise et uniquement pour le montant de ce dépassement ;
- *Portée* la portée qui correspond à l'engagement maximum du réassureur sur un sinistre.

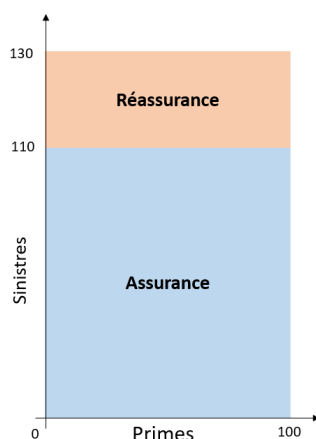


FIGURE 15 – Schéma d'un excédent de perte 20% SL 110% (*stop-loss* ou SL). Source : Wikipédia.

**Partie 2 :**  
**Modélisation des pertes dues aux**  
**tremblements de terre**

## I Brève histoire de la sismologie

L'histoire de la sismologie commence selon Jean-Paul Poirier [20] avec Aristote qui voyait dans le traité des *Météorologiques* un lien entre les vents et les tremblements de terre. On parle de théorie « pneumatique ». Sénèque développa ensuite dans *Questions Naturelles* l'idée que les séismes étaient dus à de brusques et violentes émissions de vapeur d'eau, suite au réchauffement de poches d'eau par la chaleur interne de la Terre. À la Renaissance, certains pensaient qu'ils étaient dus aux embrasements et explosions internes des volcans, vus alors comme des soupapes de sécurité de la Terre. C'est ensuite l'électricité qui au cours du 18<sup>ème</sup> siècle fut désignée responsable à cause d'accumulations de charge dans des cavités souterraines.

Il faut attendre la fin du 19<sup>ème</sup> siècle et le début du 20<sup>ème</sup> siècle, période marquée par une forte sismicité dans des régions fortement peuplées, pour que l'on commence à s'intéresser vraiment aux effets géologiques et tectoniques, à identifier les régions sismiques et à comprendre les mécanismes à la source des séismes. On comprend alors qu'ils proviennent de la rupture violente de masses rocheuses, soumises en profondeur à des contraintes mécaniques fortes. En craquant, les matériaux rocheux libèrent des ondes sismiques qui se propagent à la fois à l'intérieur de la planète mais aussi en surface. Ce sont surtout ces dernières qui provoquent des dégâts considérables dans les zones habitées.

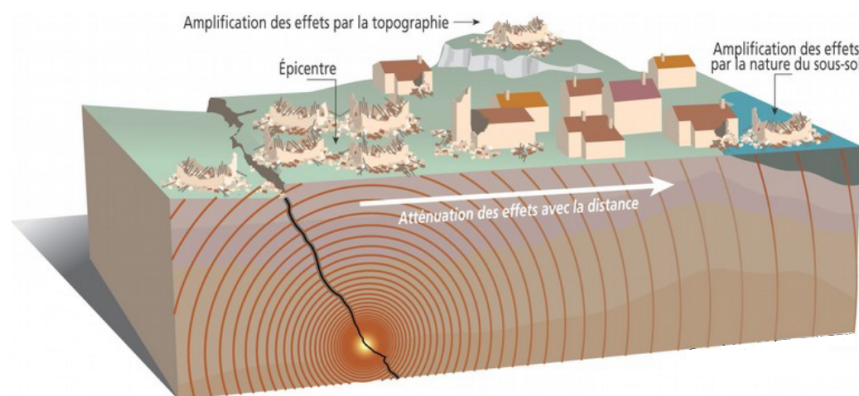


FIGURE 16 – Représentation simplifiée d'un séisme. Source : Wikipédia.

D'après Wikipédia, c'est au 19<sup>ème</sup> siècle que le géologue et sismologue français Alexis Perrey, en cataloguant tous les séismes anciens et contemporains, posa les bases de ce qui devint la sismologie historique. Aujourd'hui, elle permet grâce à des catalogues critiques d'identifier les zones à risque et les intervalles de récurrence des séismes.



## II Les étapes de la modélisation

### 1 - Module Aléa

Des modèles pointus sont utilisés pour simuler des phénomènes naturels. Le modèle peut être interne ou celui d'experts externes (AIR, EQECAT, RMS...). Il permet de générer un large catalogue stochastique d'événements réalistes et probabilisés. Un catalogue historique peut aussi être utilisé pour analyser certains résultats.



### 2 - Module Vulnérabilité

Pour chaque site constituant l'exposition de l'assureur, diverses informations sont connues : localisation, valeur assurée, garantie de contrat (bâtiment, contenu ou perte d'exploitation), caractéristique du bâtiment (structure, nombre d'étages, année de construction ...). Chaque combinaison est associée à une unique courbe de vulnérabilité qui donne la probabilité de dépasser un taux de dommage pour une intensité donnée



### 3 - Risque Brut

Le croisement du module vulnérabilité et du module aléa permet d'obtenir une estimation de la perte brute pour chacun des sites en portefeuille. On peut alors tracer la courbe OEP et AEP brutes de réassurance et réaliser les premières analyses.



### 4 - Module Financier

On applique les conditions de réassurance (traités) sur les données. Les traités de réassurance peuvent être de type proportionnels ou non proportionnels et généralement sous forme de layers (tranches). Si l'on déduit le montant de réassurance du montant brut, on peut obtenir le montant net pour chaque site.



### 5 - Risque Net

Après application des conditions de réassurance, on obtient pour chaque site en portefeuille, le risque net de réassurance. On peut de nouveau tracer les courbes OEP et AEP nettes de réassurance.



FIGURE 17 – Les étapes d'une modélisation de catastrophe naturelle en assurance.

### III Module de vulnérabilité : RISK UE LM1

#### III.1 La méthode EMS-98

L'échelle Macrosismique Européenne EMS-98 [11][12] est une méthode probabiliste qui définit une échelle d'intensité pour les séismes prenant en compte la vulnérabilité du bâtiment. Depuis les années 2000, l'EMS-98 est considérée comme un standard en Europe pour les études d'évaluation du risque ou de la vulnérabilité sismique. La méthode EMS-98 introduit plusieurs éléments majeurs : une échelle d'intensité macrosismique, une échelle de dommage aux constructions, une classification des différentes structures de bâtiment en classe de vulnérabilité distincte et des matrices de corrélation entre l'intensité sismique et l'endommagement d'une classe de bâtiments.

##### III.1.1 L'échelle d'intensité macrosismique

L'échelle d'intensité macrosismique est basée sur les observations de l'impact d'un séisme sur un lieu donné, telles que les dommages causés aux bâtiments et les effets ressentis par les populations. Elle permet de quantifier l'intensité des secousses et leur impact sur les structures et les populations.

Intensités EMS98	I	II	III	IV	V	VI	VII	VIII	IX	X+
Dégâts potentiels bâtiments vulnérables	Aucun	Aucun	Aucun	Aucun	Très léger	Modérés	Quelques effondrements partiels	Nombreux effondrements partiels	Nombreux effondrements	Effondrements généralisés
Dégâts potentiels bâtiments peu vulnérables	Aucun	Aucun	Aucun	Aucun	Aucun	Aucun	Très léger	Modérés	Effondrements partiels	Nombreux effondrements
Perception humaine	non ressentie	Très faible	Faible	Modérés	Forte	Brutale	Très brutale	Sévère	Violente	Extrême

FIGURE 18 – Échelle d'intensité macrosismique (EMS-98).

##### III.1.2 L'échelle de dommage aux constructions

L'échelle de dommage aux constructions se répartit sur 6 niveaux : aucun (*none*), négligeable à léger (*negligeable to slight*), modéré (*moderate*), substantiel à important (*substantial to heavy*), très important (*very heavy*) et destruction.

Degré	0. None	1. Negligeable to Slight	2. Moderate
Dégâts constatés	Aucun dégât structurel, aucun dégâts non structurels	Aucun dégât structurel, légers dégâts non structurels	Dégâts structurels légers, dégâts non structurels modérés
Degré	3. Substantial to Heavy	4. Very heavy	5. Destruction
Dégâts constatés	Dégâts structurels modérés, dégâts non structurels importants	Dégâts structurels importants, dégâts non structurels très importants	Dégâts structurels très importants

FIGURE 19 – Échelle de dommage aux constructions (EMS-98).

La manière dont un bâtiment se déforme sous la charge d'un tremblement de terre dépend de sa nature. Pour illustrer ce principe, il est possible de différencier les bâtiments en maçonnerie des bâtiments en béton armé :

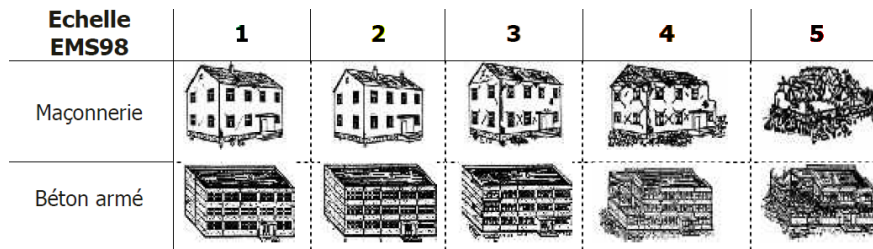


FIGURE 20 – Illustration de l'échelle des dommages EMS-98 sur des bâtiments en maçonnerie et en béton armé. Source : EMS-98

### III.1.3 Classification des différentes structures de bâtiment en classe de vulnérabilité

En parallèle, la méthode EMS-98 définit six classes de vulnérabilité, notées de A à F, ordonnées selon un niveau de vulnérabilité décroissant.

Type of Structure	Vulnerability Class					
	A	B	C	D	E	F
MASONRY	○					
	○	—				
	○					
	○	—	—			
	○	—	—	—		
	○	—	—	—	—	
	○	—	—	—	—	—
REINFORCED CONCRETE (RC)	○	—	—	—		
	○	—	—	—	—	
	○	—	—	—	—	—
	○	—	—	—	—	—
	○	—	—	—	—	—
STEEL			○	—	—	
WOOD		○	—	—		

○ most likely vulnerability class; — probable range;  
 ..... range of less probable, exceptional cases

FIGURE 21 – Différenciation des structures en classes de vulnérabilité. Source : RISK-UE. Lecture : un bâtiment en massive stone appartient très probablement à la classe C, probablement à la classe B et exceptionnellement à la classe D.

### III.1.4 Matrices de probabilité de dommage

Enfin, la méthode EMS-98 propose, pour chaque classe de vulnérabilité, une corrélation entre l'intensité sismique et l'endommagement des bâtiments. Elle utilise pour cela des indications du type « *Few* » (un peu), « *Many* » (beaucoup) et « *Most* » (la plupart). Ces indications sont comprises dans un ensemble de *Damage Probability Matrices* (DPM, matrices de probabilité de dommage).

Class A					
Damage Intensity	1	2	3	4	5
V	Few				
VI	Many	Few			
VII			Many	Few	
VIII				Many	Few
IX					Many
X					Most
XI					
XII					

Class B					
Damage Intensity	1	2	3	4	5
V	Few				
VI	Many	Few			
VII		Many	Few		
VIII			Many	Few	
IX				Many	Few
X					Many
XI					Most
XII					

Class C					
Damage Intensity	1	2	3	4	5
V					
VI	Few				
VII		Few			
VIII		Many	Few		
IX			Many	Few	
X				Many	Few
XI					Many
XII					Most

Class D					
Damage Intensity	1	2	3	4	5
V					
VI					
VII	Few				
VIII		Few			
IX		Many	Few		
X			Many	Few	
XI				Many	Few
XII					Most

Class E					
Damage Intensity	1	2	3	4	5
V					
VI					
VII					
VIII					
IX		Few			
X		Many	Few		
XI			Many	Few	
XII					

Class F					
Damage Intensity	1	2	3	4	5
V					
VI					
VII					
VIII					
IX					
X		Few			
XI		Many	Few		
XII					

FIGURE 22 – Type de bâtiment selon EMS-98 et identification de leur comportement sismique par classes de vulnérabilité. Source : RISK-UE. *Lecture : En intensité VII de nombreux bâtiments de la classe de vulnérabilité A subissent des dégâts de degré 3, quelques uns de degré 4.*

## III.2 La méthode Risk-UE

La méthodologie Risk-UE [17] a été développée de 2001 à 2004 dans le cadre d'un projet européen d'évaluation du risque sismique de grandes villes européennes. Elle comporte deux niveaux d'analyse : le premier niveau nommé LM1 se base sur les corrélations statistiques entre l'intensité macrosismique et le dommage apparent, décrit en termes de degré de dommage. Le deuxième niveau nommé LM2 est une méthode analytique qui est basée sur une analyse mécanique du comportement d'une structure face aux mouvements du sol sous forme d'un spectre d'accélération.

On ne s'intéressera ici qu'au premier des deux niveaux d'analyse. Celui-ci reprend et complète les notions présentes dans EMS-98 afin d'obtenir, pour chaque type de structure et degré d'intensité macrosismique, une distribution probabilisée de dommage. Pour ce faire, RISK UE LM1 introduit deux éléments majeurs :

- Un indice de vulnérabilité  $V_I$  qui représente et quantifie l'appartenance d'un bâtiment à une certaine classe de vulnérabilité. L'échelle de valeurs de cet indice est arbitraire ;
- Une fonction de vulnérabilité qui lie le niveau de dommage moyen  $\mu_D$  avec l'intensité macrosismique  $I$  et l'indice de vulnérabilité  $V_I$ .

### III.2.1 Formalisme et intérêt de la logique floue dans le cadre de Risk-UE LM1

Assouplissant l'algèbre de Boole, la logique floue remplace la valeur de vérité d'une proposition à choisir dans  $\{vrai, faux\}$  par un degré de vérité, à choisir par exemple dans  $[0, 1]$ . En logique floue, il y a donc des degrés dans la satisfaction d'une condition. Étudiée depuis 1920 par Łukasiewicz et Tarski, la logique floue a été formalisée en 1965 avec la théorie des ensembles flous proposée par Zadeh [36].

Un sous-ensemble flou  $\mathbf{A}$  d'un ensemble de référence  $\mathbf{E}$  est caractérisé à l'aide d'une fonction d'appartenance  $\chi_{\mathbf{A}} : \mathbf{E} \rightarrow [0; 1]$  (degré d'appartenance qui est l'extension de la fonction caractéristique d'un sous-ensemble classique). Pour un sous-ensemble flou  $\mathbf{A}$  d'un référentiel  $\mathbf{E}$  on donne les définitions suivantes :

$$\mathbf{A}_{\alpha} = \{x \in \mathbf{R} \mid \chi_{\mathbf{A}}(x) \geq \alpha\} \quad (6)$$

$$Noyau(\mathbf{A}) = \{x \in \mathbf{R} \mid \chi_{\mathbf{A}}(x) = 1\} \quad (7)$$

$$Support(\mathbf{A}) = \{x \in \mathbf{R} \mid \chi_{\mathbf{A}}(x) \neq 0\} \quad (8)$$

Autrement dit,  $Noyau(\mathbf{A})$  s'interprète comme l'ensemble des éléments «vraiment» dans  $\mathbf{A}$  et  $Support(\mathbf{A})$  comme ceux qui y sont à des degrés divers.  $\mathbf{A}_\alpha$  est appelé  $\alpha$ -coupe de  $\mathbf{A}$ .

Les fonctions d'appartenances trapézoïdales sont utilisées en particulier par Lagomarsino et Giovinazzi [10] pour caractériser l'appartenance à une classe de vulnérabilité. Elles sont définies par une valeur basse  $a$ , une valeur haute  $d$  et deux valeurs  $b$  et  $c$  qui représentent les limites de son noyau. La formule de la fonction d'appartenance trapézoïdale est représentée comme suit :

$$\chi_{\mathbf{A}}(x) = \begin{cases} 0 & \text{si } x \leq a - \alpha \text{ ou } x \geq b + \beta \\ 1 & \text{si } a < x < b \\ 1 + \frac{x-a}{\alpha} & \text{si } a - \alpha < x < a \\ 1 - \frac{b-x}{\beta} & \text{si } b < x < b + \beta \end{cases} \quad (9)$$

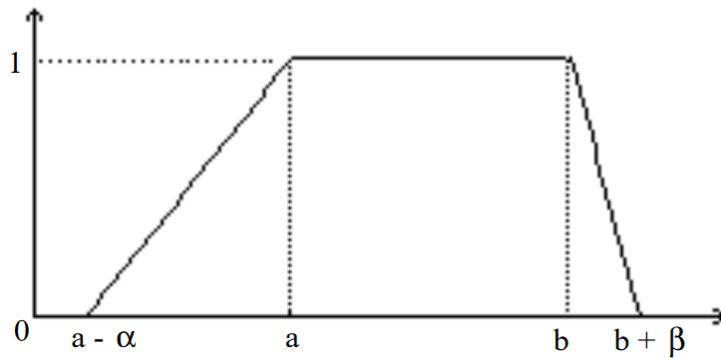


FIGURE 23 – Représentation d'une fonction d'appartenance trapézoïdale.

Ainsi, la logique floue permet de raisonner non pas sur des variables numériques, mais sur des variables linguistiques, c'est-à-dire, sur des variables qualitatives : grand, petit, moyen, loin, près, fort, etc. Raisonner sur ces variables linguistiques permet de manipuler des connaissances en langage naturel. Il devient donc possible, par la théorie des ensembles flous, de traduire quantitativement les descriptions qualitatives utilisées dans les DPM d'EMS-98 : *Many*, *Most* et *Few*.

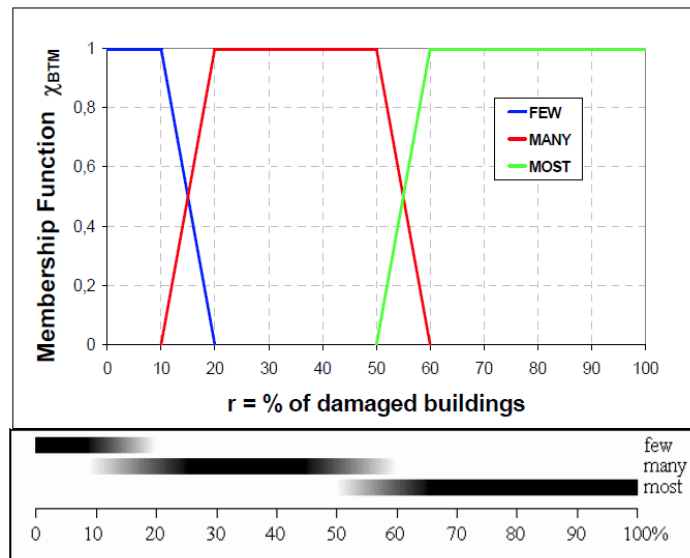


FIGURE 24 – Quantités EMS-98 « peu/*few* », « beaucoup/*many* » et « la plupart/*most* » exprimés en termes d’intervalles superposés et de fonctions d’appartenance. Source : RISK-UE. *Lecture : lorsque l’on observe peu de bâtiments endommagés, alors leur proportion est plausiblement entre 0% et 10%, et possiblement entre 10% et 20%. Cela se traduit par une fonction d’appartenance prenant constamment la valeur 1 entre 0 et 0,1 puis qui décroît linéairement jusqu’à 0 en 0,2.*

De plus, la méthode RISK-UE LM1 introduit les fonctions d’appartenance des indices de vulnérabilité pour chaque classe de vulnérabilité de A à F :

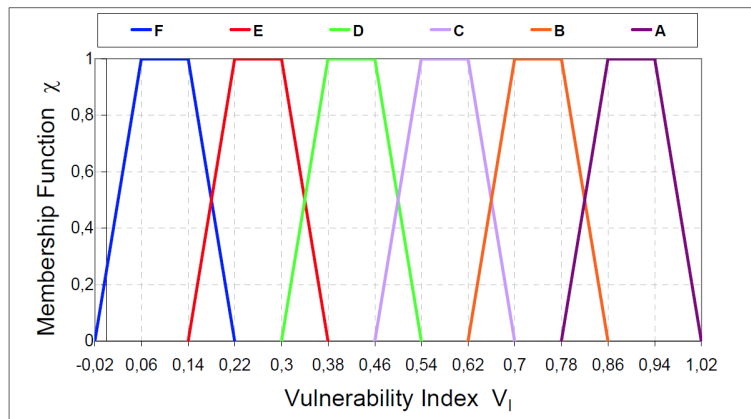


FIGURE 25 – Fonctions d’appartenance des indices de vulnérabilité par classe de vulnérabilité. Source : RISK-UE.

Enfin, RISK-UE associe également des valeurs d’appartenance aux termes introduits en figure 21 qui, pour rappel, traduisent la plausibilité de l’appartenance d’un bâtiment à une classe de vulnérabilité tels que *most likely* ( $\chi = 1$ ), *probable* ( $\chi = 0,6$ ), et *less probable* ( $\chi = 0,2$ ). Dès lors, la fonction d’appartenance d’une structure est

définie par Lagomarsino et Giovinazzi [10] comme la somme algébrique des fonctions d'appartenance des classes de vulnérabilités auxquelles appartient la structure, coéfficientées par le degré d'appartenance de la structure à la classe.

Pour illustrer ce propos, considérons un bâtiment dont la structure est en bois (*wood*). En se référant directement à la figure 21, on obtient la table suivante :

TABLE 3 – Degré d'appartenance de la structure *wood* aux classes de vulnérabilité B, C, D et E.

Classe	B	C	D	E
Appartenance à la classe	<i>Less probable</i>	<i>Probable</i>	<i>Most likely</i>	<i>Probable</i>
$\chi$	0,2	0,6	1	0,6

Dès lors, la fonction d'appartenance d'une structure en bois  $\chi_{Wood}$  est telle que :

$$\chi_{Wood} = 1 \times \chi_D + 0,6 \times (\chi_C + \chi_E) + 0,2 \times \chi_B \quad (10)$$

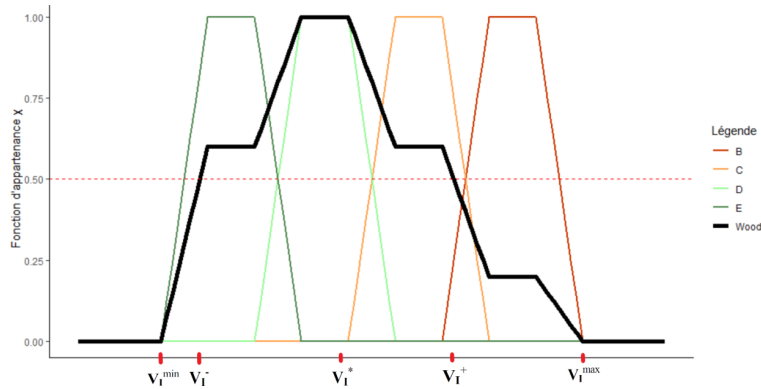


FIGURE 26 – Fonctions d'appartenance et indices de vulnérabilité remarquables d'un bâtiment dont la structure est en bois. *Lecture : les courbes B, C, D et E représentent les fonctions d'appartenance de chaque classe de vulnérabilité.*

En particulier, les indices de vulnérabilité remarquables sont :

- $V_{I,BTM}^-$  et  $V_{I,BTM}^+$ , les bornes de l'ensemble de coupure  $\alpha = 0,5$  correspondant à l'intervalle plausible de l'indice de vulnérabilité  $V_I$  ;
- $V_{I,BTM}^{min}$  et  $V_{I,BTM}^{max}$ , les bornes de l'ensemble de coupure  $\alpha = 1$  correspondant à l'intervalle possible de l'indice de vulnérabilité  $V_I$  ;
- $V_{I,BTM}^*$ , l'indice de vulnérabilité  $V_I$  le plus probable, obtenu par *défuzzification* (conversion d'une valeur floue en valeur nette) à l'aide de la méthode du centroïde définie par Ross [26] :

$$V_{I,BTM}^* = \frac{\int \chi(V) \times V dV}{\int \chi(V) dV} \quad (11)$$



TABLE 4 – Matrices de typologie Risk-UE et indices de vulnérabilité remarquables.  
Source : RISK-UE.

Typologies	Description	$V_{I,BTM}^{min}$	$V_{I,BTM}^-$	$V_{I,BTM}^*$	$V_{I,BTM}^+$	$V_{I,BTM}^{max}$
M1.1	<i>Rubble stone, fieldstone</i>	0,62	0,81	0,873	0,98	1,02
M1.2	<i>Simple stone</i>	0,46	0,65	0,74	0,83	1,02
M1.3	<i>Massive stone</i>	0,3	0,49	0,616	0,793	0,86
M2	<i>Adobe</i>	0,62	0,687	0,84	0,98	1,02
M3.1	<i>Wooden slabs</i>	0,46	0,65	0,74	0,83	1,02
M3.2	<i>Masonry vaults</i>	0,46	0,65	0,776	0,953	1,02
M3.3	<i>Composite steel and masonry slabs</i>	0,46	0,527	0,704	0,83	1,02
M3.4	<i>Reinforced concrete slabs</i>	0,3	0,49	0,616	0,793	0,86
M4	<i>Reinforced or confined masonry walls</i>	0,14	0,33	0,451	0,633	0,7
M5	<i>Overall strengthened</i>	0,3	0,49	0,694	0,953	1,02
RC1	<i>Concrete Moment Frames</i>	-0,02	0,047	0,442	0,8	1,02
RC2	<i>Concrete shear walls</i>	-0,02	0,047	0,386	0,67	0,86
RC3.1	<i>Regularly infilled walls</i>	-0,02	0,007	0,402	0,76	0,98
RC3.2	<i>Irregular frames</i>	0,06	0,127	0,522	0,88	1,02
RC4	<i>RC Dual systems (RC frame and wall)</i>	-0,02	0,047	0,386	0,67	0,86
RC5	<i>Precast Concrete Tilt-Up Walls</i>	0,14	0,207	0,384	0,51	0,7
RC6	<i>Precast C, Frames, C, shear walls</i>	0,3	0,367	0,544	0,67	0,86
S1	<i>Steel Moment Frames</i>	-0,02	0,467	0,363	0,64	0,86
S2	<i>Steel braced Frames</i>	-0,02	0,467	0,287	0,48	0,7
S3	<i>Steel frame+unreinf, mas, infill walls</i>	0,14	0,33	0,484	0,64	0,86
S4	<i>Steel frame+cast-in-place shear walls</i>	-0,02	0,047	0,224	0,35	0,54
S5	<i>Steel and RC composite system</i>	-0,02	0,257	0,402	0,72	1,02
W	<i>Wood structures</i>	0,14	0,207	0,447	0,64	0,86

Milutinovic et Trendafiloski [17] précisent que les structures en acier (S1-S5), utilisées dans un contexte hors industriel, sont assez rares en Europe. Lorsqu'elles sont utilisées, il s'agit de bâtiments caractérisés par une hauteur hors norme pour lesquels RISK-UE n'est pas adapté. De même, les structures en bois (W) ou en adobe (M2) sont exceptionnellement rares dans les zones urbaines. Celles qui existent sont utilisées soit pour des structures temporaires, soit pour des structures à fonction auxiliaire ou sont complètement abandonnées. Pour les classes de bâtiments en acier (S1-S5), en bois (W) et en maçonnerie confinée (M4), ils recommandent d'utiliser plutôt la méthodologie HAZUS.

### III.2.2 Estimation du niveau de dommage moyen $\mu_D$

Pour la mise en œuvre opérationnelle de la méthodologie, RISK-UE définit le taux de dommage moyen  $\mu_D$  tel que :

$$\mu_D = 2.5 \left( 1 + \tanh \left( \frac{I + 6.25 \times V_I - 13.1}{2.3} \right) \right) \quad (12)$$

avec :

- $V_I$  l'indice de vulnérabilité ;
- $I$  l'intensité macrosismique.

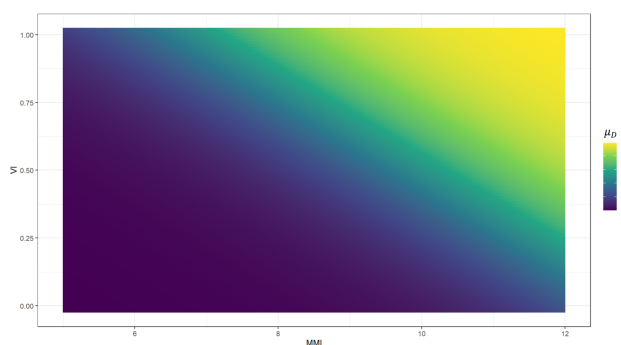


FIGURE 27 – Représentation de  $\mu_D$  en fonction de l'intensité macrosismique (MMI) et de l'indice de vulnérabilité (VI).

En particulier, connaissant l'indice de vulnérabilité le plus probable de chaque structure, il est possible de tracer les fonctions de vulnérabilité par structure :

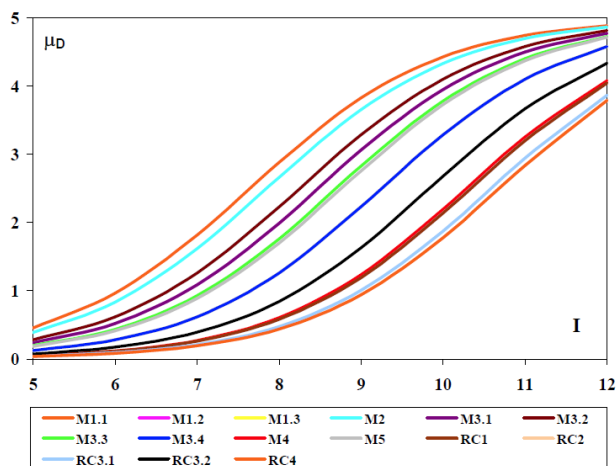


FIGURE 28 – Fonctions de vulnérabilité par structure. Source : RISK-UE.

### III.2.3 Estimation des distributions de dommage

Pour déterminer la distribution discrète des niveaux de dommage, deux distributions peuvent être considérées : la distribution binomiale utilisée par Braga et al. [3] et la distribution bêta préférée par Lagomarsino et Giovinazzi.

#### III.2.3.1 Modélisation par loi binomiale $Bin(5, \frac{\mu_D}{5})$

Dans le cadre du modèle binomial, en notant  $\mu_D$  le niveau de dommage moyen,  $p_k$  la probabilité d'être en niveau de dommage  $k \in \{0, 1, \dots, 5\}$  et  $\sigma_D$  l'écart type de la distribution, il faut considérer les relations suivantes :

$$\text{Fonction de masse : } p_k = \frac{5!}{k!(5-k)!} \left(\frac{\mu_D}{5}\right)^k \left(1 - \frac{\mu_D}{5}\right)^{5-k} \quad (13)$$

$$\sigma_D = \sqrt{\mu_D \left(1 - \frac{\mu_D}{5}\right)} \quad (14)$$

Toutefois, la simplicité de cette distribution, qui ne dépend que d'un seul paramètre, ne permet pas de définir la dispersion des niveaux de dommages autour de la valeur moyenne.

#### III.2.3.2 Modélisation par loi bêta à quatre paramètres $Beta(r, t-r, a, b)$

Sandi et Floricel [27] ont montré que la dispersion de la distribution binomiale est trop élevée lorsque l'on considère une classification détaillée des bâtiments, ce qui est le cas de la typologie développée dans RISK-UE. Selon Lagomarsino et Giovinazzi [10], cela peut conduire à une surestimation du nombre de bâtiments subissant des dommages importants dans le cas de valeurs pourtant plutôt faibles de niveaux de dommage moyens. La distribution bêta semble donc plus adaptée car plus flexible.

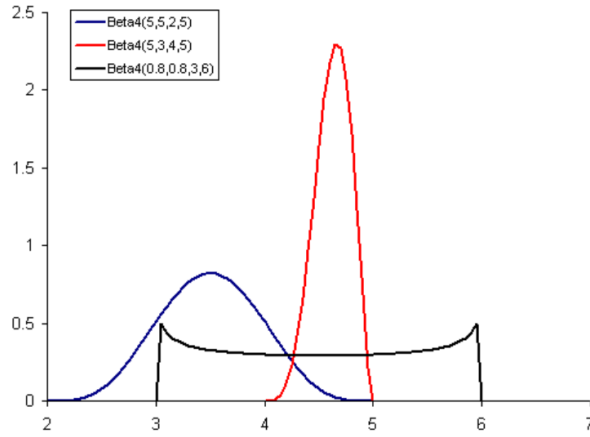


FIGURE 29 – Quelques formes de loi bêta à quatre paramètres.

Dans le cadre du modèle bêta, il faut donc considérer les relations suivantes où  $a$ ,  $b$ ,  $t$  et  $r$  sont les paramètres de la loi bêta et où  $\mu_X$  et  $\sigma_X$  sont respectivement l'espérance et l'écart type de la variable continue  $X$  suivant une telle distribution :

$$\text{Densité : } p_\beta(x) = \frac{\Gamma(t)}{\Gamma(r)\Gamma(t-r)} \frac{(x-a)^{r-1}(b-x)^{t-r-1}}{(b-a)^{t-1}} \quad a \leq x < b \quad (15)$$

$$\text{Fonction de répartition : } P_\beta(x) = \int_a^x p_\beta(\epsilon) d\epsilon \quad (16)$$

$$\mu_X = a + \frac{r}{t}(b-a) \quad (17)$$

$$\sigma_X = \sqrt{\frac{r(t-r)}{t^2(t+1)}} |b-a| \quad (18)$$

Pour utiliser la distribution bêta, il est nécessaire de se référer au niveau de dommage, qui est une variable discrète à 6 niveaux. Il est donc conseillé de choisir  $a = 0$  et  $b = 6$ . Dès lors :

$$p_k = P_\beta(k+1) - P_\beta(k) \quad (19)$$

D'après Lagomarsino et Giovinazzi [10], l'espérance de la distribution de dommage discrète  $\mu_D$ , qui est aussi le niveau de dommage moyen défini en (12), et l'espérance de la loi bêta  $\mu_X$  peuvent alors être corrélés par un polynôme de degré 3 :

$$\mu_x = 0,042 \times \mu_D^3 - 0,315 \times \mu_D^2 + 1,725 \times \mu_D \quad (20)$$

Les paramètres  $r$  et  $t$  de la distribution bêta sont alors liés avec le niveau de dommage moyen  $\mu_D$  par la relation suivante :

$$r = t (0,007 \times \mu_D^3 - 0,052 \times \mu_D^2 + 0,2875 \times \mu_D) \quad (21)$$

Il est noté dans la méthodologie RISK-UE que si  $t = 8$ , la distribution bêta ressemble à la distribution binomiale. Enfin, la courbe de fragilité définissant la probabilité d'atteindre ou de dépasser un certain niveau de dommage est obtenue directement à partir de la probabilité cumulée de la loi bêta :

$$P(D \geq D_k) = 1 - P_\beta(k) \quad (22)$$

### III.3 Une méthode pour déterminer un unique niveau de dommage

Au cours de cette section, nous avons présenté le module de vulnérabilité et de dommage RISK-UE LM1 qui permet, à partir des caractéristiques d'un bâtiment et d'une intensité macrosismique, de déterminer une distribution de probabilité de dommage.

TABLE 5 – Distribution de probabilités de dommage dans le cas d'un bâtiment d'indice de vulnérabilité 0,7 soumis à une intensité macrosismique VII, obtenue par loi bêta à quatre paramètres.

DS0	DS1	DS2	DS3	DS4	DS5
40,263%	38,867%	16,453%	3,991%	0,420%	0,007%

Dès lors, il convient de déterminer un moyen de passer de cette distribution de probabilités à une estimation de perte unique. L'approche retenue dans cette étude, basée sur la méthode de la transformée inverse, est la suivante :

1. Nous commençons par tirer uniformément un nombre entre 0 et 1 inclus. A titre d'exemple, supposons pour la suite que 0,90 soit le nombre tiré ;

2. Ensuite, il faut cumuler la distribution de probabilités de dommage. Toujours pour l'exemple, considérons la distribution de la table précédente, caractéristique d'un bâtiment d'indice de vulnérabilité 0,7 soumis à une intensité macrosismique VII. On obtient alors :

TABLE 6 – Distribution de probabilités de dommages cumulées dans le cas d'un bâtiment d'indice de vulnérabilité 0,7 soumis à une intensité macrosismique VII, obtenue par loi bêta à quatre paramètres.

DS0	DS1	DS2	DS3	DS4	DS5
40,263%	79,129%	95,582%	99,573%	99,993%	100%

3. Enfin, il convient de déterminer le premier niveau de dommage dont la probabilité cumulée dépasse le nombre précédemment tiré. Dans notre exemple :

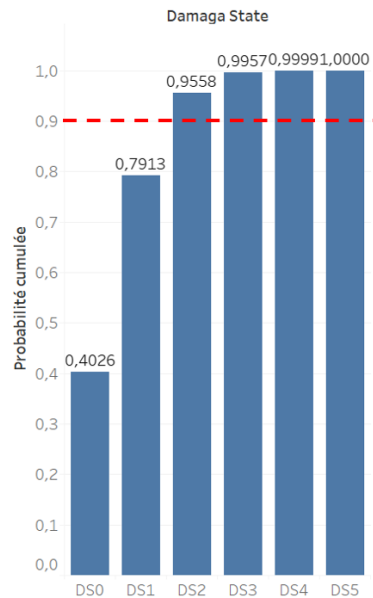


FIGURE 30 – Probabilités cumulées des niveaux de dommage dans le cas d'un bâtiment d'indice de vulnérabilité 0,7 soumis à une intensité macrosismique VII. La ligne rouge pointillée correspond au 90ème centile.

Ainsi, dans l'exemple précédent, on estime qu'un bâtiment d'indice de vulnérabilité 0,7 soumis à une intensité macrosismique VII se trouvera dans en *damage state* 2. Dès lors, il est possible d'estimer une perte brute à l'aide d'une relation dommage-coût, dit *loss ratio*. Pour diminuer la volatilité de cette estimation, une moyenne est ensuite réalisée sur 100 observations de *loss ratio* simulé.

**Partie 3 :**  
**Les sources d'incertitude dans**  
**l'estimation des pertes**

## I Les incertitudes de type « Modèle »

Ce premier type d'incertitude provient des difficultés à choisir le type de modèle approprié. Cela concerne notamment le choix de l'ensemble de *loss ratios*, de la structure du bâtiment et du modèle de distribution de pertes.

### I.1 Loss Ratios, LR

Certaines méthodologies, comme le modèle sismique HAZUS-MH MR5 développé par la *Federal Emergency Management Agency*, indiquent directement les relations dommage-coût à utiliser selon les caractéristiques du bâtiment assuré. Ce n'est cependant pas le cas de la méthode RISK-UE LM1 qui n'en fournit aucun.

Il devient alors nécessaire de choisir un ensemble de *loss ratios* pertinent. Toutefois, il peut être difficile de s'y retrouver devant l'abondance d'études existantes. La table ci-dessous présente un exemple de 9 relations dommage-coût déterminées à l'échelle d'un pays européen.

TABLE 7 – Revue de la littérature des différentes relations dommage-coût calibrées dans l'espace européen. Les auteurs et la source font respectivement référence aux scientifiques qui ont développé la relation et à ceux qui ont écrit l'article dont les valeurs sont extraites. Les *loss ratios* indiqués correspondent aux valeurs centrales. Source : A. Pothon.

Auteurs	Source (quand elle diffère des auteurs)	Zone d'étude	DS0	DS1	DS2	DS3	DS4	DS5
Meroni et al. (2017)	-	Italie	0	0,05	0,2	0,45	1	1
Riedel (2015)	-	France	0	0,03	0,14	0,34	0,65	0,9
Eleftheriadou and Karabinis (2008)	-	Grèce	0	0,005	0,15	0,65	1	1
Roca et al. (2006)	Hill and Rossetto (2008a)	Espagne	0	0,01	0,2	0,4	0,8	1
Kappos et al. (2006)	-	Grèce	0	0,01	0,09	0,29	0,63	0,77
Di Pasquale et al. (2005)	Hill and Rossetto (2008a)	Italie	0	0,01	0,1	0,35	0,75	1
Tyagunov et al. (2004)	-	Allemagne	0	0,005	0,1	0,4	0,8	1
Milutinovic et Trendafiloski (2003)	-	Europe	0	0,03	0,15	0,5	1	1
Di Pasquale et al. (2001)	Riedel (2015)	Italie	0	0,04	0,22	0,41	0,78	0,81

Par exemple les *loss ratios* définis par Meroni et al. (2017) s'interprètent comme suit :

- En DS0 : le coût de reconstruction correspond à 0% de la valeur assurée du bâtiment ;
- En DS1 : le coût de reconstruction correspond à 5% de la valeur assurée du bâtiment ;



- En DS2 : le coût de reconstruction correspond à 20% de la valeur assurée du bâtiment ;
- Etc.

Excepté Milutinovic et Trendafiloski (2003), toutes ces relations ont été calibrées à l'échelle d'un pays européen spécifique. Toutefois, il est parfois supposé qu'un ensemble de *loss ratios* peut être utilisé sans tenir compte de la zone dans laquelle il a été développé. En effet, d'après Riedel et Guéguen [23], cette hypothèse forte est souvent utilisée du fait de l'absence de relations là où les pertes dues aux séismes sont limitées : les pays à faible sismicité ou en développement.

Toutefois, il n'est pas difficile de remarquer des différences significatives entre ces relations. Elles apparaissent d'ailleurs d'autant plus clairement lorsqu'on les compare, pour chaque *damage state*, à un ensemble de *loss ratios* de référence.

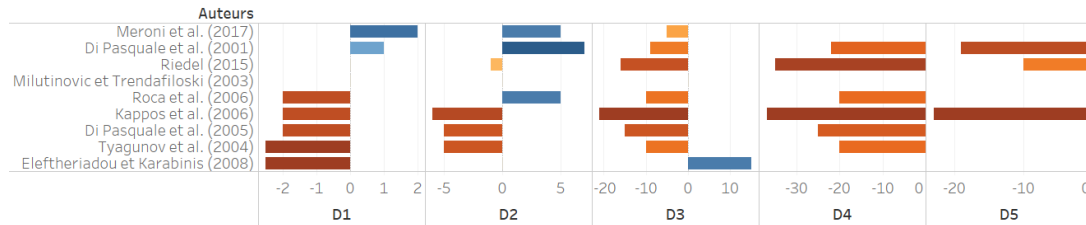


FIGURE 31 – Comparaison des relations dommage-coût de la littérature par rapport à Milutinovic et Trendafiloski (2003). *Lecture : en damage state 1, la perte déduite de Meroni et al. (2017) est 2% plus élevée que celle de Milutinovic et Trendafiloski (2003), toutes choses égales par ailleurs. A l'inverse, en damage state 5, la perte déduite de Kappos et al. (2006) est 23% plus faible que celle de Milutinovic et Trendafiloski (2003), toutes choses égales par ailleurs.*

On remarque ainsi que, toutes choses égales par ailleurs, la différence de perte entre Kappos et al. (2006) et Milutinovic et Trendafiloski (2003) peut aller jusqu'à 37% en *damage state 4*. Dans un portefeuille où la valeur assurés des biens assurés peut s'élever à plusieurs centaines de millions d'euros, il n'y a aucun doute sur le fait qu'une perte de 63% ou de 100% ne représentera pas la même charge pour l'assureur.

## I.2 Structure du bâtiment, STR

D'après le manuel Risk-UE [17], outre la présence remarquable de diverses structures en maçonnerie, les bâtiments en béton armé sont largement surreprésentés en Europe, et ce, depuis plusieurs décennies. Dans certaines zones urbaines, ils ont d'ailleurs complètement remplacé les bâtiments en maçonnerie.

Dès lors, pour la suite de notre étude, nous avons choisi de conserver trois structures en béton armé (*reinforced concrete*). Les trois structures étudiées sont :

1. Construction avec une ossature sans conception parasismique (*Frame without earthquake-resitant design*). Cette structure sera dorénavant associé au sigle *RC1* ;
2. Construction avec une ossature de niveau de conception parasismique moyen (*Frame with moderate level of earthquake-resitant design*). Cette structure sera dorénavant associé au sigle *RC2* ;
3. Construction avec ossature de niveau de conception parasismique élevé (*Frame with high level of earthquake-resitant design*). Cette structure sera dorénavant associé au sigle *RC3*.

Comme l'on connaît la fonction d'appartenance de ces structures aux classes de vulnérabilité, on peut tracer les fonctions d'appartenance de l'indice de vulnérabilité de chaque structure.

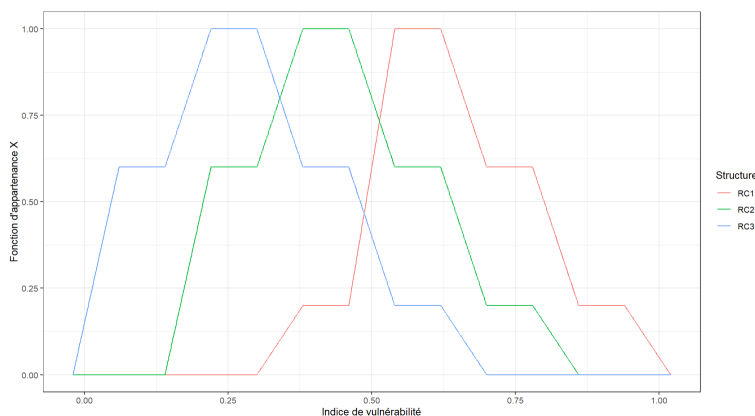


FIGURE 32 – Fonctions d'appartenance des indices de vulnérabilité des structures *RC1*, *RC2* et *RC3*.

On observe un déplacement vers la gauche des fonctions d'appartenance avec l'augmentation du niveau de conception parasismique (CPS). Cela fait sens car un bâtiment de CPS élevé sera moins vulnérable aux mouvements du sol.

Par *défuzzification* à l'aide de la méthode du centroïde, on obtient l'indice de vulnérabilité le plus probable de chaque structure :

TABLE 8 – Indice de vulnérabilité le plus probable de chaque structure.

Structure	<i>RC1</i>	<i>RC2</i>	<i>RC3</i>
$V_{I,BTM}^*$	0,644	0,447	0,287

Dès lors, à l'aide des indices  $V_{I,BTM}^*$ , on peut obtenir par l'équation 14 les fonctions de vulnérabilité les plus probables de chacune de ces structures.

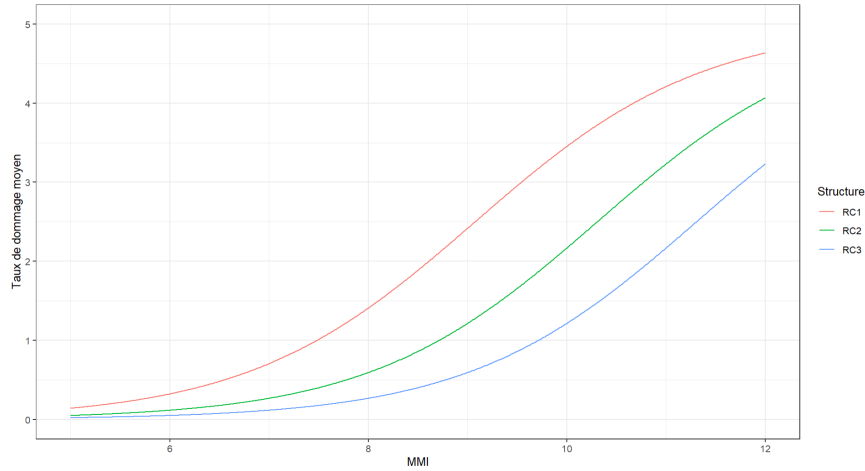


FIGURE 33 – Fonctions de vulnérabilité les plus probables des structures  $RC1$ ,  $RC2$  et  $RC3$ .

Le tracé de ces fonctions de vulnérabilité permet d'appréhender directement l'importance assurantielle de la distinction en niveau de conception parasismique. En effet, on observe à certains niveaux d'intensité macrosismique un écart significatif entre les niveaux de dommage moyens les plus probables de chaque structure.

TABLE 9 – Niveaux de dommage moyen  $\mu_D$  des structures  $RC1$ ,  $RC2$  et  $RC3$  estimés en considérant une intensité macrosismique  $X$  et l'indice de vulnérabilité le plus probable  $V_I^*$  de chaque structure.

Structure	$RC1$	$RC2$	$RC3$
$\mu_D$	3,45	2,17	1,21

Cette distinction peut donc avoir un impact majeur pour l'assureur lorsque ces niveaux de dommages sont associés à une relation dommage-coût comme Eleftheriadou et Karabinis (2008) où les pertes brutes en niveau de dommage 1 et 3 passent respectivement de 0,5% à 65% de la valeur assurée.

### I.3 Modèle de distribution de pertes, MOD

Nous avons décrit comment utiliser deux lois spécifiques pour obtenir une distribution de probabilités de dommage : la loi binomiale  $Bin(5, \frac{\mu_D}{5})$  et la loi bêta à quatre paramètres  $Beta(r, t-r, a, b)$ . Avec un choix de paramètres spécifiques, Milutinovic et Trendafiloski [17] indiquent que la distribution de la loi bêta est très similaire à celle de la loi binomiale. Toutefois, il est raisonnable de se demander dans quelle mesure cette remarque est vérifiée.

A titre d'exemple, déterminons la distribution de probabilités associée à un niveau 2 de dommage moyen. Par la méthode explicitée dans la partie précédente, nous obtenons la distribution suivante :

TABLE 10 – Distribution de probabilités associée à un niveau 2 de dommage moyen. La ligne « Delta » représente la différence entre le modèle binomial et la modèle bêta.

	DS0	DS1	DS2	DS3	DS4	DS5	Total
Binomial	7,776%	25,920%	34,560%	23,040%	7,680%	1,024%	100%
Bêta	5,280%	26,478%	35,973%	24,245%	7,532%	0,491%	100%
Delta	2,496%	-0,558%	-1,413%	-1,205%	0,148%	0,533%	0%

Pour un niveau 2 de dommage moyen, la table précédente permet deux observations intéressantes :

- La plus grande différence positive entre les probabilités issues du modèle binomial et bêta est atteinte en *damage state* 0. En effet :

$$P^{Bin}(DS0|\mu_D) - P^{Beta}(DS0|\mu_D) = 2,496\%$$

- La plus grande différence négative entre les probabilités issues du modèle binomial et bêta est atteinte en *damage state* 2. En effet :

$$P^{Bin}(DS2|\mu_D) - P^{Beta}(DS2|\mu_D) = -1,413\%$$

En effectuant cette procédure pour chaque niveau de dommage moyen  $\mu_D$  entre 0 et 5, on obtient la figure suivante :

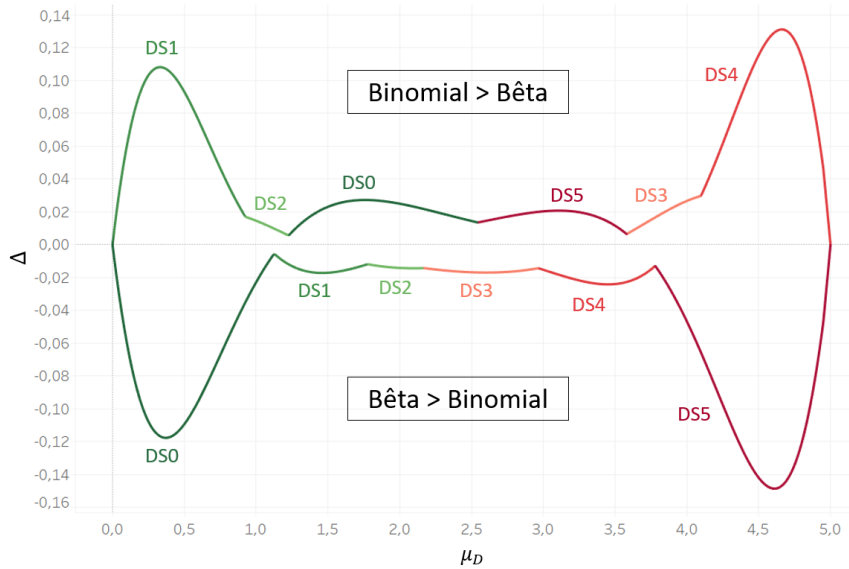


FIGURE 34 – Comparaison des lois binomiale et bêta.

Ainsi, la distribution de dommage issue de la loi bêta est très similaire à celle de la loi binomiale, mais seulement pour les niveaux de dommage moyens intermédiaires compris entre 1 et 4. Pour les niveaux de dommage moyens inférieurs à 1 ou supérieur à 4, on observe des différences significatives qui peuvent s'élever jusqu'à 15%.

## II Les incertitudes de type « Paramètre »

Ce second type d'incertitude provient des difficultés à estimer les valeurs des paramètres du modèle d'entrée. Cela concerne notamment la détermination de l'ensemble de l'intensité macrosismique et de l'indice de vulnérabilité de la structure du bâtiment.

### II.1 Intensité macrosismique, MMI

La modélisation des risques sismiques repose principalement sur l'utilisation d'équations de prédiction du mouvement du sol (GMPE). Preuve de la difficulté de résumer la propagation d'ondes sismiques par une équation, John Douglas [6] recense 53 équations différentes publiées entre 1964 et 2021 pour la mesure de l'intensité macrosismique (MMI) et 480 pour la seule mesure de l'accélération maximale au sol (PGA).

Toutefois, comme le montre les figures ci-dessous, l'écart-type du terme d'erreur (défini comme la différence entre la valeur observée et la valeur modélisée) reste constant malgré un nombre grandissant d'observations et donc de variables explicatives pour calculer le mouvement du sol.

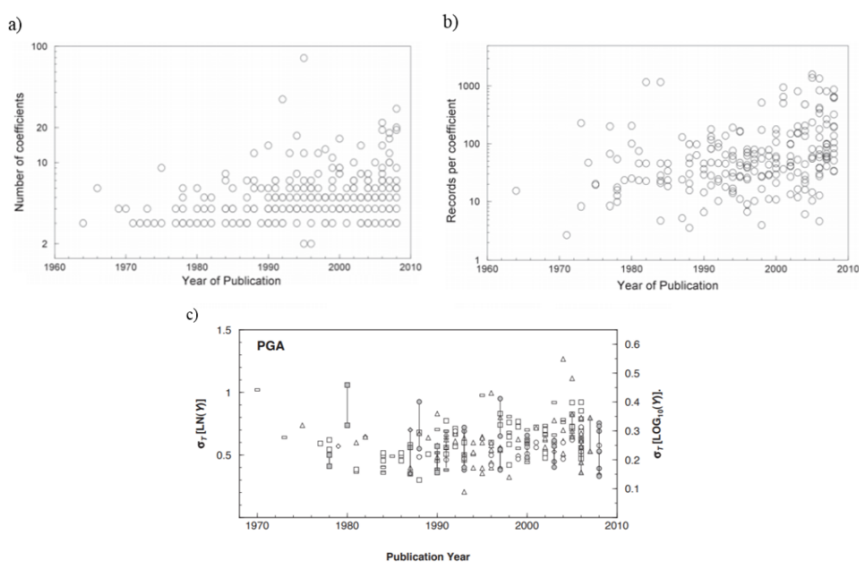


FIGURE 35 – Évolution temporelle des équations de prédiction du mouvement du sol (GMPE) en fonction : a) du nombre de coefficients dans l'équation ; du nombre de données par coefficients utilisées pour calibrer l'équation et c) de l'écart-type ( $\sigma_T [LN(Y)]$ ) de l'erreur de modélisation. Sources individuelles : figures a) et b) : Bommer et al. (2010) ; Figure c) : Strasser et al. (2009). Source générale : A. Pothon.

A titre d'exemple, observons les données de l'Institut d'études géologiques des États-Unis (*United States Geological Survey, USGS*) concernant le séisme en Italie du 24 août 2016. Ce tremblement de terre d'une magnitude 6,2 a secoué la haute vallée du Tronto, notamment la province de Rieti, celle d'Ascoli Piceno et, dans une moindre mesure, les régions voisines de l'Ombrie et des Abruzzes, en Italie centrale.

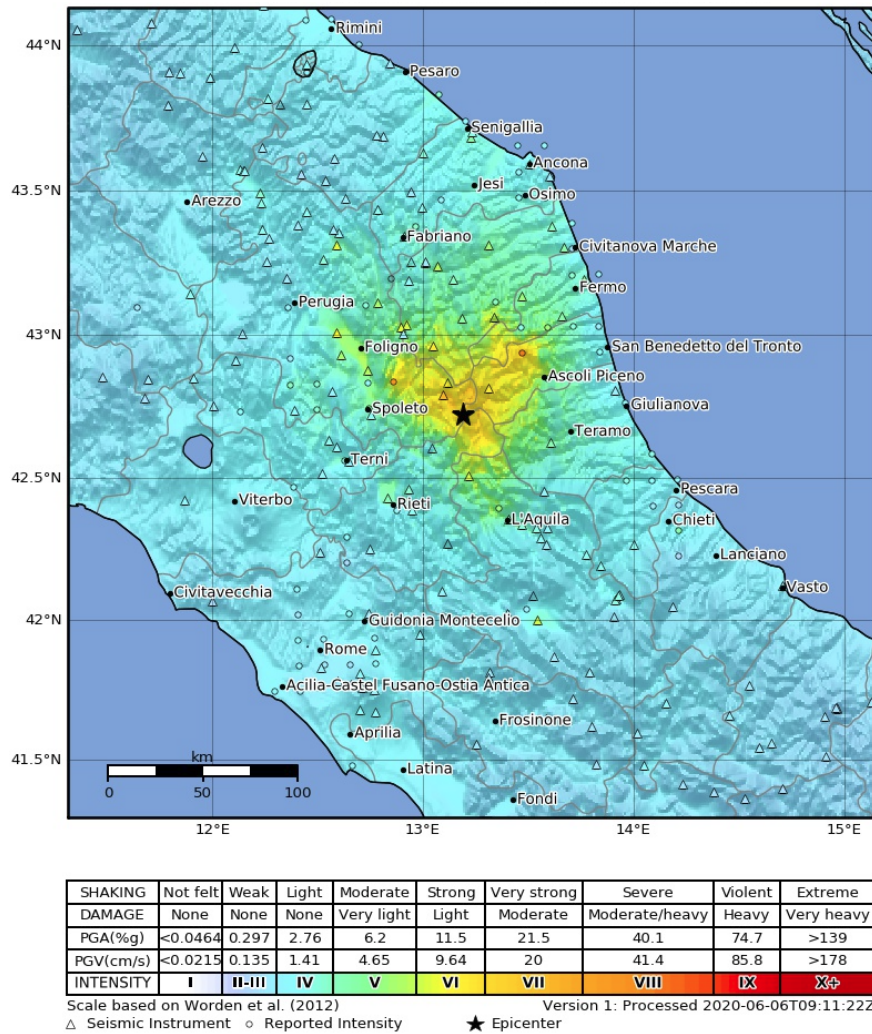


FIGURE 36 – Carte de l'intensité macrosismique du séisme du 24 août 2016 en Italie. Source : USGS.

Pour chacune des grandeurs mesurées, l'USGS associe une mesure de l'incertitude. Ci-dessous est notamment représenté l'écart type du terme d'erreur associé à l'intensité macrosismique.

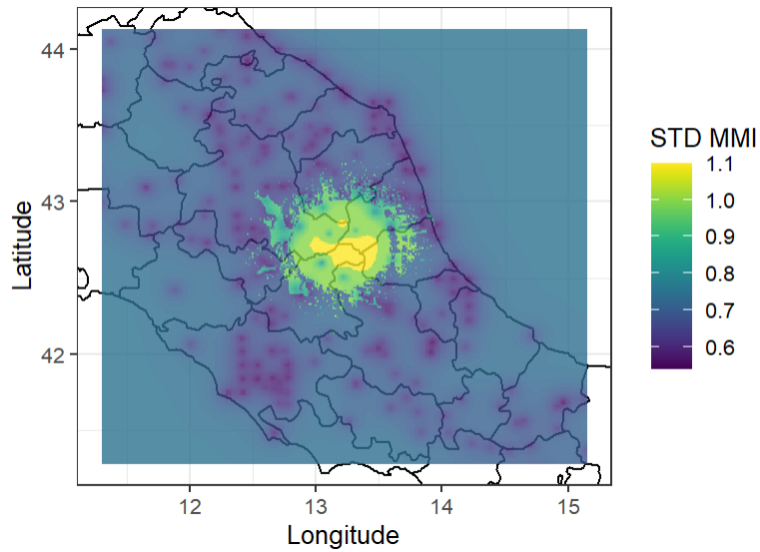


FIGURE 37 – Carte de l'écart type du terme d'erreur associé à l'intensité macrosismique du séisme du 24 août 2016 en Italie.

En moyenne, l'écart type du terme d'erreur associé à l'intensité macrosismique du séisme du 24 août 2016 en Italie est de l'ordre de 0,7 ce qui est cohérent avec les données de Strasser et al. [32]. C'est donc cette valeur qui sera retenue dans la suite de l'étude pour la modélisation de l'intensité macrosismique.

En supposant que la véritable valeur de la MMI en un point soit VIII, celle-ci sera alors représentée par une loi normale  $\mathcal{N}(8, 0, 7^2)$ . Notons qu'en 2014, Rohmer et al. [25] avaient également fait le choix de modéliser l'incertitude de la MMI par une loi normale.

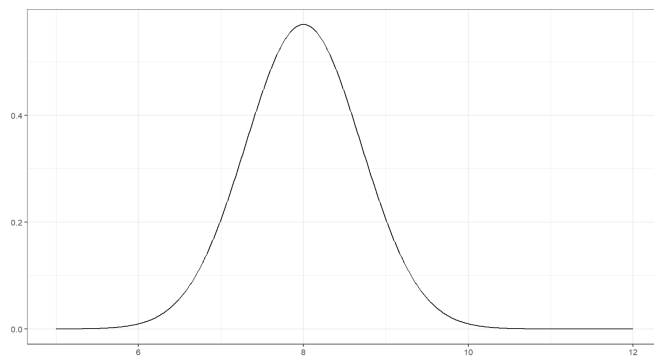


FIGURE 38 – Densité de la loi normale  $\mathcal{N}(8, 0, 7^2)$ .



## II.2 Indice de vulnérabilité, VI

Dans les sections précédentes de cette étude, nous avons montré la place centrale qu’occupait la détermination de l’indice de vulnérabilité d’une structure au sein de la méthodologie RISK-UE. En effet, l’indice de vulnérabilité étant directement lié, avec l’intensité macrosismique, au niveau de dommage moyen  $\mu_D$  d’un bâtiment par la relation (12), la précision de son estimation est un enjeu majeur.

Toutefois, l’étendue des intervalles de valeurs possibles pour l’indice de vulnérabilité de chaque structure atteste toute la difficulté de résumer en pratique la vulnérabilité d’un bâtiment à sa seule structure. Par exemple, la structure dénommée *RC2* est caractérisée par des indices de vulnérabilité minimal et maximal respectifs à 0,15 et 0,85, soit la quasi totalité de l’intervalle des valeurs possibles. Intuitivement, on peut alors penser a priori que la grande incertitude associée à l’indice de vulnérabilité d’un bâtiment sera celle qui, parmi tous les autres paramètres, aura l’impact le plus grand.

Le figure suivante illustre en effet qu’à une intensité macrosismique  $X$ , un bâtiment de structure RC2 d’indice de vulnérabilité 0,15 subira un niveau de dommage moyen égal à 0,66. Dans le même temps, un bâtiment de structure RC2 d’indice de vulnérabilité 0,85 subira un niveau de dommage moyen égal à 4,36, soit un dommage moyen environ 6,6 fois plus important.

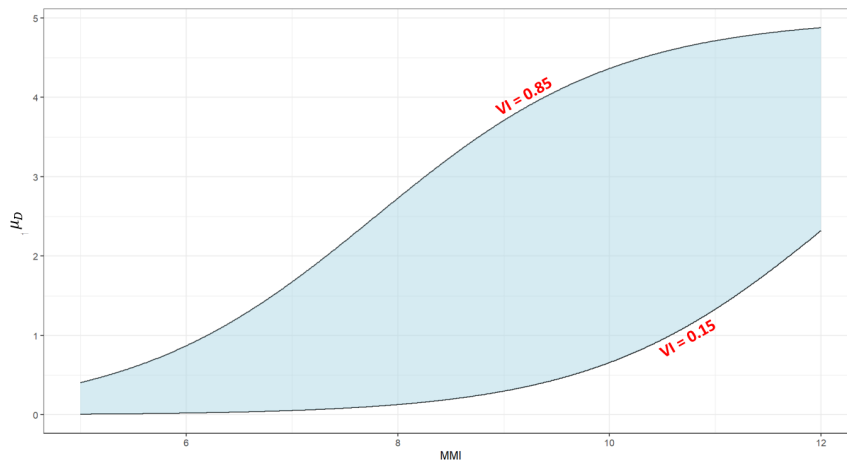


FIGURE 39 – Ensemble des courbes de vulnérabilité possibles de la structure RC2. Les courbes supérieure et inférieure noires sont respectivement associées à des indices de vulnérabilité égaux à 0,85 et 0,15.

Dès lors, il convient de déterminer comment sera modélisé l'indice de vulnérabilité d'un bâtiment dans la suite de l'étude. Le choix a été fait de réaliser directement un tirage dans la distribution des indices de vulnérabilité de chaque structure en passant par leurs fonctions d'appartenance. Soit  $X$  une variable aléatoire de densité  $f$  représentant l'indice de vulnérabilité d'une structure quelconque, alors :

$$f = \frac{\chi(V)}{\mathcal{A}} \quad (23)$$

Avec :

- $V$  l'indice de vulnérabilité ;
- $\chi(V)$  la valeur de la fonction d'appartenance en  $V$  ;
- $\mathcal{A}$  l'aire sous la courbe tel que  $\mathcal{A} = \int \chi(V) dV$ .

En particulier, la valeur de  $\mathcal{A}$  pour chacune des structures considérées est telle que :

TABLE 11 – Aire sous la courbe  $\mathcal{A}$  des fonctions d'appartenances associées à la vulnérabilité des structures  $RC1$ ,  $RC2$  et  $RC3$ .

Structure	$RC1$	$RC2$	$RC3$
Aire sous la courbe	0,32	0,384	0,384

Dès lors, les fonctions de densité des variables aléatoires représentant les indices de vulnérabilité des structures se déduisent complètement :

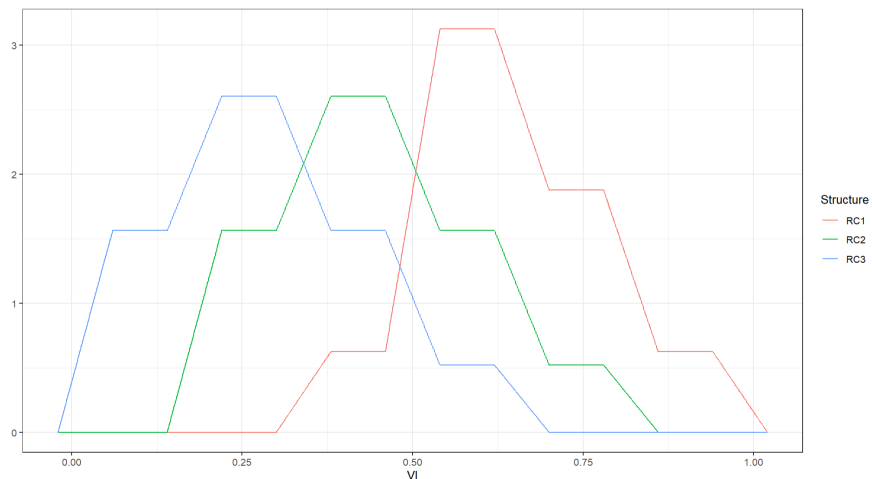


FIGURE 40 – Fonctions de densité des variables aléatoires représentant les indices de vulnérabilité des structures  $RC1$ ,  $RC2$  et  $RC3$ .

Ces variables à densité seront simulées dans la suite de l'étude par la méthode de rejet.

### Méthode de rejet

On voudrait simuler une variable aléatoire réelle  $X$  de densité de probabilité  $f$ . On suppose :

- Qu'il existe une autre densité de probabilité  $g$  telle que le ratio  $\frac{f}{g}$  soit borné, disons par  $c$  (i.e.  $f \leq c \times g$ )
- Que l'on sache simuler  $Y$  de densité  $g$ .

Dès lors, la version la plus simple de la méthode de rejet prend la forme suivante :

1. Boucler :
  - Tirer  $Y$  de densité  $g$  ;
  - Tirer  $U$  selon la loi uniforme  $\mathcal{U}(0, 1)$ , indépendamment de  $Y$  ;
2. Tant que  $U > \frac{f(Y)}{cg(Y)}$ , reprendre l'étape 1 ;
3. Accepter  $Y$  comme un tirage aléatoire de densité de probabilité  $f$ .

Partie 4 :  
Analyse globale de sensibilité

# I Constitution de la base de données

Entreprendre une analyse de sensibilité nécessite d'abord de constituer une base de données dont la qualité est essentielle afin d'obtenir des résultats significatifs et interprétables. Pour y parvenir, il est nécessaire de définir certaines hypothèses pour représenter les incertitudes des variables d'entrée du modèle considéré.

## I.1 Représentation des sources d'incertitude

En reprenant la méthodologie utilisée en 2014 par Rohmer et al. [25], ces incertitudes seront représentées comme suit :

TABLE 12 – Description des sources d'incertitude considérées et hypothèses de représentation dans le cadre de l'analyse de sensibilité.

Source d'incertitude	Type	Description	Représentation
<b>MOD</b>	Modèle	Modèle de distribution de probabilités	Variable discrète prenant uniformément ses valeurs dans $\{1, 2\}$
<b>STR</b>	Modèle	Structure du bâtiment	Variable discrète prenant uniformément ses valeurs dans $\{1, 2, 3\}$
<b>LR</b>	Modèle	Relation Dommage - Coût	Variable discrète prenant uniformément ses valeurs dans $\{1, \dots, 9\}$
<b>MMI</b>	Paramètre	Intensité macrosismique	Loi normale d'espérance 8 et d'écart type 0,7
<b>VI</b>	Paramètre	Indice de vulnérabilité	Variable continue suivant une densité « logique floue »

En particulier, les trois tables suivantes présentent les modalités des trois variables catégorielles de type « Modèle » : **MOD**, **STR** et **LR**.

TABLE 13 – Modalités de la variable catégorielles **MOD** qui représente le choix du modèle permettant d'obtenir la distribution de probabilités de dommage.

Modalités	Description
<i>MOD_1</i>	Modèle Binomial
<i>MOD_2</i>	Modèle Bêta

TABLE 14 – Modalités de la variable catégorielles **STR** qui représente le type de structure du bâtiment considéré.

Modalités	Description
<i>STR_1</i>	Construction avec une ossature sans conception parasismique ( <i>RC1</i> )
<i>STR_2</i>	Construction avec une ossature de niveau de conception parasismique moyen ( <i>RC2</i> )
<i>STR_3</i>	Construction avec ossature de niveau de conception parasismique élevé ( <i>RC3</i> )

TABLE 15 – Modalités de la variable catégorielles **LR** qui représente le choix de l'ensemble de relations dommage-coût.

Modalités	Description
<i>LR_1</i>	Meroni et al. (2017)
<i>LR_2</i>	Riedel (2015)
<i>LR_3</i>	Eleftheriadou and Karabinis (2008)
<i>LR_4</i>	Roca et al. (2006)
<i>LR_5</i>	Kappos et al. (2006)
<i>LR_6</i>	Di Pasquale et al. (2005)
<i>LR_7</i>	Tyagunov et al. (2004)
<i>LR_8</i>	Milutinovic et Trendafiloski (2003)
<i>LR_9</i>	Di Pasquale et al. (2001)

## I.2 Méthode d'échantillonnage

### I.2.1 L'échantillonnage par hypercube latin (*Latin Hypercube Sampling, LHS*)

Pour optimiser la convergence et la précision d'un algorithme d'approximation on recherche souvent à réaliser un échantillonnage qui représente au mieux l'espace étudié et qui répartit les échantillons de façon à capter les non-linéarités.

Introduit en 1979 par Beckman, Conover et McKay [16] pour évaluer numériquement les intégrales multiples, l'échantillonnage par hypercube latin (*Latin Hypercube Sampling, LHS*) est une méthode souvent utilisée dans le cadre d'une analyse de sensibilité. Il a notamment été utilisé dans plusieurs études comme Gilquin et al. [9] et par Tissot et Prieur [34] pour l'estimation des indices de Sobol.

Contrairement à l'échantillonnage aléatoire qui peut manquer de couvrir des zones importantes de l'espace de recherche, la méthode garantit que chaque échantillon soit, dans un espace  $\Omega$  de dimension  $d$ , le seul aligné sur les coordonnées qui définissent sa position dans chaque hyperplan de dimension  $d - 1$ .

Chaque échantillon est donc positionné en fonction de la position des précédents, afin d'assurer qu'il ne possèdent pas de coordonnées communes dans l'espace  $\Omega$ . En outre, ils sont positionnés de manière à maximiser la distance qui les sépare, ce qui améliore encore la couverture de l'espace de recherche.

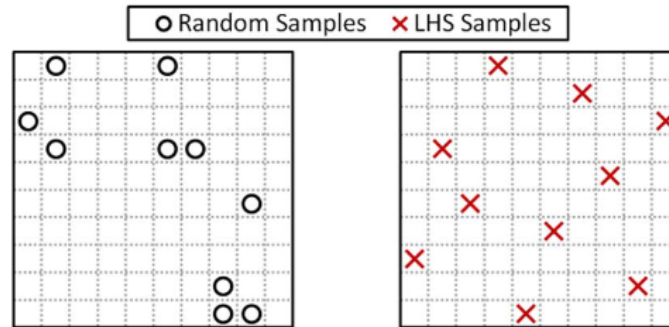


FIGURE 41 – Comparaison d'un échantillonnage aléatoire avec un échantillonnage par hypercube latin en deux dimensions. Source : Preece et Milanović.

### I.2.2 Formalisation

Un hypercube latin à  $M$  point sur  $[0, 1]^d$  est défini par l'ensemble de points  $X^i$  tel que :

$$X_j^i = \frac{\pi_j(i) + U_j^{(i)}}{M}, 1 \leq i \leq M, 1 \leq j \leq d \quad (24)$$

où :

- $\pi_j$  est une permutation de  $1, \dots, M$ ;
- $U_j^{(i)}$  est une valeur aléatoire de distribution uniforme sur  $[0, 1]$ .

Dès lors :

- $(\pi_1(i), \dots, \pi_d(i))$  représente la cellule dans laquelle se trouve le point  $X^i$ ;
- $(U_1^{(i)}, \dots, U_d^{(i)})$  désigne l'endroit où se trouve le point  $X^i$  dans cette cellule.

L'hypercube latin obtenu correspond donc à la matrice de  $n$  lignes et  $d$  colonnes à coefficients  $X_j^i$ .

### I.2.3 En pratique

Les variables **STR** et **VI** étant dépendantes, le choix a été fait de n'effectuer un échantillonnage par hypercube latin que pour les variables **MOD**, **STR**, **LR** et **MMI**. La variable **VI** sera simulée à posteriori en sachant les réalisations de **STR**.

Pour simplifier l'étude, la *véritable* valeur de l'intensité macrosismique sera supposée égale à VIII. D'après l'échelle d'intensité macrosismique, on observe à une telle intensité de nombreux effondrements partiels des bâtiments vulnérables et des dégâts modérés sur les bâtiments peu vulnérables.

Dès lors, nous allons constituer une base de données de 10 000 observations, étant un ordre de grandeur commun lorsqu'il s'agit de catalogues de risque catastrophes naturelles.

#### I.2.3.1 L'échantillonnage de la variable STR

Après l'échantillonnage par hypercube latin, nous disposons d'un ensemble de valeurs au mieux réparties dans  $[0, 1]$ . Pour passer à un ensemble représentatif de l'incertitude associée à la mesure de l'intensité macrosismique, il est nécessaire d'utiliser la fonction quantile de la loi normale d'espérance 8 et d'écart type 0,7.

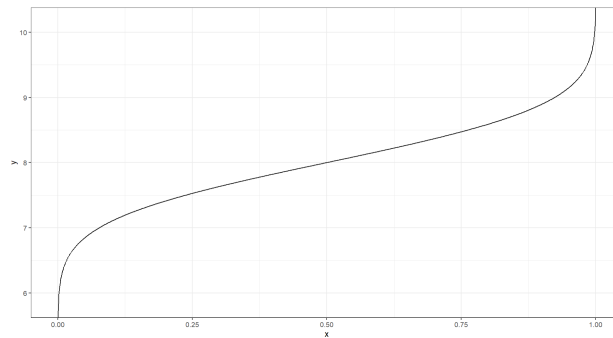


FIGURE 42 – Représentation graphique de la fonction quantile d'une loi normale d'espérance 8 et d'écart type 0,7.



### I.2.3.2 L'échantillonnage des variables catégorielles STR, MOD et LR

De nouveau, nous disposons après l'échantillonnage par hypercube latin, d'un ensemble de valeurs au mieux réparties dans  $[0, 1]$  pour chacune des variables catégorielles. Dès lors, pour passer par exemple à un ensemble représentatif de l'incertitude associée au choix du modèle de relations dommage-coût, il a été choisi d'utiliser la fonction définie par  $x \mapsto \lfloor 9x \rfloor + 1$ , la variable catégorielle **STR** possédant 9 modalités.

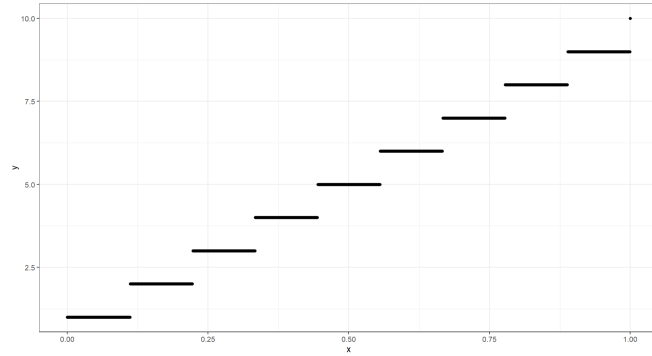


FIGURE 43 – Représentation sur  $[0, 1]$  de la fonction  $x \mapsto \lfloor 9x \rfloor + 1$ .

La probabilité d'obtenir exactement 1 lors du LHS étant négligeable, cette fonction assure des valeurs dans l'ensemble  $\{1, \dots, 9\}$  uniquement. Un raisonnement similaire sera appliqué aux autres variables catégorielles, en fonction de leur nombre de modalités.

## I.3 Statistiques descriptives des pertes observées

Après avoir réalisé l'échantillonnage des variables **STR**, **MOD**, **LR** et **MMI**, les valeurs de **VI** sont obtenues, selon chaque type de structure, par la méthode de rejet. Les différentes réalisations du vecteur de pertes **Y** sont déterminées en appliquant la méthode présentée en partie 2, comprenant le module de vulnérabilité RISK-UE.

	STR	MOD	LR	MMI	VI	Y
1	3	2	4	7.451	0.1434	0.00020
2	1	1	5	8.067	0.8869	0.34280
3	3	2	6	8.705	0.3208	0.01130
4	2	2	1	9.886	0.2740	0.09500
5	1	1	1	5.255	0.3439	0.00500
6	2	1	7	7.726	0.7109	0.12485
7	3	2	7	8.903	0.4665	0.06545
8	1	2	1	8.140	0.5399	0.07650
9	1	1	4	7.748	0.7066	0.14740
10	1	2	9	8.993	0.8916	0.70190

FIGURE 44 – Les dix premières lignes de la base de données obtenue.

À partir de ces données il est possible de déterminer certaines statistiques descriptives de  $Y$  ainsi que de tracer l'histogramme et la fonction de répartition de sa distribution.

TABLE 16 – Statistiques descriptives de  $Y$

Min	1er Quartile	Médiane	Moyenne	3ème Quartile	Max
0	0,00680	0,02850	0,08768	0,10400	0,94300

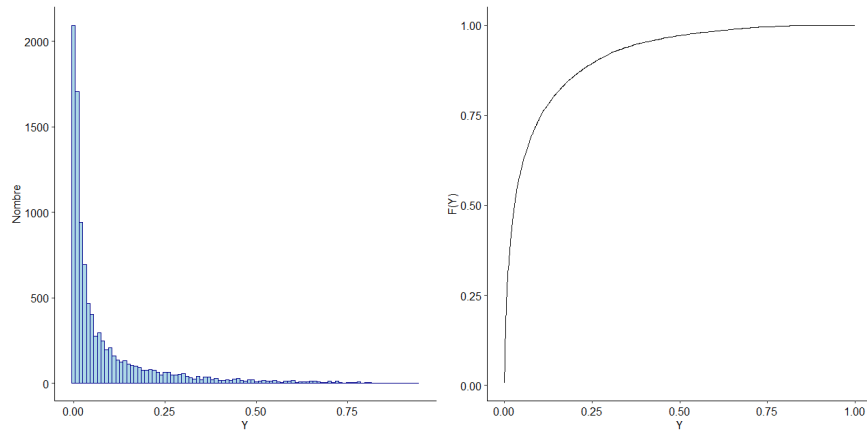


FIGURE 45 – Histogramme et fonction de répartition empirique du vecteur de pertes simulées.

La base de données étant constituée, il est alors possible de réaliser une étude de sensibilité globale. Pour contextualiser, la section suivante présentera certains concepts de décomposition fonctionnelle de la variance, de la décomposition ANOVA aux indices de Sobol. Elle justifiera ensuite pourquoi ces indices sont limités dans le cadre de notre étude où les variables d'entrée du modèle sont dépendantes et pour quelles raisons il est alors pertinent d'utiliser les indices de Shapley.

## II Décomposition fonctionnelle de la variance

Réaliser une analyse de sensibilité permet de comprendre comment l'incertitude de la sortie d'un modèle peut être attribuée à l'incertitude dans ses entrées. En tant que mesure de dispersion, la variance représente alors un moyen pertinent pour quantifier l'incertitude de la sortie d'un modèle. Pour formaliser, les notations suivantes seront utilisées dans cette section :

- $X = (X_1, \dots, X_p) \in \mathbb{R}^p$ , un vecteur de variables dont les composantes correspondent aux variables d'entrée du modèle considéré dans cette étude ;
- $Y = f(X_1, \dots, X_p)$ , la fonction de sortie du modèle ;
- $\mathcal{H}_p = \{x \in \mathbb{R} \mid 0 \leq x_i \leq 1, 1 \leq i \leq p\}$ , l'hypercube unité de dimension  $p$  ;
- $|\cdot|$ , le cardinal d'un ensemble.

### II.1 Décomposition ANOVA

Introduite en 1981 par Efron et Stein [7], la décomposition ANOVA (*ANalysis Of VAriance*) permet de décomposer une fonction  $f$  définie de  $\mathcal{H}_p$  dans  $\mathbb{R}$  par la somme de fonctions de dimension croissante. Dans la mesure où tout espace peut être transformé en  $\mathcal{H}_p$  grâce aux fonctions de répartition inverse, cette hypothèse n'est pas restrictive.

**Théorème 1** *Soit  $f$  une fonction définie de  $\mathcal{H}_p$  dans  $\mathbb{R}$ . La fonction  $f$  peut se décomposer en somme de  $2^p$  fonctions de dimensions croissantes telles que :*

$$Y = f(X) = f_0 + \sum_{i=1}^p f_i(X_i) + \sum_{1 \leq i < j \leq p} f_{ij}(X_i, X_j) + \dots + f_{1,2,\dots,p}(X_1, \dots, X_p) \quad (25)$$

avec :

- $f_0 = \mathbb{E}(Y)$ , une constante ;
- $f_i(X_i) = \mathbb{E}(Y|X_i) - \mathbb{E}(Y)$  ;
- $f_{i,j}(X_i, X_j) = \mathbb{E}(Y|X_i, X_j) - \mathbb{E}(Y|X_i) - \mathbb{E}(Y|X_j) + \mathbb{E}(Y)$  ;
- *Etc.*

Il existe une infinité de choix pour les composantes de la décomposition. Afin d'assurer son unicité, Sobol [30] ajoute des contraintes d'orthogonalité sur les composantes de la somme. Avec ces nouvelles contraintes, la décomposition ANOVA prend le nom de décomposition de Hoeffding-Sobol.

## II.2 Décomposition de Hoeffding-Sobol

La décomposition de Hoeffding-Sobol est un cas particulier de la décomposition ANOVA présentant des contraintes supplémentaires d'orthogonalité sur les composantes de la somme. Elle permet notamment d'aboutir à l'introduction des indices de Sobol.

**Théorème 2** *Soit  $f$  une fonction définie de  $\mathcal{H}_p$  dans  $\mathbb{R}$ . Il existe une unique décomposition de  $f$  de la forme :*

$$Y = f(X) = f_0 + \sum_{i=1}^p f_i(X_i) + \sum_{1 \leq i < j \leq p} f_{ij}(X_i, X_j) + \dots + f_{1,2,\dots,p}(X_1, \dots, X_p) \quad (26)$$

telle que les  $2^p - 1$  fonctions élémentaires autres que  $f_0$  soient centrées en zéro et orthogonales entre elles. Autrement dit, soient  $I, J \subseteq \mathcal{D} = 1, \dots, p$  avec  $I \neq J$  :

- $\mathbb{E}[f_I(X_I)] = 0$  ;
- $\mathbb{E}[f_I(X_I) f_J(X_J)] = 0$ .

Dès lors, en appliquant l'opérateur variance de part et d'autre de la décomposition de Hoeffding-Sobol :

$$\mathbb{V}(Y) = \sum_{i=1}^p \mathbb{V}_i + \sum_{1 \leq i < j \leq p} \mathbb{V}_{ij} + \dots + \mathbb{V}_{1,2,\dots,p} \quad (27)$$

avec :

- $\mathbb{V}_i = \mathbb{V}(f_i(X_i)) = \mathbb{V}_{X_i}(\mathbb{E}_{X_{\sim i}}[Y|X_i])$  ;
- $\mathbb{V}_{ij} = \mathbb{V}(f_{ij}(X_i, X_j)) = \mathbb{V}_{X_i, X_j}(\mathbb{E}_{X_{\sim i, j}}[Y|X_i, X_j]) - \mathbb{V}_i - \mathbb{V}_j$  ;
- Etc.

où les notations  $X_{\sim i}$  et  $X_{\sim i,j}$  sont utilisés pour désigner respectivement le vecteur contenant l'ensemble des variables de  $X = (X_1, \dots, X_p)$  sauf la variable d'entrée  $X_i$  et le vecteur contenant l'ensemble des variables de  $X = (X_1, \dots, X_p)$  sauf les variables d'entrée  $X_i$  et  $X_j$ . En reprenant ces notations, il est possible alors possible d'introduire les indices de Sobol.

### II.3 Indices de Sobol

Les indices de Sobol sont définis comme :

$$S_i = \frac{\mathbb{V}_i}{\mathbb{V}(Y)}, S_{ij} = \frac{\mathbb{V}_{ij}}{\mathbb{V}(Y)}, \dots \quad (28)$$

où  $S_i$  est l'effet de premier ordre de  $X_i$ ,  $S_{ij}$  est l'effet de second ordre de  $(X_i, X_j)$ , etc. En particulier,  $S_{ij}$  représente la contribution de l'interaction entre  $X_i$  et  $X_j$  sur l'incertitude de la variable de sortie sans tenir compte de leurs effets individuels. De plus, si la condition d'orthogonalité est respecté, on obtient que :

$$\sum_{i=1}^p S_i + \sum_i \sum_{j>i} S_{ij} + \dots + S_{1,2,\dots,p} = 1 \quad (29)$$

Reflétant la part de variabilité induite par chaque variable d'entrée  $X_i$  sur la variabilité totale du modèle  $\mathbb{V}(Y)$ , l'indice de Sobol  $S_i$  est un indicateur de l'importance de la contribution de  $X_i$  dans la variable de sortie. En effet,  $S_i$  peut s'interpréter comme la réduction attendue de la variance totale  $\mathbb{V}(Y)$  lorsque  $X_i$  est fixé à une constante. Cette propriété s'observe notamment à l'aide du théorème de la variance totale :

**Théorème 3** *Si  $X$  et  $Y$  sont deux variables aléatoires sur un même espace de probabilité, et si la variance de  $Y$  est finie, alors*

$$\mathbb{V}(Y) = \mathbb{E}_X [\mathbb{V}(Y|X)] + \mathbb{V}_X (\mathbb{E}[Y|X]) \quad (30)$$

Dès lors, l'indice de Sobol  $S_i$  peut s'écrire comme suit :

$$S_i = \frac{\mathbb{V}_{X_i}(\mathbb{E}_{X_{\sim i}}[Y|X_i])}{\mathbb{V}(Y)} = \frac{\mathbb{V}(Y) - \mathbb{E}_{X_i}(\mathbb{V}_{X_{\sim i}}[Y|X_i])}{\mathbb{V}(Y)} \quad (31)$$

En particulier, les indices de Sobol de premier ordre peuvent être utilisés pour identifier les modèles additifs où il n'y a pas d'interaction entre les variables d'entrée. Dans cette situation, nous obtenons la relation suivante :

$$\sum_{i=1}^p S_i = 1 \quad (32)$$

## II.4 Variables dépendantes : limitations de Sobol et intérêt de Shapley

Notre étude se situe dans ce cadre de non-indépendance des variables d'entrée du modèle. En effet, la détermination de l'indice de vulnérabilité d'un bâtiment considéré dépend directement de sa structure. Par exemple, si la structure du bâtiment est de type *RC1*, la variable représentant l'indice de vulnérabilité sera modélisée par la fonction de densité caractéristique de la structure *RC1*.

Or, d'après Song et al. [31], lorsque l'hypothèse d'indépendance des variables d'entrée du modèle n'est pas vérifiée, la somme des indices de Sobol n'est plus nécessairement égale à 1. Pour illustrer ce principe, considérons le modèle suivant :

$$Y = X_1 + X_2, \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathcal{N} \left( \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \right), \text{ avec } \sigma_1 = \sigma_2 > 0, \rho > 0$$

Dans ce cas :

$$\mathbb{E}[Y|X_1] = X_1 + \mu_2 + \frac{\rho\sigma_2}{\sigma_1}(X_1 - \mu_1) \Rightarrow \mathbb{V}(\mathbb{E}[Y|X_1]) = \sigma_1^2(1 + \rho)^2$$

Et :

$$\mathbb{V}(Y) = \sigma_1^2 + 2\rho\sigma_1\sigma_2 + \sigma_2^2 = 2\sigma_1^2(1 + \rho)$$

On en déduit alors l'écriture des indices de Sobol du premier ordre des variables  $X_1$  et  $X_2$  :

$$S_1 = S_2 = \frac{\sigma_1^2 (1 + \rho)^2}{2\sigma_1^2 (1 + \rho)} = \frac{1 + \rho}{2}$$

Ainsi, comme  $\rho > 0$ , on vérifie bien que  $S_1 + S_2 > 1$ .

En conclusion, lorsque l'on se trouve dans une situation de dépendance entre les variables, la pertinence des indices de Sobol est alors nettement réduite car il est alors plus compliqué d'en déduire de manière claire leur contribution à l'incertitude du modèle.

Pour garantir l'interprétabilité de notre étude, il faut alors s'intéresser aux indices de Shapley qui présentent un avantage par rapport aux indices de Sobol car ils restent interprétables, que les variables d'entrée soient dépendantes ou non. En effet, les indices de Shapley ont par construction une somme toujours égale à 1.

### III Indices de Shapley

Avant d'être utilisé en analyse de sensibilité, les indices de Shapley ont été d'abord introduits dans le cadre de la théorie des jeux coopératifs par Lloyd Shapley [28] en 1953.

#### III.1 Théorie des jeux coopératifs

Un jeu coopératif est caractérisé par un ensemble de joueurs  $\mathcal{D} = 1, \dots, p$  appelé « grande coalition » et d'une fonction caractéristique (ou fonction coût)  $\nu : 2^{\mathcal{D}} \mapsto \mathbb{R}$ . La fonction  $\nu$  associe un sous-ensemble de joueurs  $\mathcal{I} \subseteq \mathcal{D}$  à un nombre réel  $\nu(\mathcal{I})$  qui représente le *payoff* que les membres de la coalition  $\mathcal{I}$  peuvent obtenir en coopérant.

Soit  $\mathcal{J} \subseteq \mathcal{D} \setminus \{i\}$  une coalition de  $|\mathcal{J}|$  qui ne contient pas le joueur  $i$ . La contribution marginale du joueur  $i$  est donc  $\nu(\mathcal{J} \cup \{i\}) - \nu(\mathcal{J})$ . Dès lors, la valeur de Shapley d'un joueur  $i$  est défini comme :

$$\Phi_i = \sum_{\mathcal{J}} \frac{|\mathcal{J}|!(p - |\mathcal{J}| - 1)!}{p!} [\nu(\mathcal{J} \cup i) - \nu(\mathcal{J})], i = 1, \dots, p \quad (33)$$

La valeur de Shapley d'un joueur  $i$  représente donc la moyenne pondérée de la contribution marginale du joueur sur l'ensemble des coalitions possibles, ceci incluant la coalition vide associée à un *payoff* nulle. La valeur de Shapley peut également s'écrire à l'aide des permutations de  $\mathcal{D}$  :

$$\Phi_i = \frac{1}{p!} \sum_{\pi \in \Pi} [\nu(P_i(\pi) \cup i) - \nu(P_i(\pi))], i = 1, \dots, p \quad (34)$$

où :

- $\Pi$  est l'ensemble des  $p!$  permutations de joueurs possibles
- $P_i(\pi)$  est l'ensemble des joueurs qui précède le joueur  $i$  dans la permutation  $\pi \in \Pi$ .

En particulier, la somme des valeurs de Shapley est caractérisée par une propriété d'efficience qui s'avèrera particulièrement intéressante par la suite :

$$\sum_{i=1}^p \Phi_i = \nu(\{1, \dots, p\}) \quad (35)$$



Toutefois, comme la détermination des valeurs de Shapley nécessite de considérer tous les sous-ensembles de joueurs possibles, on comprend rapidement que le temps de calcul, généralement bien supérieur à celui des indices de Sobol, sera la principale difficulté. Pour réduire ce temps, on trouve notamment dans Castro et al. [4] l'estimateur des valeurs de Shapley suivant, où  $\pi_1, \dots, \pi_M$  sont  $M$  permutations aléatoires dans  $\Pi$  :

$$\hat{\Phi}_i = \frac{1}{M} \sum_{m=1}^M [\nu(P_i(\pi_m) \cup i) - \nu(P_i(\pi_m))], i = 1, \dots, p \quad (36)$$

### III.2 Application en analyse de sensibilité

Dans un cadre d'analyse de sensibilité, les variables d'entrée du modèle peuvent s'interpréter comme les joueurs participant à un jeu de *payoff* total  $\mathbb{V}(Y)$ . De même, avec un ensemble de variables d'entrée  $\mathcal{I} \subseteq \mathcal{D}$ ,  $\nu(\mathcal{I})$  représente l'incertitude de la variable de sortie du modèle associée à ces variables d'entrée. En particulier Owen [19] propose en 2014 de considérer la fonction de coût suivante :

$$\nu_1(\mathcal{I}) = \frac{\mathbb{V}_{X_{\mathcal{I}}}(\mathbb{E}[Y|X_{\mathcal{I}}])}{\mathbb{V}(Y)} \quad (37)$$

Lorsque cette fonction caractéristique est utilisée, on ne parle alors plus de valeur de Shapley mais d'indice de Shapley, noté  $Sh_i$ . En reprenant la propriété d'efficacité des valeurs de Shapley, on obtient alors la propriété fondamentale des indices de Shapley en analyse de sensibilité :

$$\sum_{i=1}^p Sh_i = 1 \quad (38)$$

Ainsi cette propriété assure une interprétation sans difficulté, avec ou sans structure de dépendance dans le modèle. Il semble aussi important de signaler que l'on trouve également dans Iooss et Prieur [14] une fonction de coût différente :

$$\nu_2(\mathcal{I}) = \frac{\mathbb{E}_{X_{\sim \mathcal{I}}}(\mathbb{V}[Y|X_{\sim \mathcal{I}}])}{\mathbb{V}(Y)} \quad (39)$$

Toutefois, même s'il est prouvé dans Song et al. [31], que les deux fonctions précédentes retournent des valeurs proches, Radaideh et al. [22] et Sun et al. [33] ont montré que l'estimateur de  $\nu_1$  peut être biaisé lorsque le nombre d'observations utilisé

pour déterminer  $\mathbb{E}[Y|X_I]$  par Monte Carlo est faible. En pratique, la fonction  $\nu_2$  est donc favorisée car son estimateur n'est jamais biaisé.

Dès lors, après avoir introduit le formalisme de l'analyse de sensibilité globale par indice de Shapley, il faut à présent constituer la base de données qui permettra de les déterminer.

### III.3 Estimation et interprétation des indices de Shapley

Il existe de nombreuses approches implémentées dans le logiciel R permettant de déterminer les indices de Shapley : *SHAFF*, *k* plus proches voisins, *ranking...* Comme elle permet la parallélisation des calculs et la prise en compte des variables catégorielles, c'est la fonction *shapleysobol\_knn* de la librairie *sensitivity* qui a été utilisée dans cette étude. Les résultats obtenus sont présentés dans la table suivante :

TABLE 17 – Indices de Shapley des variables d'entrée du modèle. *Lecture : les intervalles de confiance ont été déterminés par bootstrap non-paramétrique à partir de 1000 simulations.*

Variable	Indice de Shapley	Intervalle de confiance à 95%
<b>VI</b>	56,3%	[55,9% ; 56,9%]
<b>MMI</b>	32,0%	[31,7% ; 32,8%]
<b>MOD</b>	4,9%	[4,6% ; 5,3%]
<b>LR</b>	4,7%	[4,5% ; 4,9%]
<b>STR</b>	2,1%	[2,1% ; 2,4%]

Comme l'on pouvait s'y attendre intuitivement, l'incertitude de la variable de sortie du modèle est principalement associée aux variables **VI** et **MMI** représentant respectivement l'indice de vulnérabilité d'une structure et l'intensité des mouvements du sol. Toutefois, on remarque que **VI** est la variable qui contribue en majorité à l'erreur globale de prédiction. Dès lors, la précision de sa détermination représente un enjeu de la plus haute importance pour un assureur souhaitant diminuer la volatilité et augmenter la précision de son modèle.

Partie 5 :  
Importance des variables et  
analyse locale de sensibilité

## I Importance d'une variable et modèle de substitution

En introduction de ce mémoire, deux questions ont été soulevées. Tout d'abord : dans quelle mesure chaque paramètre est-il responsable de l'incertitude associée à la valeur de la perte estimée ? C'est ce qui a été déterminé dans la partie précédente par détermination des indices de Shapley. Dès lors, il convient de répondre à la seconde interrogation : quelles sont les variables qui contribuent le plus à cette même valeur ? Notons que cette contribution pourra être dénommée *poids* ou *importance*.

Dans le cas le plus simple du modèle linéaire, le poids d'une variable dans une prédiction peut être associé dans une quelconque mesure à son coefficient de régression. Idéalement, bien que le modèle de risque sismique ici étudié soit plus complexe, c'est cette simplicité d'interprétation qu'il est souhaitable d'obtenir. Pour cela, il a été choisi ici d'entreprendre cette étude en passant par un modèle de substitution (*surrogate model*).

En particulier, en considérant que le choix du modèle de substitution soit adapté, cette approche peut aider à comprendre les propriétés et les comportements du modèle original en respectant au mieux sa physique propre. De plus, ils peuvent être plus faciles à expliquer et à communiquer à un public non spécialisé, ce qui est utile lors de la présentation des résultats à la direction.

On pourrait penser que cette recherche d'interprétabilité écarte de fait les modèles *machine learning* qui, malgré leur performance, sont limités par l'effet boîte noire qui veut que l'on puisse seulement observer les données en entrée et en sortie mais pas le fonctionnement interne. Cependant, les dernières années ont vu le fort développement des concepts d'IA explicable, ou XAI (*eXplainable Artificial Intelligence*).

XAI est un domaine de recherche en intelligence artificielle qui vise à rendre les décisions prises par des modèles de *machine learning black box* compréhensibles et transparentes. En particulier, l'IA explicable permet de mettre en oeuvre différentes concepts d'analyse de sensibilité locale comme les valeurs SHAP (*SHAPley Additive exPlanations*).

Dès lors, cette dernière partie propose de chercher un modèle de substitution interprétable au mieux adapté au modèle original. Pour cela, il est d'abord nécessaire de définir un critère de sélection entre les différents modèles étudiés.

## II Critères de sélection de modèle statistique

Lorsqu'il s'agit de modéliser des données, il est primordial de sélectionner la méthode statistique la plus appropriée. Le choix du modèle aura évidemment un impact considérable sur la qualité des prévisions, des interprétations ou des conclusions que l'on peut tirer de l'analyse. Pour déterminer le meilleur modèle à utiliser dans une situation donnée, il est nécessaire d'évaluer différentes options en fonction des critères de sélection appropriés.

### II.1 La *Mean Squared Error* (MSE) et la *Root Mean Squared Error* (RMSE)

La MSE et la RMSE sont deux mesures courantes pour évaluer la précision d'un modèle de régression. La MSE mesure l'erreur moyenne au carré entre les valeurs prédites et les valeurs réelles, tandis que la RMSE mesure la racine carrée de cette erreur moyenne. En notant  $y$  les valeurs observées (ou réelles),  $\hat{y}$  les valeurs prédites et  $n$  le nombre d'observations :

$$MSE = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n} \quad (40)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}} \quad (41)$$

Par définition, ces métriques pénalisent les erreurs importantes plus fortement que les plus faibles. Elles sont donc toutes les deux sensibles aux valeurs aberrantes. Toutefois, comme la RMSE s'exprime dans la même unité que la variable à prédire, elle est plus simple à interpréter que la MSE et est donc plus utilisée en pratique.

### II.2 La *Mean Absolute Error* (MAE)

Contrairement aux métriques précédentes qui mesurent l'erreur quadratique moyenne, la MAE mesure simplement l'erreur moyenne absolue entre les valeurs prédites et les valeurs réelles. En reprenant les mêmes notations que précédemment, la MAE se définit comme :

$$MAE = \frac{\sum_{i=1}^n |y_i - \hat{y}_i|}{n} \quad (42)$$

Par définition, cette métrique pénalise donc autant les erreurs importantes que les erreurs les plus faibles.

### II.3 Le critère d'information d'Akaike

Lorsque l'on estime un modèle, il est possible de diminuer son biais en ajoutant un ou plusieurs paramètres. Le critère d'information d'Akaike (*Akaike Information Criterion*, AIC) pénalise les modèles en fonction du nombre de paramètres utilisés et cherche ainsi à satisfaire le critère de parcimonie ou, autrement dit, cherche à décrire les données avec le plus petit nombre de paramètres possibles.

En notant  $k$  le nombre de variables du modèle et  $L$  le maximum de la fonction de vraisemblance du modèle, le critère AIC est défini par :

$$AIC = 2k - 2\log(L) \quad (43)$$

Pour les échantillons de petite taille, on peut noter qu'il existe une définition corrigée de l'AIC notée AICc définie, avec  $n$  le nombre d'observations, comme :

$$AICc = AIC + \frac{2k(k+1)}{n-k-1} \quad (44)$$

### II.4 Le critère d'information bayésien

Critère d'information dérivé de l'AIC, le *Bayesian Information Criterion* (BIC) pénalise non seulement en fonction du nombre de paramètres du modèle mais aussi en fonction de la taille de l'échantillon. En reprenant les notations précédentes où  $k$  est le nombre de variables du modèle,  $n$  le nombre d'observations dans l'échantillon et  $L$  le maximum de la fonction de vraisemblance du modèle, le BIC se définit comme :

$$BIC = k\log(n) - 2\log(L) \quad (45)$$

En pratique, il est donc considéré que l'AIC est utilisé pour retenir les variables pertinentes lors de prévisions, et que le critère BIC vise la sélection de variables statistiquement significative dans le modèle.

## II.5 Choix des critères d'entraînement et de sélection des modèles

Les critères bayésien et d'Akaike, se basant sur la notion de vraisemblance, sont particulièrement adaptés aux modèles paramétriques comme les GLM ou semi-paramétriques comme les GAMLSS. Morsqu'il s'agit de modèles non-paramétriques basés par exemple sur des arbres de régression, il n'existe cependant pas de fonction de vraisemblance explicite. L'AIC ou le BIC ne peuvent donc pas être utilisés pour évaluer leurs performances.

Par rapport à la MAE, rappelons qu'elle pénalise autant les erreurs importantes que les plus faibles. Pour cette raison, elle ne sera également pas envisagée car il est estimé que, dans un modèle appliqué à l'assurance, une erreur importante importante est plus grave qu'une erreur faible. Dès lors, dans le cadre de cette étude, le critère RMSE semble être le plus approprié pour sélectionner un modèle, qu'il soit paramétrique ou non-paramétrique. Le modèle retenu sera celui avec le RMSE le plus proche de zéro.

De plus, notons que d'après Vrieze [35], lorsque le modèle original n'est pas présent dans l'ensemble des modèles comparés, l'AIC est *efficace* dans le sens où il va asymptotiquement choisir le modèle qui minimise la MSE, ce qui n'est pas le cas du BIC. Ainsi, même si le critère de sélection *in fine* sera la RMSE, le critère choisi lors de l'entraînement du modèle sera l'AIC pour les modèles possédant une fonction de vraisemblance explicite.

Par ailleurs, pour ces types de modèle, il a été choisi d'effectuer une sélection par algorithme pas à pas permettant pour limiter le nombre de variables explicatives et le risque de sur-apprentissage. Il en existe trois types : l'approche *forward* ajoute progressivement les variables qui réduisent le plus l'AIC, tandis que l'approche *backward* les supprime progressivement. Ici, l'approche mixte combinant une étape d'élimination de variable après chaque étape de sélection qui a été choisie.

En reprenant la base de données constituée en partie précédente, la pratique veut alors que celle-ci soit séparée en deux parties : 80% servira à l'entraînement du modèle et 20% à son évaluation. Pour limiter le risque de sur-apprentissage, un processus de validation croisée sera effectuée sur la base d'entraînement. Ce processus, appelé *k-fold cross-validation*, répète k-fois le processus de partage du set de données, d'entraînement et de validation. Dès lors, la RMSE du modèle sera égale à la moyenne de celle des k sous-modèles.

### III Modèles Paramétriques : du modèle linéaire multiple à la régression bêta

#### III.1 Régression linéaire multiple

Le modèle linéaire multiple est un modèle statistique qui, contrairement au modèle linéaire simple qui utilise une seule variable indépendante pour prédire une variable dépendante, utilise deux ou plusieurs variables indépendantes pour prédire une même variable dépendante. En particulier, le modèle linéaire multiple peut être utilisé pour déterminer l'importance relative de chaque variable indépendante dans la prédiction de la variable dépendante et examiner leurs interactions.

##### III.1.1 Présentation du modèle

Considérons un échantillon  $(Y_i, X_{i1}, \dots, X_{in})$  avec  $N$  le nombre d'observations et  $i \in \{1, \dots, N\}$ . L'objectif de la régression linéaire multiple est d'expliquer les valeurs de la variable endogène  $Y$  grâce aux valeurs des variables explicatives  $X_1, \dots, X_n$ . Le modèle est défini comme :

$$Y_i = \beta_0 + \beta_1 X_{i1} + \dots + \beta_n X_{in} + \epsilon_i, \text{ avec } i \in \{1, \dots, N\} \quad (46)$$

Sous forme matricielle, le modèle s'écrit :

$$Y = X\beta + \epsilon \quad (47)$$

Avec :

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_N \end{pmatrix}, X = \begin{pmatrix} x_{1,0} & \dots & x_{1,p-1} \\ \vdots & & \vdots \\ x_{N,1} & \dots & x_{N,p-1} \end{pmatrix}, \beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_{p-1} \end{pmatrix}, \epsilon = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_N \end{pmatrix}$$

Dans ce modèle, les coefficients de régression  $\beta_0, \beta_1, \dots, \beta_n$  sont les paramètres à estimer et les  $\epsilon_i$  sont des variables aléatoires représentant l'erreur ou le bruit du modèle telles que, pour tout  $i \in \{1, \dots, N\}$  et  $j \in \{1, \dots, N\}$  différent de  $i$  :

- $(\mathcal{C}_1) : \mathbb{E}(\epsilon_i) = 0$  (Centrage);



- $(\mathcal{C}_2) : \mathbb{V}(\epsilon_i) = \sigma^2$  (Homoscedasticité) ;
- $(\mathcal{C}_3) : Cov(\epsilon_i, \epsilon_j) = 0$  (Non corrélation)

En particulier, lorsque l'on souhaite faire de l'inférence statistique, il faut considérer le modèle de régression linéaire multiple sous hypothèse gaussienne :

- $(\mathcal{C}_4) : \epsilon$  est un vecteur gaussien.

Dans ce cadre, les variables  $\epsilon_i$  sont supposées i.i.d. et les variables  $Y_i$  sont supposées indépendantes. En effet, les conditions  $(\mathcal{C}_1)$  à  $(\mathcal{C}_4)$  sont équivalentes à :

- $\epsilon \sim \mathcal{N}(0, \sigma^2 I_N)$  ;
- $Y \sim \mathcal{N}(X\beta, \sigma^2 I_N)$  ;

Notons que les coefficients du modèle sont obtenus par méthode des moindres carrés ordinaires, consistant à déterminer les valeurs des coefficients qui minimisent la somme des carrés des résidus du modèle.

### III.1.2 Mise en place et critique du modèle

Il est possible d'effectuer une régression linéaire multiple directement sous R avec la fonction *lm*. Les résultats obtenus sont présentés ci-dessous :

TABLE 18 – RMSE du modèle linéaire multiple sur 10-fold et sur l'échantillon test.

Modèle	$RMSE_{cv}$	$RMSE_{test}$
Modèle linéaire multiple	0,0320	0,0376

Pour vérifier que le modèle obtenu vérifie les hypothèses du modèle linéaire, il est généralement nécessaire d'observer différents graphiques de diagnostic. Ceux-ci peuvent notamment aider à visualiser les résidus, les valeurs aberrantes et les observations influentes qui peuvent affecter la validité des résultats de la régression. En pratique, on s'intéresse souvent aux graphiques :

- *Residuals vs Fitted* pour évaluer l'hypothèse d'homoscédasticité : on doit observer une dispersion égale des résidus autour de zéro, indépendamment de la valeur prédite. Autrement dit, il ne faut pas remarquer de schéma apparent dans leur dispersion ;
- *Scale-Location* qui représente, contrairement au graphique précédent, la racine carrée des résidus standardisés en fonction des valeurs prédites ;

- *Normal Q-Q* permet de vérifier la normalité des résidus en comparant les quantiles de la distribution avec ceux de la loi normale ;
- *Residuals vs Leverage* qui montre les résidus standardisés en fonction de la mesure de l'influence de chaque observation, appelée *leverage*. Autrement dit, le levier représente la mesure dans laquelle les coefficients du modèle changeraient si une observation particulière était supprimée de l'ensemble de données. Il est donc très utile pour repérer d'éventuelles observations aberrantes.

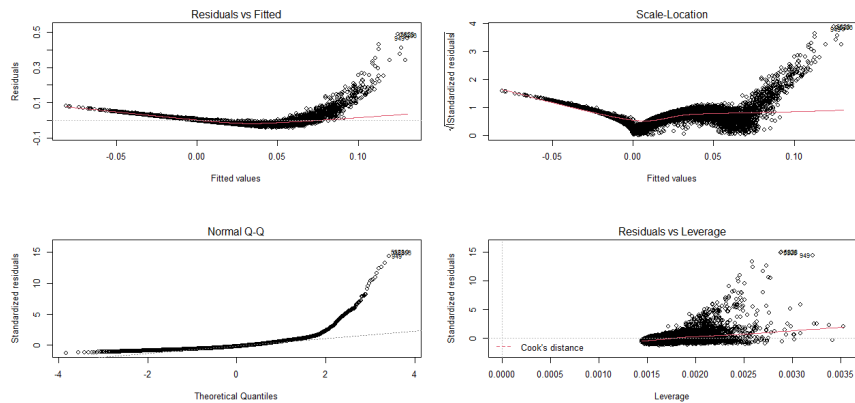


FIGURE 46 – Graphiques de diagnostic du modèle linéaire multiple. *Lecture : le modèle ne suit pas les hypothèses du modèle linéaire. Même si l'on ne remarque pas de valeur aberrante particulière, le modèle ne respecte pas les hypothèses de normalité et d'homoscédasticité des résidus.*

Pour confirmer les déductions faites après l'analyse des divers graphiques de diagnostic, il est intéressant d'effectuer également certains tests statistiques. En particulier, on peut avoir recours aux test suivants :

- Le test d'Anderson-Darling est utilisé pour vérifier si les résidus suivent une distribution normale, ce qui est l'hypothèse nulle de ce test ;
- Le test de Breusch-Pagan est utilisé pour tester si les variances des résidus sont constantes. La version studentisée de ce test tient compte des erreurs standard des résidus. L'hypothèse nulle de ce test est que les variances des résidus sont constantes ;
- Le test de Durbin-Watson permet de vérifier la présence d'autocorrélation des résidus du modèle. L'hypothèse nulle de ce test est qu'il n'y a pas d'autocorrélation.

Le modèle linéaire étant caractérisé par une certaine rigidité, si une des hypothèses ( $\mathcal{C}_1$ ) à ( $\mathcal{C}_4$ ) n'est pas vérifiée alors son interprétation et ses prédictions peuvent être grandement faussées. De plus, en pratique, une variable aléatoire symétrique nor-

TABLE 19 – Résultats des divers tests permettant de vérifier les hypothèses du modèle linéaire.

Nom du test	Statistique de test	p-value	Conclusion
Anderson-Darling	A = 517,74	<2,2e-16	L'hypothèse nulle normalité est rejetée
Breusch-Pagan	BP = 496,02	<2,2e-16	L'hypothèse nulle d'homoscédasticité est rejetée
Durbin-Watson	DW = 2,021038	0,39	L'hypothèse nulle d'absence d'autocorrélation ne peut pas être rejetée : il existe une faible corrélation négative entre les résidus

malement distribuée avec une variance fixe ne permet généralement pas de décrire des situations réelles avec précision.

Le modèle linéaire multiple n'est donc clairement pas adapté à la structure des données étudiées. Pour tenter de « normaliser » les résidus, une pratique répandue consiste à appliquer à  $\mathbf{Y}$  une transformation comme la transformée de Box-Cox. Le choix a été ici fait de ne pas étudier cette option, ces transformées compliquant l'interprétation des résultats et n'étant pas toujours efficaces. Il est donc nécessaire de chercher un nouveau modèle plus adapté.

En partie 2, il a été expliqué en quoi la variable  $\mathbf{Y}$  que l'on cherche ici à modéliser est fortement liée à la distribution bêta, ce qui est d'ailleurs confirmé par le graphe de Cullen-Frey de la distribution empirique :

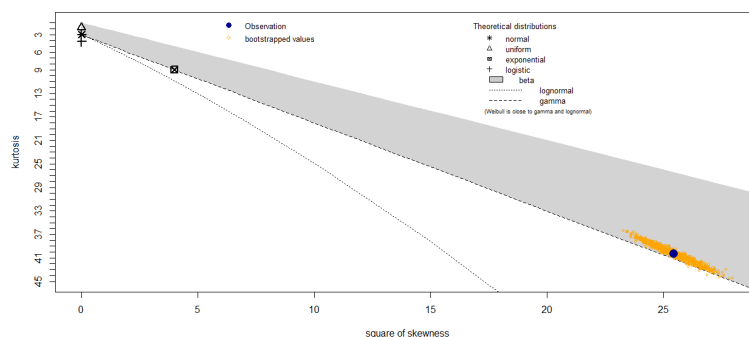


FIGURE 47 – Graphe de Cullen-Frey de la distribution empirique de  $\mathbf{Y}$ .

L'intérêt de ce graphique est de pouvoir de comparer la distribution empirique à certaines distributions théoriques en termes de coefficient d'asymétrie et de kurtosis. Pour rappel, étant donnée une variable aléatoire réelle  $X$  d'espérance  $\mu$  et d'écart type  $\sigma$ , on définit :

- Son coefficient d'asymétrie comme le moment d'ordre trois de la variable centrée réduite :

$$\nu = \mathbb{E} \left[ \left( \frac{X - \mu}{\sigma} \right)^3 \right] \quad (48)$$

— Son kurtosis non normalisé comme le moment d'ordre quatre de la variable centrée réduite :

$$\tau = \mathbb{E} \left[ \left( \frac{X - \mu}{\sigma} \right)^4 \right] \quad (49)$$

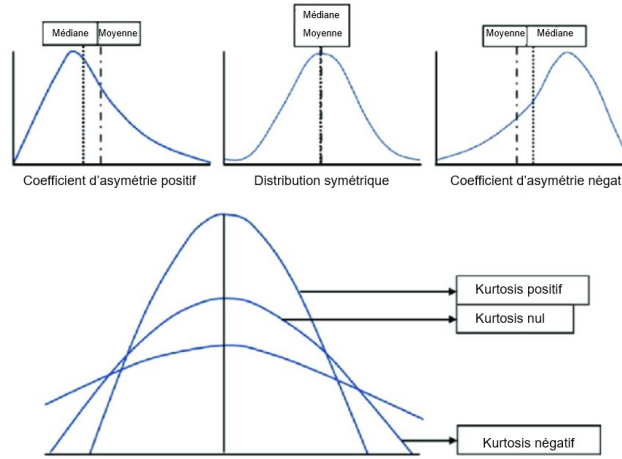


FIGURE 48 – Illustration des différents configurations du coefficient d’asymétrie et de kurtosis. Notons qu’un coefficient d’asymétrie nul n’indique pas nécessairement que la distribution est symétrique, mais une distribution symétrique a un coefficient nul.

La section suivante sera alors dédiée au modèle de régression bêta, introduit en 2004 par Ferrari et Cribari-Neto [8], qui étend à la distribution bêta le modèle linéaire généralisé de Nelder et Wedderburn réservé à la famille exponentielle.

## III.2 Régression Bêta

### III.2.1 Introduction au modèle linéaire généralisé

Le modèle linéaire généralisé (*generalized linear model*, GLM) se distingue du modèle linéaire classique par le fait qu’au lieu de modéliser directement la variable à expliquer, il considère plutôt une fonction de l’espérance de cette variable, appelée fonction lien. Un GLM se compose de trois éléments :

1. Une composante stochastique qui précise que les  $Y_i$  sont indépendantes, d’espérance  $\mathbb{E}[Y_i] = \mu_i$  avec une densité appartenant à la famille exponentielle. Autrement dit, cette densité peut s’écrire sous la forme :

$$f(y, \theta, \phi) = \exp \left( \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \right) \quad (50)$$

avec  $\theta \in \mathbb{R}$  le paramètre canonique (ou paramètre de la moyenne),  $\phi \in \mathbb{R}$  le paramètre de dispersion,  $a$  une fonction non nulle définie sur  $\mathbb{R}$ ,  $b$  une fonction définie sur  $\mathbb{R}$  deux fois dérivable et  $c$  une fonction définie sur  $\mathbb{R}^2$ .

La famille exponentielle se caractérise notamment par deux propriétés :

- $\mathbb{E}(Y_i) = b'(\theta_i)$ ,  $i = 1, \dots, n$
- $\mathbb{V}(Y_i) = b''(\theta) a(\phi)$ ,  $i = 1, \dots, n$

TABLE 20 – Exemples de lois de la famille exponentielle.

Distribution	$\theta$	$\phi$	$a(\phi)$	$b(\phi)$	$c(y, \phi)$
Normale( $\mu, \sigma^2$ )	$\mu$	$\sigma^2$	$\phi$	$\frac{\theta^2}{2}$	$-\frac{1}{2} \left( \frac{y^2}{\sigma^2} + \log(2\pi\sigma^2) \right)$
Bernouilli( $\mu$ )	$\log\left(\frac{\mu}{1-\mu}\right)$	1	1	$\log(1 + \exp(\theta))$	0
Poisson( $\mu$ )	$\log(\mu)$	1	1	$\exp(\theta)$	$-\log(y!)$
Gamma( $\mu, \alpha$ )	$\frac{-1}{\mu}$	$\alpha^{-1}$	$\phi$	$-\log(-\theta)$	$\alpha \log(\alpha y) - \log(y) - \log(\Gamma(\alpha))$
Inverse Gaussienne( $\mu, \sigma^2$ )	$\frac{-1}{2\mu^2}$	$\sigma^2$	$\phi$	$-(-2\theta)^{1/2}$	$-\frac{1}{2} \left( \log(2\pi\phi y^3) + \frac{1}{y\phi} \right)$

2. Une composante systématique qui attribue à chaque observation un prédicteur linéaire :

$$\eta_i = \beta_0 + \beta_1 X_{i1} + \dots + \beta_n X_{in} \quad (51)$$

3. Une fonction lien  $g$ , monotone et différentiable, qui lie l'espérance  $\mu_i$  de  $Y_i$  au prédicteur linéaire  $\eta_i$  par :

$$\eta_i = g(\mu_i) = \beta_0 + \beta_1 X_{i1} + \dots + \beta_n X_{in} \quad (52)$$

TABLE 21 – Exemples de fonctions de lien.

Nom	Fonction de lien
Identité	$g(\mu) = \mu$
Log	$g(\mu) = \log \mu$
Logit	$g(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$
Probit	$g(\mu) = \Phi^{-1}(\mu)$
Inverse	$g(\mu) = \frac{-1}{\mu}$
Log-log	$g(\mu) = \log(-\log(\mu))$
Complementary log-log	$g(\mu) = \log(-\log(1-\mu))$

### III.2.2 Extension à la distribution bêta

La fonction de densité d'une variable  $Y$  suivant une loi bêta  $Beta(\alpha, \beta)$ , avec  $\alpha > 0$  et  $\beta > 0$ , est donnée par :

$$f_Y(y) = \frac{1}{B(\alpha, \beta)} y^{\alpha-1} (1-y)^{\beta-1}, 0 < y < 1 \quad (53)$$

Où  $B$  est la fonction bêta, définie pour tous nombres complexes  $x$  et  $y$  de parties réelles strictement positives, telle que :

$$B(x, y) = \int_0^1 t^{x-1} (1-t)^{y-1} dt \quad (54)$$

Dès lors, l'espérance et la variance de  $Y \sim \text{Beta}(\alpha, \beta)$  sont égales à :

$$\mathbb{E}(Y) = \frac{\alpha}{\alpha + \beta} \text{ et } \mathbb{V}(Y) = \frac{\alpha\beta}{(\alpha + \beta)^2 (\alpha + \beta + 1)} \quad (55)$$

La loi bêta n'appartenant pas à la fonction exponentielle, il n'est pas possible d'utiliser directement la théorie du modèle linéaire généralisé. Toutefois, pour utiliser des techniques similaires, il est possible de reparamétriser la loi bêta en posant  $\mu = \alpha / (\alpha + \beta)$  et  $\phi = \alpha + \beta$ . Dès lors, l'espérance et la variance de la loi s'écrivent :

$$\mathbb{E}(Y) = \mu \text{ et } \mathbb{V}(Y) = \frac{\mu(1-\mu)}{1+\phi} \quad (56)$$

Les paramètres  $\mu$  et  $\phi$  peuvent alors s'interpréter respectivement comme la valeur moyenne et le paramètre de précision car, à  $\mu$  constant,  $\mathbb{V}(Y)$  diminue avec la croissance de  $\phi$ . En reprenant les notations de la section précédente, le modèle de régression bêta s'écrit également avec une fonction lien  $g$ , monotone et différentiable, qui lie l'espérance  $\mu_i$  de  $Y_i$  au prédicteur linéaire  $\eta_i$  par :

$$\eta_i = g(\mu_i) = \beta_0 + \beta_1 X_{i1} + \dots + \beta_n X_{in} \quad (57)$$

### III.2.3 Mise en place et critique du modèle

Dans le cadre de cette étude, la principale limitation de la régression bêta est qu'elle ne permet de considérer qu'une variable réponse prenant ses valeurs dans  $]0, 1[$ , la fonction de log-vraisemblance de la loi bêta contenant des termes non-bornés en 0 et 1. En effet, en considérant  $y_i$  une composante de  $\mathbf{Y}$ , alors la log-vraisemblance s'écrit :

$$\begin{aligned} \log L(y_i) &= \log \Gamma(\alpha + \beta) - \log \Gamma(\alpha) - \log \Gamma(\beta) \\ &\quad + (\alpha - 1) \log(y_i) + (\beta - 1) \log(1 - y_i) \end{aligned} \quad (58)$$

Où  $\Gamma$  est la fonction, définie pour tous nombres complexes  $z$  de parties réelles strictement positives, telle que :

$$\Gamma(z) = \int_0^{+\infty} t^{z-1} e^{-t} dt \quad (59)$$

Notons que, pour tous nombres complexes  $x$  et  $y$  de parties réelles strictement positives, les fonctions gamma et bêta sont liées par la relation suivante :

$$B(x, y) = \frac{\Gamma(x) \Gamma(y)}{\Gamma(x + y)} \quad (60)$$

Ainsi, comme dans cette étude  $\mathbf{Y}$  est distribuée dans  $[0, 1]$ , Smithson et Verkuilen [29] propose de considérer  $\mathbf{Y}'$  telle que  $Y' = (Y \times (n - 1) + 0.5)/n$  où  $n$  est le nombre d'observations. Cela permet à  $\mathbf{Y}'$  d'être distribuée dans  $]0, 1[$ .

Dès lors, il est possible d'effectuer une régression bêta directement sous R avec la fonction *betareg*. Les résultats obtenus sont présentés ci-dessous :

TABLE 22 – RMSE du modèle de régression bêta sur 10-fold et sur l'échantillon test transformé.

Modèle	$RMSE_{cv}$	$RMSE_{test}$
Régression bêta	0,00890	0,0105

Le *summary* et divers graphiques de diagnostic du modèle sont présentés en annexe B. Ainsi, le modèle de régression bêta semble particulièrement bien ajusté aux données avec une RMSE trois fois inférieure à celle du modèle linéaire. Reste toutefois le problème d'avoir à passer par une transformation pour regrouper les observations de  $\mathbf{Y}$  dans  $]0, 1[$ . Pour y répondre, il est intéressant d'étudier les GAMLSS qui étendent notamment les possibilités de distribution de la variable réponse à la distribution bêta inflatée en 0 et en 1.

## IV Modèle semi-paramétrique : GAMLSS (*Generalized Additive Model for Location, Scale and Shape*)

Bien que des modèles plus souples où les relations non linéaires entre la variable de réponse et les variables explicatives sont traitées en utilisant des fonctions de lissage non paramétriques comme les modèles additifs (*Additive Model*, AM) et additifs généralisés (*Generalized Additive Model*, GAM), ces derniers supposent également que la variable de réponse a une distribution qui appartient à la famille exponentielle.

### IV.1 Présentation du modèle

Proposés en 2005 par Rigby et Stasinopoulos [24], les GAMLSS assouplissent l'hypothèse de la distribution de la variable réponse en ne la limitant plus à la famille exponentielle. De plus, ils permettant la modélisation des paramètres d'écart type  $\sigma$ , d'asymétrie  $\nu$  et de kurtosis  $\tau$  en plus du paramètre de la moyenne  $\mu$  à l'aide de quatre fonctions de lien :

$$g_1(\mu) = \nu_1 = X_1\beta_1 + \sum_{j=1}^{J_1} Z_{j1}\gamma_{j1} \quad (61)$$

$$g_2(\sigma) = \nu_2 = X_2\beta_2 + \sum_{j=2}^{J_2} Z_{j2}\gamma_{j2} \quad (62)$$

$$g_3(\nu) = \nu_3 = X_3\beta_3 + \sum_{j=3}^{J_3} Z_{j3}\gamma_{j3} \quad (63)$$

$$g_4(\tau) = \nu_4 = X_4\beta_4 + \sum_{j=4}^{J_4} Z_{j4}\gamma_{j4} \quad (64)$$

Avec  $\mu$ ,  $\sigma$ ,  $\nu$ ,  $\tau$  et  $\nu_k$  des vecteurs de longueur  $n$ . Dans ce modèle,  $X\beta$  représente les parties paramétriques du modèle et  $Z\gamma$  les parties non paramétriques. La grande flexibilité d'un GAMLSS provient ainsi, d'une part, de la possibilité d'ajouter le nombre de termes additifs désirés  $J_k$  à chacune des fonctions de lien  $g_k$  et, d'autre part, de ses parties non paramétriques  $Z_{jk}\gamma_{jk}$  pouvant introduire une gamme variée de termes additifs comme des effets aléatoires ou des fonctions de lissage.



Les GAMLSS sont directement accessibles dans le logiciel R à l'aide de la librairie *gamlss* comprenant une large gamme de distributions possibles pour la variable réponse :

TABLE 23 – Fonctions de lien des paramètres de quelques familles de distributions GAMLSS. Source : Rigby et Stasinopoulos. *Lecture : la fonction logit est telle que, pour  $x \in ]0, 1[$ ,  $\text{logit}(x) = \log\left(\frac{x}{1-x}\right)$*

Distribution	Nom dans R	$\mu$	$\sigma$	$\nu$	$\tau$
Bêta	BE()	logit	logit	-	-
Bêta inflatée (en 0 et 1)	BEINF()	logit	logit	log	log
Weibull	WEI()	log	log	-	-
Box-Cox-t	BCT()	identité	log	identité	log
Poisson Inverse Gaussienne	PIG()	log	log	-	-

En particulier, dans la suite de cette étude, c'est la distribution bêta inflatée en 0 et 1 qui sera utilisée. Tout d'abord, parce qu'elle est appropriée lorsque la variable réponse prend ses valeurs entre  $[0, 1]$  mais surtout car, par construction, la variable loss ratio  $\mathbf{Y}$  que l'on cherche à modéliser est fortement liée à la distribution bêta.

Contrairement à la distribution bêta à 2 paramètres qui ne permet de modéliser qu'une variable réponse dans  $]0, 1[$ , la distribution bêta inflatée en 0 et 1 permet de modéliser les probabilités pour  $\mathbf{Y}$  de prendre les valeurs 0 et 1, notées respectivement  $p_0$  et  $p_1$ , par une régression logistique. La proportion restante,  $1 - p_0 - p_1$ , est alors modélisée par une distribution bêta standard.

Dès lors, lorsque  $Y \sim \text{Beinf}(\mu, \sigma, \nu, \tau)$ , la fonction de densité de la distribution bêta inflatée en 0 et 1 est définie par :

$$f_Y(y|\mu, \sigma, \nu, \tau) = \begin{cases} p_0 & \text{si } y = 0 \\ (1 - p_0 - p_1) f_W(y) & \text{si } 0 < y < 1 \\ p_1 & \text{si } y = 1 \end{cases} \quad (65)$$

Caractérisée par une espérance  $\mathbb{E}(W) = \alpha/(\alpha + \beta)$ ,  $W$  suit une distribution bêta  $\text{Beta}(\mu, \sigma)$  de fonction de densité :

$$f_W(y) = \frac{1}{B(\alpha, \beta)} y^{\alpha-1} (1-y)^{\beta-1}, 0 < y < 1 \quad (66)$$

Où  $\alpha, \beta, p_0$  et  $p_1$  sont définis comme :

- $\alpha = \mu (1 - \sigma^2) / \sigma^2, \alpha > 0;$
- $\beta = (1 - \mu) (1 - \sigma^2) / \sigma^2, \beta > 0;$
- $p_0 = \nu (1 + \nu + \tau)^{-1}, 0 < p_0 < 1;$
- $p_1 = \tau (1 + \nu + \tau)^{-1}, 0 < p_1 < 1 - p_0.$

Autrement dit les paramètres de la fonction  $Beinf(\mu, \sigma, \nu, \tau)$  peuvent s'écrire de la manière suivante :

- $\mu = \alpha / (\alpha + \beta), 0 < \mu < 1;$
- $\sigma = (1 + \alpha + \beta)^{-1/2}, 0 < \sigma < 1;$
- $\nu = p_0 / (1 - p_0 - p_1), \nu > 0;$
- $\tau = p_1 / (1 - p_0 - p_1), \tau > 0.$

Ainsi,  $\nu$  et  $\tau$  peuvent s'interpréter dans cette situation comme les paramètres modélisant respectivement les probabilités en 0 et 1. Notons aussi que  $\mathbb{E}(Y) = \frac{\tau + \mu}{1 + \nu + \tau}$ .

## IV.2 Mise en place et critique du modèle

Il est possible de mettre en place un GAMLSS directement sous R avec la fonction `gamlss`. Les résultats obtenus sont présentés ci-dessous :

TABLE 24 – RMSE du GAMLSS bêta inflaté en 0 et 1 sur 10-fold et sur l'échantillon test.

Modèle	$RMSE_{cv}$	$RMSE_{test}$
GAMLSS	0,00825	0,00957

Le *summary* et divers graphiques de diagnostic du modèle sont présentés en annexe C. Ainsi, le GAMLSS bêta inflaté en 0 et 1 est donc bien plus efficace que la régression bêta selon le critère RMSE. Toutefois, avec sa grande flexibilité de paramètres, il est plus complexe à entraîner et à interpréter. À titre informatif, notons qu'ils peuvent même incorporer diverses fonctions de lissage comme, par exemple, des fonctions polynomiales cubiques (*cubic splines*) ou des régressions locales (*locally estimated scatterplot smoothing*) qui augmentent le risque de sur-ajustement.

Dès lors, il peut être intéressant de comparer un GAMLSS à des modèles non-paramétriques comme, par exemple, le modèle MARS et les arbres de régression qui permettent notamment de détecter et reproduire des relations non-linéaires complexes sans avoir besoin de spécifier une forme fonctionnelle a priori.

## V Modèles non-paramétriques : du modèle MARS aux arbres de régression

### V.1 MARS (*Multivariate adaptive regression splines*)

Introduite en 1991 par Friedman et Silverman, la régression multivariée par spline adaptative (*Multivariate adaptive regression splines*, MARS) est une technique de régression non paramétrique. Cette méthode permet notamment de modéliser automatiquement les interactions et les non-linéarités, tout en conservant une interprétabilité équivalente au modèle linéaire.

#### V.1.1 Présentation du modèle

En notant  $B_i$  les des fonctions de base et  $c_i$  les coefficients constants, un modèle MARS s'exprime sous la forme :

$$\hat{f}(x) = \sum_{i=1}^k c_i B_i(x) \quad (67)$$

Chaque fonction de base peut prendre l'une des trois formes suivantes :

- Une constante. Il n'y a qu'un seul terme de ce type : l'intersection avec l'axe ;
- Une fonction qui s'écrit comme  $\max(0, x - \text{constante})$  ou  $\max(0, \text{constante} - x)$ . Une fonction de cette forme est appelée « charnière » ;
- Un produit de fonctions charnières.

Un modèle MARS se comporte comme un algorithme glouton : il ajoute à chaque étape le couple fonction de base - réciproque qui diminue le plus la somme des carrés des résidus. Les  $c_i$  sont ensuite déterminés par régression linéaire. Pouvant aussi être combinée avec un GLM, la régression multivariée par spline adaptative se démarque par sa grande flexibilité et sa facilité d'implémentation car elle ne demande pas de pré-sélection des variables.

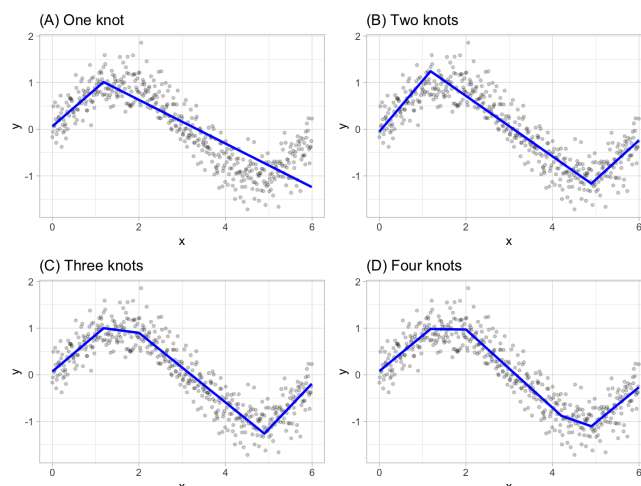


FIGURE 49 – Exemple de construction d’un modèle MARS étape par étape.

### V.1.2 Mise en place et critique du modèle

Il est possible d’ajuster un modèle MARS directement sous R avec la fonction *earth*. Les résultats obtenus sont présentés ci-dessous :

TABLE 25 – RMSE du modèle MARS sur 10-fold et sur l’échantillon test.

Modèle	$RMSE_{cv}$	$RMSE_{test}$
MARS	0,0100	0,0115

Le *summary* et divers graphiques de diagnostic du modèle sont présentés en annexe D. Même s’il est plus efficace qu’un modèle linéaire multiple, il est moins adapté qu’une régression bêta ou GAMLSS. Dès lors, il est intéressant d’étudier d’autres modèles non-paramétriques. Ceux qui seront abordés dans la suite cette étude sont basés sur la notion d’arbres de régression.

## V.2 CART (*Classification And Regression Trees*)

### V.2.1 Présentation du modèle

Le principe des arbres de classification et de régression, introduit par Breiman en 1984, consiste à partitionner de façon binaire le jeu de données en classes homogènes par rapport à la variable de sortie. A chaque étape, un algorithme glouton choisit une variable et un seuil de partition qui minimise la déviance (pour un arbre de classification) ou la somme des carrés (pour un arbre de régression) de la nouvelle partition.

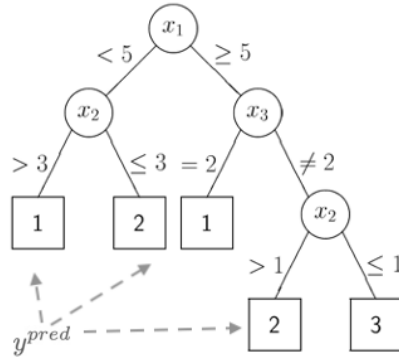


FIGURE 50 – Exemple d’arbre déterminé par algorithme CART.

Le principe général d’un arbre de régression est de lier une observation de la variable réponse  $y_i$  et une observation des variables explicatives  $x_i = (x_{i1}, \dots, x_{iN})$  avec une fonction  $f$  telle que :

$$f(x) = \sum_{m=1}^N c_m \mathbb{1}_{\{x \in R_m\}} \quad (68)$$

Une coupure est un élément de la forme :

$$\{X_j \leq d\} \cup \{X_j > d\}, \text{ avec } j \in \{1, \dots, n\} \quad (69)$$

Dans le cas d’une variable explicative catégorielle, une coupure est de la forme

$$\{X_j \in d\} \cup \{X_j \in \hat{d}\}, \text{ avec } j \in \{1, \dots, n\} \quad (70)$$

### V.2.2 Limitations

Malgré sa grande facilité d’interprétation et des méthodes développées pour éviter le surapprentissage comme *l’élagage*, il s’avère que la construction d’un arbre optimal peut varier fortement quand bien même le jeu de données initial varie peu.

Deux méthodes d’agrégation ont été développées pour corriger l’instabilité de l’arbre optimal déterminé par l’algorithme CART. Une stratégie d’agrégation aléatoire (*Bagging*) et une stratégie d’apprentissage incrémental (*Boosting*).

## V.3 Gradient Boosting

### V.3.1 Présentation du modèle

Le *boosting* permet d'améliorer la précision et les performances prédictives des modèles de *machine learning* en convertissant plusieurs apprenants faibles (i.e. capables de reconnaître deux classes au moins aussi bien que le hasard ne le ferait) en un seul modèle d'apprentissage fort. Contrairement au *bagging*, le *boosting* permet de réduire le biais.

Le principe général de cette méthode est de construire une famille de modèles de façon itérative où chaque modèle cherche à corriger les faiblesses du précédent en donnant plus d'importance aux observations difficiles à prédire. Ces pondérations sont ainsi ajustées au fur et à mesure de l'apprentissage.

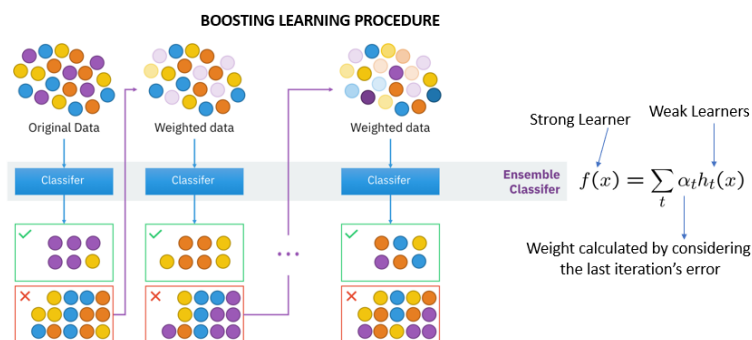


FIGURE 51 – Principe général du *boosting*. Source : Wikipédia

Ainsi, le *boosting* permet de construire un apprenant fort  $F$  comme une combinaison linéaire d'apprenants faibles  $f_i$  avec  $1 \leq i \leq M$ , où chacun des  $f_i$  est pondéré par un  $\alpha_i$  déterminé par l'erreur de l'itération précédente. Lorsque cette méthode est combiné à une descente de gradient, on parle alors de *gradient boosting*. Le principe est le suivant :

1. Un modèle est initialisé par une valeur constante  $f_0(x)$  ;
2. En notant  $L$  la fonction de perte quadratique, les pseudo-résidus  $r_{im}$  sont calculés à chaque itération  $m$  selon :

$$r_{im} = -\frac{\partial L(y_i, f_{m-1}(x_i))}{\partial f_{m-1}(x_i)} \quad (71)$$

Pour rappel, la fonction de perte quadratique  $L$  est telle que  $L(y, \hat{y}) = (y - \hat{y})^2$

3. Un apprenant faible  $h_m$  est alors entraîné sur les données  $(x_i, r_{im})$  ;

4. La pondération  $\gamma_m$  de  $h_m$  est alors déterminée par :

$$\gamma_m = \underset{\gamma}{\operatorname{argmin}} \sum_{i=1}^M L(y_i, f_{m-1}(x_i) + \gamma h_m(x_i)) \quad (72)$$

5. Le modèle est alors mis à jour en considérant un hyperparamètre  $\eta$  appelé taux d'apprentissage :

$$f_m = f_{m-1}(x) + \eta \gamma_m h_m(x) \quad (73)$$

Ce paramètre permet ainsi de contrôler la vitesse à laquelle les paramètres sont ajustés dans la direction de la descente de gradient.

### V.3.2 Mise en place et critique du modèle

Il est possible d'ajuster une méthode de *gradient boosting* directement sous R avec la fonction *xgboost*, une implémentation optimisée de l'algorithme. Dès lors, les résultats obtenus sont présentés ci-dessous :

TABLE 26 – RMSE du modèle XGBoost sur 10-fold et sur l'échantillon test.

Modèle	$RMSE_{cv}$	$RMSE_{test}$
XGBoost	0,00860	1,05

Le *tuning* des hyperparamètres du modèle, le *summary* et divers graphiques de diagnostic du modèle sont présentés en annexe E. Bien qu'il soit hautement flexible, même avec des techniques telles que la validation croisée et le *tuning* rigoureux des hyperparamètres, il est possible que le modèle XGBoost sur-apprenne, en particulier sur des ensembles de données complexes ou avec un grand nombre de variables. Ici, la différence entre  $RMSE_{cv}$  et  $RMSE_{test}$  montre que le modèle entraîné est clairement sujet à un sur-apprentissage. Dès lors, il est intéressant de comparer les résultats du *boosting* à ceux d'une stratégie *bagging*.

## V.4 Random Forest

Le principe du *bagging* (*Bootstrap aggregating*) est de créer, par tirage aléatoire et avec remise dans le jeu de données initial,  $n$  échantillons bootstrap sur lesquels est construit un arbre de décision par méthode CART. Pour obtenir la prédiction finale, les  $n$  estimateurs obtenus sont alors moyennés (pour un arbre de régression) ou utilisés pour un vote à la majorité (pour un arbre de classification). Même s'il n'est pas très efficace pour réduire le biais, le *bagging* permet d'éviter le surajustement.

Étant donné son principe, le *bagging* fait qu'une partie des données sont inutilisées pour l'apprentissage de chaque arbre. Ces données sont dites *Out Of Bag* (OOB). On considère généralement l'erreur OOB qui est l'erreur de prédiction calculée sur les données qui ne sont pas dans l'échantillon bootstrap. Cette erreur peut donc s'interpréter comme une approximation de l'erreur de généralisation de la forêt.

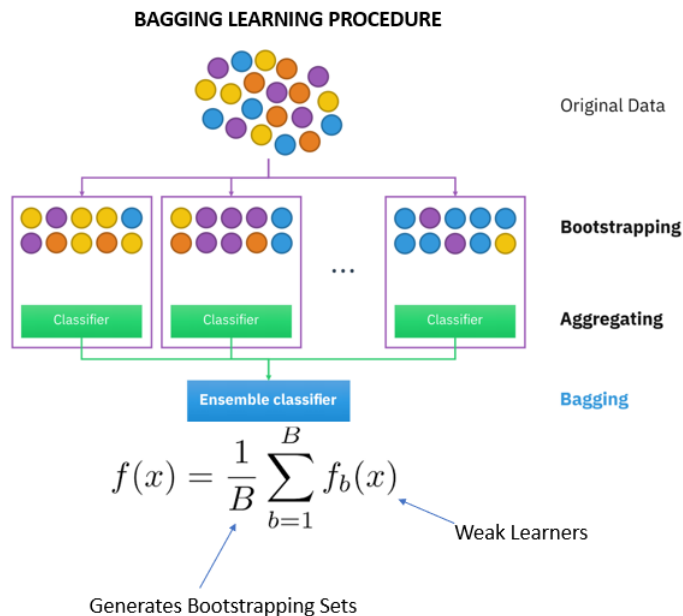


FIGURE 52 – Principe général du *bagging*. Source : Wikipédia

Les forêts aléatoires sont une amélioration du *bagging* pour les arbres de décision CART dans le but de rendre les arbres utilisés moins corrélés. Pour chaque échantillon bootstrap, on construit un arbre CART selon un algorithme légèrement modifié : à chaque fois qu'un noeud doit être coupé on tire au hasard une partie des variables et on choisit le meilleur découpage dans ce sous-ensemble.

#### V.4.1 Mise en place et critique du modèle

Il est possible d'ajuster une méthode *random forest* directement sous R avec la fonction *ranger*. Les résultats obtenus sont présentés ci-dessous :

TABLE 27 – RMSE du modèle Random Forest sur 10-fold et sur l'échantillon test.

Modèle	$RMSE_{cv}$	$RMSE_{test}$
Random Forest	0,00799	0,00922



Ainsi, contrairement au modèle XGBoost, il semble que Random Forest ne présente pas de sur-apprentissage. Parmi tous les modèles étudiés précédemment, c'est celui dont la RMSE est la plus faible. Il convient alors de poursuivre cette étude en conservant cette implémentation de Random Forest.

Un problème se pose toutefois : son manque apparent d'interprétabilité. Pour apporter une solution à ce problème, les dernières années ont vu le développement des concepts d'IA explicable, également appelé XAI (*eXplainable Artificial Intelligence*). Ces concepts cherchent à fournir des explications compréhensibles et intuitives sur la façon dont les modèles d'apprentissage automatique prennent des décisions, permettant ainsi aux utilisateurs d'en comprendre les raisons sous-jacentes.

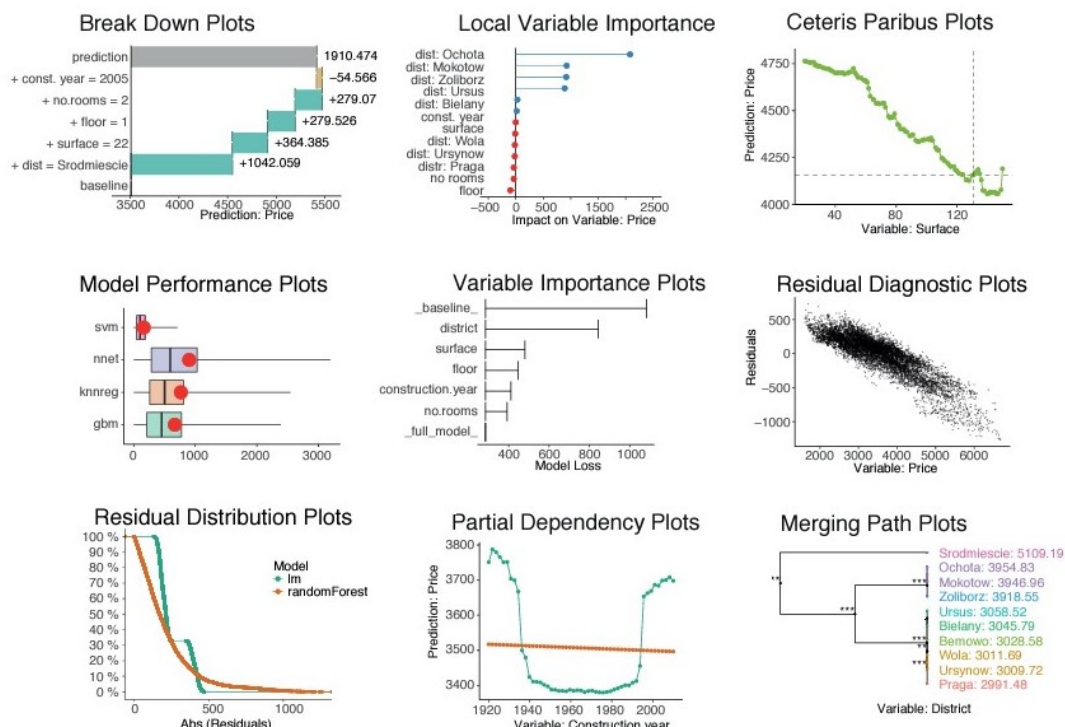


FIGURE 53 – Quelques outils graphiques d'IA Explicable.

En particulier, les valeurs SHAP (*SHapley Additive exPlanation*) reposant sur la théorie des méthodes additives d'attribution de caractéristique permettent de réaliser une analyse locale de sensibilité et d'obtenir un indicateur d'importance de contribution des variables du modèle. Celui permettra alors de répondre à la deuxième question posée en introduction de cette étude : quelles sont les variables qui contribuent le plus à la valeur des pertes estimée ?

## VI Méthodes additives d'attribution de caractéristique

Pour accroître l'interprétabilité des modèles de *machine learning*, Lundberg et Lee [15] ont développé en 2017 le formalisme des méthodes additives d'attribution de caractéristique (*additive feature attribution methods*). Toutefois, avant d'y parvenir, il est d'abord nécessaire d'introduire la notion de méthode locale.

### VI.1 Méthodes locales

Soit  $f$  le modèle original que l'on cherche à expliquer. Une méthode locale cherche à expliquer une prédiction  $f(x)$ , généralement à partir d'un input vecteur binaire simplifié  $x'$  que l'on peut retrouver dans la littérature sous le nom de « masque ».

Les vecteurs  $x$  et  $x'$  sont liés par une fonction  $h_x$  telle que  $x = h_x(x')$  qui peut ainsi être considérée comme la fonction de « démasquage » spécifique de  $x$ . Étant donnée une telle fonction, il est possible d'évaluer  $f(h_x(z'))$  et de déterminer l'impact qu'aura la présence ou l'absence d'une ou plusieurs variables selon  $z'_i = 1$  ou  $z'_i = 0$ . En particulier, une méthode locale garantie que :

$$g(z') \approx f(h_x(z')), \text{ lorsque } z' \approx x' \quad (74)$$

Dès lors, en notant  $p$  le nombre de variables d'entrées et  $\phi_i$  l'effet de la variable  $X_i$ , un modèle  $g$  appartenant à la classe des méthodes additives d'attribution de caractéristique est défini comme une fonction linéaire de variables binaires :

$$g(z') = \phi_0 + \sum_{i=1}^p \phi_i z'_i \quad (75)$$

Plus précisément,  $\phi_i$  est un nombre réel indiquant dans quelle mesure une sortie  $f(x)$  du modèle original dépend de la variable  $X_i$ . Ainsi, par définition, un tel modèle  $g$  attribue un effet à chaque variable d'entrée et lie la somme de tous les effets à une quantité d'intérêt étant une approximation de  $f(x)$ . Bien qu'ils ne soient pas étudiés ici, notons à titre d'information qu'il existe de nombreux modèles respectant cette propriété : LIME, DeepLIFT...

## VI.2 Propriétés souhaitables

Pour qu'une méthode additive d'attribution de caractéristique soit satisfaisante, il est souhaitable qu'elle vérifie un ensemble de trois propriétés :

1. Précision locale (*local accuracy*) : l'évaluation du modèle original  $f$  en  $x$  coïncide avec celle du modèle explicatif  $g$  en  $x'$ .

$$f(x) = g(x') = \phi_0 + \sum_{j=1}^p \phi_j x'_j \quad (76)$$

2. Absence (*missingness*) : l'effet d'une variable absente de l'input original est nul.

$$x'_i = 0 \implies \Phi_i = 0 \quad (77)$$

3. Cohérence (*consistency*) : effectuer un changement de modèle tel qu'une variable a un impact plus important sur celui-ci ne fera jamais décroître l'effet qui lui est attribué. En notant  $f_x(z') = f(h_x(z'))$  et  $z' \setminus i$  pour  $z'_i = i$ , alors pour deux modèles  $f$  et  $f'$  et  $z' \in \{0, 1\}^p$  :

$$f'_x(z') - f'_x(z' \setminus i) \geq f_x(z') - f_x(z' \setminus i) \implies \phi_i(f', x) \geq \phi_i(f, x) \quad (78)$$

En particulier, il a été démontré par Lundberg et Lee que le seul modèle d'explication  $g \in \mathcal{G}$  qui vérifie les trois propriétés précédentes est celui où :

$$\phi_i(f, x) = \sum_{z' \subseteq x'} \frac{|z'|! (p - |z'| - 1)!}{p!} [f_x(z') - f_x(z' \setminus i)] \quad (79)$$

Où  $|z'|$  est le nombre de bits non nuls de  $z'$  et  $z' \subseteq x'$  représente l'ensemble des vecteurs  $z'$  dont les entrées non nulles sont un sous-ensemble des entrées non nulles de  $x'$ . On reconnaît bien l'écriture des valeurs de Shapley introduites dans la section précédente, qui dépendent ici de l'observation  $x$ .

Autrement dit, étant déjà caractérisées par la propriété 2, cela signifie que les méthodes additives d’attribution de caractéristique ne reposant pas sur les valeurs de Shapley ne respectent pas au moins une des deux propriétés de précision locale et de cohérence.

### VI.3 Valeurs SHAP (*SHapley Additive exPlanation*)

Dès lors, il est possible d’introduire les valeurs SHAP, qui sont solutions de l’équation précédente avec  $f_x(z') = f(h_x(z')) = \mathbb{E}[f(z) | z_S]$ , avec  $S$  l’ensemble des indices non nuls de  $z'$ .

Les valeurs SHAP permettent d’attribuer à chaque variable d’entrée le changement de prédiction attendue du modèle original lorsque l’on conditionne par rapport à cette variable. Il est important de noter que, lorsque le modèle est non linéaire ou que les entités en entrée sont dépendantes, l’ordre dans lequel les variables sont ajoutées à son importance. Dans ce cas, les valeurs SHAP sont la moyenne des  $\phi_i$  sur toutes les combinaisons possibles.

$$f(x) = \phi_0 + \sum_{i=1}^p \phi_i x'_i$$

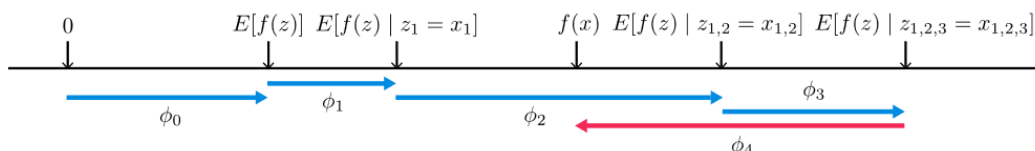


FIGURE 54 – Principe de la méthode SHAP. Source : Lundberg et Lee.

En pratique, il existe actuellement différents moyens d’approximer les valeurs SHAP : KernelSHAP, TreeSHAP... La plupart de ces méthodes ont pour hypothèses que les variables du modèles sont indépendantes, ce qui n’est pas le cas de cette étude.

Pour déterminer les valeurs SHAP, la méthode établie par Aas, Jullum, et Løland [1] qui sera appliquée. Celle-ci permet notamment d’étendre la méthode KernelSHAP à des modèles à variables dépendantes. Elle est directement applicable dans le logiciel R grâce à la fonction *shapr*. Voici les résultats obtenus sur deux observations particulières :

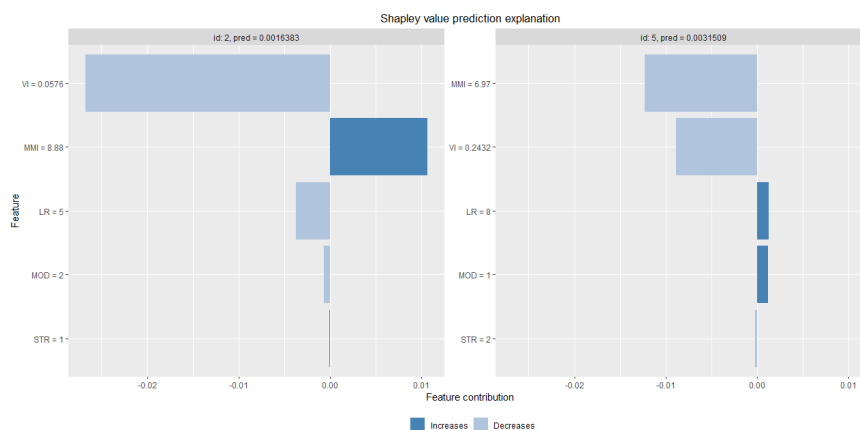


FIGURE 55 – SHAP *breakdown* de deux observations de la base de données.

## VI.4 Indicateur SHAP d'importance de contribution

Dès lors, il est possible de définir un indicateur d'importance de contributions en moyennant les valeurs SHAP de la base d'apprentissage pour une variable. En notant  $n$  le nombre d'observations et  $\phi_j^{(i)}$  la valeur SHAP de la variable d'entrée  $X_j$  dans l'observation  $i$ , on définit alors :

$$I_j = \frac{1}{n} \sum_{i=1}^n |\phi_j^{(i)}| \quad (80)$$

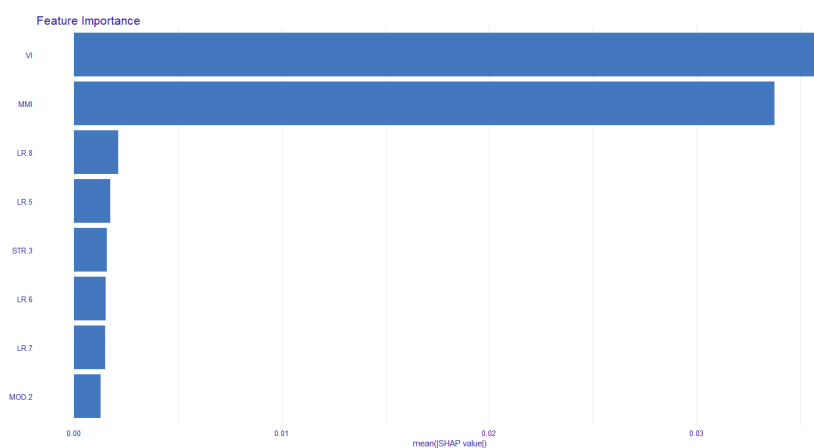


FIGURE 56 – Valeur de l'indicateur SHAP pour chacune des 8 variables les plus importantes.

Ainsi, contrairement à l'analyse de sensibilité globale où il a été démontré que l'indice de vulnérabilité était de loin la variable la plus responsable de la variabilité de l'estimation des pertes, cet indicateur montre que les variables **VI** et **MMI** contribuent à la valeur du ratio de pertes estimé d'une façon relativement semblable.

Dans une moindre mesure, il semble que le choix de l'ensemble de loss ratios représenté par **LR** ait son importance, en particulier les ensembles 8 et 5 étant respectivement ceux définis par Milutinovic et Kappos. Notons qu'il avait été précédemment remarqué que la différence entre ces deux ratios était la plus élevée de tous en *damage state* 4 et 5.

De même, les choix de modèle bêta ou binomial et de structure n'apparaissent avoir qu'une contribution mineure dans la valeur de perte. Remarquons toutefois l'impact d'une structure de type 3 sur cette estimation, étant représentatrice d'une construction avec ossature de niveau de conception parasismique élevé.

Rappelons toutefois que ces conclusions doivent être extrapolées au modèle original avec précaution. En effet, même si les valeurs SHAP ont été déterminées sur un modèle Random Forest qui semble particulièrement bien ajusté, elles restent une approximation d'un concept basé sur une approximation de modèle.

Pour poursuivre cette étude, il semble à présent particulièrement intéressant de comparer les résultats précédents des analyses de sensibilité globale et locale avec ceux obtenus lorsque l'on considère, non pas un seul et unique bâtiment, mais un portefeuille de police d'assurances dans le cadre d'un séisme historique.

**Partie 6 :**  
**Extension de l'étude à un portefeuille**  
**d'assurances**

# I Contexte de l'étude

Dans les parties précédentes, nous avons réalisé une analyse globale de sensibilité à partir des valeurs de Shapley, puis obtenu un indicateur d'importance de contribution grâce à une analyse locale de sensibilité et aux valeurs SHAP. Toutefois, rappelons que ces résultats ont été obtenus dans un cadre précis. Il s'agissait en effet d'étudier un seul et unique bâtiment, sur lequel s'appliquait une intensité macrosismique centrée en VIII. Dès lors, pour accroître l'étendue de cette étude et son intérêt assurantiel, il est intéressant d'observer l'évolution de ces résultats en considérant désormais un ensemble de bâtiments assurés dans le cadre d'une catastrophe naturelle historique : le séisme du 23 novembre 1980 dans la région italienne d'Irpinia qui se distingue par son intensité macrosismique extrême de niveau X.

## I.1 Impacts socio-économiques du séisme

Le séisme du 23 novembre 1980 s'est produit à Irpinia en Italie à 19 h 34 heure locale, suivie d'une seconde secousse après 40 secondes. Le séisme a duré en totalité près d'une minute et demie. Il a été ressenti dans toute la région ainsi qu'en Ligurie au nord, provoquant des dégâts dans plus de 800 localités réparties entre les régions de Campanie et de Basilicate.

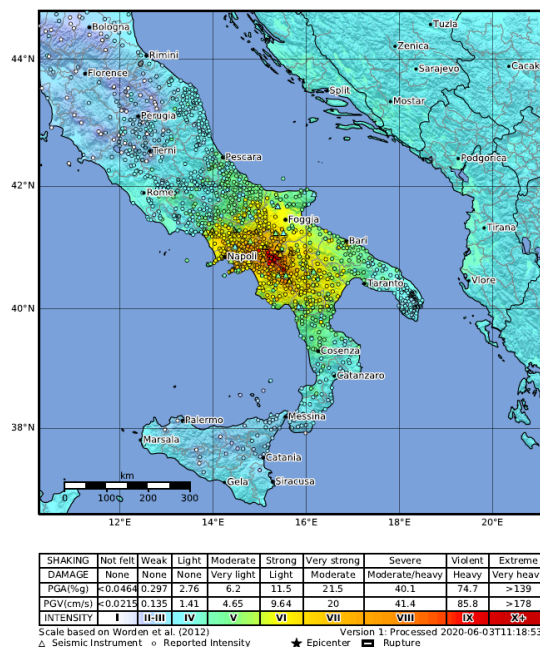


FIGURE 57 – Carte de l'intensité macrosismique du séisme du 23 novembre 1980 en Irpinia. Source : USGS.



Cette catastrophe est depuis restée dans la mémoire des populations locales, non seulement en raison de sa puissance, mais aussi en raison de la dévastation causée, entraînant la perte de nombreuses vies humaines, la destruction de régions entières et la disparition d'un patrimoine culturel précieux dans la zone épacentrale. Au total, environ 75 000 maisons ont été détruites et 275 000 ont subi de graves dommages. On estime que près de 3 000 personnes ont perdu la vie et 10 000 ont été blessées. D'après le Swiss Re Institute [13], les pertes économiques sont estimées à environ 40 milliards d'euros, en corrigeant de l'inflation jusqu'en 2023. Par ailleurs, le niveau d'*industry loss* estimé par RMS s'élève à 5 milliards d'euros en ajustant de l'inflation jusqu'en 2023.



FIGURE 58 – Photographies prises après le séisme du 23 novembre 1980 en Irpinia.

De plus, Porfido et al. [21] notent que le séisme a également eu des conséquences significatives sur l'environnement naturel : plus de 200 glissements de terrain se sont produits et des changements importants dans le débit d'eau de certaines sources ont été signalés. Par ailleurs, des failles de surface étendues, d'une longueur totale d'environ 40 km, sont encore visibles aujourd'hui.

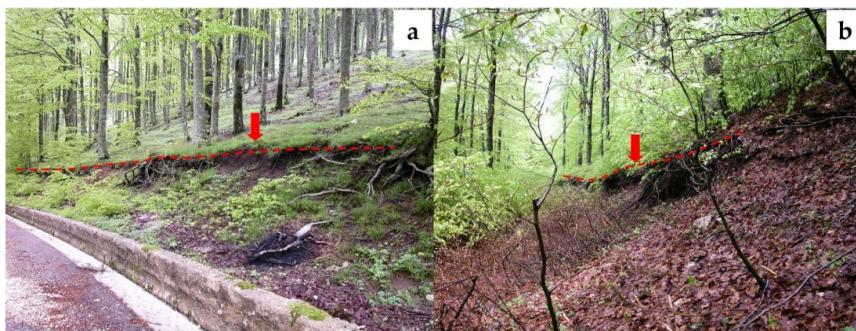


FIGURE 59 – Failles de surface visibles en 2004 résultant du séisme du 23 novembre 1980 en Irpinia. Source : S. Porfido

Enfin, Moscaritolo [18] montre comme le séisme d'Irpinia témoigne parfaitement des interactions complexes entre l'Homme et la nature, à travers les processus de reconstruction et d'adaptation à l'environnement qui se sont par la suite déroulés sur le court et long terme. En cela, son raisonnement est similaire à celui de Bousquet [2] cité en introduction de cette étude.

## I.2 Description du portefeuille d'assurances

Il convient désormais de présenter le portefeuille d'assurances avec lequel sera réalisée cette nouvelle étude. Celui-ci est issu du portefeuille d'assurances Axa, réduit à 500 sites de structure *RC1*, *RC2* et *RC3* présents dans la zone touchée par le séisme d'Irpinia. En particulier, il contient pour chaque police les informations suivantes :

- L'identifiant du contrat ;
- La longitude et la latitude du bâtiment assuré (en degrés) ;
- La structure du bâtiment assuré (*RC1*, *RC2* ou *RC3*) ;
- La valeur assurée par le contrat (en euros) ;

Notons que pour des raisons évidentes de confidentialité, les informations relatives à chaque contrat ont été anonymisées. Un court extrait de ce portefeuille est présentée dans la table ci-dessous :

TABLE 28 – Les 5 premières lignes du portefeuille d'assurances.

ID Contrat	Longitude	Latitude	Structure	Valeur Assurée
XXXX4761	14,6087°	40,9589°	RC2	2 763 586 €
XXXX6863	14,2914°	40,9088°	RC3	26 550 587 €
XXXX6473	14,1578°	40,8086°	RC3	905 978 €
XXXX8646	14,2079°	40,8253°	RC2	43 524 €
XXXX3998	14,4584°	41,66034°	RC1	466 216 €

Dès lors, il est possible de produire certains graphiques descriptifs permettant de représenter les caractéristiques des bâtiments assurés. Celui ci-dessous illustre par exemple la fonction de répartition de la valeur assurée des bâtiments en portefeuille. En particulier, avec une somme assurée totale d'environ 511,3 millions d'euros, la médiane des valeurs assurées s'élève à 284 192€ et la moyenne à 1 022 780€.

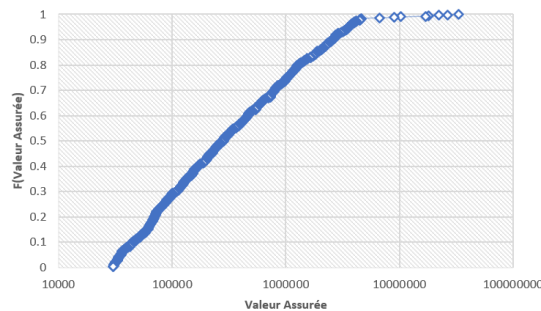


FIGURE 60 – Fonction de répartition de la valeur assurée des bâtiments en portefeuille.

De plus, il est intéressant d'observer le graphique de la répartition de la structure des bâtiments selon la valeur assurée. En effet, il permet clairement d'illustrer la tendance suivante : plus la valeur assurée est élevée, plus la proportion de bâtiment possédant un niveau de conception parasismique supérieur est élevée.

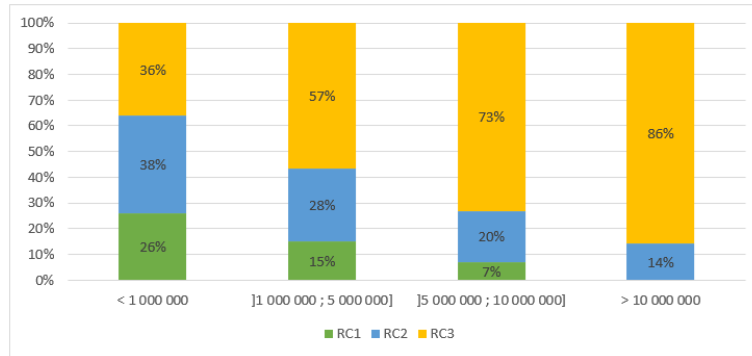


FIGURE 61 – Répartition de la structure des bâtiments assurés selon la valeur assurée.

Enfin, on observe la forte concentration des bâtiments assurés dans les zones urbanisées, notamment dans l'aire urbaine napolitaine. En moyenne, c'est également dans cette région que se trouvent les valeurs assurées les plus importantes.

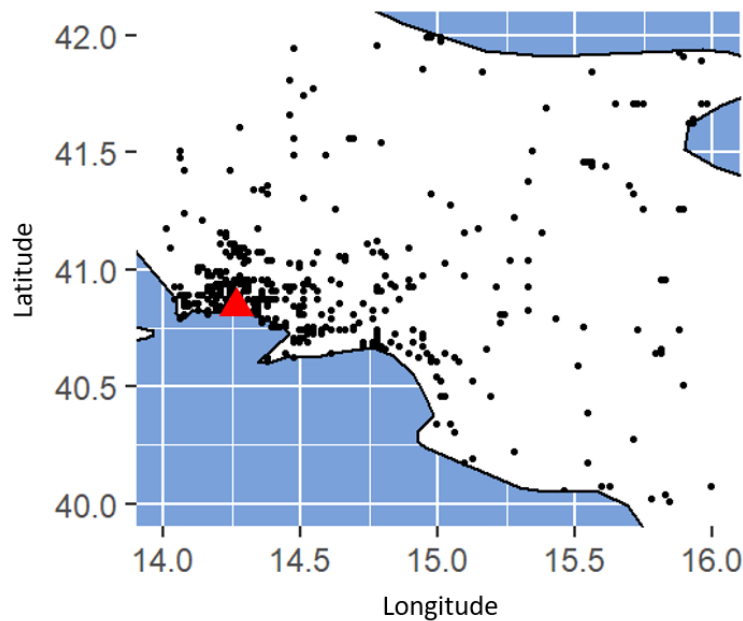


FIGURE 62 – Carte des sites en portefeuille. *Lecture : chaque point noir correspond à la localisation d'un site assuré. Le triangle rouge indique la ville de Naples.*

Il convient alors, à partir des informations de ce portefeuille, de construire la base de données qui permettra de réaliser les analyses de sensibilité de ce modèle étendu.

## II Construction de la base de données du modèle étendu

Contrairement aux parties précédentes, où il s'agissait d'étudier un modèle restreint où s'appliquait une intensité macrosismique centrée en VIII sur un seul et unique bâtiment, on considère désormais de multiples bâtiments, soumis à des niveaux d'intensité macrosismique variés et caractérisés par des valeurs assurées distinctes.

En cela, il est nécessaire d'appliquer certaines modifications à la méthode utilisée plus tôt permettant de construire la base de données. Ainsi, la méthode actuelle est constituée de trois grandes étapes : l'association des bâtiments assurés et des niveaux de MMI estimés en chaque point, puis la construction de la base de données par un échantillonnage représentatif des incertitudes des variables d'entrée du modèle, et enfin l'application de la méthode Risk-UE LM1 pour obtenir le *loss ratio* au niveau portefeuille.

### II.1 Association du portefeuille et des niveaux d'intensité macrosismique

La *shakemap* USGS et les sites assurés étant géocodés, il est directement possible d'associer à chaque bâtiment le niveau d'intensité macrosismique moyen estimé en chaque point.

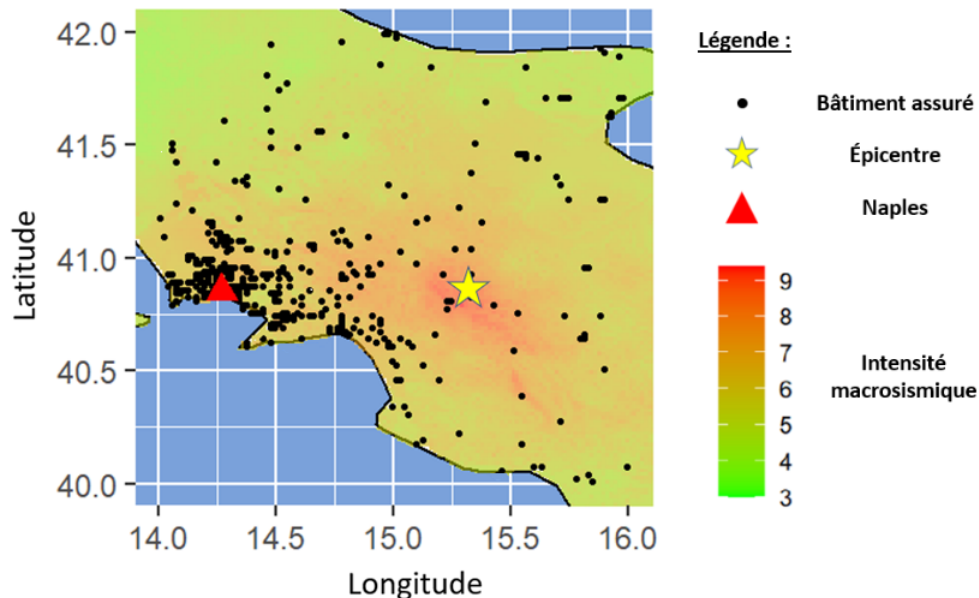


FIGURE 63 – Carte des sites en portefeuille associés à l'intensité macrosismique du séisme du 23 novembre 1980 en Irpinia.

À partir de ces données, il est possible de déterminer certaines statistiques descriptives sur les niveaux d'intensité macrosismique s'appliquant aux bâtiments assurés :

TABLE 29 – Statistiques descriptives de l'intensité macrosismique.

Min	1er Quartile	Médiane	Moyenne	3ème Quartile	Max
4,4	6	6,4	6,3	6,8	9

En utilisant un nuage de points, représentant pour chaque bâtiment sa valeur assurée et le niveau d'intensité macrosismique subi, on observe également que les biens dont les valeurs assurées sont les plus importantes sont principalement soumis à des niveaux d'intensité macrosismique moyens entre VI et VII.

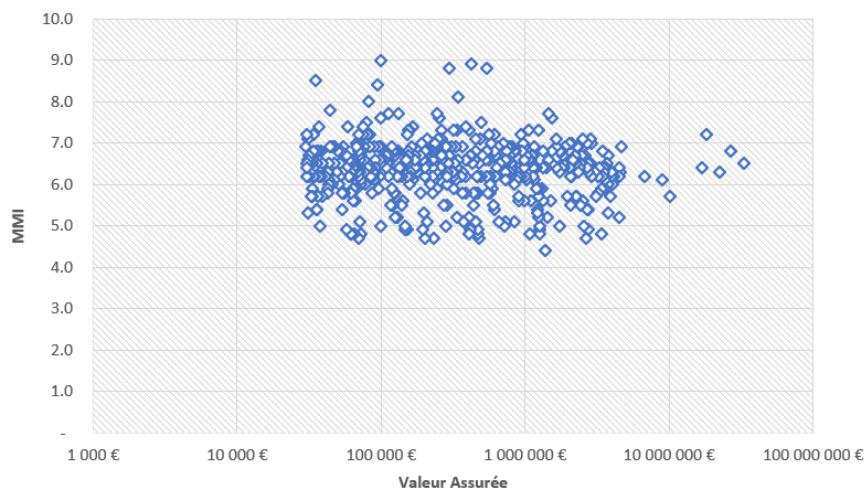


FIGURE 64 – Représentation pour chaque site en portefeuille de la valeur assurée et de l'intensité macrosismique subie.

## II.2 Représentation des incertitudes et échantillonnage

Dès lors, pour chaque bâtiment, un échantillonnage par hypercube latin quasi-similaire à celui utilisé pour constituer la base de données du modèle restreint est mis en oeuvre. Toutefois, deux différences majeures sont à relever :

- Le niveau d'intensité macrosismique moyen n'est plus supposé égal à VIII. Comme illustré en section précédente, chaque bâtiment subit désormais l'intensité estimée en un point par la *shakemap* USGS du séisme d'Irpinia. Pour simplifier l'étude et permettre une meilleure comparaison avec les résultats obtenus par les analyses de sensibilité du modèle restreint, le choix a été fait de conserver une représentation de l'incertitude par une loi normale d'écart type égal à 0.7.

- Pour permettre au logiciel R de réaliser les analyses de sensibilité, il est nécessaire de concaténer chaque réalisation des processus d'échantillonnage. Dès lors, chaque cellule de la matrice des variables d'entrée contient une chaîne de caractères représentant les réalisations du processus d'échantillonnage pour les 500 bâtiments.

La base de données obtenue est présentée ci dessous. Elle est constituée de 10 000 lignes et de 5 colonnes, à l'image de celle utilisée dans le cadre du modèle restreint.

STR	MOD	LR	MMI	VI
3, 1, 2, 2, 1, 1, 1, ...	2, 2, 2, 2, 2, 1, 2, ...	6, 2, 8, 9, 2, 9, 7, ...	6.390, 7.018, 7.024, 6.224, 6.302, 5.998, 4.822, ...	0.247, 0.096, 0.257, 0.373, 0.346, 0.311, 0.529, ...
2, 1, 3, 1, 1, 1, 1, ...	1, 2, 2, 1, 2, 1, 2, ...	8, 3, 9, 5, 5, 7, 3, ...	6.265, 7.286, 6.723, 6.703, 5.631, 5.885, 6.233, ...	0.221, 0.415, 0.203, 0.163, 0.268, 0.334, 0.337, ...
3, 2, 3, 2, 3, 1, 1, ...	2, 2, 1, 1, 1, 2, 1, ...	4, 4, 8, 1, 2, 8, 4, ...	6.286, 6.409, 6.857, 7.682, 5.588, 6.034, 5.656, ...	0.172, 0.536, 0.187, 0.206, 0.388, 0.466, 0.283, ...
1, 1, 1, 2, 3, 3, 3, ...	1, 1, 2, 1, 1, 2, 2, ...	9, 6, 5, 1, 1, 6, 4, ...	7.742, 6.763, 6.781, 7.657, 7.413, 5.161, 5.599, ...	0.374, 0.053, 0.365, 0.162, 0.218, 0.326, 0.481, ...
1, 2, 2, 1, 2, 2, 2, ...	1, 1, 1, 1, 1, 1, 1, ...	8, 1, 5, 5, 3, 6, 1, ...	7.056, 7.200, 5.950, 7.086, 5.639, 5.482, 5.044, ...	0.395, 0.312, 0.455, 0.483, 0.019, 0.295, 0.149, ...

FIGURE 65 – Les cinq premières lignes de la matrice des variables d'entrée du modèle étendu.

### II.3 Application de la méthode Risk-UE LM1 et obtention du *loss ratio* au niveau portefeuille

A partir de cette matrice obtenue à la suite des processus d'échantillonnages, il est désormais possible de déterminer les différentes réalisations du vecteur de pertes  $\mathbf{Y}$ . Celles-ci sont calculées en appliquant la méthode présentée en partie 2 : le module de vulnérabilité RISK-UE.

Toutefois, on s'intéresse désormais au *loss ratio* global du portefeuille et non plus à celui d'un seul et unique bâtiment. Il est donc nécessaire de prendre en compte la valeur assurée de chaque construction. Pour cela, il faut d'abord introduire la matrice intermédiaire suivante :

Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	...	Y_500
0,0002	0,0012	0,0039	0,0016	0,0006	0,0036	...	0,00015
0,0021	0,0064	0,0008	0,0017	0	0,0002	...	0,0242
0,0001	0,0009	0,003	0,004	0,0015	0,003	...	0,003
0,0146	0,0004	0,0034	0,0075	0,0055	0	...	0,0027
0,0066	0,0075	0,0014	0,0054	0,0001	0,0003	...	0,0421

FIGURE 66 – Les cinq premières lignes de la matrice des *loss ratios* de chaque bâtiment.

Ainsi, cette matrice intermédiaire, constituée de 10 000 lignes et de 500 colonnes, représente les différentes réalisations du vecteur de pertes pour chacun des 500 bâtiments assurés.

Dès lors, le vecteur final de pertes  $\mathbf{Y}$ , représentant les réalisations du *loss ratio* global du portefeuille, est obtenu par le calcul suivant :

$$Y_i = \frac{1}{TIV} \sum_{j=1}^{500} Y_{i,j} \times IV_j, \text{ pour chaque } i \in \{1, \dots, 10000\} \quad (81)$$

avec :

- $Y_{i,j}$  le *loss ratio* du bâtiment  $j$  lors de la réalisation  $i$  ;
- $IV_j$  la valeur assurée du bâtiment  $j$  ;
- $TIV$  la valeur assurée totale du portefeuille telle que  $TIV = \sum_j IV_j$ .

À partir de ces données il est possible de déterminer certaines statistiques descriptives de  $\mathbf{Y}$  ainsi que que tracer l'histogramme et la fonction de répartition de sa distribution.

TABLE 30 – Statistiques descriptives de  $\mathbf{Y}$

Min	1er Quartile	Médiane	Moyenne	3ème Quartile	Max
0,0173	0,0296	0,0342	0,0373	0,0414	0,169

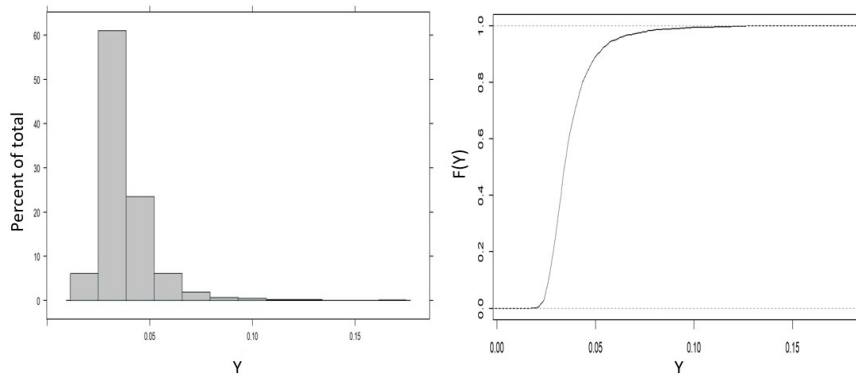


FIGURE 67 – Histogramme et fonction de répartition empirique du vecteur de pertes simulées.

La base de données étant constituée, il est alors possible de réaliser les études de sensibilité globale et locale en reprenant les méthodes présentées précédemment.

### III Résultats des analyses de sensibilité

En suivant les méthodes utilisées dans le cadre du modèle restreint, il est désormais possible de réaliser, dans le cadre du modèle étendu, une analyse globale de sensibilité à partir des indices de Shapley, puis d'obtenir un indicateur d'importance de contribution grâce à une analyse locale de sensibilité et aux valeurs SHAP.

#### III.1 Indices de Shapley

Comme elle permet la parallélisation des calculs et la prise en compte des variables catégorielles, c'est à nouveau la fonction *shapleysobol\_knn* de la librairie *sensitivity* qui a été utilisée afin d'obtenir les indices de Shapley. Les résultats obtenus sont présentés dans la table suivante :

TABLE 31 – Indices de Shapley des variables d'entrée du modèle étendu. *Lecture : les intervalles de confiance ont été déterminés par bootstrap non-paramétrique à partir de 1000 simulations.*

Variable	Indice de Shapley	Intervalle de confiance à 95%
<b>VI</b>	53,6%	[52,8% ; 53,7%]
<b>MMI</b>	27,2%	[26,8% ; 27,5%]
<b>MOD</b>	5,9%	[5,8% ; 6,3%]
<b>LR</b>	5,5%	[5,3% ; 5,6%]
<b>STR</b>	7,8%	[7,3% ; 7,8%]

Comme l'on pouvait s'y attendre, l'incertitude de la variable de sortie du modèle est à nouveau principalement associée aux variables **VI** et **MMI** représentant respectivement l'indice de vulnérabilité d'une structure et l'intensité des mouvements du sol. En particulier, la valeur de l'indice de Shapley associé à **VI** confirme également l'enjeu important que doit représenter sa détermination pour un assureur souhaitant diminuer la volatilité et augmenter la précision de son modèle.

Toutefois, par rapport au modèle restreint, on observe une relative diminution de l'indice de Shapley de la variable **VI** ainsi qu'une plus forte baisse encore de celui de la variable **MMI**. Ces diminutions sont compensées par une contribution plus importante de la variable **STR** dans l'incertitude du modèle. Pour ce qui est des contributions respectives des variables **MOD** et **LR**, elles ne présentent qu'une légère augmentation par rapport au modèle restreint.



De plus, comme l'on dispose désormais de réalisations des vecteurs de pertes pour des bâtiments exposés à des niveaux variés d'intensité macrosismique, il est possible d'observer l'évolution des résultats du modèle restreint à d'autres niveaux que VIII. En raison du temps de calcul important nécessaire à l'obtention des indices de Shapley, il a été décidé de restreindre leurs déterminations à quatre sites sur lesquels s'appliquent des niveaux d'intensité macrosismique égal à V, VI, VII et VIII.

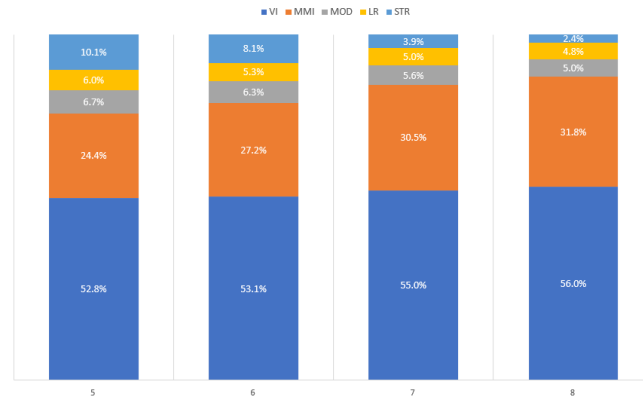


FIGURE 68 – Indices de Shapley des variables d'entrée du modèle étendu à différents niveaux d'intensité macrosismique.

Ainsi, avec la réduction du niveau d'intensité macrosismique, on observe clairement une diminution de la part d'incertitude du modèle associée à la variable **MMI**. Cette diminution semble principalement compensée par la croissance notable de l'indice de Shapley associée à la variable **STR** avec la baisse du niveau d'intensité macrosismique.

Enfin, avant de conclure cette étude, il semble aussi particulièrement intéressant de réaliser une analyse de sensibilité locale en considérant le modèle étendu.

### III.2 Indicateur d'importance de contribution

Dans la section précédente, il a été possible de comparer directement les résultats obtenus en considérant le modèle restreint et le modèle étendu. Malheureusement, dans le cas de l'analyse locale de sensibilité, cette comparaison est plus complexe. En effet, pour construire la nouvelle base de données des variables d'entrée du modèle, il a été nécessaire d'allouer à chaque cellule une chaîne de caractères représentant les réalisations du processus d'échantillonnage pour les 500 bâtiments.

Par conséquent, chaque variable est désormais catégorielle et possède au plus 10 000 modalités distinctes. Mais, comme le montre la figure 56, présentant les Valeurs de l'indicateur SHAP pour chacune des 8 variables les plus importantes du modèle restreint, l'algorithme permettant de déterminer les valeurs SHAP ne s'applique qu'au niveau des modalités et non à celui des variables catégorielles en tant qu'ensemble. Une modalité pouvant par exemple être de la forme "1, 2, 2, 1, 3, 1, ..." pour la variable **STR**, l'intérêt de déterminer et de comparer les indicateurs SHAP d'importance de contribution est donc relativement limité.

Toutefois, il existe une question à laquelle une nouvelle analyse de sensibilité locale permet de répondre : quels sont les sites qui contribuent le plus à la perte au niveau portefeuille? Pour y répondre, il est nécessaire de ne plus considérer  $X = (STR, MOD, LR, MMI, VI)$  comme matrice d'entrées mais bien  $Y = (Y_1, Y_2, \dots, Y_{500})$ , étant la matrice intermédiaire définie plus tôt représentant les différentes réalisations du vecteur de pertes pour chacun des 500 bâtiments assurés.

Dès lors, l'extension de la méthode KernelSHAP a de nouveau été appliquée sur un modèle *random forest* entraîné sur la matrice des vecteurs de pertes de chacun des sites assurés. La détermination de l'indicateur d'importance de contribution se fait ensuite directement en moyennant les valeurs SHAP de la base d'apprentissage pour chaque variable. La table ci-dessous présente alors, pour les huit contributions les plus importantes, le rang de la valeur assurée du bâtiment au niveau du portefeuille :

TABLE 32 – Comparaison des rangs de l'indicateur d'importance de contribution et de la valeur assurée du bâtiment.

Rang de l'indicateur d'importance de contribution	Rang de la valeur assurée du bâtiment dans le portefeuille
1	6
2	4
3	9
4	1
5	16
6	7
7	2
8	12

Ainsi, les résultats obtenus semblent confirmer l'intuition suivante : en moyenne, l'incertitude sur le *loss ratio* au niveau du portefeuille est principalement attribuée aux incertitudes sur les dommages aux bâtiments dont les valeurs assurées sont les plus importantes. En cela, ce sont eux qui contribuent le plus à la perte au niveau portefeuille.

## IV Conclusion

L'objectif de cette étude était de réaliser une analyse de sensibilité d'un modèle de risque sismique, utilisé pour estimer les expositions des assureurs et réassureurs aux tremblements de terre et ainsi optimiser leurs traités de réassurance et besoins en capital réglementaire.

Plus précisément, il s'agissait ici d'étudier la méthodologie Risk-UE, développée dans le cadre d'un projet européen d'évaluation du risque sismique de grandes villes européennes. Plusieurs variables ont été retenues : la structure du bâtiment **STR**, l'ensemble de *loss ratios* **LR**, la distribution des niveaux de dommage **MOD**, l'indice de vulnérabilité d'une structure **VI** et l'intensité macrosismique **MMI**. Dès lors, il s'agissait de répondre à deux questions soulevées en introduction :

1. Dans quelle mesure chaque variable est responsable de l'incertitude associée à la valeur de la perte estimée ?
2. Quelles sont les variables qui contribuent le plus à cette même valeur ?

Tout d'abord, la détermination des indices de Shapley a permis de montrer que la variabilité de la perte estimée pouvait être attribuée en majorité à l'incertitude sur l'indice de vulnérabilité basé sur la théorie de la logique floue. Ainsi, pour accroître la précision du modèle, les efforts devraient être principalement concentrés sur les moyens de réduire cette incertitude et produire des courbes de vulnérabilité représentatives contrairement à l'intensité macrosismique qui, même si elle responsable d'une partie conséquente de l'incertitude, est plus complexe à réduire.

Les incertitudes associées à la structure du bâtiment, à l'ensemble de *loss ratios* et à la distribution des niveaux de dommage ne représentent qu'une part négligeable de l'incertitude des pertes. Ces conclusions ont par la suite été confirmées à l'échelle globale d'un portefeuille d'assurance lors de l'étude d'un sinistre historique, mais aussi à l'échelle individuelle à différents niveaux d'intensité macrosismique. Notons toutefois que ce constat aurait pu être différent si l'on avait considéré d'autres structures comme, par exemple, le bois et l'acier. Rappelons que cette approche n'a pas été retenue dans cette étude, comme il s'agissait plutôt de représenter l'erreur opérationnelle associée à la mauvaise catégorisation d'une structure.

Ensuite, pour déterminer les variables qui contribuaient le plus à la valeur estimée des pertes, il s'agissait de déterminer un modèle de substitution interprétable. Plusieurs ont été étudiés : régression linéaire multiple, régression bêta, GAMLSS, MARS, arbres de régression... Il s'est avéré que celui répliquant la physique du modèle le plus fidèlement était basé sur la notion de forêts aléatoires, méthode d'apprentissage en-

sembliste de type boîte noire. Pour l'interpréter, il était donc nécessaire d'utiliser une méthode issue du domaine de l'IA Explicable : les valeurs SHAP. Elles permettent notamment de réaliser des études de sensibilité locale et de construire un indicateur d'importance de contribution des variables.

Bien que la mesure de ces conclusions doive être prise avec précaution, cet indicateur a permis de montrer que l'indice de vulnérabilité avait le plus d'influence sur la valeur de perte, suivi de près par l'intensité macrosismique. Exception faite d'une structure et de quelques ensembles de loss ratios, l'influence des autres variables peut être considérée comme minime. Enfin, au niveau portefeuille, il apparaît clairement que l'effort de réduction des incertitudes doit être concentré sur les bâtiments dont les valeurs assurées sont les plus importantes.

Pour poursuivre cette étude, il peut être pertinent de vérifier si ces conclusions sont toujours vraies en considérant différents types de structure. De même, l'influence de jugements d'experts permettant de ne plus avoir à considérer des structures équiprobables devrait être approfondi. Enfin, il peut être particulièrement intéressant de déterminer l'effet de diverses structures de réassurance, n'ayant pas été prises en compte dans ce mémoire. En particulier, la considération de traités en excédent de pertes ou d'obligations catastrophes paramétriques pourrait sans doute modifier l'importance de l'intensité macrosismique.

## Annexes

### Annexe A : Détails de la mise en place de la régression linéaire multiple.

```

Coefficients:
      Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.1145132  0.0118814   9.638 < 2e-16 ***
STR2         0.0002463  0.0014632   0.168  0.86630
STR3        -0.0027068  0.0014890  -1.818  0.06911 .
MOD2         0.0111133  0.0065800   1.689  0.09127 .
LR2          0.1152813  0.0139593   8.258 < 2e-16 ***
LR3          0.0148784  0.0136114   1.093  0.27439
LR4          0.0336482  0.0135160   2.489  0.01281 *
LR5          0.1437231  0.0136449  10.533 < 2e-16 ***
LR6          0.1090460  0.0138671   7.864 4.22e-15 ***
LR7          0.1254389  0.0136775   9.171 < 2e-16 ***
LR8          0.0106064  0.0137854   0.769  0.44168
LR9          0.0341739  0.0136140   2.510  0.01209 *
MMI         -0.0181024  0.0014677 -12.334 < 2e-16 ***
VI          -1.1988479  0.0218671 -54.824 < 2e-16 ***
STR2:VI     -0.0001936  0.0045168  -0.043  0.96581
STR3:VI      0.0119145  0.0045297   2.630  0.00855 **
MOD2:LR2     0.0000364  0.0023873   0.015  0.98784
MOD2:LR3     0.0044477  0.0023620   1.883  0.05973 .
MOD2:LR4     0.0018569  0.0023627   0.786  0.43195
MOD2:LR5     0.0002766  0.0024036   0.115  0.90838
MOD2:LR6    -0.0017454  0.0023882  -0.731  0.46489
MOD2:LR7     0.0014737  0.0024123   0.611  0.54128
MOD2:LR8    -0.0056047  0.0023867  -2.348  0.01888 *
MOD2:LR9    -0.0051650  0.0023685  -2.181  0.02923 *
MOD2:MMI    -0.0016812  0.0007968  -2.110  0.03488 *
LR2:MMI     -0.0127200  0.0017142  -7.420 1.29e-13 ***
LR3:MMI     -0.0017706  0.0016779  -1.055  0.29133
LR4:MMI     -0.0036995  0.0016633  -2.224  0.02616 *
LR5:MMI     -0.0162027  0.0016756  -9.670 < 2e-16 ***
LR6:MMI     -0.0120163  0.0017097  -7.028 2.26e-12 ***
LR7:MMI     -0.0143263  0.0016767  -8.544 < 2e-16 ***
LR8:MMI     -0.0002172  0.0016993  -0.128  0.89829
LR9:MMI     -0.0037491  0.0016732  -2.241  0.02507 *
LR2:VI      -0.0915988  0.0077224 -11.861 < 2e-16 ***
LR3:VI      -0.0462628  0.0074915  -6.175 6.93e-10 ***
LR4:VI      -0.0528884  0.0075694  -6.987 3.03e-12 ***
LR5:VI      -0.1175755  0.0077690 -15.134 < 2e-16 ***
LR6:VI      -0.0998316  0.0074841 -13.339 < 2e-16 ***
LR7:VI      -0.1006083  0.0078110 -12.880 < 2e-16 ***
LR8:VI      -0.0455534  0.0076064  -5.989 2.21e-09 ***
LR9:VI      -0.0210935  0.0076868  -2.744  0.00608 **
MMI:VI       0.1775635  0.0026400  67.260 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.0251 on 7958 degrees of freedom
Multiple R-squared:  0.6881,    Adjusted R-squared:  0.6865
F-statistic: 428.2 on 41 and 7958 DF,  p-value: < 2.2e-16

```

FIGURE 69 – *Summary* du modèle linéaire multiple.

## Annexe B : Détails de la mise en place de la régression bêta.

```

Coefficients (mean model with logit link):
      Estimate Std. Error z value Pr(>|z|)
(Intercept) -10.75143    0.57046  -18.847 < 2e-16 ***
VI           -0.30495    0.90306   -0.338 0.735599
MMI          0.61631    0.06764    9.111 < 2e-16 ***
LR2         -0.03885    0.51515   -0.075 0.939889
LR3         -3.86020    0.73856  -5.227 1.73e-07 ***
LR4         -2.32256    0.52131  -4.455 8.38e-06 ***
LR5         -2.51053    0.52530  -4.779 1.76e-06 ***
LR6         -2.33142    0.80227  -2.906 0.003661 **
LR7         -3.00237    0.65730  -4.568 4.93e-06 ***
LR8         -1.77557    0.47777  -3.716 0.000202 ***
LR9         -1.03753    0.47671  -2.176 0.029523 *
MOD2        -1.88122    0.30444  -6.179 6.44e-10 ***
VI:MMI      0.85225    0.11026    7.729 1.08e-14 ***
VI:LR2     -0.40090    0.27460  -1.460 0.144309
VI:LR3      1.77146    0.39284    4.509 6.50e-06 ***
VI:LR4      0.02193    0.35195    0.062 0.950306
VI:LR5     -0.09837    0.30795   -0.319 0.749396
VI:LR6     -0.36958    0.38590   -0.958 0.338208
VI:LR7      1.24813    0.32415    3.851 0.000118 ***
VI:LR8      0.39066    0.28135    1.388 0.164986
VI:LR9     -0.38522    0.28024  -1.375 0.169258
MMI:LR2    -0.04296    0.05950   -0.722 0.470340
MMI:LR3     0.31878    0.08017    3.976 7.00e-05 ***
MMI:LR4     0.20415    0.05968    3.421 0.000624 ***
MMI:LR5     0.19126    0.05786    3.306 0.000947 ***
MMI:LR6     0.19287    0.09023    2.138 0.032548 *
MMI:LR7     0.20438    0.07338    2.785 0.005350 **
MMI:LR8     0.15224    0.05532    2.752 0.005920 **
MMI:LR9     0.13279    0.05558    2.389 0.016875 *
VI:MOD2     0.86513    0.16260    5.321 1.03e-07 ***
MMI:MOD2    0.16258    0.03478    4.674 2.95e-06 ***
VI:STR2     0.12525    0.08106    1.545 0.122341
VI:STR3    -0.05320    0.07382   -0.721 0.471100
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

FIGURE 70 – *Summary* du modèle de régression bêta : Paramètre de moyenne.

```

Phi coefficients (precision model with log link):
      Estimate Std. Error z value Pr(>|z|)
(Intercept)  10.2814    1.9174    5.362 8.23e-08 ***
STR2         -4.3118    1.6338   -2.639 0.00831 **
STR3         1.5578    1.7870    0.872 0.38337
MOD2         0.1949    0.2773    0.703 0.48198
LR2          1.3829    0.5776    2.394 0.01665 *
LR3         -0.6886    0.6267   -1.099 0.27186
LR4          0.5516    0.5917    0.932 0.35119
LR5         -0.2661    0.6435   -0.414 0.67917
LR6          0.3327    0.5908    0.563 0.57334
LR7          0.2013    0.5803    0.347 0.72869
LR8         -0.4291    0.6033   -0.711 0.47694
LR9         -0.4662    0.6244   -0.747 0.45532
MMI         -0.3623    0.2281   -1.588 0.11227
VI           8.2114    4.7424    1.731 0.08337 .
STR2:MMI     0.6131    0.2026    3.027 0.00247 **
STR3:MMI    -0.1497    0.2205   -0.679 0.49726
MOD2:VI     -1.9685    0.8396   -2.345 0.01904 *
LR2:VI      -3.8877    1.6115   -2.413 0.01584 *
LR3:VI      -0.2366    1.6942   -0.140 0.88892
LR4:VI      -2.4248    1.7838   -1.359 0.17403
LR5:VI       1.3140    1.7653    0.744 0.45667
LR6:VI      -1.8068    1.7316   -1.043 0.29676
LR7:VI      -0.3706    1.6247   -0.228 0.81958
LR8:VI       3.0981    1.7925    1.728 0.08392 .
LR9:VI       1.1810    1.8204    0.649 0.51651
MMI:VI      -1.2450    0.5738   -2.170 0.03001 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

FIGURE 71 – *Summary* du modèle de régression bêta : Paramètre de précision.

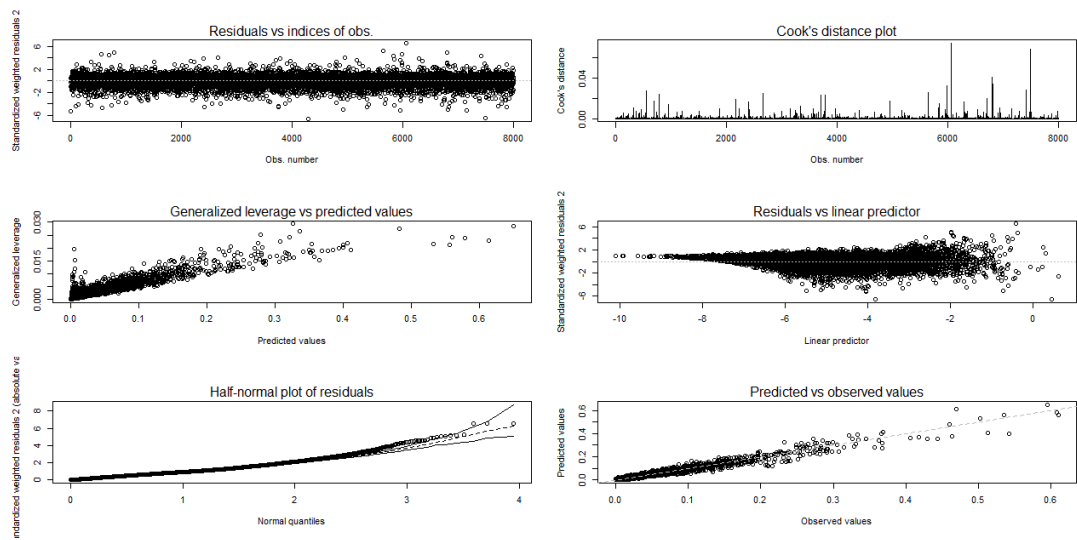


FIGURE 72 – Graphiques de diagnostic du modèle de régression bêta. *Lecture : contrairement aux diagnostic plots d'un GLM, ceux d'une régression bêta ne permettent pas de tirer de conclusions définitives sur l'ajustement du modèle. De façon générale, les résidus d'une régression bêta n'ont pas à être normalement distribués.*

## Annexe C : Détails de la mise en place du GAMLSS bêta inflaté en 0 et 1.

```

Family: c("BEINF", "Beta Inflated")

Mu Coefficients:
(Intercept)      MMI      VI      LR.1      LR.9
-17.70473      1.27172      6.01997      2.36108      1.93301
  LR.8      LR.5      LR.7      LR.6      MOD.1
  0.80350     -0.73699     -2.22887     -0.43080     2.03318
VI:MOD.1  VI:LR.1  MMI:MOD.1  VI:LR.9  MMI:LR.1
-1.07342     -0.91359     -0.16729     -0.75925     -0.18781
MMI:LR.9  MMI:VI  VI:LR.7  MMI:LR.7  VI:LR.8
-0.15397     0.31904     0.94080     0.16283     -0.34902
VI:LR.5  MMI:LR.8  MMI:LR.5  LR.5:MOD.1
  0.07285     -0.05886     0.02312     -0.05170

Sigma Coefficients:
(Intercept)      VI      LR.3      MMI      LR.4
-5.81581      1.59345     0.74440     0.24542     0.96025
  MOD.2      LR.5      LR.2      LR.6      LR.9
  0.25399     -0.21385     -0.13784     -0.66436     1.07442
LR.3:MOD.2  VI:MOD.2  VI:LR.3  VI:LR.4  LR.4:MOD.2
-0.19196     -0.28850     -0.35287     -0.33378     -0.12308
MMI:LR.6  MOD.2:LR.2  VI:LR.2  VI:LR.6  MMI:LR.9
  0.06156     -0.11839     0.32502     0.30600     -0.12972
MMI:LR.4
-0.07316

Nu Coefficients:
(Intercept)      MMI      VI      MOD.1      STR.2
20.0893     -2.8948     -14.7407     -2.9699     0.5364
  LR.6      LR.9      LR.7
  0.7605     0.7519     0.5582

Tau Coefficients:
(Intercept)
-25.65

Degrees of Freedom for the fit: 54 Residual Deg. of Freedom 7946
Global Deviance: -69743.8
AIC: -69635.8
SBC: -69258.5

```

FIGURE 73 – *Summary* du GAMLSS bêta inflaté en 0 et 1.

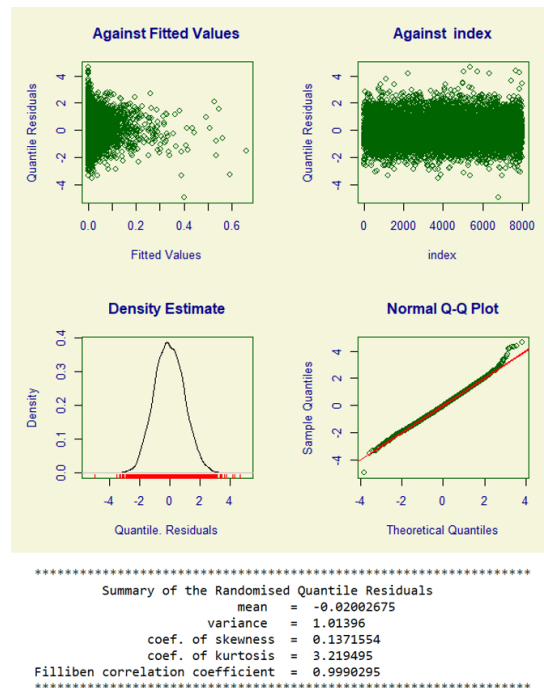


FIGURE 74 – Graphiques de diagnostic du GAMLSS bêta inflaté en 0 et 1.



## Annexe D : Détails de la mise en place du modèle MARS

Le modèle MARS possède un certain nombre de paramètres qui doivent être ajustés pour optimiser ses performances et limiter le risque de sur-apprentissage. En particulier, il est possible d'ajuster les paramètres suivants :

- Le degré d'interaction, généralement un nombre entier compris entre 1 et 3, indiquant le nombre maximal de variables qui peuvent interagir en même temps ;
- Le nombre de termes du modèle final après un processus de régularisation permettant de supprimer les termes non significatifs. Ce processus est appelé *pruning*.

Le choix des paramètres optimaux a été effectué par une approche *grid search*. Celle-ci permet de parcourir de manière systématique toutes les combinaisons possibles de valeurs de paramètres et de sélectionner la combinaison qui donne les meilleures performances. Bien qu'elle soit assez coûteuse en termes de temps de calcul, cette méthode est très appréciée car elle est simple à mettre en place et permet d'obtenir une combinaison optimale sans nécessiter de connaissances spécifiques.

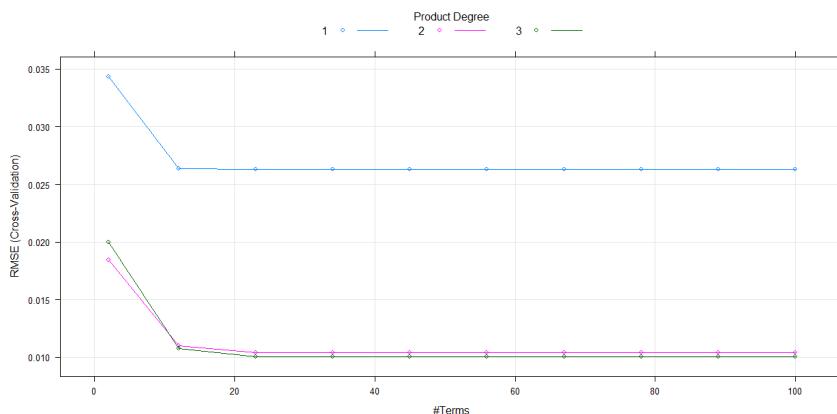


FIGURE 75 – Exemple de construction d'un modèle MARS étape par étape.

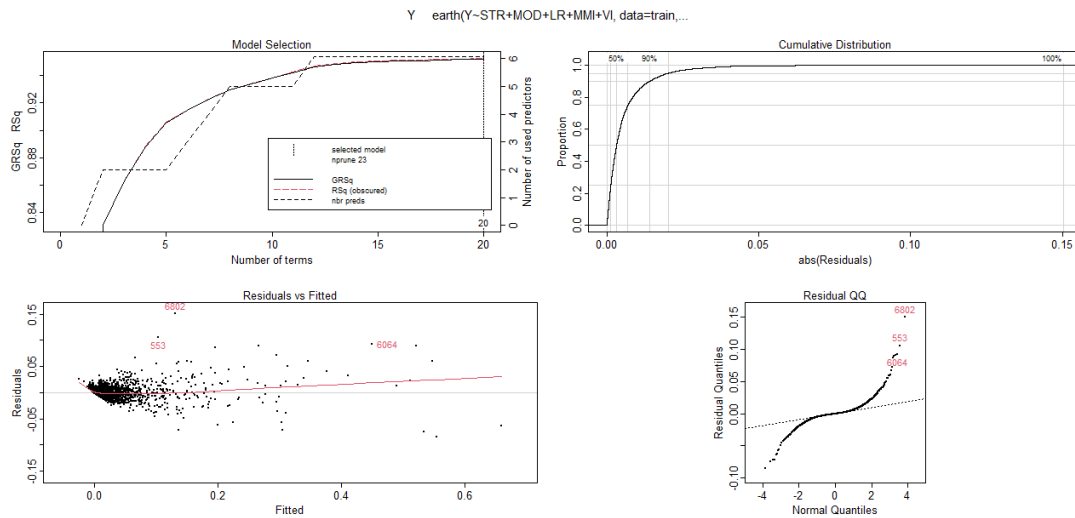


FIGURE 76 – Graphiques de diagnostic de modèle MARS.

	Coefficients
(Intercept)	0.02184568
h(VI-0.493582)	0.47220492
h(0.493582-VI)	-0.04765758
h(MMI-8.17898)*h(VI-0.493582)	0.75852427
h(8.17898-MMI)*h(VI-0.493582)	-0.27942398
h(MMI-8.61728)	0.20177661
h(8.61728-MMI)	-0.01183717
h(MMI-8.61639)*h(0.493582-VI)	-0.76458368
h(8.61639-MMI)*h(0.493582-VI)	0.02633507
h(MMI-8.61728)*h(VI-0.283181)	-0.29858951
h(MMI-8.61728)*h(0.283181-VI)	0.59883539
LR5*h(MMI-8.61728)*h(VI-0.283181)	-0.41968434
h(VI-0.336606)	0.06217394
LR6*h(VI-0.493582)	-0.14013233
h(MMI-7.51662)*h(VI-0.336606)	0.25849250
h(7.51662-MMI)*h(VI-0.336606)	0.04994168
LR5*h(VI-0.336606)	-0.16011277
LR7*h(VI-0.336606)	-0.15284325
LR6*h(MMI-7.51662)*h(VI-0.336606)	-0.17602570
LR2*h(MMI-7.51662)*h(VI-0.336606)	-0.16117672

GLM (family gaussian, link identity):

nulldev	df	dev	df	devratio	AIC	iters	converged
16.0741	7999	0.764908	7980	0.952	-51300	2	1

Earth selected 20 of 20 terms, and 6 of 13 predictors (nprune=23)  
Termination condition: Reached nk 27  
Importance: MMI, VI, LR5, LR6, LR7, LR2, STR2-unused, STR3-unused, ...  
Number of terms at each degree of interaction: 1 5 11 3  
Earth GCV 9.67834e-05 RSS 0.7649077 GRSq 0.9518435 RSq 0.9524137 CVRSq 0.9467706

FIGURE 77 – Summary du modèle MARS.

## Annexe E : Optimisation des hyperparamètres du modèle XGBoost

Les hyperparamètres d'un modèle XGBoost sont des variables qui déterminent comment le modèle est entraîné et ajusté. Les plus importants du modèle XGBoost sont notamment :

- *max\_depth* : la profondeur maximale de chaque arbre de décision dans le modèle. Augmenter cette valeur permet de capturer des relations plus complexes entre les variables, mais peut également aboutir à du sur-apprentissage ;
- *eta* : la taille de chaque pas d'ajustement des poids du modèle. Une valeur faible indique que les poids seront mis à jour lentement, ce qui permet au modèle de converger avec une meilleure précision ;
- *min\_child\_weight* : le poids minimum qu'un enfant doit avoir pour être créé lors de la construction de l'arbre. Augmenter cette valeur permet au modèle de réduire la complexité de l'arbre en évitant de créer des nœuds qui n'ont pas suffisamment de poids.
- *subsample* : la fraction de l'ensemble de données d'entraînement qui est utilisée pour construire chaque arbre de décision ;
- *colsample\_bytree* : la fraction de variables qui sont sélectionnées au hasard pour chaque arbre de décision ;
- *gamma* : contrôle la complexité des arbres en éliminant les branches qui ne contribuent pas suffisamment à la réduction de l'erreur de prédiction.

Comme pour le modèle MARS, le choix des hyperparamètres optimaux a été effectué par une approche *grid search*.

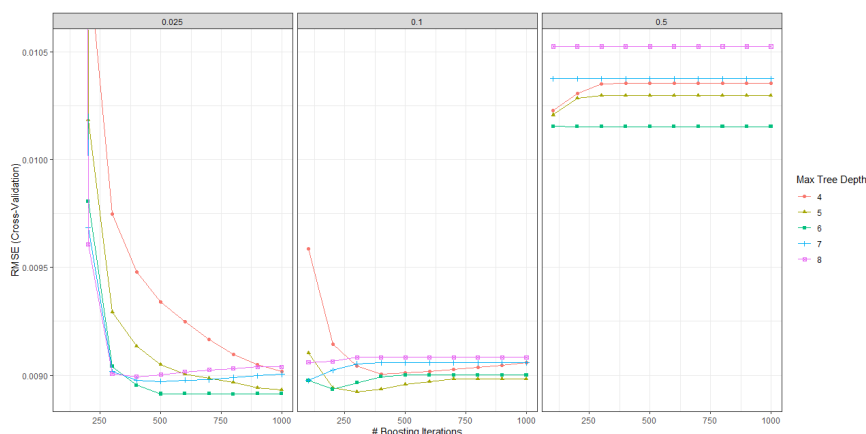


FIGURE 78 – Optimisation des hyperparamètres *max\_depth* et *eta* du modèle XGBoost.

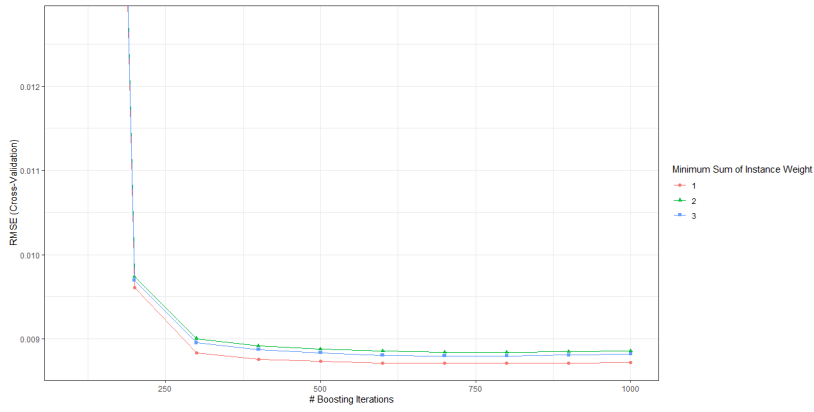


FIGURE 79 – Optimisation de l’hyperparamètre *min\_child\_weight* du modèle XGBoost.

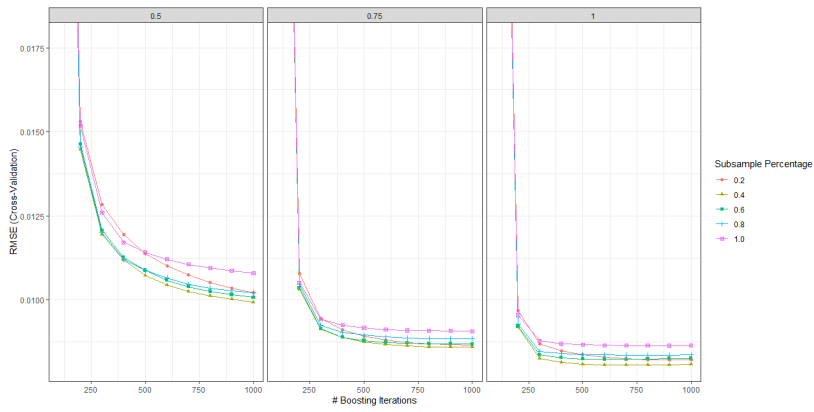


FIGURE 80 – Optimisation des hyperparamètres *colsample\_bytree* et *subsample* du modèle XGBoost.

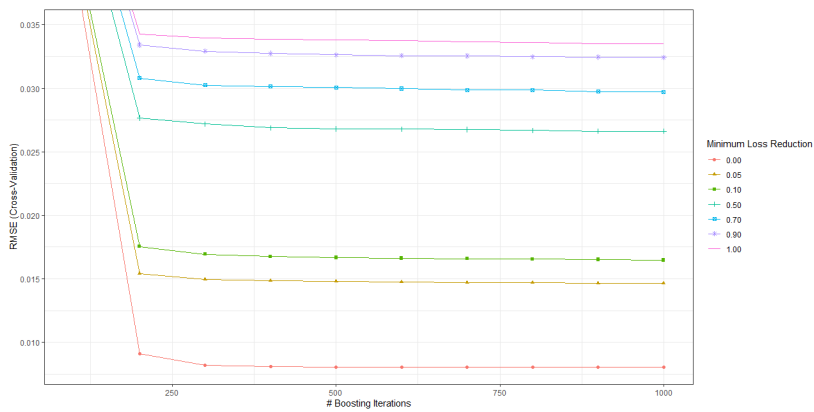


FIGURE 81 – Optimisation de l’hyperparamètre *gamma* du modèle XGBoost.

## Annexe F : Quelques outils graphiques d'IA Explicable

### Partial Dependence Plot (PDP)

Les *Partial Dependence Plots* sont des graphiques qui permettent de visualiser la relation entre une variable et la sortie d'un modèle, tout en maintenant les autres variables explicatives constantes. En particulier, ils peuvent aider à comprendre comment les différentes variables contribuent à la prédiction globale du modèle.

La théorie repose sur la notion de fonction de dépendance partielle  $\hat{f}_{S,PDP}$ . Celle-ci est calculée en moyennant les prédictions du modèle sur toutes les observations tout en faisant varier la valeur de la variable d'intérêt. En notant  $S$  l'ensemble des variables d'intérêt et  $C$  les autres variables présentes dans le modèle, un *Partial Dependence plot* est une représentation de la fonction :

$$\hat{f}_{S,PDP}(x) = \mathbb{E}_{X_C} [\hat{f}(x_S, X_C)] = \int_{X_C} \hat{f}(x_S, X_C) d\mathbb{P}(X_C) \quad (82)$$

Ainsi, le PDP montre comment la réponse moyenne du modèle varie en fonction de la variable explicative d'intérêt. Toutefois, un tel outil est biaisé dans le sens où il peut considérer des combinaisons ne respectant pas la logique du modèle initial. Dans le cas de cette étude, si la variable d'intérêt est l'indice de vulnérabilité, cela pourrait être de considérer une structure *RC3* à  $VI = 0,25$  qui n'appartient pas au support de la fonction d'appartenance de cette structure.

En présence d'interaction ou de corrélation importante, il est donc primordial de ne déduire des conclusions d'un PDP qu'avec la plus grande prudence.

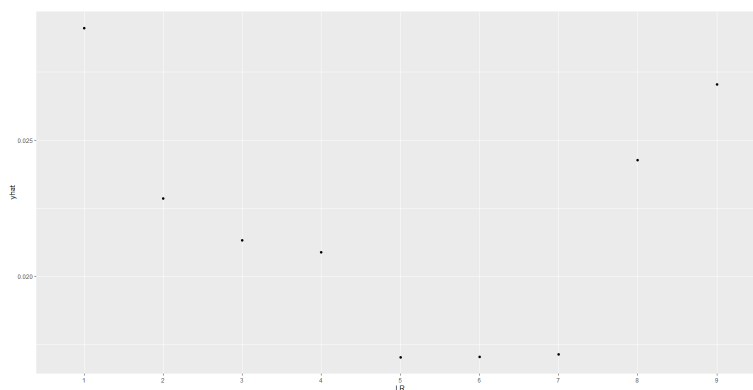


FIGURE 82 – Partial dependance plot de la variable **LR**.

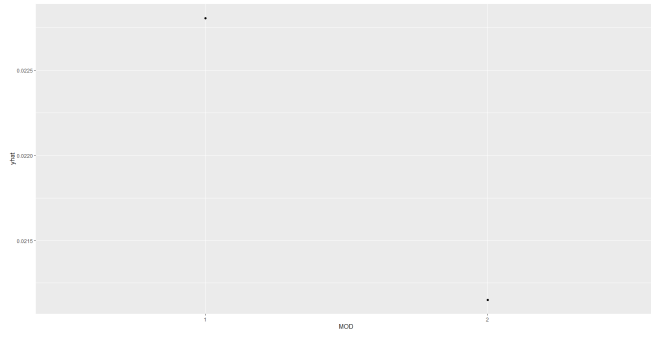


FIGURE 83 – Partial dependance plot de la variable **MOD**.

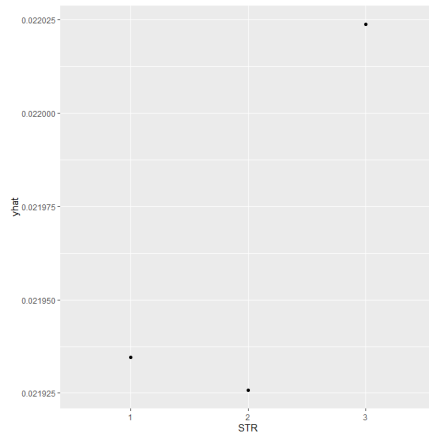


FIGURE 84 – Partial dependance plot de la variable **STR**.

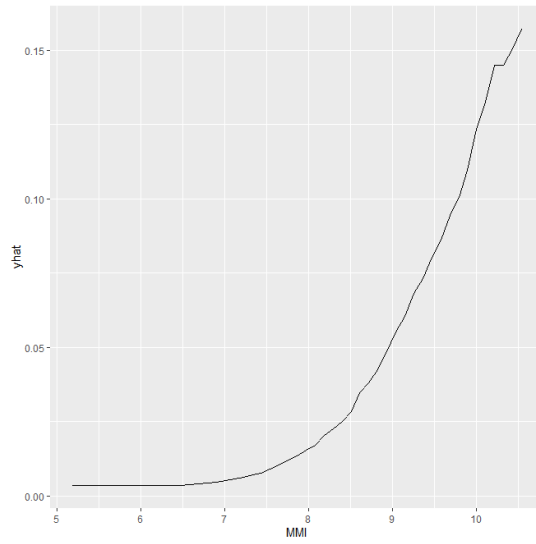


FIGURE 85 – Partial dependance plot de la variable **MMI**.

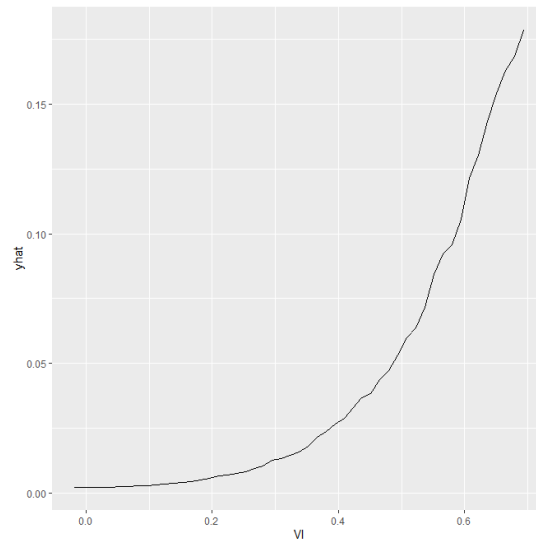


FIGURE 86 – Partial dependance plot de la variable **VI**.

Il est aussi possible de générer un PDP bivarié permettant d’observer l’éventuel effet d’interaction de deux variables sur la sortie du modèle :

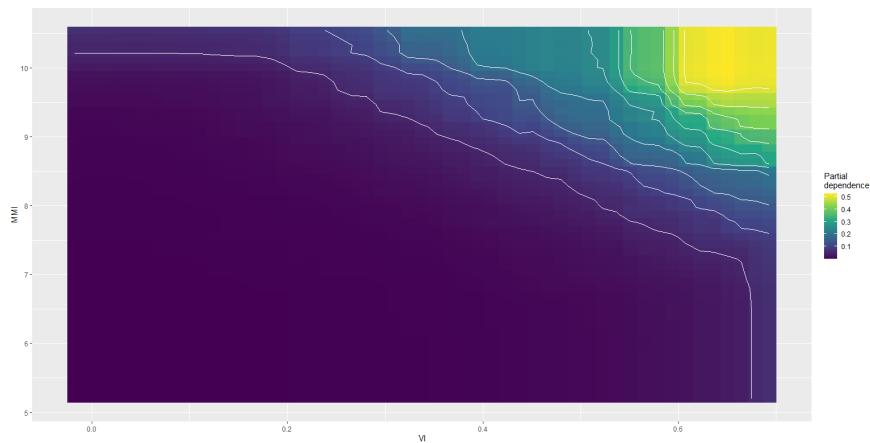


FIGURE 87 – Partial dependance plot des variables **VI** et **MMI**.

### *Accumulated Local Effects (ALE) plot*

Un ALE plot est un outil complémentaires aux PDP pour comprendre les relations entre les variables explicatives et la réponse dans un modèle *machine learning*. Il est une représentation de la fonction suivante :

$$\begin{aligned}\hat{f}_{S,ALE}(x_S) &= \int_{z_{0,S}}^{x_S} \mathbb{E}_{X_C|X_S=x_S} \left[ \hat{f}^S(X_s, X_c|X_S = z_S) \right] dz_S - constant \\ &= \int_{z_{0,S}}^{x_S} \left( \int_{x_C} \hat{f}^S(z_s, X_c) d\mathbb{P}(X_C|X_S = z_S) \right) dz_S - constant\end{aligned}\tag{83}$$

Dès lors, on remarque trois différences principales entre la fonction précédente et celle d'un PDP :

1. On considère ici la moyenne du changement des prédictions au lieu des prédictions elles-mêmes. Ce changement est traduit par la dérivée partielle  $\hat{f}^S(x_s, x_c) = \frac{\partial \hat{f}(x_S, x_C)}{\partial x_S}$  ;
2. La présence d'une seconde intégration sur  $z$  qui permet d'accumuler les dérivées partielles locale sur les éléments de  $S$  ;
3. L'ajout d'une constante permet de centrer l'effet en 0.

En pratique, comme un modèle de *machine learning* ne possède pas forcément de gradient, les dérivées partielles sont remplacées par des intervalles, ce qui permet d'aboutir à une approximation de la fonction ALE.

Dès lors, l'ALE peut s'interpréter comme l'effet principal d'une variable à une certaine valeur comparé à la prédiction moyenne de l'ensemble de données. Contrairement à un PDP, un ALE plot n'est pas biaisé et devrait être préféré en cas d'interactions ou de corrélations fortes.



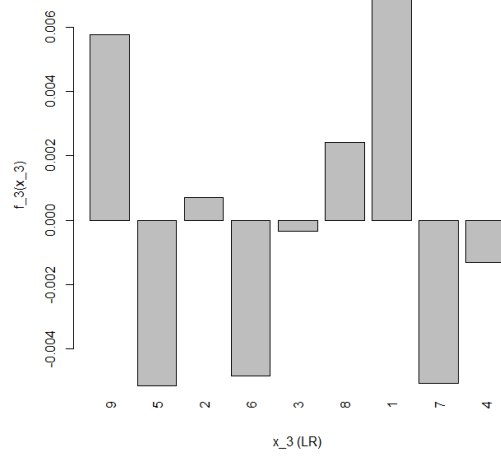


FIGURE 88 – ALE plot de la variable **LR**.

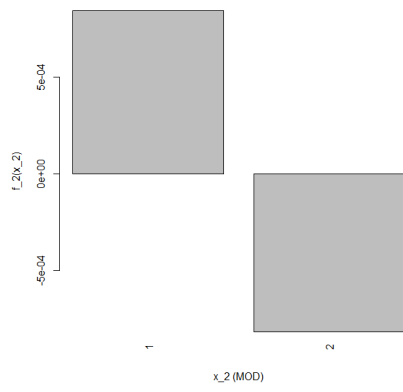


FIGURE 89 – ALE plot de la variable **MOD**.

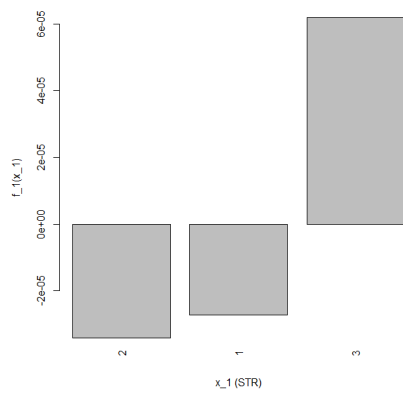


FIGURE 90 – ALE plot de la variable **STR**.

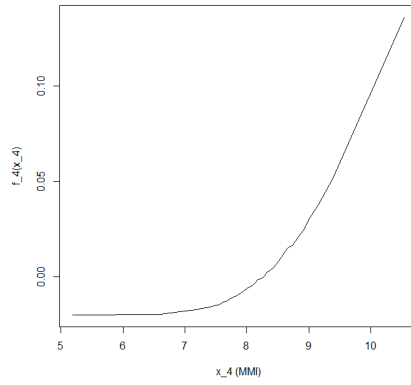


FIGURE 91 – ALE plot de la variable **MMI**.

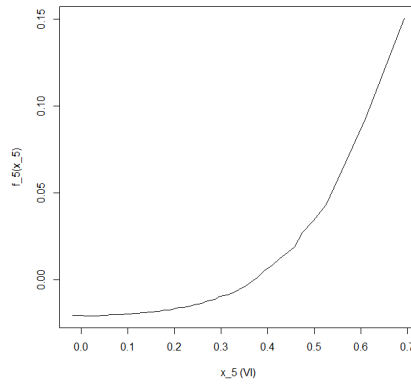


FIGURE 92 – ALE plot de la variable **VI**.

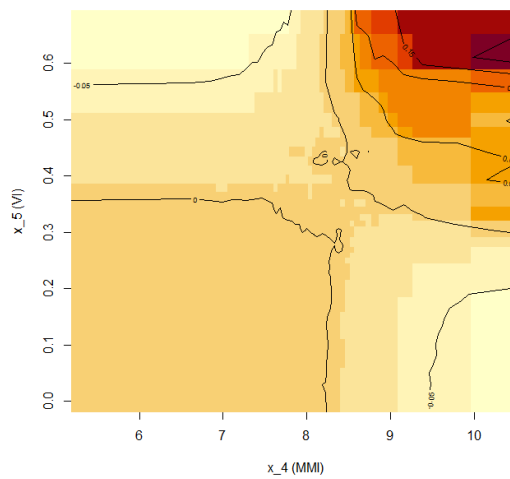


FIGURE 93 – ALE plot des variables **VI** et **MMI**.

## Interaction plot

Développée en 2008 par Friedman et Popescu, la H-statistique estime de combien la variation de la prédiction du modèle dépend de l'interaction entre les variables initiales. Sa théorie repose notamment sur les fonctions de dépendance partielle.

Considérant d'abord deux variables  $x_j$  et  $x_k$  sans interaction. En supposant que les fonctions de dépendance de ces deux variables soient centrées en 0, il est possible de décomposer la fonction de dépendance partielle bivariée comme suit :

$$PD_{jk}(x_j, x_k) = PD_j(x_j) + PD_k(x_k) \quad (84)$$

De même, si une variable donnée  $x_j$  n'interagit avec aucune des autres variables, la fonction de prédiction  $\hat{f}$  peut s'écrire comme :

$$\hat{f}(x) = PD_j(x_j) + PD_{-j}(x_{-j}) \quad (85)$$

Où  $PD_{-j}$  est la fonction de dépendance partielle qui dépend des de toutes les variables sauf  $x_j$ .

En particulier, la H-statistique vaut 0 s'il n'y a pas d'interaction et 1 si la variance des  $PD_{jk}$  ou  $\hat{f}$  est totalement expliquée par la somme des fonctions de dépendance partielle. Plus formellement, la H-statistique de l'interaction entre  $x_j$  et  $x_k$  est :

$$H_{jk}^2 = \frac{\sum_{i=1}^n \left[ PD_{jk}(x_j^{(i)}, x_k^{(i)}) - PD_j(x_j^{(i)}) - PD_k(x_k^{(i)}) \right]^2}{\sum_{i=1}^n PD_{jk}^2(x_j^{(i)}, x_k^{(i)})} \quad (86)$$

De la même façon, la H-statistique de l'interaction entre  $x_j$  avec les autres variables est :

$$H_j^2 = \frac{\sum_{i=1}^n \left[ \hat{f}(x^{(i)}) - PD_j(x_j^{(i)}) - PD_{-j}(x_{-j}^{(i)}) \right]^2}{\sum_{i=1}^n \hat{f}^2(x^{(i)})} \quad (87)$$

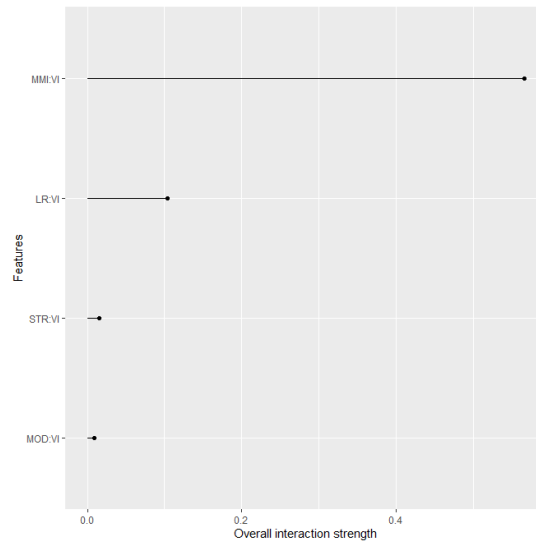


FIGURE 94 – Interaction plot d’ordre 2 des variables par rapport à **VI**.

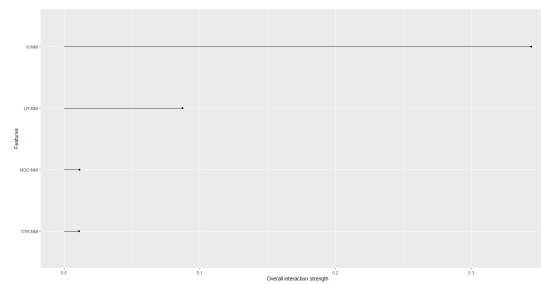


FIGURE 95 – Interaction plot d’ordre 2 des variables par rapport à **MMI**.

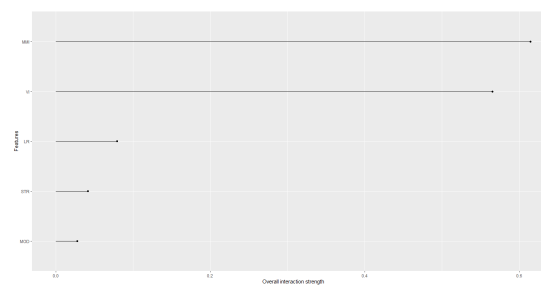


FIGURE 96 – Interaction plot d’ordre 1 des variables du modèle.

## Références

- [1] AAS, K., JULLUM, M. et LØLAND, A. (2021), “Explaining individual predictions when features are dependent: More accurate approximations to Shapley values”.
- [2] BOUSQUET, B. (2006), “Les séismes de l’Antiquité, entre nature et société”.
- [3] BRAGA, F., DOLCE, M. et LIBERATORE, D. (1982), “A statistical study on damaged buildings and an ensuing review of the MSK-76 scale”.
- [4] CASTRO, J., GÓMEZ, D. et TEJADA, J. (2009), “Polynomial calculation of the Shapley value based on sampling”.
- [5] DELACROIX, A. (2021), “Réassurance Non Vie”.
- [6] DOUGLAS, J. (2021), “Ground motion prediction equations 1964–2021”.
- [7] EFRON, B. et STEIN, C. (1981), “The jackknife estimate of variance”.
- [8] FERRARI, S. et CRIBARI-NETO, F. (2004), “Beta regression for modelling rates and proportions”.
- [9] GILQUIN, L. et al. (2019), “Making the best use of permutations to compute sensitivity indices with replicated orthogonal arrays”.
- [10] GIOVINAZZI, S. et LAGOMARSINO, S. (2004), “A macroseismic method for the vulnerability assessment of buildings”.
- [11] GRÜNTAL, G. (1998), “European macroseismic scale 1998 (EMS-98)”.
- [12] GRÜNTAL, G. et LEVRET, A. (2001), “L’echelle macrosismique européenne= European macroseismic scale 1998:(EMS-98)”.
- [13] INSTITUTE, Swiss Re (2019), “L’Aquila, 10 years on”.
- [14] IOOSS, B. et PRIEUR, C. (2019), “Shapley effects for sensitivity analysis with correlated inputs: comparisons with Sobol’indices, numerical estimation and applications”.
- [15] LUNDBERG, S. M. et LEE, S.-I. (2017), “A unified approach to interpreting model predictions”.
- [16] MCKAY, M. D., BECKMAN, R. J. et CONOVER, W. J. (1979), “A comparison of three methods for selecting values of input variables in the analysis of output from a computer code”.
- [17] MILUTINOVIC, Z. V. et TRENDAFILOSKI, G. S. (2003), “Risk-UE An advanced approach to earthquake risk scenarios with applications to different european towns”.
- [18] MOSCARITOLO, G. I. (2020), “Reconstruction as a Long-Term Process. Memory, Experiences and Cultural Heritage in the Irpinia Post-Earthquake (November 23, 1980)”.
- [19] OWEN, A. B. (2014), “Sobol’indices and Shapley value”.
- [20] POIRIER, J.-P. (2008), “Histoire de la sismologie”.
- [21] PORFIDO, S. et al. (2022), *40 years later: new perspectives on the 23 November 1980, Ms 6.9, Irpinia-Lucania earthquake.*

- [22] RADAIDEH, M. II et al. (2019), “Shapley effect application for variance-based sensitivity analysis of the few-group cross-sections”.
- [23] RIEDEL, I. et GUÉGUEN, P. (2018), “Modeling of damage-related earthquake losses in a moderate seismic-prone country and cost–benefit evaluation of retrofit investments: application to France”.
- [24] RIGBY, R. A. et STASINOPOULOS, D. M. (2005), “Generalized additive models for location, scale and shape”.
- [25] ROHMER, J. et al. (2014), “Weighing the importance of model uncertainty against parameter uncertainty in earthquake loss assessments”.
- [26] ROSS, T. J. (2010), “Fuzzy logic with engineering applications, Third Edition”.
- [27] SANDI, H. et FLORICEL, I. (1995), “Analysis of seismic risk affecting the existing building stock”.
- [28] SHAPLEY, L. S. (1953), “A value for n-person games”.
- [29] SMITHSON, M. et VERKUILEN, J. (2006), “A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables.”
- [30] SOBOL, I. M. (1993), “Sensitivity analysis for non-linear mathematical models”.
- [31] SONG, E., NELSON, B. L et STAUM, J. (2016), “Shapley effects for global sensitivity analysis: Theory and computation”.
- [32] STRASSER, F. O., ABRAHAMSON, N. A. et BOMMER, J. J. (2009), “Sigma: Issues, insights, and challenges”.
- [33] SUN, Y., APLEY, D. W. et STAUM, J. (2011), “Efficient nested simulation for estimating the variance of a conditional expectation”.
- [34] TISSOT, J.-Y. et PRIEUR, C. (2015), “A randomized orthogonal array-based procedure for the estimation of first-and second-order Sobol’indices”.
- [35] VRIEZE, S. I. (2012), “Model selection and psychological theory: a discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC).”
- [36] ZADEH, L. A. (1965), “Fuzzy Sets”.

## Table des figures

1	Source : Répartition des événements dommageables d'origine naturelle par gravité survenus dans le monde en 2021. Source : <i>catnat.net</i> . . . . .	6
2	Carte des événements et catastrophes naturels remarquables dans le monde en 2021. Source : Munich Re. . . . .	7
3	Les étapes du régime Cat-Nat. Source : CCR. . . . .	8
4	Les principaux éléments de la réforme du régime Cat-Nat. . . . .	10
5	Les grands piliers de Solvabilité II. . . . .	11
6	Les modules du SCR formule standard et le sous-module risque catastrophe. Source : ACPR. . . . .	12
7	Impact de la réassurance sur le résultat technique représenté en pourcentage des primes avant et après réassurance. Source : Fédération Française de l'Assurance (FFA). . . . .	13
8	Evolution du capital de la réassurance. Source : Aon. . . . .	15
9	Evolution du ratio combiné des réassureurs. Source : A. Delacroix. . . . .	16
10	Le cycle de la réassurance. . . . .	16
11	Les différentes catégories de réassurance. . . . .	17
12	Résultat d'une réassurance en quote-part sur un portefeuille de sinistres. <i>Lecture : dans cet exemple, le taux de cession est fixé à 30%</i> . . . . .	17
13	Résultat d'une réassurance en excédent de plein sur un portefeuille de sinistres. <i>Lecture : dans cet exemple, la cédante paye un montant maximal par sinistre de 500€ et obtient une capacité supplémentaire auprès d'un réassureur correspondant à 3 pleins, soit 1500€</i> . . . . .	18
14	Schéma d'un excédent de sinistre 500 000 € XS 2 500 000 € ( <i>excess of loss</i> ou XL). Source : Wikipédia. . . . .	19
15	Schéma d'un excédent de perte 20% SL 110% ( <i>stop-loss</i> ou SL). Source : Wikipédia. . . . .	20
16	Représentation simplifiée d'un séisme. Source : Wikipédia. . . . .	22
17	Les étapes d'une modélisation de catastrophe naturelle en assurance. . . . .	23
18	Échelle d'intensité macrosismique (EMS-98). . . . .	24

19	Échelle de dommage aux constructions (EMS-98). . . . .	24
20	Illustration de l'échelle des dommages EMS-98 sur des bâtiments en maçonnerie et en béton armé. Source : EMS-98 . . . . .	25
21	Différenciation des structures en classes de vulnérabilité. Source : RISK-UE. <i>Lecture : un bâtiment en massive stone appartient très probablement à la classe C, probablement à la classe B et exceptionnellement à la classe D.</i> . . . . .	25
22	Type de bâtiment selon EMS-98 et identification de leur comportement sismique par classes de vulnérabilité. Source : RISK-UE. <i>Lecture : En intensité VII de nombreux bâtiments de la classe de vulnérabilité A subissent des dégâts de degré 3, quelques uns de degré 4.</i> . . . . .	26
23	Représentation d'une fonction d'appartenance trapézoïdale. . . . .	28
24	Quantités EMS-98 « peu/few », « beaucoup/many » et « la plupart/most » exprimés en termes d'intervalles superposés et de fonctions d'appartenance. Source : RISK-UE. <i>Lecture : lorsque l'on observe peu de bâtiments endommagés, alors leur proportion est plausiblement entre 0% et 10%, et possiblement entre 10% et 20%. Cela se traduit par une fonction d'appartenance prenant constamment la valeur 1 entre 0 et 0,1 puis qui décroît linéairement jusqu'à 0 en 0,2.</i> . . . . .	29
25	Fonctions d'appartenance des indices de vulnérabilité par classe de vulnérabilité. Source : RISK-UE. . . . .	29
26	Fonctions d'appartenance et indices de vulnérabilité remarquables d'un bâtiment dont la structure est en bois. <i>Lecture : les courbes B, C, D et E représentent les fonctions d'appartenance de chaque classe de vulnérabilité.</i> . . . . .	30
27	Représentation de $\mu_D$ en fonction de l'intensité macrosismique (MMI) et de l'indice de vulnérabilité (VI). . . . .	32
28	Fonctions de vulnérabilité par structure. Source : RISK-UE. . . . .	32
29	Quelques formes de loi bêta à quatre paramètres. . . . .	34
30	Probabilités cumulées des niveaux de dommage dans le cas d'un bâtiment d'indice de vulnérabilité 0,7 soumis à une intensité macrosismique VII. La ligne rouge pointillée correspond au 90ème centile. . . . .	36



31	Comparaison des relations dommage-coût de la littérature par rapport à Milutinovic et Trendafiloski (2003). <i>Lecture : en damage state 1, la perte déduite de Meroni et al. (2017) est 2% plus élevée que celle de Milutinovic et Trendafiloski (2003), toutes choses égales par ailleurs. A l'inverse, en damage state 5, la perte déduite de Kappos et al. (2006) est 23% plus faible que celle de Milutinovic et Trendafiloski (2003), toutes choses égales par ailleurs.</i> . . . . .	39
32	Fonctions d'appartenance des indices de vulnérabilité des structures <i>RC1</i> , <i>RC2</i> et <i>RC3</i> . . . . .	40
33	Fonctions de vulnérabilité les plus probables des structures <i>RC1</i> , <i>RC2</i> et <i>RC3</i> . . . . .	41
34	Comparaison des lois binomiale et bêta. . . . .	43
35	Évolution temporelle des équations de prédiction du mouvement du sol (GMPE) en fonction : a) du nombre de coefficients dans l'équation ; du nombre de données par coefficients utilisées pour calibrer l'équation et c) de l'écart-type ( $\sigma_T [LN(Y)]$ ) de l'erreur de modélisation. Sources individuelles : figures a) et b) : Bommer et al. (2010) ; Figure c) : Strasser et al. (2009). Source générale : A. Pothon. . . . .	44
36	Carte de l'intensité macrosismique du séisme du 24 août 2016 en Italie. Source : USGS. . . . .	45
37	Carte de l'écart type du terme d'erreur associé à l'intensité macrosismique du séisme du 24 août 2016 en Italie. . . . .	46
38	Densité de la loi normale $\mathcal{N}(8, 0, 7^2)$ . . . . .	46
39	Ensemble des courbes de vulnérabilité possibles de la structure <i>RC2</i> . Les courbes supérieure et inférieure noires sont respectivement associées à des indices de vulnérabilité égaux à 0,85 et 0,15. . . . .	47
40	Fonctions de densité des variables aléatoires représentant les indices de vulnérabilité des structures <i>RC1</i> , <i>RC2</i> et <i>RC3</i> . . . . .	48
41	Comparaison d'un échantillonnage aléatoire avec un échantillonnage par hypercube latin en deux dimensions. Source : Preece et Milanović. . . . .	53
42	Représentation graphique de la fonction quantile d'une loi normale d'espérance 8 et d'écart type 0,7. . . . .	54
43	Représentation sur $[0, 1]$ de la fonction $x \mapsto [9x] + 1$ . . . . .	55
44	Les dix premières lignes de la base de données obtenue. . . . .	55

45	Histogramme et fonction de répartition empirique du vecteur de pertes simulées. . . . .	56
46	Graphiques de diagnostic du modèle linéaire multiple. <i>Lecture : le modèle ne suit pas les hypothèses du modèle linéaire. Même si l'on ne remarque pas de valeur aberrante particulière, le modèle ne respecte pas les hypothèses de normalité et d'homoscédasticité des résidus.</i> . . . . .	72
47	Graphe de Cullen-Frey de la distribution empirique de $\mathbf{Y}$ . . . . .	73
48	Illustration des différents configurations du coefficient d'asymétrie et de kurtosis. <i>Notons qu'un coefficient d'asymétrie nul n'indique pas nécessairement que la distribution est symétrique, mais une distribution symétrique a un coefficient nul.</i> . . . . .	74
49	Exemple de construction d'un modèle MARS étape par étape. . . . .	82
50	Exemple d'arbre déterminé par algorithme CART. . . . .	83
51	Principe général du <i>boosting</i> . Source : Wikipédia . . . . .	84
52	Principe général du <i>bagging</i> . Source : Wikipédia . . . . .	86
53	Quelques outils graphiques d'IA Explicable. . . . .	87
54	Principe de la méthode SHAP. Source : Lundberg et Lee. . . . .	90
55	SHAP <i>breakdown</i> de deux observations de la base de données. . . . .	91
56	Valeur de l'indicateur SHAP pour chacune des 8 variables les plus importantes. . . . .	91
57	Carte de l'intensité macrosismique du séisme du 23 novembre 1980 en Irpinia. Source : USGS. . . . .	94
58	Photographies prises après le séisme du 23 novembre 1980 en Irpinia. . . . .	95
59	Failles de surface visibles en 2004 résultant du séisme du 23 novembre 1980 en Irpinia. Source : S. Porfido . . . . .	95
60	Fonction de répartition de la valeur assurée des bâtiments en portefeuille. . . . .	96
61	Répartition de la structure des bâtiments assurés selon la valeur assurée. . . . .	97
62	Carte des sites en portefeuille. <i>Lecture : chaque point noir correspond à la localisation d'un site assuré. Le triangle rouge indique la ville de Naples.</i> . . . . .	97
63	Carte des sites en portefeuille associés à l'intensité macrosismique du séisme du 23 novembre 1980 en Irpinia. . . . .	98

64	Représentation pour chaque site en portefeuille de la valeur assurée et de l'intensité macrosismique subie. . . . .	99
65	Les cinq premières lignes de la matrice des variables d'entrée du modèle étendu. . . . .	100
66	Les cinq premières lignes de la matrice des <i>loss ratios</i> de chaque bâtiment. . . . .	100
67	Histogramme et fonction de répartition empirique du vecteur de pertes simulées. . . . .	101
68	Indices de Shapley des variables d'entrée du modèle étendu à différents niveaux d'intensité macrosismique. . . . .	103
69	<i>Summary</i> du modèle linéaire multiple. . . . .	107
70	<i>Summary</i> du modèle de régression bêta : Paramètre de moyenne. . . . .	108
71	<i>Summary</i> du modèle de régression bêta : Paramètre de précision. . . . .	108
72	Graphiques de diagnostic du modèle de régression bêta. <i>Lecture : contrairement aux diagnostic plots d'un GLM, ceux d'une régression bêta ne permettent pas de tirer de conclusions définitives sur l'ajustement du modèle. De façon générale, les résidus d'une régression bêta n'ont pas à être normalement distribués.</i> . . . . .	109
73	<i>Summary</i> du GAMLSS bêta inflaté en 0 et 1. . . . .	110
74	Graphiques de diagnostic du GAMLSS bêta inflaté en 0 et 1. . . . .	110
75	Exemple de construction d'un modèle MARS étape par étape. . . . .	111
76	Graphiques de diagnostic de modèle MARS. . . . .	112
77	Summary du modèle MARS. . . . .	112
78	Optimisation des hyperparamètres <i>max_depth</i> et <i>eta</i> du modèle XGBoost. . . . .	113
79	Optimisation de l'hyperparamètre <i>min_child_weight</i> du modèle XGBoost. . . . .	114
80	Optimisation des hyperparamètres <i>colsample_bytree</i> et <i>subsample</i> du modèle XGBoost. . . . .	114
81	Optimisation de l'hyperparamètre <i>gamma</i> du modèle XGBoost. . . . .	114
82	Partial dependance plot de la variable <b>LR</b> . . . . .	115
83	Partial dependance plot de la variable <b>MOD</b> . . . . .	116

84	Partial dependance plot de la variable <b>STR</b> . . . . .	116
85	Partial dependance plot de la variable <b>MMI</b> . . . . .	116
86	Partial dependance plot de la variable <b>VI</b> . . . . .	117
87	Partial dependance plot des variables <b>VI</b> et <b>MMI</b> . . . . .	117
88	ALE plot de la variable <b>LR</b> . . . . .	119
89	ALE plot de la variable <b>MOD</b> . . . . .	119
90	ALE plot de la variable <b>STR</b> . . . . .	119
91	ALE plot de la variable <b>MMI</b> . . . . .	120
92	ALE plot de la variable <b>VI</b> . . . . .	120
93	ALE plot des variables <b>VI</b> et <b>MMI</b> . . . . .	120
94	Interaction plot d'ordre 2 des variables par rapport à <b>VI</b> . . . . .	122
95	Interaction plot d'ordre 2 des variables par rapport à <b>MMI</b> . . . . .	122
96	Interaction plot d'ordre 1 des variables du modèle. . . . .	122

## Liste des tableaux

1	Franchises minimales légales. Source : CCR. . . . .	9
2	Les 10 premiers réassureurs mondiaux selon les primes émises brutes en 2020. Source : A.M. Best. . . . .	14
3	Degré d'appartenances de la structure <i>wood</i> aux classes de vulnérabilité B, C, D et E. . . . .	30
4	Matrices de typologie Risk-UE et indices de vulnérabilité remarquables. Source : RISK-UE. . . . .	31
5	Distribution de probabilités de dommage dans le cas d'un bâtiment d'indice de vulnérabilité 0,7 soumis à une intensité macrosismique VII, obtenue par loi bêta à quatre paramètres. . . . .	35
6	Distribution de probabilités de dommage cumulées dans le cas d'un bâtiment d'indice de vulnérabilité 0,7 soumis à une intensité macrosismique VII, obtenue par loi bêta à quatre paramètres. . . . .	36
7	Revue de la littérature des différentes relations dommage-coût calibrées dans l'espace européen. Les auteurs et la source font respectivement référence aux scientifiques qui ont développé la relation et à ceux qui ont écrit l'article dont les valeurs sont extraites. Les <i>loss ratios</i> indiqués correspondent aux valeurs centrales. Source : A. Pothon. . . . .	38
8	Indice de vulnérabilité le plus probable de chaque structure. . . . .	40
9	Niveaux de dommage moyen $\mu_D$ des structures <i>RC1</i> , <i>RC2</i> et <i>RC3</i> estimés en considérant une intensité macrosismique X et l'indice de vulnérabilité le plus probable $V_I^*$ de chaque structure. . . . .	41
10	Distribution de probabilités associée à un niveau 2 de dommage moyen. La ligne « Delta » représente la différence entre le modèle binomial et la modèle bêta. . . . .	42
11	Aire sous la courbe $\mathcal{A}$ des fonctions d'appartenances associées à la vulnérabilité des structures <i>RC1</i> , <i>RC2</i> et <i>RC3</i> . . . . .	48
12	Description des sources d'incertitude considérées et hypothèses de représentation dans le cadre de l'analyse de sensibilité. . . . .	51
13	Modalités de la variable catégorielles <b>MOD</b> qui représente le choix du modèle permettant d'obtenir la distribution de probabilités de dommage. . . . .	51

14	Modalités de la variable catégorielles <b>STR</b> qui représente le type de structure du bâtiment considéré. . . . .	52
15	Modalités de la variable catégorielles <b>LR</b> qui représente le choix de l'ensemble de relations dommage-coût. . . . .	52
16	Statistiques descriptives de <b>Y</b> . . . . .	56
17	Indices de Shapley des variables d'entrée du modèle. <i>Lecture : les intervalles de confiance ont été déterminés par bootstrap non-paramétrique à partir de 1000 simulations.</i> . . . . .	64
18	RMSE du modèle linéaire multiple sur 10-fold et sur l'échantillon test. .	71
19	Résultats des divers tests permettant de vérifier les hypothèses du modèle linéaire. . . . .	73
20	Exemples de lois de la famille exponentielle. . . . .	75
21	Exemples de fonctions de lien. . . . .	75
22	RMSE du modèle de régression bêta sur 10-fold et sur l'échantillon test transformé. . . . .	77
23	Fonctions de lien des paramètres de quelques familles de distributions GAMLSS. Source : Rigby et Stasinopoulos. <i>Lecture : la fonction logit est telle que, pour <math>x \in ]0, 1[</math>, <math>\text{logit}(x) = \log\left(\frac{x}{1-x}\right)</math></i> . . . . .	79
24	RMSE du GAMLSS bêta inflaté en 0 et 1 sur 10-fold et sur l'échantillon test. . . . .	80
25	RMSE du modèle MARS sur 10-fold et sur l'échantillon test. . . . .	82
26	RMSE du modèle XGBoost sur 10-fold et sur l'échantillon test. . . . .	85
27	RMSE du modèle Random Forest sur 10-fold et sur l'échantillon test. . .	86
28	Les 5 premières lignes du portefeuille d'assurances. . . . .	96
29	Statistiques descriptives de l'intensité macrosismique. . . . .	99
30	Statistiques descriptives de <b>Y</b> . . . . .	101
31	Indices de Shapley des variables d'entrée du modèle étendu. <i>Lecture : les intervalles de confiance ont été déterminés par bootstrap non-paramétrique à partir de 1000 simulations.</i> . . . . .	102
32	Comparaison des rangs de l'indicateur d'importance de contribution et de la valeur assurée du bâtiment. . . . .	104