Application de l'apprentissage automatique pour la construction des courbes de fragilité sismique

Anh-Dung Tran

université PARIS-SACLAY

Application de l'apprentissage automatique pour la construction des courbes de fragilité sismique

Use of machine learning for the construction of seismic fragility curves

École doctorale n° n°579 : sciences mécaniques et énergétiques, matériaux et géosciences (SMEMaG)
Spécialité de doctorat : Génie Civil
Graduate School : Sciences de l'ingénierie et des systèmes. Référent : Université d'Évry Val-d'Essone

Thèse préparée dans l'unité de recherche **LMEE** (Université Paris-Saclay, Univ Evry), sous la direction de **Thien-Phu LE**, Maître de conférences-HDR, le co-encadrement de **Michael BURMAN**, Maître de conférences et le co-encadrement de **Gérard PORCHER**, Maître de conférences

Thèse soutenue à Paris-Saclay, le 22 mars 2024, par

Anh-Dung TRAN

Composition du jury

Membres du jury avec voix délibérative

Patrick PAULTRE Professeur, Université de Sherbrooke	Président du jury
Luigi GARIBALDI Professeur, Politecnico di Torino	Rapporteur et Examinateur
Roger SERRA Professeur, INSA Centre Val de Loire	Rapporteur et Examinateur
Amer CHPOUN Professeur, Université d'Evry - Paris-Saclay	Examinateur

THESE DE DOCTORAT

NNT : 2024UPAST017

université PARIS-SACLAY

ÉCOLE DOCTORALE
Sciences mécaniques et énergétiques, matériaux et géosciences (SMEMAG)

Titre : Application de l'apprentissage automatique pour la construction des courbes de fragilité sismique

Mots clés : Apprentissage automatique, Simulation numérique, Courbe de fragilité sismique, Spectre de réponse, Système linéaire, Système non-linéaire.

Résumé : L'évaluation des risques sismiques pour les structures est nécessaire pour prévenir les pertes humaines et matérielles en cas de catastrophes naturelles. La construction des courbes de fragilité sismique, l'outil qui définit la probabilité de défaillance d'une structure en fonction de l'intensité des séismes, joue un rôle important dans cette évaluation. En général, la méthode de construction de cette courbe demande des procédures coûteuses, très souvent par appel des analyses par éléments finis, nécessitant des ressources considérables en termes de temps et d'informatique. Cette difficulté empêche l'utilisation de cet outil en temps réel ou l'application de cet outil pour les structures plus populaires. L'apprentissage automatique a connu un développement spectaculaire avec son l'application dans plusieurs domaines. Il est reconnu comme un outil puissant pour modéliser une relation complexe entre les entrées et les sorties à partir des données. Une approche novatrice émerge : l'utilisation des modèles d'apprentissage automatique pour prédire la réponse des structures sous séismes. Par conséquent, la motivation principale de la thèse est d'étudier l'application de l'apprentissage automatique à la génération de réponses sismiques pour l'évaluation des risques sismiques, plus spécifiquement pour construction de la courbe de fragilité. L'étude débute par une revue de la littérature, qui présente les courbes de fragilité sismique et le défi lié à la charge de calculs pour leur construction. Après une brève présentation de l'apprentissage automatique, le premier chapitre se concentre sur son application pour modéliser les réponses sismiques des structures. Devant un très large choix parfois contradictoire des caractéristiques, il est nécessaire de proposer des procédures pratiques avec les caractéristiques les plus pertinentes et faciles à la mise en œuvre. La thèse examine ce problème et donne une proposition innovante consiste à utiliser et sélectionner efficacement le spectre de réponse en accélération échantillonnées aux différentes périodes pour construire les modèles d'apprentissage automatique. Deux procédures, nommées PRO-LIN et PRO-NONLIN, sont proposées pour des structures linéaires et non-linéaires respectivement. Pour valider ces propositions, des tests sont effectués sur des structures linéaires et non-linéaires en combinant avec des enregistrements synthétiques, démontrant une réduction significative du temps de simulation tout en maintenant la précision dans la construction des courbes de fragilité. Bien que les premières validations des procédures soient réalisées avec des enregistrements synthétiques, une validation avec des enregistrements réels est indispensable pour valider ces propositions. Les enregistrements sont sélectionnés suivant le spectre conditionnel depuis des bases de données. Finalement, une autre validation vise à tester les procédures proposées avec des bases de données existantes dans en littérature. La première base de données concerne une structure linéaire en béton armé. La deuxième base de données concerne des portiques non-linéaires résistantes aux moments en acier. En conclusion, basant sur les résultats obtenus dans les études, ce travail met en lumière l'efficacité des procédures proposées de la thèse. Ces procédures améliorent de façon efficace la construction des courbes de fragilité sismique et l'évaluation des risques sismiques.

Title : Use of machine learning for the construction of seismic fragility curves
Keywords : Machine learning, Numerical simulation, Seismic fragility curve, Response spectrum, Linear system, Nonlinear system.

Abstract : The evaluation of seismic risks for structures is necessary to prevent human and material losses in the event of natural disasters. The construction of seismic fragility curves, the tool that defines the probability of failure of a structure based on earthquake intensity, plays a significant role in this assessment. Generally, the method of constructing these curves requires costly procedures, often involving finite element analyses, which require considerable time and computing resources. This difficulty hinders the real-time use of this tool or its application for more common structures. Machine learning has experienced a remarkable development with its application in various fields. It is recognized as a powerful tool for modeling complex relationships between inputs and outputs from data. An innovative approach is emerging : the use of machine learning models to predict structural responses to earthquakes. Therefore, the main motivation of the thesis is to study the application of machine learning to generate seismic responses for seismic risk assessment, specifically for constructing fragility curves. The study begins with a literature review, which presents seismic fragility curves and the challenge related to the computational burden for their construction. After a brief introduction to machine learning, the first chapter focuses on its application to model seismic responses of structures. Given the sometimes contradictory wide range of characteristics, it is necessary to propose practical procedures with the most relevant and easy-to-implement features. The thesis addresses this issue and proposes an innovative approach to efficiently use and select the acceleration response spectrum sampled at different periods to build machine learning models. Two procedures, named PRO-LIN and PRO-NONLIN, are proposed for linear and nonlinear structures respectively. To validate these proposals, tests are performed on linear and nonlinear structures combined with synthetic records, demonstrating a significant reduction in simulation time while maintaining accuracy in constructing fragility curves. Although the initial validations of the procedures are conducted with synthetic records, validation with real records is essential to validate these proposals. The records are selected according to the conditional spectrum from databases. Finally, another validation aims to test the proposed procedures with existing databases in the literature. The first database concerns a linear reinforced concrete structure. The second database concerns nonlinear moment-resistant steel frames. In conclusion, based on the results obtained in the studies, this work highlights the effectiveness of the proposed procedures in the thesis. These procedures effectively improve the construction of seismic fragility curves and seismic risk assessment.

Remerciements

Les travaux de thèse ont été réalisés au laboratoire LMEE (Laboratoire de Mécanique et d'Énergétique d'Evry) de l'Université d'Evry - Paris-Saclay. Je remercie tous les membres du laboratoire pour leur accueil chaleureux.

Je suis très reconnaissant envers M. Luigi GARIBALDI, Professeur à Politecnico di Torino, et M. Roger SERRA, Professeur à l'INSA Centre Val de Loire, qui ont accepté d'être rapporteurs de cette thèse. Leurs commentaires et appréciations m'ont permis d'apporter d'importantes améliorations à mon manuscrit. J'en suis honoré et je les remercie sincèrement.

Je tiens à remercier M. Patrick PAULTRE, Professeur à l'Université de Sherbrooke au Canada, et M. Amer CHPOUN, Professeur à l'Université d'Evry, pour leur participation à l'évaluation de cette thèse, en qualité de membre du Jury.

Je tiens à remercier Monsieur Pierre ARGOUL, Professeur à l'École des Ponts ParisTech, Monsieur Jean-Michel CROS, Professeur à l'Université d'Evry et directeur de Laboratoire LMEE, Monsieur Olivier QUÉMÉNER, Professeur à l'Université d'Evry et Monsieur Jean-Michel GÉNEVAUX, Professeur à l'Université du Maine, qui m'ont fait le grand plaisir de participer dans mon comité de suivi de thèse. Leurs remarques et leurs critiques m'ont donné des bonnes réflexions pour cette thèse.

Ma profonde gratitude s'adresse particulièrement à Monsieur Thien-Phu LE, Maître de conférences-HDR à l'Université d'Evry - Paris-Saclay, pour m'avoir accepté de travailler sous sa direction. Je lui suis très reconnaissant de m'avoir transmis ses précieux conseils avec beaucoup de patiences, de soutiens et d'encouragements. Il n'a jamais hésité de prendre le temps de m'expliquer, de discuter, d'échanger et de partager ses grandes connaissances.

Je n'oublie jamais de remercier à mes co-encadrant de thèse, Monsieur Gérard PORCHER et Monsieur Michael BURMAN, Maître de conférences à l'Université d'Evry - Paris-Saclay, pour leur aide et leur accompagnent. Ils était toujours disponibles pour m'accompagner durant la thèse. Ils sont aussi passé bon nombre de soirées à corriger les innombrables fautes grammaires présentées dans ce manuscrit.

J'ai eu la chance de pouvoir effectuer mon travail sous leur direction pendant trois ans. Je les ai appelés et je les appellerai "Professeur" avec tout mon respect jusqu'à la fin de ma vie.

Je voudrais remercier aussi mes amis au Vietnam et en France. Que Mme Hai-Que PHAN, Minh-Tu, Thuy-Linh, Viet-Quang, Thanh-Trung, Minh-Hang, Thao-Phuong, Tuan-Thanh, Mai-Anh, Viet-Anh, et spécialement Van-Anh soient remerciés. Les uns et les autres, chacun de sa manière, merci pour leur patience et leur encouragement tout au long de ma vie.

Enfin, je voudrais profiter de cette occasion pour remercier mes parents pour leur soutien. Les mots ne suffisent pas à exprimer ma gratitude pour leur soutien et leur abnégation. Je remercie aussi mon frère Hoai-Anh. Il m'a apporté un soutien émotionnel et psychologique tout au long de ces années. Je tiens également à le féliciter pour son mariage avec Thu-Thao.

Table des matières

Abréviations

PRO-LIN	Procédure à base de l'apprentissage automatique pour structures linéaires
PRO-NONLIN	Procédure à base de l'apprentissage automatique pour structures non-linéaires
FEM	Finite element models
ML	Machine learning
LS	Limit state
MCS	Monte Carlo simulation method
ML-MCS	Machine Learning based Monte Carlo simulation method
SSI	Scaled seismic intensity method
MLE	Maximum likelihood estimation method
ML-MLE	Machine Learning based Maximum likelihood estimation method
PSDM/PSCM	Probabilistic seismic demand model/ Probabilistic seismic capacity model method
IM	Intensity measure
PGA	Peak ground acceleration
PGV	Peak ground velocity
PGD	Peak ground displacement
S_a	Spectre de réponse en accélération
S_v	Velocity response spectrum
S_d	Displacement response spectrum
ASI	Intensité du spectre en accélération
DSI	Intensité du spectre en déplacement
D_{5-75}	Durée significative de 5 à 75 %
D_{5-95}	Durée significative de 5 à 95 %
CAV	Énergie absolue cumulée
I_A	Intensité d'Arias
ω_{eqk}	Fréquence angulaire du séisme
V/A	Rapport entre la vitesse du sol et l'accélération
a_{RMS}	Accélération quadratique moyenne
T_D	Durée totale d'enregistrement sismique

ChI	Intensité caractéristique
ASA	Accélération spectrale moyenne
T_P	Période prédominante
LR	Regression algorithm
KNN	K-Nearest Neighbor
DT	Decision tree
RF	Random forest
ANN	Artificial Neural Network
SVMR	Support vector machine for regression
AdaBoost	Adaptive boosting
LightGBM	Light Gradient Boosting Machine
XGBoost	Extreme Gradient Boosting Machine
SHAP	Shapley Additive explanations
ESM	Engineering Strong-Motion Database
NGA-West2	Next Generation Attenuation Relationships for Western US Database
RFE	Élimination Récursive des Caractéristiques (Recursive Feature Elimination)
1ddl	Un degré de liberté
2ddl	Deux degrés de liberté
NTHA	Analyse temporelle non-linéaire (Nonlinear time history analysis)
UHS	Uniform Hazard Spectrum
CS	Conditional Response Spectrum
GMM	Ground Motion Model
PSHA	Probabilistic Seismic Hazard Assessment
GMPE	Ground Motion Prediction Equation
SMRF	Steel moment-resisting frame

Symbols

$Fr(a)$	Courbe de fragilité
$\mathbf{Y}$	Réponse du système
y_0	Limite critique de la réponse du système
A	Intensité caractéristique du mouvement du sol
$\mathbf{x}$	Vecteur des caractéristiques d'entrée
ϵ	Erreur en raison du manque d'ajustement du ML modèle
a	Valeur de l'intensité caractéristique du mouvement du sol
D	Demande sismique imposée à la structure
C	Capacité sismique de la structure
A_m	Médiane
β	Ecart-type
$\Phi(\cdot)$	Fonction de répartition de probabilité de la loi normale centrée réduite
$\mathbf{M}$	Matrice de masse
$\mathbf{C}$	Matrice d'amortissement
$\mathbf{K}$	Matrice de rigidité
$\mathbf{r}$	Vecteur d'influence
$y(t)$	Déplacement de la structure en temps
$\dot{y}(t)$	Vitesse de la structure en temps
$\ddot{y}(t)$	Accélération de la structure en temps
$a_g(t)$	Accélération du sol en temps

Introduction

La sûreté sismique des structures est primordiale pour éviter des pertes humaines et limiter des dégâts matériels en cas de désastres. La conception parasismique cherche à proposer des dimensionnements permettant aux structures de résister aux efforts générés par les mouvements du sol. En général, plus une structure est solide, plus elle résiste à un séisme. Toutefois, comme les tremblements de terre sont de nature aléatoire, une sureté totale de 100 % n'existe pas. En plus, une probabilité de défaillance très faible, correspondant souvent à des événements rares, peut conduire à une conception économiquement non faisable. Il est ainsi nécessaire de considérer le risque acceptable relative par rapport aux structures.

Un outil d'aide à la décision peut être défini à l'aide des courbes de fragilité sismique qui présentent la probabilité de défaillance d'une composante, d'un système ou d'une structure, en fonction d'une intensité sismique du sol comme l'accélération maximale du sol par exemple. Initialement proposées dans le domaine nucléaire, les courbes de fragilité sismique ont trouvé dans les dernières années une application grandissante dans le domaine de génie civil avec des ouvrages de grande dimension tels que bâtiments, ponts, barrages, ...

Les courbes de fragilité sismique, utiles pour la validation des conceptions parasismiques, pour la planification des aides pendant les séismes et pour la réhabilitation après les séismes, peuvent être obtenues par des simulations numériques par la méthode des éléments finis. La difficulté majeure dans cette démarche est liée au coût important nécessaire des réponses sismiques.

L'apprentissage automatique, ou machine learning en anglais (ML), a connu un développement spectaculaire ces dernières années. Son application dans plusieurs domaines tels que la reconnaissance vocale, la reconnaissance faciale, le contrôle automatique, ..., présente des résultats impressionnants. L'apprentissage automatique est reconnu comme un outil puissant pour modéliser une relation complexe, non explicite entre les entrées et les sorties à partir des données. La question immédiate est donc : peut-on appliquer l'apprentissage automatique pour traduire la relation entre les séismes (entrées) et les réponses sismiques des structures (sorties) ?

La motivation principale de la thèse est donc d'étudier l'application de l'apprentissage automatique à la génération d'une base de données de qualité de réponses sismiques qui sera ensuite utilisée pour l'évaluation de la sûreté sismique sans faire appel de façon excessive aux calculs temporels souvent non-linéaires par la méthode des éléments finis. La réussite de cette proposition nécessite de lever deux verrous scientifiques liés respectivement (i) au choix des intensités sismiques pertinentes pour les réponses structurales et (ii) à une stratégie de sélection des caractéristiques afin de rendre les modèles d'apprentissage automatique plus performants. L'aboutissement du travail de thèse allégera le coût des calculs sismiques des structures et il rendra l'évaluation de la sûreté sismique plus robuste grâce aux modèles de l'apprentissage automatique.

Le travail de la thèse est organisé en cinq chapitres.

- **Chapitre 1** : *Études biographiques et problématique générale*, présente la définition d'une courbe de fragilité sismique, son utilité lors de la phase de conception, d'exploitation et de réhabilitation des ouvrages. Parmi les trois méthodes de construction existantes basées respectivement sur les données empiriques, sur les avis d'expertise ou sur les simulations numériques, la dernière est la plus précise, la plus générique et son domaine d'application est le plus large. Cette méthode rencontre toutefois un challenge lié à la charge des calculs dynamiques non-linéaires par des simulations numériques. Deux voies de recherche sont proposées pour le pallier : (i) le développement de méthodes purement statistiques avec un nombre de simulations mécaniques limité et (ii) le remplacement des calculs mécaniques par des méta-modèles. Le travail de la thèse se trouve dans la deuxième voie de recherche par l'apprentissage automatique. L'apprentissage automatique a été largement développé et appliqué les derniers temps, accompagné par le développement rapide des moyens informatiques. Après une présentation rapide des méthodes d'apprentissage automatique les plus populaires disponibles, le chapitre 1 se concentre sur leur application pour modéliser les réponses sismiques des structures. Devant un très large choix, parfois contradictoire, des caractéristiques des études existantes pour les séismes, il est donc nécessaire de proposer des procédures pratiques avec les caractéristiques les plus pertinentes, faciles à metttre en œuvre. Les procédures proposées doivent être en plus, applicables pour les données de natures variées telles que : séismes synthétiques/réels ; structures linéaires/non-linéaires ;...Ces points correspondent aux objectifs des chapitres suivants de la thèse.
- **Chapitre 2** : *Apprentissage automatique pour les structures linéaires*, propose d'utiliser les spectres de réponse en accélération échantillonnés en différentes périodes comme les caractéristiques des modèles de l'apprentissage automatique. Les réponses dynamiques des systèmes linéaires sont la superposition des réponses modales et cette propriété suggère des zones prioritaires de périodes à prendre en compte dans les modèles. Il est également possible de réduire le nombre de spectres en utilisant les scores obtenus par l'étude SHAP des modèles. Une procédure pratique complète pas-à-pas PRO-LIN a été proposée. Elle est validée pour un système à un degré de liberté et un système à deux degrés de liberté.
- **Chapitre 3** : *Apprentissage automatique pour les structures non-linéaires*, cherche à étendre l'utilisation des spectres de réponse en accélération comme caractéristiques des modèles de l'apprentissage automatique aux structures non-linéaires. L'analyse préliminaire montre qu'une zone plus large et un échantillonnage plus fin des spectres par rapport à ceux définis pour des structures linéaires, sont nécessaires. Afin de rendre les modèles plus performants, seules les caractéristiques les plus pertinentes sont retenues. Trois méthodes disponibles en littératures sont étudiées : (i) méthode de filtrage, rapide mais moins précis due à la non prise en compte de l'interaction entres les caractéristiques, (ii) méthode d'enveloppe, précis mais couteux en temps de calculs et (iii) méthode intégrée disponible seulement dans les modèles de forêt aléatoire (RF). Cette analyse permet de proposer une nouvelle procédure hybride PRO-NONLIN combinant les avantages des méthodes de filtrage et d'enveloppe. La procédure pratique complète est détaillée pas-à-pas. Elle est validée avec les oscillateurs non-linéaires d'amortissement de Coulomb et de rigidité de Bouc-Wen et avec un bâtiment non-linéaire à 8 étages. A partir des

modèles d'apprentissage automatique, il est alors facile de générer des simulations permettant de construire des courbes de fragilité sismique dans le contexte d'évaluation des risques sismiques.

- **Chapitre 4** : *Validation avec les enregistrements sismiques réels*, a pour objectif de tester les procédures proposées dans les chapitres précédents aux séismes réels. En prenant une pratique courante des études de l'évaluation des risques sismiques, les séismes réels sont sélectionnés suivant le spectre conditionnel à partir de la base de données des enregistrements réels comme NGA West (Next Generation Attenuation for the Western United States database). La base de la méthode et la procédure pratique de la sélection des séismes réels sont d'abord présentées. Ces séismes sont ensuite utilisés dans les procédures proposées respectivement aux structures linéaires et non-linéaires. Leur validation et leur limite sont enfin relevées.
- **Chapitre 5** : *Validation avec les données disponibles en littérature*, vise à tester les procédures proposées avec des bases de données provenant des études déjà publiées aux communautés scientifiques. La première base de données concerne une structure linéaire en béton armé. Elle est mesurée expérimentalement et recalée sur une maquette réduite au laboratoire. Le recalage du modèle permet d'obtenir des matrices de masse, de rigidité et d'amortissement. Les réponses sismiques simulées de ce système sont utilisées pour tester la procédure proposée pour les structures linéaires. La deuxième base de données concerne des portiques non-linéaires résistants aux moments en acier dont les réponses temporelles sont simulées par la méthode des éléments finis. Les réponses et les enregistrements sismiques réels correspondants, mis à disposition par les auteurs aux communautés, sont extraits pour tester la procédure hybride proposée pour les structures non-linéaires. La validité des procédures est vérifiée et évaluée en utilisant les métriques et les validations croisées de l'apprentissage automatique

Enfin, les conclusions et perspectives synthétisent les contributions principales de la thèse et elles proposent des idées de poursuite de ce travail.

1 Synthèse bibliographique et problématique générale

Sommaire

1.1. Introduction

L'évaluation des risques sismiques des structures est cruciale pour éviter des pertes humaines et matérielles en cas de désastres. La conception parasismique vise à proposer des dimensionnements permettant aux structures de résister aux efforts générés par les mouvements du sol. En général, plus une structure est solide, plus elle a de chances de survivre lors des séismes. Toutefois, comme les tremblements de terre sont de nature aléatoire, une sureté à 100 % n'existe pas. De plus, une probabilité de défaillance très faible, correspondant souvent à des événements rares de tremblements de terre, peut rendre une conception économiquement non faisable et non rentable. Il est ainsi nécessaire de considérer le risque acceptable relativement à l'importance des structures. Un outil d'aide à la décision incontournable dans cette évaluation repose sur les courbes de fragilité sismique, qui représentent la probabilité de défaillance d'une composante, d'un système ou d'une structure, en fonction d'une intensité sismique du sol, comme par exemple l'accélération maximale du sol.

La section 1.2 présente la définition d'une courbe de fragilité sismique, son utilité et les différentes approches de construction. Considérée comme l'approche la plus efficace, celle basée sur les simulations numériques est présentée avec deux méthodes souvent utilisées pour établir des courbes de fragilité sismique : la méthode de Monte-Carlo et la méthode du maximum de vraisemblance. Les défis des méthodes basées sur les simulations numériques sont identifiés : la charge de calcul très importante associée à la détermination des réponses sismiques par la méthode des éléments finis. Une des pistes pour dépasser cet obstacle est d'utiliser des méta-modèles dont l'apprentissage automatique.

La section 1.3 présente une analyse rapide de l'application des méthodes d'apprentissage automatique dans le domaine du génie civil. Son application spécifique dans l'évaluation des risques sismiques des structures est examinée dans la section 1.4. L'analyse bibliographique montre que le succès de l'apprentissage automatique nécessite un bon choix des caractéristiques des excitations sismiques. C'est le sujet abordé dans la section 1.5, où plusieurs modèles d'apprentissage automatique ont été recensés avec différentes caractéristiques sismiques sélectionnées. Face à ce constat, la potentialité de l'utilisation des spectres de réponse comme caractéristiques est soulignée.

Les mouvements du sol sont des données essentielles dans les études sismiques des structures. La section 1.6 est consacrée à une description rapide de deux sources de signaux sismiques : les enregistrements réels et les enregistrements synthétiques. Ces signaux sont utilisés pour valider les propositions de la thèse, dont les objectifs sont résumés dans la section 1.7, avec le contenu principal des chapitres suivants.

En conclusion, la synthèse des points essentiels du chapitre est présentée.

1.2. Courbe de fragilité sismique

1.2.1. Définition

La courbe de fragilité sismique revêt une importance majeure dans les études probabilistes de la sécurité sismique, car elle évalue la probabilité de défaillance ou d'endommagement d'une structure

ou d'un de ses composants en fonction d'une intensité sismique du sol comme l'accélération maximale du sol (peak ground acceleration, PGA) par exemple. Initialement utilisée dans les études de sûreté des centrales nucléaires pour prévenir les fuites radioactives, cette courbe a été introduite pour la première fois par Kennedy *et al.* [1]. Son application est par la suite étendue à divers domaines et types de structures, englobant notamment l'équipement des centrales nucléaires et les systèmes de refroidissement en génie nucléaire [2], les bâtiments [3, 4], ainsi que les ponts [5, 6] dans le domaine du génie civil.

Cette courbe est devenue populaire dans le domaine du génie parasismique, parce qu'elle est également très utile pour la validation des conceptions parasismiques, pour la planification des aides pendant les séismes et pour la réhabilitation après les séismes. Cette utilisation est possible par l'avancée rapide des performances informatiques au cours des dernières années.

Selon Kafali [7], une courbe de fragilité sismique d'une structure est définie comme la probabilité conditionnelle de défaillance pour une intensité d'excitation du sol A donnée :

$$Fr(a) = P\left[\mathbf{Y} \geq y_0 | A = a\right] \tag{1.1}$$

où A est la mesure d'intensité sismique du mouvement du sol et a est sa valeur. La défaillance est considérée comme se produisant si la réponse $\mathbf{Y}$ dépasse la limite critique y_0 de la structure. Les courbes de fragilité sont construites en supposant que leur forme suit la fonction de probabilité cumulative de la loi log-normale. Les courbes de fragilité relatives aux différents niveaux d'endommagement peuvent être aussi déterminées en utilisant la notion d'endommagement plutôt que de défaillance. Par exemple, la Figure 1.1 présente un exemple de trois courbes de fragilité obtenues par l'hypothèse de la loi log-normale pour trois niveaux d'endommagement. Le pic d'accélération du sol a été choisi comme l'intensité caractéristique A du mouvement du sol.

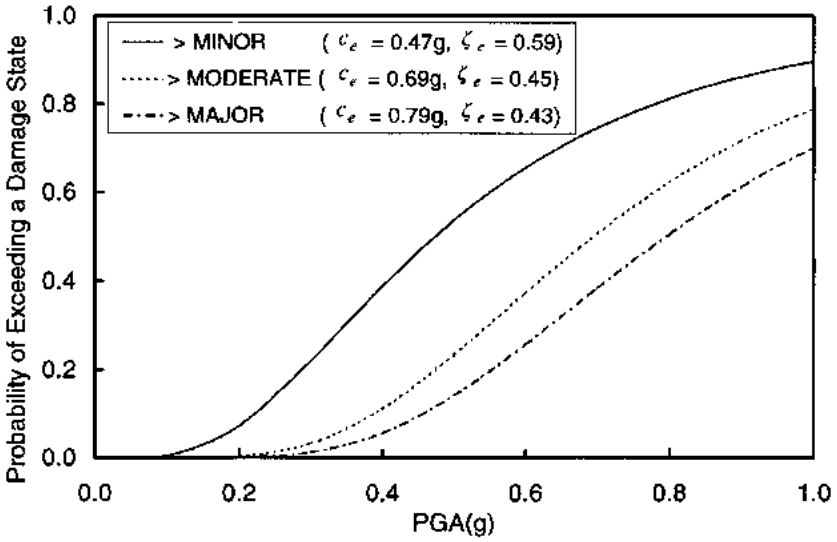


Figure 1.1 – Courbes de fragilité sismique selon PGA [8]

1.2.2. Approches de construction des courbes de fragilité sismique

Il existe plusieurs approches pour établir les courbes de fragilité d'une structure. La classification des approches est réalisée suivant l'origine des données. Celles-ci peuvent être obtenues à partir d'observations post-séisme, d'avis d'expertise ou de résultats d'analyse numérique par des modèles mécaniques. Une comparaison de ces trois approches a été réalisée par Dang [9]. Le Tableau 1.1 présente leurs forces et faiblesses. L'approche numérique a été montrée comme l'approche la plus performante entre les trois. L'étude de la thèse se limite donc à la construction des courbes de fragilité sismique par simulation numérique seulement.

Approche	**Forces**	**Faiblesses**
Empirique	- Observation réelle post-sismique	- Ne tient pas en compte les changements des paramètres structuraux et les mouvements du sol - Rareté des observations
Avis d'expertise	- Informations relatives aux dommages issues des évaluations d'experts et de chercheurs spécialisés - Rapide	- Dépendance de l'expérience individuelle des experts
Numérique	- Informations sur les dommages tirées des analyses numériques	- Analyse temporelle non-linéaire avec plusieurs variables aléatoires - Construction coûteuse
	- Prise en compte de la variation des paramètres structuraux et du mouvement du sol	- Sensible aux choix de méthode d'analyse, à l'analyse des structures et à la définition des états de dommage

Table 1.1 – Forces et faiblesses des approches de construction de la courbe de fragilité sismique

1.2.3. Construction des courbes de fragilité sismique par approche numérique

Cette partie présente brièvement des méthodes de construction de la courbe de fragilité. Une présentation détaillée est trouvée dans le travail de Le *et al.* [10].

Basées sur la simulation et l'hypothèse log-normale de la forme de la courbe de fragilité, il existe quatre méthodes de construction de la courbe de fragilité : la méthode de simulation de Monte Carlo (MCS), la méthode de normalisation de l'intensité sismique (SSI), la méthode du maximum de vraisemblance (MLE), et la méthode basée sur le modèle de demande et de capacité sismiques (PSDM/PSCM). Ici, nous considérons les méthodes les plus performantes : la méthode de simulation de Monte Carlo et la méthode du maximum de vraisemblance.

La méthode par simulation de Monte Carlo est considérée comme la méthode de référence pour comparer des courbes de fragilité construites par les autres méthodes. Cette méthode est réalisée à partir d'un nombre N_{MCS} suffisamment grand des accélérations du sol synthétiques ou sélectionnées. Leurs intensités sismiques sont organisées dans N_I intervalles de $[a_j - da_j, a_j + da_j]$, où da_j est un incrément autour de l'intensité à calculer la probabilité a_j. Il faut noter qu'il y a N_I intervalles et l'intervalle j contient N_j observations. Donc, la totalité des observations N_{MCS} est égale à $\sum_{j=1}^{N_I} N_j$. Après avoir simulé les réponses de la structure sous ces séismes, l'état de la structure est déterminé en comparant la réponse $\mathbf{Y}$ avec sa limite y_0. Chaque simulation numérique présente deux états possibles de la structure sous séisme : soit la structure est défaillie $(y_i = 1)$ quand $\mathbf{Y} \geq y_0$, soit elle est sécurisée $(y_i = 0)$ au contraire. Cette relation est présenté par la fonction d'indicateur **1**, attribuée à 1 au cas de défaillance et à 0 sinon.

Finalement, la fragilité, $Fr^{\text{MCS}}(.)$, et la probabilité de défaillance, $p_f^{\text{MCS}}(.)$, peut être estimée par la méthode MCS selon les formules :

$$Fr^{\text{MCS}}(a_j) \cong \frac{\sum \mathbf{1}\left[\mathbf{Y} \geq y_0 | A = a \in [a_j - da_j, a_j + da_j]\right]}{N_j}, \quad j = 1 \ldots N_I \tag{1.2}$$

$$p_f^{\text{MCS}} \cong \frac{\sum \mathbf{1}\left[Y \geq y_0\right]}{N_{\text{MCS}}} \tag{1.3}$$

À la place d'évaluer plusieurs points sur la courbe de fragilité sismique comme dans la méthode précédente, la méthode du maximum de vraisemblance identifie directement les deux paramètres de la loi log-normale de la courbe de fragilité. Cette approche consiste à modéliser les résultats des expériences numériques à l'aide d'une variable aléatoire de Bernoulli. Supposons le nombre d'accélérations est N_{MLE}, la fonction de vraisemblance maximale à résoudre s'exprime comme suit :

$$L(A_m, \beta; y_i) = \prod_{i=1}^{N_{\text{MLE}}} \left[Fr(a_i)\right]^{y_i} \left[1 - Fr(a_i)\right]^{1-y_i} \tag{1.4}$$

La médiane A_m et l'écart-type logarithme β dans l'équation (1.4) sont estimés grâce au problème d'optimisation associé :

$$(\hat{A}_m, \hat{\beta}) = \arg\min_{A_m, \beta} (-\ln L) \tag{1.5}$$

Cette approche, développée par Shinozuka et al. [8], a été initialement employée pour établir des courbes de fragilité pour les ponts à partir de données empiriques, telles que les observations des dommages réels après des séismes [8]. Par la suite, ces chercheurs ont étendu cette méthode à des résultats obtenus via des analyses numériques [11, 6], où une transformation directe des réponses structurelles en données binaires représentant les états de dommage est réalisée.

1.2.4. Comparaison des méthodes de construction

Chaque méthode présentée dans le paragraphe 1.2.3 a ses avantages et ses inconvénients. Ceux-ci sont résumés dans le Tableau 1.2.

L'avantage de la méthode MCS réside dans sa précision. Elle permet de déterminer des points de défaillance pour différents niveaux d'intensité sismique sur la courbe de fragilité, indépendante

de l'hypothèse de la loi log-normale. Cette approche peut être chronophage, surtout lorsqu'elle est employée pour évaluer des probabilités faibles, car elle requiert un grand nombre de calculs numériques.

D'autre part, la méthode MLE ne nécessite pas de nombreuses simulations. Elle s'accorde bien avec l'hypothèse lognormale, comme le montre le test d'adéquation de Shinozuka [8]. Cette méthode est basée sur des données simples, limitées à des informations binaires sur la défaillance ou non de la structure. Cependant, elle dépend de la qualité des observations disponibles. L'incertitude associée à ces observations, comme le manque d'événements enregistrés, la précision limitée des observations des dommages survenus après un tremblement de terre, etc. peut réduire l'efficacité de la méthode dans ce cas.

Méthodes	**Avantages**	**Inconvénients**
Monte Carlo	- Plus fiable - Indépendance de la loi log-normale	- Plus coûteux en temps
Maximum de vraisemblance	- Coût relativement faible (quelques centaines de calculs suffisent) - Données d'entrées simples (défaillance ou non défaillance)	- Forte dépendance de l'hypothèse log-normale - Pas de prise en compte de la corrélation entre les paramètres de mouvements du sol et les réponses structurelles

Table 1.2 – Avantages et inconvénients des méthodes de construction des courbes de fragilité

1.2.5. Challenges sur la construction des courbes de fragilité sismique

Il convient de noter qu'il peut exister des incohérences significatives entre les courbes de fragilité obtenues par différentes méthodes, même pour une même structure soumise aux mêmes conditions d'aléa sismique. La méthode MLE est la méthode d'approximation la plus prometteuse [10]. Par contre, elle est encore dépendante de l'hypothèse log-normale et des données observées. Lorsque l'hypothèse log-normale est valide, et à condition qu'un grand nombre de simulations soit effectué, la méthode MLE peut fournir des résultats excellents. Cependant, dans la pratique, il peut être difficile, voire impossible, de satisfaire simultanément ces deux conditions. Il est envisageable d'explorer d'autres approches qui permettent de s'affranchir de l'hypothèse de la loi log-normale en dehors de la méthode classique de simulation de Monte Carlo. La thèse de Dang [9] propose une amélioration de la méthode MLE selon cette direction, en incluant la densité de probabilité de l'intensité sismique.

Le calcul probabiliste utilisé pour construire la courbe de fragilité ou pour calculer la probabilité de défaillance repose aussi sur des simulations, nécessitant l'analyse de structures. La méthode de simulation de Monte Carlo est reconnue comme la plus fiable et est souvent considérée comme la méthode de référence pour élaborer ces courbes. Cependant, il convient de noter que cette méthode

exige une quantité considérable de ressources de calcul. Par exemple, selon Ghosh *et al.* [12], il peut être nécessaire d'effectuer plusieurs milliers de simulations basées sur la méthode des éléments finis pour obtenir une estimation suffisamment précise de la probabilité de défaillance d'une structure. Comme l'a souligné Bourinet [13], pour obtenir une estimation de la probabilité de défaillance avec une précision acceptable, il est impératif de réaliser des milliers, voire des dizaines de milliers de simulations. Par conséquent, l'application de cette méthode en général et en particulière aux structures réelles, dans le but de créer un outil efficace, se heurte à un obstacle majeur lié au coût des calculs. Cette problématique est particulièrement prégnante dans le domaine de la sécurité sismique, où les simulations, généralement basées sur la méthode des éléments finis, s'avèrent extrêmement coûteuses, principalement en raison du comportement dynamique non-linéaire des structures soumises à des excitations sismiques sévères. Les simulations nécessitent une charge de calcul importante.

Par conséquent, il reste encore un défi principal pour améliorer la construction de la courbe de fragilité. Ce défi est d'augmenter le nombre d'observations des structures sous séismes, notamment dans la zone de faible probabilité de défaillance. La question se pose donc de savoir s'il est possible de trouver un outil plus performant pour remplacer les simulations mécaniques, dans le but de rendre le processus de construction des courbes de fragilité plus efficace. Cette démarche devrait réduire la charge de calcul associée à l'analyse des structures en conditions sismiques tout en préservant un niveau de précision satisfaisant. Explorer de telles alternatives représente un domaine de recherche prometteur dans le contexte de l'analyse de la fragilité sismique des structures. Dans la suite de la thèse, l'apprentissage automatique émerge comme une solution prometteuse pour relever ce défi.

1.3. Apprentissage automatique dans le domaine de génie civil

1.3.1. Présentation de l'apprentissage automatique

L'apprentissage automatique est une branche de l'intelligence artificielle qui se concentre sur le développement de modèles et d'algorithmes capables d'apprendre à partir des données et d'améliorer leurs performances au fil du temps, sans être explicitement programmés. Il s'agit d'un processus par lequel un système informatique est capable de détecter des schémas et des structures dans les données, et d'utiliser ces informations pour prendre des décisions ou effectuer des prédictions. Depuis l'Antiquité, l'idée de machines pensantes a suscité l'intérêt des esprits. Cela a jeté les bases de l'intelligence artificielle, notamment de l'apprentissage automatique. Maintenant, l'application de l'apprentissage automatique est trouvée facilement dans la vie tels que la reconnaissance d'images, la traduction automatique, l'analyse prédictive, la détection de fraudes, la recommandation de produits, l'analyse des sentiments et bien d'autres.

L'un des avantages clés de l'apprentissage automatique est sa capacité à traiter des quantités massives d'observations pour découvrir des modèles complexes qui seraient difficiles à détecter entre les caractéristiques ("features" en anglais) et les réponses.

Selon les données disponibles pour l'entraînement, il existe différentes catégories d'apprentissage. Il s'agit d'un apprentissage supervisé lorsque les données sont étiquetées, c'est-à-dire que la

réponse cible est connue à partir de ces données. Si les étiquettes sont discrètes, on parle de classification, tandis que si les étiquettes sont continues, on parle de régression. L'apprentissage par renforcement se produit lorsque le modèle est appris en fonction d'une récompense donnée par le programme pour chacune des actions. Dans le cas sans étiquette, on cherche à déterminer la structure sous-jacente des données, comme une densité de probabilité par exemple, et cela s'appelle un apprentissage non supervisé. L'apprentissage automatique peut être appliqué à différents types de données, tels que des photos, des images, des courbes, ou plus simplement des vecteurs numériques, qui peuvent être des variables qualitatives ou quantitatives continues ou discrètes. De nombreux modèles de ML sont développés, selon différents types comme illustré par la Figure 1.2. Chaque modèle de ML a ses points forts et ses points faibles, donc il est utilisé pour différents problèmes. Pour ce travail, les modèles de ML de type régression sont considérés pour prédire la réponse des structures. Dans la section suivante, les modèles de ML les plus souvent utilisés sont présentés et ensuite utilisés dans ce travail. À côté de ces modèles de machine learning, il y en a encore beaucoup d'autres, selon Thai [14]. Une présentation détaillée de chaque méthode est disponible dans le travail de Géron [15]. Dans ce travail, les bibliothèques Tensorflow [16] et scikit-learn [17] sont utilisées.

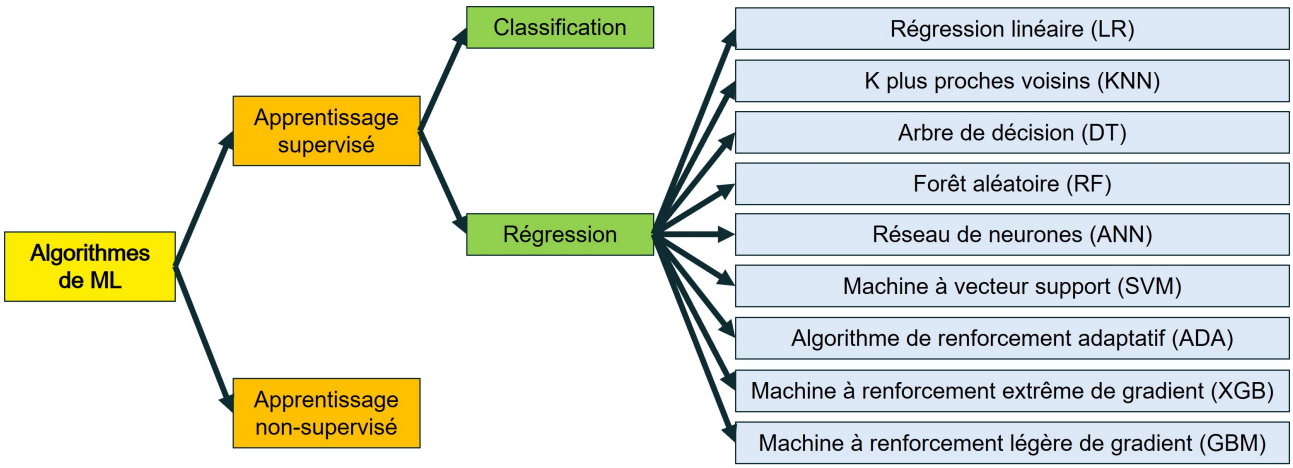


Figure 1.2 – Présentation des modèles d'apprentissage automatique

1.3.2. Algorithmes d'apprentissage automatique

Dans le cadre du domaine du génie civil, la réponse de la structure, qui est le résultat des analyses dynamiques non-linéaires complexes par éléments finis, demande très souvent des temps de calcul prohibitifs. Donc, l'apprentissage automatique décrit une série de méthodes qui permettent d'apprendre à partir de données, c'est-à-dire de déterminer les relations existant entre les quantités d'intérêt. Lors de la réalisation d'études probabilistes liées à la fiabilité des structures, il est prometteur de remplacer la simulation d'éléments finis par un modèle de ML construit sur un ensemble de réponses simulées, pour réduire le temps de calcul. Alors, dans le contexte de cette étude, l'idée de base d'un modèle de ML est de déduire une fonction d'approximation des résultats simulés pour

trouver la réponse à partir des caractéristiques. On suppose que la vraie relation entre la réponse $\mathbf{Y}$ et un vecteur des caractéristiques $\mathbf{x}$, est présentée par la formule (1.6).

$$\mathbf{Y} = f(\mathbf{x}), \qquad \mathbf{x} = \{x_1, x_2, \ldots, x_n\} \tag{1.6}$$

Alors, un modèle d'apprentissage automatique $\mathbf{ML}(\mathbf{x})$ est recherché par faire une approximation de la vraie relation. La prédiction de la réponse, désigné $\mathbf{Y}'$, et sa relation avec la réponse $\mathbf{Y}$ et les caractéristiques $\mathbf{x}$ devient les équations (1.7) et (1.8).

$$\mathbf{Y}' = \mathbf{ML}(\mathbf{x}), \qquad \mathbf{x} = \{x_1, x_2, \ldots, x_n\} \tag{1.7}$$

$$\mathbf{Y} = \mathbf{Y}' + \epsilon = \mathbf{ML}(\mathbf{x}) + \epsilon, \qquad \mathbf{x} = \{x_1, x_2, \ldots, x_n\} \tag{1.8}$$

où ϵ représente l'erreur en raison du manque d'ajustement du ML modèle.

a. Algorithme de régression

L'algorithme de régression est une technique de modélisation prédictive qui a été développée en premier lieu en statistique pour étudier la relation entre les caractéristiques $\mathbf{x}$ et le réponse $\mathbf{Y}$. Cette méthode a ensuite été appliquée en apprentissage automatique sous l'algorithme d'apprentissage supervisé pour prédire les réponses en fonction des valeurs des caractéristiques. Il existe différents types de modèles de régression développés en apprentissage automatique en fonction (i) du nombre de caractéristiques, (ii) du type de caractéristiques et (iii) de la forme de la fonction de régression. Les modèles de régression couramment utilisés en génie civil comprennent :

- Régression linéaire (Linear regression, LR) : La régression linéaire est le modèle de ML de type régression le plus simple. Il ajuste une ligne droite (fonction linéaire) pour déterminer les réponses à partir des caractéristiques. Si une seule caractéristique est utilisée, le modèle est appelé régression linéaire simple. Dans le cas de plusieurs caractéristiques, il est appelé régression linéaire multiple. La Figure 1.3 présente un modèle de régression linéaire simple (ligne rouge) sur l'ensemble des observations (points bleus).
- Régression multivariée : La régression multivariée est une extension de la régression linéaire multiple qui traite des problèmes ayant plus d'une caractéristique. Le mot "multivarié" fait référence à plus d'une réponse, tandis que le mot "multiple" fait référence à plus d'une caractéristique. Le mérite de cette méthode est qu'elle aide à comprendre la corrélation entre les caractéristiques et les réponses. Cette méthode est également largement utilisée en apprentissage automatique pour les problèmes de régression.
- Régression polynomiale : La différence entre régression polynomiale et la régression linéaire réside dans la forme de la ligne de régression. La meilleure ligne d'ajustement du modèle est une courbe (fonction polynomiale) avec une puissance supérieure à un pour les caractéristiques.
- Régression Lasso : La régression Lasso est une version régularisée de la régression linéaire utilisée lorsque les caractéristiques sont fortement corrélées. Dans ce cas, l'utilisation de la technique de régression linéaire peut entraîner un surajustement (overfitting). Par conséquent, ce modèle est proposé pour réduire le problème en ajoutant un terme de régularisation dans la fonction de coût pendant l'entraînement. Le terme de régularisation utilisé est la norme L1 (valeur absolue de la pondération).

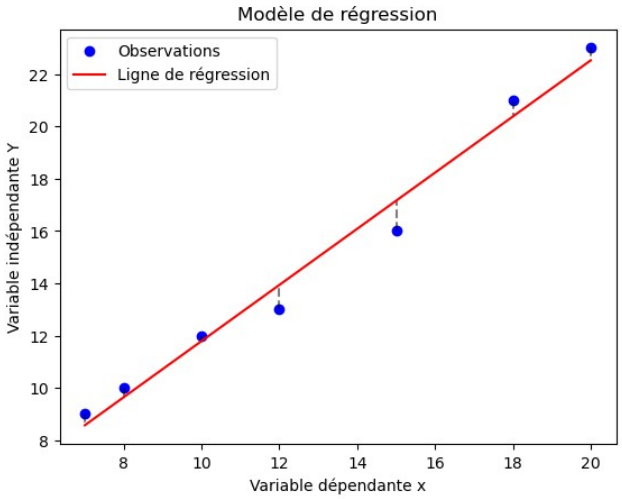


Figure 1.3 – Exemple d'un modèle de régression linéaire

- Régression Ridge : Comme la régression Lasso, la régression ridge est également une version régularisée de la régression linéaire. Par contre, son terme de régularisation est la norme L2 (valeur au carré de la pondération). Le but de cette technique est d'essayer d'éliminer la pondération des caractéristiques les moins importantes.

b. K-Plus Proches Voisins

Le modèle des k plus proches voisins (K-Nearest Neighbor, KNN) [18, 19] est une méthode non paramétrique qui prédit une réponse par une fonction de distance. L'idée principale de cette méthode est de trouver les k observations les plus proches de l'observation considérée en utilisant une mesure de distance appropriée. La classe ou valeur de cette observation est attribuée en voyant celle la plus fréquente parmi les k données plus proches. La Figure 1.4 illustre un exemple du modèle de KNN. Supposons qu'une observation est définie par deux caractéristiques et elle appartient dans une de deux classes différentes. Ces deux classes sont marquées par deux couleurs, cyan et violet. Tous les points en cercle sont les observations d'entrainement du modèle. Une nouvelle observation est marquée par le symbole en rouge, et les lignes pointillées illustrent sa distance vers les observations les plus proches. Selon le nombre de points les plus proches, le modèle décide la classe de cette observation. Dans ce cas d'exemple, cette nouvelle observation est dans la classe violète. Alors, la distance et le nombre de voisins sont les clés pour ce modèle KNN. Seules les mesures de distance suivantes seront considérées dans ce travail, bien qu'il en existe plusieurs autres parmi lesquelles choisir. Pour deux points dont les coordonnées sont respectivement x_i et y_i), la distance est calculée par :

- Distance euclidienne : Cette mesure de distance est la plus reconnue. Elle est applicable uniquement aux vecteurs numériques. Elle calcule la distance euclidienne en suivant une ligne droite

entre deux points, en utilisant la formule ci-dessous.

$$\mathrm{D}_{\text{Euclide}} = \sqrt{\sum_{i=1}^{k}(x_i - y_i)^2} \tag{1.9}$$

- Distance de Manhattan : Cette distance mesure simplement la somme de la différence absolue entre deux points.

$$\mathrm{D}_{\text{Manhattan}} = \left(\sum_{i=1}^{k}|x_i - y_i|\right) \tag{1.10}$$

- Distance de Minkowski : Cette mesure de distance est la forme généralisée de deux mesures. Le paramètre p dans la formule ci-dessous permet de varier la mesure de distance selon la formule suivante.

$$\mathrm{D}_{\text{Minkowski}} = \left(\sum_{i=1}^{k}|x_i - y_i|^p\right)^{1/p} \tag{1.11}$$

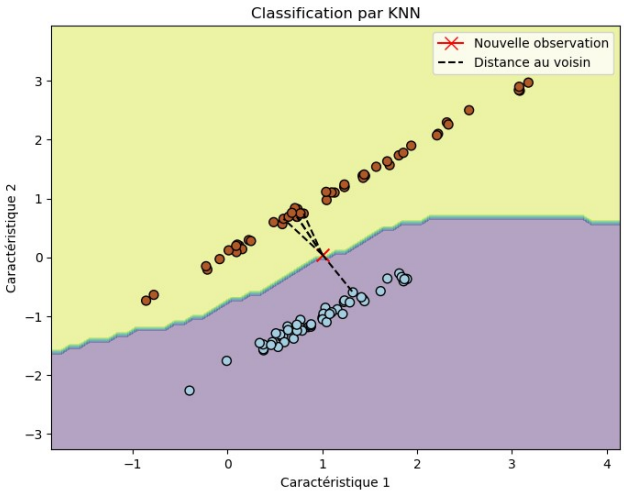


Figure 1.4 – Exemple d'un modèle KNN

Les avantages du modèle KNN sont :

- Simplicité : KNN est un algorithme simple et intuitif. Il ne nécessite pas de phase d'apprentissage coûteuse, car il stocke simplement les observations d'apprentissage pour les utilisations ultérieures.
- Explicabilité des résultats : Étant basé sur des observations réelles, KNN peut fournir des explications sur les prédictions en identifiant les voisins les plus proches.

Cependant, KNN a également certaines limites :

- Sensibilité à l'échelle : Les caractéristiques avec des échelles différentes peuvent avoir un impact disproportionné sur les calculs de distance. Il est souvent nécessaire de normaliser les caractéristiques avant d'appliquer KNN.

- Choix de k : La sélection du bon nombre de voisins k est une décision importante. Un k trop petit peut conduire à une sensibilité excessive aux valeurs aberrantes, tandis qu'un k trop grand peut conduire à une dilution des informations discriminantes.

c. Arbre de décision

Les arbres de décision (Decision tree, DT) [20, 21] sont une autre méthode de régression non paramétrique qui fonctionne en formant un réseau arborescent par l'apprentissage de règles de décision simples déduites des caractéristiques des données. Cette méthode devient très populaire grâce à sa simplicité.

La Figure 1.5 illustre la structure d'un arbre de décision typique et simple, qui contient quatre éléments différents : un nœud racine, des branches, des nœuds internes et les nœuds terminaux (feuilles). Le nœud racine est le nœud le plus élevé d'un arbre, tandis qui les feuilles se situent à l'extrémité de la branche, elles n'ont pas de descendant et elles indiquent une décision à prendre. Les nœuds internes représentent une condition qui permet de diviser l'ensemble de données.

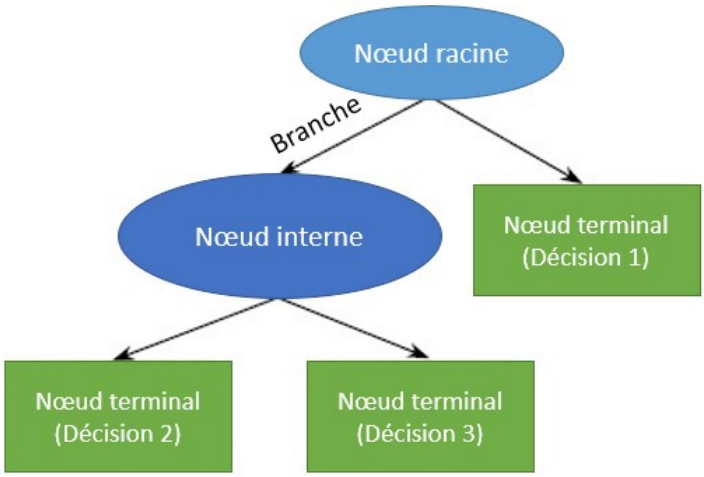


Figure 1.5 – Structure d'un arbre de décision

L'arbre de décision commence par un nœud racine qui représente l'ensemble des données d'entraînement. À chaque nœud, une caractéristique est sélectionnée pour effectuer une division en sous-ensembles plus petits en fonction de la valeur de cette caractéristique. Cette division est basée sur des critères tels que le nombre des observations mal classées (pour la classification) ou la réduction de l'erreur (pour la régression). Le processus de division se poursuit de manière récursive jusqu'à ce qu'une condition d'arrêt soit atteinte. Par exemple lorsque toutes les observations appartiennent à la même classe ou lorsque la profondeur maximale de l'arbre est atteinte. La Figure 1.6 illustre un exemple d'un arbre de décision pour la classification. Dans cet exemple, on essaie de classifier les animaux sauvages. Le premier critère de classification est leur couleur et la deuxième est leur hauteur. Ces deux critères sont des caractéristiques pour entrainer le modèle. Les deux sous-groupes au deuxième niveau sont les nœuds internes. Finalement, ces animaux sont classifiés et mis dans les quatre nœuds terminaux. Cet exemple présente clairement le fonctionnement du modèle DT.

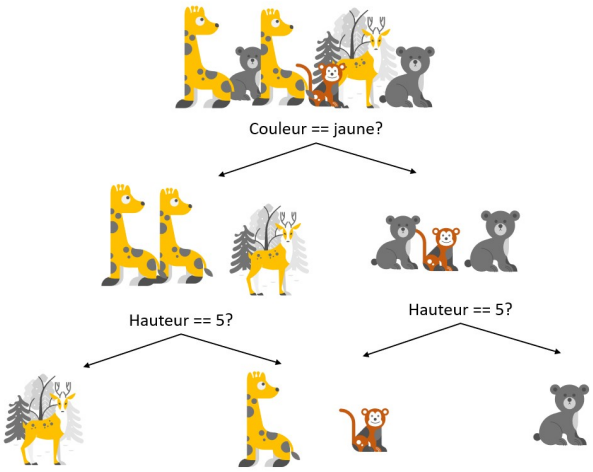


Figure 1.6 – Exemple d'un arbre de décision

Les avantages des arbres de décision incluent :

- Interprétabilité : Les arbres de décision peuvent être facilement visualisés et interprétés, ce qui permet de comprendre les décisions prises par le modèle.
- Vitesse de calcul : Les prédictions d'un arbre de décision sont généralement rapides, car le temps de calcul dépend du nombre de divisions plutôt que de la taille des données.

Cependant, les arbres de décision peuvent souffrir de surajustement, si la profondeur de l'arbre est trop grande ou si le nombre d'exemples d'entraînement est insuffisant. Pour atténuer ce problème, des techniques de régularisation, telles que la limitation de la profondeur de l'arbre ou l'élagage, peuvent être utilisées.

d. Forêt aléatoire

La forêt aléatoire (Random forest, RF) [22, 23] est une méthode qui est créée à partir de multiples structures composées d'arbres de décision. C'est-à-dire que l'idée de ce modèle de ML est de construire une forêt des DTs individuelles basées sur différentes combinaisons des caractéristiques, qui sont sélectionnées de façon aléatoire, ensuite de combiner les résultats des DTs individuelles selon la majorité des votes (pour un problème de classification) ou par la moyenne (pour un problème de régression), comme le montre la Figure 1.7.

La Figure 1.8 présente un exemple du modèle RF. Dans cet exemple, une classification des fruits est réalisée. Chaque arbre de décision donne la classe à une observation. Les résultats de chaque arbre sont possiblement différents. Par contre, le résultat final est réalisé selon la majorité des décisions. Dans ce cas, la classe finale est une pomme, parce que plus d'arbres trouvent ce résultat.

Les forêts aléatoires offrent plusieurs avantages, notamment :

- Robustesse aux données bruitées : Ces algorithmes sont moins sensibles aux valeurs aberrantes et aux erreurs de mesure dans les données.

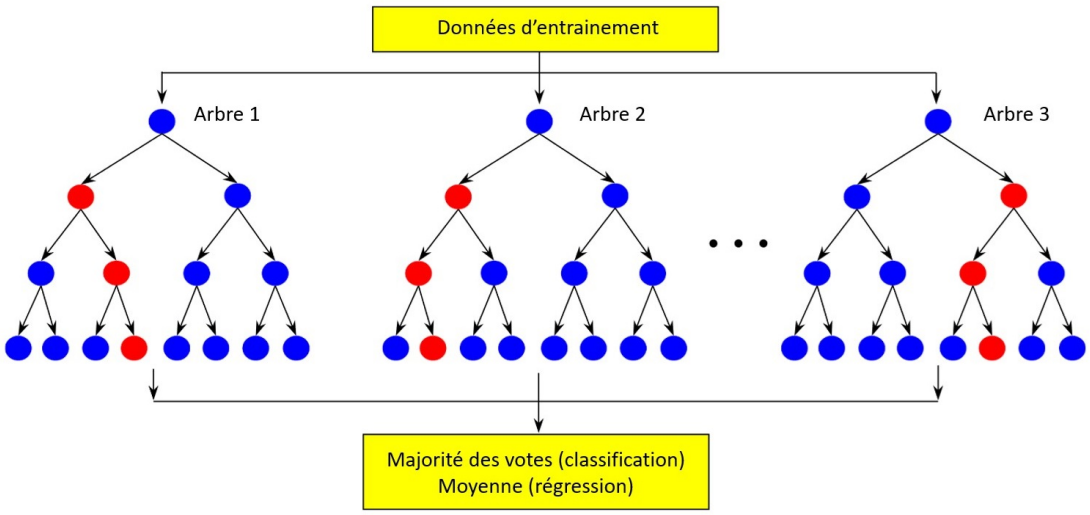


Figure 1.7 – Illustration d'algorithme d'une forêt aléatoire

- Gestion de grandes dimensions : Ils peuvent gérer efficacement les ensembles de données avec un grand nombre de caractéristiques.
- Évaluation de l'importance des caractéristiques : Ils fournissent des mesures d'importance des caractéristiques, ce qui permet d'identifier les caractéristiques les plus influentes dans le modèle.
- Prédiction précise : Ils ont généralement une bonne précision de prédiction, grâce à l'agrégation des prédictions de plusieurs arbres.

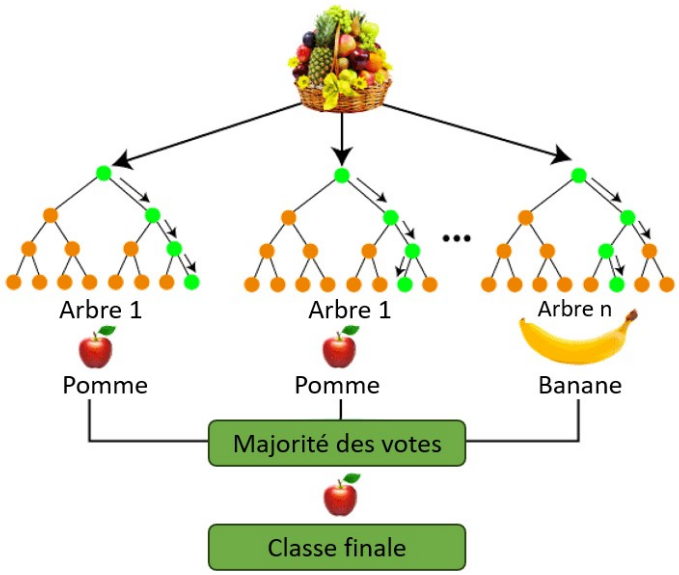


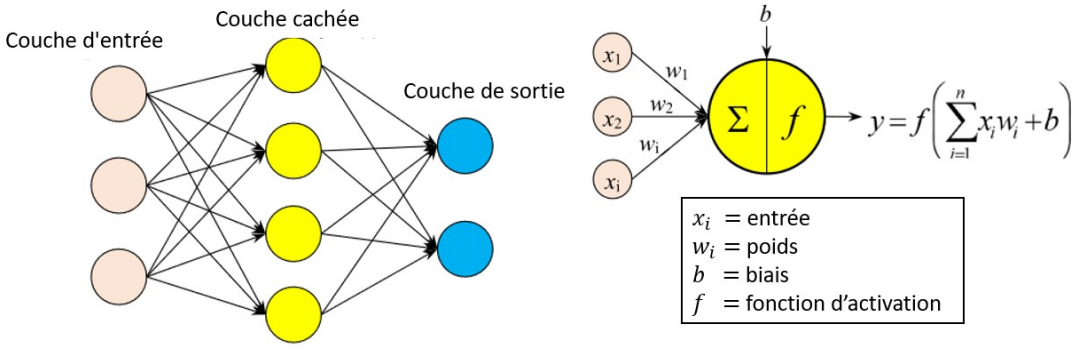
Figure 1.8 – Exemple d'une forêt aléatoire

Elles ont aussi les limites, comme :

- Complexité du modèle : RF est constituée d'un grand nombre d'arbres de décision, ce qui rend son modèle plus complexe et difficile à interpréter que certains autres algorithmes. Il peut être plus difficile d'expliquer les relations entre les caractéristiques dans le modèle final.
- Besoin de ressources computationnelles : La construction d'une forêt peut nécessiter plus de moyens de calcul par rapport à certains autres algorithmes. Cela peut être un inconvénient lorsque l'on travaille avec des ensembles de données très volumineux ou sur des systèmes avec des ressources limitées.
- Sensibilité aux paramètres : La performance de l'algorithme peut dépendre de la sélection des paramètres, tels que le nombre d'arbres, la profondeur maximale des arbres et le nombre de caractéristiques sélectionnées pour chaque arbre. Une mauvaise sélection de ces paramètres peut conduire à des résultats sous-optimaux.

e. Réseau de neurones artificiels

Les réseaux de neurones artificiels (Artificial Neural Network, ANN) [24] sont un type de modèle de régression qui fonctionne en imitant la fonction des neurones en biologie. Ces neurones sont entièrement connectées et arrangées dans les couches. Les réseaux neuronaux typiques sont composés d'une couche d'entrée, de plusieurs couches cachées et d'une seule couche de sortie, comme le montre la Figure 1.9.



(a) Architecture du réseau (a) Neurone de traitement

Figure 1.9 – Exemple d'un réseau de neurones

Le comportement d'un neurone est basé sur une valeur de poids attribuée w_i. Une fois qu'une observation x_i est placée dans la couche d'entrée, elle est multipliée par son poids w_i correspondant. En outre, une fonction de transfert calcule la somme pondérée des observations, avec un biais b, qui est ajusté afin de minimiser la différence entre la prédiction et la réponse actuelle. Ensuite, la valeur calculée par la fonction de transfert passe par une fonction d'activation f pour déterminer si un neurone doit transmettre ses données vers la couche de sortie ou non. Chaque neurone d'une couche cachée a un signal de sortie comme le montre l'équation (1.12). Tous les poids sont attribués aux neurones lorsque l'entraînement du réseau est terminé. La différence entre les prédictions du

réseau et les valeurs actuelles est évaluée afin d'être minimisée, en ajustant les valeurs de poids des neurones.

$$y_k = f\left[\sum_{j=1}^{m} w_{kj}x_j + b_k\right] \tag{1.12}$$

Trois types de fonctions d'activation de régression sont couramment explorés, comprenant la sigmoïde, la tangente hyperbolique ($\tanh$) et l'unité linéaire rectifiée (ReLU).

$$f_{\text{sigmoïde}}(v) = \frac{1}{1+e^{-v}} \tag{1.13}$$

$$f_{\tanh}(v) = \tanh(v) \tag{1.14}$$

$$f_{\text{ReLU}}(v) = \max(0, v) \tag{1.15}$$

ANNs présentent plusieurs avantages et inconvénients dans leur utilisation en apprentissage automatique. Leurs avantages sont :

- Capacité à modéliser des relations complexes : Les ANNs sont capables de capturer et de modéliser des relations complexes entre les caractéristiques et la réponse. Ils peuvent apprendre à reconnaître des motifs et des caractéristiques non-linéaires dans les données.
- Adaptabilité à différents types de données : Ils peuvent être utilisés avec différents types de données, tels que des données numériques, des images, des séquences temporelles, etc. Ils peuvent être appliqués à une large gamme de problèmes.
- Robustesse face au bruit : Les ANNs sont relativement robustes face aux données bruitées. Ils sont capables de généraliser et de faire des prédictions même lorsque les données contiennent des imperfections ou des erreurs.

Les inconvénients de l'ANN sont :

- Complexité du modèle et paramétrage : Les ANNs peuvent être complexes à construire et à paramétrer. Ils nécessitent de choisir l'architecture du réseau, le nombre de couches et de neurones, les fonctions d'activation, les poids initiaux, etc. Une mauvaise configuration peut entraîner des performances médiocres.
- Besoin de grandes quantités de données : Ils ont généralement besoin d'un grand nombre d'exemples d'entraînement pour apprendre efficacement. Si les données d'entraînement sont insuffisantes, il peut y avoir un risque de surajustement.
- Interprétabilité limitée : En raison de leur nature complexe et non linéaire, il peut être difficile à interpréter. Il peut être compliqué d'expliquer pourquoi un réseau de neurones prend une décision spécifique, ce qui peut limiter leur applicabilité dans des domaines où l'interprétabilité est cruciale.
- Temps de calcul et ressources informatiques : Les modèles de ce type peuvent nécessiter des ressources informatiques importantes, en particulier pour les réseaux de grande taille. L'entraînement et l'inférence peuvent être gourmands en temps de calcul, ce qui peut être un inconvénient dans des environnements avec des contraintes de temps ou de ressources limitées.

f. Machine à vecteur de support

La machine à vecteur de support pour régression (Support vector machine for regression, SVMR) [25, 26] est aussi un modèle de ML qui est populaire et puissant. Ce type de modèle de ML est premièrement introduit pour le problème de classification dans les années 1990s, puis développé pour le problème de régression.

L'idée de modèle de machine à vecteur de support (SVM) est de distinguer les observations, dans ce cas appelées vecteurs, et de trouver une séparation optimale, aussi appelée hyperplan, dont la marge est maximale. Les données trouvées sur la marge sont appelées vecteur de support, qui influencent la position et l'orientation de l'hyperplan. Pour le but de régression, ce modèle de ML cherche une fonction qui ajuste mieux les données dans un périmètre de décision par régression linéaire. La ligne qui ajuste mieux est cet hyperplan qui ont le nombre maximal des données dans une limite ε. Ces deux idées sont illustrées par la Figure 1.10.

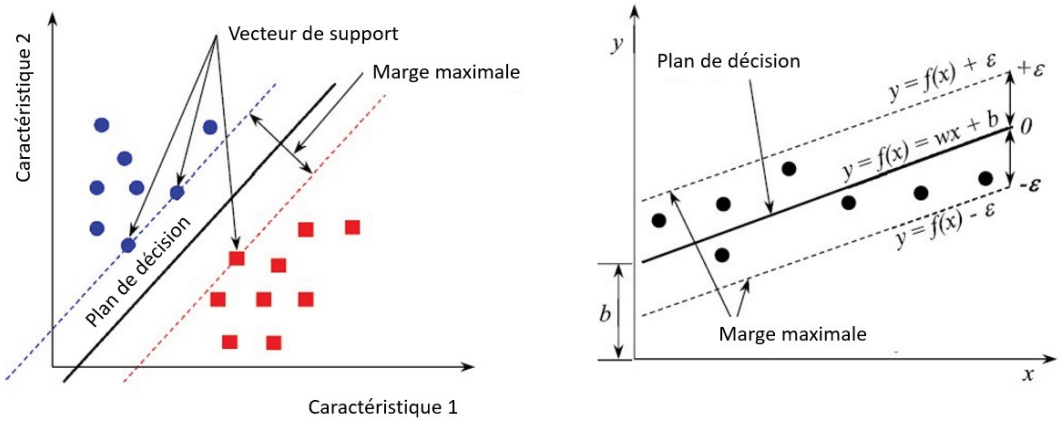


Figure 1.10 – Exemple d'une machine à vecteur de support pour classification et régression

Dans la plupart des cas, les données ne sont pas linéairement séparables, donc, ce n'est pas possible de trouver un hyperplan. Dans ce cas-là, un paramètre de pénalty et une fonction de noyau (kernel function), sont introduits et ensuite illustrés dans la Figure 1.11. Le paramètre de pénalty fait compromis entre maximiser la marge d'hyperplan et minimiser la distance totale des données qui sont mal-classifiées par rapport à son propre groupe. Donc, cette technique permet aux SVMRs de faire certaines erreurs mais de créer des marges suffisamment larges pour que les autres points soient classifiés correctement. En parallèle, la fonction de noyau est utilisée pour que les données originales soient transmises dans un nouvel espace où ces données sont linéairement séparables. Les fonctions de noyau les plus populaires sont une fonction polynomiale linéaire ou non-linéaire, sigmoïdales et radiales. Ces deux techniques influencent fortement la performance du modèle de ML.

Ce modèle possède les avantages :

- Efficacité dans les espaces de grande dimension : Les SVMRs fonctionnent bien lorsque le nombre de dimensions des données est élevé. Cela les rend adaptées à des problèmes où les données peuvent avoir de nombreuses caractéristiques.

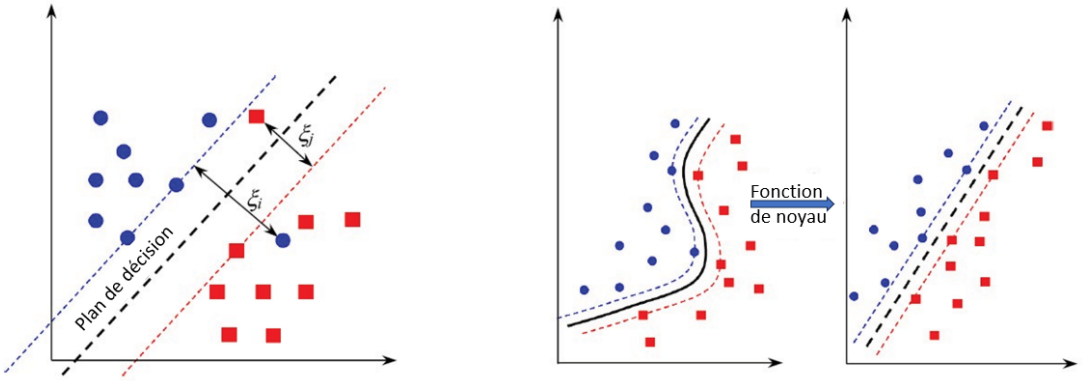


Figure 1.11 – Illustration de paramètre de pénalty ξ et la fonction de noyau

- Résistance aux valeurs aberrantes : Les SVMRs sont relativement résistantes aux valeurs aberrantes dans les données d'entraînement. Les vecteurs de support, qui sont les exemples les plus proches de la frontière de décision, sont les seuls à influencer la construction de l'hyperplan de séparation.

Cependant, il y a aussi quelques inconvénients à prendre en compte :

- Sensibilité aux paramètres : Les SVMRs ont des paramètres, tels que le paramètre de régularisation et le choix du noyau, qui doivent être réglés de manière appropriée. Un mauvais réglage des paramètres peut entraîner une mauvaise performance du modèle.
- Difficulté d'interprétation : Les SVMRs ne fournissent pas une interprétation directe des relations entre les caractéristiques et les classes. Il peut être difficile de comprendre comment les caractéristiques contribuent à la décision de classification.
- Temps d'entraînement plus long pour de grands ensembles d'observations : L'entraînement d'un SVMR peut être plus lent pour de grands ensembles d'observations en raison de la complexité du calcul et de l'optimisation de l'hyperplan de séparation.

g. Algorithme de renforcement adaptatif

L'algorithme de renforcement adaptatif (Adaptive boosting, AdaBoost) [27] est un algorithme dans la famille des algorithmes de renforcement de gradient, et également l'un des algorithmes les plus utilisés. Cette technique fonctionne en combinant plusieurs modèles d'apprentissage automatique de faible performance pour créer un modèle dont la performance est plus forte. AdaBoost est construit sur la base des algorithmes DT, donc les modèles d'apprentissages de faible performance sont les DTs avec un nœud et deux feuilles appelées les tronçons de décision (decision stumps). L'idée derrière AdaBoost est d'améliorer des modèles de faible performance en utilisant des observations pondérées de manière adaptative obtenues sur le résultat des modèles précédents. La Figure 1.12 illustre la procédure de mise en œuvre d'un modèle AdaBoost. Le premier modèle est entraîné en utilisant un poids uniforme pour toutes les observations d'entraînement. Ensuite, le deuxième modèle est entraîné en utilisant les observations pondérées avec les coefficients de poids mis à jour pour tenir compte de l'erreur du premier modèle. C'est-à-dire en augmentant les poids des observations mal

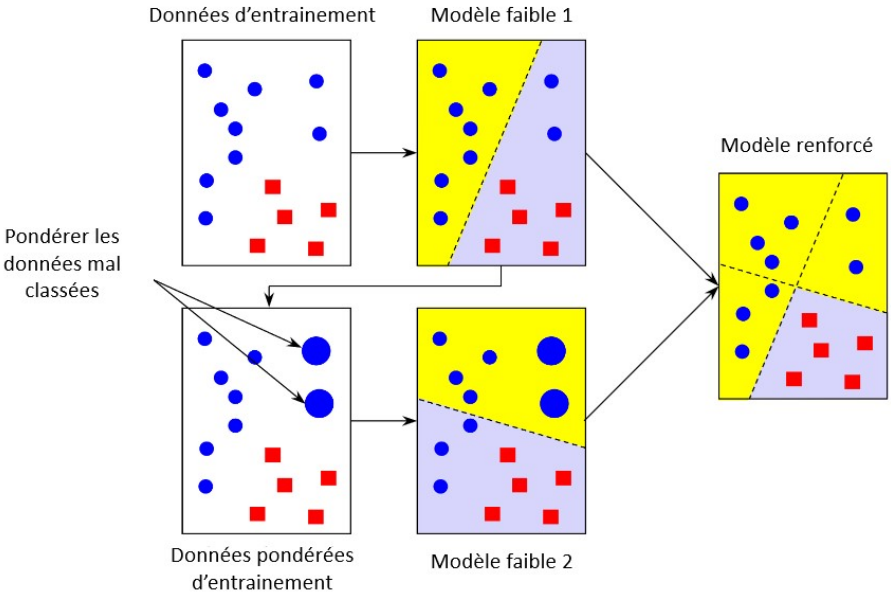


Figure 1.12 – Implémentation de AdaBoost avec deux modèles faibles

classées et en diminuant les poids des observations correctement classées. Ce processus est répété jusqu'au dernier modèle de faible performance. Enfin, le modèle renforcé est formé en combinant les frontières de décision apprises par tous les modèles de faible performance.

Voici quelques avantages de l'algorithme Adaptive boosting :

- Amélioration des performances : AdaBoost peut améliorer considérablement les performances des modèles individuels de faible performance. En combinant plusieurs modèles d'apprentissage de faible performance, il est capable de capturer des motifs complexes dans les observations et de prendre des décisions plus précises.
- Adaptabilité : Cet algorithme est un algorithme adaptatif qui ajuste les poids des observations à chaque itération. Il met davantage l'accent sur les observations mal classées, ce qui permet de se concentrer sur les cas difficiles et d'améliorer progressivement la précision du modèle final.
- Utilisation de modèles simples : Il utilise des modèles d'apprentissage de faible performance, tels que des arbres de décision peu profonds, comme base. Alors, ces modèles sont généralement simples et rapides à entraîner, ce qui réduit le temps de calcul nécessaire pour construire le modèle global.

Cependant, il y a aussi quelques limitations à prendre en compte :

- Sensibilité au bruit et aux valeurs aberrantes : AdaBoost est sensible au bruit et aux valeurs aberrantes dans les observations d'entrainement. Ces éléments peuvent entraîner une mauvaise classification des observations et affecter les performances du modèle global.
- Risque de surajustement : Si le nombre d'itérations du modèle est trop élevé, il existe un risque de surajustement du modèle aux observations d'entrainement. Cela peut entraîner une performance médiocre sur de nouvelles données.

- Dépendance aux observations d'entrainement : Il dépend fortement des données d'entrainement. Si les observations d'apprentissage ne représentent pas fidèlement la distribution des données réelles, cela peut affecter les performances du modèle.

h. Machine à renforcement extrême de gradient

La machine à renforcement extrême de gradient (Extreme Gradient Boosting Machine, XGBoost) [28] est également un algorithme faisant partie de la famille des algorithmes de renforcement de gradient. Il a été proposé par Chen et Guestrin [29] pour améliorer la vitesse et les performances de l'algorithme de renforcement de gradient, qui est souvent lent à mettre en œuvre en raison de son entraînement de modèle séquentiel. XGBoost met en œuvre plusieurs techniques qui le rendent plus rapide que les modèles précédents. Par exemple, une technique de randomisation est utilisée pour réduire le surajustement et accélérer l'entraînement. Une structure basée sur les colonnes compressées est utilisée pour stocker les données, réduisant ainsi le coût du tri, qui est la partie la plus chronophage de l'entraînement du modèle. Le calcul parallèle et distribué est également mis en œuvre pour utiliser tous les processeurs disponibles pendant l'entraînement et la recherche des branches de décision. L'implémentation du traitement parallèle rend XGBoost extrêmement puissant pour résoudre rapidement et avec précision de grands problèmes avec de vastes ensembles de données. Il est donc considéré comme l'une des méthodes d'apprentissage automatique les plus efficaces.

Le modèle de XGBoost possède les avantages suivants :

- Haute performance : XGBoosts est connu pour sa haute performance prédictive. Il est souvent considéré comme l'un des meilleurs algorithmes d'apprentissage automatique en raison de sa capacité à produire des modèles précis.
- Régularisation intégrée : XGBoost offre plusieurs techniques de régularisation qui aident à contrôler la complexité du modèle et à prévenir le surajustement.
- Parallélisme et évolutivité : XGBoost est conçu pour tirer parti du parallélisme et peut fonctionner efficacement sur des configurations à plusieurs cœurs ou sur des clusters. Cela permet d'accélérer le temps de calcul et de gérer de grandes quantités de données.

Cependant, il y a quelques points à prendre en compte avec XGBoost :

- Complexité des paramètres : XGBoost dispose d'un large nombre de paramètres à régler, ce qui peut rendre sa configuration complexe. Trouver les meilleurs paramètres pour un ensemble de données spécifique peut nécessiter des essais et des ajustements supplémentaires.
- Temps d'apprentissage plus long : Par rapport à certains autres algorithmes d'apprentissage automatique, XGBoost peut nécessiter un temps d'apprentissage plus long en raison de la complexité de son processus de construction des arbres de décision.
- Sensibilité aux valeurs aberrantes : Comme la plupart des modèles basés sur les arbres de décision, XGBoost peut être sensible aux valeurs aberrantes dans les données. Les valeurs aberrantes peuvent influencer négativement les performances du modèle.

i. Machine à renforcement léger de gradient

La machine à renforcement léger de gradient (Light Gradient Boosting Machine, LightGBM) [30] est aussi un algorithme dans la famille des algorithmes de renforcement de gradient. Il est développé

par Ke *et al.* [31] pour se concentrer sur l'efficacité computationnelle avec un niveau de précision acceptable. Contrairement à d'autres algorithmes comme XGBoost, LightGBM utilise une stratégie de croissance de l'arbre axée sur les feuilles pour développer le modèle. Cet algorithme choisit la feuille qui donne la plus de perte de prédiction pour développer les branches de décision, à la place de développer les branches pour tous les nœuds internes. Deux fonctionnalités importantes sont implémentées dans LightGBM : l'échantillonnage d'un côté basé sur le gradient (Gradient-based One-Side Sampling) et le regroupement exclusif de caractéristiques (Extreme Feature Boosting). L'idée de l'échantillonnage d'un côté basé sur le gradient est de se concentrer sur les observations mal entrainées et de garder les observations bien entrainées pour les itérations suivantes. Le regroupement exclusif de caractéristiques permet de regrouper des caractéristiques dont les valeurs sont simultanément égales à zéro, donc réduit le nombre des caractéristiques. L'utilisation de ces fonctionnalités peut accélérer le processus d'entraînement en réduisant le nombre de nœuds internes et les branches de décision. LightGBM est conçu pour être utilisé avec de grandes quantités d'observations et de caractéristiques. Par contre, le développement de cette méthode peut causer un surajustement sans limiter le développement des branches de décision.

Voici quelques avantages spécifiques de LightGBM :

- Efficacité en termes de mémoire : LightGBM utilise des fonctionnalités, qui permettent de réduire considérablement l'utilisation de la mémoire. Cela permet de travailler avec des ensembles de données plus volumineux sans sacrifier les performances.
- Vitesse de calcul élevée : LightGBM est optimisé pour une vitesse de calcul élevée. Il utilise des techniques telles que le regroupement des caractéristiques, l'élagage de l'arbre et l'optimisation des calculs parallèles pour accélérer le processus d'apprentissage et de prédiction.
- Précision et performance : LightGBM offre une bonne performance en utilisant des arbres de décision peu profonds et en optimisant les critères de scission.

Cependant, il y a quelques points à prendre en compte avec LightGBM :

- Besoin de données suffisantes : Comme tout algorithme de renforcement de gradient, LightGBM peut souffrir de surajustement si les observations d'entrainement sont insuffisantes ou non représentatives de la population générale. Il est important d'avoir un ensemble d'observations d'entrainement de taille adéquate.
- Sensibilité aux paramètres : La performance de LightGBM peut être influencée par les paramètres choisis. Il est nécessaire de procéder à une recherche et à une optimisation appropriée des paramètres pour obtenir les meilleurs résultats.

j. Avantages et inconvénients des algorithmes d'apprentissage automatique

Après avoir considéré ces méthodes, le Tableau 1.3 montre un bref résumé de ces algorithmes. Dans ce tableau, les avantages et inconvénients de chaque algorithme sont présentés.

Les algorithmes LR, KNN et DT sont les algorithmes les plus simples et très faciles à mettre en œuvre. Par contre, ils ne sont pas très adaptés pour les cas où la relation entre les données et la réponse est très complexe. Dans ce cas-là, les algorithmes de RF, ANN, SVMR ou les algorithmes à renforcement donnent les résultats plus optimistes. Par contre, pour mettre en œuvre ces algorithmes, il y a plusieurs paramètres à régler. Cela pose une question par rapport à la sélection d'un algorithme de machine learning adaptée pour modéliser la relation entre les données et la réponse.

Algorithme	**Avantages**	**Inconvénients**
LR	- Simplicité - Explicabilité des résultats	- Sensibilité aux valeurs aberrantes - Non optimal pour des relations complexes
KNN	- Simplicité - Explicabilité des résultats	- Sensibilité au choix de k - Non optimal pour des relations complexes
DT	- Interprétabilité - Vitesse de calcul	- Surajustement
RF	- Robustesse aux données bruitées - Gestion de grandes dimensions - Évaluation de l'importance des caractéristiques - Prédiction précise	- Complexité du modèle - Besoin de ressources computationnelles - Sensibilité aux paramètres
ANN	- Capacité à modéliser des relations complexes - Adaptabilité aux données différentes - Robustesse face au bruit	- Besoin de grandes quantités de données - Complexité du modèle - Interprétabilité limitée - Temps de calcul
SVMR	- Efficacité dans espaces de grande dimension - Résistance aux valeurs aberrantes	- Sensibilité aux paramètres - Difficulté d'interprétation - Temps d'entraînement plus long pour de grands ensembles de données
AdaBoost	- Amélioration des performances - Adaptabilité - Utilisation de modèles simples	- Sensibilité au bruit et aux valeurs aberrantes - Risque de surajustement - Dépendance aux données
LightGBM	- Efficacité en termes de mémoire - Vitesse de calcul élevée	- Besoin de données suffisantes - Sensibilité aux paramètres
XGBoost	- Haute performance - Régularisation intégrée - Parallélisme et évolutivité	- Complexité des paramètres - Temps d'apprentissage plus long - Sensibilité aux valeurs aberrantes

Table 1.3 – Avantages et inconvénients des algorithmes d'apprentissage automatique

1.3.3. Entrainement d'un modèle d'apprentissage automatique

En générale, afin d'obtenir un modèle de machine learning, l'apprentissage implique trois étapes principales. La première étape est de construire un jeu de données à apprendre. Ces données, aussi appelées observations, sont disponibles en nombre fini lors de la phase de conception d'un système. La deuxième étape consiste à construire un modèle à partir des données. Cette étape est de construire un modèle qui montre la relation dans la base de données, par exemple, relation entre les caractéristiques (features) et les étiquettes. Celle-ci, communément appelée "apprentissage" ou "entraînement", est généralement effectuée avant l'utilisation pratique du modèle. La troisième étape consiste à mettre le modèle en production : une fois le modèle déterminé, de nouvelles données peuvent être soumises pour obtenir le résultat correspondant à la tâche souhaitée. Dans la pratique, certains systèmes peuvent continuer à s'améliorer une fois en production s'ils peuvent obtenir un retour sur la qualité des résultats produits. La Figure 1.13 illustre une procédure typique d'entrainement d'un modèle de ML. Les méthodes d'apprentissage automatique utilisées dans cette étude sont mises en œuvre à l'aide du langage de programmation Python avec la bibliothèque TensorFlow [16] via JupyterLab.

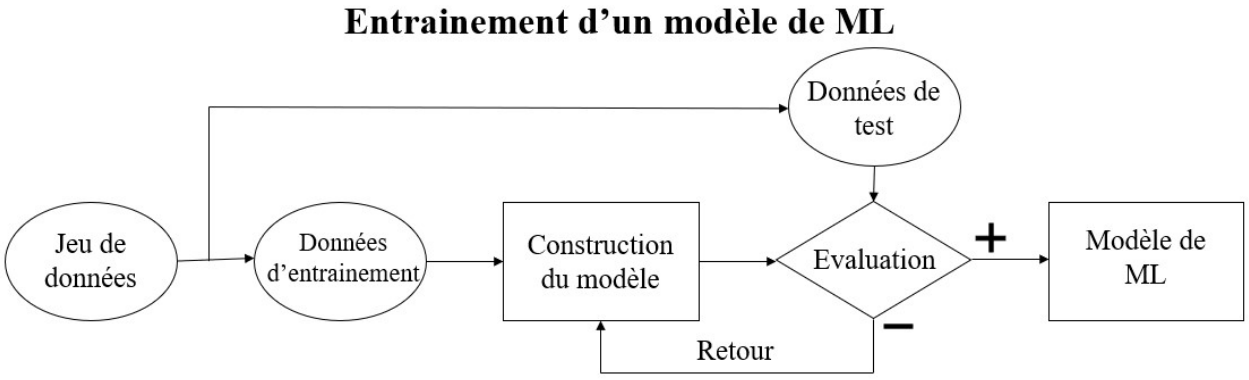


Figure 1.13 – Procédure typique d'entrainement un modèle d'apprentissage automatique

- **Première étape :** La base de données est normalement présentée sous forme d'un ensemble des caractéristiques $\mathbf{x}$ et des réponses correspondantes $\mathbf{Y}$. Les caractéristiques d'entrée sont aussi appelées *features* et les réponses sont aussi appelées *étiquette* ou *label*, selon les terminologies d'apprentissage automatique en anglais. Cette base de données est divisée de façon aléatoire en deux parties, l'une partie d'entrainement et l'autre partie de test. La partie d'entrainement contient habituellement 70-80 % des données de la base initiale.
- **Deuxième étape :** Le but de cette étape est d'entrainer le modèle de ML avec la partie d'entrainement des données. Il existe de nombreux algorithmes de ML qui sont développés, et ils sont présentés dans la section 1.3.2.
- **Troisième étape :** Une fois le modèle de ML entrainé, sa performance est évaluée. Le rôle de ces critères est important, donc il est nécessaire de bien les choisir. Les critères d'évaluation sont nombreux. Dans l'étude présentée, la performance du modèle est évaluée par quatre critères couramment utilisés qui sont des indicateurs de l'exactitude et de la précision d'un modèle,

notamment le coefficient de corrélation de Pearson r, le coefficient de détermination R^2, l'erreur absolue moyenne symétrique en pourcentage $SMAPE$, et l'erreur quadratique moyenne $RMSE$. Les équations (1.16), (1.17), (1.18) et (1.19) présentent les formules de définition des métriques r, R^2, $SMAPE$ et $RMSE$ respectivement :

$$\text{Pearson's} \quad r = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(Y_i' - \bar{Y}')}{\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}\sqrt{\sum_{i=1}^{n}(Y_i' - \bar{Y}')^2}} \tag{1.16}$$

$$R^2 = 1 - \frac{\sum_i (Y_i' - \bar{Y})^2}{\sum_i (Y_i - \bar{Y})^2} \tag{1.17}$$

$$SMAPE = \frac{100\%}{n}\sum_{i=1}^{n}\frac{|Y_i - Y_i'|}{(|Y_i| + |Y_i'|)/2} \tag{1.18}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i' - Y_i)^2} \tag{1.19}$$

où Y_i est la valeur réelle et Y_i' est la prédiction correspondante de l'observation i. $\bar{Y}, \bar{Y}'$ sont leurs valeurs moyennes. Elles sont calculées par :

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \tag{1.20}$$

$$\bar{Y}' = \frac{1}{n}\sum_{i=1}^{n} Y_i' \tag{1.21}$$

où n est le nombre d'observations.
Le coefficient de Pearson r est une mesure de la relation statistique entre deux variables continues, dont les valeurs sont comprises entre -1 et 1, les valeurs les plus élevées proches de 1 signifiant une corrélation linéaire parfaite entre l'ensemble $\mathbf{Y}$ et $\mathbf{Y}'$. Le coefficient de détermination R^2 quantifie la qualité de l'ajustement du modèle sur la base de la méthode de la somme des carrés, où les valeurs proches de 1 traduisent une meilleure performance, et les valeurs inférieures à 0 signifient une performance objectivement plus mauvaise que l'ajustement des valeurs à leur propre valeur médiane. Les valeurs de $SMAPE$ et $RMSE$ quantifient respectivement le pourcentage d'erreur du modèle et l'écart du modèle. Ainsi, les modèles les plus performants verront des valeurs d'erreur plus faibles. Non seulement ces métriques sont utilisées pour évaluer les modèles de machine learning, mais aussi une combinaison de ces métriques est utilisée. De plus, un indice de référence est possiblement employé. Cet indice personnalisé est calculé, par exemple en utilisant les valeurs normalisées des quatre mesures de performance avec un schéma de pondération de 10 % pour r, 30 % pour R^2, 30 % pour $SMAPE$ et 30 % pour $RMSE$, comme suggéré par Todorov *et al.* [32]. Cet indice de référence permet de comparer directement les modèles d'apprentissage automatique entraînés.

1.3.4. Bilan de l'apprentissage automatique en génie civil

La combinaison de la simulation numérique, du calcul probabiliste et de l'apprentissage automatique a suscité un fort intérêt dans la littérature ces dernières années dans le domaine de génie civil.

Thai [14] a fait une enquête bibliométrique dans la littérature actuelle de l'utilisation des méthodes ML pour les applications d'ingénierie structurelle, donc la Figure 1.14 montre l'augmentation d'intérêt du sujet dans le domaine. Ce sujet a reçu plus d'attention au cours des cinq dernières années, comme en témoigne la croissance du nombre de publications. Cela peut être expliquée par les progrès récents des algorithmes de ML, de la puissance du calcul ainsi que par la disponibilité de grands ensembles de données collectées à partir de l'expérience ou de la modélisation numérique.

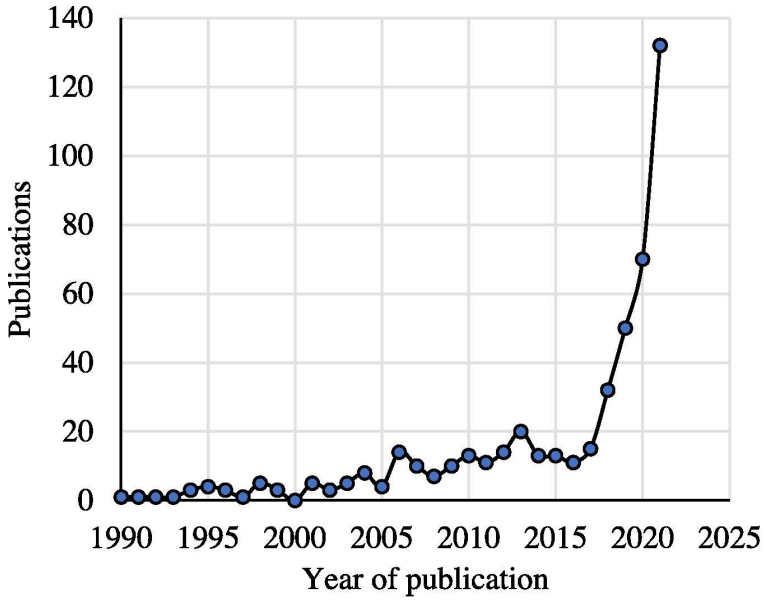


Figure 1.14 – Distribution annuelle d'articles relatifs aux applications du ML dans l'ingénierie des structures [14]

L'apprentissage automatique es utilisé pour différents objectifs. Quatre sujets suivants représentent les principales applications de ML dans le domaine du génie civil.

- **Prédiction des réponses structurales** : Ce sujet de recherche obtient le plus d'attention dans le cadre de l'application des méthodes de ML. La prédiction est réalisée pour différentes réponses de la structures, par exemple la résistance de cisaillement des membres structuraux [33, 34], la résistance axiale [35, 36], la résistance à la flexion et torsion, la force d'adhérence entre béton et armature [37, 38] ou la résistance au flambement des structures [39, 40].
- **Prédiction des propriétés des matériaux** : De la même manière, de grandes efforts sont réalisés afin d'utiliser l'apprentissage automatique pour prédire les propriétés des matériaux, spécialement du béton et de l'acier dans le domaine du génie civil. La plupart des applications de ML sur le béton se concentrent sur la prédiction de la résistance à la compression [41, 42, 43], la résistance à la traction [44, 45], le module d'Young [46, 47]. Les propriétés de l'acier ont aussi attiré les intérêts de la recherche [48, 49, 50].

- **Surveillance de la santé des structures et la détection des endommagements** : La surveillance de la santé des structures est un processus réalisé par la détection et l'évaluation de leur état. L'utilisation de ML dans le processus de surveillance de la santé des structures devient plus populaire, comme il est possible d'appliquer ces algorithmes pour deux méthodes de surveillance : l'une basée sur la vibration et l'autre basée sur l'analyse des images. Les références [51, 52, 53] montrent l'utilisation de l'apprentissage automatique dans ce sujet de recherche.
- **Analyse et dimensionnement des structures** : L'utilisation du ML dans l'analyse et le dimensionnement de la structure a été initiée en 1989 par Adeli et Yeh [54]. Ensuite, ces algorithmes de ML sont utilisés pour dimensionner et analyser différents éléments structuraux ou différentes structures. Par exemple, l'apprentissage automatique est utilisé pour dimensionner des structures en béton armé [55, 56], ou analyser la capacité des structures [57, 58].

Les Tableaux 1.4 et 1.5 présentent un résumé des études sur la prédiction des réponses structurales et la prédiction des propriétés des matériaux. Ces études sont divisées selon leurs bases de données : l'un est par les simulations numériques et l'autre est par les résultats expérimentaux. Ces références résument les recherches utilisant les modèles d'apprentissage automatique de régression et aussi de classification. Pour chaque étude, la structure, les données, la réponse souhaitée, les caractéristiques, les algorithmes d'apprentissage automatique adoptés ainsi que les métriques d'évaluation de la performance du modèle sont présentés.

Le Tableau 1.4 résume les études qui ont utilisé des modèles d'apprentissage automatique pour estimer les réponses des structures différentes basées sur la simulation. La base de données simulées permet d'étudier des problèmes plus complexes, comme le dynamique et la vulnérabilité des structures. Ces études se sont appuyées fortement sur des mesures d'intensité des signaux sismiques, IMs, en tant que caractéristiques du modèle de machine learning, telles que Kiani *et al.* [59], Wang *et al.* [60], etc. Ces études sont analysées plus en détail dans la section suivante. En outre, des recherches ont aussi adopté une approche plus complexe en intégrant des paramètres liés au dimensionnement structural, tels que le nombre d'étages du bâtiment, le chargement appliqué dans la conception des structures, ainsi que des éléments liés à la modélisation numérique, comme l'amortissement, la loi de comportement linéaire ou non-linéaire, ou bien d'autres caractéristiques relatives aux propriétés des matériaux, comme la limite d'élasticité de l'acier ou du béton. Cela permet d'ouvrir la possibilité d'évaluation d'un ensemble de structures.

Le Tableau 1.5 résume les études qui ont développé des modèles d'apprentissage automatique en utilisant des données expérimentales. Pour les recherches conduites à partir des expériences, la taille des données est très limitée. Leurs bases de données expérimentales sont souvent des résultats des études antérieures. Par exemple, Mangalathu *et al.* [61] emploient les observations depuis 393 murs résistants au cisaillement en béton armé à un étage et une portée, dont les sections transversales sont rectangulaires et autres. La plupart des spécimens utilisés dans cette base de données ont été extraits, à partir de deux études de Usta *et al.* [62] et de Grammatikou *et al.* [63], avec des autres tests expérimentaux précédents.

Peu importe la nature des données, les modèles de machine learning sont similaires entre les études. Ces modèles qui sont très souvent utilisés dans le domaine d'apprentissage automatique comme : régression linéaire (LR), k-plus proches voisins (KNN), forêt aléatoire (RF), arbre de décision (DT), machines à vecteurs de support (SVMR), réseaux de neurones artificiels (ANN), algorithme de

renforcement adaptatif (AdaBoost), machine à renforcement léger de gradient (LightGBM) et machine à renforcement extrême de gradient (XGBoost). Il y a aussi quelques autres modèles comme surface de réponse polynomiale (PRS), fonction adaptative de base (ABFC), splines régressives adaptatives multivariées (MARS), fonction radiale de base (RBF), etc. . De plus, les métriques d'évaluation sont aussi similaires pour ces études : r, R^2 et $RMSE$ pour les algorithmes de régression, ou la matrice de confusion pour les algorithmes de classification.

Une différence notable entres les études pour ces deux bases de données est le nombre d'observations. Pour les études basées sur les expériences, le nombre d'observation reste seulement à centaine, ce qui n'est pas vraiment optimal dans le domaine de machine learning.

Par conséquent, une difficulté, qui a empêché le développement de machine learning dans le domaine de l'ingénierie, est le manque de bases de données pour assurer la validation des modèles prédictifs. Toutefois, ces dernières années, la recherche a pris les mesures nécessaires pour surmonter cet obstacle en créant des plateformes de partage des bases de données. Par exemple, Datacenterhub [64] est une plateforme pour le stockage, la publication et le partage de jeux de données scientifiques et d'ingénierie principalement liés aux contraintes subies par les structures en béton lors de conditions sismiques. Similairement, la base de données du Pacific Earthquake Engineering Research Center [65] est aussi une ressource précieuse pour la recherche en génie parasismique et sismologie. Le PEER, basé à l'Université de Californie, à Berkeley, offre un accès à une large base de données liées aux séismes, notamment des données expérimentales, des enregistrements de mouvements sismiques et des modèles de calcul. Les chercheurs et les ingénieurs utilisent souvent cette base de données pour étudier le comportement des structures et des matériaux sous l'effet des séismes. La base de données cybernétique DesignSafe [66] a été développée pour permettre et faciliter la recherche en génie des risques naturels, qui nécessite nécessairement de couvrir plusieurs disciplines et peut tirer parti des avancées en matière de calcul, d'expérimentation et d'analyse de données. DesignSafe permet aux chercheurs de partager et de trouver plus efficacement des données grâce à des services cloud, d'effectuer des simulations numériques à l'aide de calculs haute performance et d'intégrer divers ensembles de données de manière à ce que les chercheurs puissent faire des découvertes qui étaient auparavant inaccessibles.

1.3.5. Conclusion

L'utilisation de l'apprentissage automatique dans le domaine du génie civil a évidemment suscité un intérêt croissant au cours des dernières années. Cela découle en grande partie de la capacité de l'apprentissage automatique à extraire des informations utiles à partir de données complexes et volumineuses, ce qui présente un potentiel significatif pour ce domaine. Les avantages de l'application de l'apprentissage automatique dans ce domaine sont nombreux.

Tout d'abord, l'apprentissage automatique permet une analyse plus approfondie des données, ce qui peut conduire à une meilleure compréhension des comportements structuraux, des charges, des contraintes et des réponses aux différents facteurs. Cela peut aider les ingénieurs à concevoir des structures plus efficaces et plus durables. De plus, l'apprentissage automatique offre la possibilité de prédire le comportement des structures, ce qui peut être particulièrement utile pour évaluer la sécurité et la performance des bâtiments et des infrastructures sous diverses conditions, y compris

les séismes. Ces prédictions peuvent servir de base à des décisions éclairées en matière de conception et de gestion des actifs.

Cependant, l'utilisation de l'apprentissage automatique dans le génie civil n'est pas sans défi. Les modèles d'apprentissage automatique dépendent fortement de la qualité et de la quantité des données d'entraînement, ce qui peut être un défi dans le domaine du génie civil, où les données expérimentales sont souvent rares. De plus, les modèles d'apprentissage automatique peuvent être complexes et difficiles à interpréter, ce qui soulève des questions de transparence et de confiance [67].

Malgré ces défis, l'apprentissage automatique offre de nouvelles perspectives passionnantes pour le génie civil. Il peut contribuer à une conception plus efficace, à une maintenance préventive, à une gestion des risques améliorée et à une prise de décision plus éclairée. Par conséquent, de nombreux acteurs du génie civil investissent dans la recherche et le développement de solutions d'apprentissage automatique pour améliorer la durabilité, la sécurité et la résilience des infrastructures, ce qui ouvre la voie à des avancées significatives dans le domaine.

Étude	Structure	Taille de la base de données	Réponse ciblée	Caractéristiques	Modèles de ML	Critère d'évaluation
Sun *et al.* [68]	Ossatures de contreventement retenues par flambage de l'acier	60 observations (1 structure x 60 GMs [a])	Déplacement relatif maximal	PGA, PGV, PGD, $S_a(T_1^0)$, ASA, CAV, I_A	ANN	MSE
Lagoros *et al.* [69]	Portiaue en acier	NA (95 GMs)	Déplacement relatif maximal	PGA, PGV, PGD, V/A, I_A, $S_D(T_1^0)$, a_{RMS}, ChI, CAV, SI, T_D, $S_v(T_1^0)$, S_aC	ANN	Erreur médiane
Zhou *et al.* [70]	Barrage en enrochement à haute teneur en béton	100 observations (1 structure x 100 GMs)	Déplacement vertical, déformation au cisaillement plastique, état d'endommagement	PGA, CAV, I_A, T_P	SVM	MSE, R^2, MAE
Mitropoulou *et al.* [71]	Portique en béton armé	200 observation (2 structures $\times$ 100 GMs)	Déplacement relatif maximal	I_A, I_C, CAV	ANN	NA
Wang *et al.* [60]	Équipement des centrales nucléaires	100 observations (1 structure x 100 GMs)	Défaut d'ancrage de l'armoire électrique	PGA, PGV, PGD, I_A, CAV, PS_a, ASA, T_P	ANN	$RMSE, R^2, RMAE$
R. Segura *et al.* [72]	Barrage-poids	250 observations	Glissement de base	$S_{aH}(T_1^0)$, $S_{aV}(T_1^0)$, $S_{vH}(T_1^0)$, PGA, PGA_V, PGV, PGD, SI, ω_{eqk}, I_A, Ds_{5-95}	PRS, ABFC, MARS, RBF, RF, SVM	$RMSE, R^2, RMAE$
Rezaei *et al.* [73]	Pont à poutres-caissons en béton	1440 observations (9 structures $\times$ 160 GMs)	Paramètres de la demande d'ingénierie	$S_a(1.0\text{s})$	Régression symbolique	$RMSE$, MAE, r

Table 1.4 – Bilan des études de ML par simulation

a. GMs : enregistrements sismiques

Noureldin *et al.* [74]	Système à un degré de liberté	61200 observations (400 structures $\times$ 153 GMs)	Paramètres de la demande d'ingénierie	PGA, $S_a(T_i)$ et paramètres du systèmes	ANN	$RMSE$, R^2
Kiani *et al.* [59]	Portique en acier	2000 observations (1 structure $\times$ 2000 GMs)	État d'endommagement de la structure	$S_a(T_i)$, ASI, SI, DSI, PGA, CAV, D_{s5-75}, D_{s5-95}	Logistic, LASSO, SVM, Naïve Bayes, DT, RF, KNN, DA, ANN	Rappel, précision et F-mesure
Zhang *et al.* [75]	Portique en béton armé	9900 observations (199 structures $\times$ 240 GMs)	État d'endommagement de la structure	PGA, EPA, PGV, SI, $S_a(T_i)$ et paramètres de conception	RF, XGB	Rappel, précision et F-mesure
Nguyen *et al.* [76]	Portiques résistants aux moments en acier	112320 observations (468 structures $\times$ 240 GMs)	$S_a(T_i)$ et paramètres de la structure	État d'endommagement de la structure	KNN, DT, RF, ADA, XGB, LGB, CatBoost	Rappel, précision et F-mesure
Guan *et al.* [77]	Portiques résistants aux moments en acie	149040 observations (621 structures $\times$ 240 GMs)	Déplacement relatif maximal	$S_a(T_i)$, $S_d(T_i)$, et paramètres de la structure	RA, RF, XGB et ANN	Différence relative, MARD, R^2, MSE
Demertzis *et al.* [78]	Bâtiment en béton armé	5850 observation (90 structures $\times$ 65 GMs)	Déplacement relatif maximal	PGA, PGV, PGD, I_A, SED, CAV, ASI, HI, EPA, $PGV \backslash PGA$, PP, UD, BD, SD et paramètres de la structure	LGB, GBR, RF, KNN, LR, Ridge, DT, ADA, Elastic Net, LASSO	$RMSE$, R^2, MAE, MSE, $MAPE$

Table 1.4 – Bilan des études de ML par simulation(continué)

Étude	Structure	Taille de la base de données	Réponse ciblée	Caractéristiques du signal	Modèles de ML	Critère d'évaluation
Mangalathu *et al.* [61]	Murs de cisaillement	393 observations (3 types de section)	Mode de défaillance	9 paramètres de conception	NB, KNN, DT, RF, ADA, XGB, LGB, CatBoost	Rappel, précision
Mangalathu *et al.* [79]	Colonnes de pont en béton	311 observations (2 types de section)	Mode de défaillance des poteaux	4 paramètres de conception	DA, KNN, DT, RF, NB, ANN	Rappel, précision
Vu *et al.* [80]	Dalle en béton de fibres	83 observations (3 types de section)	Capacité de cisaillement au poinçonnage	6 paramètres de conception	SVMR	$RMSE$, $MAPE$, R^2
Hoang *et al.* [81]	Composants en béton armé	218 observations (3 types de sections)	Résistance ultime à la liaison acier-béton sous corrosion	Propriétés de l'acier et du béton	SVM	$RMSE$, $MAPE$, R^2
Lou *et al.* [82]	Poteau en béton armé	160 observations	Déplacement relatif maximal	12 paramètres liées à la structure	SVM	$RMSE$, $MAPE$, R^2
Gajan [83]	Fondation superficielle	200 observations	Basculement de la fondation	PGA, I_A, type de sol et 3 paramètres de la structure	KNN, SVM, RF	MAE, $MAPE$
Huang *et al.* [84]	Poteau en béton armé	498 observations	Différentes paramètres du modèle hystérésis	7 paramètres de conception, le chargement et le mode de défaillance	ANN	R^2, $RMSE$

Table 1.5 – Bilan des études de ML par expériences

1.4. Apprentissage automatique pour l'évaluation des risques sismiques

1.4.1. Bilan bibliographique

L'utilisation de l'apprentissage automatique pour l'évaluation des risques sismiques des structures a aussi gagné en intérêt, avec des applications plus récentes afin d'évaluer la sécurité de différentes structures, telles que des ponts, des barrages ou des bâtiments.

Dans les études récentes présentées dans le Tableau 1.4 qui utilisent différents modèles de ML pour déterminer la résistance sismique des structures, la principale différence (outre que l'utilisation de différentes méthodes d'apprentissage automatique) réside dans le choix des caractéristiques (features) pour entraîner ces modèles. Bien entendu, si l'on souhaite connaître la réponse d'une structure à une excitation extérieure, par exemple à un signal sismique, il faut bien prendre en compte ce signal. Les mesures d'intensité (Intensity measure, IM) sont utilisées pour décrire les signaux sismiques. Ces mesures d'intensité sismique sont les propriétés des enregistrements sismiques qui reflètent leur capacité à affecter les bâtiments ou autres structures.

Sudret [85] examine les méta-modèles de machine learning de type surface de réponse polynomiale, d'expansion polynomiale du chaos et de krigeage pour construire une étude de la fiabilité structurelle et de la quantification de l'incertitude de la structure. Dans cette application, les caractéristiques sont l'amplitude de la charge appliquée au système et la réponse est le déplacement horizontal du système sous séisme. Il utilise une structure de trois portées, cinq étages avec expansion polynomiale du chaos comme exemple d'application pour montrer l'efficacité des modèles de ML.

Mangalathu *et al.* [79] propose une méthodologie pour la génération de courbes de fragilité spécifiques aux ponts, en utilisant le modèle de forêts aléatoires. Une étude pour analyser la fragilité de deux ponts en Californie, modélisés par OpenSees [86], est réalisée. Cette méthodologie permet d'estimer la fragilité pour un nouvel ensemble de paramètres d'entrée sans simulation coûteuse. Parmi les caractéristiques, $S_a(T = 1.0 \text{ s})$ a également été utilisé comme caractéristique sismiques dans cette étude.

Kiani *et al.* [59] mettent en œuvre des outils de ML basés sur la classification afin de prédire les réponses structurelles et puis déduire les courbes de fragilité d'un bâtiment de huit étages. A cet égard, dix méthodes différentes basées sur la classification sont explorées : régression logistique, régression lasso, machine à vecteur de support, Naïve Bayes, arbre de décision, forêt aléatoire, analyses discriminantes linéaires et quadratiques, réseau de neurones et K-plus proches voisins. Pour l'entraînement du modèle, les IMs suivantes ont été utilisées comme caractéristiques : $S_a(T_i)$ avec des ordonnées sur quinze périodes, l'intensité du spectre en vitesse (SI), l'intensité du spectre en accélération (ASI), l'intensité du spectre en déplacement (DSI), la durée significative de 5 à 75 % et de 5 à 95 % (D_{5-75} et D_{5-95}), l'énergie absolue cumulée (CAV) et PGA. Les auteurs ont effectué une comparaison entre les algorithmes d'apprentissage automatique étudiés et ils ont discuté de l'importance des différentes caractéristiques des mouvements du sol pour prédire les réponses structurelles. Ils trouvent que le modèle RF est le plus efficace, tandis que la caractéristique SI est la plus importante.

Segura *et al.* [72] font une étude sur les techniques pour construire des modèles de ML, afin de proposer une méthodologie pour utiliser ces techniques pour l'évaluation sismique des barrages-poids

pour générer des fonctions de fragilité multivariées. Dans cette étude, ils examinent les techniques de surface de réponse polynomiale (Polynomial Response Surface), construction de la fonction de base adaptative (Adaptive Basis Function Construcion), splines régressives adaptatives multivariées (Multivariate Adaptive Regression Splines), fonctions de base radiales (Radial Basis Functions), machine à vecteurs de support et forêt aléatoire pour la régression. Ils construisent quatorze différents modèles de ML pour prédire le glissement relatif maximal à la base d'un barrage poids en béton à Quebec, Canada. Le choix des caractéristiques comprenait des paramètres de modèle, tels que la résistance du matériau, l'efficacité de drainage et l'amortissement du béton. Plusieurs mesures d'intensité sismique IMs ont également été prises en compte, telles que le spectre de réponse en accélération à la période fondamentale $S_a(T_1^0)$, le spectre de réponse en vitesse à la période fondamentale $S_v(T_1^0)$, la pulsation angulaire du séisme (ω_{eqk}), l'intensité d'Arias (I_A), l'accélération maximale du sol dans la direction verticale PGA_V, l'accélération spectrale à la période fondamentale dans la direction verticale $S_{aV}(T_1^0)$, PGD, PGV, PGA, D_{5-95} et ASI. Le modèle de ML de surface de réponse polynomiale d'ordre 4 est le plus performant dans leur étude et il est utilisé pour générer des surfaces de fragilité de la structure.

Lagaros *et al.* [69] ont proposé une procédure efficace en termes de calcul basée sur le modèle ANN pour l'évaluation rapide de la réponse sismique non-linéaire d'une structure en portique à moment résistant en acier. Différentes mesures d'intensité sismique ont été prises en compte, comprenant le rapport entre la vitesse du sol et l'accélération V/A, l'accélération quadratique moyenne a_{RMS}, la durée totale T_D, l'intensité caractéristique ChI, I_A, D_{5-95}, SI, CAV, $S_v(T_1^0)$, PGA, PGV, PGD et une mesure d'intensité personnalisée proposée par Cordova et al. [87], SaC. Dix-neuf ensembles différents de combinaisons de ces IMs ont également été explorés pour déterminer le choix optimal des caractéristiques pour le modèle ANN. La conclusion est que l'utilisation de la totalité des caractéristiques est plus facile pour entraîner son modèle, mais la sélection des caractéristiques est importante et ce choix a une forte influence sur le modèle final.

Sun *et al.* [68] montent une étude où une analyse de fragilité sismique basée sur l'apprentissage automatique est établie pour évaluer efficacement le risque pour les structures sous des conditions de chargement sismique. Un modèle optimal de réseau de neurones artificiels peut être formé en utilisant des mesures de dommages et d'intensité calculées, une technique qui sera utilisée pour calculer les courbes de fragilité des cadres en acier équipés de contreventements à enrobage limité au lieu d'utiliser la simulation non-linéaire par éléments finis. Sept mesures d'intensité, contenants PGA, PGV, PGD, $S_a(T_1)$, ASA, CAV et IA, ont été considérées comme des caractéristiques. Les résultats numériques montrent qu'une évaluation de la probabilité de défaillance instantanée hautement efficace peut être réalisée avec le cadre proposé pour des structures de bâtiments à grande échelle et réalistes.

Wang *et al.* [60] ont proposé une méthodologie pour le calcul des courbes de fragilité des équipements des centrales nucléaires, basée sur un réseau de neurones ANN. La sélection des IMs les plus pertinents en tant que caractéristiques pour le modèle ANN a été réalisée avec une approche de sélection progressive. Parmi les huit mesures d'intensité différentes considérées, comprenant accélération pseudo-spectrale à la première période fondamentale $PS_a(f_1^0)$, accélération spectrale moyenne (ASA), la période prédominante (T_P) et PGA, PGV, PGD, IA, CAV. Seulement les deux mesures les plus efficaces, ASA et IA, ont été sélectionnées.

Zhou *et al.* [70] évaluent la fragilité sismique d'un barrage en enrochement à haute teneur en

béton en mettant en oeuvre un modèle de ML de machine à vecteur de support pour prédire trois réponses différentes du barrage. Dans cet article, 100 mouvements synthétiques du sol sont générés pour obtenir les mesures d'intensité correspondantes. Trois réponses sont considérées, y compris le déplacement vertical, la déformation plastique de cisaillement, ainsi que l'indice de dommage de la dalle de façade. Ensuite, les modèles SVM correspondant aux trois états d'endommagement sont formés en fonction des IMs, contenant PGA, CAV, I_A et T_P. Enfin, les courbes de fragilité des différents états limites basés sur les trois états sont construites.

Rezaei *et al.* [73] utilisent une analyse de régression symbolique pour développer des modèles de prédiction de la réponse sismique et des courbes de fragilité pour les ponts à poutres-caissons en béton. Les incertitudes géométriques, matérielles, structurelles et de mouvement du sol ont été prises en compte pour améliorer la fiabilité des modèles mathématiques dérivés. Les réponses ont été enregistrées en tant que paramètres de demande d'ingénierie par une analyse non-linéaire en temps des ponts simulés et elles ont été utilisées comme cibles pour les algorithmes de prédiction. Plusieurs paramètres structurels des ponts ainsi que le spectre de réponse en accélération à une seconde, $S_a(T = 1.0\ \mathrm{s})$, ont été pris en compte comme caractéristiques.

1.4.2. Construction du modèle d'apprentissage automatique pour l'évaluation des risques sismiques des structures

Les modèles de machine learning sont construits à partir des caractéristiques $\mathbf{x}$, qui contiennent souvent les paramètres des structures ainsi que les mesures d'intensité. Si on ne considère que les IMs comme caractéristiques, l'équation (1.6) devient :

$$\mathbf{Y} = \mathbf{ML}(\mathbf{x}) + \epsilon, \qquad \mathbf{x} = \{\mathrm{IM}_1, \mathrm{IM}_2, \ldots, \mathrm{IM}_n\} \tag{1.22}$$

Ces IMs sont utilisées pour décrire les signaux sismiques pour chaque observation afin d'entrainer les modèles d'apprentissage automatique. La réponse $\mathbf{Y}$ est la réponse des structures sous excitation sismique, qui permet de définir la défaillance dans l'équation (1.1).

De plus, une procédure de référence pour l'étude de la fragilité sismique des structures est illustrée dans la Figure 1.15. Cette procédure implique généralement plusieurs étapes :

- **Collecte des données** : Réunir des enregistrements sismiques contenant des informations sur les structures touchées par des séismes. Ces enregistrements peuvent inclure diverses caractéristiques telles que les propriétés des mouvements du sol ou les mesures d'intensité, les paramètres sismiques, les types de bâtiments, etc. .
- **Ingénierie des caractéristiques** : Sélectionner et préparer les caractéristiques pertinentes pour la modélisation. Cela peut impliquer des analyses statistiques pour choisir les caractéristiques les plus significatives et leur transformation pour l'entraînement du modèle. Ces ensembles de caractéristiques sont notées $\mathbf{x}$, et la réponse de la structure est notée $\mathbf{Y}$.
- **Choix des modèles** : Identifier les modèles d'apprentissage automatique adaptés pour construire la courbe de fragilité. Ces modèles peuvent être des réseaux de neurones, des méthodes de forêt aléatoire, des machines à vecteurs de support, etc. .
- **Entraînement des modèles** : Utiliser les données collectées pour entraîner les modèles sélectionnés.

- **Validation du modèle de ML** : Évaluer les performances des modèles en utilisant des métriques telles que r, R^2, $RMSE$, ou d'autres métriques appropriées. Ces métriques aident à mesurer la précision et l'ajustement des modèles entrainés.
- **Utilisation du modèle** : Cela implique l'utilisation du modèle avec des nouveaux signaux sismiques pour prédire la réponse de la structure, puis réaliser une étude d'évaluation des risques sismiques en général ou de la fragilité des structures en particulier.

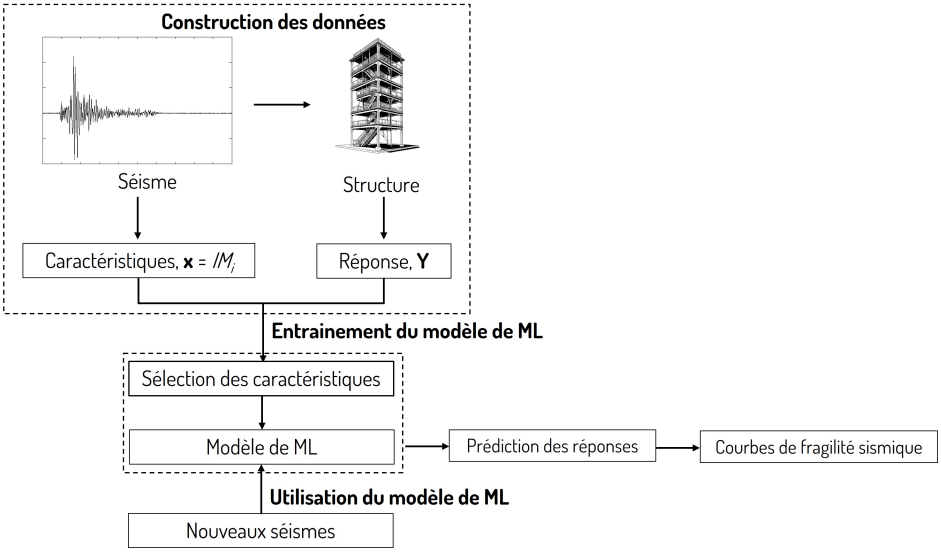


Figure 1.15 – Procédure de construction de la courbe de fragilité basé sur l'apprentissage automatique

1.4.3. Conclusion

Dans le contexte de l'étude de la fragilité sismique des structures, l'application de techniques de machine learning apparaît comme une solution prometteuse pour rendre le processus plus efficace et économique par rapport à la méthode de Monte Carlo, qui est coûteuse en termes de temps de calcul. De plus, l'utilisation de modèles de machine learning peut permettre le développement de modèles de substitution rapides et précis reproduisant le comportement dynamique non-linéaire des structures, sans nécessiter de nouvelles simulations mécaniques coûteuses. L'application de l'apprentissage automatique pour construire des courbes de fragilité sismique réduit considérablement le temps de calcul tout en maintenant un niveau de précision acceptable. La Figure 1.15 présente la procédure de référence pour la construction des modèles de ML destinés à la construction des courbes de fragilité sismiques.

Cependant, il est à noter que le succès de cette approche rencontre encore un défi majeur. Celui-ci est lié au choix des caractéristiques $\mathbf{x}$ pour entrainer les modèles de machine learning. Le nombre de caractéristiques sélectionnées est très variable, d'une seule jusqu'à une dizaine. Cela pose la question sur le choix et le nombre optimal de caractéristiques pour entrainer des modèles. Selon les références, le choix des caractéristiques $\mathbf{x}$ liées à l'enregistrement sismique à partir d'IMs est important.

Une IM est une la représentation d'une propriété différente d'un séisme donc le nombre d'IMs est large. Devant un très large choix, parfois contradictoire des caractéristiques des études existantes pour les séismes, il est donc nécessaire de proposer des procédures pratiques avec les caractéristiques les plus pertinentes et faciles à mise en œuvre. Le choix des caractéristiques adaptées est aussi une motivation pour ce travail de thèse. La section 1.5 détaille plus sur la variabilité des mesures d'intensité, qui montre la difficulté de les utiliser pour entrainer les modèles de machine learning à l'état actuel. Une proposition pour le choix d'une IM pertinente est donc présentée.

1.5. Caractérisation des signaux sismiques

1.5.1. Mesure d'intensité sismique

Comme présenté dans la section 1.4, les auteurs utilisent les mesures d'intensité pour décrire leurs signaux sismiques utilisés. Ces mesures d'intensité sismique représentent les propriétés du mouvement du sol qui reflètent leur capacité à affecter les structures. Il existe de nombreuses mesures d'intensité sismique qui peuvent être obtenues du mouvement du sol. Selon Hariri-Ardebili *et al.* [88], il existe environ soixante-dix IMs différentes. Elles peuvent être divisées en sept catégories telles que les IMs non échelonnables (magnitude, durée du séisme, etc.), les IMs scalaire dépendants du mouvement du sol (PGA, PGV, PGD, etc.), les IMs composés dépendants du mouvement du sol (V/A, ChI), les IMs spectrales indépendantes de la structure (Intensité du spectre en accélération ASI, en vitesse VSI ou en déplacement DSI, etc.), les IMs spectrales dépendantes de la structure ($S_a(T)$,$S_a(T)$,$S_a(T)$ pour une valeur d'amortissement ξ, etc.), les IMs de type vectoriel (intensité de Baker [89], intensité de Bojorquez [90], etc.) et les IMs pour les mouvements du sol à composantes multiples (valeur moyenne des IMs de chaque composante). De Biasio [91] a étudié les IMs du mouvement du sol pour l'analyse probabiliste de risque sismique et il a divisé ces IMs en trois catégories, à savoir : les IMs basées sur le pic, les IMs basées sur la durée et les IMs basées sur la réponse en fréquence. Avec ces catégories définies, plusieurs groupes de mesures d'intensité peuvent être définis :

- Les IMs basées sur les pics : ce groupe contient les mesures d'intensité qui représentent les valeurs absolues maximales des caractéristiques de mouvement du sol au fil du temps. PGA, PGV et PGD sont les IM les plus fréquemment utilisés de ce groupe.
- Les IMs basées sur la durée : ce groupe contient les mesures d'intensité qui sont l'intégration d'une caractéristique de mouvement du sol sur la durée du signal. Les IMs fréquemment utilisés dans ce groupe sont l'intensité d'Arias I_A, vitesse absolue cumulée CAV par exemple.
- Les IMs basées sur la réponse en fréquence : ce groupe contient les mesures d'intensité qui sont basées sur la réponse des oscillateurs élastiques au mouvement du sol, comprenant les spectres de réponse $S_a(T)$, $S_v(T)$ et $S_d(T)$, les spectres pseudo $PS_v(T)$ et $PS_a(T)$. En outre, les IMs comme l'intensité spectrale d'accélération ASI, l'accélération spectrale moyenne relative ASA sont également fréquemment utilisées.

Le Tableau 1.6 présente les IMs les plus courantes de chacune de ces classes. Les mesures d'intensité sélectionnées sont utilisées comme caractéristiques $\mathbf{x}$.

Cette liste de mesures d'intensité n'est pas exhaustive. Il donc convient de se poser la question comment choisir les caractéristiques afin de pouvoir bien entraîner les modèles de machine learning.

L'un des candidats les plus prometteurs est le spectre de réponse, une mesure d'intensité sismique qui contient non seulement des informations du signal sismique mais aussi celles de la structure.

Types	**Classification**	**Définition**	**Note**
Basées sur les pics	Dépendantes du séisme	$PGA = \max_t \lvert a(t) \rvert$	$\lvert a(t) \rvert$: accélération temporelle du sol
		$PGV = \max_t \lvert v(t) \rvert$	$\lvert v(t) \rvert$: vitesse temporelle du sol
		$PGD = \max_t \lvert u(t) \rvert$	$\lvert d(t) \rvert$: déplacement temporel du sol
Basées sur la durée	Dépendantes du séisme	$I_A = \frac{\pi}{2g} \int_0^{t_f} a(t)^2 dt$	t_f : durée totale du mouvement du sol
		$CAV = \int_0^{t_f} \lvert a(t) \rvert dt$	
		$E_{cum} = \int_0^{t_f} a(t)^2 dt$	
Basées sur la fréquence	Dépendantes du séisme et la structure	$S_a(T,\xi), S_v(T,\xi), S_d(T,\xi)$	T : période, ξ : taux d'amortissement
		$PS_a(T) = \omega_n^2 S_d$	ω_n : pulsation naturelle
		$PS_v(T) = \omega_n S_d$	
		$SI = \int_{0.1}^{2.5} S_v(T_1^0) dT$	T_1^0 : période fondamentale
		$ASI = \int_{0.1}^{0.5} PS_a(T_1^0) dT$	
		$ASA = \frac{1}{f_1^0(1-X_f)} \int_{X_f f_1^0}^{f_1^0} PS_a(f) df$	f_1^0 : fréquence fondamentale

Table 1.6 – Les IMs fréquemment utilisées

1.5.2. Spectre de réponse - Une mesure importante

Le concept de spectre de réponse a été utilisé pour la première fois par Biot [92] et popularisé par Housner [93] pour son utilisation en génie parasismique. L'importance de ce concept vient du fait que, lors d'un dimensionnement, on ne s'intéresse surtout qu'à la valeur maximum de la réponse d'une structure au séisme [94]. Il est obtenu en calculant la valeur maximale de la réponse de la structure (en accélération, vitesse et déplacement) pour un séisme spécifique, un coefficient d'amortissement et une large gamme de fréquences d'un oscillateur.

Pour une réponse quelconque de la structure Y, on a

$$\mathbf{Y} = \max_t \left(\lvert Y(t) \rvert \right) \tag{1.23}$$

où l'indice de max désigne la valeur maximale de la réponse dans le temps. Les spectres sont définis par :

$$\text{Spectre de réponse en déplacement relatif} \qquad S_d(\omega,\xi) = \max_t |u(t,\omega,\xi)| \tag{1.24}$$

$$\text{Spectre de réponse en vitesse relative} \qquad S_v(\omega,\xi) = \max_t |\dot{u}(t,\omega,\xi)| \tag{1.25}$$

$$\text{Spectre de réponse en accélération totale} \qquad S_a(\omega,\xi) = \max_t |\ddot{u}(t,\omega,\xi)| \tag{1.26}$$

où $u(t)$ est le déplacement d'un oscillateur linéaire.

La Figure 1.16 montre un exemple de construction d'un spectre de réponse en accélération pour l'enregistrement sismique d'El Centro.

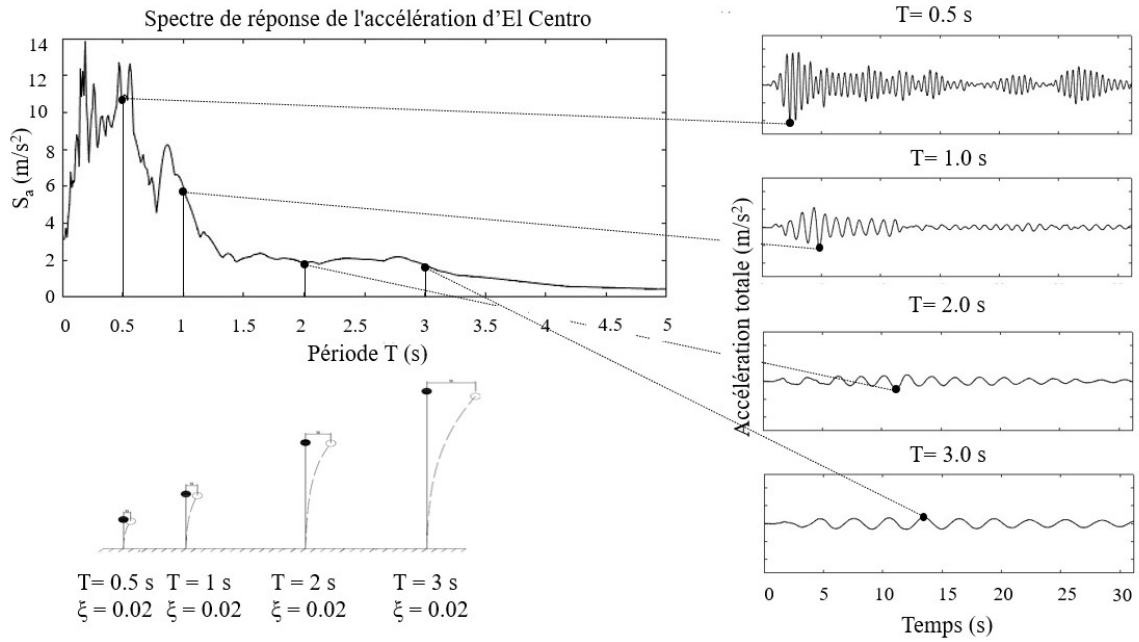


Figure 1.16 – Spectre de réponse d'accélération totale pour le séisme d'El Centro

En génie civil, pour un système à un degré de liberté avec de faibles valeurs d'amortissement, étant donné que les pulsations non amorties et amorties sont proches, c'est-à-dire $\omega \cong \omega_d$, on observe que :

$$S_a \cong \omega_d^2 S_d \cong \omega^2 S_d \tag{1.27}$$

où S_a et S_d sont le spectre de réponse en accélération et déplacement respectivement. L'équation (1.27) montre la relation linéaire entre le spectre de réponse en accélération à la période naturelle et le déplacement du système. Un concept similaire de spectre de réponse peut être trouvé dans le domaine aérospatial lorsque l'excitation est sous forme d'un choc. Il s'agit donc de spectre de réponse au choc (Shock Response Spectrum, SRS) qui indique la réponse maximal des oscillateurs pour un type de choc spécifique.

Par conséquent, le spectre d'accélération en réponse à une période égale à la période naturelle du système peut être une caractéristique importante pour prédire le déplacement maximal du système de 1ddl. De même, le spectre d'accélération en réponse aux périodes égales aux périodes naturelles du système à plusieurs degrés de liberté sont des caractéristiques importantes pour prédire son déplacement. Donc, cette mesure d'intensité montre une potentialité d'être utilisée comme la caractéristique $\mathbf{x}$ pour entrainer des modèles de machine learning.

1.5.3. Potentialité de l'utilisation des spectres de réponse

Les références citées dans la section 1.4 exposent diverses utilisations de modèles de ML pour évaluer des risques sismiques des structures. Bien que les réponses des systèmes soient similaires, les auteurs sélectionnent les caractéristiques $\mathbf{x}$ de manière différente à partir des mesures d'intensités sismiques. Dans ces études, les caractéristiques visent à représenter et/ou décrire l'enregistrement sismique dans les modèles de ML. Devant une pluralité d'IMs, le choix des caractéristiques pour élaborer des modèles d'apprentissage automatique varie d'une seule mesure à un ensemble de plusieurs dizaines de mesures. Ces choix des caractéristiques, basés sur différentes combinaisons de mesures d'intensité, sont-ils admissibles ou optimaux pour l'entraînement des modèles de ML? La question du choix des caractéristiques $\mathbf{x}$ pour représenter les signaux sismiques demeure complexe et l'efficacité de ces choix suscite également des interrogations majeures. Cela soulève la question de savoir comment choisir la caractéristique la plus efficace pour entraîner un modèle de ML.

D'autre part, les spectres de réponse en déplacement, en vitesse et en accélération contiennent des informations précieuses de l'enregistrement sismique ainsi que celles de la structure. Par conséquent, ils peuvent être utilisés comme caractéristiques d'entrée $\mathbf{x}$ pour entraîner des modèles d'apprentissage automatique.

Cette potentialité est démontrée par certaines études. Kim *et al.* [95] ont entraîné des réseaux de neurones artificiels pour prédire la réponse d'un système non-linéaire hystérésis à un degré de liberté. Une combinaison de mesures d'intensité des mouvements du sol, comprenant PGA, PGV et PGD, ainsi que les accélérations spectrales amorties à 5 % dans l'intervalle de période de 0.005 s à 10 s ($S_a(T_i)$ avec $i = 1, 2, ..., 110$), a été utilisée comme caractéristiques $\mathbf{x}$. Noureldin *et al.* [74] ont utilisé plus de 60000 analyses non-linéaires comme jeu de données pour entraîner un modèle ANN afin de prédire la réponse d'une structure à un degré de liberté.Ce modèle a pris en compte le $S_a(T_i)$ avec la période d'échantillonage T_i jusqu'à 5 s comme caractéristiques principales. Cependant, aucune stratégie de construction de ces caractéristiques n'est détaillée.

Bien que ces travaux montrent la possibilité d'utiliser le spectre de réponse comme caractéristiques pour former un modèle d'apprentissage automatique, malheureusement, la stratégie de sélection des périodes et des limites d'intervalle du spectre de réponse pour construire l'ensemble $\mathbf{x}$ optimal n'est pas spécifiée. Cela soulève la nécessité de mener une étude sur la construction des caractéristiques $\mathbf{x}$ en tenant compte des valeurs du spectre de réponse.

1.6. Mouvement du sol

La réponse des structures aux séismes dépend à la fois des caractéristiques des structures et des mouvements du sol. Ainsi, le mouvement du sol est un composant important pour les études parasismiques. Dans cette étude, la structure est analysée en utilisant les simulations par éléments finis. Les réponses dynamiques sont obtenues par l'intégration directe en temps puisque le comportement des structures est souvent non-linéaire sous excitation sismique.

Un enregistrement décrit comment l'accélération du sol fluctue autour de zéro tout au long du séisme. Un enregistrement peut durer de quelques secondes à quelques dizaines de secondes. La Figure 1.17 présente un exemple d'un enregistrement sismique de trois composants, enregistré à la station Sous Sol BRGM Marseille, par le Réseau Accélérometrique Permanent (French Accelerometrique Network, RAP) en 2012.

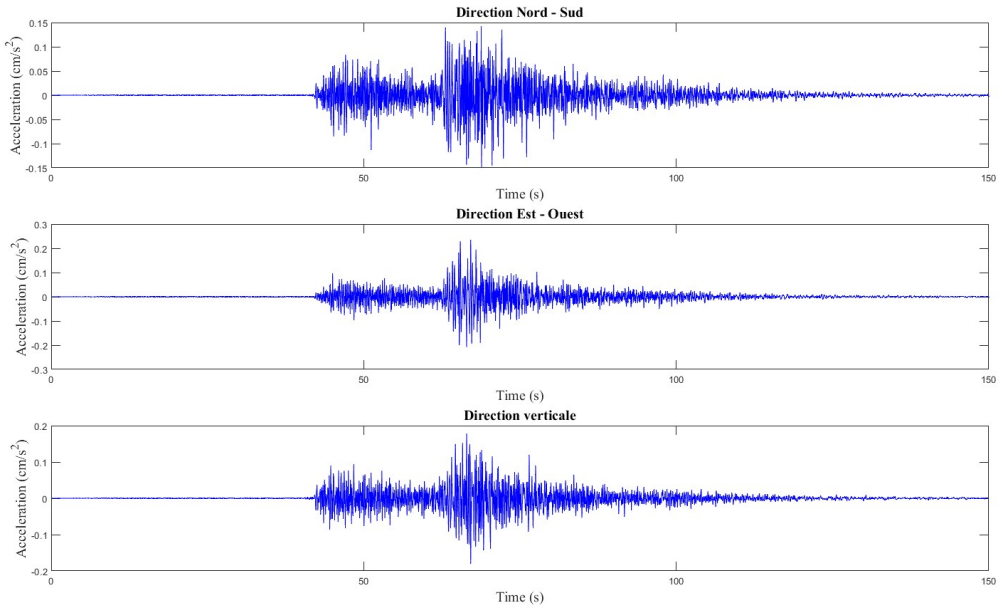


Figure 1.17 – Exemple d'un accélérogramme d'enregistrements réels avec trois composants

Un enregistrement sismique peut être obtenu à partir de deux sources différentes : réel par enregistrement ou synthétique par un générateur sismique. Chaque source d'enregistrement sismique présente des avantages et inconvénients, qui sont présentés dans la suite.

1.6.1. Enregistrement sismique réel

Des stations de mesures partout dans le monde enregistrent les excitations sismiques réelles. Souvent, ces enregistrements sont présentés sous forme temporelle. Ils sont disponibles dans les zones où il y a des stations de mesure et ils reflètent exactement la nature d'un séisme des sites de mesure. Des bases de données sont créées afin de partager ces enregistrements.

Par exemple, le Centre de recherche en génie parasismique du Pacifique (Pacific Earthquake Engineering Research Center) [96] constitue la plus grande base de données de mouvements du sol enregistrés et traités, Next Generation Attenuation Relationships for Western US Database (NGA-West2) [65]. La base de données NGA-West est spécifique à la région de l'ouest des États-Unis. Elle a été développée pour fournir des informations précieuses aux ingénieurs et chercheurs travaillant sur des projets parasismiques dans cette région. Elle contient des enregistrements de mouvements forts causés par des séismes. La base de données NGA-West est utilisée pour développer des modèles d'atténuation sismique spécifiques adaptés pour le site considéré. Ces modèles sont essentiels pour évaluer les risques sismiques et concevoir des infrastructures résistantes aux séismes. Sur cette base de données, on peut également trouver les informations associées, telles que : PGA, PGV, S_a, magnitude, distance à la source, etc. . Les valeurs comme PGA, PGV, etc. , sont les mesures d'intensité attributs d'un séisme. Segura *et al.* [72] ont sélectionné 250 enregistrements sismiques à partir de la base de données NGA-West2. Utilisant la même base, Kiani *et al.* [59] ont sélectionné 2000 enregistrements, contenant 100 enregistremens de séisme pour 20 niveaux de risque.

Un autre exemple, Engineering Strong-Motion Database (ESM) [97] est aussi une base de données des enregistrements. ESM a été spécifiquement conçu pour fournir aux utilisateurs finaux uniquement des données de mouvements sismiques de qualité vérifiée, uniformément traitées, ainsi que des paramètres pertinents, et ce depuis 1969 dans la région euro-méditerranéenne. La base de données a été conçue pour un large ensemble de parties prenantes (sismologues experts, ingénieurs en génie sismique, étudiants et professionnels) avec une interface Web conviviale et intuitive. La base de données ESM contient 23 000 formes d'onde tri-composantes. Environ 60 % d'entre elles (13 191) sont traitées manuellement, 15 % sont traitées automatiquement et nécessitent une révision manuelle, et 25 % sont considérées de mauvaise qualité. Les données de mauvaise qualité sont conservées dans la base de données, car des mesures d'intensité telles que les valeurs de PGA peuvent être utilisées en toute confiance. Les formes d'onde traitées manuellement proviennent de 1929 événements sismiques dont la magnitude $M_w \geq 4.0$ et ont été enregistrées par 1901 sites exploités par 38 réseaux. La base de données EMS est une ressource majeure pour les ingénieurs civils, les chercheurs en génie parasismique et les professionnels de la construction. Elle contient des enregistrements de mouvements forts générés par des séismes à travers le monde. Ces enregistrements sont essentiels pour évaluer comment les bâtiments et les infrastructures réagissent aux séismes réels. La base de données EMS est utilisée pour étudier les effets des séismes sur les structures, développer des normes parasismiques, et concevoir des bâtiments et des infrastructures résistants aux séismes. Les données dans la base de données EMS proviennent de différentes sources, notamment des réseaux de capteurs sismiques, des instruments installés dans des bâtiments, et des enregistrements de séismes passés. Cette base de données est utilisée par Zentner *et al.* [98] par exemple.

Ces deux bases de données sont des ressources précieuses pour la compréhension des séismes, l'ingénierie parasismique et la conception de bâtiments résistants aux séismes, en particulier dans des régions à risque sismique élevé.

Cependant, malgré les efforts déployés pour enrichir les bases de données, le nombre d'enregistrements disponible et fourni aux utilisateurs reste limité. Lorsque de nombreux enregistrements sont nécessaires, ils doivent provenir de différentes stations sismiques. Toutefois, ces séismes sont souvent hétérogènes et présentent des aléas associés à la région où les données sont collectées. Cela

rend les études paramétriques sur l'influence des paramètres plus complexes. C'est la raison par laquelle un générateur sismique artificielle est aussi nécessaire pour pouvoir faire ces études pour les régions à faible activité sismique.

1.6.2. Enregistrement sismique synthétique

Afin de surmonter le manque des enregistrements sismiques réels, le concept des enregistrements synthétiques à partir d'un générateur est introduit. Un générateur sismique est utile lorsqu'une population importante de séismes est nécessaire, ou dans le cas des régions étudiés qui ont des activités sismiques faibles, donc les enregistrements sont peu nombreux. Un générateur sismique peut produire une quantité illimitée d'enregistrements synthétiques dont les propriétés (spectre, intensité) sont similaires aux séismes réels.

Il existe plusieurs modèles développés par divers chercheurs tels que Clough et Penzien [99], Rodolfo [100], ou Rezaeian [101] qui visent à générer des accélérations sismiques basées sur la densité spectrale de puissance. Ces modèles caractérisent le séisme par une fonction d'enveloppe déterministe et un processus aléatoire Gaussien stationnaire dont la densité dépend de plusieurs facteurs tels que la source sismique, le chemin de la source au site, l'effet du site, etc. .

Parmi ces modèles, celui de Clough et Penzien [99], par exemple, est largement utilisé pour l'analyse de la réponse sismique stochastique des structures. Ces modèles sont souvent intégrés à des représentations spectrales [102], des décompositions de Karhunen-Loève [103] ou des décompositions orthogonales pour représenter explicitement le mouvement du sol sous forme de fonctions aléatoires. Cependant, ces modèles, bien qu'utiles, restent des approximations phénoménologiques [104].

Certains chercheurs, comme Hwang [105] et Shinozuka *et al.* [8, 11], ont développé des générateurs artificiels basés sur des scénarios sismiques pour l'évaluation des risques sismiques des structures. La série MCEER [106] comprend 100 accélérations synthétiques générées depuis un modèle de barrière spécifique. Ces séries sont réparties en quatre groupes selon des périodes de retour spécifiques.

L'impact de la qualité du générateur sismique sur l'exactitude des résultats de l'analyse est indéniable. Dans les exemples d'application de cette thèse, le modèle de mouvement du sol élaboré par Boore [107] a été choisi pour simuler les séismes. Une implémentation de ce modèle a été réalisée sous MATLAB. Les détails du modèle sont présentés dans l'annexe A. La Figure 1.18 illustre une simulation d'un séisme de magnitude $M = 7$ à une distance de $R = 9$ km de l'épicentre. Ces signaux synthétiques sont des mouvements du sols provoqués par les ondes de cisaillement. Il faut aussi noter que les signaux sont les mouvements du sols dans n'importe quelle direction. Dans le cadre de la thèse, les signaux sismiques sont appliqués dans la direction horizontale pour l'exemple des bâtiments.

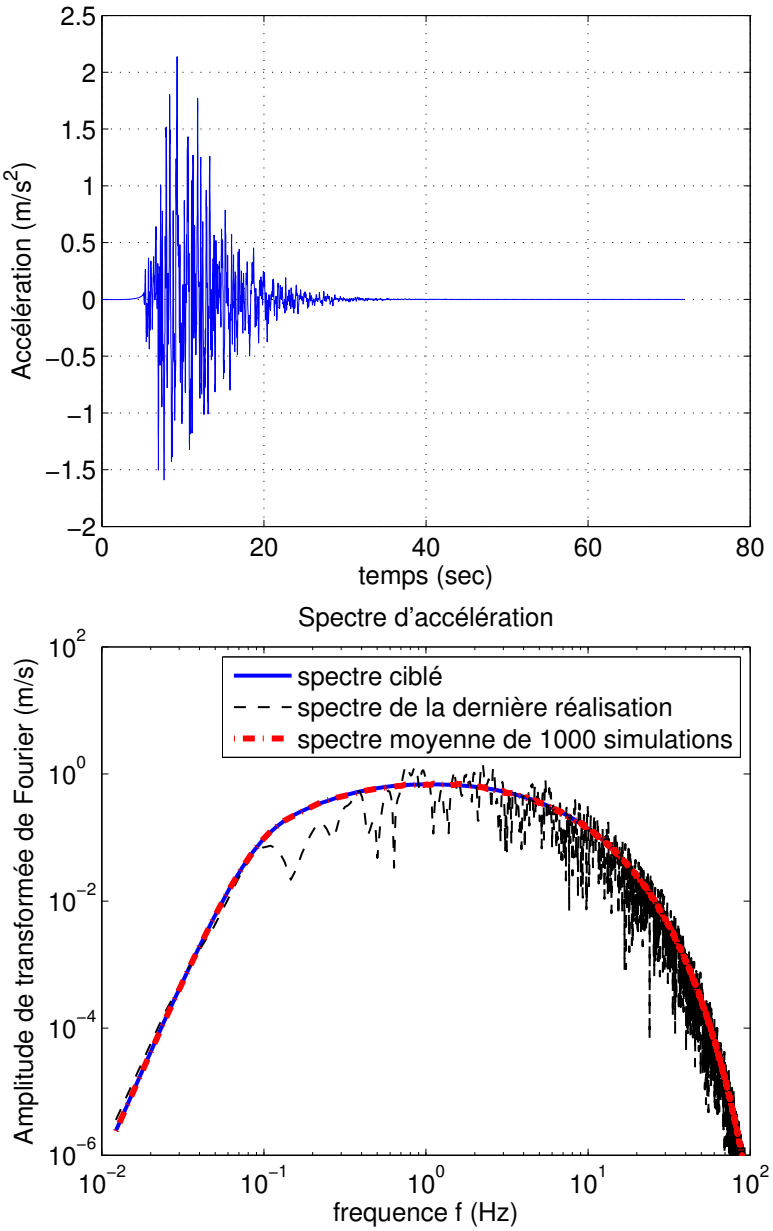


Figure 1.18 – Une simulation de séisme par le modèle de Boore

1.6.3. Résumé sur les signaux sismiques

Les enregistrements sismiques réels sont limités en quantité et en disponibilité, ce qui peut poser des défis pour les études sismiques. Lorsqu'une quantité significative d'enregistrements est nécessaire, il est souvent requis d'utiliser des données provenant de différentes stations sismiques. Cependant, ces enregistrements ne proviennent pas tous de la même origine sismique et ils sont sujets à des variations en fonction des régions d'où ils sont collectés. Cette disparité rend difficile les études

paramétriques concernant l'effet des paramètres sismiques.

Ainsi, en raison de la rareté et de la variabilité des enregistrements réels, il est impératif d'étendre ces observations à des régions à faible activité sismique, en utilisant des générateurs d'excitations sismiques artificielles. Ces générateurs permettent de simuler des séismes possédant des caractéristiques équivalentes à ceux observés, ce qui est particulièrement crucial lorsque de nombreuses données sismiques sont requises ou lorsque les régions étudiées ont une faible activité sismique, conduisant à un nombre limité d'enregistrements réels. En utilisant les enregistrements sismiques synthétiques, le problème de pénurie d'informations est contourné, car ils ont la capacité de produire une quantité illimitée de séismes artificiels partageant des propriétés similaires aux séismes réels, notamment en termes de spectre et d'intensité.

Par rapport à l'apprentissage automatique, un nombre important d'observations d'un phénomène est crucial pour son entrainement. Les enregistrements synthétiques sont donc plus adaptés pour l'apprentissage automatique, grâce à une disponibilité illimitée et une similarité entre les observations. Par contre, l'utilisation des enregistrements réels est inévitable. Les modèles de ML formés uniquement sur des données synthétiques pourraient être biaisés par les caractéristiques spécifiques de ces enregistrements. La vérification avec les données réelles diversifie l'ensemble de données, améliorant la capacité du modèle à traiter à des situations inédites. C'est la raison pour laquelle dans nombreuses études, un ensemble de structures est étudié en accompagnement d'un ensemble de centaines enregistrements réels. Cette utilisation permet d'obtenir plus d'observations sur l'influences des séismes aux structures. De plus, les données réelles peuvent représenter des scénarios extrêmes ou des événements rares, desquels il peut être difficile d'obtenir des enregistrements synthétiques générés à partir des modèles mathématiques. L'utilisation des enregistrements adaptés pour chaque étude aide à trouver des résultats souhaités.

Les avantages et inconvénients de l'utilisation des signaux sont résumées dans le Tableau 1.7.

Signal	**Avantages**	**Avantages**
Réel	- Représentation fidèle des événements sismiques réels	- Disponibilité limitée
	- Disponibles dans des régions dotées de stations de mesure	- Nombre restreint de stations et des zones sismiquement actives
Synthétique	- Pallie le manque de données réelles	- Ne reflète pas exactement la complexité des séismes réels
	- Simule des séismes ayant des propriétés équivalentes	- Dépend de la précision du générateur

Table 1.7 – Avantages et inconvénients des séismes

1.7. Objectifs et organisation de la thèse

1.7.1. Objectifs de la thèse

Les analyses bibliographiques des sections précédentes relèvent des challenges clés pour l'application de l'apprentissage automatique pour les études d'évaluation des risques sismiques des structures. Premièrement, l'apprentissage automatique montre sa potentialité pour améliorer ces études, en diminuant le coût de calculs et des analyses numériques. Par contre, cette potentialité ne se présente que dans les études récentes. L'application de l'apprentissage automatique n'est pas encore bien formée. Cela est présenté par le choix très différent des caractéristiques à entrainer les modèles de ML, qui donnera une difficulté pour reproduire les autres études dans le même contexte. Il faut trouver un ensemble de caractéristiques qui soit pertinent et disponible pour tous les enregistrements sismiques observés. Deuxièmement, comme le choix des caractéristiques est encore différent, la procédure pour appliquer les modèles de ML n'est pas encore efficace. Cela demande une/des procédures concrètes pour pouvoir utiliser l'apprentissage automatique de façon plus efficace. Une fois ces procédures sont proposées, il faut avoir des études pour pouvoir les valider. Les objectifs principaux de cette thèse sont de trouver des résolutions pour ces challenges :

- *Proposer une nouvelle méthode d'application des modèles de l'apprentissage automatique en prenant les spectres de réponse en accélération comme caractéristiques.*
  Avec de nombreuses options pour le choix des caractéristiques selon les mesures d'intensité disponibles présentées dans la section 1.5, les références explorent l'utilisation de ces caractéristiques pour arriver à entraîner les modèles d'apprentissage automatique et à prédire la réponse structurale. Cependant, ils ne discutent pas l'impact du choix des caractéristiques sur l'obtention d'un résultat. Le choix des caractéristiques pour entrainer des modèles d'apprentissage automatique de façon efficace reste à être approfondi. De plus, la question du choix qualitatif des caractéristiques et de la taille optimale de l'ensemble de données d'entraînement d'un modèle d'apprentissage automatique reste sans réponse. La thèse propose d'utiliser le spectre de réponse en accélération pour l'entraînement des modèles. Plus précisément, en se basant sur les valeurs du spectre de réponse en accélération, un ensemble efficace de caractéristiques $\mathbf{x}$ est construit en utilisant le spectre de réponse en accélération échantillonné à différentes périodes pour entraîner un modèle d'apprentissage automatique. Cette proposition sera d'abord explorée, puis testée avec des structures à comportement linéaire et non-linéaire. Ensuite, pour chaque cas, une étape supplémentaire sera appliquée pour sélectionner les caractéristiques les plus impactantes pour entraîner le modèle de machine learning. Cette étape supplémentaire est nécessaire non seulement pour améliorer l'efficacité du modèle d'apprentissage automatique, mais aussi pour fournir une perception sur les caractéristiques à utiliser pour l'entraînement du modèle. Ces résultats seront finalement utilisés pour déduire des procédures pratiques pour l'application de l'apprentissage automatique afin d'évaluer les risques sismiques des structures. Enfin, des procédures d'application de l'apprentissage automatique seront proposées pour les systèmes linéaires et non-linéaires, afin de construire des modèles de ML de manière plus efficace.
- *Améliorer la construction des courbes de fragilité sismique par l'apprentissage automatique.*

Le deuxième but de la thèse est d'améliorer la méthode existante pour évaluer des risques sismiques des structures, plus particulièrement pour construire la courbe de fragilité sismique en utilisant les modèles d'apprentissage automatique. Selon la section 1.2, le nombre d'observations important est la clé pour construire la courbe de fragilité. Pour atteindre un nombre suffisamment grand, la thèse vise à utiliser l'apprentissage automatique pour éviter d'avoir à exécuter un grand nombre de simulations par éléments finis pour construire la courbe de fragilité. Avec les procédures pratiques proposées, la thèse présente la possibilité d'améliorer la construction de la courbe de fragilité par l'apprentissage automatique.

- *Vérifier la proposition de la nouvelle méthode par des enregistrements réels*
  La thèse propose une procédure basée sur l'apprentissage automatique en utilisant les valeurs du spectre de réponse en accélération échantillonnées aux différentes périodes. Les premières applications utilisent les signaux sismiques synthétiques, puisque les générateurs sismiques permettent d'obtenir un nombre illimité d'observations. Par contre, dans la réalité, des études de la fragilité sismique demandent l'utilisation des enregistrements réels. Une vérification en utilisant des enregistrements réels est donc nécessaire. Une étude est mise en place pour examiner la compatibilité de la proposition de la thèse avec ces enregistrements.
- *Vérifier la proposition de la nouvelle méthode par des bases de données existantes*
  Finalement, les procédures proposées sont vérifiées par l'utilisation des bases de données disponibles dans la littérature. Cette validation est indispensable pour montrer le caractère universel de la proposition de la méthode proposée dans le cadre de la thèse.

Ces objectifs sont les contenus des chapitres suivants. Ils sont présentés selon une organisation décrite dans la section suivante.

1.7.2. Organisation de la thèse

La thèse est organisée en fonction des objectifs majeurs identifiés. Le schéma présenté dans la Figure 1.19 illustre l'organisation globale de la thèse, découpée en cinq chapitres distincts. Chaque chapitre est défini selon son contenu principal, structurant ainsi le travail de recherche. Le premier chapitre, "Synthèse bibliographique et problématique générale", expose le contexte de la thèse en présentant une revue de la littérature, la méthodologie adoptée, ainsi que les objectifs principaux de l'étude. Les chapitres suivants, allant du deuxième au cinquième, abordent des aspects spécifiques de l'étude. Le deuxième chapitre se concentre sur l'utilisation du spectre de réponse pour les structures linéaires et l'application des modèles d'apprentissage automatique sur celles-ci. Les structures linéaires sont choisies parce qu'elles sont simples et la charge demandée pour obtenir des observations dans ce cas n'est pas excessive. Cependant, il est important de noter que dans des conditions extrêmes, la linéarité peut ne plus être respectée, nécessitant ainsi l'utilisation de modèles non-linéaires pour une analyse plus précise et réaliste. Les éléments ou systèmes structuraux non-linéaires sont souvent rencontrés dans des cas tels que la plasticité des matériaux, les grandes déformations, les sollicitations dynamiques extrêmes comme les séismes, ou lorsque des matériaux tels que le béton, l'acier ou les assemblages subissent des dégradations. Donc, le troisième chapitre considère l'utilisation du spectre de réponse en accélération pour entrainer les modèles de ML pour les structures non-linéaires, explorant leur compréhension ainsi que l'adaptation du spectre pour ces structures, suivie

de l'expérimentation des modèles d'apprentissage automatique appropriés. Pour ces deux chapitres, les procédures efficaces pour l'utilisation du spectre de réponse en accélération pour l'apprentissage automatique sont proposées. Le quatrième chapitre se concentre sur une validation des procédures en utilisant les séismes réels. La sélection des enregistrements sismiques réels est aussi présentée. Ce chapitre montre les différences et similitudes dans les résultats des modèles de machine learning, analysant leur pertinence par rapport aux deux types de données sismiques. Le cinquième chapitre s'attèle à la validation et à l'analyse des résultats des modèles de machine learning sur les différentes structures dans les bases de données disponible dans la littérature, mettant en lumière les limites, les résultats significatifs et les recommandations qui en découlent. Enfin, dans "Conclusions et Perspectives", un récapitulatif des principaux résultats ainsi que des recommandations pour des études futures sont présentés.

1.8. Conclusion

Les courbes de fragilité sismique se sont prouvées très utiles pour l'évaluation des risques sismiques. Leur construction par les simulations numériques est la plus efficace mais elle rencontre un grand défi lié à la charge de calcul importante des simulations sismiques par la méthode des éléments finis. L'étude bibliographique du chapitre montre qu'il est possible de dépasser cet obstacle par l'application de l'apprentissage automatique. Pourtant, il est difficile de réutiliser les études existantes puisqu'il n'existe pas une méthode et/ou une procédure systématique. En plus, le choix des caractéristiques des modèles est très large et aucune recommandation est donnée pour une structure nouvelle. Dans ce contexte, l'objectif de la thèse est de proposer une nouvelle méthode d'application des modèles d'apprentissage automatique avec les spectres de réponse en accélération comme les caractéristiques. Elle est détaillée sous forme des procédures pratiques pas-à-pas avec une stratégie d'échantillonnage et de sélection des caractéristiques les plus pertinentes. Le chapitre suivant présente la méthode pour les structures linéaires.

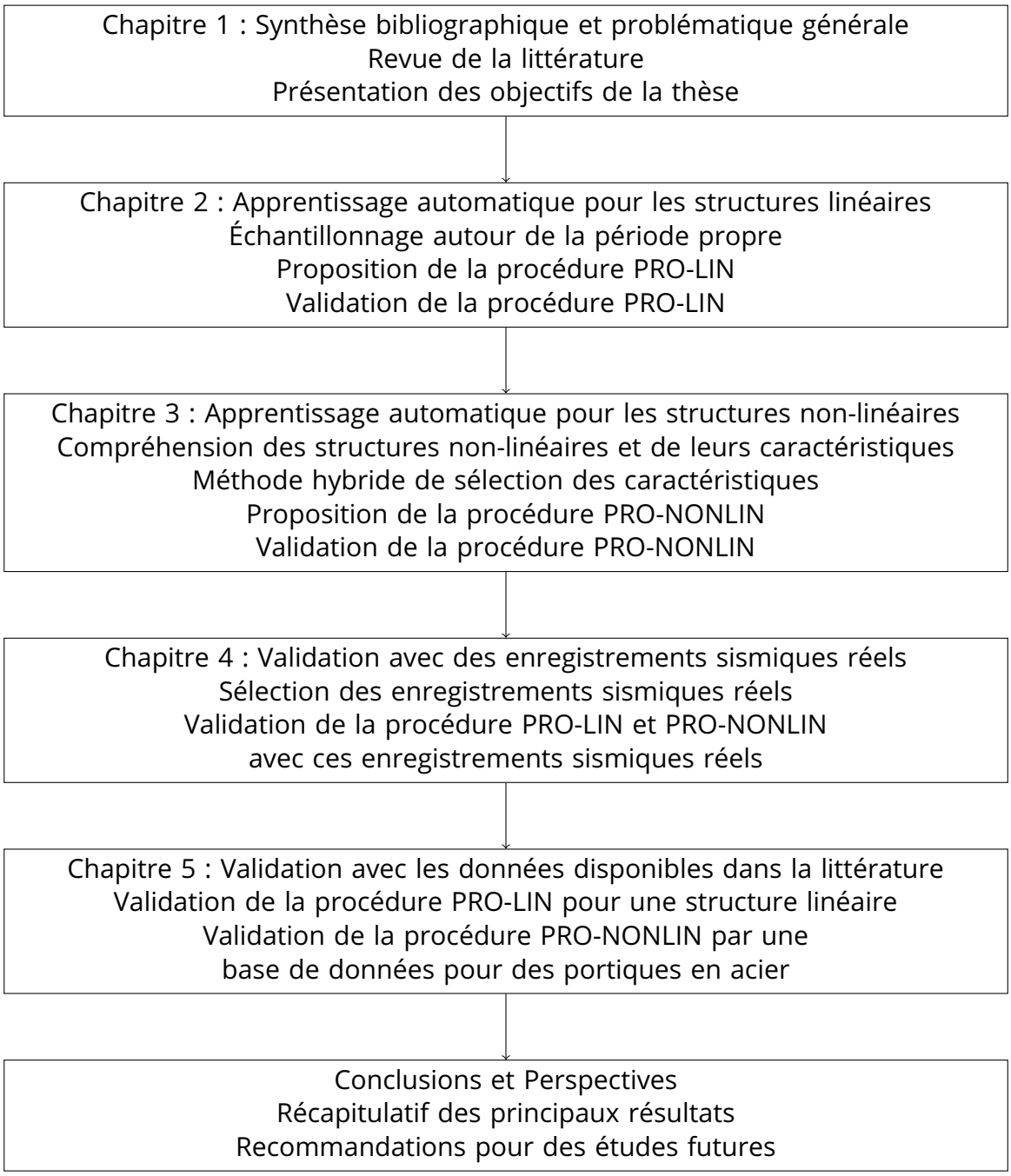


Figure 1.19 – Organisation de la thèse

2 Apprentissage automatique aux structures linéaires

Sommaire

2.1. Introduction

L'étude bibliographique présentée dans le chapitre 1 montre que l'application des modèles d'apprentissage automatique est possible pour substituer les déterminations des réponses sismiques par la méthode des éléments finis. Toutefois, son application pour une nouvelle structure est loin d'être évidente à cause de plusieurs difficultés. Tout d'abord, le choix des caractéristiques des modèles n'est pas clair et il n'est donc pas immédiat. Les caractéristiques varient d'une étude à l'autre tant pour la nature que par la quantité. D'un point de vue pratique, pour l'instant aucune procédure concrète a été recommandée pour appliquer l'apprentissage automatique et pour optimiser les modèles.

L'objectif de ce chapitre est de donc proposer une méthode d'application de l'apprentissage automatique pour modéliser les réponses sismiques en considérant les spectres de réponse en accélération comme les caractéristiques des modèles. Ce chapitre se concentre sur les structures linéaires. Il est à noter que des études existantes [95, 74] ont déjà utilisé des valeurs du spectre de réponse dans le choix des caractéristiques. Toutefois, aucune précision sur l'étendue et sur les périodes à échantillonner, est indiquée.

Après une analyse préliminaire en section 2.2, les zones les plus importantes des spectres de réponses en accélération dans les voisinages des modes propres des structures, sont identifiées. En prenant en compte cette propriété et en combinant avec l'analyse SHAP (Shapley Additive exPlanations) pour mesurer l'importance des caractéristiques des modèles, une procédure pratique complète pas-à-pas PRO-LIN est proposée et détaillée en section 2.3. Elle est accompagnée par un schéma algorithmique pour faciliter la mise en œuvre pratique.

La validation de la procédure PRO-LIN est présentée en section 2.4. Deux systèmes linéaires à un degré de liberté puis à deux degrés de liberté sont utilisés pour illustrer l'application de la procédure. Un modèle d'apprentissage automatique entrainé et validé peut être utilisé pour déduire ensuite des courbes de fragilité sismique.

Des conclusions sur la validité de la méthode proposée et de la procédure PRO-LIN sont données en fin du chapitre sur la base de la performance des modèles obtenus et la précision des courbes de fragilité sismique établies.

2.2. Analyse préliminaire du choix des spectres de réponse en accélération

L'utilisation du spectre de réponse pour entrainer les modèles de machine learning est très prometteuse. L'une des mesures d'intensité la plus importante et couramment utilisée pour représenter les signaux sismiques est le spectre de réponse. Plus spécifiquement le spectre de réponse en accélération. Il contient les informations du signal sismique ainsi que du système considéré. Un spectre de réponse est obtenu en calculant la valeur maximale de la réponse de la structure (accélération, vitesse et déplacement), un taux d'amortissement et une large zone de fréquences d'un oscillateur linéaire pour un séisme spécifique. Par conséquence, le spectre de réponse en accélération à une

période égale à la période fondamentale du système peut être une caractéristique importante pour prédire le déplacement maximal du système à un degré de liberté. De même, pour un système à plusieurs degrés de liberté, il est probablement important pour prédire le déplacement structurel. De plus, les spectres de réponse en accélération sont toujours disponibles dans les bases de données des enregistrements sismiques. Sa disponibilité offre un emploi plus facile que certaines autres mesures d'intensités, qui doivent être calculées en amont de l'étude.

Cependant, concernant l'utilisation du spectre de réponse en accélération comme caractéristiques pour l'entraînement d'un modèle d'apprentissage automatique, l'utilisation de l'ensemble du spectre de réponse ne peut pas être pratique en raison de sa nature continue. Cette continuité du spectre peut créer une quantité excessive de caractéristiques pour le processus d'entraînement. Une approche potentielle pour incorporer le spectre de réponse en accélération en tant que caractéristique consiste à échantillonner un nombre fini de périodes et à les regrouper. Comme illustré dans la Figure 2.1, le spectre de réponse en accélération est échantillonné dans l'intervalle de 0 à 5 s, avec un incrément de 0.5 s. De cette manière, chaque observation comprend une série de valeurs du spectre de réponse en accélération échantillonnées aux périodes différentes, $S_a(T_i)$. Par conséquence, la quantité de caractéristiques dépendra de la largeur et de la résolution de l'intervalle d'échantillonnage du spectre.

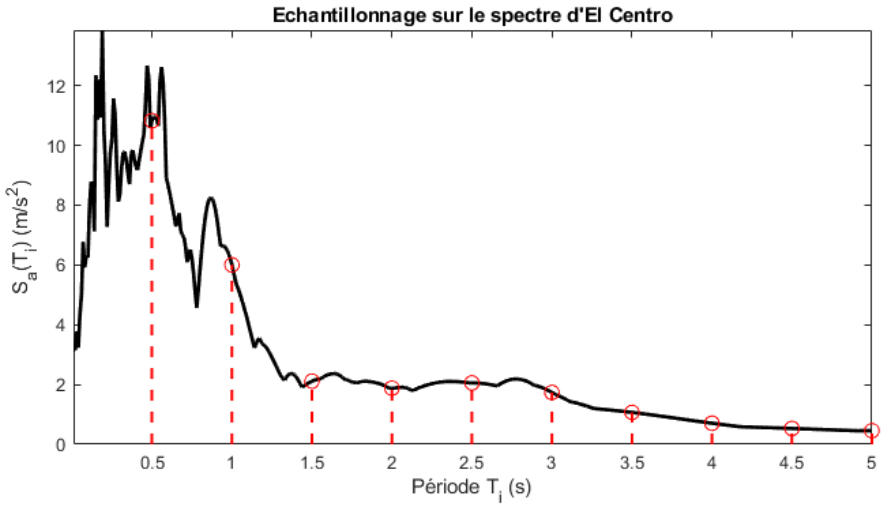


Figure 2.1 – Exemple d'échantillonnage de $S_a(T_i)$

De plus, l'utilisation des spectres de réponse est plus raisonnable que l'utilisation des accélérations temporelles du sol comme caractéristiques. En effet, un spectre de réponse en accélération contient non seulement l'information du signal sismique mais aussi celle de la réponse structurelle, tandis qu'une accélération temporelle du sol contient seulement la première information. Deuxièmement, la longueur des signaux sismiques est différente, elle dépend de la durée d'un séisme, du pas de temps utilisé pour enregistrer des signaux de chaque système ou station de mesure, etc. De plus, la durée significative de chaque signal est différente, d'où la difficulté de définir une caractéristique propre à partir de ces valeurs. En revanche, les spectres de réponse en accélération n'ont pas ces points faibles. Les spectres ont théoriquement des valeurs sur tous les périodes et donc elles sont similaires pour tous les signaux. Un spectre présente aussi l'influence différente du signal sur des

oscillateurs linéaires. C'est la raison pour laquelle l'utilisation des spectres de réponse en accélération comme caractéristiques est plus intéressante que les accélérations temporelles.

Dans une première étape, nous allons considérer les valeurs du spectre de réponse en accélération comme caractéristiques. En regardant les structures linéaires à considérer dans ce chapitre, nous allons échantillonner autour des périodes propres des structures. Par exemple, la Figure 2.2 présente la superposition du spectre de réponse en accélération et de la fonction de réponse en fréquence pour un système linéaire à trois degrés de liberté. La structure est une poutre simple d'une longueur de 29 m sur deux appuis, d'un module d'élasticité égal à 3.85 $\times$ 10^{10}N/m^2, d'un moment d'inertie égal à 2.4 m^4 et d'une masse linéique égale à 19.7 kg/m. Ses trois premières fréquences propres sont 4.039, 16.043 et 34.063 Hz. Les calculs des fonctions de réponse en fréquence ont été réalisés dans la plage de fréquences de 0 à 40 Hz avec un pas de 0.01 Hz. L'amplitude de la fonction de réponse en fréquence dont la réponse et l'excitation est au premier degré de liberté est présentée sur cette figure. Le spectre de réponse, pour le but d'illustration, est le spectre moyen des 100 séismes synthétiques de Boore. Les pics de réponse en fréquence permettent de définir un choix des intervalles pour l'échantillonnage du spectre de réponse en accélération. La stratégie d'échantillonnage définit une attention plus forte autour des fréquences propres de la structure que des autres fréquences. Il est important de noter que l'intervalle d'échantillonnage est corrélé à la largeur des pics de la réponse en fréquence, qui dépend à son tour du taux d'amortissement du système.

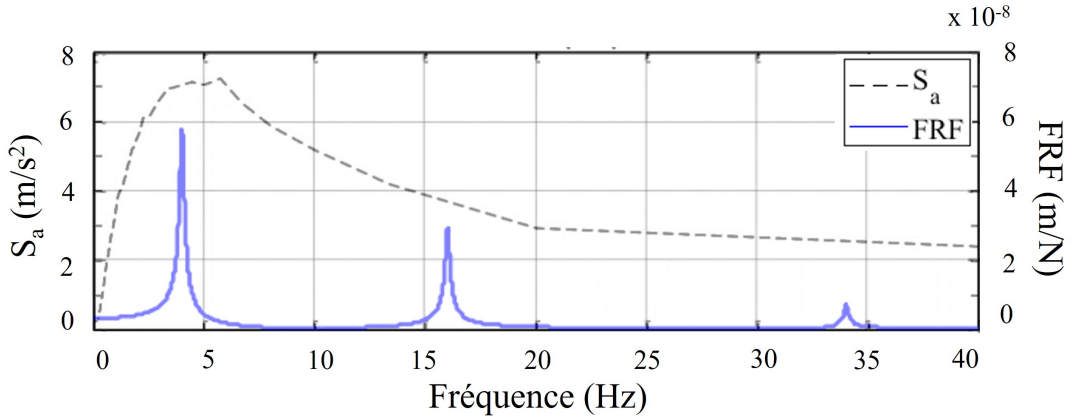


Figure 2.2 – Spectre de réponse en accélération avec la fonction de réponse en fréquence

Par conséquence, en utilisant le spectre de réponse en accélération comme caractéristiques, la relation entre la réponse de la structure $\mathbf{Y}$ et le vecteur de caractéristiques $\mathbf{x}$ dans Équation (1.8) peut être modifiée ainsi :

$$\mathbf{Y} = \mathbf{Y}' + \epsilon = \mathbf{ML}(\mathbf{x}) + \epsilon, \qquad \mathbf{x} = \{x_1, x_2, \ldots, x_n\}$$

$$\mathbf{Y} = \mathbf{ML}(\mathbf{x}) + \epsilon, \qquad \mathbf{x} = \{S_a(T_i), S_a(T_j^0)\} \tag{2.1}$$

où T_i représente la i-ème période et T_j^0 représente la j-ème période propre de la structure à échantillonner.

La stratégie d'échantillonnage du spectre de réponse en accélération en considérant les valeurs des périodes naturelles du système, en combinaison avec le choix d'un pas d'échantillonnage, sera présentée dans les sections suivantes. Ces échantillonnages seront ensuite utilisés comme caractéristiques $\mathbf{x}$ pour les modèles d'apprentissage automatique.

2.3. Procédure PRO-LIN pour les structures linéaires

Une procédure basée sur le modèle d'apprentissage automatique en utilisant un spectre de réponse en accélération est proposée, nommée PRO-LIN, pour des structures linéaires afin de construire les courbes de fragilité. Cette section décrit en détail la procédure PRO-LIN de construction des modèles de l'apprentissage automatique avec un spectre de réponse en accélération en tant que caractéristiques, l'utilisation des algorithmes d'apprentissage automatique et l'évaluation de la performance du modèle. Enfin, les courbes de fragilité sismique sont construites par les observations générées par les modèles de ML.

Pour prendre en compte les valeurs du spectre de réponse en accélération échantillonnées aux différentes périodes comme la caractéristiques $\mathbf{x}$, la procédure de référence dans la Figure 1.15 est modifiée pour les systèmes linéaires. La nouvelle procédure, illustrée par la Figure 2.3, comprend quatre étapes principales, qui seront présentées plus en détail dans les sous-sections suivantes.

- **Première étape :** Création du jeu de données : Pour entraîner un modèle d'apprentissage automatique, il est nécessaire de créer un jeu de données. Ce jeu de données comprend les observations des enregistrements sismiques, présentées par des caractéristiques $\mathbf{x}$ et la réponse correspondante $\mathbf{Y}$. Pour les structures linéaires, la procédure PRO-LIN propose d'échantillonner autour de leurs périodes propres. La réponse correspond à la réaction de la structure aux excitations sismiques, comme le déplacement maximal, le déplacement relatif entre étage maximal, etc. Cette réponse est obtenue par les mesures sur les structures réelles, ou par les analyses à l'aide des simulations par éléments finis.
- **Deuxième étape :** Entraînement du modèle d'apprentissage automatique : Le modèle d'apprentissage automatique est entraîné à l'aide du jeu de données créé à l'étape 1. On commence par diviser le jeu de données en deux sous-ensembles, l'un d'entraînement et l'autre de test. Puis on sélectionne un algorithme d'apprentissage automatique approprié, et on ajuste les paramètres pour obtenir les meilleures performances. Il faut noter que l'entrainement du modèle est accompagnement avec la sélection des caractéristiques pour améliorer l'efficacité du modèle. Une fois que le modèle d'apprentissage automatique est entraîné, il doit être validé à l'aide du jeu de données de test pour garantir la généralité du modèle.
- **Troisième étape :** Utilisation du modèle d'apprentissage automatique : Une fois que le modèle d'apprentissage automatique est validé, de nouvelles excitations sismiques, représentées par leurs spectres de réponse en accélération correspondants, peuvent être transmises aux modèles d'apprentissage automatique entraînés pour prédire les réponses structurelles.
- **Quatrième étape :** Construction de la courbe de fragilité : Les réponses prédites par les modèles d'apprentissage automatique sont utilisées pour construire la courbe de fragilité en utilisant des méthodes existantes, présentées dans la section 1.2. Ici on présente la construction

de la courbe de fragilité comme une étude spécifique, mais cette procédure est applicable pour les autres études.

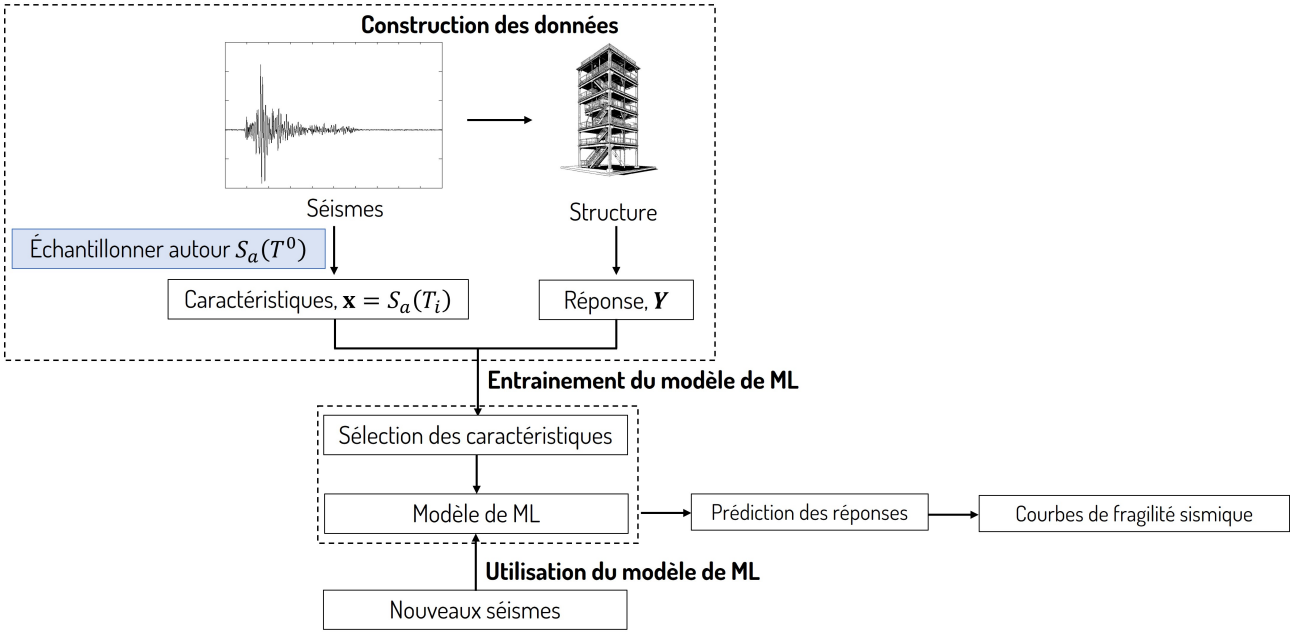


Figure 2.3 – [PRO-LIN] Procédure de l'application de ML pour des structures linéaires

2.3.1. Construction du jeu de données

Généralement, selon le bilan des études de machine learning 1.4, les caractéristiques $\mathbf{x}$ contiennent les mesures d'intensité sismique des mouvements du sol et les paramètres de la structure si nécessaire.

Selon la proposition de ce travail, pour prédire la réponse $\mathbf{Y}$ des structures, les caractéristiques $\mathbf{x}$ sont des valeurs du spectre de réponse en accélérations échantillonnées aux différentes périodes.

Ces caractéristiques ont pour but de décrire les enregistrements sismiques aux modèles de machine learning. Selon les circonstances, ces enregistrements de mouvement du sol peuvent être des enregistrements réels provenant d'une station sismique ou des signaux sismiques générés de manière synthétique. Le choix des signaux sismiques dépend de chaque étude.

Les spectres de réponse en accélération des mouvements sismiques sélectionnés ou générés sont calculés par une analyse spectrale et ensuite échantillonnés de manière appropriée à différentes périodes pour former les caractéristiques, $\mathbf{x} = \{Sa(T_i), Sa(T_i^0)\}$. Les valeurs spectrales à échantillonner dépendent de la relation entre les caractéristiques et la réponse souhaitée. Il faut aussi noter que la sélection des caractéristiques est nécessaire pour cette étape. Cette sélection est nécessaire pour améliorer l'efficacité du modèle, en supprimant des caractéristiques insignifiantes pour l'entrainement du modèle. Pour les structures linéaires, la procédure PRO-LIN propose échantillonner autour de leurs périodes propres.

Ces mouvements du sol sont aussi utilisés pour calculer la réponse souhaitée de la structure $\mathbf{Y}$ à l'aide de simulations par éléments finis.

Finalement, le résultat de cette étape est un ensemble de données contenant les caractéristiques $\mathbf{x}$ et la réponse structurelle $\mathbf{Y}$.

2.3.2. Entrainement des modèles d'apprentissage automatique

Les algorithmes d'apprentissage automatique sont résumés dans la section 1.3.2. Ces techniques montrent leur potentiel d'utilisation dans le domaine du génie civil en général et dans les études de l'évaluation des risques sismiques en particulier.

Dans ce chapitre, les études sont réalisées pour examiner et comparer la faisabilité des algorithmes de machine learning de régression, nommées régression linéaire (LR), k-plus proches voisins (KNN), forêt aléatoire (RF), arbre de décision (DT), machines à vecteurs de support (SVMR), réseaux de neurones artificiels (ANN), algorithme de renforcement adaptatif (AdaBoost), machine à renforcement légère de gradient (LightGBM) et machine à renforcement extrême de gradient (XGBoost) pour différentes structures. Les méthodes d'apprentissage automatique utilisées dans cette étude sont mises en œuvre à l'aide du langage de programmation Python avec la bibliothèque TensorFlow [108] via JupyterLab.

De plus, parmi plusieurs métriques pour évaluer les modèles de machine learning, il est important de choisir des métriques faciles à interpréter et suffisamment polyvalentes pour permettre aux utilisateurs d'ajuster le seuil d'exactitude acceptable.

L'une des métriques couramment utilisés pour évaluer les performances des modèles de régression est le coefficient de détermination (R^2), qui reflète la proportion de la variance entre la réponse réelle calculée ou mesurée sur la structure, $\mathbf{Y}$, et la réponse prédite par le modèle de ML, $\mathbf{Y}'$. Bien que ce coefficient mesure la qualité de l'ajustement d'un modèle, il ne quantifie pas explicitement sa précision. Des valeurs plus proches de 1 indiquent une meilleure performance.

Un autre indicateur est le coefficient de corrélation de Pearson (r), qui est une mesure statistique évaluant la relation linéaire entre deux variables continues. Il varie de -1 à 1, où -1 indique une relation linéaire négative parfaite, 0 indique l'absence de relation linéaire et 1 indique une relation linéaire positive parfaite. Il est couramment utilisé dans l'analyse de données et la recherche pour déterminer la force et la direction de la relation entre deux variables.

L'erreur de pourcentage absolue moyenne symétrique ($SMAPE$) est une mesure utilisée pour évaluer l'exactitude d'une prévision en comparant les valeurs réelles et prédites. Elle tient compte de l'échelle des valeurs réelles et prédites en utilisant leur moyenne. La SMAPE est exprimée en pourcentage et des valeurs plus proches de 0 indiquent une meilleure précision.

L'erreur quadratique moyenne ($RMSE$) est une mesure de performance largement utilisée pour les modèles de régression qui mesure la différence entre les valeurs prédites et réelles d'une variable cible. Elle est calculée en prenant la racine carrée de la moyenne des carrés des différences entre les valeurs prédites et réelles. La valeur de $RMSE$ indique à quel point les valeurs prédites sont proches des valeurs réelles en moyenne. Elle est exprimée dans les mêmes unités que la variable de réponse, ce qui facilite son interprétation. Une $RMSE$ plus basse indique une meilleure performance du modèle.

2.3.3. Utilisation du modèle d'apprentissage automatique

Après avoir réussi à entrainer les modèles de machine learning, ces modèles sont prêts à être déployés. Les nouvelles excitations sismiques, représentées par leurs spectres de réponse en accélération correspondant, peuvent être transmises aux modèles d'apprentissage automatique. Ces modèles prédisent la réponse structurelle, sans analyses par éléments finis qui sont couteuses en temps et matériel. La capacité de donner une réponse immédiate par l'utilisation des modèles de machine learning facilite les réactions contre ces désastres, pour éviter des pertes humaines et matérielles, ainsi que pour préparer la planification d'aides et de réhabilitions durant et après ces désastres.

On peut facilement construire les courbes de fragilité sismique. Puisque la construction des courbes de fragilité demande un grand nombre d'observations, il très difficile à atteindre avec les simulations par éléments finis. L'utilisation du modèle d'apprentissage automatique ouvre la possibilité d'obtenir suffisamment d'observations sans demande trop de charges supplémentaires.

2.4. Validation de la procédure PRO-LIN

2.4.1. Introduction

Dans cette section, la procédure proposée a été appliquée à deux structures présentant des comportements linéaires, à un degré de liberté et à deux degrés de liberté. L'oscillateur linéaire utilisé dans cette étude est similaire à celui utilisé par Le *et al.* [10]. Kafali [7] a réalisé une analyse de fragilité pour cette structure, en utilisant 1000 différentes excitations sismiques. De même, Le et al. [10] ont comparé différentes méthodes de construction de courbes de fragilité en utilisant jusqu'à 2×10^5 simulations numériques. Cette étude permet de comparer les courbes de fragilité construites par des méthodes conventionnelles avec celles obtenues par l'utilisation des modèles d'apprentissage automatique. En général, le choix des structures linéaires permet d'obtenir un grand nombre de réponses structurales sous excitation sismique avec une charge de calcul réduite.

Pour obtenir le nombre souhaité d'observations de l'ordre 10^5 mouvements du sol pour la simulation de Monte Carlo, les mouvements du sol dans cette application sont générés par le modèle de Boore. Le générateur de séismes est implémenté en utilisant le logiciel MATLAB.

Les algorithmes d'apprentissage automatique mentionnés, notamment LR, KNN, RF, DT, SVMR, ANN, AdaBoost, LightGBM et XGBoost, sont utilisés dans cette étude. Afin de trouver la configuration optimale des modèles de machine learning, une optimisation de la configuration des modèles est réalisée. Cette optimisation est définie par une recherche par grille (ou grid search en anglais). Les configurations pour chaque modèle de machine learning sont détaillées pour chaque étude.

La base de données a été divisée en sous-ensembles d'entraînement et de test avec un ratio de 80 % - 20 % respectivement. Les données dans ensembles d'entraînement et de test sont normalisés entre [0, 1]. La performance des modèles entraînés est évaluée à l'aide des quatre mesures : le coefficient de corrélation de Pearson r, le coefficient de détermination R^2, l'erreur de pourcentage absolue moyenne symétrique $SMAPE$, et l'erreur quadratique moyenne $RMSE$. De plus, un indice de référence est calculé en utilisant les valeurs normalisées des quatre mesures de performance avec un schéma de pondération de 10 % pour r, 30 % pour R^2, 30 % pour $SMAPE$ et 30 % pour $RMSE$,

comme suggéré par Todorov *et al.* [32]. Plus précisément, les valeurs de r et R^2 sont normalisées par les valeurs maximales et minimales, tandis que les valeurs de $SMAPE$ et $RMSE$ sont normalisées dans le sens inverse, comme présenté par les formules (2.2) et (2.3) respectivement. Cet indice de référence permet de comparer les modèles d'apprentissage automatique entraînés.

$$\bar{x_i} = \frac{x_i - \min(x)}{\max(x) - \min(x)} \tag{2.2}$$

$$\bar{z_i} = 1 - \frac{z_i - \min(z)}{\max(z) - \min(z)} \tag{2.3}$$

Des courbes de fragilité sont construites à partir des résultats prédits une fois les modèles sont bien entrainés, selon la procédure PRO-LIN. Ces courbes seront comparées à celles construites par des approches conventionnelles basées sur la simulation par éléments finis.

2.4.2. Système à un degré de liberté

a. Présentation du problème

On étudie l'oscillateur linéaire, illustré à la Figure 2.4.

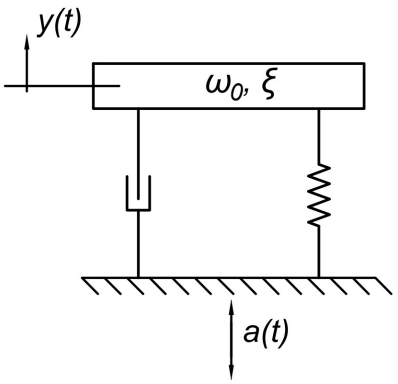


Figure 2.4 – Structure à un degré de liberté

Les paramètres numériques de cet oscillateur sont inspirés de Le *et al.* [10], avec une pulsation, ω_0, égale à 5.9 rad/s et un taux d'amortissement, ξ, de 2 %. La période fondamentale de la structure, T^0, est de 1.05 s. Le déplacement de la masse est calculé par le logiciel MATLAB, selon l'équation (2.4) :

$$\ddot{y}(t) + 2\xi\omega_0\dot{y} + \omega_0^2 y(t) = -a(t) \tag{2.4}$$

La réponse de la structure considérée pour la courbe de fragilité dans cet exemple est le déplacement maximal de la masse, $\mathbf{Y} = \max_t |y(t)|$.

b. Capacité de prédiction du spectre de réponse

On utilise la stratégie d'échantillonnage initiale pour le spectre de réponse en accélération $S_a(T_i)$, avec T_i = 0.1 à 2.0 s avec un pas de 0.1 s, qui pourra être ajustée par la suite pour améliorer son efficacité. Cette sélection correspond approximativement à la plage de 0.1 T^0 à 2 T^0, où T^0 représente la période fondamentale de la structure. Il est à noter que cette sélection évite l'utilisation de la valeur exacte de $S_a(T^0)$, qui est corrélée linéairement avec la réponse structurale. La Figure 2.5 illustre la position relative des périodes échantillonnées de $S_a(T_i)$ avec le spectre moyen des excitations sismiques considérées, qui est la moyenne des spectres de réponse en accélération de 100 séismes synthétiques.

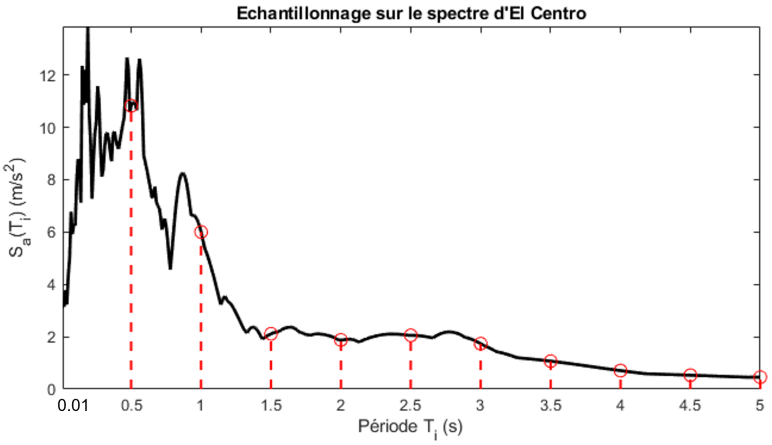


Figure 2.5 – [1ddl] Échantillonnage initiale du spectre de réponse en accélération moyenne

Neuf modèles d'apprentissage automatique ont été entraînés en utilisant les caractéristiques $\mathbf{x}$ de $S_a(T_i)$ et la réponse $\mathbf{Y}$. Les mesures de performance de ces modèles sont présentées dans le Tableau 2.1. Dans cette étude, les modèles sont testés avec les différentes configurations comme suivantes :

- KNN : Le nombre de voisins, k, variant de 2 à 25 sont recherchées pour déterminer la valeur optimale.
- DT : La profondeur maximale des arbres, $\max_{\text{profondeur}}$, dans ce cas est testée entre 1 et 10.
- RF : L'optimisation pour ce modèle est effectuée en recherchant un nombre maximal d'échantillons, m, entre 50 et 100 avec un incrément de 10, et un nombre d'arbres dans chaque forêt, n_{arbre}, entre 50 et 1000 avec un incrément de 50.
- ANN : A côté des fonctions d'activations $f_{\text{activation}}$, les configurations des réseau de neurones contient de 0 à 3 couches, dont le nombre de neurones varie entre 0 et 80.
- SVMR : Plusieurs fonctions de noyau f_{noyau}, telles que linéaire, polynomiale (avec des degrés de 3, 4 et 5), fonction de base radiale et sigmoïdale, ont été recherchés pour ce modèle. En outre, des valeurs de prix, C, allant de de 50 à 500 (incrément de 50), ainsi que des valeurs de pénalty, ϵ, allant de 0.001 à 1000 ont été explorées.
- AdaBoost : Pour ce modèle, l'optimisation est effectuée en modifiant le nombre des tronçons de décision, n étant compris entre 50 et 500 par incrément de 50.

- LightGBM : Pour ce modèle, l'optimisation est effectuée en modifiant le nombre de feuilles dans chaque arbre, n_{leaf} entre 2 et 30 sont considérés. En outre, des taux d'apprentissage, v, de 0.5 à 0.005 et des profondeurs d'arbre entre 2 et 25, $\max_{\text{profondeur}}$, sont également étudiés de manière exhaustive.
- XGBoost : L'optimisation est effectuée en modifiant le nombre des arbres, n_{arbre} entre 50 et 500 avec un incrément de 50 sont considérés. En outre, des taux d'apprentissage, v, compris entre 0.5 et 0.005 et des profondeurs d'arbre, $\max_{\text{profondeur}}$, comprises entre 2 et 10 sont également recherchées de manière exhaustive.

Les critères d'évaluation sur l'ensemble de test indiquent que ces modèles sont capables de prédire avec précision les réponses structurales. La valeur élevée du coefficient de corrélation de Pearson r et du coefficient de détermination R^2 suggère la cohérence entre les valeurs réelles et prédites des réponses. Sauf les modèles KNN, DT et LightGBM, tous les autres modèles ont un R^2 supérieur à 0.9. On peut observer que les modèles LR et ANN ont les valeurs les plus élevées de r, ce qui indique une forte corrélation entre les valeurs prédites et réelles. De même, ces modèles ont les valeurs les plus élevées de R^2, ce qui indique un bon ajustement entre les valeurs prédites et réelles. Le modèle ANN a les valeurs les plus faibles de $SMAPE$ et $RMSE$, ce qui indique qu'il a la meilleure performance globale parmi les modèles. Le modèle KNN, quant à lui, a les valeurs les plus élevées de $SMAPE$ et $RMSE$, ce qui indique une moins bonne performance par rapport aux autres modèles.

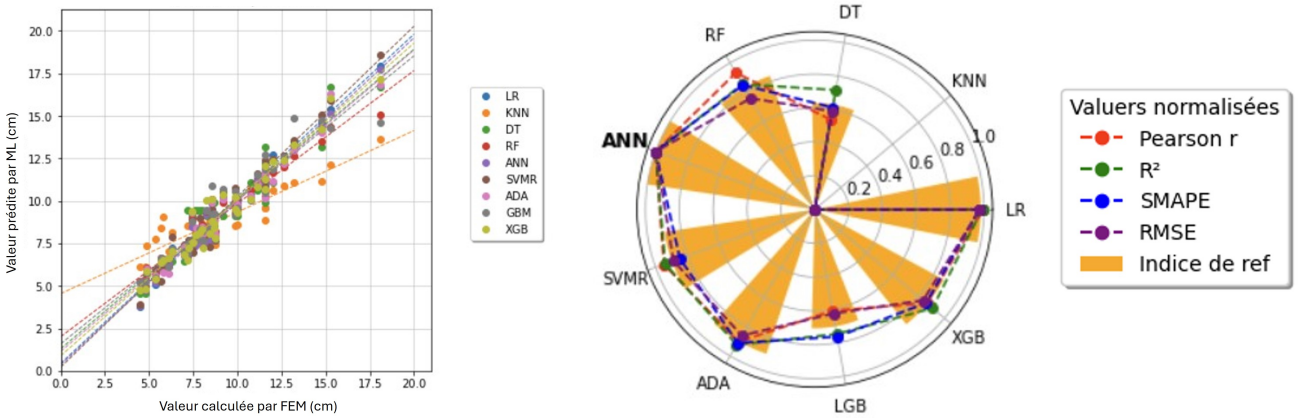
Modèle	**Configuration optimale**	r	R^2	$SMAPE$ (%)	$RMSE$ (cm)
LR	Non applicable	0.983	0.964	5.561	0.563
KNN	k = 7	0.896	0.671	14.680	1.708
DT	$\max_{\text{profondeur}}$ = 10	0.943	0.883	8.939	1.016
RF	m = 50, n_{arbre} = 650	0.978	0.924	6.671	0.818
ANN	$N_1 = 60, N_2 = 0, N_3 = 0, f_{\text{activation}} = ReLU$	0.983	0.968	5.294	0.535
SVMR	$f_{\text{noyau}} =$ RBF, C = 1000, $\epsilon = 0.001$	0.979	0.940	6.611	0.667
AdaBoost	$n = 400$	0.975	0.945	6.131	0.698
LightGBM	n_{feuille} = 4, v = 0.050, $\max_{\text{profondeur}}$ = 10	0.949	0.893	7.511	0.975
XGBoost	n_{arbre} = 450, v = 0.050, $\max_{\text{profondeur}}$ = 2	0.972	0.941	6.611	0.728

Table 2.1 – [1ddl] Évaluation de la performance des modèles de ML

La Figure 2.6 présente également une comparaison des modèles ML obtenus. Le diagramme de dispersion et la ligne de tendance dans la Figure 2.6a montrent que les lignes de tendance de la plupart des modèles sont proches de la pente de 45 degrés, ce qui indique des prédictions précises. La Figure 2.6b compare les modèles en utilisant les critères d'évaluation normalisés et l'indice de référence, où le modèle ANN est le plus performant, tandis que les modèles LR et SVMR ont également une performance approximativement équivalente. Il faut noter que Figure 2.6b ne montre seulement que le modèle KNN est moins performant que les autres, du fait que les critères sont normalisés par les valeurs maximales et minimales. Les critères de qualité r, R^2 du modèle sont minimales, donc

les valeurs normalisées égales à 0. Au contraire, les erreurs $SMAPE$ et $RMSE$ du modèle sont maximales, qui sont traduites à 0 par la normalisation des erreurs.

Les résultats montrent qu'avec cette sélection initiale en utilisant le spectre de réponse en accélération comme caractéristiques, il est possible de former des modèles d'apprentissage automatique avec une grande précision. De plus, ils présentent des performances différentes pour cette tâche. Cela souligne l'importance d'évaluer et de comparer différents modèles en utilisant plusieurs mesures pour sélectionner le modèle le plus efficace.



(a) Comparaison des valeurs prédites et actuelles **(b)** Comparaison des critères d'évaluation

Figure 2.6 – [1ddl] Comparaison des modèles de ML utilisés

c. Influence du nombre d'observations

L'objectif de l'utilisation des modèles d'apprentissage automatique est de réduire la charge de calcul des simulations par la méthode des éléments finis, qui sont intensives en temps et en ressources. Il est crucial d'étudier l'impact du nombre d'observations sur les performances du modèle. À cette fin, une analyse a été réalisée en utilisant des modèles de régression linéaire avec respectivement 100, 200, 300 et 400 observations. Les modèles LR et ANN ont été choisis pour leur précision. La performance a ensuite été évaluée et comparée afin de déterminer le nombre optimal d'observations nécessaires pour atteindre le niveau de précision souhaité.

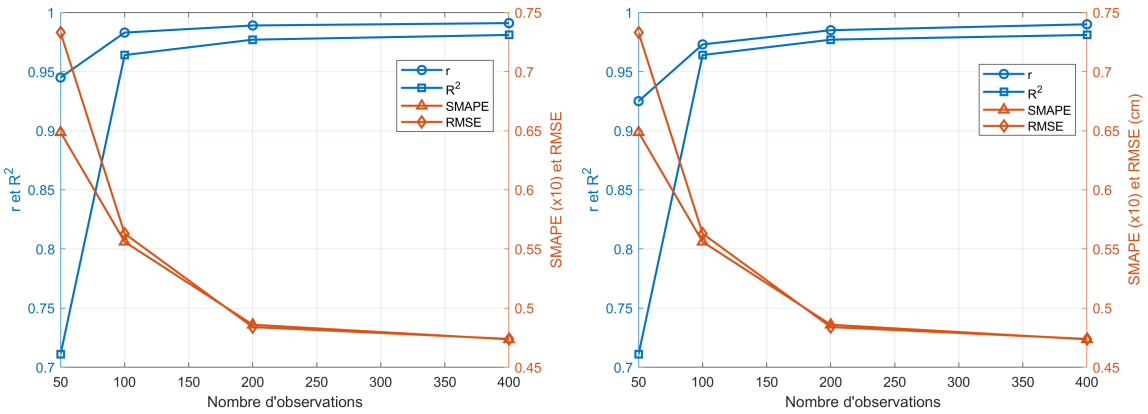
Le Tableau 2.2 présente les résultats de l'étude de l'impact du nombre d'observations sur les performances du modèle d'apprentissage automatique. Dans cet exemple, le modèle LR et ANN sont entraînés dix fois avec dix portions différentes des données. Les métriques de performance indiquées dans le Tableau 2.2 sont les moyennes obtenues. Les résultats démontrent qu'une augmentation du nombre d'observations de 50 à 400 entraîne une amélioration significative des performances des modèles.

Avec 50 observations, le modèle LR et ANN obtenu sont considérés comme insuffisants pour prédire avec précision le déplacement maximum de la structure, comme en témoignent les faibles valeurs de r et de R^2 par rapport aux autres jeux de données de taille plus élevée. Cependant, à mesure

que le nombre d'observations augmente à 100, les modèles obtenus montrent une amélioration progressive de la précision, avec une valeur de R^2 dépassant 0.9, une valeur de $SMAPE$ d'environ 5 % et une valeur de $RMSE$ de 0.5. Le modèle LR avec 400 observations est le plus précis de tous les modèles, obtenant de meilleures valeurs pour tous les critères d'évaluation. Ces résultats suggèrent qu'une utilisation d'un plus grand nombre d'observations peut améliorer la précision du modèle LR, en faisant un outil plus efficace et plus performant pour prédire le déplacement maximum des structures. Cependant, 200 observations peuvent être considérées comme la taille optimale des observations, car elles offrent un bon compromis entre les performances du modèle LR et le nombre des observations. Ce nombre est similaire avec le modèle ANN et donc utilisé pour les autres entrainements.

	LR				**ANN**			
Nombre	50	100	200	400	50	100	200	400
r	0.945	0.983	0.989	0.991	0.940	0.978	0.981	0.983
R^2	0.711	0.964	0.977	0.981	0.749	0.978	0.975	0.983
$SMAPE$ (%)	6.487	5.561	4.860	4.736	6.508	5.570	4.853	4.736
$RMSE$ (cm)	0.733	0.563	0.484	0.474	0.728	0.564	0.509	0.484

Table 2.2 – [1ddl] Modèles LR et ANN pour différents nombres d'observations



(a) Modèle LR **(b)** Modèle ANN

Figure 2.7 – [1ddl] Influence du nombre d'observations

d. Sélection des valeurs du spectre de réponse d'accélération à échantillonner

Comme mentionné dans la section précédente, la taille de l'intervalle d'échantillonnage et la résolution d'échantillonnage du spectre de réponse en accélération sont des paramètres qui affectent la quantité de données utilisées comme caractéristiques pour former un modèle d'apprentissage automatique. Pour tester l'idée proposée d'échantillonner le spectre d'accélération près des périodes

naturelles du système, au lieu de prélever presque tout le spectre, l'étude SHAP a été appliquée. SHAP (SHapley Additive exPlanations) par Lundberg *et al.* [109] est une méthode qui vise à expliquer des modèles en attribuant une valeur d'importance à chaque caractéristique du modèle de machine learning.

Pour la première approche, le modèle LR a été entraîné sur des données provenant de 200 spectres de réponse en accélération échantillonnés directement sur un intervalle de période de 0.1 s à 2.0 s avec incrément de 0.1 s. Le résultat de l'étude SHAP est présenté dans la Figure 2.8, montrant que la contribution de $S_a(T_i)$ avec T_i de 1.0 s et 1.1 s est plus importante que pour les autres périodes, suivie par les périodes entre 0.9 s et 1.3 s. Étant donné que la période fondamentale de la structure est de 1.05 s, ce résultat est en accord avec la littérature qui indique que le spectre de réponse en accélération à la période fondamentale de l'oscillateur est linéairement corrélé avec son déplacement maximal. Cela corrobore l'idée présentée dans la Figure 2.8 selon laquelle des données optimales peuvent être obtenues en réduisant l'intervalle d'échantillonnage $S_a(T_i)$ autour de la fréquence naturelle du système, c'est-à-dire de 0.8 à 1.2 T^0 de l'oscillateur.

Pour confirmer le résultat de l'étude SHAP, un modèle LR est entraîné en utilisant 200 observations, échantillonnées seulement à $S_a(T_i)$ avec T_i = 0.9, 1.0, 1.1, 1.2, 1.3 s correspondant de 0.8 à 1.2 T^0 (égale à 1.05 s pour ce cas) de l'oscillateur comme caractéristiques. Le modèle obtenu a une performance similaire à celle du modèle initial, comme le montre le Tableau 2.3, ce qui permet d'utiliser les T_i échantillonnées entre 0.8 T^0 et 1.2 T^0 comme caractéristiques. L'étude de SHAP confirme que les valeurs $S_a(T_i)$ autour de la période propre sont les plus importantes.

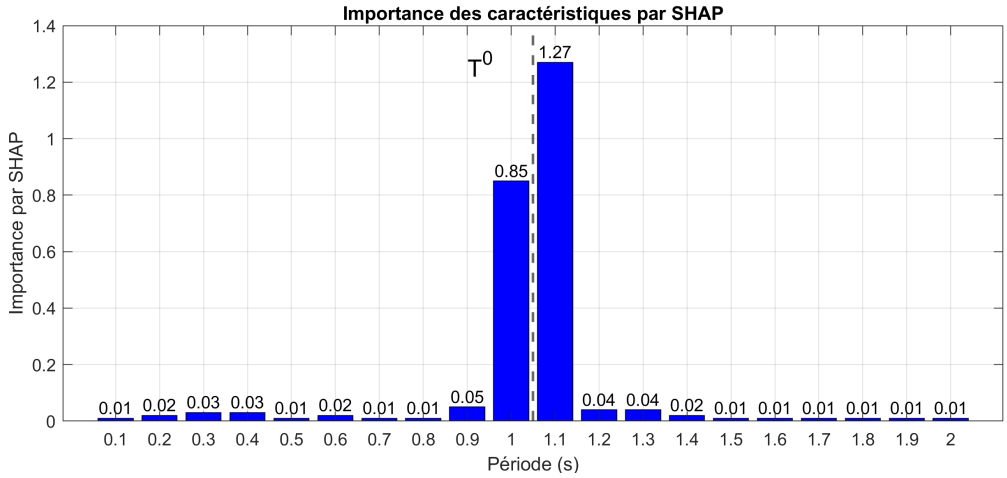


Figure 2.8 – [1ddl] Importance des caractéristiques par une étude de SHAP

Modèle	$S_a(T_i)$	r	R^2	$SMAPE$ (%)	$RMSE$ (cm)
LR1	0.1 : 2.0 s	0.989	0.977	4.860	0.484
LR2	0.9 : 1.3 s	0.983	0.962	5.089	0.543

Table 2.3 – [1ddl] Comparaison du modèle LR avec différentes caractéristiques

e. Construction des courbes de fragilité

Dans la dernière étape de cet exemple, des courbes de fragilité ont été construites à l'aide des réponses structurelles prédites par le modèle d'apprentissage automatique. Sur la base des résultats antérieurs, le modèle LR a été choisi et entraîné en utilisant 200 observations échantillonnées sur cinq périodes de 0.9 s à 1.3 s avec un incrément de 0.1 s. Après un entraînement réussi, le modèle LR a été utilisé pour prédire la réponse de la structure sous de nouvelles excitations sismiques. Les réponses prédites peuvent ensuite être utilisées pour construire les courbes de fragilité.

Trois états d'endommagement ont été sélectionnés correspondant au seuil de déplacement maximal de la masse y_0 égale à 7, 10 et 13 cm respectivement.

L'analyse de fragilité sismique de la structure en question a été réalisée à l'aide des méthodes d'estimation du maximum de vraisemblance (MLE) et de simulation de Monte Carlo (MCS). La procédure détaillée pour construire ces courbes de fragilité est présentée par Le *et al.* [10].

L'évaluation de la probabilité de défaillance pour chaque niveau de mesure d'intensité a par la méthode de Monte Carlo permet de déduire chaque point de la courbe de fragilité $Fr(a)$ selon l'équation (1.2). La qualité du calcul dépend fortement du nombre d'observations qui doit être suffisamment grand par rapport à la probabilité cible. La procédure de la méthode comporte 4 étapes :

- Étape 1 : Générer (ou sélectionner) un nombre N_{MCS} suffisamment grand de séismes indépendants.
- Étape 2 : Arranger ces signaux selon N_I niveaux de mesure d'intensité a_j dans les intervalles $[a_j - da_j, a_j + da_j]$ correspondantes. Dans cette thèse, la mesure d'intensité est le PGA.
- Étape 3 : Calculer les réponses sous excitations sismiques et vérifier l'état de la structure par rapport à son état limite y_0. Il faut noter que dans cette thèse, les réponses structurelles sont aussi générées par le modèle de l'apprentissage automatique. La méthode dont la courbe de fragilité sismique est déduite à partir des calculs par éléments finis est notée par MCS ; tandis que celle obtenue à partir des observations générées depuis les modèles de machine learning est notée par ML-MCS.
- Étape 4 : Estimer la probabilité de défaillance par la méthode de Monte Carlo pour chaque intervalle de a_j de PGA.

La procédure de construction des courbes de fragilité sismique selon la méthode MCS est considérée comme la référence des méthodes, pour un nombre N_{MCS} suffisamment grand.

En revanche, les courbes de fragilité par la méthode du maximum de vraisemblance sont considérées comme des approximations par rapport à celles de MCS. Cette méthode se base sur l'hypothèse de la forme log-normal des courbes. La médianne A_m et l'écart-type logarithme β sont estimés selon les équations (1.4) et (1.5). La procédure complète de la méthode contient 5 étape :

- Étape 1 : Générer (ou sélectionner) un nombre N_{MLE} séismes indépendants.
- Étape 2 : Calculer les réponses $\mathbf{Y}$ sous ces séismes et vérifier l'état de la structure par rapport à l'état limite y_0.
- Étape 3 : Établir la variable aléatoire de Bernoulli $Y_{\text{Bernoulli}}$, qui est attribuée à 1 ($y_i = 1$) au cas de défaillance, et à 0 ($y_i = 0$) sinon.
- Étape 4 : Définir la fonction de vraisemblance à maximiser avec deux paramètres A_m et β par l'équation (1.4).

- Étape 5 : Estimer ces deux paramètres de la courbe de fragilité en optimisant l'équation (1.5). Il faut aussi noter que la méthode dont la courbe de fragilité sismique est déduite à partir des calculs par éléments finis est notée par MLE ; tandis que celles obtenues à partir des observations générées depuis les modèles de machine learning est notée par ML-MLE.

Un ensemble de 10 000 enregistrements d'accélération du sol a été utilisé pour construire la courbe de fragilité par la méthode MLE. La méthode MCS demande un plus grand nombre d'accélérations du sol, soit 2×10^5 dans cette étude. Les courbes de fragilité par approche ML-MLE et ML-MCS basées sur l'apprentissage automatique ont été construites également avec le même nombre de mouvements du sol que les méthodes conventionnelles, soit 10 000 et 2×10^5 respectivement. Contrairement à la méthode de simulation numérique, les observations de ML-MLE et ML-MCS sont générées par le modèle d'apprentissage automatique. Les 200 simulations par éléments finis requises sont celles utilisées pour entraîner le modèle, tandis que les autres observations pour les courbes de fragilité sont prédites par ML. Le Tableau 2.4 présente le nombre d'observations pour construire les courbes de fragilité et la durée de simulation pour ces observations. Cela signifie que la méthode basée sur le modèle de ML est considérablement plus rapide que les méthodes conventionnelles, ne nécessitant que seulement 4 % et 0.2 % du temps de simulation des méthodes conventionnelles. Cela est seulement le cas de l'oscillateur linéaire. Pour les structures plus complexes, par exemple à plusieurs degrés de liberté et/ou non-linéaires, le facteur de temps réduit est encore plus élevé. Par conséquent, la méthode basée sur le modèle de ML présente un avantage clair par rapport à la méthode conventionnelle et offre une alternative efficace pour construire les courbes de fragilité.

Méthode	**MLE**	**MCS**	**ML-MLE**	**ML-MCS**
Nombre d'observations	10000	2×10^5	2×10^5	2×10^5
Nombre de simulations par éléments finis	10000	2×10^5	200	200
Durée de simulation (second)	2500	50000	100	100

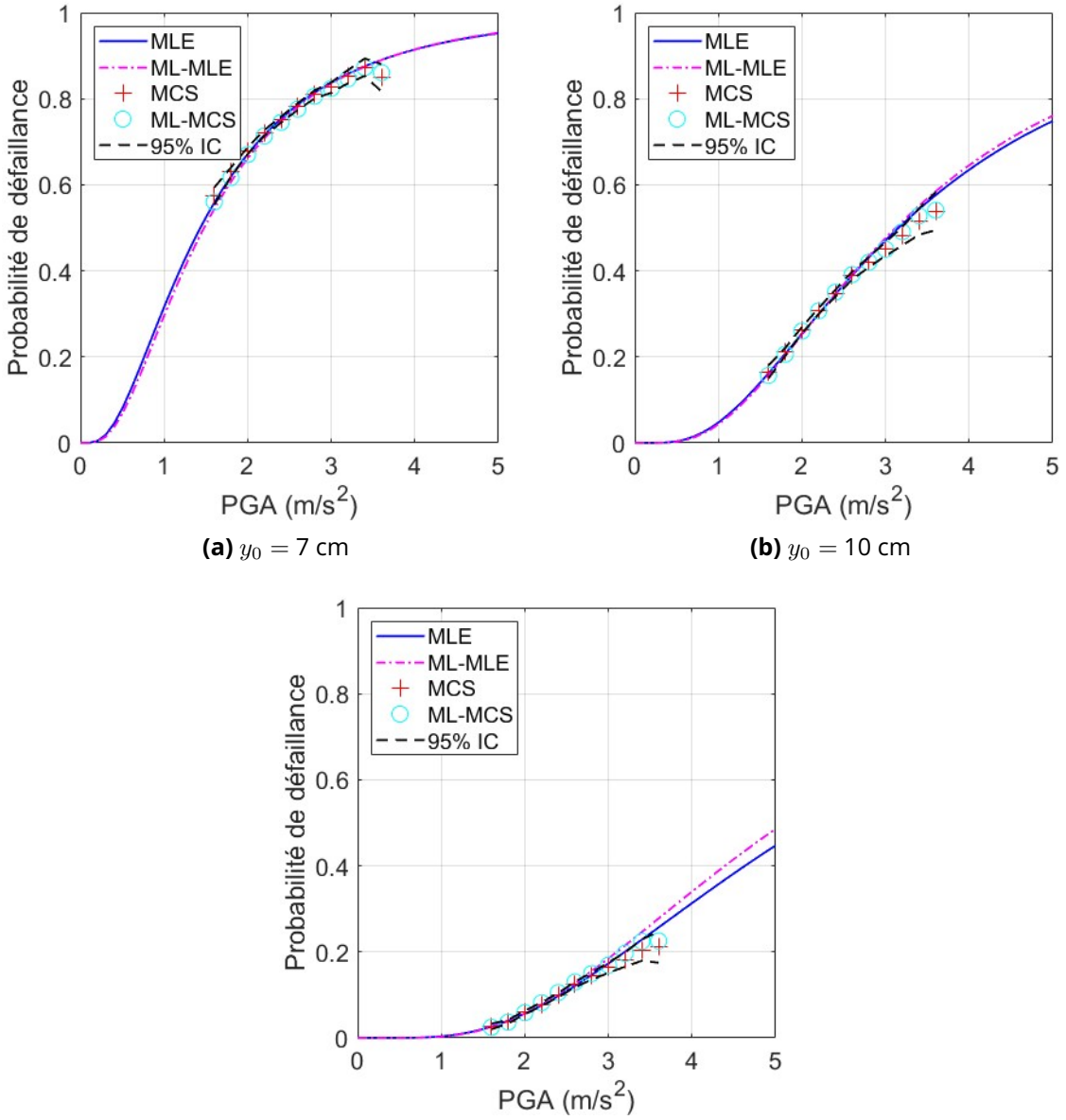
Table 2.4 – [1ddl] Nombres d'observations pour la construction de la courbe de fragilité

Un critère est utilisé pour comparer la performance de ces méthodes d'évaluation de fragilité. En utilisant les résultats de la méthode MCS comme référence, ce critère est basé sur la distance entre la courbe de fragilité et les résultats de référence. Ce critère est appelé l'erreur quadratique moyenne (MSE) et il est défini par :

$$MSE = \frac{1}{N_{\text{int}}} \sum_{i=1}^{N_{\text{int}}} \left[F_r^X(a_i) - F_r^{\text{MCS}}(a_i) \right]^2 \tag{2.5}$$

où N_{int} est le nombre total de niveaux d'intensité considérés dans l'analyse de fragilité, $F_r^X(a_i)$ représente la courbe de fragilité construite à l'aide de différentes méthodes à l'intensité a_i et $F_r^{\text{MCS}}(a_i)$ représente la courbe de fragilité de référence obtenue à partir de la méthode MCS au même niveau d'intensité a_i.

La valeur de MSE est une mesure utilisée pour évaluer la proximité de la courbe de fragilité générée par la méthode d'approximation (MLE, ML-MLE ou ML-MCS) et celle de référence obtenue à partir de la méthode MCS. Une valeur plus petite de MSE indique un meilleur accord entre les deux



(a) $y_0 = 7$ cm

(b) $y_0 = 10$ cm

(c) $y_0 = 13$ cm

Figure 2.9 – [1ddl] Courbes de fragilité par les méthodes différentes

courbes de fragilité, ce qui signifie que la méthode correspondante est plus précise pour prédire la probabilité de dépassement des différents états d'endommagement lors d'événements sismiques.

La Figure 2.9 illustre quatre courbes de fragilité construites à l'aide de quatre méthodes différentes. Les courbes par la méthode MCS sont représentés par des marqueurs de signe plus (+) ainsi que leurs intervalles de confiance à 95 %, représentées par deux lignes pointillées en noir. L'apparition des résultats ML-MCS dans l'intervalle de confiance à 95 % du MCS montre que les prédictions du ML sont fiables.

La Figure 2.10 présente le critère de comparaison MSE de ces courbes de fragilité pour chaque état limite. L'approche ML-MCS donne les résultats les plus précis dans cet exemple. Cela peut s'expliquer par le fait que les approches MLE et ML-MLE sont des méthodes approximatives basées sur l'hypothèse de distribution log-normale des courbes de fragilité, tandis que la méthode ML-MCS suit une approche similaire à la méthode MCS. Les courbes de fragilité obtenues par les méthodes MLE et ML-MLE sont aussi en bon accord avec les résultats de la méthode MCS de référence (dans l'intervalle de confiance).

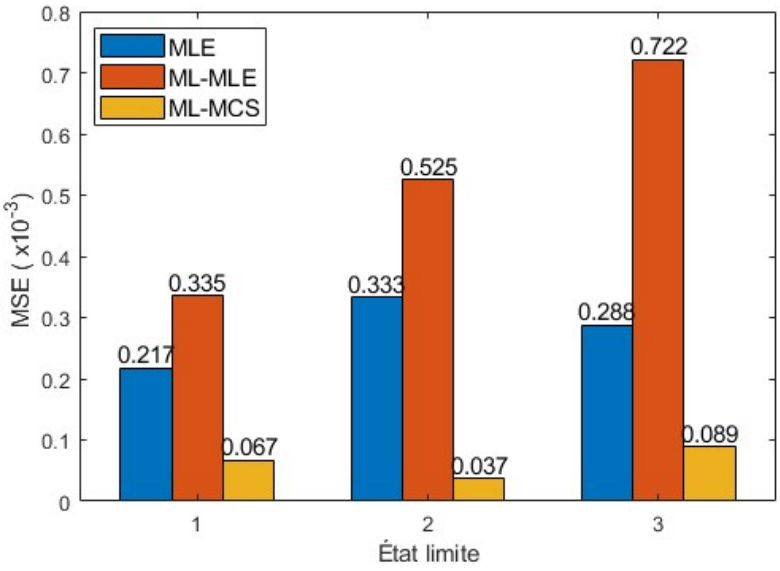


Figure 2.10 – [1ddl] Critère de comparaison pour trois états limites

2.4.3. Système à deux degrés de liberté

a. Présentation du problème

La même approche présentée dans la section précédente a été utilisée pour construire des modèles d'apprentissage automatique et des courbes de fragilité pour une structure à deux degrés de liberté, présentée par la Figure 2.11. Les matrices de rigidité et de masse ont été calculées avec les masses de plancher correspondantes m_1, m_2 et les raideurs latérales d'étage k_1, k_2 indiquées dans le Tableau 2.5. Les deux périodes propres de la structure sont de 1.34 s et 0.98 s. Le rapport d'amortissement de la structure a été choisi à 5 % pour les deux modes.

L'équation dynamique du système s'écrit sous forme matricielle dans l'équation (2.6). Elle sera résolue par la méthode de Newmark dans le logiciel MATLAB.

$$\mathbf{M}\ddot{\mathbf{u}}(t) + \mathbf{C}\dot{\mathbf{u}}(t) + \mathbf{K}\mathbf{u}(t) = -\mathbf{M}a(t) \tag{2.6}$$

Trois réponses structurelles peuvent être considérées comme la réponse souhaitée. Elles représentent, respectivement, le déplacement maximum en haut de la structure, noté $u_{top} = \max_t |u_2(t)|$, le

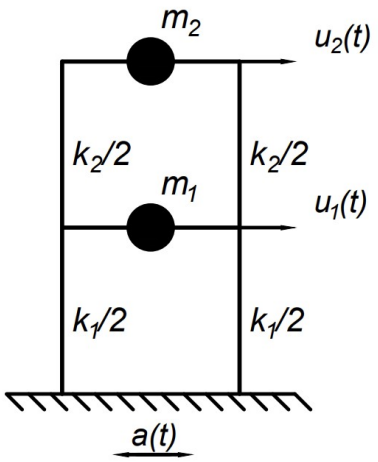


Figure 2.11 – Structure à deux degrés de liberté

Propriété	m_1 (kg)	m_2 (kg)	k_1 (N/m)	k_2 (N/m)
Valeur	20000	2000	6×10^6	0.6×10^6

Table 2.5 – [2ddl] Masses concentrées et rigidités du système

déplacement relatif maximum entre deux étages de la construction, noté $d_u^i = \max_t |u_{i+1}(t) - u_i(t)|$, et le taux maximum de déformation inter-étages de la structure, noté $\delta = \max_i |d_u^i / h_i| \times 100\%$, où h_i représente la hauteur de chaque étage. Dans cet exemple, seule la déformation inter-étages est utilisée comme réponse structurelle.

b. Potentialité du spectre de réponse comme caractéristiques de ML

Dans cet exemple, le spectre de réponse en accélération avec amortissement de 5 % a été utilisé comme caractéristiques, correspondant au rapport d'amortissement de la structure. En appliquant la procédure PRO-LIN, le spectre de réponse en accélération avec amortissement de 5 % a été échantillonné de 0.8 T_2^0 à 1.2 T_1^0, où T_1^0 et T_2^0 sont les périodes naturelles de la structure. Ainsi, les valeurs échantillonnées des périodes sont de 0.8 s à 1.6 s avec un incrément de 0.1 s. La Figure 2.12 représente le spectre moyen des enregistrements considérés dans cet exemple avec la position des périodes échantillonnées ainsi que les périodes de la structure. Cette stratégie d'échantillonnage réduit la charge informatique pour entrainer un modèle de ML, en comparaison de l'utilisation de la totalité du spectre, spécialement avec une haute résolution.

400 observations sont utilisées pour entrainer les modèles de ML.

En utilisant le spectre de réponse en accélération comme caractéristique $\mathbf{x}$ et le taux de déformation relative maximale entre étages comme réponse $\mathbf{Y}$, différents modèles d'apprentissage automatique ont été entraînés.

Le Tableau 2.6 présente les mesures de performance des modèles entraînés. Le coefficient de corrélation r et le coefficient de détermination R^2 les plus élevés sont obtenus par le modèle LR,

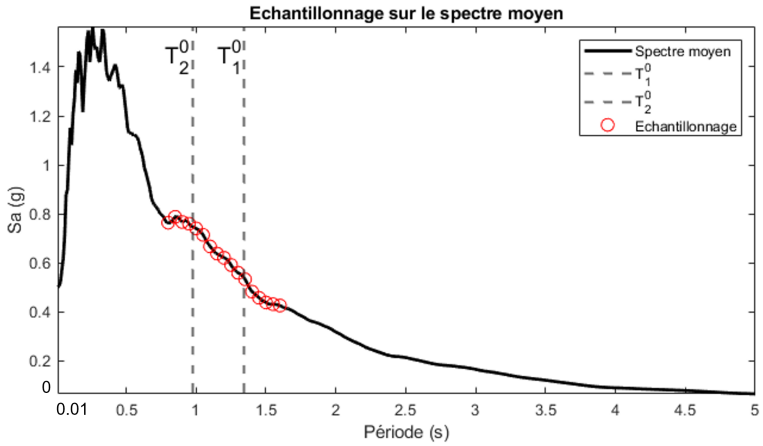
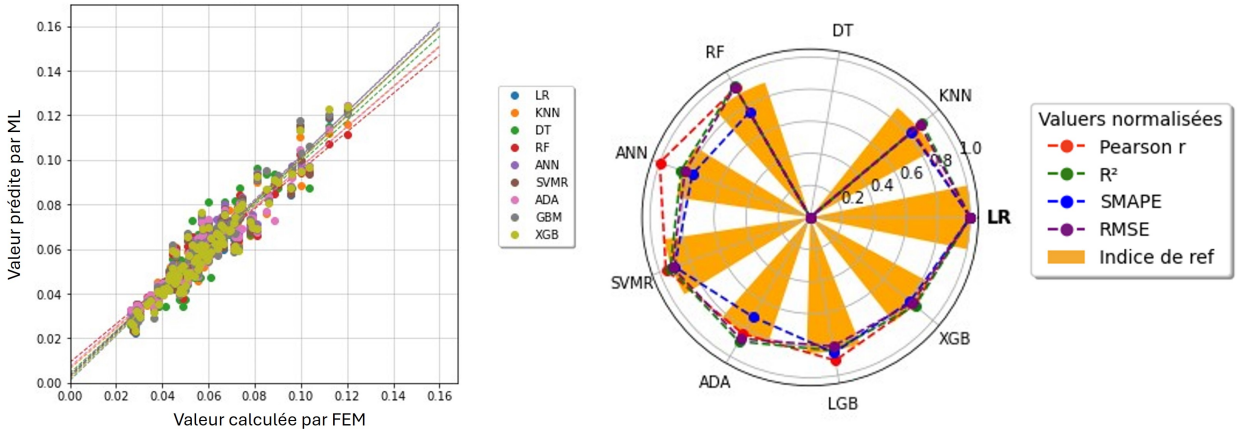


Figure 2.12 – [2ddl] Échantillonnage du spectre de réponse en accélération moyen et périodes propres du système

Modèle	r	R^2	$SMAPE$ (%)	$RMSE$
LR	0.960	0.915	7.056	5.578
KNN	0.953	0.906	7.740	5.838
DT	0.915	0.821	10.996	8.077
RF	0.957	0.910	8.009	5.732
ANN	0.960	0.901	7.909	5.997
SVMR	0.958	0.908	7.445	5.790
AdaBoost	0.953	0.905	8.174	5.896
LightGBM	0.956	0.900	7.615	6.037
XGBoost	0.953	0.901	7.786	5.999

Table 2.6 – [2ddl] Évaluation de la performance des modèles de ML

suivi de près par ANN et SVMR. Ces résultats sont également retrouvés dans l'erreur des modèles, le modèle LR ayant les valeurs les plus faibles d'erreurs $SMAPE$ et $RMSE$ parmi les modèles entraînés. Les critères d'évaluation sur l'ensemble de test indiquent que les modèles d'apprentissage automatique sont capables de prédire avec précision les réponses structurelles dans cet exemple. La valeur élevée du coefficient de Pearson r et du coefficient de détermination R^2 démontre la cohérence entre les réponses de la structure calculées à l'aide de ces modèles et de l'analyse par éléments finis, ce qui est illustré aussi dans la Figure 2.13. Cette figure présente la comparaison des modèles entraînés à l'aide du diagramme de dispersion et du graphique polaire. On observe une tendance similaire à l'exemple précédent. La Figure 2.13a illustre les lignes de tendance basées sur le diagramme de dispersion des valeurs réelles et prédites, tandis que la Figure 2.13b montre la comparaison de ces modèles d'apprentissage automatique à l'aide de quatre mesures d'évaluation et de l'indice de référence. Le modèle LR est le plus performant dans cet exemple, tandis que DT est le moins efficace.



(a) Comparaison des valeurs prédites et actuelles **(b)** Comparaison des critères d'évaluation

Figure 2.13 – [2ddl] Comparaison des modèles de ML utilisés

c. Influence du nombre d'observations

De manière similaire au système à un degré de liberté, cette section examine l'impact du nombre d'observations sur la précision des modèles d'apprentissage automatique. Pour étudier cette influence, cinq nombres différents d'observations ont été pris en compte, à savoir 50, 100, 200, 400 et 500. Les modèles LR et ANN ont été utilisés pour cette analyse, après avoir confirmé leur capacité dans l'exemple précédent. Plus précisément, la performance des modèles LR et ANN avec le nombre différent d'observations a été examinée.

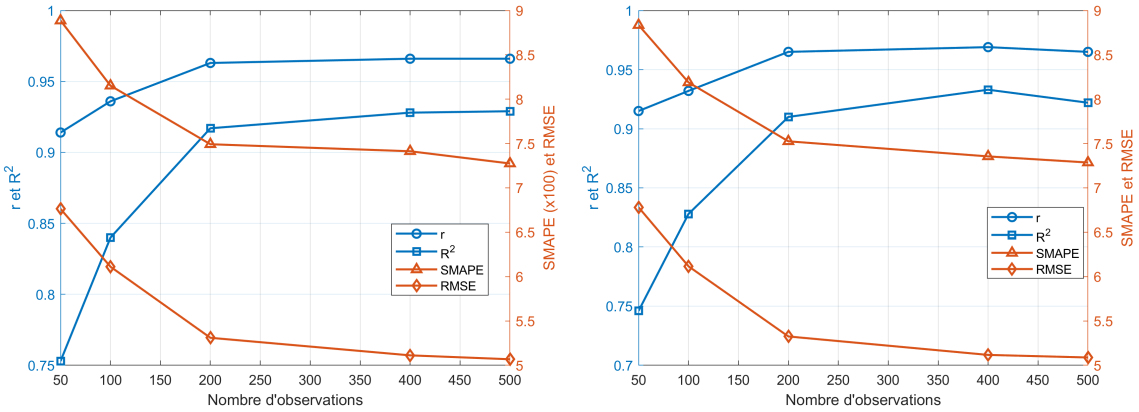
Le Tableau 2.7 présente les mesures de performance de cinq modèles qui ont été entraînés en utilisant les différents nombres d'observations définis. Ces mesures ont été évaluées à l'aide d'une validation croisée sur 10 fois d'entrainement, et leurs valeurs ont été moyennées entre les modèles. Globalement, la performance des modèles s'améliore à mesure que le nombre d'observations augmente, comme en témoignent les valeurs croissantes des mesures de performance. Les modèles LR et ANN entraînés avec seulement 50 observations se sont révélés moins capable de prédire le taux de déformation maximum de la structure, avec une valeur de R^2 de seulement 0.753 et 0.746 respectivement. Lorsque le nombre d'observations a augmenté à 200, les modèles obtenus ont montré une meilleure performance, avec des valeurs de R^2 augmentant progressivement au-dessus de 0.9. La valeur de $SMAPE$ et $RMSE$ diminue progressivement lorsque le nombre d'observations augmente. Les modèles avec 500 observations sont considérés comme les plus performants dans ce cas, atteignant les valeurs les plus élevées des critères d'évaluation. Cependant, il est à noter qu'après 400 observations, les améliorations de performance deviennent marginales, comme en témoignent les valeurs relativement proches des critères d'évaluation entre les tailles d'entrée de 400 et 500.

d. Sélection du spectre de réponse d'accélération à échantillonner

Pour examiner l'importance de l'échantillonnage des caractéristiques dans la prédiction de la réponse structurelle, dans cette section, un modèle de régression linéaire a été entraîné en utilisant 202

	LR					ANN				
Nombre	50	100	200	400	500	50	100	200	400	500
r	0.914	0.936	0.963	0.966	0.966	0.915	0.932	0.965	0.969	0.965
R^2	0.753	0.840	0.917	0.928	0.929	0.746	0.828	0.910	0.933	0.922
$SMAPE$ (%)	8.886	8.152	7.492	7.413	7.275	8.835	8.189	7.525	7.355	7.286
$RMSE$	6.763	6.111	5.310	5.112	5.068	6.781	6.116	5.325	5.117	5.088

Table 2.7 – [2ddl] Modèles LR et ANN pour différents nombres d'observations



(a) Modèle LR **(b)** Modèle ANN

Figure 2.14 – [2ddl] Influence du nombre d'observations

caractéristiques, obtenues par l'échantillonnage du spectre de réponse en accélération échantillonné entre 0.05 s et 10.0 s avec un incrément de 0.05s et deux périodes naturelles de la structure. Le résultat de cette étude SHAP est présenté dans la Figure 2.15, où les valeurs sont multipliées par 1000 pour une interprétation plus facile. Le résultat montre que la contribution de $S_a(T_i)$ échantillonné de 0.8 s à 1.6 s est plus importante que les autres.

De plus, une étude du coefficient de corrélation a également été réalisée. La corrélation entre une caractéristique quelconque, désigné x_i, dans l'ensemble des caractéristiques $\mathbf{x}$ avec la réponse Y, est calculée avec la formule (2.7).

$$r_{x_iY} = \frac{\sum_j (x_i^j - \bar{x}_i)(Y^j - \bar{Y})}{\sqrt{\sum_j (x_i^j - \bar{x}_i)^2 \sum_j (Y^j - \bar{Y})^2}} \tag{2.7}$$

La Figure 2.15 illustre les coefficients de corrélation (la ligne orage) et l'importance des valeurs du spectre de réponse (le diagramme à barres en bleu) par rapport aux taux de déformation relative maximal entre étages pour les échantillonnages jusqu'à 2 s. Les deux périodes propres de la structure sont marquées avec des lignes verticales (T_1^0 et T_2^0). Le résultat montre que les coefficients de corrélation sont faibles pour les périodes loin des périodes propres, tandis que les coefficients de corrélation

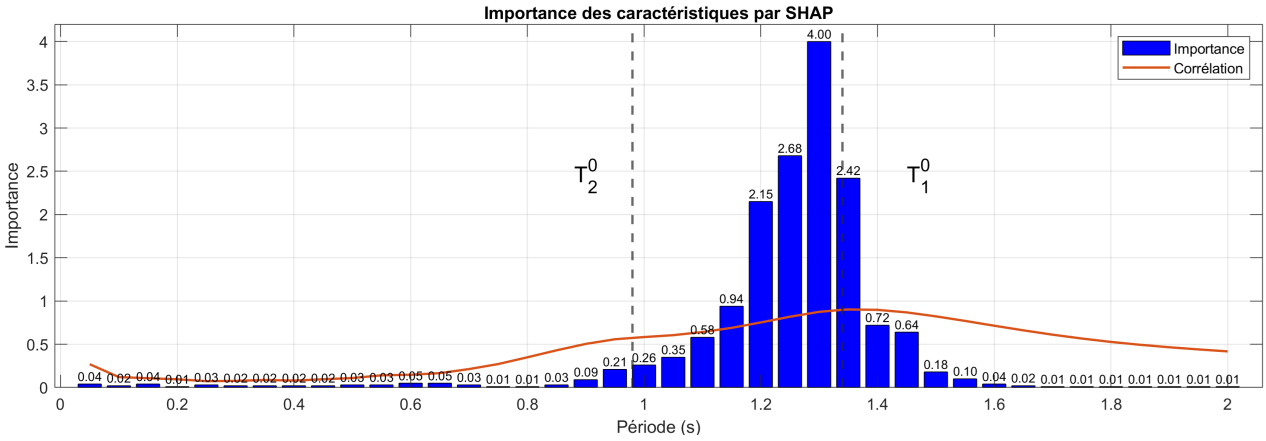


Figure 2.15 – [2ddl] Valeurs SHAP des caractéristiques du spectre de réponse en accélération

dans la plage de 1.0 s à 2.0 s sont plus importants. L'étude de SHAP aussi montre la même tendance, c'est-à-dire les caractéristiques proches des périodes propres de la structure contribue plus que les autres pour la prédiction de la réponse souhaitée.

Les résultats obtenus à partir de l'étude SHAP et de l'étude du coefficient de corrélation sont cohérents, indiquant qu'il est possible d'échantillonner les caractéristiques à partir du spectre de réponse en accélération autour des deux périodes de la structure.

e. Construction des courbes de fragilité

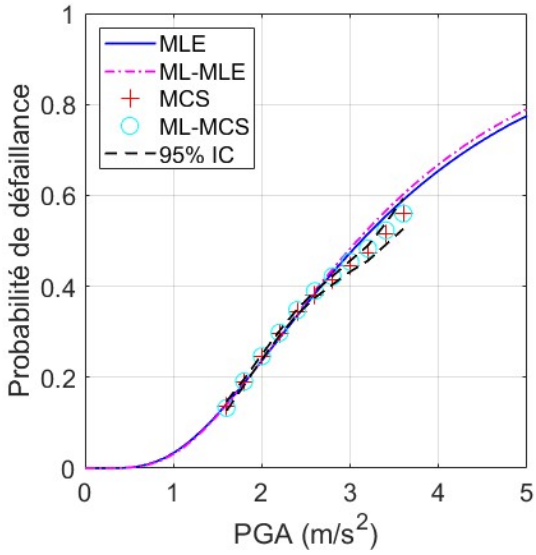
Enfin, des courbes de fragilité ont été construites en utilisant les réponses prédites par le modèle d'apprentissage automatique. Un modèle LR a été entraîné en utilisant 400 observations du spectre de réponse en accélération échantillonné sur 20 périodes, de 0.8 s (0.8 T_2^0) à 1.6 s (1.2 T_1^0). La réponse considérée de la structure était le taux de déformation maximum entre les étages.

L'état d'endommagement du système dépend du taux de déformation relative maximale entre étages. Le critère est atteint lorsque la réponse structurelle dépasse la limite critique y_0. Trois états limites y_0 ont été considérés égaux à 1 %, 1.2 % et 1.5 %, correspondants respectivement aux états de dommage mineur, modéré et grave.

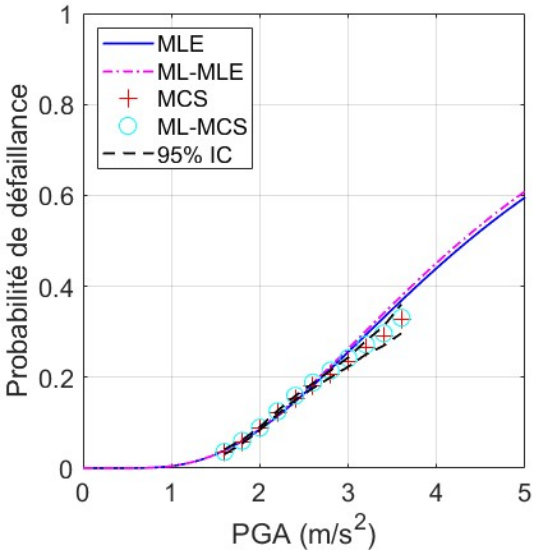
De manière similaire à l'exemple du système 1ddl, les courbes de fragilité de la structure obtenues par les méthodes MLE, ML-MLE, MCS et ML-MCS sont présentées dans la Figure 2.16 pour chaque état limite.

Les courbes de fragilité construites à l'aide de l'approche basée sur le modèle de ML sont comparées à celles construites à l'aide des méthodes conventionnelles MCS et MLE. Les résultats indiquent que les courbes de fragilité obtenues à l'aide des approches basées sur le modèle de ML sont fiables, parce qu'elles se trouvent dans l'intervalle de confiance de celles obtenues à l'aide de la méthode conventionnelle MCS.

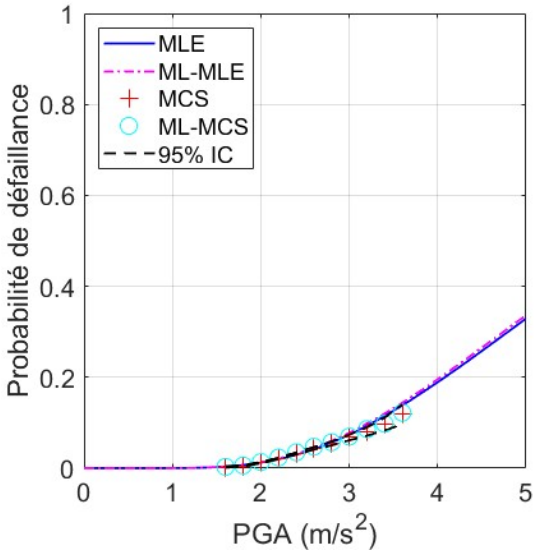
Il est important de noter que les deux courbes de MLE et ML-MLE ont été construites en utilisant 5000 observations, tandis que la méthode MCS a utilisé un plus grand nombre de 10^5 observations. Il est également important de noter que la méthode MCS conventionnelle a nécessité 10^5 simulations



(a) $y_0 = 1$ %



(b) $y_0 = 1.2$ %



(c) $y_0 = 1.5$ %

Figure 2.16 – [2ddl] Courbes de fragilité par les méthodes différentes

par éléments finis, tandis que la méthode ML-MCS n'a nécessité que 400 simulations pour l'entrainement des modèles de ML.

La comparaison de l'erreur quadratique moyenne de ces méthodes est présentée dans la Figure 2.17. Le résultat montre l'erreur quadratique moyenne pour trois états limites différents calculés par trois méthodes. L'approche ML-MLE donne une erreur plus élevée par rapport à l'approche MLE pour le premier état limite. On observe la même tendance pour l'état limite y_0 égal à 1.2 % et 1.5 %, mais la différence est plus faible. Dans les deux exemples présentés de systèmes à 1ddl et 2ddl, la méthode

ML-MCS présente de plus petites erreurs par rapport aux méthodes MLE et ML-MLE.

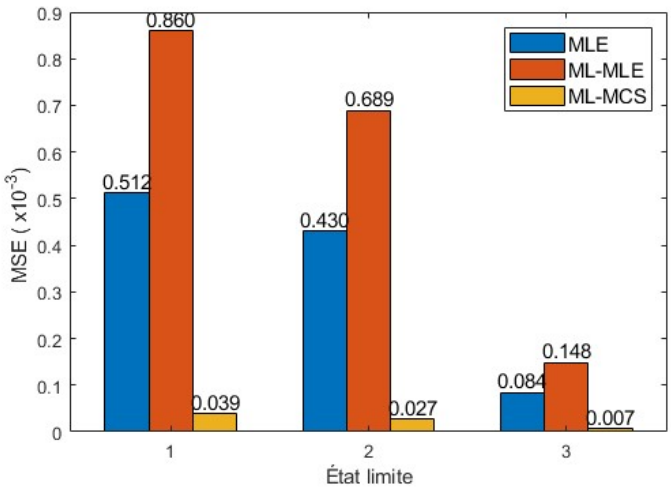


Figure 2.17 – [2ddl] Critère de comparaison pour trois états limites

2.5. Conclusion

Une nouvelle méthode d'application d'apprentissage automatique pour modéliser les réponses sismiques des structures linéaires a été proposée dans ce chapitre. Elle repose sur les spectres de réponse en accélération. Une procédure complète pas-à-pas PRO-LIN a été détaillée avec la précision en zones et en périodes à échantillonner. Elle comporte également une mesure d'importance des caractéristiques basé sur SHAP pour optimiser les modèles.

La procédure PRO-LIN a été validée avec un système à un degré de liberté et avec un système à deux degrés de liberté. Quelques remarques importantes sont à souligner.

Premièrement, le caractère novateur de la proposition réside dans l'utilisation unique du spectre de réponse en accélération échantillonné à différentes périodes comme caractéristiques des modèles d'apprentissage automatique. Il est important de noter qu'il n'est pas encore possible d'utiliser le spectre de réponse en accélération entier en tant que caractéristique continue, en raison de la surcharge de données. Il est tout à fait logique de se limiter à des zones de périodes dans le voisinage des modes propres des structures linéaires.

Deuxièmement, l'application de la procédure PRO-LIN à deux systèmes linéaires permet d'obtenir plusieurs modèles d'apprentissage automatique. La validation croisée montre que ces modèles sont capables de prédire la réponse $\mathbf{Y}$ avec les spectres de réponse en accélération comme les caractéristiques $\mathbf{x}$ des modèles de l'apprentissage automatique. Le modèle de réseau de neurone (ANN) est le plus précis pour le système à un degré de liberté tandis que le modèle de régression linéaire (LR) se montre le plus efficace en raison de sa simplicité et de sa précision pour le système à deux degrés de liberté.

Troisièmement, la précision des prédictions des modèles dépend de la taille des données d'entrainement. Pour chaque structure, une analyse de convergence a été menée afin de déterminer le nombre optimal d'observations. La structure à un degré de liberté a demandé 200 observations pour entraîner les modèles d'apprentissage automatique tandis que la structure à deux degrés de liberté a nécessité 400 observations. Cela indique que le nombre d'observations nécessaire pour entrainer des modèles peut varier en fonction de la complexité de la relation entre les caractéristiques et la réponse du système étudié.

Enfin, les courbes de fragilité sismiques basées sur l'apprentissage automatique par simulation de Monte Carlo (ML-MCS) et par estimation du maximum de vraisemblance (ML-MLE) ont été construites en utilisant les réponses prédites des modèles d'apprentissage validés. Elles sont ensuite comparées à celles obtenues par les approches conventionnelles de MCS et de MLE. Les probabilités obtenues par l'approche ML-MCS montrent une précision équivalente à celles obtenues par la méthode MCS traditionnelle. Les méthodes ML-MCS et ML-MLE à base d'apprentissage automatique montrent une amélioration par rapport aux méthodes conventionnelles par une réduction significative du temps de simulation. Pour les deux systèmes linéaires testés, la construction des courbes de fragilité sismique basée sur les modèles d'apprentissage automatique nécessite moins de 5 % du temps de simulation par rapport aux méthodes conventionnelles.

L'ensemble des résultats du chapitre confirme la validité de la méthode et de la procédure PRO-LIN proposées pour les structures linéaires. Les structures non-linéaires sont traitées dans le chapitre suivant.

3 Apprentissage automatique aux structures non-linéaires

Sommaire

3.1. Introduction

L'application de l'apprentissage automatique a été proposée pour les structures linéaires au chapitre 2 avec une nouvelle méthode et sa procédure pratique PRO-LIN pour laquelle les spectres de réponse en accélération ont été utilisés comme caractéristiques des modèles. Toutefois, sous l'excitation sismique souvent sévère, les structures présentent en général un comportement non-linéaire. L'objectif de ce chapitre est donc d'étendre la méthode déjà applicable aux structures linéaires pour les structures non-linéaires.

Les études préliminaires présentées en section 3.2 mettent en évidence la différence entre l'application des modèles d'apprentissage automatique aux structures non-linéaires et celle aux structures linéaires. En prenant toujours les spectres de réponse en accélération comme caractéristiques des modèles, les zones de périodes à considérer deviennent plus larges et la résolution en période d'échantillonnage devient plus fine. Le challenge est donc de trouver un nombre de caractéristiques raisonnable sans compromettre la précision et la performance des modèles.

La section 3.3 se concentre ensuite sur les différentes méthodes de sélection des caractéristiques des modèles d'apprentissage automatique. Les avantages et les inconvénients des méthodes sont comparés. Cette analyse permet de proposer une méthode de sélection hybride pour les spectres de réponse en accélération. La méthode hybride est introduite dans une nouvelle procédure pratique pas-à-pas PRO-NONLIN pour les structures non-linéaires.

La section 3.4 est réservée à la description de la procédure PRO-NONLIN. En commençant par un échantillonnage du spectre de réponse en accélération assez fin dans les voisinages assez étendus des modes propres, la procédure PRO-NONLIN conduit à un nombre restreint des spectres pour garantir la performance et aussi la précision des modèles d'apprentissage automatique.

La validation de la procédure PRO-NONLIN est ensuite réalisée en section 3.5. Pour cela on utilise des oscillateurs non-linéaires de Coulomb et de Bouc-Wen ainsi qu'un portique de bâtiment à 8 étages. Des courbes de fragilité sismique basées sur des modèles d'apprentissage automatique sont également établies. La section 3.6 présente une vérification étendue de la procédure PRO-NONLIN avec une étude de l'influence de la non-linéarité de Bouc-Wen.

Enfin, une conclusion sur la validité de la méthode proposée avec la procédure PRO-NONLIN est donnée.

3.2. Études préliminaires pour appliquer l'apprentissage automatique aux structures non-linéaires

Pour appliquer les modèles de machine learning aux structures non-linéaires, les études préliminaires sont nécessaires. Il faut comprendre l'influence de la non-linéarité sur l'entrainement des modèles de machine learning. Dans ce chapitre, nous étudierons deux modèles de non-linéarité de Coulomb puis celui de Bouc-Wen. Ces modèles sont très souvent utilisés dans le domaine du génie

civil. De plus, ils représentent deux types de non-linéarité : une non-linéarité d'amortissement pour le modèle de Coulomb et une non-linéarité de rigidité pour le modèle de Bouc-Wen.

Ces structures sont inspirées par le travail de Le *et al.* [10]. Les propriétés de ces oscillateurs restent similaires dans ce chapitre, seulement leur non-linéarité est variée. La réponse dans l'équation (1.1) est le déplacement maximal de la masse, $\mathbf{Y} = \max_t |y(t)|$. Cette réponse est déterminée par la méthode numérique de Runge-Kutta dans MATLAB.

3.2.1. Présentation des structures non-linéaires

a. Oscillateur de Coulomb

L'oscillateur de Coulomb, également connu sous le nom d'oscillateur à frottement sec, est un modèle d'un système oscillant non-linéaire avec frottement. Il est utilisé pour décrire le comportement d'un grand nombre de systèmes mécaniques qui présentent un frottement sec, c'est-à-dire un frottement qui ne dépend pas de la vitesse de déplacement. Le modèle est basé sur la loi de Coulomb de frottement sec, qui définit que la force de frottement est proportionnelle à la force normale appliquée au système. Le modèle est utilisé en génie mécanique et en génie civil pour analyser le comportement des structures soumises à des charges dynamiques, telles que les ponts, les bâtiments.

Dans cette étude, on considère un oscillateur de Coulomb, représenté sur La Figure 3.1. L'équation qui régit la dynamique de l'oscillateur s'écrit :

$$\ddot{y}\left(t\right) + \mu g \text{sgn}(\dot{y}\left(t\right)) + \omega_0^2 y\left(t\right) = -a(t) \tag{3.1}$$

où ω_0 est la pulsation propre non amortie. Dans le reste du chapitre, ω_0 est choisie égale à 5.97 rad/s. De plus, pour ce système, μ est le coefficient de frottement sec, g égale à 9.81 m/s^{-2} est l'accélération de la pesanteur . Pour cet oscillateur de Coulomb, le coefficient μ caractérise sa non-linéarité.

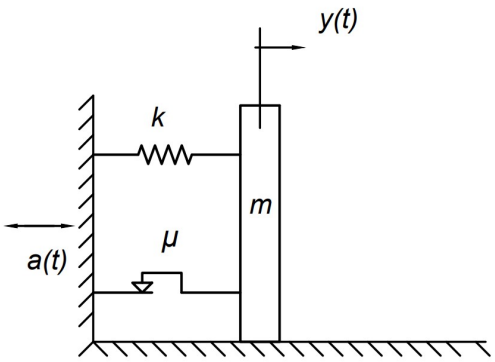


Figure 3.1 – Oscillateur non-linéaire de Coulomb

b. Oscillateur de Bouc-Wen

En ingénierie structurelle, le modèle d'hystérésis de Bouc-Wen est l'un des modèles les plus utilisés pour décrire les systèmes non-linéaires. Il a été introduit par Robert Bouc [110] et étendu par

Yi-Kwei Wen [111], qui a démontré sa polyvalence en produisant une variété de formes d'hystérésis. Ce modèle est capable de traiter, sous forme analytique, une gamme de formes de cycles correspondant au comportement d'une large classe de systèmes hystérétiques. En raison de sa polyvalence et de sa faisabilité mathématique, le modèle de Bouc-Wen est appliqué à une grande variété de problèmes d'ingénierie. Le modèle de Bouc-Wen, ses variantes et ses extensions ont été utilisées dans le contrôle des structures, en particulier dans la modélisation du comportement des amortisseurs magnéto-rhéologiques.

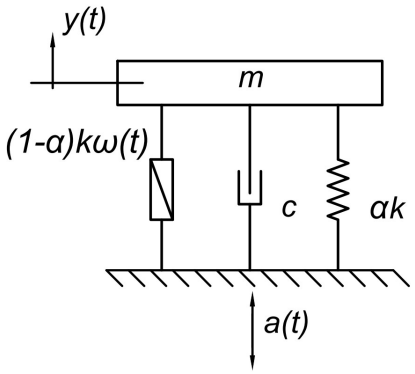


Figure 3.2 – Oscillateur non-linéaire de Bouc-Wen

Ici nous étudions un oscillateur de Bouc-Wen défini sur la Figure 3.2. L'équation de la dynamique de l'oscillateur s'écrit :

$$\ddot{y}(t) + 2\xi\omega_0\dot{y}(t) + \omega_0^2(\alpha y(t) + (1-\alpha)\omega(t)) = -a(t) \tag{3.2}$$

$$\text{avec} \quad \dot{\omega}(t) = C_1\dot{y}(t) - C_2|\dot{y}(t)||\omega(t)|^{n_d-1}\omega(t) - C_3\dot{y}(t)|\omega(t)|^{n_d} \tag{3.3}$$

où $\omega(t)$ est le déplacement d'hystérésis. Pour l'oscillateur de Bouc-Wen, la pulsation propre non-amortie ω_0 est choisie égale à 5.97 rad/s, le taux d'amortissement ξ est de 2 %. De plus, pour ce système, les constantes choisies sont : $C_1 = 1$, $C_2 = C_3 = 0.5/\mathrm{cm}^{n_d}$ et $n_d = 1$.

Pour cet oscillateur de Bouc-Wen, la non-linéarité est influencée par le coefficient α. Quand la valeur de α est égale à 1, la structure est linéaire amortie. Plus la valeur de α est petit, plus la non-linéarité apparait dans l'oscillateur de Bouc-Wen.

3.2.2. Influence de la non-linéarité sur la réponse de la structure

Dans cette section, la non-linéarité des oscillateurs varie avec le coefficient μ pour Coulomb et α pour Bouc-Wen. Un jeu de données est généré pour chaque valeur de coefficient de non-linéarité de l'oscillateur. Par exemple, 1000 signaux sismiques, générés par le modèle de Boore et préparés en avance, seront utilisés pour calculer la réponse d'un oscillateur étudié. Une observation dans le jeu de données contient les valeurs des spectres de réponse en accélération représentant le signal sismique et la réponse maximale correspondant de l'oscillateur.

a. Oscillateur de Coulomb

La non-linéarité de l'oscillateur de Coulomb est définie par le coefficient μ. Six valeurs de μ sont étudiées à 0, 0.005, 0.010, 0.015, 0.025 et 0.030. Il faut rappeler que lorsque μ est égale à 0, l'oscillateur est linéaire non amorti. 6 jeux de données contenant 1000 observations pour chaque jeu de données sont préparés.

La Figure 3.3 illustre la corrélation entre des spectres d'accélération de chacun des oscillateurs en fonction de μ et la réponse $\mathbf{Y}$. Quand la non-linéarité de la structure augmente, les coefficients de corrélation diminuent. Plus la non-linéarité de la structure est forte, moins il y a de corrélation entre les spectres d'accélération et le déplacement maximal de la structure. La valeur de corrélation maximale diminue graduellement de 1 (correspondant à μ = 0). Quand le coefficient μ est plus grand (0.030 dans cette étude), la corrélation au pic est la plus faible, égale approximativement 0.8. Bien que la corrélation diminue, ses valeurs maximales persistent à la période propre de la structure, marquée par la ligne verticale.

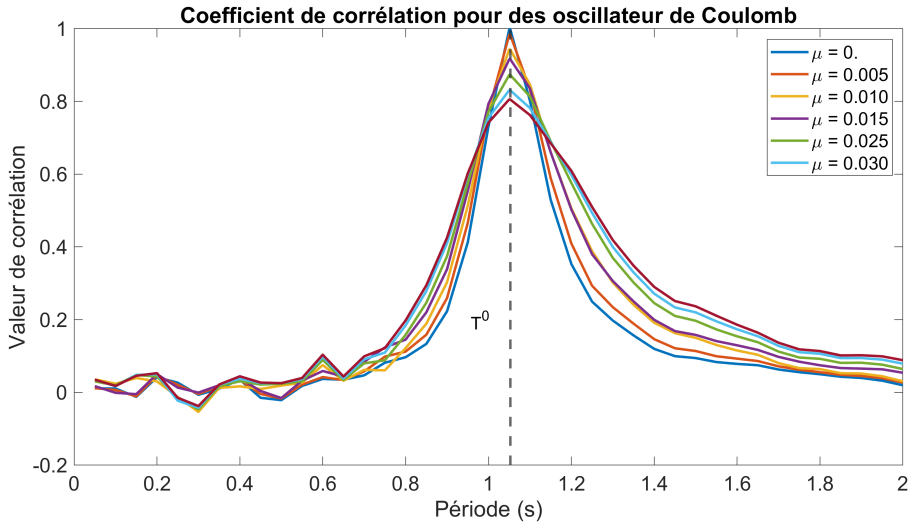


Figure 3.3 – [Coulomb] Coefficient de corrélation par rapport à μ

b. Oscillateur de Bouc-Wen

De la même manière, la non-linéarité de l'oscillateur de Bouc-Wen est examinée par la variation du coefficient α entre 0.1 et 1 avec un incrément de 0.1. Il faut aussi rappeler que lorsque α égale à 1, la structure est linéaire amortie (il n'y a que le ressort et l'amortisseur). Un jeu de données comprenant 2000 observations est généré pour chaque oscillateur. Le coefficient de corrélation entre les caractéristiques et la réponse de l'oscillateur est calculée afin de comprendre l'influence de la non-linéarité du système, présenté par La Figure 3.4. Quand α est égale à 1, la valeur maximale de corrélation est égale à 1 à la période propre de la structure. En effet, la structure devient linéaire donc :

$$S_a(T^0, \alpha = 1) \cong \omega_0^2 S_d(T^0, \alpha = 1) \tag{3.4}$$

avec le spectre en accélération à la période propre du système linéaire $S_a(T^0, \alpha = 1)$, le spectre en déplacement du même système $S_d(T^0, \alpha = 1)$ et la pulsation propre ω_0. L'équation (3.4) montre la corrélation linéaire entre le spectre en accélération à la période propre de la structure avec son déplacement maximal.

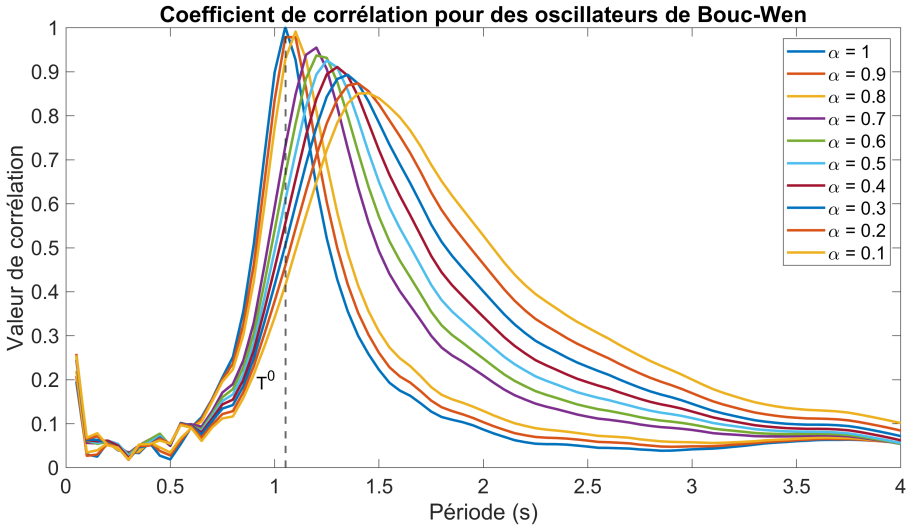


Figure 3.4 – [Bouc-Wen] Coefficient de corrélation par rapport à α

Lorsque la non-linéarité de la structure augmente, le coefficient de corrélation diminue. Plus il y a de non-linéarité, moins il y a de corrélation entre les spectres de réponse en accélération et le déplacement maximal $\mathbf{Y}$ de la structure. La valeur de corrélation maximale diminue graduellement de 1 (correspondant à α = 1). Quand le coefficient α considéré est minimal (0.1 dans cette étude), la corrélation est la plus faible, égale seulement à 0.8 approximativement. Un décalage de la position du pic de corrélation vers les périodes plus grandes que la période propre de la structure linéaire. Ce déplacement peut être expliqué par la non-linéarité de Bouc-Wen, induite par une force de rappel. La présence de comportements non-linéaires, tels que l'hystérésis, modifie la dynamique de la structure, entraînant des changements dans sa réponse temporelle.

3.2.3. Première application des modèles de machine learning

Pour ces deux oscillateurs non-linéaires, une première application est réalisée pour voir si la procédure PRO-LIN avec l'échantillonnage des valeurs spectrales autour de la période propre du système est valide. Pour ce test, des observations générées sont réutilisées pour chaque oscillateur. Les caractéristiques sont les valeurs du spectre de réponse en accélération, prises autour de la période propre du système. Pour ces deux oscillateurs, la pulsation propre non-amortie ω_0 est choisie égale à 5.97 rad/s soit la période propre T^0 égale à 1.05 s. Les caractéristiques sont initialement échantillonnées sur le spectre de réponse en accélération $S_a(T_i)$, avec T_i entre 0.1 T^0 et 2 T^0 avec un incrément de 0.1 T^0. Cet échantillonnage est approximativement entre 0.1 s et 2.0 s avec un incrément de 0.1 s. La réponse de la structure est le déplacement maximal de l'oscillateur donc $\mathbf{Y} = \max_t |y(t)|$. L'équation (3.5) représente les modèles d'apprentissage automatique :

$$\mathbf{Y} = \mathbf{ML}(\mathbf{x}) + \epsilon, \quad \mathbf{x} = \{S_a(T_i), S_a(T^0)\}, \quad T_i = [0.1 : 0.1 : 2.0] \tag{3.5}$$

Chaque jeu de donnés est utilisé pour la construction et le test des modèles d'apprentissage automatique, en allouant 80 % de la base des données à l'entraînement, tandis que les 20 % restants sont dédiés au test. Cette répartition est conforme aux standards en apprentissage automatique, permettant ainsi une évaluation objective des performances des modèles sur des données non vues lors de l'entraînement. Dans cette étude, le modèle RF est utilisé, grâce à sa facilité d'utilisation et à ses performances universelles.

Les résultats trouvés ne sont pas suffisamment concluants et ils varient en fonction de la non-linéarité du système. Les Tableaux 3.1 et 3.2 fournissent une comparaison des performances du modèle en fonction de la variation du coefficient de non-linéarité. Pour ces deux types d'oscillateur, les tableaux montrent qu'une augmentation de la non-linéarité entraîne une diminution de la performance du modèle, comme l'indiquent la baisse de R^2 et l'augmentation de $RMSE$. Pour l'oscillateur de Coulomb, quand le comportement est linéaire (μ = 0), le modèle a une performance excellente, mais cette performance se dégrade progressivement à mesure que μ augmente. Quand μ atteint 0.03, la performance du modèle est faible. Pour l'oscillateur de Bouc-Wen, la même tendance est observée.

μ	R^2	$RMSE$ (cm)
0.000	1.000	0.005
0.005	0.957	0.841
0.010	0.921	1.032
0.015	0.893	1.490
0.025	0.822	1.719
0.030	0.794	1.780

Table 3.1 – [Coulomb] Comparaison des modèles de ML en fonction de μ

α	R^2	$RMSE$ (cm)
1.0	1.000	0.041
0.9	0.977	0.298
0.8	0.964	0.495
0.7	0.942	0.766
0.6	0.891	0.860
0.5	0.847	1.238
0.4	0.812	1.485
0.3	0.751	1.762
0.2	0.663	2.037
0.1	0.576	2.126

Table 3.2 – [Bouc-Wen] Comparaison des modèles de ML en fonction de α

3.2.4. Complément d'étude sur l'application de l'apprentissage automatique

Suite à l'application décrite dans le paragraphe au-dessus, nous décidons d'étudier la sélection des caractéristiques selon la nature de l'oscillateur.

Pour l'oscillateur de Coulomb, comme sa valeur maximale de corrélation reste inchangée par rapport à la non-linéarité, une stratégie d'échantillonnage plus fine est considérée. C'est-à-dire on échantillonne le spectre de réponse en accélération $S_a(T_i)$, avec T_i de 0.05 à 2.0 s avec un incrément de 0.05 s. Cette stratégie d'échantillonnage double le nombre de caractéristiques. L'équation (3.6) représente les modèles d'apprentissage automatique :

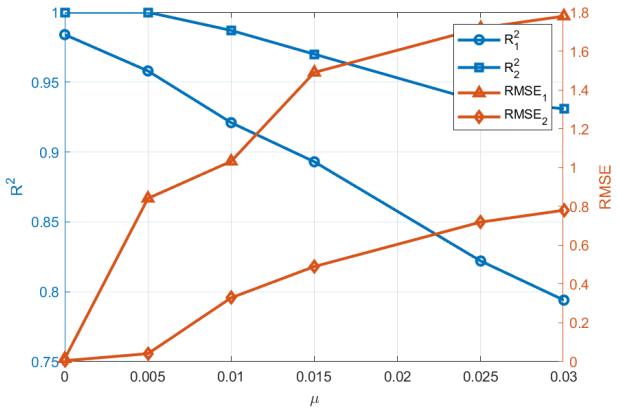
$$\mathbf{Y} = \mathbf{ML}(\mathbf{x}) + \epsilon, \quad \mathbf{x} = \{S_a(T_i), S_a(T^0)\}, \quad Ti = [0.05 : 0.05 : 2] \tag{3.6}$$

Le Tableau 3.5a compare les performances des modèles RF de chaque oscillateur de Coulomb avec ce nouvel échantillonnage. Les métriques utilisées sont le coefficient de détermination R^2 et l'erreur quadratique moyenne $RMSE$. La même tendance est observée en comparaison avec la première tentative. Avec la valeur la plus basse de μ égale à 0.005, le modèle de ML affichent un ajustement meilleur ($R^2 = 1.0$) avec une erreur $RMSE$ de 0.041. À mesure que μ augmente, les performances du modèle diminuent, indiquant une réduction de la qualité des modèles à capturer les variations dans le système à mesure que la non-linéarité μ augmente. Ces données montrent l'influence du coefficient de frottement μ sur la capacité des modèles de ML à prédire le comportement du système.

Ainsi un échantillonnage plus fine du spectre de réponse en accélération conduisant à plus de caractéristiques, donne un meilleur résultat. Ce résultat est représenté sur la Figure 3.5b. Cette figure montre une comparaison des métriques R^2 et $RMSE$ pour les deux échantillonnages. On rappelle que les critères R_2^2 et $RMSE_2$ correspondent à la deuxième stratégie d'échantillonage présentée dans l'équation (3.6), tandis que les critères R_1^2 et $RMSE_1$ correspondent à la première stratégie présentée dans l'équation (3.5). Avec l'augmentation du nombre de caractéristiques par une échantillonnage plus fin, les modèles de machine learning obtenus sont plus performants.

μ	R^2	$RMSE$ (cm)
0.000	1.000	0.005
0.005	1.000	0.041
0.010	0.987	0.329
0.015	0.970	0.490
0.025	0.936	0.719
0.030	0.931	0.780

(a) Comparaison des modèles de ML en fonction de μ



(b) Illustration de la comparaison

Figure 3.5 – [Coulomb] Influence de μ et du nombre de caractéristiques

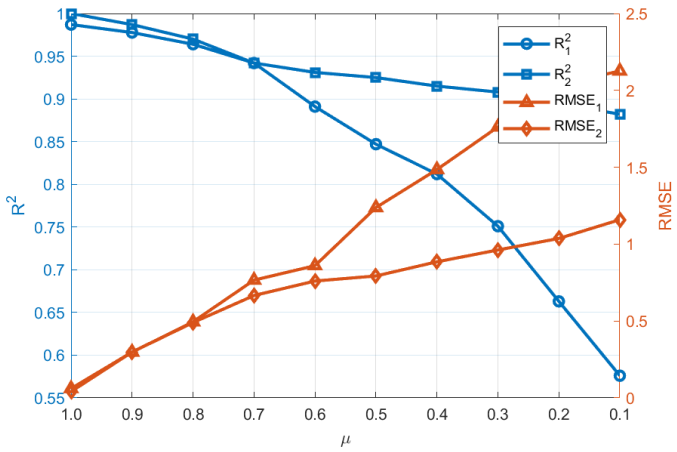
Similairement, pour l'oscillateur de Bouc-Wen, la même stratégie de l'application de machine learning est réalisée. Il faut noter que le pic de corrélation est décalé, donc, un intervalle plus grand des valeurs spectrales doit être échantillonné. Les caractéristiques se composent de spectres de réponse en accélération $S_a(T_i)$, où chaque période T_i est soigneusement échantillonnée de 0.05 T^0 à 4 T^0 par incréments de 0.05 T^0. De surcroît, la valeur du le spectre de réponse à la période propre du système $S_a(T^0)$ est pris en compte pour une meilleure modélisation de l'oscillateur. Ce choix d'échantillonnage définit 81 caractéristiques distinctes et il assure une description complète de la réponse spectrale de la structure. L'équation (3.7) représente le modèle d'apprentissage automatique :

$$\mathbf{Y} = \mathbf{ML}(\mathbf{x}) + \epsilon, \quad \mathbf{x} = \{S_a(T_i), S_a(T^0)\}, \quad T_i = [0.05 : 0.05 : 4] \tag{3.7}$$

Le Tableau 3.6a présente les résultats de comparaison des modèles RF pour différents α de l'oscillateur de Bouc-Wen. Une variation des performances des modèles en fonction de la valeur de α est constatée. Encore une fois, la même tendance est observée. Lorsque α est proche de 1.0, le coefficient de détermination R^2 est approximativement de 1.0, ce qui indique une excellente qualité du modèle. De plus, l'erreur quadratique moyenne est minimale, atteignant une valeur de 0.041. Cela traduit une excellente précision dans la prédiction des résultats par rapport aux données observées. A mesure que la non-linéarité augmente, une diminution du R^2 ainsi qu'une augmentation de $RMSE$ sont observées. Ce tableau montre encore une moindre qualité du modèle aux données, avec des performances prédictives moins précises sous l'influence de la non-linéarité. Cette tendance est aussi illustrée par la Figure 3.6b.

μ	R^2	$RMSE$ (cm)
1.0	1.000	0.041
0.9	0.987	0.298
0.8	0.970	0.490
0.7	0.942	0.666
0.6	0.931	0.560
0.5	0.925	0.793
0.4	0.915	0.885
0.3	0.908	0.962
0.2	0.895	1.037
0.1	0.882	1.159

(a) Comparaison des modèles de ML en fonction de α



(b) Illustration de la comparaison

Figure 3.6 – [Bouc-Wen] Influence de μ et du nombre de caractéristiques

En résumé, ce tableau met en évidence l'effet de la non-linéarité sur les performances des modèles de machine learning. Par contre, l'augmentation du nombre de caractéristiques par un échantillonnage plus fin du spectre de réponse en accélération sur un intervalle plus grand de période donne un meilleur résultat, en particulier pour une non-linéarité plus importante. Cela signifie que ces nouvelles caractéristiques capturent mieux la réponse $\mathbf{Y}$.

3.2.5. Discussions

Les différentes applications réalisées sur les oscillateurs de Coulomb et de Bouc-Wen mettent en évidence des résultats importants.

Premièrement, la non-linéarité influence la performance du modèle de ML. Pour le système de Coulomb, le coefficient de frottement μ a une forte influence sur les performances des modèles de ML. Plus la non-linéaire est importante, plus les performances des modèles diminuent. Concernant le système de Bouc-Wen, l'étude menée en fonction du coefficient α a également montré cette tendance. Cela indique que la non-linéarité introduite dans l'oscillateur de Coulomb et de Bouc-Wen impacte significativement la qualité des prédictions des modèles.

Deuxièmement, il est à noter que la non-linéarité influence aussi le nombre des caractéristiques. La comparaison des métriques R^2 et $RMSE$ pour les deux échantillonnages du spectre de réponse utilisés, les Figures 3.5b et 3.6b montrent l'influence du nombre de caractéristiques sur la réponse $\mathbf{Y}$. L'augmentation du nombre de caractéristiques est obtenue par l'élargissement de l'intervalle d'étude et la réduction de l'incrément de période.

Cependant, il convient de souligner que le nombre de caractéristiques utilisées pour élaborer ces modèles est excessive. Il est remarqué que le nombre de caractéristiques n'est pas optimisé. Il s'avère donc impératif d'intégrer une phase de sélection des caractéristiques avant la construction des modèles d'apprentissage automatique. Cette étape cruciale vise à optimiser le nombre de caractéristiques en ne conservant que celles les plus pertinentes. Cette sélection des caractéristiques peut améliorer la performance des modèles tout en réduisant la complexité computationnelle et le risque de surajustement.

En résumé, ces analyses mettent en évidence l'influence significative des coefficients de non-linéarité, plus spécifiquement μ et α, sur les performances des modèles de ML dans la prédiction de $\mathbf{Y}$ des oscillateurs non-linéaires de Coulomb et de Bouc-Wen. Une augmentation du nombre des caractéristiques est nécessaire pour pouvoir construire des modèles de machine learning suffisamment performants. Par contre, cela nécessite une phase de sélection des caractéristiques, présentée dans le paragraphe suivant.

3.3. Sélection des caractéristiques

3.3.1. Différentes méthodes de sélection des caractéristiques

La sélection des caractéristiques est une étape cruciale dans le processus d'analyse de données et de construction des modèles de machine learning. Elle consiste à identifier les caractéristiques les plus pertinentes et informatives pour résoudre un problème donné. Cette étape permet de réduire le temps et les ressources informatiques par une gestion de la surcharge des informations. Il existe trois méthodes principales pour la sélection des caractéristiques selon les références [112, 113, 114] que nous détaillons dans les sous sections suivantes.

a. Méthode de filtrage

La méthode de filtrage (filter en anglais) est l'une des méthodes couramment utilisées pour la sélection des caractéristiques. Elle consiste à évaluer les caractéristiques individuellement, indépendamment du modèle d'apprentissage, en utilisant des mesures statistiques ou des tests de corrélation. Elle fonctionne généralement en utilisant les techniques suivantes :

- Mesure de corrélation : la corrélation entre chaque caractéristique et la variable cible peut être calculée. Des mesures couramment utilisées sont le coefficient de corrélation pour les caractéristiques numériques et le test du chi-carré pour les caractéristiques catégorielles. Les caractéristiques fortement corrélées avec la variable cible sont considérées comme importantes.
- Test de signification : Des tests statistiques peuvent être employés pour évaluer si la distribution des valeurs d'une caractéristique diffère significativement entre les différentes classes de la variable cible. Des tests de corrélation, f-régression, ou des tests non paramétriques peuvent être utilisés en fonction de la nature des données. Les caractéristiques qui présentent une différence significative sont considérées comme importantes.
- Sélection basée sur la variance : La variance est calculée entre les caractéristiques et la réponse cible. La variance inférieure au seuil choisi est considérée comme peu informative et les caractéristiques correspondantes peuvent être exclues.

L'avantage de la méthode de filtrage est sa simplicité et sa rapidité, car elle ne nécessite pas d'entraîner un modèle de machine learning. Cependant, elle ne tient pas compte des interactions entre les caractéristiques et elle peut ne pas être optimale pour des problèmes complexes. De plus, elle peut exclure des caractéristiques potentiellement importantes si elles ne présentent pas de corrélation ou de relation évidente avec la variable cible.

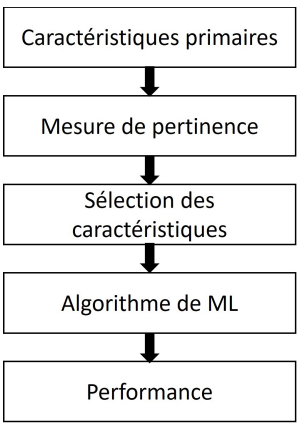


Figure 3.7 – Procédure de sélection des caractéristiques par la méthode de filtrage

Les étapes, illustrées sur la Figure 3.7, sont les étapes de base de cette procédure :

- Calcul des mesures de pertinence : Les caractéristiques sont évaluées individuellement en utilisant des mesures de pertinence, par exemple la corrélation. Ces mesures quantifient la relation entre chaque caractéristique et la variable cible.

- Sélection des caractéristiques : Les caractéristiques sont classées en fonction de leur pertinence, généralement de manière descendante. Les caractéristiques les plus pertinentes sont considérées comme les plus importantes pour le modèle. Un seuil est choisi pour sélectionner les caractéristiques à conserver ou à réduire. Les caractéristiques ayant une pertinence supérieure au seuil sont sélectionnées et les autres sont éliminées.
- Construction du modèle : Le modèle d'apprentissage automatique est construit en utilisant uniquement les caractéristiques sélectionnées. Cela réduit la dimension du problème et peut améliorer les performances du modèle en réduisant le bruit et la redondance des caractéristiques.
- Évaluation du modèle : Finalement, la performance du modèle est évaluée pour le valider ou non.

b. Méthode d'enveloppe

La méthode d'enveloppe (wrapper en anglais) est une méthode de sélection des caractéristiques pour un modèle de machin learning. Cette méthode consiste à évaluer différents sous-ensembles de caractéristiques en entraînant et en évaluant avec ces sous-ensembles pour chaque itération.

L'avantage de la méthode d'enveloppe est qu'elle prend en compte les interactions entre les caractéristiques et elle peut identifier les sous-ensembles optimaux qui conduisent à de meilleures performances prédictives. Elle permet également de prendre en compte des critères spécifiques de sélection des caractéristiques, tels que la complexité du modèle ou la stabilité. Cependant, la méthode d'enveloppe peut être coûteuse en termes de temps de calcul, en particulier lorsque l'ensemble de données est volumineux ou que le nombre de caractéristiques est élevé. De plus, elle peut être sensible au surajustement si le nombre d'observations est limité par rapport au nombre de caractéristiques. Il est donc important de prendre des mesures pour éviter le surajustement, telles que l'utilisation de techniques de validation croisée.

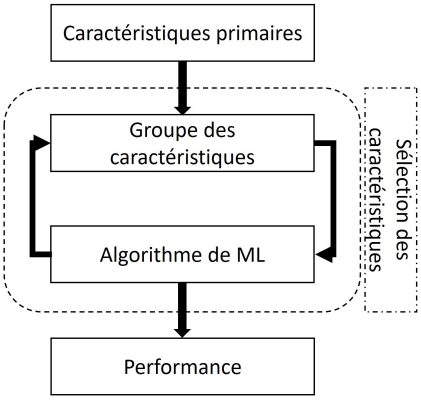


Figure 3.8 – Procédure de sélection des caractéristiques par par la méthode d'enveloppe

Cette méthode suit la procédure représentée sur la Figure 3.8 pour sélectionner les caractéristiques :

- Génération des sous-ensembles : On génère différents sous-ensembles de caractéristiques à partir de l'ensemble de données initial. Cela peut être fait de manière exhaustive en considérant tous les sous-ensembles possibles, ou de manière heuristique en utilisant des algorithmes de recherche, tels que la recherche en avant, la recherche en arrière, ou la recherche bidirectionnelle.
- Entraînement des modèles : Pour chaque sous-ensemble de caractéristiques, on entraîne un modèle prédictif en utilisant un algorithme d'apprentissage approprié. Cela peut être un algorithme de régression, de classification ou toute autre méthode d'apprentissage automatique en fonction du problème.
- Évaluation des performances : On évalue les performances de chaque modèle en utilisant des métriques d'évaluation telles que R^2, $RMSE$, ou d'autres mesures appropriées en fonction du problème. Les modèles qui obtiennent les meilleures performances sont considérés comme les plus prometteurs.
- Sélection du meilleur sous-ensemble : On sélectionne le sous-ensemble de caractéristiques qui produit le meilleur modèle prédictif en fonction des performances évaluées. Ce sous-ensemble est considéré comme le plus approprié pour la prédiction du problème.
- Évaluation du modèle : Finalement, la performance du modèle est évaluée pour le valider ou non.

c. Méthode intégrée

La méthode intégrée consiste à incorporer le processus de sélection des caractéristiques directement dans l'algorithme d'apprentissage automatique. Contrairement aux méthodes de filtrage et d'enveloppe qui sont appliquées en amont ou en aval de l'apprentissage, la méthode intégrée évalue et sélectionne les caractéristiques pendant le processus d'apprentissage. Les caractéristiques sont évaluées et attribuées à une métrique automatiquement pendant l'apprentissage, en fonction de leur contribution au modèle. Les caractéristiques jugées moins importantes peuvent être pénalisées ou éliminées, tandis que celles jugées plus importantes sont favorisées.

Les avantages de la méthode intégrée résident dans sa capacité à sélectionner automatiquement les caractéristiques pertinentes sans nécessiter d'étapes supplémentaires de sélection. Elle peut également prendre en compte les interactions entre les caractéristiques et fournir des modèles plus stables et plus performants. Cependant, la méthode intégrée peut présenter quelques inconvénients. Elles dépendent des modèles de machine learning et elles ne sont disponibles que pour quelques modèles seulement, par exemple pour le modèle RF. Elles peuvent être aussi sensibles aux paramètres de l'algorithme et peuvent avoir des biais inhérents dans la sélection des caractéristiques.

Le processus de sélection des caractéristiques est le suivant :

- Entrainement du modèle de machine learning : Le modèle est entraîné sur les données d'apprentissage. Pendant l'entraînement, le modèle attribue une importance pour chaque caractéristique.
- Sélection des caractéristiques : Les caractéristiques sont triées en fonction de leur importance. Une boucle est nécessaire pour réduire progressivement leur nombre. Pour chaque itération, l'importance entre les caractéristiques par rapport la variable cible est calculée, et un certain pourcentage des caractéristiques les moins importantes est supprimé. Le modèle est entraîné

de nouveau avec les caractéristiques restantes et leur performance est évaluée à l'aide du score pour chaque itération.

- Obtention du modèle : Le processus continue jusqu'à ce que la performance du modèle soit stable.

La procédure de sélection des caractéristiques par cette méthode est présentée dans la Figure 3.9. La boucle de sélection des caractéristiques est à l'intérieur de l'entrainement du modèle, c'est-à-dire il faut entrainer le modèle pour commencer la sélection des caractéristiques. La méthode intégrée attribue un score à toutes les caractéristiques initialement choisies après cet entrainement.

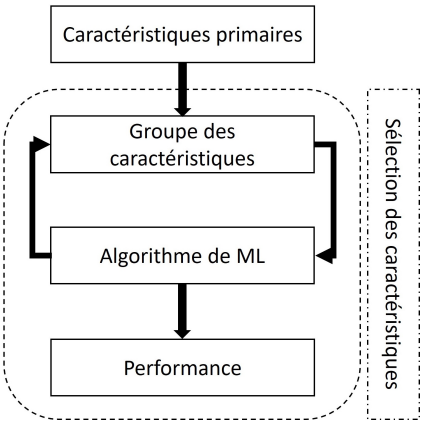


Figure 3.9 – Procédure de sélection des caractéristiques par la méthode intégrée

3.3.2. Application de la sélection des caractéristiques

Comme explicité, l'objectif de la sélection des caractéristiques est d'identifier les caractéristiques les plus pertinentes et informatives, afin d'améliorer la précision, la simplicité, la transparence et l'efficacité des modèles de machine learning. Cela permet de mieux comprendre les relations entre les caractéristiques $\mathbf{x}$ et la réponse $\mathbf{Y}$.

Une application est réalisée pour montrer les forces et les lacunes de chaque méthode de sélection des caractéristiques. En observant les résultats, les caractéristiques les plus pertinentes pour construire un modèle de machine learning correspondant sont identifiées. Finalement, une méthode hybride pour sélectionner les caractéristiques pour ce cas d'études est proposée, en combinant les méthodes de filtrage et d'enveloppe. Cette combinaison correspond bien aux perspectives d'avenir de la sélection des caractéristiques, selon Venkatesh *et al.* [114].

Pour cette application, un oscillateur non-linéaire de Bouc-Wen est considéré. À titre d'illustration, la non-linéarité moyenne, décrit par le coefficient α égale à 0.7, est choisie. Le coix de la non-linéarité moyenne est motivée par le fait que le nombre de caractéristiques nécessaires aux modèles de ML est suffisamment faible. Cela facilite la comparaison des méthodes de sélection des caractéristiques.

Nous utilisons la stratégie développée dans le paragraphe précédent, qui conduit à un jeu de données contenant 81 caractéristiques, $\mathbf{x} = \{S_a(T_i), S_a(T^0)\}$, avec $Ti = [0.05 : 0.05 : 4]$. La réponse à prédiction est le déplacement maximal, alors $\mathbf{Y} = \max_t |y(t)|$.

Dans cette application, trois modèles de prédiction sont utilisés, qui sont la forêt aléatoire (RF), le réseau de neurones (ANN) et la machine à vecteur de support (SVMR).

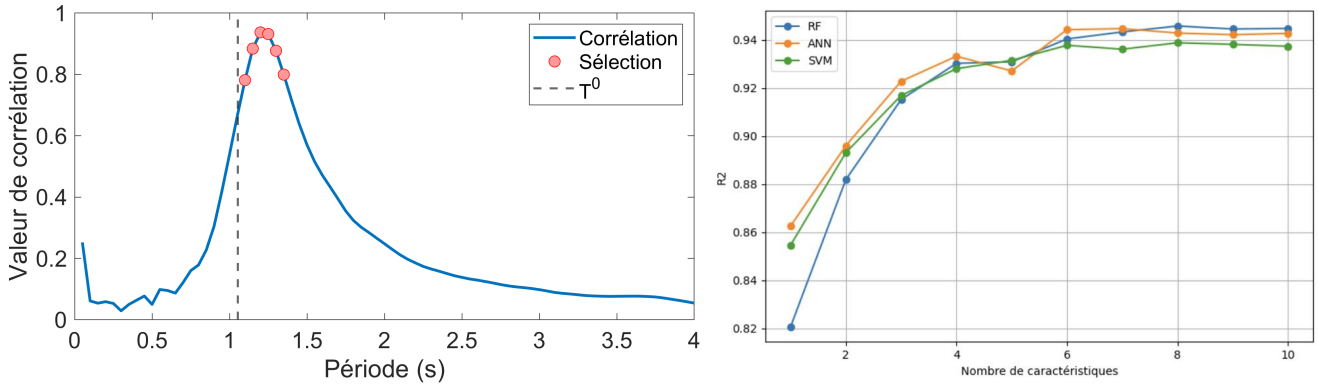
a. Méthode de filtrage

Dans le contexte de l'utilisation de la méthode de filtrage, une mesure de pertinence, qui est la corrélation, est utilisée. Pour rappeler, la corrélation entre une caractéristique quelconque, désignée x_i, dans l'ensemble des caractéristiques $\mathbf{x}$ avec la réponse $\mathbf{Y}$, est calculée avec la formule (2.7).

$$r_{x_iY} = \frac{\sum_j (x_i^j - \bar{x}_i)(Y^j - \bar{Y})}{\sqrt{\sum_j (x_i^j - \bar{x}_i)^2 \sum_j (Y^j - \bar{Y})^2}}$$

La Figure 3.10a illustre les valeurs de corrélation normalisées entre $\mathbf{x}$ et $\mathbf{Y}$. Elle montre que la corrélation maximale est à proximité de la période propre du système linéaire à 1.05 s, marquée par la ligne verticale.

Pour cette méthode, après avoir calculé la mesure de pertinence, les caractéristiques les moins importantes sont enlevées dans l'ensemble total. Dans cette application, pour chaque itération, 5 % des caractéristiques entre eux sont enlevées. Malgré l'utilisation de modèles de machine learning différents, pour chaque itération, les caractéristiques enlevées sont similaires (selon l'ordre calculé par corrélation). L'application parcours toutes les caractéristiques. Les performances, la valeurs de R^2 dans ce cas, sont calculées pour chaque modèle de prédiction pour chaque itération et puis illustrées dans la Figure 3.10b.



(a) Valeurs normalisées de corrélation **(b)** R^2 des modèles selon sélection

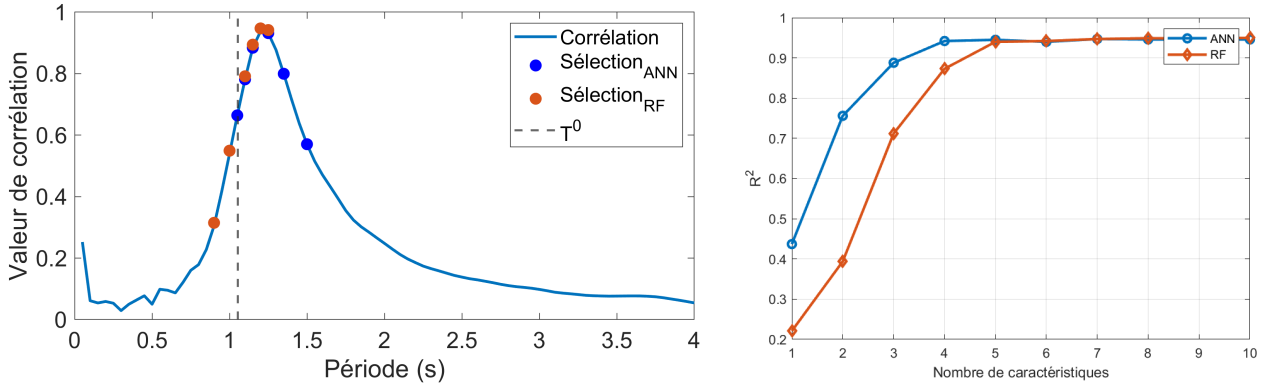
Figure 3.10 – [Filtrage] Sélection par la méthode de filtrage

Le graphique montre la variation du score R^2 pour quatre algorithmes choisis au fil des itérations en rajoutant des caractéristiques. Pour les premières itérations, la performance du modèle n'est pas encore stable et elle s'améliore en ajoutant plus de caractéristiques. Cela indique qu'il faut un nombre suffisant des caractéristiques pour obtenir un modèle de ML bien entrainé. Dans ce cas, à partir de 6 caractéristiques, les erreurs des modèles sont stables.

Le modèle de prédiction de type RF montre la stabilité de performance, avec la valeur de R^2 reste supérieure à 0.94. Les modèles de type ANN, SVMR le suivent respectivement. La caractéristique la plus importante selon cette méthode est la valeur de spectre de réponse en accélération au pic de la corrélation. Les caractéristiques aux itérations finales à choisir sont autour de cette caractéristique, comme illustrée par la Figure 3.10a. Cette sélection est systématique et ne prend pas en compte le modèle de ML à utiliser.

b. Méthode d'enveloppe

Pour la méthode d'enveloppe, deux modèles de type ANN et RF sont utilisés. Le nombre de caractéristiques pour la sélection est exploré entre 1 et 10. La Figure 3.11 montre la variation de R^2 après la sélection des caractéristiques par cette méthode pour deux modèles de prédiction. A partir de 6 caractéristiques, la performance des modèles reste similaire, autour de la valeur de 0.95.



(a) Sélection des caractéristiques

(b) R^2 des modèles selon sélection

Figure 3.11 – [Enveloppe] Sélection par la méthode de filtrage

Pour atteindre cette performance, ces modèles de machine learning utilisent des caractéristiques différentes, comme le montre la Figure 3.11a. En fait, la méthode d'enveloppe pour sélectionner des caractéristiques est sensible aux modèles de machine learning utilisés. La sélection est non seulement différente entre les modèles mais aussi différente entre les méthodes de ML. Cette différence est mise en évidence par la comparaison de la Figure 3.10b et 3.11b. La méthode enveloppe non seulement sélectionne les caractéristiques dont la corrélation avec $\mathbf{Y}$ est importante, mais aussi celles qui sont moins corrélées.

La sélection des caractéristiques pour les modèles RF et ANN montre des ensembles partiellement différents de périodes de spectre de réponse en accélération (T_i) considérées comme importantes pour la prédiction.

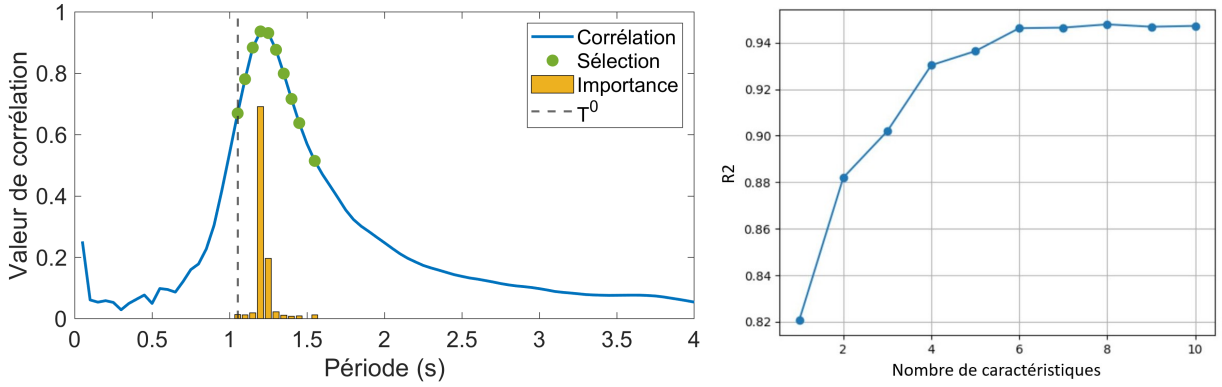
La différence entre les sélections entre ces deux modèles peut être dues à la manière dont chaque modèle traite les informations. Les RF sont connus pour leur capacité à capturer des interactions non-linéaires et des hiérarchies dans les données, tandis que les ANN sont capables de modéliser des relations complexes et abstraites grâce à leur architecture de neurones, de couches multiples et de lien entre eux.

La présence de caractéristiques communes suggère qu'il y a des caractéristiques intrinsèquement informatives pour les deux modèles, malgré leurs différentes capacités de modélisation.

c. Méthode intégrée

Pour la sélection des caractéristiques selon la méthode intégrée, le modèle RF est employé. Ce modèle est choisi parce que la méthode intégrée est disponible pour cette méthode sans algorithme supplémentaire. Le nombre de caractéristiques pour la sélection est exploré entre 1 et 10 comme précédemment. Afin de sélectionner selon cette méthode, le modèle RF est entrainé avec la totalité des caractéristiques. Ce modèle évalue et propose une mesure d'importance des caractéristiques selon la majorité des arbres dans le modèle. La sélection est donc réalisée selon ce calcul. C'est la raison pour laquelle les modèles comme ANN et SVMR ne sont pas utilisées dans ce cas, comme ils ne possèdent pas de mesure d'importance intégrée.

La Figure 3.12 montre le résultat de la sélection des caractéristiques par la méthode intégrée avec le modèle RF. A partir de 6 caractéristiques sélectionnées, le score de R^2 obtenu devient presque constant, à 0.95 comme pour les autres cas. Les caractéristiques choisies pour chaque itération sont aussi notées à coté de leur importance pour le modèle dans la Figure 3.12. Cette figure illustre l'importance des caractéristiques calculée à l'intérieur du modèle RF pour donner plus de raison pour ces sélections. Le spectre de réponse en accélération à la période propre, marquée par la ligne pointillée, présente une valeur d'importance faible. Cela montre que la caractéristique $S_a(T^0)$ n'est pas la plus importante pour tous les cas.

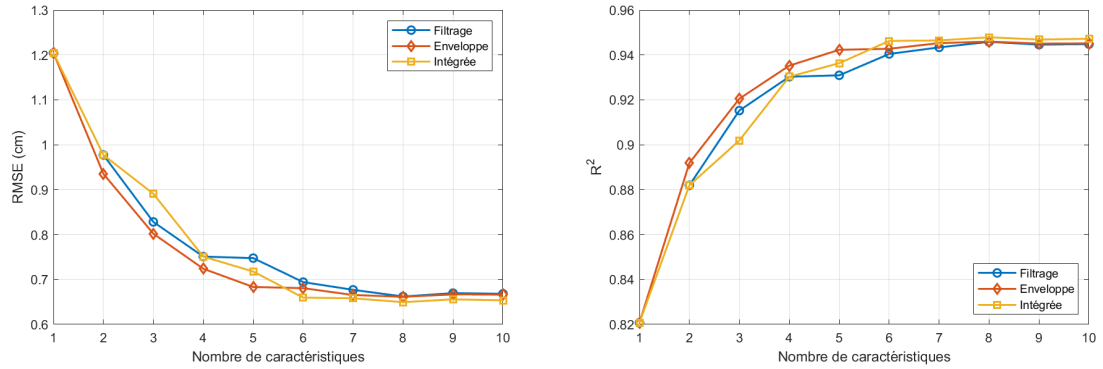


(a) Importance des caractéristiques **(b)** Variation de R^2

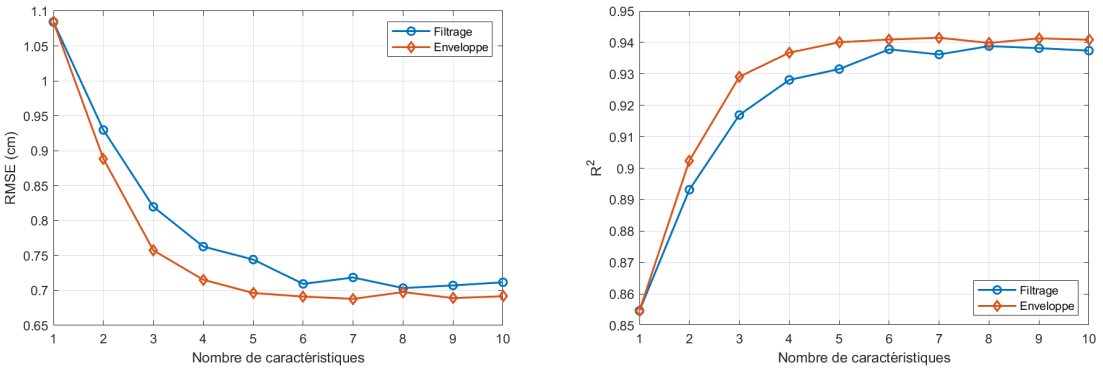
Figure 3.12 – [Intégrée] Résultat de la sélection des caractéristiques récursive

d. Comparaison des méthodes de sélection

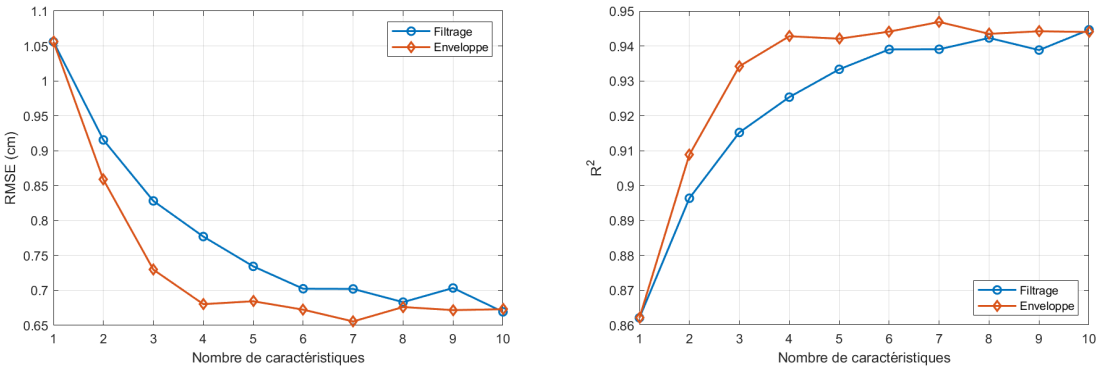
Un bilan des études de sélection des caractéristiques pour cet oscillateur est réalisé. Le nombre de caractéristiques varie de 1 jusqu'à 10. Trois modèles de machine learning (RF, SVMR et ANN) sont considérés. Les critères d'évaluation sont $RMSE$ et R^2. La Figure 3.13 compare les méthodes de sélection par rapport au nombre de caractéristiques. Il faut noter que la méthode intégrée est seulement disponible pour le modèle RF.



(a) Forêt aléatoire



(b) Machine à vecteur de support



(c) Réseau de neurones

Figure 3.13 – Évolution de $RMSE$ (gauche) et R^2 (droit) des modèles de ML

Ces comparaisons montrent une tendance similaire entre les méthodes de sélection et des modèles de ML. A la première itération, les critères d'évaluation sont similaires, c'est-à-dire les méthodes choisissent la même caractéristique pour le départ. Pour ce système de Bouc-Wen, les caractéristiques autour du pic de la courbe de corrélation sont les plus importantes. Ensuite, la méthode d'enveloppe montre la capacité supérieure par rapport aux autres. Cela est compréhensible parce que

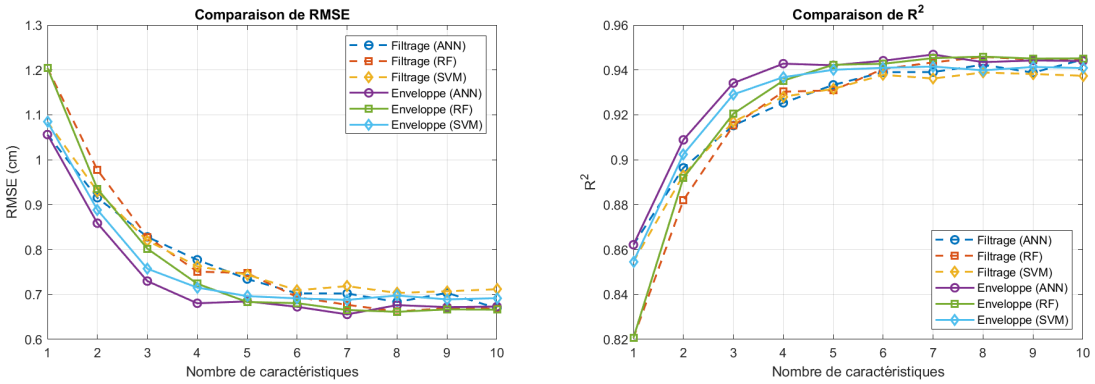
cette méthode a choisi les sous-ensembles les plus efficaces pour l'entrainement du modèle. A partir de six caractéristiques, la performance des modèles est presque similaire et reste inchangée. C'est-à-dire que six caractéristiques est le nombre optimal pour entrainer des modèles de ML pour cet exemple.

Les trois méthodes de sélection des caractéristiques fonctionnent de manière très différente pour donner leur choix. Le choix par la méthode de filtrage est donné immédiatement, parce qu'il est fait selon la mesure de pertinence. Le choix par la méthode intégrée demande plus de temps, parce qu'il faut entrainer une fois le modèle pour trouver l'importance des caractéristiques. Cette importance est la base de la sélection. La méthode d'enveloppe, au contraire, demande plus de temps. Pour sélectionner un nombre souhaité de caractéristiques avec un modèle de machine learning choisi, il faut entrainer ce modèle avec tous les choix un par un. C'est la raison pour laquelle le temps nécessaire pour cette méthode varie avec le nombre de caractéristiques et le modèle à utiliser. Pour ce cas d'exemple, le modèle ANN prend plus de temps, suivi par le modèle de RF et SVMR.

La Figure 3.14 compare la performance des modèles de ML pour ce jeu de données. Comme la méthode intégrée n'est disponible que pour le modèle RF, elle n'est pas représentée sur le figure. Les deux graphiques présentent l'évolution des performances de trois modèles de machine learning différents (RF, SVMR et ANN) en fonction du nombre de caractéristiques sélectionnées par deux méthodes de sélection de caractéristiques : la méthode de filtrage (illustrée par les courbes en trait pointillé) et la méthode d'enveloppe (illustrée par les courbes en trait continu). Les modèles construits par la méthode de filtrage et par la méthode d'enveloppe montrent des tendances similaires, avec une baisse rapide du $RMSE$ qui se stabilise pour un certain nombre de caractéristiques. Le modèle ANN semble offrir les meilleures performances (le plus faible $RMSE$), surtout avec la méthode d'enveloppe, suivie de près par le RF et SVMR. L'indice de R^2 augmente avec le nombre de caractéristiques, présenté par les courbes se rapprochant de 1 puis se stabilisant. Cette tendance suggère qu'ajouter plus de caractéristiques au-delà d'un certain point n'améliore plus significativement la qualité du modèle.

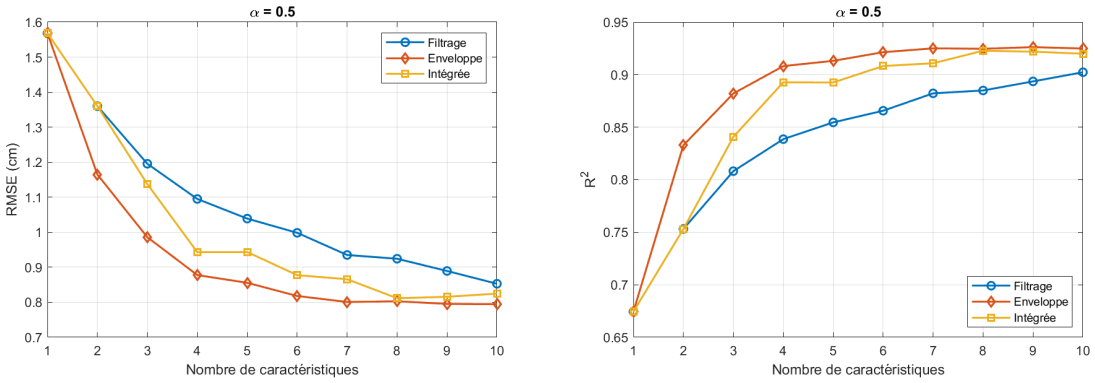
Généralement, les courbes se stabilisent avec seulement la sélection de 10 % des caractéristiques dans cette étude (8 par rapport à 81 caractéristiques initialement), indiquant que peu de caractéristiques sont nécessaires pour atteindre une performance optimale. Chaque méthode sélectionne une combinaison différente pour entrainer les modèles. En tout cas, la méthode d'enveloppe est la méthode la plus performante, spécialement pour les premières itérations. Cependant, cette méthode est la plus couteuse en temps.

Les autres valeurs de α sont aussi considérées pour voir la performance de chaque méthode par rapport à la non-linéarité de Bouc-Wen. Le modèle RF est utilisé pour pouvoir comparer les trois méthodes de sélection parce que ces méthodes ne sont pas fortement influencées par les modèles de ML. Les Figures 3.15, 3.16 et 3.17 présentent le nombre de caractéristiques pour les cas où α est respectivement égale à 0.5, 0.3 et 0.1. Plus l'oscillateur de Bouc-Wen est non-linéaire, plus les performances des méthodes sont différentes. Le nombre de caractéristiques augmente en fonction de la non-linéarité. La méthode de filtrage ne présente plus sa capacité de trouver un modèle RF dont la performance converge. Tandis que la méthode d'enveloppe et la méthode intégrée graduellement trouvent le nombre suffisant de caractéristiques. Malgré que la méthode de filtrage soit facile à réaliser et qu'elle ne soit pas couteuse en temps, le nombre élevé de caractéristiques n'est pas optimal pour les modèles de ML.



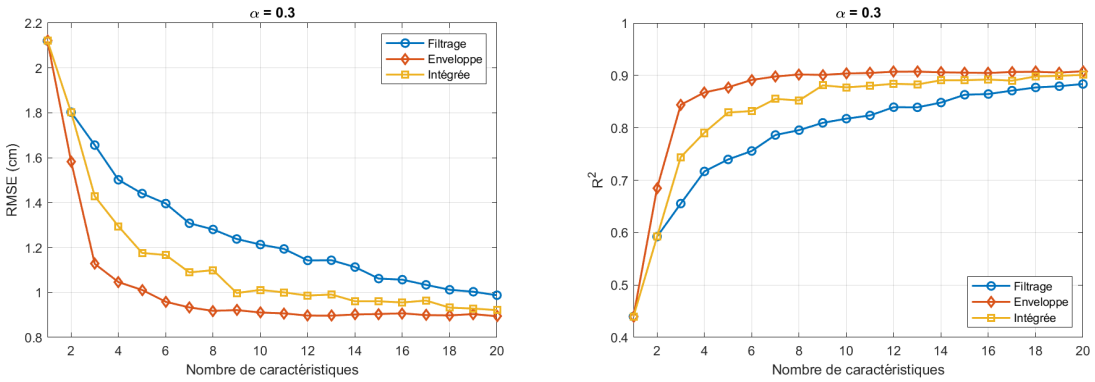
(a) $RMSE$ **(b)** R^2

Figure 3.14 – Évolution de $RMSE$ et R^2 selon les méthodes de sélections et les modèles de ML



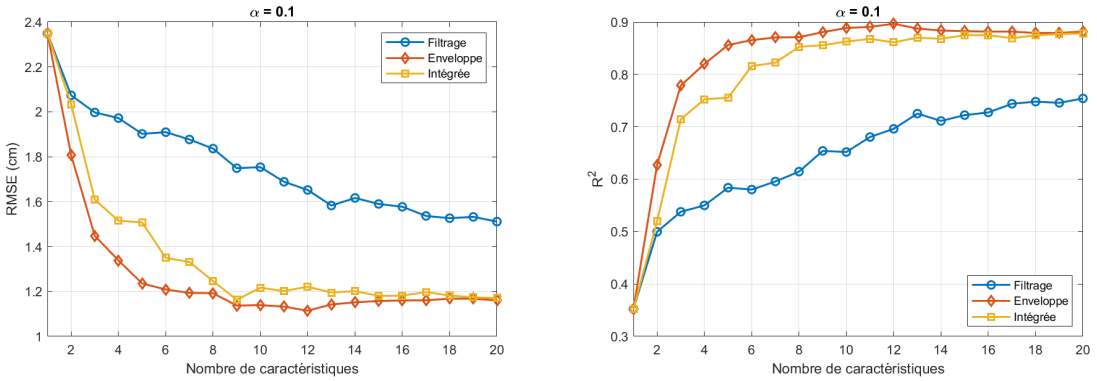
(a) $RMSE$ **(b)** R^2

Figure 3.15 – [α = 0.5] Évolution de $RMSE$ et R^2 des méthodes de sélection



(a) $RMSE$ **(b)** R^2

Figure 3.16 – [α = 0.3] Évolution de $RMSE$ et R^2 des méthodes de sélection



(a) $RMSE$ **(b)** R^2

Figure 3.17 – [α = 0.1] Évolution de $RMSE$ et R^2 des méthodes de sélection

Le Tableau 3.3 résume les points forts et les points faibles de chaque méthode de sélection des caractéristiques.

Méthode	**Description**		
	Filtrage	**Enveloppe**	**Intégrée**
Implémentation	Ne fait pas appel à un algorithme d'apprentissage automatique spécifique.	Effectue en observant chaque itération de la phase d'entraînement du modèle.	Évalue sur un algorithme spécifique d'apprentissage automatique.
Avantages	Simple à mettre en œuvre, calcul rapide, peut éliminer les caractéristiques redondantes ou peu informatives.	Prend en compte les interactions entre les caractéristiques, peut être adaptée aux modèles complexes.	Fournit une mesure directe de l'importance des caractéristiques.
Désavantages	Ne prend pas en compte les interactions entre les caractéristiques, moins efficace pour des modèles complexes.	Plus coûteuse en termes de temps de calcul, risque de surajustement.	Dépend du modèle d'apprentissage automatique, ce qui limite la flexibilité.

Table 3.3 – Comparaison des méthodes pour la sélection des caractéristiques

3.4. Procédure PRO-NONLIN pour structures non-linéaires

Une étude sur les méthodes de sélection des caractéristiques selon les méthodes de filtrage, d'enveloppe et intégrée est réalisée. Ces méthodes sont définies pour réduire le nombre des caractéristiques pour l'entrainement des modèles de machine learning. Chaque méthode a ses points forts et ses points faibles.

Dans l'objectif de rendre l'application des modèles de ML efficace. Il y a deux stratégies possibles. La première stratégie est de sélectionner plus de caractéristiques pour l'entrainement du modèle, basé sur la méthode de filtrage. Les caractéristiques choisies sont autour de la valeur maximale de cette technique de filtrage (corrélation par exemple). Cette technique est plus rapide à réaliser, cependant, elle doit prendre en compte un nombre important de caractéristiques. La deuxième stratégie est de sélectionner les caractéristiques selon la méthode d'enveloppe. Cette méthode assure que le nombre optimal des caractéristiques est choisi. Elle aide à réduire la complexité du modèle. Cependant, cette méthode a besoin de plus de temps pour sélectionner la combinaison optimale. La méthode intégrée est considérée comme la méthode intermédiaire, mais elle est limitée à l'utilisation du modèle de RF seulement.

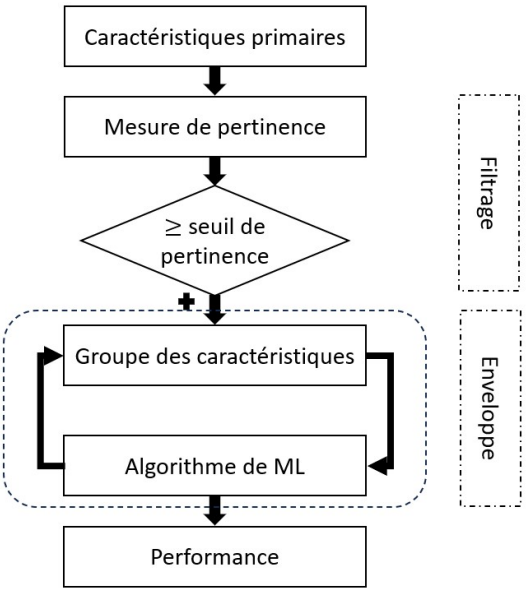


Figure 3.18 – Procédure de sélection des caractéristiques par la méthode hybride

A partir de l'analyse ci-dessus, il est possible et nécessaire de pouvoir proposer une stratégie hybride de sélection des caractéristiques. La Figure 3.18 présente la proposition de la méthode hybride de sélection des caractéristiques. Il est important de noter que la sélection des caractéristiques est un processus itératif et dépendant des données. Différentes techniques de filtrage et d'enveloppe peuvent être combinées en fonction du type de données, de la taille de l'ensemble de données et des objectifs du modèle prédictif. L'objectif ultime est de parvenir à un modèle avec des caractéristiques

significatives qui fournissent une prédiction précise de la réponse cible $\mathbf{Y}$ en économisant le temps de sélection.

Cette stratégie comprend les étapes suivantes :

- **Étape de filtrage :** Utiliser une méthode pour calculer la valeur de pertinence des caractéristiques $\mathbf{x}$ pour évaluer initialement leurs impact par rapport à la réponse cible $\mathbf{Y}$. Par exemple, le coefficient de corrélation. Une sélection préliminaire est réalisée, pour sélectionner les caractéristiques spectrales les plus pertinentes selon le filtrage en prenant un seuil acceptable. Cette sélection préliminaire réduit un nombre important des caractéristiques spectrales, en gardant celles qui ont plus grand impact par rapport à la variable cible.
- **Étape d'enveloppe :** Utiliser la méthode d'enveloppe pour sélectionner un sous-ensemble optimal des caractéristiques à partir des caractéristiques restantes de l'étape précédente en prenant en compte leurs impacts et leurs interactions dans l'entrainement du modèle.
- **Validation :** Après avoir sélectionné un sous-ensemble optimal des caractéristiques, évaluer la performance du modèle pour assurer que le modèle est bien entrainé et ne souffre pas de surajustement. Si la performance du modèle est insatisfaisante, il faut reprendre à l'étape de filtrage et ajuster les critères de sélection pour trouver un plus grand et meilleur sous-ensemble des caractéristiques.

Cette méthode hybride de sélection des caractéristiques fait partie de la procédure PRO-NONLIN, qui est basée sur l'apprentissage automatique pour l'évaluation des risques sismiques des structures non-linéaires. Cette procédure PRO-NONLIN est présentée dans la Figure 3.19.

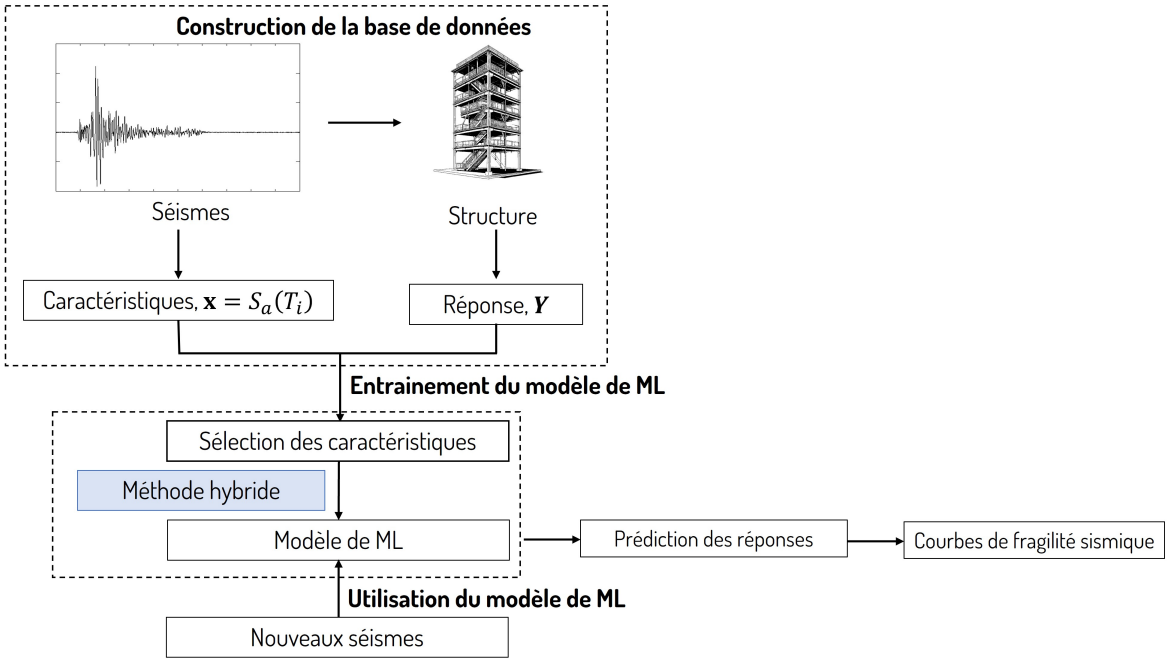


Figure 3.19 – [PRO-NONLIN] Procédure de l'application de ML pour des structures non-linéaires

3.5. Validation de la procédure PRO-NONLIN

3.5.1. Introduction

La stratégie hybride est appliquée dans cette section pour valider cette procédure d'abord et ensuite pour améliorer les études d'évaluation des risques sismiques par la construction des courbes de fragilité à partir de l'utilisation de modèles de ML optimisés.

Dans cette phase de validation, un oscillateur non-linéaire de Coulomb, un oscillateur non-linéaire de Bouc-Wen et un bâtiment à huit degrés de liberté sont étudiés. Le jeu de données sera détaillé pour chaque étude. Cette section examine la faisabilité des algorithmes suivants : régression linéaire (LR), k-Plus proches voisins (KNN), forêt aléatoire (RF), arbre de décision (DT), machines à vecteurs de support (SVMR), réseaux de neurones artificiels (ANN), algorithme de renforcement adaptatif (AdaBoost), machine à renforcement léger de gradient (LightGBM) et machine à renforcement extrême de gradient (XGBoost). Les critères d'évaluation sont la corrélation (r), le coefficient de détermination (R^2), l'erreur quadratique moyenne ($RMSE$) et l'erreur de pourcentage absolue moyenne symétrique ($SMAPE$). Ensuite, avec le modèle de ML le plus performant, la méthode hybride est appliquée pour pouvoir sélectionner des caractéristiques optimales pour l'entrainement. Finalement, la courbe de fragilité est construite pour montrer l'avantage de l'utilisation de l'apprentissage automatique dans l'évaluation des risques sismiques des structures.

3.5.2. Oscillateur non-linéaire de Coulomb

Une étude d'un oscillateur non-linéaire de Coulomb est réalisée. Un coefficient de frottement μ est choisi à 0.025 pour cette étude. La procédure PRO-NONLIN est appliquée pour construire des modèles de l'apprentissage automatique.

Un jeu de données comprenant 1000 observations est utilisé. Chaque observation représente la réponse de la structure soumise à une excitation sismique. Les caractéristiques $\mathbf{x}$ se composent de valeurs de spectre de réponse en accélération $S_a(T_i)$, où chaque période T_i est échantillonnée de 0.05 s à 2 s par incrément de 0.05 s et la valeur du spectre de réponse à la période propre du système $S_a(T^0)$. Ce choix d'échantillonnage définit 41 caractéristiques distinctes et assure une description des signaux sismiques. La réponse $\mathbf{Y}$ est le déplacement maximal de la masse de l'oscillateur, calculée selon l'équation (3.8). 80 % des observations du jeu de données sont utilisés pour l'entrainement, les 20 % restants permettent de valider.

$$\mathbf{Y} = \mathbf{ML}(\mathbf{x}) + \epsilon, \quad \mathbf{x} = \{S_a(T_i), S_a(T^0)\}, T_i = [0.05 : 0.05 : 2] \tag{3.8}$$

a. Sélection des caractéristiques

La sélection des caractéristiques par la méthode hybride est réalisée pour le modèle de machine learning de type RF. La sélection des caractéristiques par la méthode hybride est illustrée par la Figure 3.20. Le coefficient de corrélation entre les caractéristiques $\mathbf{x}$ et la réponse $\mathbf{Y}$ est calculé pour la première étape de la méthode hybride, et illustré par les barres dans cette figure. Un seuil de filtrage

est appliqué à la valeur de corrélation de 0.2, pour supprimer les caractéristiques les moins intéressantes. Après avoir appliqué ce filtrage, il reste 15 caractéristiques dans l'ensemble $\mathbf{x}$, par rapport à 41 initialement. Ensuite, la méthode d'enveloppe est employée pour sélectionner des caractéristiques optimales. Le nombre des caractéristiques entre 1 et 10 est exploré. L'ordre de sélection est numéroté en rouge pour les caractéristiques sélectionnées.

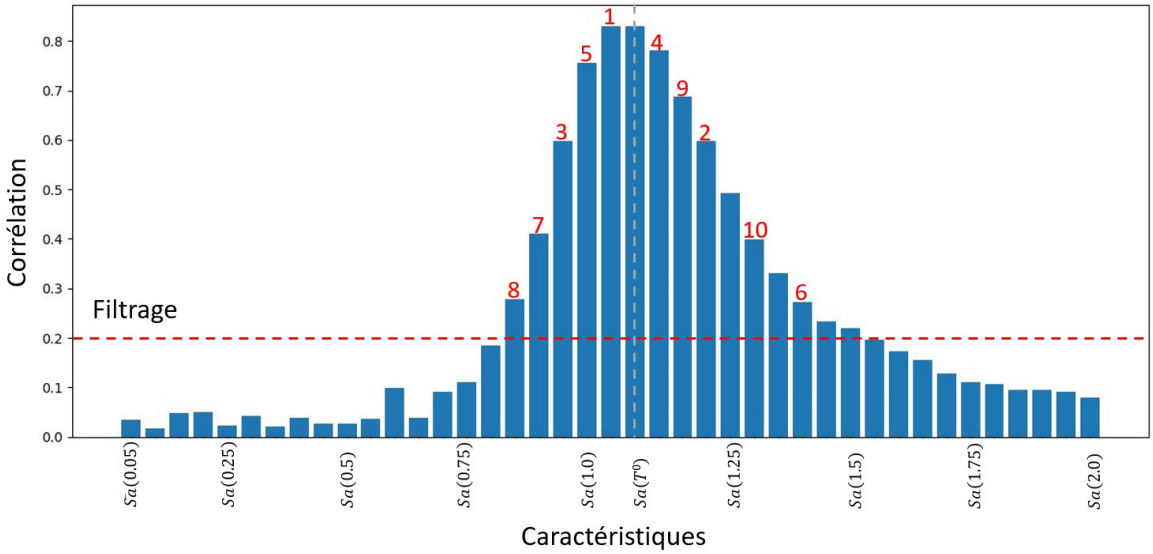


Figure 3.20 – [Coulomb] Application de la méthode hybride

La Figure 3.21 présente l'évolution des mesures d'évaluation, comprenant R^2 et $RMSE$, selon le nombre de caractéristiques. A partir de 6 caractéristiques, le modèle RF montre une performance assez stable, l'augmentation de R^2 entre deux itérations est inférieure à 1 %.

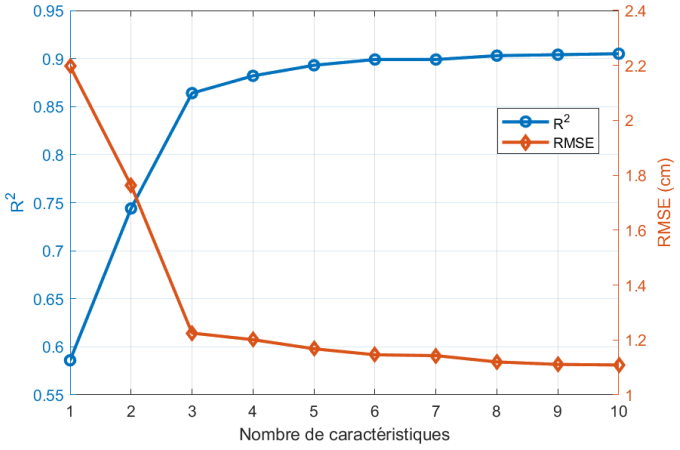


Figure 3.21 – [Coulomb] Évolution des métriques d'évaluation selon le nombre de caractéristiques

b. Entrainement des modèles de machine learning

Les différents modèles de ML sont construits en utilisant la méthode hybride pour sélectionner des caractéristiques. La performance du modèle est évaluée par rapport aux critères habituels. Ces critères sont la moyenne de 10 entrainements du modèle sur une répartition des données. Ceux-ci sont présentés dans le Tableau 3.4.

Modèle	r	R^2	$SMAPE$ (%)	$RMSE$ (cm)
LR	0.946	0.892	11.687	1.006
KNN	0.906	0.785	15.278	1.420
DT	0.900	0.800	15.284	1.367
RF	0.951	0.894	11.382	0.999
ANN	0.971	0.939	8.044	0.712
SVMR	0.950	0.902	10.814	0.960
AdaBoost	0.942	0.879	12.419	1.067
LightGBM	0.971	0.931	7.181	0.717
XGBoost	0.955	0.909	10.346	0.925

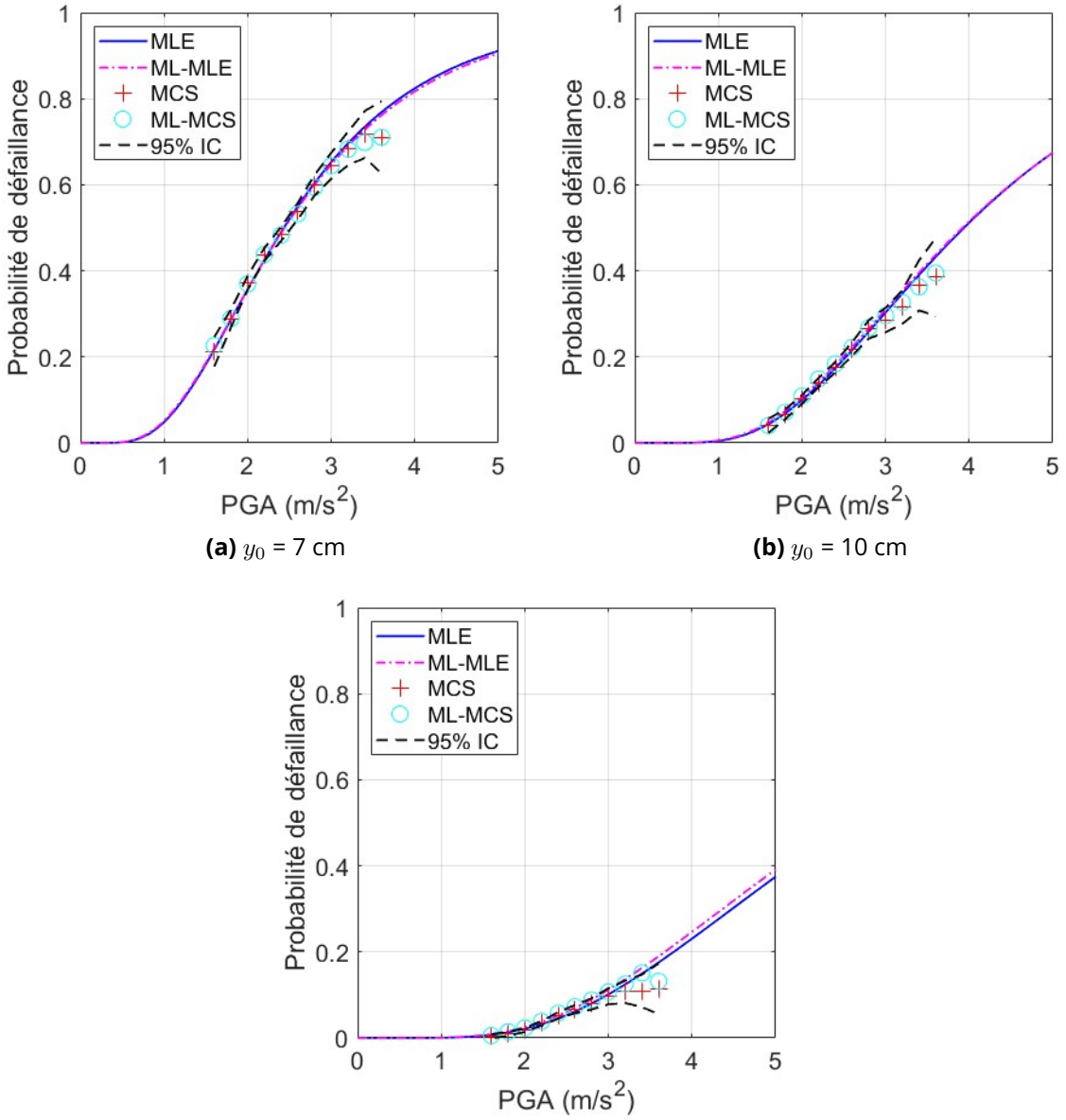
Table 3.4 – [Coulomb] Évaluation de performance des modèles de ML

Les modèles basés sur les réseaux de neurones artificiels ANN et les méthodes de boosting de gradient (LightGBM et XGBoost) sont les plus performants pour cette tâche spécifique, indiquant leur aptitude à modéliser avec succès les non-linéarités et les complexités du système non-linéaire de Coulomb. Leurs faibles valeurs de $RMSE$ et de $SMAPE$ montrent qu'ils sont robustes et précis.

c. Courbes de fragilité sismique

Après avoir trouvé le modèle d'apprentissage automatique optimal qui est le ANN pour ce cas, il est utilisé pour prédire le déplacement $\mathbf{Y}$ de la structure. Les réponses prédites sont utilisées pour construire la courbe de fragilité. Dans ce cas, après avoir entrainé avec 1000 observations, 50000 nouvelles observations sont utilisées pour la prédiction. Les déplacements limites pour ce cas sont égaux à 7, 10 et 13 cm, donc y_0 = [7, 10, 13] cm. Les méthodes pour construire la courbe de fragilité sont la simulation de Monte Carlo (MCS), la simulation de Monte Carlo par l'apprentissage automatique (ML-MCS), la méthode du maximum de vraisemblance (MLE) et finalement la méthode du maximum de vraisemblance par l'apprentissage automatique (ML-MLE).

La Figure 3.22 présente les courbes de fragilité obtenues par les quatre méthodes énoncées, avec un intervalle de confiance à 95 % pour la méthode MCS. Les quatre méthodes montrent un bon accord, et les courbes sont très proches de celle de référence obtenue par la méthode MCS. De plus, les probabilités calculées par la ML-MCS se trouvent dans l'intervalle de confiance à 95 % de celle de MCS. La Figure 3.23a compare l'erreur MSE pour chaque méthode par rapport à la méthode ML-MCS. La méthode ML-MCS a une erreur de MSE faible, ce qui présente la cohérence entre les méthodes MCS et ML-MCS. De plus, la probabilité de défaillance est aussi calculée et représentée sur la Figure 3.23b.



(a) y_0 = 7 cm

(b) y_0 = 10 cm

(c) y_0 = 13 cm

Figure 3.22 – [Coulomb] Courbes de fragilité par différentes méthodes

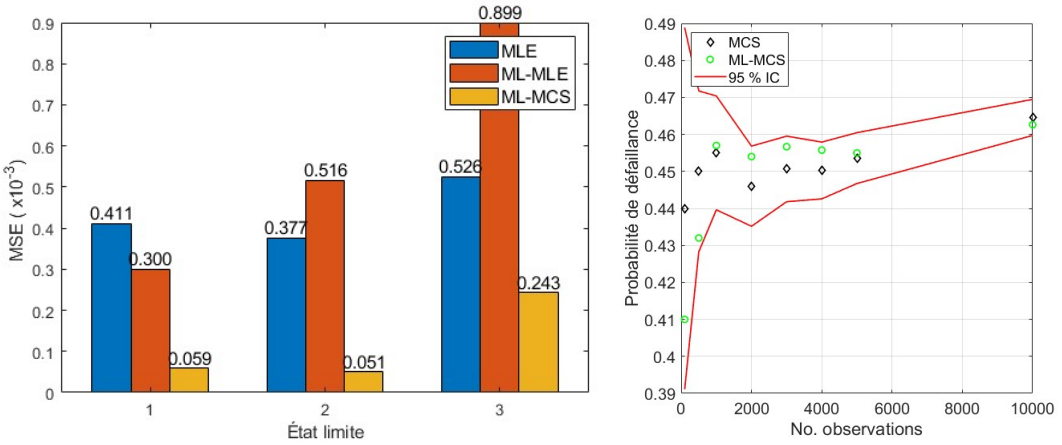
La probabilité calculée par la méthode ML-MCS reste dans l'intervalle de confiance à 95 % de celle calculée par la méthode de référence MCS, signifiant ainsi la cohérence entre ces 2 méthodes.

De la même manière avec le cas de l'oscillateur linéaire, le bilan de temps réservé pour la construction des observations de la structure est établi dans le Tableau 3.5. Les méthodes MLE et MCS traditionnelles utilisent respectivement 10000 et 2×10^5 observations obtenues par les calculs FEM. Les méthodes ML-MLE et ML-MCS utilisent le même nombre d'observations mais elles sont générées par

les modèles de machine learning, qui ne demandent que 1000 observations par FEM pour leur entrainement. Pour cet oscillateur de Coulomb, un calcul nécessite une minute. Le temps réservé pour la méthode approximative MLE est 10 fois plus grand que la méthode ML-MLE, tandis que celui de la méthode MCS est 2000 fois plus grand. Cette réduction présente aussi un avantage de la méthode basée sur les modèles de machine learning.

Méthode	MLE	MCS	ML-MLE	ML-MCS
Nombre d'observations	10000	2×10^5	2×10^5	2×10^5
Nombre de simulations par éléments finis	10000	2×10^5	1000	1000
Durée de simulation (minute)	10000	2×10^5	1000	1000

Table 3.5 – [Coulomb] Nombres d'observations pour la construction des courbes de fragilité



(a) Comparaison des méthodes **(b)** Comparaison des critères d'évaluation

Figure 3.23 – [Coulomb] Comparaisons des méthodes de construction de la courbe

3.5.3. Oscillateur non-linéaire de Bouc-Wen

Pour la construction des modèles de machine learning, nous utilisons la procédure PRO-NONLIN sur un oscillateur caractérisé par un comportement non-linéaire de Bouc-Wen. Cet oscillateur est présenté dans l'étude préliminaires. Pour notre étude, la constante α est choisie égale à 0.7. On utilise 2000 observations dont 80 % sont utilisés pour entrainer les modèles et 20 % pour les tester. Les caractéristiques sont les spectres de réponse en accélération $S_a(T_i)$, avec T_i échantillonné entre 0.05 s et 4 s avec un incrément de 0.05 s, T_i = [0.05 : 0.05 : 4] s ainsi que le spectre de réponse à la période propre, $S_a(T^0)$. 81 caractéristiques sont définies au total. La réponse $\mathbf{Y}$ est le déplacement maximal de la structure, $\mathbf{Y} = \max_t |y(t)|$. L'équation (3.9) représente les modèles de ML :

$$\mathbf{Y} = \mathbf{ML}(\mathbf{x}) + \epsilon, \quad \mathbf{x} = \{S_a(T_i), S_a(T^0)\}, T_i = [0.05 : 0.05 : 4] \tag{3.9}$$

a. Sélection des caractéristiques

La sélection des caractéristiques par la méthode hybride dans la procédure PRO-NONLIN est réalisée pour le modèle de machine learning de type RF. Premièrement, le coefficient de corrélation entre les caractéristiques $\mathbf{x}$ et la réponse $\mathbf{Y}$ est calculé. Un seuil de filtrage est appliqué à la valeur de corrélation de 0.2, pour supprimer les caractéristiques les moins intéressantes. Après avoir appliqué ce filtrage, il reste 26 caractéristiques dans l'ensemble $\mathbf{x}$, par rapport aux 81 initiales. Deuxièmement, nous employons la méthode d'enveloppe pour sélectionner entre 1 et 10 caractéristiques à partir de 26 restantes. La Figure 3.24 montre la sélection des caractéristiques pour chaque itération. La caractéristique choisie est identifiée avec son ordre de sélection (chiffre rouge). Pour ce cas, le spectre au pic de corrélation est choisi premièrement, il est donc considéré comme le plus important. De plus, une dispersion des sélections est aussi observée. La Figure 3.24 représente l'application de la méthode hybride dans notre cas. Le choix du filtrage aide à la réduction du nombre de caractéristiques considérées par l'enveloppe. Le fait d'augmenter ou de baisser le seuil de filtrage influence les caractéristiques sélectionnées. Dans cette étude, le choix est fait pour avoir un bon compromis de temps pour la sélection finale. Des études paramétriques peuvent possiblement être faites pour améliorer ce choix.

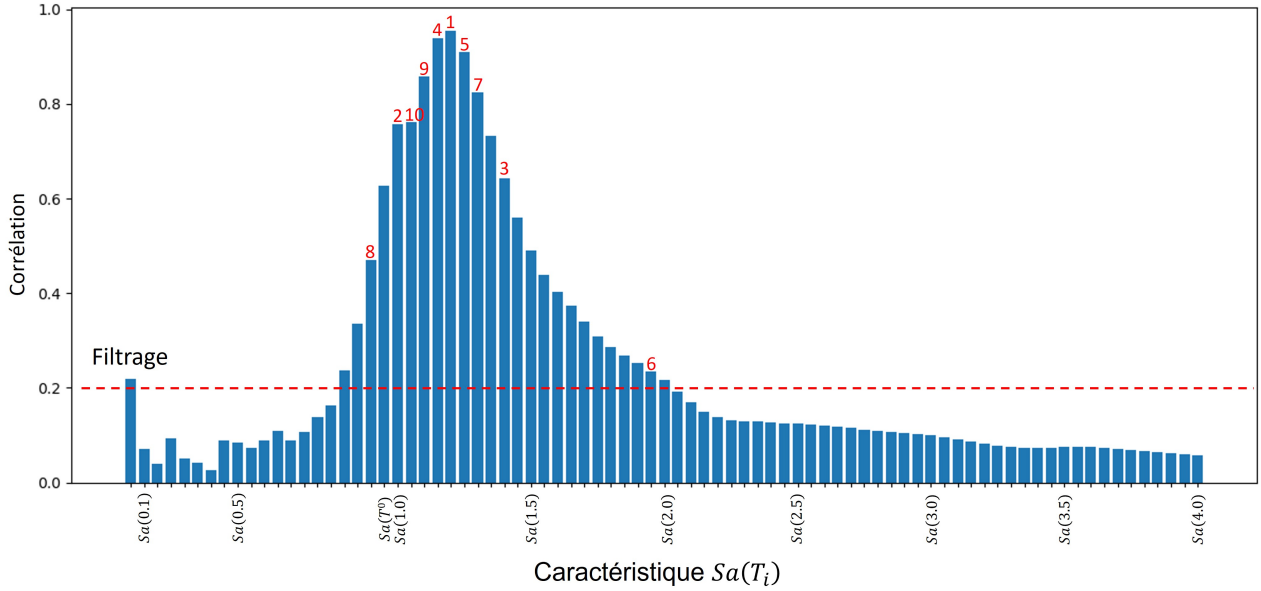


Figure 3.24 – [Bouc-Wen] Application de la méthode hybride

La Figure 3.25 présente l'évolution des mesures d'évaluation, comprenant R^2 et $RMSE$, selon le nombre de caractéristiques. A partir de 6 caractéristiques, un modèle d'apprentissage automatique de type RF est bien validé. La sélection des caractéristiques par la méthode hybride permet d'optimiser le nombre nécessaire pour entrainer un modèle de machine learning. Ainsi, seulement 6 caractéristiques sont nécessaires pour avoir un modèle de ML de bonne qualité de prédiction. De plus, en

combinant la méthode de filtrage avec la méthode d'enveloppe, le méthode hybride permet d'optimiser le temps de sélection. Cette méthode hybride est possible parce que la nature de caractéristiques est similaire, ceux sont des valeurs du spectre de réponse en accélération.

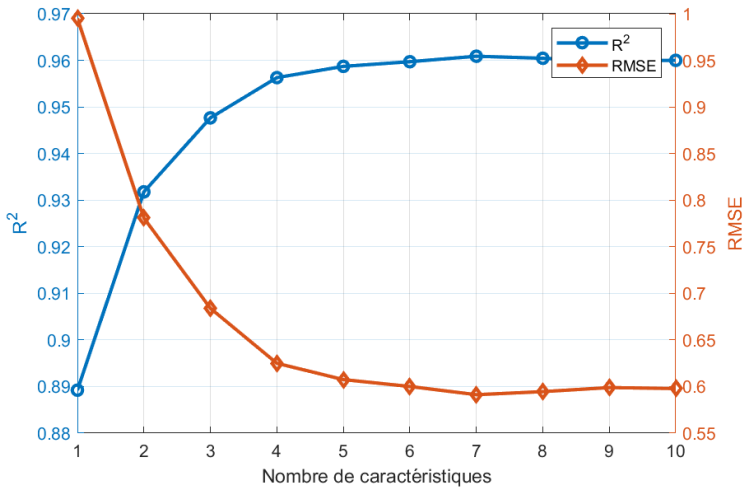


Figure 3.25 – [Bouc-Wen] Évolution des métriques d'évaluation selon le nombre de caractéristiques

b. Entrainement des modèles de machine learning

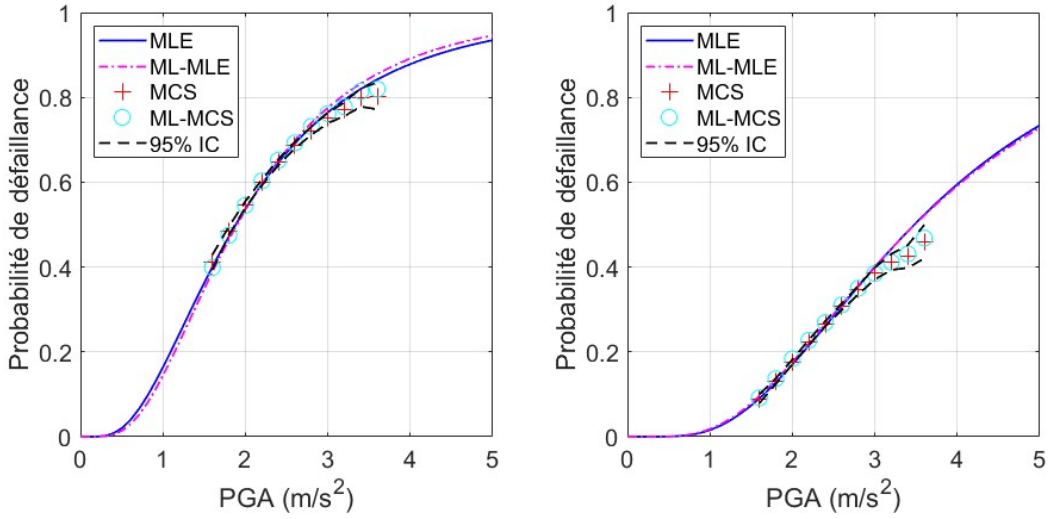
D'autres modèles de ML sont construits pour ce système en utilisant la sélection des caractéristiques par la méthode hybride proposée. Le Tableau 3.6 montre les résultats obtenus. Sauf le modèle KNN, les modèles d'apprentissage automatique présentent de bonnes performances. Le modèle de type ANN est le plus efficace dans ce cas-là, suivi par les modèles de type XGBoost, LightGBM et SVMR.

Modèle	r	R^2	$SMAPE$ (%)	$RMSE$ (cm)
LR	0.976	0.951	6.738	0.643
KNN	0.912	0.797	12.753	1.306
DT	0.956	0.911	8.435	0.865
RF	0.976	0.952	6.176	0.632
ANN	0.983	0.964	5.287	0.509
SVMR	0.978	0.955	6.135	0.607
AdaBoost	0.974	0.946	6.889	0.671
LightGBM	0.977	0.954	6.546	0.625
XGBoost	0.980	0.960	5.929	0.583

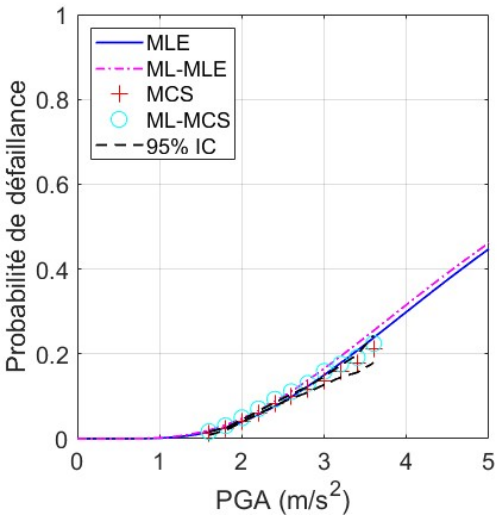
Table 3.6 – [Bouc-Wen] Évaluation de la performance des modèles de ML

c. Courbes de fragilité sismique

Après avoir entrainé les modèles, ils sont utilisés pour prédire le déplacement de la structure. Dans ce cas, le modèle ANN est employé. Les réponses prédites sont utilisées pour construire la courbe de fragilité basée sur l'apprentissage automatique. Dans ce cas, après avoir entrainé avec 2000 observations, 50000 nouvelles observations sont générées. Les déplacements limites y_0 sont égaux à 7, 10 et 13 cm. Les méthodes pour construire la courbe de fragilité sont MCS, ML-MCS, MLE et ML-MLE. Ces courbes sont illustrées dans la Figure 3.26.



(a) y_0 = 7 cm **(b)** y_0 = 10 cm



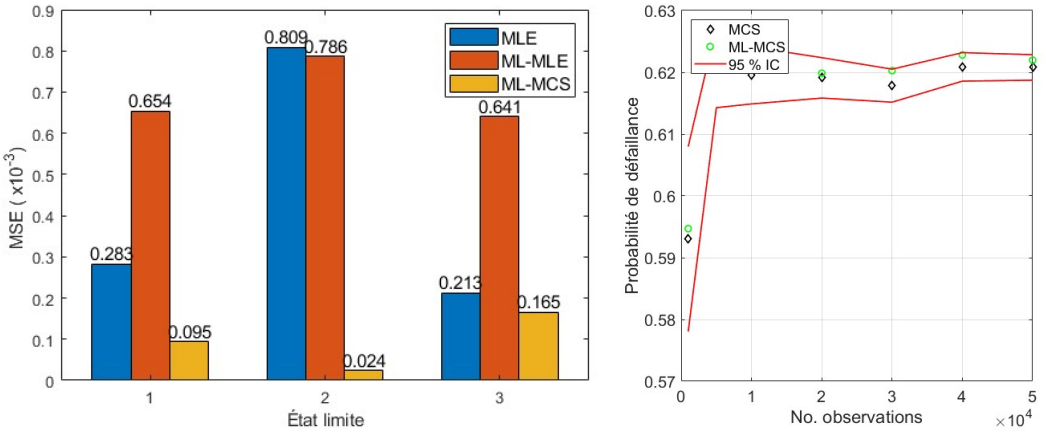
(c) y_0 = 13 cm

Figure 3.26 – [Bouc-Wen] Courbes de fragilité par différentes méthodes

La ligne en pointillés représente un intervalle de confiance à 95 % de la méthode MCS, indiquant

la plage dans laquelle les vraies valeurs des probabilités d'endommagement sont estimées se trouver avec une confiance de 95 %. Les quatre méthodes montrent un accord entre elles. Les courbes obtenues sont proches entre elles et par rapport à celle de la méthode MCS de référence.

La Figure 3.27 présente la comparaison des courbes de fragilité par l'utilisation de MSE et la probabilité de défaillance en fonction du nombre d'observations. Les probabilités calculées par la ML-MCS se trouvent dans l'intervalle de confiance à 95 % défini. Cela montre la performance de cette méthode à base d'apprentissage automatique.



(a) Comparaison des méthodes **(b)** Comparaison des critères d'évaluation

Figure 3.27 – [Bouc-Wen] Comparaisons des méthodes de construction de la courbe

3.5.4. Système non-linéaire à plusieurs degrés de liberté de Bouc-Wen

a. Description du système

Nous considérons ici un système à 8 degrés de liberté avec le comportement non-linéaire de Bouc-Wen, illustré par la Figure 3.28. L'équation dynamique du système s'écrit :

$$\left\{\begin{array}{lcl} m_1\ddot{y}_1 + (c_1+c_2)\dot{y}_1 - c_2\dot{y}_2 + \alpha(k_1+k_2)y_1 - \alpha k_2 y_2 + (1-\alpha)k_1 w_1 - (1-\alpha)k_2 w_2 & = & -m_1 a \\ m_2\ddot{y}_2 - c_2\dot{y}_1 + (c_2+c_3)\dot{y}_2 - c_3\dot{y}_3 - \alpha k_2 y_1 + & & \\ +\alpha(k_2+k_3)y_2 - \alpha k_3 y_3 + (1-\alpha)k_2 w_2 - (1-\alpha)k_3 w_3 & = & -m_2 a \\ \dots & = & \dots \\ m_7\ddot{y}_7 - c_7\dot{y}_6 + (c_7+c_8)\dot{y}_7 - c_8\dot{y}_8 - \alpha k_7 y_6 + & & \\ +\alpha(k_7+k_8)y_7 - \alpha k_8 y_8 + (1-\alpha)k_7 w_7 - (1-\alpha)k_8 w_8 & = & -m_7 a \\ m_8\ddot{y}_8 - c_8\dot{y}_7 + c_8\dot{y}_8 - \alpha k_8 y_7 + \alpha k_8 y_8 + (1-\alpha)k_8 w_8 & = & -m_8 a \end{array}\right. \quad (3.10)$$

où $w_{i=1,\dots,8}$ présentent le comportement de Bouc-Wen et sont définis par :

$$\left\{\begin{array}{lcl} \dot{w}_1 & = & C_1\dot{y}_1 - C_2\,|\dot{y}_1|\,|w_1|^{n_d-1}\,w_1 - C_3\dot{y}_1\,|w_1|^{n_d} \\ \dot{w}_2 & = & C_1(\dot{y}_2-\dot{y}_1) - C_2\,|(\dot{y}_2-\dot{y}_1)|\,|w_2|^{n_d-1}\,w_2 - C_3(\dot{y}_2-\dot{y}_1)\,|w_2|^{n_d} \\ \dots & & \\ \dot{w}_8 & = & C_1(\dot{y}_8-\dot{y}_7) - C_2\,|(\dot{y}_8-\dot{y}_7)|\,|w_8|^{n_d-1}\,w_8 - C_3(\dot{y}_8-\dot{y}_7)\,|w_8|^{n_d} \end{array}\right.$$

Les paramètres du système sont repris de l'étude de Chen et Li [115] et de Dang [9], où $\alpha = 0.04$, $C_1 = 1$, $C_2 = 30/\mathrm{m}^{n_d}$, $C_3 = 10/\mathrm{m}^{n_d}$, $n_d = 1$. Les périodes propres de la structure sont 0.942, 0.338, 0.205, 0.153, 0.124, 0.108, 0.098 et 0.092 s.

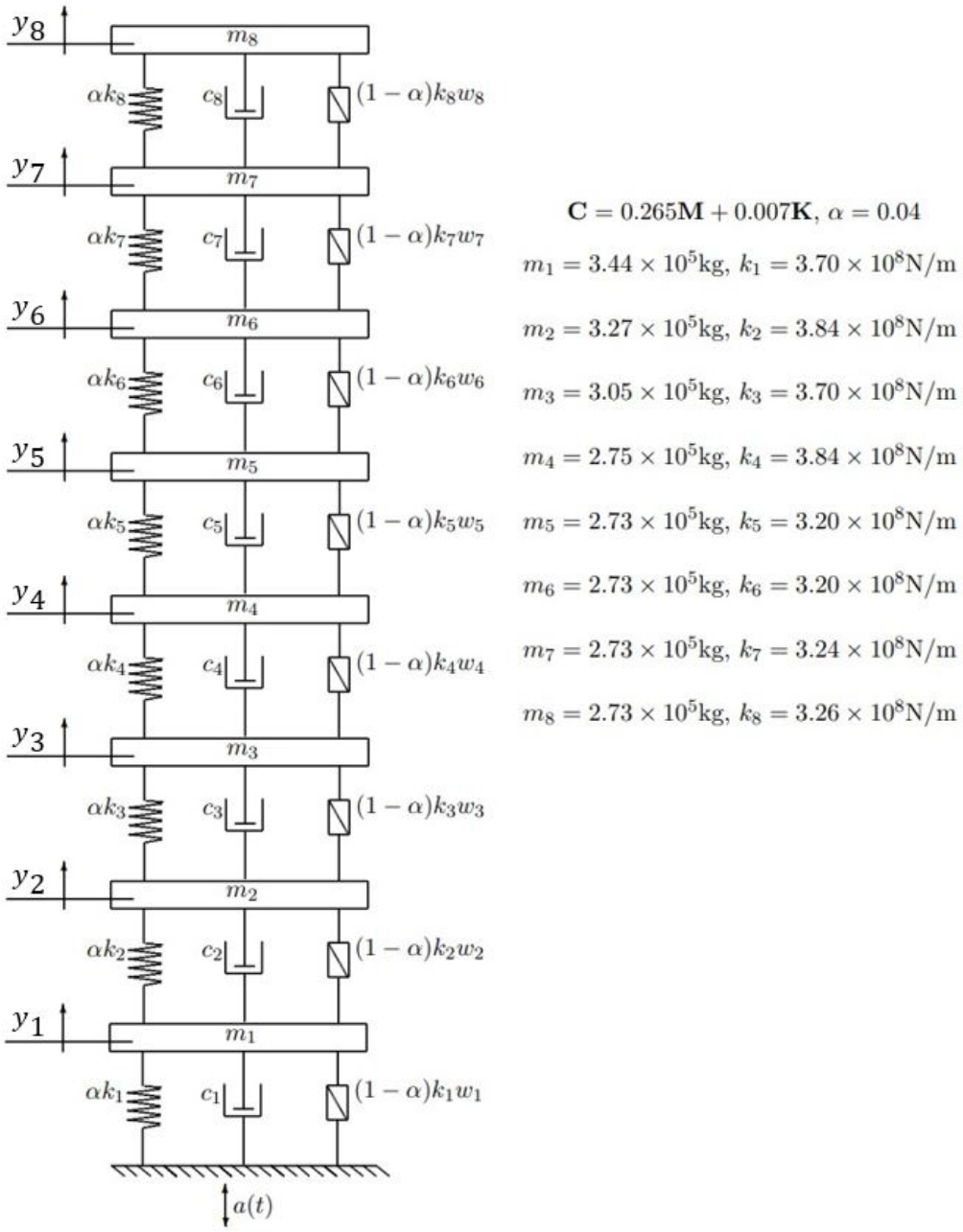


Figure 3.28 – Structure à huit degrés de liberté non-linéaire de Bouc-Wen

L'équation dynamique du système s'écrit sous forme matricielle :

$$\mathbf{M}\ddot{\mathbf{y}}(t) + \mathbf{C}\dot{\mathbf{y}}(t) + \alpha\mathbf{K}\mathbf{x}(t) + (1-\alpha)\mathbf{G}\mathbf{w}(t) = -\mathbf{M}\mathbf{i}a(t) \tag{3.11}$$

où :

$$\mathbf{M} = \begin{pmatrix} m_1 & 0 & \dots & 0 \\ 0 & m_2 & \dots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & m_8 \end{pmatrix}; \mathbf{i} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$$

$$\mathbf{C} = \begin{pmatrix} c_1 + c_2 & -c_2 & \dots & 0 & 0 \\ -c_2 & c_2 + c_3 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & -c_7 & 0 \\ 0 & 0 & -c_7 & c_7 + c_8 & -c_8 \\ 0 & 0 & 0 & -c_8 & c_8 \end{pmatrix};$$

$$\mathbf{K} = \begin{pmatrix} k_1 + k_2 & -k_2 & \dots & 0 & 0 \\ -k_2 & k_2 + k_3 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & -k_7 & 0 \\ 0 & 0 & -k_7 & k_7 + k_8 & -k_8 \\ 0 & 0 & 0 & -k_8 & k_8 \end{pmatrix};$$

$$\mathbf{G} = \begin{pmatrix} k_1 & -k_2 & \dots & 0 & 0 \\ 0 & k_2 & \dots & 0 & 0 \\ 0 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & 0 & 0 \\ 0 & 0 & 0 & k_7 & -k_8 \\ 0 & 0 & 0 & 0 & k_8 \end{pmatrix}$$

Dans cette étude, on utilise 4000 observations dont 80 % des données sont utilisés pour entrainer les modèles et 20 % pour les tester. Les caractéristiques pour cette étude sont les valeurs du spectre de réponse d'accélération $S_a(T_i)$, avec T_i échantillonné suivant T_i = [0.05 : 0.05 : 2] s, équivalent à 2 fois T_1^0. De plus, les valeurs du spectre de réponse aux périodes propres du système, de $S_a(T_1^0)$ à $S_a(T_8^0)$, sont aussi utilisées. Nous obtenons 48 caractéristiques. Les déplacements des masses, $y_i(t)$, sont obtenus par la méthode de Runge-Kutta par MATLAB. La réponse à modéliser par machine learning est le déplacement maximal relatif entre étages de la structure, $\mathbf{Y} = \max_t |\delta(t)|$, avec $\delta(t) = (y_i(t) - y_{i-1}(t))/h_i$ avec h_i est la hauteur de l'étage i.

b. Sélection des caractéristiques

La sélection des caractéristiques par la méthode hybride est réalisée pour le modèle de machine learning de type RF. La procédure est aussi appliquée pour d'autres modèles de ML.

Premièrement, le coefficient de corrélation entre les caractéristiques $\mathbf{x}$ et la réponse $\mathbf{Y}$ est calculé. Les spectres aux périodes propres sont marqués par les lignes pointillées verticales. Un seuil de

filtrage est appliqué à la valeur de corrélation de 0.25, pour supprimer les caractéristiques les moins intéressantes. Après avoir appliqué ce filtrage, il reste 33 caractéristiques dans l'ensemble $\mathbf{x}$ des 48 initiales. Les lignes verticales pointillées marquent les caractéristiques échantillonnées aux périodes propres. Ces caractéristiques sont aussi supprimées par cette étape. Cela signifie que ces caractéristiques ne sont pas importantes dans ce cas. Deuxièmement, la méthode d'enveloppe est employée pour sélectionner les caractéristiques. Le nombre des caractéristiques entre 1 et 10 est exploré. La Figure 3.29 montre la sélection obtenue des caractéristiques pour chaque itération avec son ordre (chiffre rouge) selon la méthode hybride proposée. La Figure 3.30 présente l'évolution des mesures d'évaluation, comprenant R^2 et $RMSE$, selon le nombre de caractéristiques. A partir de 6 caractéristiques, la performance du modèle RF est assez stable. Cela signifie que c'est le nombre optimal de caractéristiques dans ce cas. En outre, non seulement les caractéristiques autour de la période avec une haute corrélation avec la réponse $\mathbf{Y}$ sont considérées, les autres qui participent à décrire les signaux sismiques sont aussi importantes pour l'entrainement du modèle. La sélection des caractéristiques par la méthode hybride permet d'optimiser le nombre nécessaire pour entrainer un modèle de machine learning. Nous avons montré que seulement 6 caractéristiques sont nécessaires pour réussir l'entrainement du modèle de ML sur les 48 initialement définies.

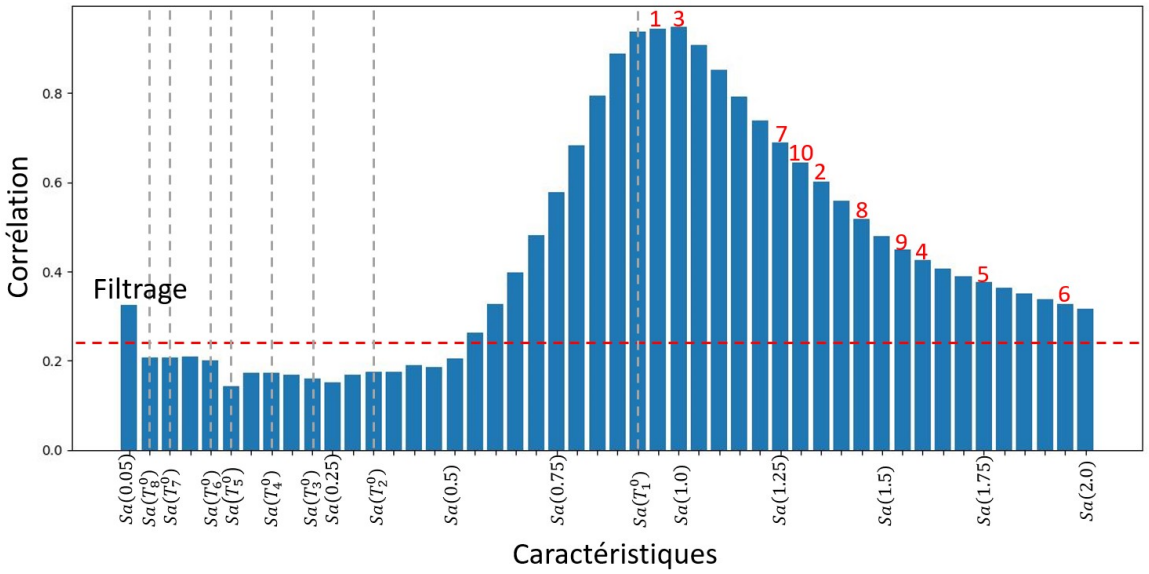


Figure 3.29 – [Bouc-Wen 8ddl] Application de la méthode hybride

c. Entrainement des modèles de machine learning

En utilisant la procédure PRO-NONLIN, les différents modèles de ML sont aussi entrainés pour le système à plusieurs degrés de liberté. Dans cet exemple, le modèle RF est choisi grâce à sa précision. Le Tableau 3.7 présente la performance des modèles selon les critères. Les critères d'évaluation sont respectivement le coefficient de Pearson r, le coefficient de détermination R^2, l'erreur de pourcentage absolue moyenne symétrique $SMAPE$ et l'erreur quadratique moyenne $RMSE$. En se basant sur les résultats obtenus, détaillés dans le Tableau 3.7, les modèles d'apprentissage automatique

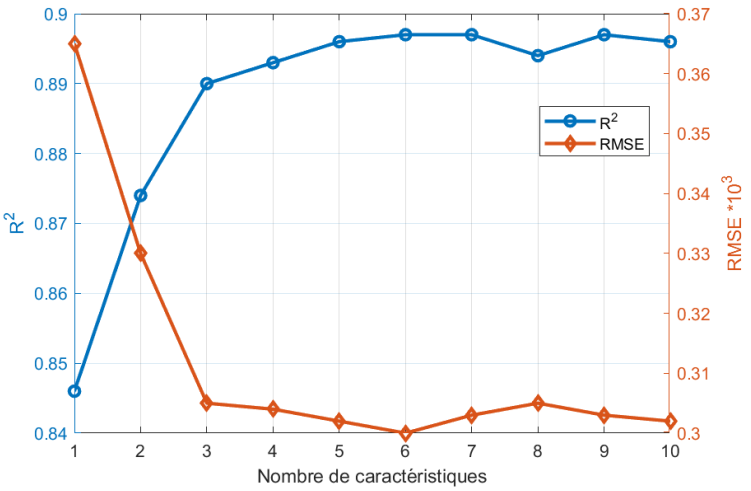


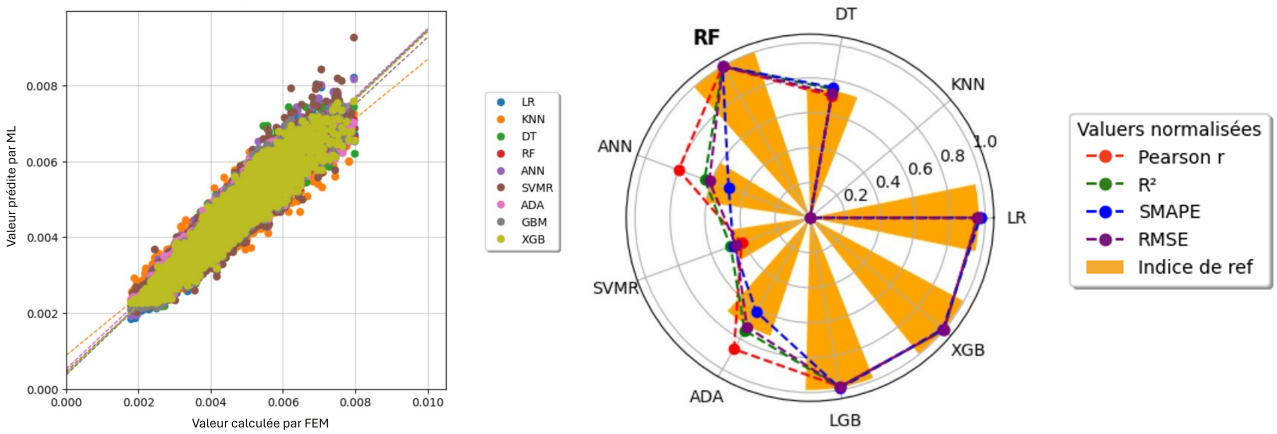
Figure 3.30 – [Bouc-Wen 8ddl] Évolution des métriques d'évaluation selon le nombre de caractéristiques

sont acceptables. Le modèle de type RF est le plus efficace, suivi par les modèles de type XGBoost et LightGBM. Les performances des modèles sont assez proches dans ce cas-là.

Modèle	r	R^2	$SMAPE$ (%)	$RMSE \times 10^{-3}$
LR	0.953	0.909	5.640	0.304
KNN	0.929	0.854	7.271	0.379
DT	0.947	0.897	6.002	0.326
RF	0.955	0.911	5.594	0.297
ANN	0.949	0.891	6.442	0.367
SVMR	0.940	0.882	6.490	0.373
AdaBoost	0.951	0.897	6.236	0.316
LightGBM	0.954	0.910	5.610	0.296
XGBoost	0.955	0.911	5.611	0.298

Table 3.7 – [Bouc-Wen 8ddl] Évaluation de la performance des modèles de ML

Une autre comparaison est présentée par la Figure 3.31. L'indice de référence montre clairement que le modèle de RF est le plus performant dans ce cas. De plus, les prédictions par des modèles, illustrées par la Figure 3.31b, sont proches de celles du modèle RF.



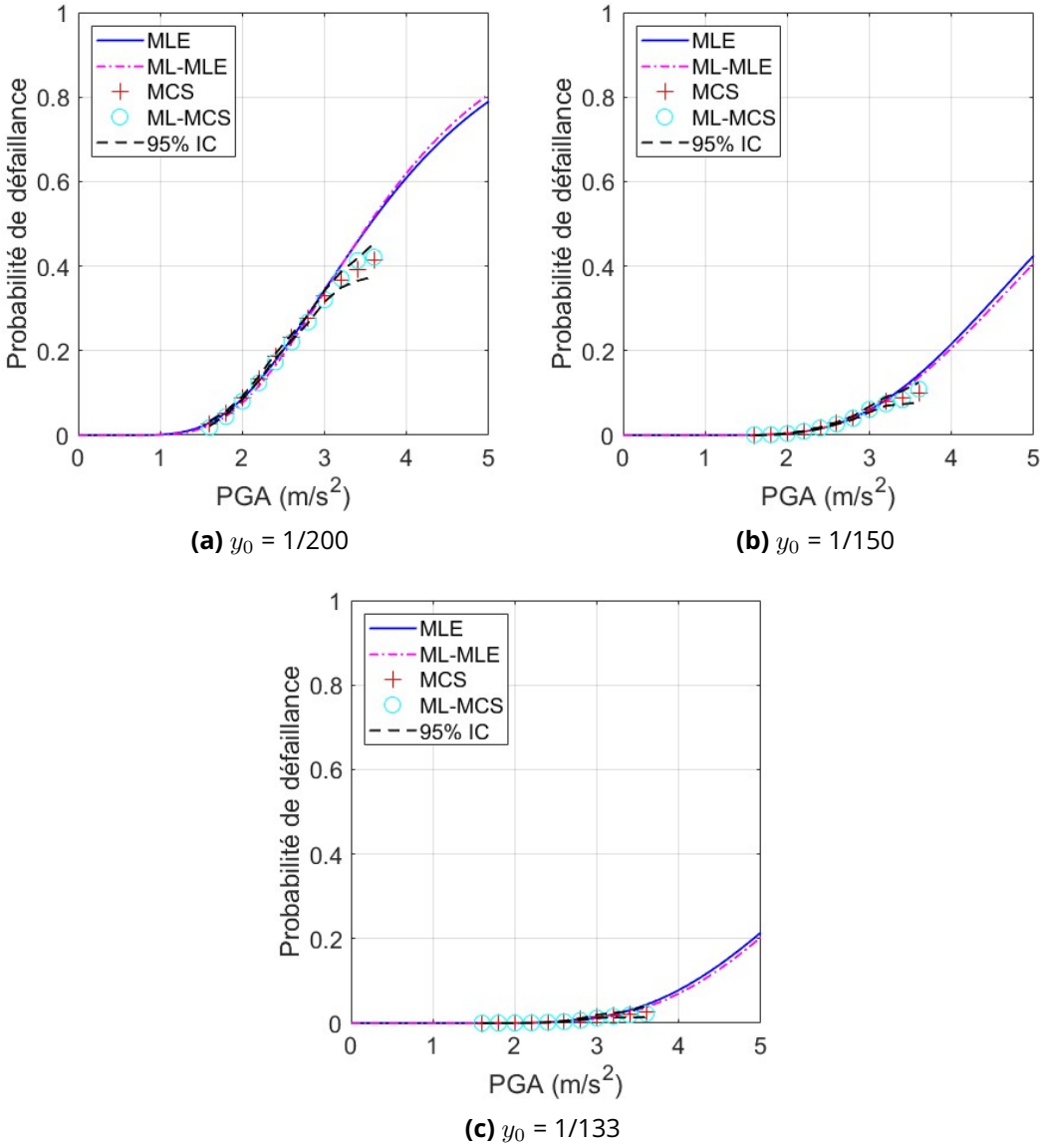
(a) Comparaison des valeurs prédites et réelles **(b)** Comparaison des modèles

Figure 3.31 – [Bouc-Wen 8ddl] Comparaisons des modèles de machine learning

d. Courbes de fragilité sismique

Après avoir construit les modèles de machine learning, ils sont utilisés pour prédire le ratio maximal de déplacement inter-étage du système de huit degrés de liberté avec le modèle non-linéaire de Bouc-Wen. Plus précisément, le modèle RF est employé dans ce cas. Les réponses prédites sont utilisées pour faire la courbe de fragilité basée sur l'apprentissage automatique. Dans ce cas, après avoir entrainé avec 4000 observations, 50000 nouvelles observations sont utilisées pour la prédiction. Les ratios de déplacement inter-étage maximaux limites pour ce cas sont égaux à 1/200, 1/150 et 1/133, donc y_0 = [1/200, 1/150, 1/133]. Les méthodes pour construire la courbe de fragilité sont respectivement la simulation de Monte Carlo (MCS), la simulation de Monte Carlo par l'apprentissage automatique (ML-MCS), la méthode du maximum de vraisemblance (MLE) et la méthode du maximum de vraisemblance par l'apprentissage automatique (ML-MLE).

La Figure 3.32 illustre les courbes de fragilité obtenues par les quatre méthodes mentionnées avec l'intervalle de confiance à 95 % de la méthode MCS. Les quatre méthodes montrent un bon accord entre elles. Les courbes sont proches de celle obtenue par la méthode MCS. De plus, les probabilités calculées par la ML-MCS se trouvent dans l'intervalle de confiance à 95 % de celle de MCS, qui montre la fiabilité de la méthode à base d'apprentissage automatique.



(a) y_0 = 1/200

(b) y_0 = 1/150

(c) y_0 = 1/133

Figure 3.32 – [Bouc-Wen 8ddl] Courbes de fragilité par différentes méthodes

3.6. Influence de la non-linéarité de Bouc-Wen sur la sélection des caractéristiques

La section précédente a mis en lumière la capacité de la méthode hybride pour sélectionner des caractéristiques. La non-linéarité de l'oscillateur influence la construction des modèles de machine learning et donc la sélection des caractéristiques. Spécialement pour l'oscillateur de Bouc-Wen,

quand la non-linéarité apparait dans la rigidité, comme le montre la Figure 3.4. Cependant, le modèle de Bouc-Wen considéré était caractérisé par une non-linéarité modérée avec le coefficient α de 0.7, conduisant à des performances de sélection de caractéristiques relativement similaires entre les méthodes. Cette similitude des résultats soulève des questions quant à leur applicabilité à des systèmes présentant une non-linéarité plus prononcée. Pour aborder cette problématique, cette section entreprend une exploration plus approfondie. Elle vise à évaluer l'impact de la non-linéarité sur la procédure de sélection des caractéristiques. L'objectif est de déterminer comment ce facteur influence l'efficacité la méthode de sélection les caractéristiques en utilisant la procédure PRO-NONLIN.

Pour chaque oscillateur de Bouc-Wen étudié, l'ensemble de données en question contient 81 caractéristiques, exprimées par $\mathbf{x} = \{S_a(T_i), S_a(T^0)\}$, où $S_a(T_i)$ désigne les valeurs du spectre d'accélération pour une série de périodes qui varient de 0.05 à 4 secondes par incrément de 0.05. La réponse souhaitée $\mathbf{Y}$ pour ce cas d'étude est le déplacement maximal de la masse. Dans le cadre de cette analyse, le modèle RF est utilisé. Pour cette étude, l'impact du coefficient α est examiné en détail. Le coefficient α, qui est utilisé pour modifier l'intensité de la non-linéarité du système, varie de 0.9 à 0.1. Le nombre d'observations pour chaque cas est de 2000. Cette variation de α est nécessaire pour évaluer son influence sur la sélection des caractéristiques et la performance subséquente du modèle de machine learning.

Les figures (de 3.33 à 3.41) présentent les résultats de la sélection des caractéristiques par la méthode hybride pour des oscillateurs de Bouc-Wen selon la variation de α. Pour chaque oscillateur, la sélection des caractéristiques par la méthode hybride est détaillée avec son filtrage et avec la sélection pour chaque itération (par le numéro en rouge sur celles sélectionnées). De plus, les évolutions de R^2 et $RMSE$ selon le nombre de caractéristiques pour chaque itération sont aussi présentées. On observe qu'à partir de 81 caractéristiques, la sélection hybride permet d'obtenir le nombre optimal des caractéristiques. Avec plus de caractéristiques, le modèle RF présente une meilleure performance. Par contre, le nombre optimal de caractéristiques et les caractéristiques sélectionnées dépendent des oscillateurs.

Pour α = 0.9, présenté dans la Figure 3.33, on remarque que la sélection est autour du pic de la corrélation entre $\mathbf{x}$ et $\mathbf{Y}$, indiquant que le système se comporte de manière presque linéaire. Similairement pour les autres oscillateurs comme α = 0.8 (présenté par la Figure 3.34) ou α = 0.7 (présenté par la Figure 3.35). Les cas de non-linéaire intermédiaire nécessitent une sélection des caractéristiques plus éloignées du pic de corrélation. À mesure que la valeur de α diminue, on constate une dispersion plus marquée des sélections, qui présente la non-linéarité plus signifiante. Particulièrement pour les oscillateurs de forte non-linéarité, par exemple dans le cas où α est égale à 0.3 (Figure 3.39), 0.2 (Figure 3.40) et 0.1 (Figure 3.41), l'étendue de la sélection est plus large. Cela suggère que cette méthode est particulièrement sensible à la capture de la non-linéarité accrue du système.

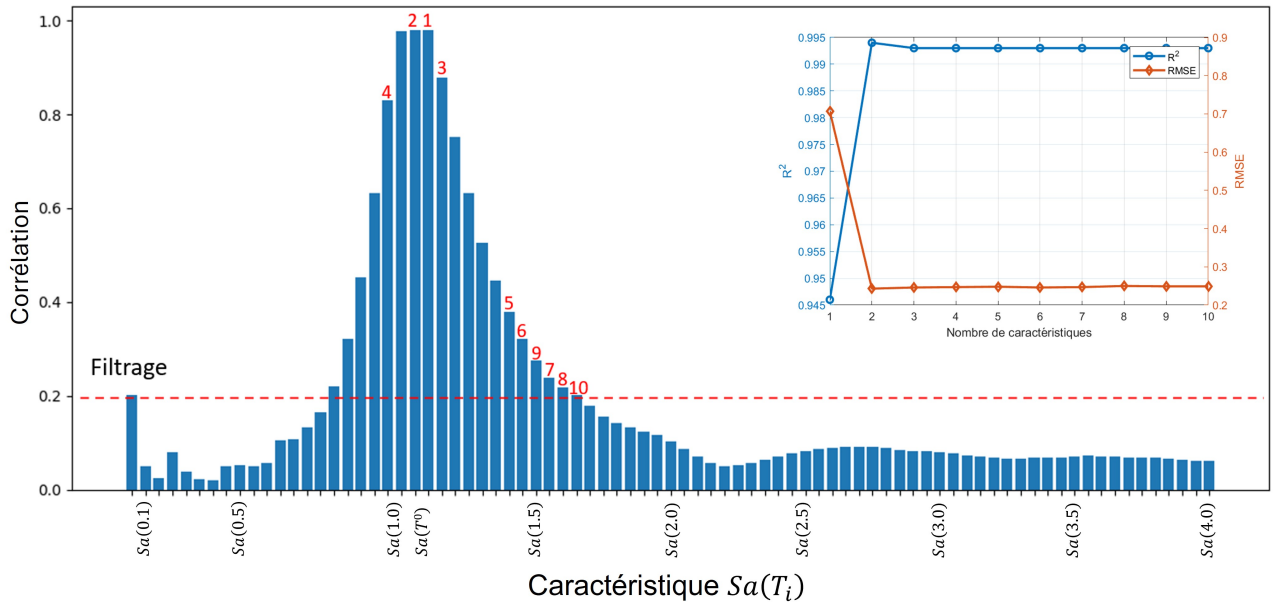


Figure 3.33 – [Bouc-Wen] Sélection pour α = 0.9

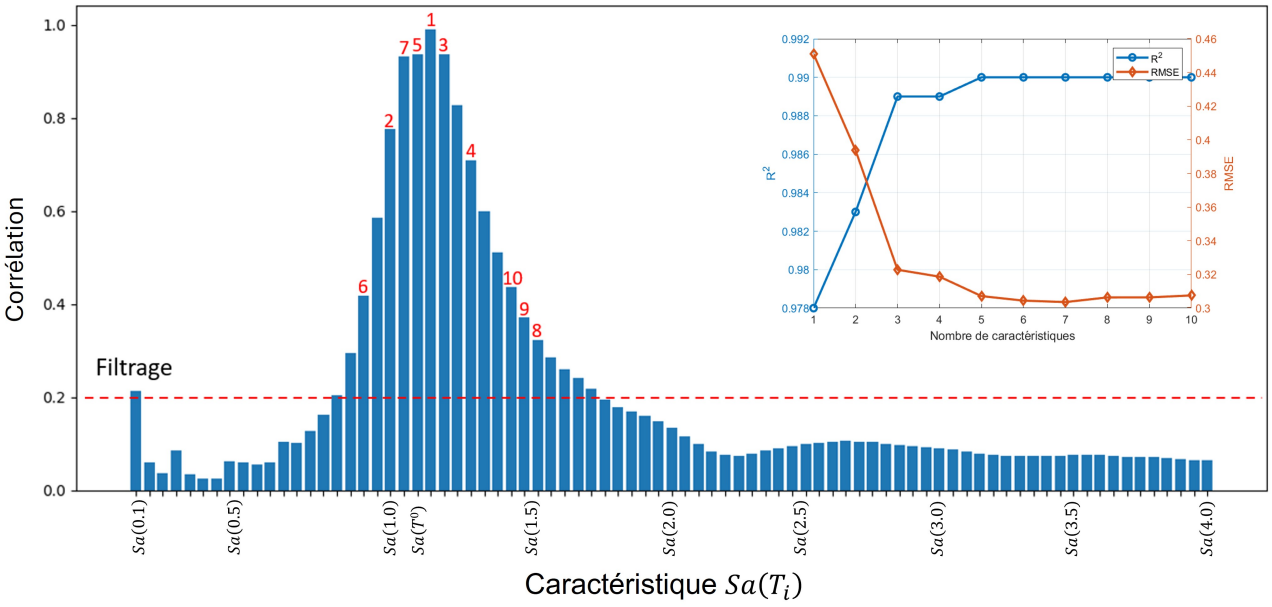


Figure 3.34 – [Bouc-Wen] Sélection pour α = 0.8

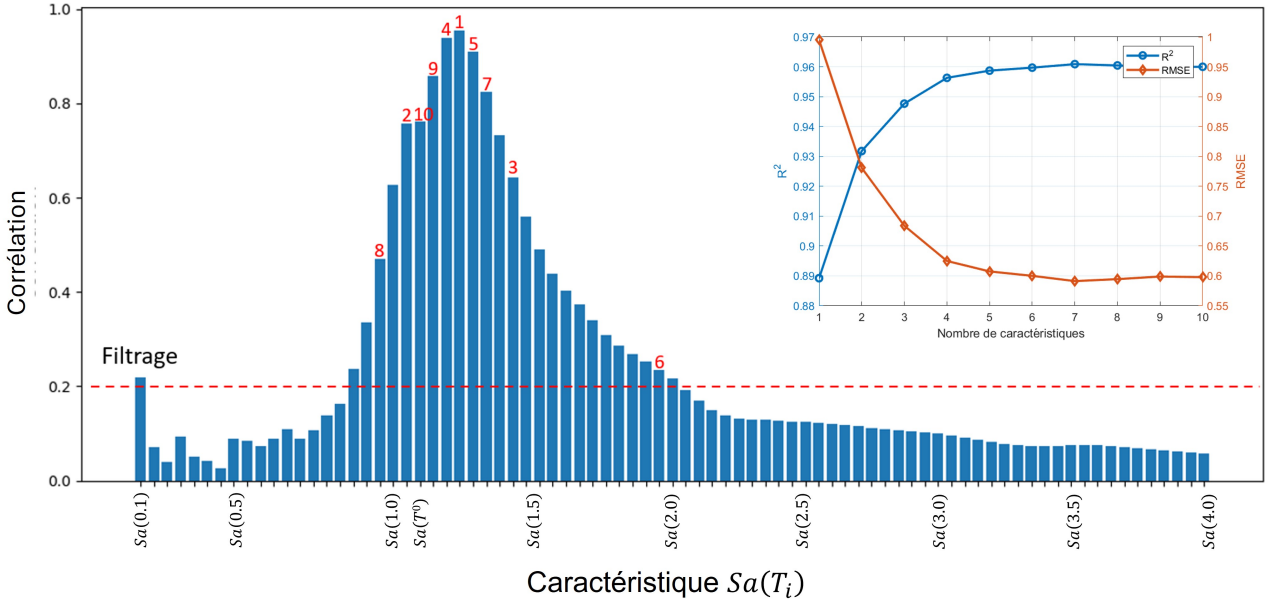


Figure 3.35 – [Bouc-Wen] Sélection pour α = 0.7

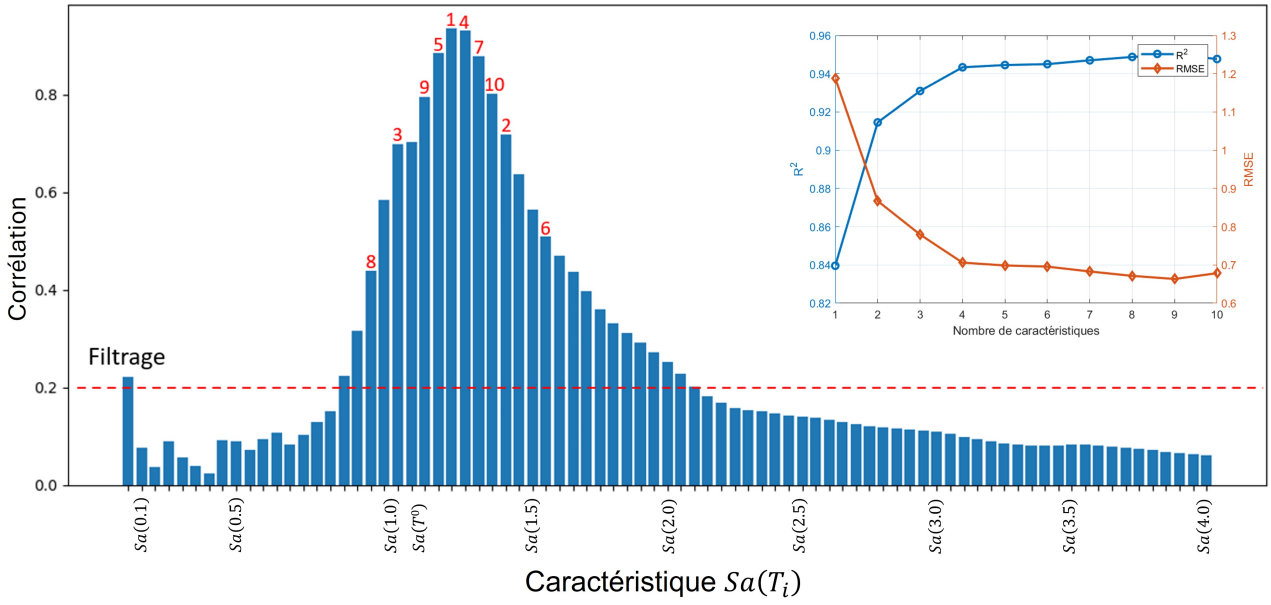


Figure 3.36 – [Bouc-Wen] Sélection pour α = 0.6

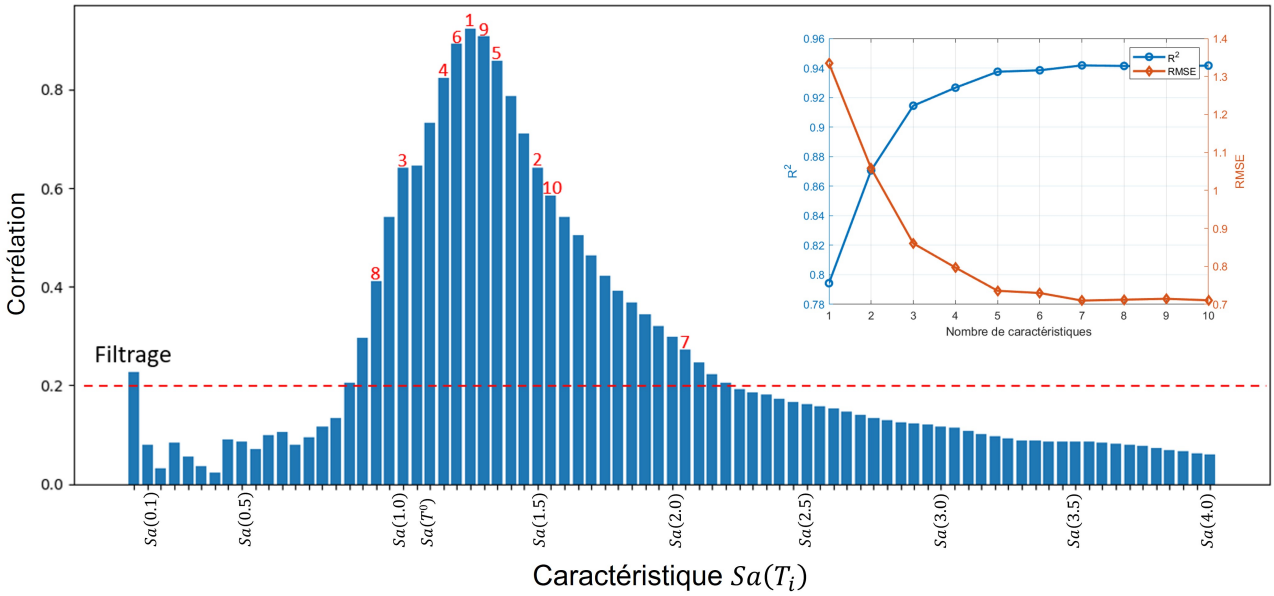


Figure 3.37 – [Bouc-Wen] Sélection pour α = 0.5

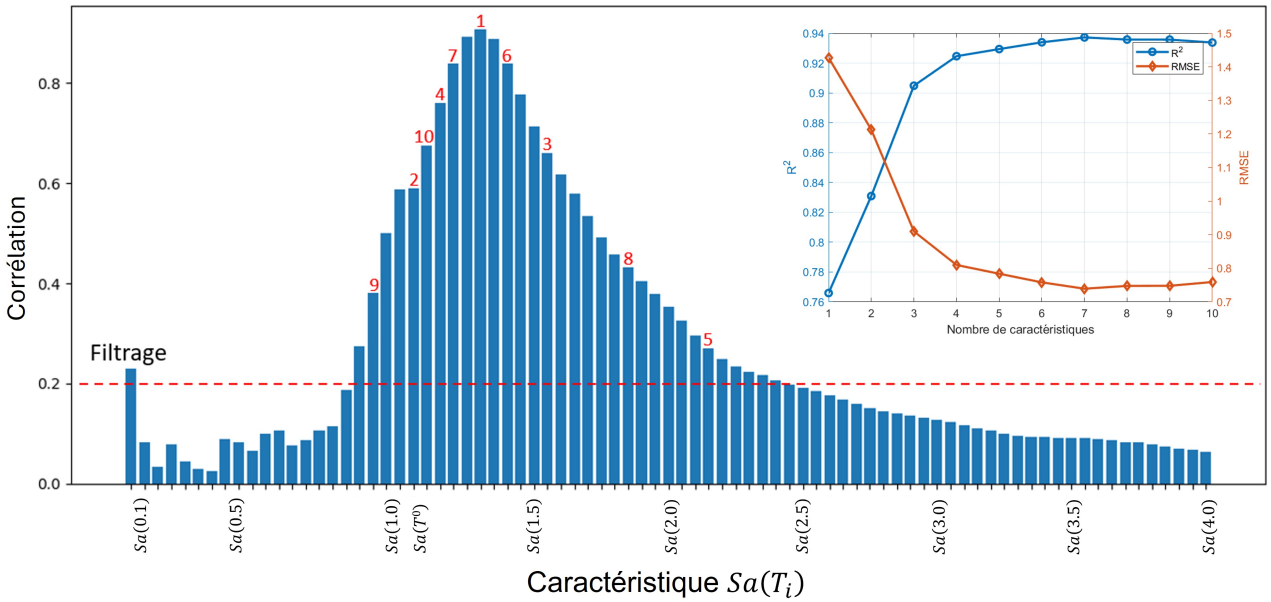


Figure 3.38 – [Bouc-Wen] Sélection pour α = 0.4

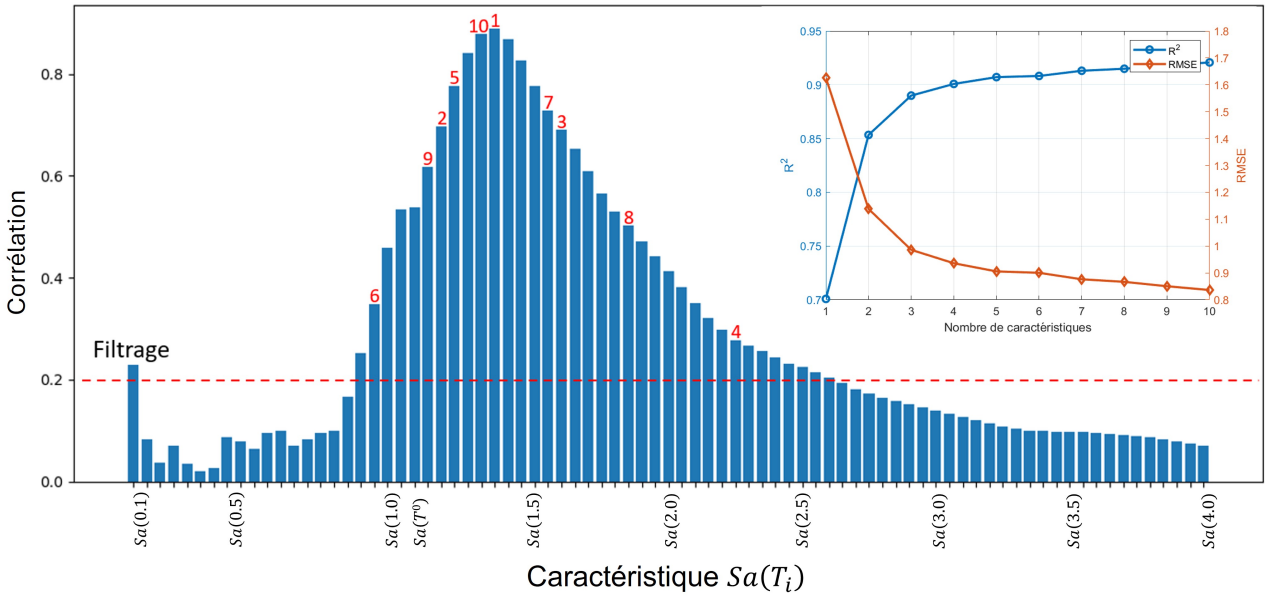


Figure 3.39 – [Bouc-Wen] Sélection pour α = 0.3

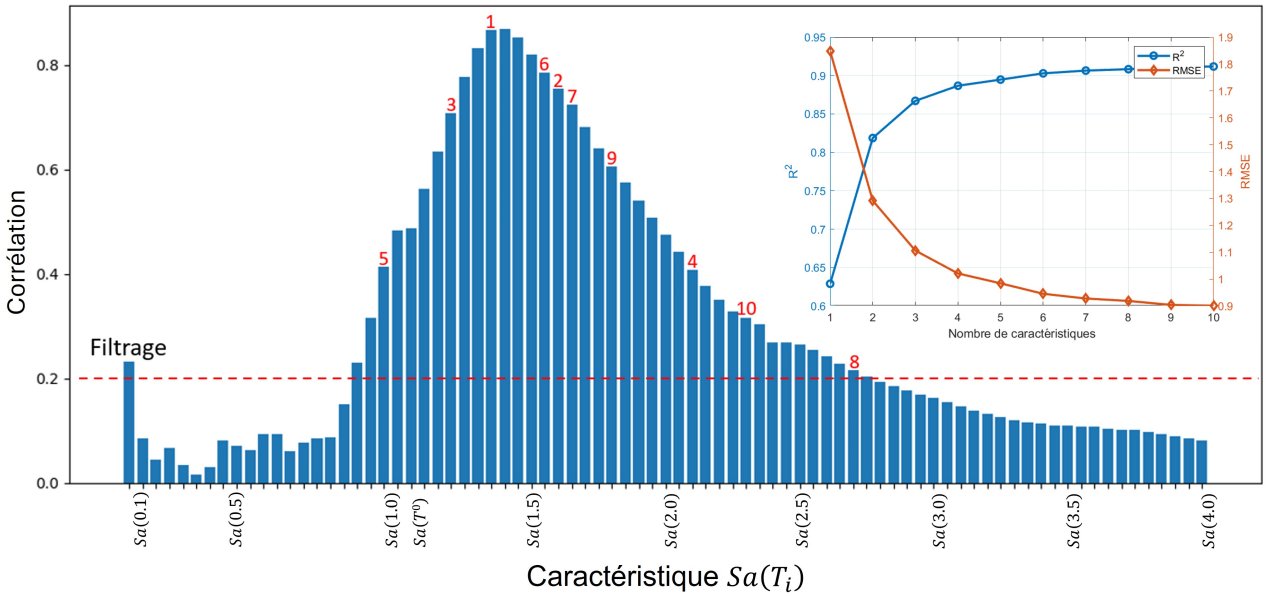


Figure 3.40 – [Bouc-Wen] Sélection pour α = 0.2

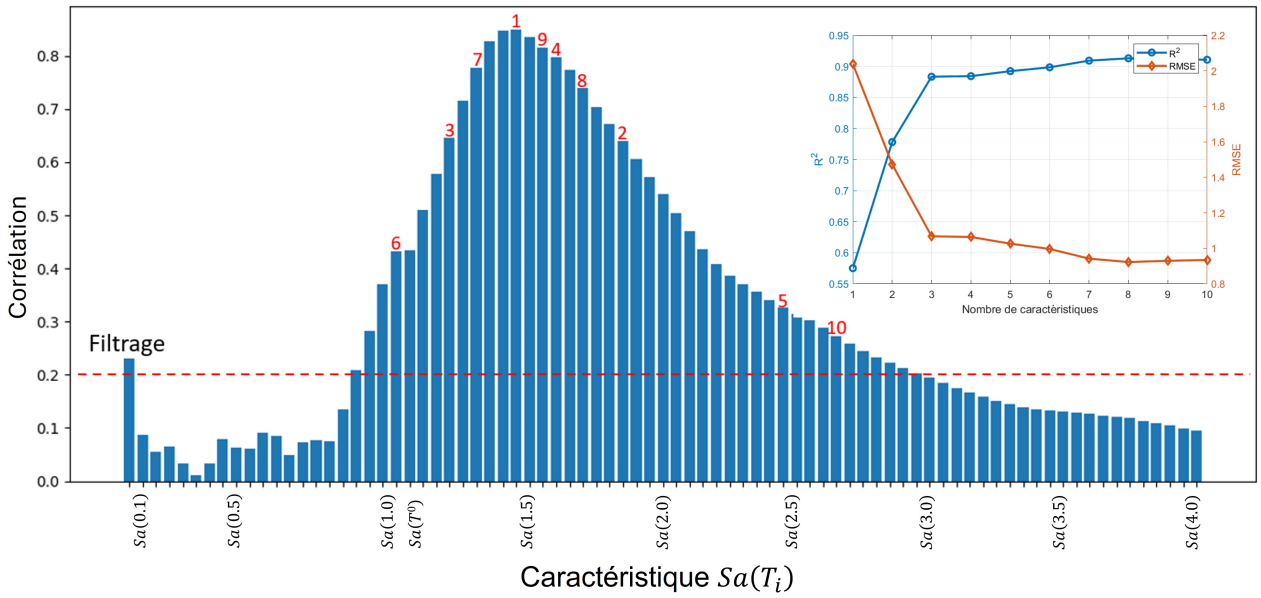


Figure 3.41 – [Bouc-Wen] Sélection pour α = 0.1

3.7. Conclusion

Une nouvelle méthode d'application d'apprentissage automatique pour modéliser les réponses sismiques des structures non-linéaires a été proposée. Elle repose sur les spectres de réponse en accélération considérés comme les caractéristiques des modèles. Au contraire des structures linéaires, les spectres des structures non-linéaires sont d'abord échantillonnés dans des zones plus étendues aux voisinages des modes propres et avec une résolution plus fine. Ils sont ensuite affinés grâce à la méthode de sélection hybride filtrage-enveloppe des caractéristiques.

La méthode proposée est détaillée sous forme d'une procédure pratique complète pas-à-pas PRO-NONLIN. Cette procédure a été validée avec des oscillateurs non-linéaires d'amortissement sec de Coulomb et de rigidité de Bouc-Wen. Elle a été également testée sur un exemple d'un portique de bâtiment non-linéaire à 8 étages. Plusieurs modèles d'apprentissage automatique ont été obtenus avec la procédure PRO-NONLIN. Il est possible de les classer en se basant sur les métriques d'ajustement et de précision. Il est à noter que la procédure PRO-NONLIN offre une réduction significative de temps grâce à la méthode de sélection hybride des spectres inclue dans la procédure.

En se basant sur les modèles d'apprentissage automatique obtenus, des courbes de fragilité sismique ont été construites par la méthode de Monte-Carlo et par la méthode de maximum de vraisemblance basée sur les observations de machine learning. Ces courbes sont comparées avec celles dérivées des méthodes conventionnelles en termes de précision et de temps de construction. Une réduction de temps très significative a été notée pour une précision meilleure ou équivalente des courbes de fragilité sismique à base de modèles d'apprentissage automatique.

L'ensemble des résultats du chapitre confirme la validité de la méthode proposée avec la procédure PRO-NONLIN aux structures non-linéaires.

4 Validation avec les enregistrements sismiques réels

Sommaire

4.1. Introduction

L'application de l'apprentissage automatique aux structures linéaires via la procédure PRO-LIN et aux structures non-linéaires via la procédure PRO-NONLIN a été proposée respectivement aux chapitres 2 et 3. Ces procédures ont été testées avec plusieurs exemples à un degré de liberté et à plusieurs degrés de liberté. Toutefois, les mouvements du sol dans ces tests de validation ne sont que des signaux sismiques synthétiques, tirés à partir d'un générateur sismique basé sur le modèle de Boore [107]. Pourtant, plusieurs études d'évaluation de risque sismique et des courbes de fragilité sismique dans la littérature se basent sur des enregistrements réels de mouvement du sol. L'objectif du chapitre est donc de vérifier la validité des procédures PRO-LIN et PRO-NONLIN aux cas des enregistrements sismiques réels.

La section 4.2 présente brièvement la sélection des signaux sismiques réels. En utilisant la base des enregistrements réels NGA-West2, les accélérations temporelles des mouvements réels du sol sont obtenues avec la méthode de spectre conditionnel. Ces signaux sont ensuite utilisés dans la section 4.3 pour tester d'abord la procédure PRO-LIN avec un oscillateur linéaire et puis la procédure PRO-NONLIN avec des oscillateurs non-linéaires de Coulomb et de Bouc-Wen.

Enfin, des commentaires sur la validité et la précision des modèles d'apprentissage automatique par rapport aux enregistrements réels sont formulés.

4.2. Sélection des enregistrements sismiques réels

4.2.1. Base de la sélection des enregistrements sismiques réels

La sélection des enregistrements sismiques dépend du spectre de réponse cible. Les enregistrements sélectionnés à partir du spectre cible ont une forme spectrale correspondant à celle du site d'intérêt. Le spectre d'aléa uniforme (uniform hazard spectrum, UHS) et le spectre de réponse conditionnel (conditionel response spectrum, CS) sont les spectres cibles les plus couramment utilisés.

Le spectre UHS est une enveloppe des accélérations spectrales avec un taux constant de dépassement à toutes les périodes. L'algorithme de sélection basé sur le spectre UHS sélectionne des enregistrements de manière à ce que le spectre moyen des enregistrements sélectionnés corresponde étroitement au spectre cible. Comme aucune variation par rapport à ce spectre cible UHS est attendue, le spectre UHS est conservateur. Selon Baker [116], en utilisant les enregistrements sismiques selon le spectre UHS, la simulation par éléments finis produit une estimation conservatrice de la demande sismique moyenne.

Le spectre CS a été proposé pour répondre au caractère conservateur du spectre UHS. Contrairement au spectre UHS, l'idée du spectre CS est de construire un spectre cible conditionné sur l'accélération spectrale à une période, $S_a(T^\star)$, avec une variance des autres périodes. La période $T^\star$ est souvent liée à la période propre de la structure considérée, qui affecte de manière significative sa réponse. Il est aussi possible de conditionner à une moyenne d'un vecteur de périodes, notée par

$AvgS_a(T^\star)$, selon Kohrangi *et al.* [117]. Il est de plus en plus approprié pour les applications dans l'évaluation des risques [118]. La procédure de sélection des enregistrements basée sur le spectre CS doit s'adapter à la moyenne et à la variance, contrairement à la procédure basée sur le spectre UHS correspondante, qui ne s'adapte qu'à la moyenne.

Dans le cadre de cette thèse, le spectre UHS est utilisé comme le spectre cible pour la génération des enregistrements synthétiques. Pour diversifier le choix des enregistrements, le spectre CS est utilisé pour la sélection des enregistrements réels. Plus précisément, la méthode consistant à des enregistrements selon le spectre CS introduite par Baker et Cornell [116] est employée.

Un algorithme, proposé par Jayaram *et al.* [119], est utilisé pour sélectionner des enregistrements sismiques selon le spectre CS. Premièrement, un spectre conditionnel doit être défini. Ce spectre conditionnel est la distribution de probabilité des valeurs logarithmiques du $S_a(T_i)$, qui sont conditionnées par rapport au spectre de réponse à la période d'intérêt $S_a(T^\star)$. Ensuite, des enregistrements sont sélectionnés à partir de la base de données. Ils sont ensuite mis à l'échelle afin de minimiser l'erreur entre le spectre de réponse des enregistrements sélectionnés et le spectre cible en termes de moyenne et de variance. Cependant, la moyenne et la variance des spectres sélectionnés peuvent rester légèrement différentes du spectre cible. Une technique d'optimisation est appliquée pour améliorer l'ensemble initialement sélectionné en termes de discordance. Selon cette direction, chaque mouvement du sol sélectionné précédemment est remplacé par un nouvel enregistrement dans la base de données, qui possiblement donne une amélioration du spectre moyen par rapport au spectre cible. Si les remplacements ne conduisent à aucune amélioration, l'ensemble d'enregistrements initialement sélectionné est conservé. Ainsi, l'ensemble des enregistrements le plus proche au spectre cible est obtenu.

4.2.2. Sélection des enregistrements

Selon le principe proposé par Jayaram *et al.* [119], la sélection des enregistrements a été effectuée à l'aide d'EzGM, une boîte à outils créée par Ozsarac [120].

L'objectif du logiciel EzGM est de fournir une boîte à outils libre (open-source) pour aider les ingénieurs à effectuer la sélection et le traitement des enregistrements sismiques. À cette fin, EzGM a été développé en utilisant le langage de programmation Python. Il comporte deux modules principaux, *EzGM.selection* pour la sélection et *EzGM.signal* pour le traitement des enregistrements. Le module de sélection *EzGM.selection* permet aux utilisateurs d'accéder directement aux enregistrements sélectionnés à partir des bases de données des mouvements sismiques comme NGA-West2 [65] ou ESM-2018 [121]. Il sélectionne des enregistrements sismiques en fonction d'un spectre cible. Ce module peut également être utilisé pour effectuer une sélection d'enregistrements basée sur les exigences de plusieurs codes du bâtiment : EN 1998-1 :2004, ASCE/SEI 7-16 et TBEC-2018. De plus, il propose aussi l'utilisation des modèles de mouvement du sol (ground motion model, GMM) pour construire le spectre cible. Ces modèles sont différents et disponibles dans la bibliothèque de OpenQuake [122]. Ces modèles de mouvement du sol définissent tous les scénarios causatifs dans la génération du spectre conditionnel cible. Pour EzGM, un GMM est défini en spécifiant un minimum de paramètres de rupture, de distance et de site : magnitude (M), angle de glissement, distance Joyner-Boore (R_{JB}),

vitesse moyenne des ondes de cisaillement sur une profondeur de 30 mètres (V_{S30}). Cet outil permet à l'utilisateur de définir aussi des limites concernant ces paramètres. Ensuite, le spectre cible est déduit à partir des paramètres fournis avec sa moyenne logarithmique et ses écarts-types sur tous les intervalles. Les enregistrements sismiques dans la base de données sont filtrés en fonction des limites prescrites. À partir de la base de données filtrée, selon l'algorithme proposé par Jayaram et al [119], les enregistrements sont sélectionnés et ajustés de manière à minimiser l'erreur par rapport au spectre cible. Leurs fonctionnalités sont expliquées plus en détail par Ozsarac [120].

Dans la suite du chapitre, la sélection des enregistrements est faite par le logiciel EzGM. Elle repose sur le modèle du mouvement du sol de Boore *et al.* [123] et le modèle de corrélation des ordonnées spectrales de Baker et Jayaram [124]. Dans le cadre de la thèse, la base de données NGA-West2 est employée. Cette sélection conditionne la période de $T^\star$ égal à 1 s. Les paramètres de rupture, de distance et de site sont la magnitude M égale à 7.0, l'angle de glissement égale à 0, la distance Joyner-Boore R_{JB} égale à 10 km, et la vitesse V_{S30} égale à 500 m/s. Cela représente un scénario de fort séisme sur un sol dur, selon Eurocode 8 [125]. La comparaison entre les enregistrements sélectionnés et le spectre cible selon ce scénario est représentée dans la Figure 4.1. La Figure 4.2 présente l'enregistrement RSN987-NORTHR-CEN245 sélectionné à partir de la base NGA-West2, selon le spectre cible présenté. Cet enregistrement est à l'origine du séisme Northridge-01 en 1994, mesuré par la station LA-Centinela St.

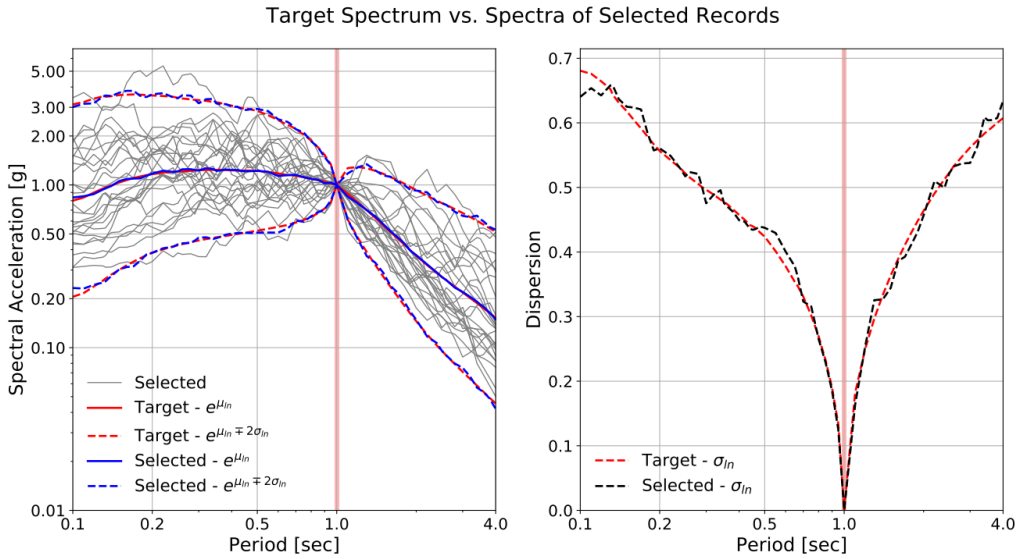


Figure 4.1 – Spectres des enregistrements sélectionnés par EzGM

4.3. Validation avec des structures sous séismes réels

Avec les enregistrements sismiques réels sélectionnés, différentes structures sont étudiées pour valider la procédure PRO-LIN et PRO-NONLIN. L'oscillateur linéaire, l'oscillateur non-linéaire de Coulomb, et l'oscillateur de Bouc-Wen sont utilisés. Leurs propriétés sont similaires avec celles présentées

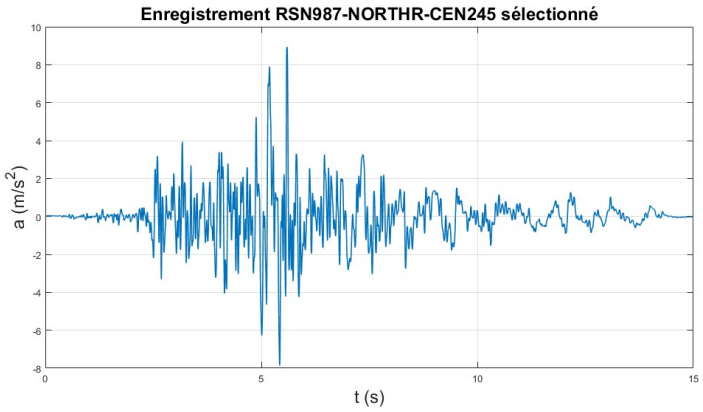


Figure 4.2 – Exemple de l'enregistrement sélectionné

dans le chapitre 2 et 3. Les modèles d'apprentissage automatique appliqués sont la régression linéaire (LR), les k-Plus proches voisins (KNN), la forêt aléatoire (RF), l'arbre de décision (DT), les machines à vecteurs de support (SVMR), les réseaux de neurones artificiels (ANN), l'algorithme de renforcement adaptatif (AdaBoost), la machine à renforcement léger de gradient (LightGBM) et la machine à renforcement extrême de gradient (XGBoost). Les critères d'évaluation sont la corrélation (r), le coefficient de détermination (R^2), l'erreur quadratique moyenne ($RMSE$) et l'erreur de pourcentage absolue moyenne symétrique ($SMAPE$) et l'indice de référence comme suggéré par Todorov *et al.* [32]. Les modèles d'apprentissage automatique utilisés dans cette étude sont mis en œuvre à l'aide du langage de programmation Python avec la bibliothèque TensorFlow [108] via JupyterLab.

4.3.1. Oscillateur linéaire

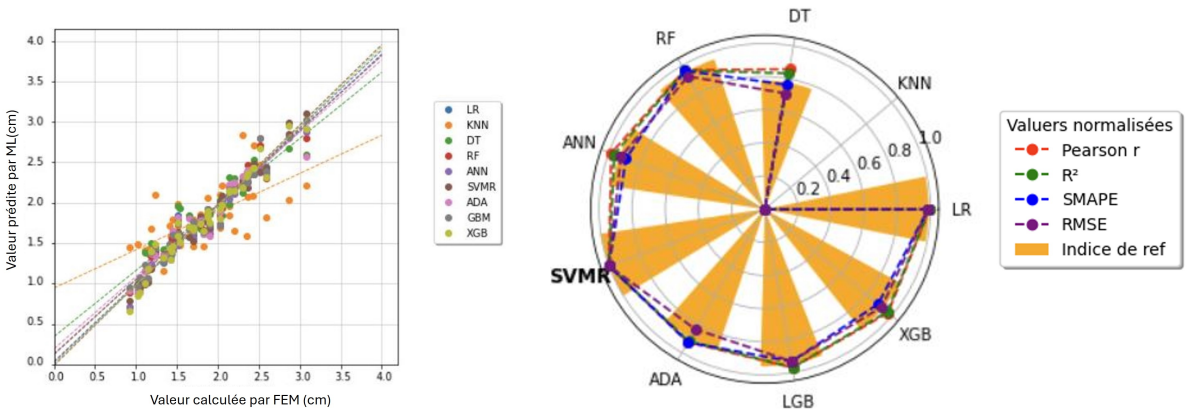
Dans ce cas, l'oscillateur linéaire présenté par la Figure 2.4 est utilisé. Les paramètres numériques de cet oscillateur sont la pulsation propre égale à 5.9 rad/s et un taux d'amortissement de 2 %. Les enregistrements sismiques sélectionnés à partir de la base de données NGA-West2 sont utilisés comme source d'excitation. La période fondamentale T^0 de l'oscillateur est égale à la période conditionnelle des signaux sismiques. La réponse de l'oscillateur est le déplacement maximal de la masse sous séisme, $\mathbf{Y} = \max_t |y(t)|$. 200 observations sont utilisées afin d'entrainer et tester les modèles.

Comme constaté dans le chapitre 2, les valeurs spectrales $S_a(T_i)$ autour de la période fondamentale du système sont les caractéristiques les plus importantes pour l'entrainement des modèles d'apprentissage automatique. Dans ce cas d'étude, une réduction des caractéristiques par l'échantillonnage autour de la période propre est réalisée. Cette sélection des caractéristiques présente non seulement un effort pour améliorer les modèles, mais aussi une correspondance entre l'apprentissage automatique et le mécanisme du système. Selon la procédure PRO-LIN, l'échantillonnage du spectre de réponse en accélération est réalisé sur l'intervalle de T_i entre 0.8 T^0 et 1.2 T^0 avec un pas de 0.1 T^0. Comme la période propre de l'oscillateur est de 1.05s, le choix des caractéristiques $S_a(T_i)$ est donc réalisé avec T_i entre 0.9 s et 1.3 s avec un pas de 0.1 s. Ces $S_a(T_i)$ sont utilisés comme caractéristiques pour entrainer les modèles de machine learning.

Modèle	r	R^2	$SMAPE$ (%)	$RMSE$ (cm)
LR	0.981	0.960	2.897	0.121
KNN	0.654	0.423	5.184	0.371
DT	0.936	0.872	3.696	0.228
RF	0.965	0.925	2.984	0.210
ANN	0.974	0.947	3.054	0.199
SVMR	0.983	0.963	2.730	0.110
AdaBoost	0.974	0.943	3.078	0.218
LightGBM	0.959	0.919	3.637	0.240
XGBoost	0.972	0.941	3.560	0.259

Table 4.1 – [Linéaire] Évaluation de la performance des modèles de ML

Le Tableau 4.1 montre la performance des modèles de machine learning sur les données utilisées. Chaque modèle est évalué en fonction des métriques r, R^2, $SMAPE$ et $RMSE$. En se basant sur ces métriques, le modèle SVMR est le plus performant parmi ceux évalués dans ce tableau, suivi par le modèle LR. Ils présentent des valeurs élevées pour r et R^2 et faibles pour $SMAPE$ et $RMSE$. Leurs métriques indiquent un bon ajustement des modèles aux données. Ce résultat est confirmé par les comparaisons dans la Figure 4.3. Les modèles d'apprentissage automatique, à part du type KNN, montrent une tendance proche à la tendance de référence (ligne inclinée à 45 degrés). Cela confirme la fiabilité des modèles de machine learning entrainés. Il faut noter que la performance des modèles dans ce cas est marginalement inférieure à celle des modèles utilisant des signaux synthétiques, présentée dans le Tableau 2.1 dans le chapitre 2. Cela s'explique par le fait que les signaux sismiques sélectionnés à partir des bases de données en fonction du spectre conditionnel sont moins conservateurs que ceux choisis en se basant sur le spectre d'aléa uniforme. Pour les résultats obtenus, l'utilisation des enregistrements réels est satisfaisante.



(a) Comparaison des valeurs prédites et actuelles **(b)** Comparaison des critères d'évaluation

Figure 4.3 – [Linéaire] Comparaison des modèles de ML utilisés

Afin de pouvoir valider la procédure PRO-LIN dans ce cas, un échantillonnage plus complet est aussi utilisé pour construire des modèles de machine learning. Les caractéristiques sont $S_a(T_i)$ avec T_i entre 0.1 s et 2.0 s avec un pas de 0.1 s. Ces $S_a(T_i)$ sont aussi utilisés comme caractéristiques pour entrainer les modèles de machine learning. L'étude de SHAP montre la contribution de chaque caractéristique, ici les valeurs spectrales $S_a(T_i)$, dans une prédiction de la réponse. Cette étude montre clairement que des valeurs spectrales des périodes autour de la période propre donnent plus de contribution que les autres périodes. Selon le résultat présenté dans la Figure 4.4, la stratégie de la procédure PRO-LIN est validée. L'importance des valeurs de spectre de réponse $S_a(T_i)$ autour de la période propre est confirmée par la physique et maintenant par des modèles d'apprentissage automatique.

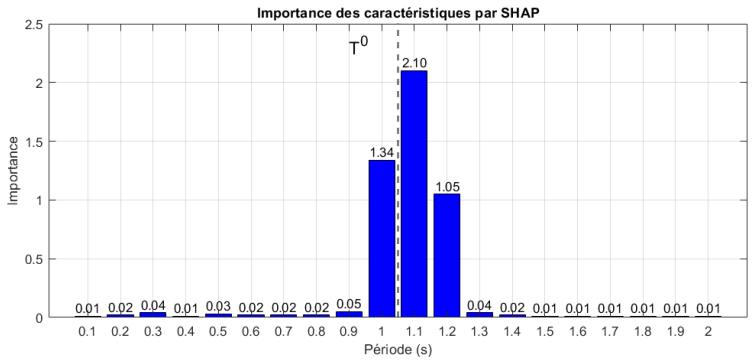


Figure 4.4 – [Linéaire] Contribution des caractéristiques selon l'étude de SHAP

De plus, le modèle SVMR est choisi pour voir la performance de la procédure PRO-LIN. Ce modèle est entrainé en utilisant le même nombre d'observations, échantillonnées seulement à $S_a(T_i)$ selon deux stratégies. La performance obtenue pour le modèle avec les caractéristiques optimisées SVMR2 est similaire avec celle entrainée les caractéristiques initiales SVMR1, montrée par le Tableau 4.2. Ce résultat permet de valider le choix des T_i proposé comme caractéristiques. De plus, cette procédure PRO-LIN réduit le nombre des caractéristiques de 20 à 5 et elle aide à améliorer l'efficacité du modèle.

Modèle	$S_a(T_i)$	r	R^2	$SMAPE$ (%)	$RMSE$
SVMR1	0.9 : 1.3 s	0.973	0.954	2.689	0.123
SVMR2	0.1 : 2.0 s	0.983	0.963	2.730	0.110

Table 4.2 – [Linéaire] Comparaison du modèle SVMR pour différentes nombres de caractéristiques

4.3.2. Oscillateur non-linéaire de Coulomb

L'oscillateur de Coulomb est étudié pour valider la procédure PRO-NONLIN avec la sélection par la méthode hybride illustrée par la Figure 3.18 pour cette étude.

Cet oscillateur est détaillé sur la Figure 3.1. Pour cette validation, la pulsation propre non-amortie ω_0 de l'oscillateur est choisie égale à 2 π rad/s, donc T^0 égale à 1 s. Avec un nombre limité de mouvements du sol, un comportement non-linéaire moyen, avec le coefficient μ égale à 0.01, est choisi.

Pour conduire cette analyse, un jeu de données comprenant des observations, générées avec les enregistrements sélectionnés depuis la base de données NGA-West2, est construit. Chaque observation représente la réponse de l'oscillateur de Coulomb soumis à une excitation sismique réelle. 80 % des données sont utilisées pour l'entrainement et 20 % pour le test des modèles.

Les caractéristiques $\mathbf{x}$ sont les valeurs du spectre de réponse en accélération $S_a(T_i)$, dont chaque période T_i est échantillonnée de 0.05 s à 2 s par incrément de 0.05 s ; ainsi que la valeur du spectre de réponse à la période propre du système, $S_a(T^0)$. Ce choix d'échantillonnage, conduisant à 41 caractéristiques distinctes, assure une description des signaux sismiques.

La réponse de l'oscillateur est le déplacement maximal de la masse, donc $\mathbf{Y} = \max_t |y(t)|$. Le nombre de caractéristiques est semblable à celui utilisé dans la validation avec des enregistrements synthétiques. La procédure PRO-NONLIN est appliquée pour pouvoir optimiser le nombre de caractéristiques pour entrainer des modèles. L'équation (4.1) présente l'algorithme de machine learning avec les caractéristiques initialement échantillonnées $\mathbf{x}$ et la réponse $\mathbf{Y}$.

$$\mathbf{Y} = \mathbf{ML}(\mathbf{x}) + \epsilon, \quad \mathbf{x} = \{S_a(T_i), S_a(T^0)\}, T_i = [0.05 : 0.05 : 2] \tag{4.1}$$

Le nombre d'observations requis pour l'entraînement du modèle est étudié. Le modèle RF est choisi du fait de sa précision et de sa facilité d'application. La Figure 4.5 présente l'évolution des métriques, $RMSE$ et R^2, selon le nombre d'observations. Il est apparu que le nombre de 50 et 100 observations est manifestement insuffisant pour obtenir un apprentissage convergent. À mesure que le nombre d'observations augmente, une amélioration des métriques est observée. Cependant, au-delà de 500 observations, ces métriques atteignent un plateau, ne démontrent plus de variations significatives. Pour cet oscillateur de Coulomb, 500 observations constituent un jeu de données adéquat pour l'entraînement du modèle.

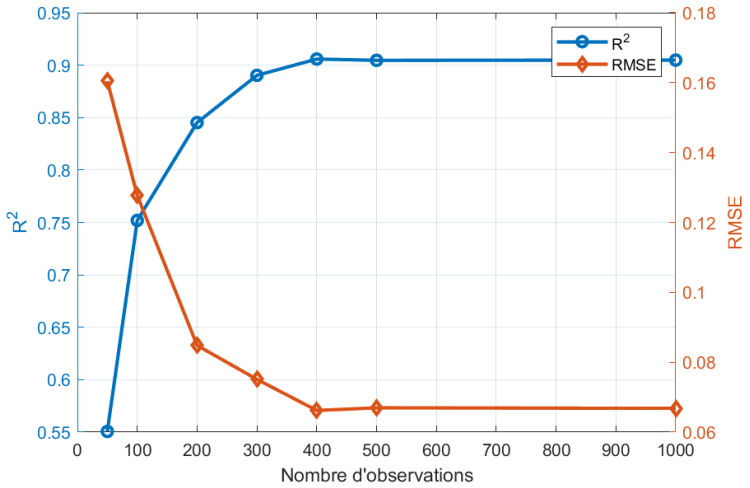


Figure 4.5 – [Coulomb] Évolution des métriques selon le nombre d'observations

Pour trouver le meilleur ensemble de caractéristiques afin d'entraîner le modèle, la procédure PRO-NONLIN, présentée dans la Figure 3.18, a été appliquée à l'ensemble initial des 41 caractéristiques. Dans cette démarche, la sélection par méthode de filtrage a été utilisée pour éliminer les caractéristiques $\mathbf{x}$ les moins corrélées par rapport à la réponse $\mathbf{Y}$. Ensuite, la sélection par méthode d'enveloppe a été appliquée pour déterminer le nombre optimal de caractéristiques à retenir. La Figure 4.6 illustre l'évolution de deux métriques, R^2 et $RMSE$, en fonction du nombre de caractéristiques sélectionnées par la méthode hybride. Cette analyse permet de choisir 6 caractéristiques pour entraîner le modèle RF final. Une comparaison de performance a été effectuée entre ce modèle final avec seulement six caractéristiques et celui avec l'ensemble des 41 caractéristiques. Les résultats présentés dans le Tableau 4.3 démontrent des performances similaires entre ces deux modèles, confirmant l'efficacité de la procédure PRO-NONLIN proposée dans cette thèse.

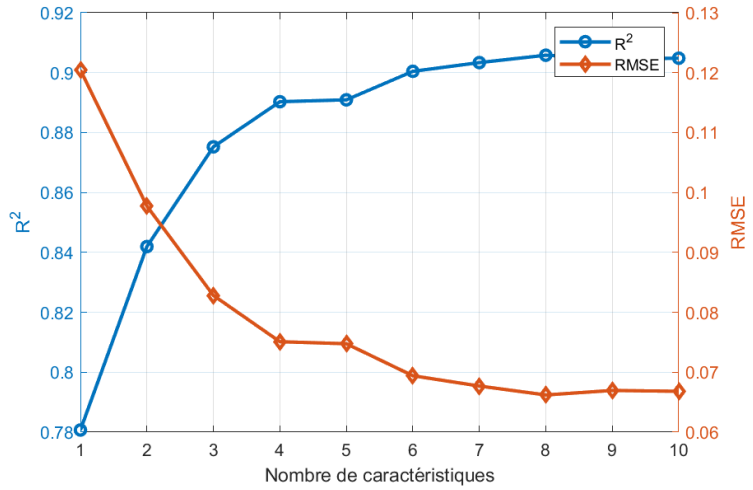


Figure 4.6 – [Coulomb] Évolution des métriques selon le nombre de caractéristiques

Modèle	**Nombre de caractéristiques**	r	R^2	$SMAPE$ (%)	$RMSE$
RF1	41	0.963	0.909	2.103	0.063
RF2	6	0.953	0.901	2.189	0.084

Table 4.3 – [Coulomb] Comparaison du modèle LR pour différentes nombres de caractéristiques

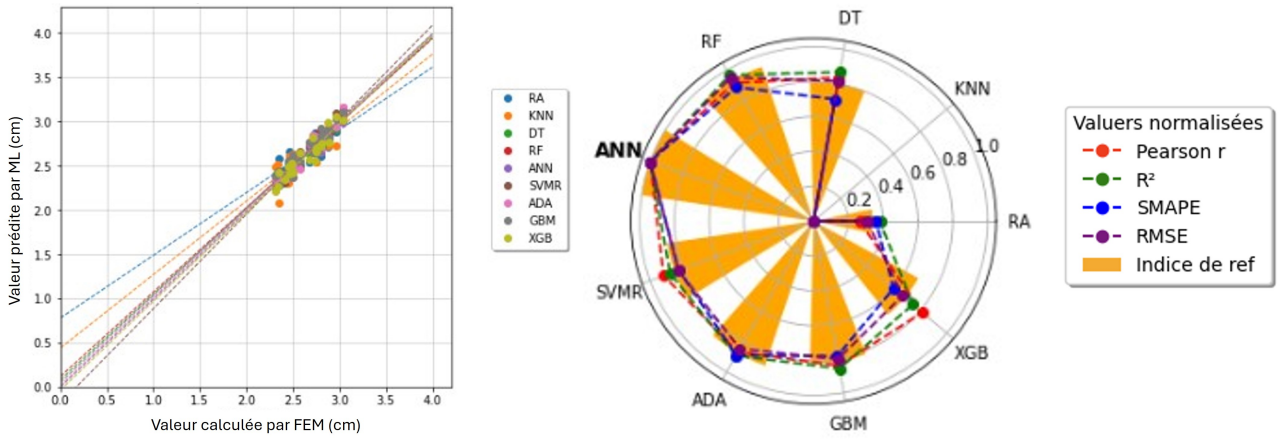
Finalement, la procédure PRO-NONLIN est appliquée pour tous les modèles de machine learning. La performance des modèles est détaillée dans le Tableau 4.4. Les modèles ANN, RF et XGBoost ont les valeurs de r et R^2 plus élevées, indiquant une bonne capacité à prédire les réponses. Le modèle ANN a les valeurs de $SMAPE$ et $RMSE$ les plus faibles, ce qui suggère une meilleure précision du modèle. En résumé, le modèle ANN est le meilleur modèle pour ce cas. Les modèles RF, XGBoost et aussi RF sont aussi efficaces en raison des valeurs de R^2 élevées, de $SMAPE$ et de $RMSE$ faibles.

En outre, les modèles sont comparés selon la Figure 4.7. La comparaison des modèles est réalisée en se basant sur ces quatre métriques d'évaluation et l'indice de référence, illustrée par un graphe

Modèle	r	R^2	$SMAPE$ (%)	$RMSE$
LR	0.851	0.723	3.485	0.109
KNN	0.811	0.609	4.256	0.130
DT	0.938	0.869	2.727	0.075
RF	0.950	0.899	2.345	0.066
ANN	0.963	0.909	2.103	0.063
SVMR	0.950	0.872	2.479	0.075
XGBoost	0.943	0.876	2.342	0.073
AdaBoost	0.939	0.866	2.561	0.076
LightGBM	0.935	0.832	2.964	0.085

Table 4.4 – [Coulomb] Évaluation de la performance des modèles de ML

polaire dans la Figure 4.7b. Cette comparaison signifie que le modèle de type ANN est le plus efficace. Ces modèles utilisent aussi les enregistrements de test pour prédire des réponses correspondantes de l'oscillateur. Ces réponses prédites par machine learning sont comparées avec celles par simulation mécanique. Cette comparaison est illustrée dans la Figure 4.7a. La comparaison montre que les tendances des prédictions sont proches à la tendance de référence, qui est inclinée à 45 degrés. C'est-à-dire les prédictions sont proches aux valeurs actuelles. Ce résultat signifie que les modèles de machine learning sont bien validés et ils sont capables de donner des prédictions fiables.



(a) Comparaison des valeurs prédites et actuelles **(b)** Comparaison des critères d'évaluation

Figure 4.7 – [Coulomb] Comparaison des modèles de ML

4.3.3. Oscillateur non-linéaire de Bouc-Wen

Dans cette troisième validation, l'oscillateur de Bouc-Wen, présenté par la Figure 3.2, est utilisé. Pour cet oscillateur, la pulsation propre non amortie ω_0 est définie à 2 π rad/s, le taux d'amortissement

ξ est égale à 2 %. De plus, les constantes sont choisies avec C_1 = 1, C_2 = C_3 = 0.5 cm^{n_d}, et n_d = 1. Le coefficient α qui caractérise la non-linéarité du système est choisi égale à 0.7.

De la même manière, le jeu de données généré avec les enregistrements sélectionnés depuis la base de données NGA-West2, est construit. Chaque observation représente la réponse de l'oscillateur soumis à une excitation sismique. 80 % de données sont réservées pour entrainer les modèles et 20 % pour les tester.

Les caractéristiques $\mathbf{x}$ sont les valeurs du spectre de réponse en accélération $S_a(T_i)$. Comme constaté dans le chapitre 3, la non-linéarité de Bouc-Wen est une non-linéarité de rigidité vers les périodes plus grandes. Cette non-linéarité cause un décalage du pic de corrélation entre les caractéristiques $\mathbf{x}$ et la réponse $\mathbf{Y}$. C'est la raison pour laquelle il faut échantillonner sur une intervalle de $S_a(T_i)$ plus étendu. Les périodes T_i sont échantillonnées de 0.05 s à 4 s par incrément de 0.05 s. Le spectre de réponse à la période propre du système $S_a(T^0)$ est aussi échantillonné. Ce choix d'échantillonnage, conduisant à 81 caractéristiques distinctes, assure une description complète de la réponse spectrale de la structure.

Le nombre d'observations requis pour l'entraînement des modèles d'apprentissage automatique est aussi étudié pour cet oscillateur de Bouc-Wen. Le modèle de type RF est réutilisé. La Figure 4.8 présente l'évolution des métriques d'évaluation selon le nombre d'observations. Il s'est avéré que le nombre initial de 50 observations est clairement insuffisant pour obtenir un modèle efficace. Une amélioration des métriques est observée à mesure que le nombre d'observations augmente. 500 observations sont considérées comme suffisantes pour entraîner le modèle RF pour cet oscillateur de Bouc-Wen. Il faut aussi noter que la non-linéarité de l'oscillateur est moyenne, donc le nombre d'observations requis est faible. Pour les oscillateurs plus non-linéaires, le nombre requis est possiblement plus élevé.

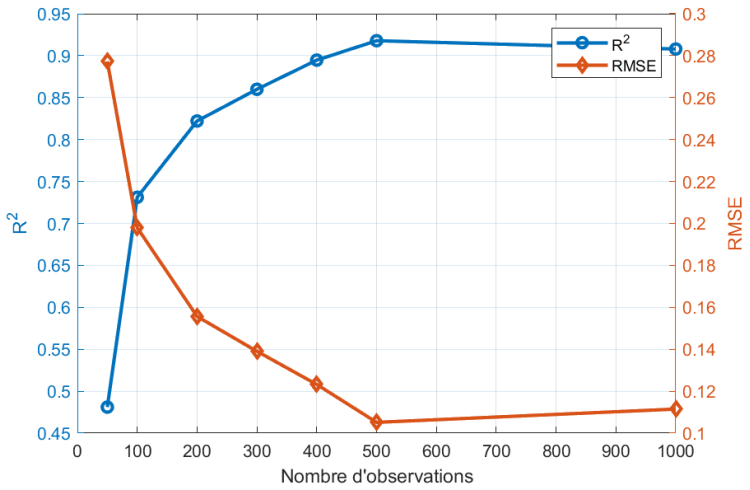


Figure 4.8 – [Bouc-Wen] Évolution des métriques selon le nombre d'observations

La procédure PRO-NONLIN est appliquée pour trouver un meilleur ensemble des caractéristiques pour entrainer le modèle RF. A partir de 81 caractéristiques, un filtrage est appliqué pour supprimer celles les moins corrélées. Ensuite, la sélection des caractéristiques par la méthode d'enveloppe est

réalisée pour obtenir un nombre optimal des caractéristiques. La Figure 4.9 montre l'évolution de deux métriques $RMSE$ et R^2 selon le nombre de caractéristiques sélectionnées par la méthode hybride. En observant cette figure, 8 caractéristiques sont choisies pour entrainer le modèle RF final. Ce modèle final RF2 est comparé avec le modèle complet RF1 qui utilise toutes les caractéristiques. Le Tableau 4.5 montre la comparaison des métriques pour ces deux modèles. Les performances similaires montrent l'efficacité de la proposition de la thèse.

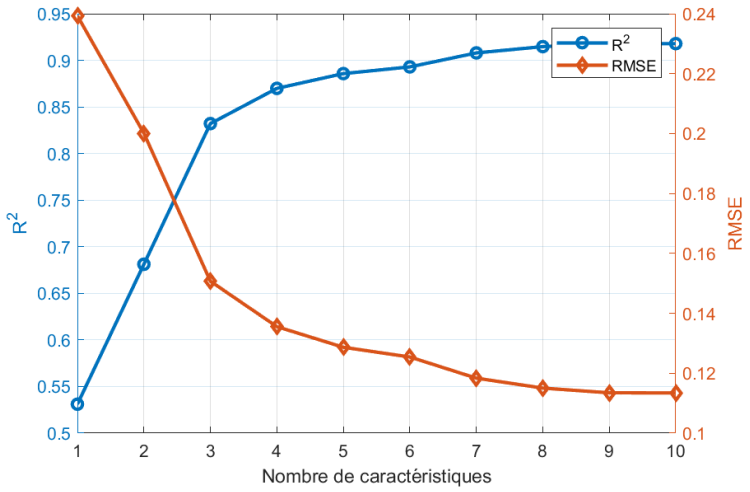


Figure 4.9 – [Bouc-Wen] Évolution des métriques selon le nombre de caractéristiques

Modèle	Nombre de caractéristiques	r	R^2	$SMAPE$ (%)	$RMSE$
RF1	81	0.964	0.915	3.743	0.114
RF2	8	0.961	0.915	3.859	0.115

Table 4.5 – [Bouc-Wen] Comparaison du modèle LR selon le nombre de caractéristiques

Les différents modèles de ML sont finalement construits pour ce système, en utilisant le nombre d'observations obtenu et la procédure PRO-NONLIN proposée. Les résultats sont détaillés dans le Tableau 4.6.

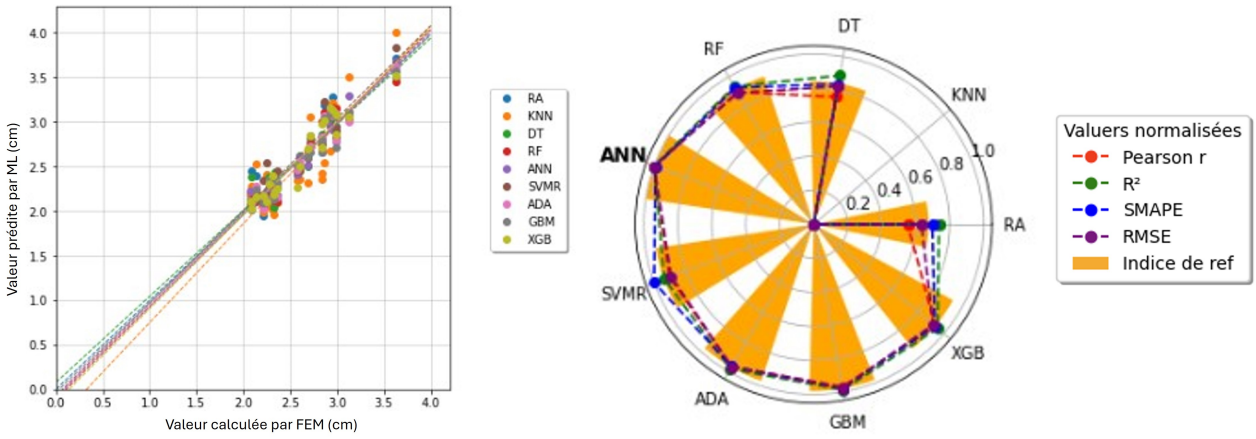
Les modèles ANN, XGBoost et SVMR ont les meilleures performances globales, avec des valeurs de r et R^2 élevées, des erreurs $SMAPE$ et $RMSE$ faibles. Les modèles RF et LightGBM montrent également de bonnes performances, bien qu'un peu moins bonnes. Le choix du modèle dépendra des exigences spécifiques de l'application, mais les modèles ANN, SVMR, XGBoost et AdaBoost semblent être de bons candidats pour des prévisions précises.

En outre, les modèles sont comparés selon la Figure 4.10. La comparaison des modèles est réalisée en se basant sur ces quatre métriques d'évaluation et l'indice de référence qui est la somme pondérée des métriques. Cette comparaison est illustrée par un graphe polaire dans la Figure 4.10b. Cette comparaison signifie que le modèle de type ANN est le plus efficace dans ce cas. Après avoir

Modèle	r	R^2	$SMAPE$ (%)	$RMSE$
LR	0.947	0.946	4.892	0.180
KNN	0.873	0.725	7.025	0.196
DT	0.934	0.920	5.863	0.146
RF	0.951	0.963	4.213	0.133
ANN	0.974	0.965	3.743	0.114
SVMR	0.962	0.954	3.872	0.133
XGBoost	0.965	0.960	3.727	0.120
AdaBoost	0.960	0.949	3.832	0.117
LightGBM	0.966	0.951	3.187	0.129

Table 4.6 – [Bouc-Wen] Évaluation de performance des modèles de ML

bien entrainé les modèles de machine learning, ils sont utilisés pour prédire des réponses de l'oscillateur sous chaque enregistrement de test. Ces réponses prédites par machine learning sont comparées avec celles obtenues par simulation. Cette comparaison est illustrée dans la Figure 4.10a. Cette comparaison montre que les tendances des prédictions sont proches de la tendance de référence inclinée à 45 degrés, c'est-à-dire les prédictions sont proches des valeurs actuelles. Par conséquent, les modèles de machine learning construits pour ce cas sont fiables.



(a) Comparaison des valeurs prédites et actuelles **(b)** Comparaison des critères d'évaluation

Figure 4.10 – [Bouc-Wen] Comparaison des modèles de ML utilisés

4.3.4. Oscillateurs non-linéaires de Bouc-Wen avec la non-linéarité variable

a. Présentation du cas d'étude

Dans cette validation finale, un autre type d'application d'apprentissage automatique est testé. Un jeu de données est construit en observant un ensemble de structures excitées par une ensemble de séismes. Ce type d'étude est observé dans certaines références. Par exemple, Zhang *et al.* [75] ont utilisé 9900 observations pour entrainer différents modèles de machine learning afin de prédire l'état d'endommagement des portiques en béton armé. Cette étude a porté sur 198 structures en béton armé représentatives des bâtiments existants en Chine. Elles comprenaient 90 structures de trois étages, 72 de six étages et 36 de huit étages, couvrant divers paramètres de conception structurelle comme les périodes de construction, le nombre d'étages, le nombre de travées, la longueur de travées, la hauteur d'étage. 50 enregistrements sismiques horizontaux recommandés par FEMA [126] ont été choisis pour évaluer la réponse dynamique des structures. Ces enregistrements de mouvements sismiques comprenaient 28 enregistrements de champs proches et 22 enregistrements de champs lointains. Guan *et al.* [127] construisent une base de données avec 621 structures sous 240 signaux sismiques par la simulation d'éléments finis. Ces structures sont dans des archétypes de ces portiques en acier, sous les séismes recommandés par Miranda [128].

Dans notre cas, l'étude est dédiée à la validation de la procédure PRO-NONLIN avec des enregistrements réels en utilisant un ensemble d'oscillateurs de Bouc-Wen. La non-linéarité des oscillateurs varie en fonction du coefficient α. Le jeu de données contient les observations sur 10 oscillateurs de Bouc-Wen dont la valeur de α est entre 0.1 et 1.0 avec un pas de 0.1. Cette utilisation conduit à un ensemble de 10 oscillateurs de Bouc-Wen à étudier. 100 signaux sismiques sont sélectionnés depuis la base de données NGA-West2 pour solliciter ces oscillateurs. Le jeu de données contient alors 1000 observations. Ces observations sont divisées en deux, 80 % des données sont utilisées pour entrainer les modèles et 20 % sont utilisés pour les tester. Les caractéristiques des modèles de ML sont liées à la structure et aux séismes. La caractéristique de type structure est α. Les caractéristiques de type spectrale $\mathbf{x}$ sont les valeurs du spectre de réponse en accélération $S_a(T_i)$, avec T_i est échantillonné entre 0.1 s et 4.0 s par incrément de 0.1 s. C'est-à-dire $\mathbf{x}$ contient $S_a(T_i)$, avec T_i = [0.1 : 0.1 : 4] s. La réponse $\mathbf{Y}$ est le déplacement maximal des oscillateurs, donc $\mathbf{Y} = \max_t |y^{\alpha}(t)|$. Les caractéristiques initiales sont résumées sur Figure 4.11.

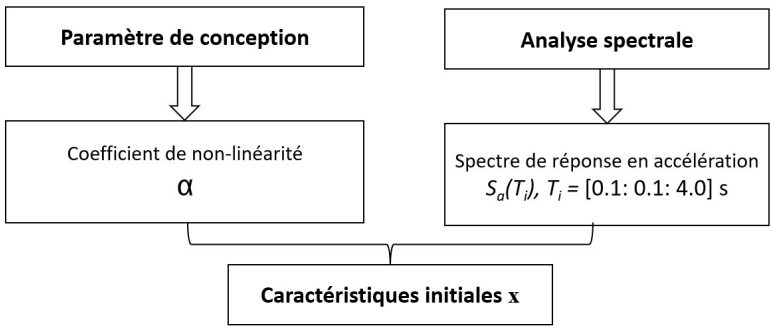


Figure 4.11 – Caractéristiques initiales $\mathbf{x}$ pour les oscillateurs de Bouc-Wen

Il contient le paramètre de conception des oscillateurs, α, et ceux de mouvement du sol, $S_a(T_i)$. L'équation (4.2) représente les modèles d'apprentissage automatique pour cette validation :

$$\mathbf{Y} = \mathbf{ML}(\mathbf{x}) + \epsilon, \quad \mathbf{x} = \{\alpha, S_a(T_i)\}, \quad T_i = [0.1 : 0.1 : 4] \tag{4.2}$$

b. Application de la procédure PRO-NONLIN pour cette base de données

Le modèle d'apprentissage automatique de type RF est utilisé comme illustration. Une étude sur le nombre d'observations pour entrainer le modèle RF est réalisée. Cette étude a pour but de trouver le nombre optimal d'observations pour entrainer ce modèle. Le nombre d'observations à considérer est entre 100 et 1000 observations. La Figure 4.12 illustre l'évolution des métriques d'évaluation, qui sont R^2 et $RMSE$, selon le nombre d'observations testées. La tendance de R^2 à partir de 800 observations semble converger. La valeur de $RMSE$, pour un nombre d'observations de 800 est suffisante pour entrainer le modèle.

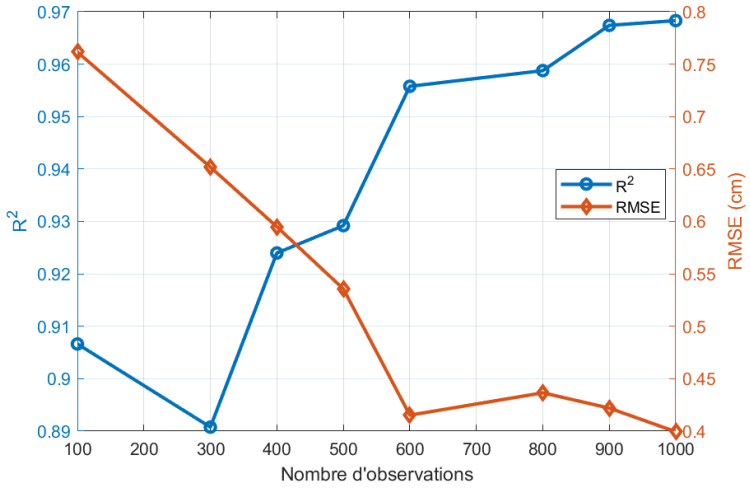


Figure 4.12 – Évolution du modèle selon le nombre d'observations

La sélection des caractéristiques par la méthode hybride est employée. Cette sélection des caractéristiques commence avec 41 caractéristiques, comme le montre Figure 4.11. Il faut noter que la méthode hybride est appliquée pour réduire les caractéristiques liées aux excitations sismiques, illustrée par Figure 4.13. Une première sélection par filtrage est appliquée. Cette sélection ne s'applique que sur les caractéristiques $S_a(T_i)$. Selon les valeurs de corrélation, un seuil de filtrage de 0.4 est appliqué pour pouvoir enlever les caractéristiques les moins importantes. C'est la raison pour laquelle les caractéristiques entre 0.1 s et 1.2 s sont éliminées par cette étape. Il faut aussi noter que la période propre de l'oscillateur est de 1 s. Cela montre que le spectre de réponse en accélération à la période propre, $S_a(T^0)$, très souvent utilisé comme caractéristique, n'est certainement pas la caractéristique la plus importante pour tous les cas.

La sélection par la méthode d'enveloppe est utilisée pour trouver les caractéristiques les plus optimales pour ce cas après la première sélection. Le coefficient α est aussi considéré dans cette

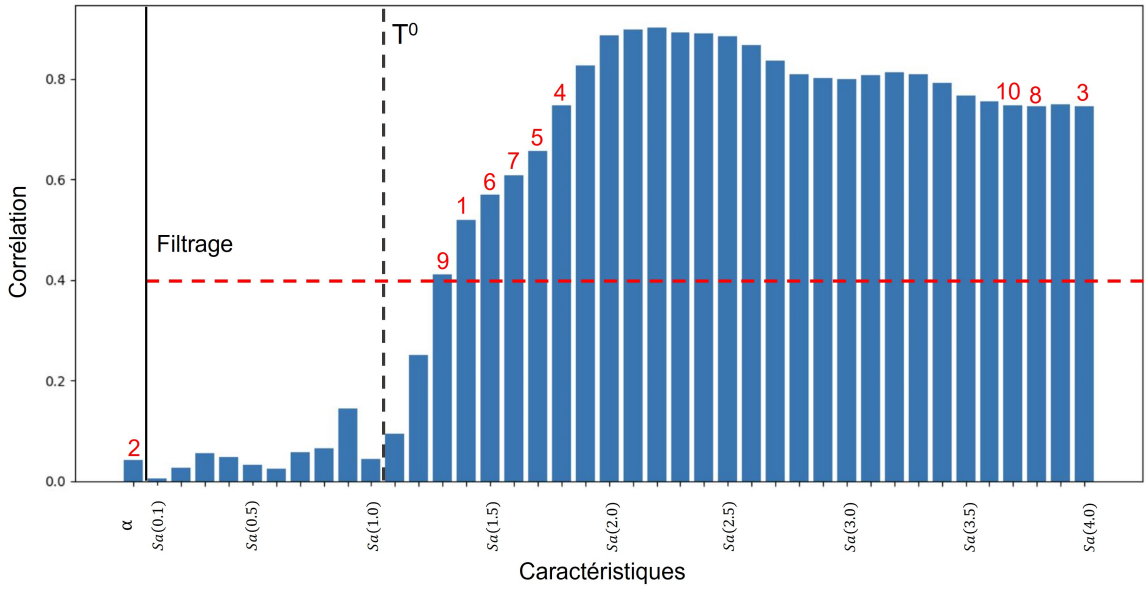


Figure 4.13 – Sélection des caractéristiques par méthode hybride

étape. Le nombre de caractéristiques à explorer dans cette étude est entre 1 et 10. La Figure 4.13 montre les caractéristiques sélectionnées pour chaque itération avec leurs ordres. Selon l'ordre de sélection, les caractéristiques de type spectrale sont celles les plus importantes. Cela montre que les caractéristiques liées aux enregistrements du sol ont plus d'impact lors de l'entrainement du modèle. Le coefficient α est aussi sélectionné, ce qui indique que la représentation du système contribue aussi dans l'entrainement. Les barres montrent la corrélation entre les caractéristiques $\mathbf{x}$ et la réponse $\mathbf{Y}$. Cette corrélation entre $\mathbf{x}$ et $\mathbf{Y}$ est linéaire, tandis que la relation exacte entre eux est non-linéaire. C'est la raison pour laquelle les caractéristiques les plus corrélées ne sont pas toujours prises en compte dans ce cas. La Figure 4.14 montre qu'avec 4 caractéristiques, le modèle de type RF est entrainé avec une bonne performance. A partir de 4 caractéristiques, l'ajout des caractéristiques n'augmente plus significativement la performance du modèle.

Finalement, les différents modèles de machine learning sont entrainés, en utilisant 1000 observations et la procédure PRO-NONLIN. Le Tableau 4.7 présente la performance des modèles selon les quatre critères d'évaluations r, R^2, $RMSE$ et $SMAPE$. En comparant les modèles à partir de ces résultats, le modèle de type ANN obtient la meilleure performance pour presque toutes les mesures. Il a une valeur de R^2 élevée, avec des valeurs de $SMAPE$ et de $RMSE$ faibles, ce qui suggère qu'il fait de meilleures prédictions que les autres modèles pour cette tâche. RF, DT et SVMR suivent de près le modèle ANN en termes de performance, mais il a légèrement plus d'erreurs que le modèle ANN. De plus, le modèle KNN montre aussi une bonne performance. Elle diffère de celle des études précédentes, car la nature du modèle est plus adaptée pour cette étude. Le modèle LR a la performance la moins bonne parmi les modèles examinés. Cela est dû à la non-linéarité des données.

La Figure 4.15 montre une comparaison entre des modèles de machine learning entrainés en utilisant les mêmes outils que les études précédentes. La Figure 4.15b présente une comparaison

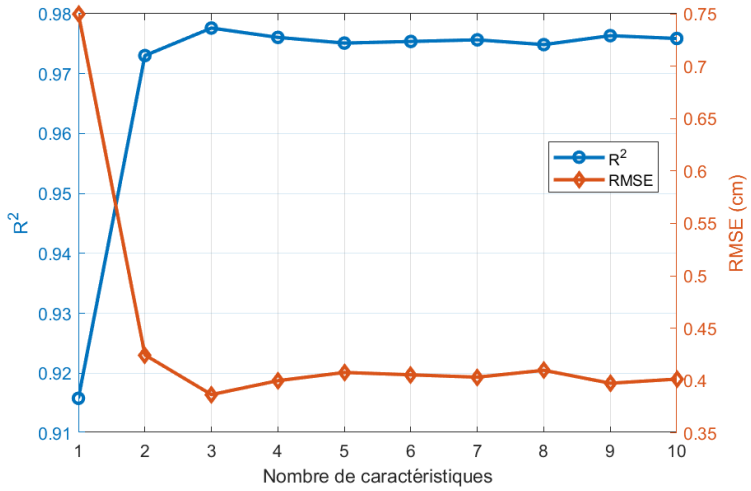
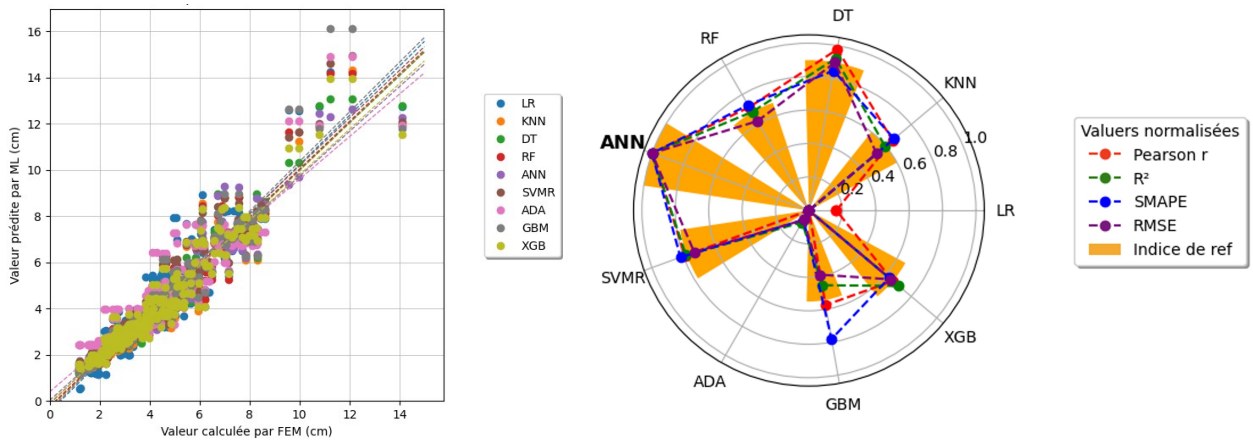


Figure 4.14 – Évolution des métriques selon le nombre de caractéristiques

Modèle	r	R^2	$SMAPE$ (%)	$RMSE$ (cm)
LR	0.934	0.821	17.39	0.87
KNN	0.962	0.908	9.604	0.624
DT	0.971	0.931	8.527	0.541
RF	0.975	0.935	8.39	0.511
ANN	0.974	0.941	7.795	0.503
SVMR	0.958	0.913	9.064	0.604
XGBoost	0.947	0.883	11.443	0.702
AdaBoost	0.927	0.830	16.627	0.848
LightGBM	0.932	0.835	13.282	0.833

Table 4.7 – Évaluation de la performance des modèles de ML

entre des modèles sur le graphe polaire, où les métriques r, R^2, $SMAPE$ et $RMSE$ sont normalisées entre 0 et 1. L'indice de référence, comme présenté dans les chapitres précédents, est calculé comme la somme pondérée des métriques avec un taux respectivement égal à 10 %, 30 %, 30 %, 30 %. Ce graphe montre clairement la performance supérieure du modèle ANN pour ce cas d'étude en comparaison avec les autres modèles. Une comparaison directe entre les valeurs réelles (obtenues par simulation par éléments finis) et les valeurs prédites (obtenues par des modèles de machine learning) est réalisée avec un nuage de points, présentée sur la Figure 4.15a. Une ligne droite de tendance est tracée pour chaque comparaison. Selon cette figure, toutes les lignes sont inclinées de presque 45 degrés, ce qui signifie une bonne précision des modèles. Par contre, les valeurs aberrantes sont quelques fois mal prédites, mêmes pour les modèles performants comme l'arbre de décision ou la forêt aléatoire. Cela signifie un manque potentiel d'observations dont la valeur est critique.



(a) Comparaison des valeurs prédites et actuelles **(b)** Comparaison des critères d'évaluation

Figure 4.15 – Comparaison des modèles de ML utilisés

4.4. Conclusion

Les procédures PRO-LIN et PRO-NONLIN ont été vérifiées et validées dans ce chapitre avec les séismes réels qui sont sélectionnés à partir de la base de données des enregistrements réels NGA-West2, par la méthode de spectre conditionnel.

La validation a été réalisée sur les oscillateurs linéaires et non-linéaires. Les résultats montrent que des modèles d'apprentissage automatique de bonne précision ont été obtenus. Ils sont similaires à ceux obtenus à partir des enregistrements synthétiques mais avec une précision plus faible. Cela pourrait s'expliquer par le processus de sélection des séismes réels par la méthode de spectre conditionnel qui est moins conservatrice que l'approche génératrice basée sur le spectre uniforme.

En utilisant les enregistrements réels, la procédure PRO-LIN reconfirme l'importance des spectres autour des périodes propres, tandis la procédure PRO-NONLIN prouve l'efficacité de la sélection des spectres les plus pertinents pour les modèles d'apprentissage automatique. A cause d'un nombre d'enregistrements réels limité, un modèle d'apprentissage automatique validé ne peut générer qu'une base de données de taille limitée. C'est une des limites en utilisant les enregistrements sismiques réels.

5 Validation avec les données disponibles dans la littérature

Sommaire

5.1. Introduction

La nouvelle méthode d'application d'apprentissage automatique a été proposée avec deux procédures pratiques PRO-LIN et PRO-NONLIN dédiées respectivement aux structures linéaires et non-linéaires.

Ces procédures ont été validées pour des oscillateurs et des systèmes à plusieurs degrés de liberté soumis à des séismes synthétiques et réels. Les exemples de validation sont assez variés mais ils restent pourtant académiques. Il est donc temps de confronter les procédures PRO-LIN et PRO-NONLIN à des données nouvelles, soit par le type de structures, soit par la nature des données.

Dans cette optique, une maquette préalablement recalée par des mesures expérimentales d'un bâtiment à cinq étages en béton armé est d'abord utilisée pour vérifier la procédure PRO-LIN. Il s'agit d'une structure linéaire proposée par O. Sahin *et al.* [129]. La validation de la procédure est décrite à la section 5.2.

La section 5.3 est réservée à la validation de la procédure PRO-NONLIN pour laquelle une base de données élaborée par Guan *et al.* [127] est utilisée. Il convient de noter que cette base de données a déjà été employée pour construire les modèles d'apprentissage automatique dans les références [76, 77].

Quelques conclusions sur la validité des procédures par rapport à ces nouvelles données sont finalement exprimées en fin de chapitre.

5.2. Validation de la procédure PRO-LIN par un modèle analytique

5.2.1. Présentation de la structure

Pour valider la procédure PRO-LIN pour les structures linéaires, la maquette proposée par O. Sahin *et al.* [129] est employée. Il s'agit d'un bâtiment de cinq étages en béton armé, présenté sur la Figure 5.1. La Figure 5.2 illustre sa maquette construite dans le laboratoire avec sa modélisation mécanique. Chaque étage du bâtiment est représenté par une plaque de 800 mm (longueur) $\times$ 600 mm (largeur) $\times$ 15 mm (épaisseur). Chaque plaque de la maquette est soutenue par quatre colonnes en acier de section rectangulaire de 6 mm $\times$ 15 mm. La dimension la plus petite des colonnes est colinéaire avec la direction de la force d'excitation. La hauteur des colonnes est de 300 mm. Les colonnes sont soudées à la plaque. Les limites d'élasticité du matériau des colonnes et de la plaque sont respectivement de 750 MPa et 235 MPa. Des boulons ont été utilisés pour fixer la base à la table vibrante. Deux barres d'attache sont articulées entre toutes les plaques pour minimiser le mouvement horizontal (direction z) de la maquette. Des masses d'acier supplémentaires sont aussi ajoutées à chaque plaque de la maquette pour avoir les mêmes fréquences que celles du bâtiment en béton armé de référence.

Figure 5.1 – Bâtiment en béton armé de référence

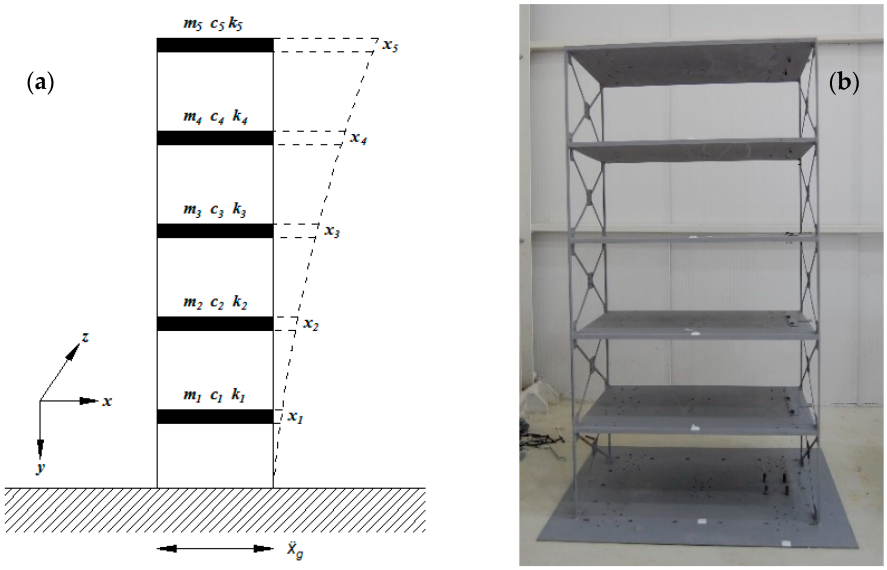


Figure 5.2 – (a) Modèle mécanique de la structure et (b) sa maquette

A partir de la dimension de la maquette, la matrice de masse, qui considère la masse des plaques, des colonnes et des masses supplémentaires, s'exprime :

$$\mathbf{M} = \begin{bmatrix} 70.57 & 0 & 0 & 0 & 0 \\ 0 & 70.57 & 0 & 0 & 0 \\ 0 & 0 & 70.57 & 0 & 0 \\ 0 & 0 & 0 & 70.57 & 0 \\ 0 & 0 & 0 & 0 & 70.57 \end{bmatrix} \text{ kg}$$

La rigidité en flexion de chaque colonne est calculée par l'équation (5.1) :

$$k_i = \frac{12EI}{H^3} \tag{5.1}$$

Dans cette équation, le module d'élasticité ($E = 2 \times 10^{11}$ N/m^2), le moment d'inertie ($I = 2.7 \times 10^{-10}$ m^4), et la hauteur de chaque colonne ($H = 0.3$ m) sont représentés par des variables E, I et H respectivement. A partir de ces propriétés, la matrice de rigidité de la maquette est construite :

$$\mathbf{K} = \begin{bmatrix} 192000 & -96000 & 0 & 0 & 0 \\ -96000 & 192000 & -96000 & 0 & 0 \\ 0 & -96000 & 192000 & -96000 & 0 \\ 0 & 0 & -96000 & 192000 & -96000 \\ 0 & 0 & 0 & -96000 & 96000 \end{bmatrix} \text{ N/m}$$

Un essai vibratoire libre est réalisé pour déterminer le taux d'amortissement de la maquette. La méthode du décrément logarithmique [94] est appliquée sur la réponse de la structure. Le taux d'amortissement est calculé par la formule (5.2) :

$$\xi = \frac{1}{2m\pi} \frac{u_n}{u_{n+m}} \tag{5.2}$$

où u_n et u_{n+m} sont respectivement les amplitudes de déplacement à l'instant t quelconque et à l'instant $t + mT$, où m est le nombre de périodes T de vibration libre.

Finalement, la matrice d'amortissement est déduite et présentée comme suite :

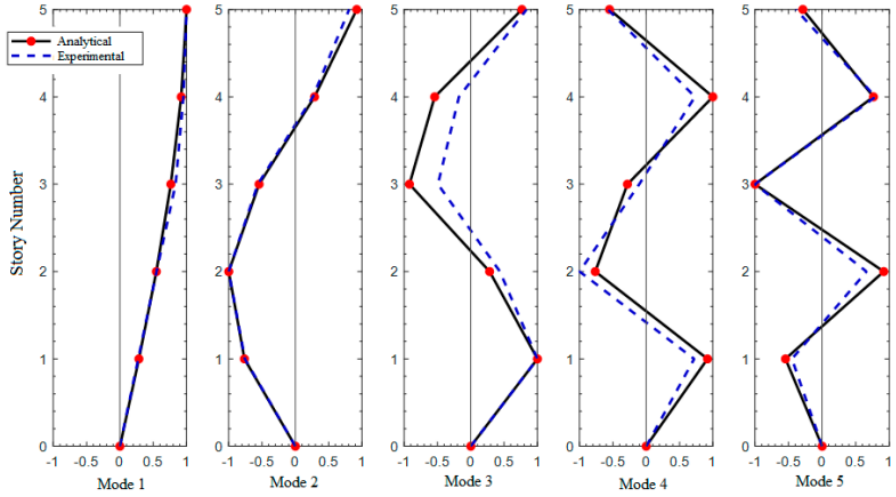
$$\mathbf{C} = \begin{bmatrix} 50.72 & -23.59 & 0 & 0 & 0 \\ -23.59 & 50.72 & -23.59 & 0 & 0 \\ 0 & -23.59 & 50.72 & -23.59 & 0 \\ 0 & 0 & -23.59 & 50.72 & -23.59 \\ 0 & 0 & 0 & -23.59 & 27.108 \end{bmatrix} \text{ Ns/m}$$

Les fréquences naturelles sont calculées analytiquement par l'équation dynamique. Puis, elles sont obtenues expérimentalement par la décomposition fréquentielle à partir de la réponse libre. Les imprécisions ou incertitudes au sein des modèles analytiques peuvent ou non représenter avec précision le comportement structurel réel. Étant donné la fiabilité des tests expérimentaux, il est conseillé de recaler le modèle analytique du système de bâtiment en intégrant les données dérivées des résultats de mesure. Ce recalage permet de corriger les écarts entre les modèles analytiques et les résultats expérimentaux en modifiant les matrices analytiques originales, telles que la rigidité et l'amortissement. Le logiciel MATLAB/Simulink est utilisé pour établir une corrélation entre le modèle analytique et la réponse dynamique mesurée du modèle de bâtiment. Cette corrélation permet de rapprocher la déformée modale entre le modèle analytique et la maquette. Les paramètres dynamiques ont éventuellement été modifiés en employant cette approche pour produire des résultats plus proches de la dynamique structurelle réelle de la maquette. La matrice de rigidité et d'amortissement modifiées sont :

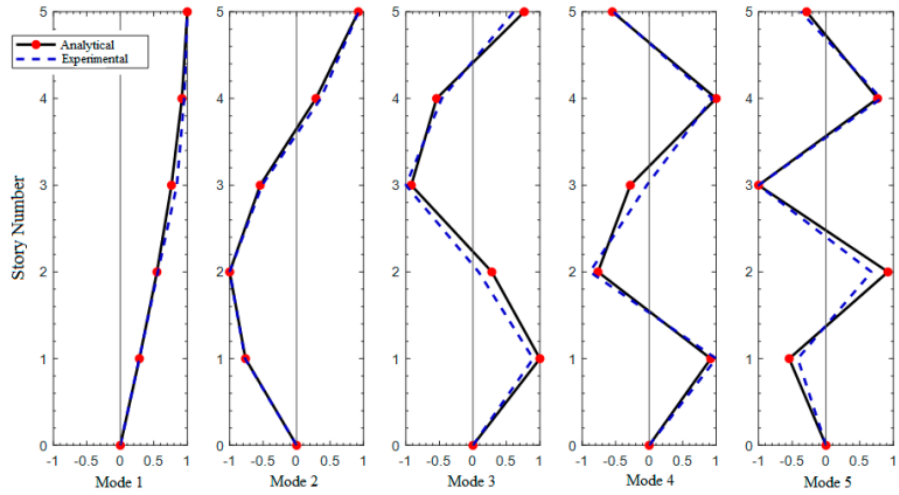
$$\mathbf{K} = \begin{bmatrix} 176632 & -88316 & 0 & 0 & 0 \\ -88316 & 176632 & -88316 & 0 & 0 \\ 0 & -88316 & 176632 & -88316 & 0 \\ 0 & 0 & -88316 & 176632 & -88316 \\ 0 & 0 & 0 & -88316 & 88316 \end{bmatrix} \text{ N/m}$$

$$\mathbf{C} = \begin{bmatrix} 152.752 & -76.376 & 0 & 0 & 0 \\ -76.376 & 152.752 & -76.376 & 0 & 0 \\ 0 & -76.376 & 152.752 & -76.376 & 0 \\ 0 & 0 & -76.376 & 152.752 & -76.376 \\ 0 & 0 & 0 & -76.376 & 152.752 \end{bmatrix} \text{ Ns/m}$$

La Figure 5.3 montre les déformées modales avant et après le recalage.



(a) Avant



(b) Après

Figure 5.3 – Comparaison de la déformée modale entre le modèle et la maquette [129]

Les fréquences naturelles des cinq premiers modes de la maquette, obtenues par le calcul analytique et par les essais, sont comparées dans le Tableau 5.1. Le Tableau 5.1 présente aussi les fréquences de recalage de la maquette.

	Fréquences naturelles (Hz)		
Mode (Selon x)	Analytique	Expérimentale	Recalage
1	1.67	1.61	1.60
2	4.88	4.83	4.68
3	7.69	8.00	7.38
4	9.87	9.95	9.48
5	11.27	11.38	10.81

Table 5.1 – Fréquences naturelles analytiques, expérimentales et recalées de la maquette

5.2.2. Validation de la procédure PRO-LIN

Pour entrainer des modèles de machine learning dans ce cas d'étude, un jeu de données contenant 500 observations est généré en utilisant 500 enregistrements réels. Ces enregistrements sont similaires à ceux du chapitre 4, choisis à partir de la base de données NGA-West2 selon le modèle de Boore [123] et le modèle de corrélation spectrale de Baker et Jayaram [124]. Parmi ces 500 observations, 400 observations sont utilisées pour entrainer les modèles de ML et les 100 observations restantes sont utilisées pour tester les modèles.

Selon les résultats obtenus dans le chapitre 2, les valeurs de spectre de réponse en accélération autour des périodes propres, spécialement de la période fondamentale, sont plus importantes pour construire l'ensemble des caractéristiques. Le Tableau 5.2 détaille les périodes propres T_i, les facteurs de participation modale, L_i, et les masses modales de la maquette, L_i^2. Selon Paultre [94], le calcul de L_i par superposition modale est donné à l'équation (5.3) :

$$L_i = \phi_i^{\mathsf{T}} \mathbf{M} \mathbf{r} \tag{5.3}$$

où ϕ_i est la déformée modale du mode i et $\mathbf{r}$ est la vecteur d'influence qui contient pour ce cas les composantes égales à 1. Les valeurs de L_i négatives présentent une direction inverse du chargement du mode i par rapport à la direction de l'excitation sismique au sol.

Mode (Selon x)	ω_i (rad/s)	T_i (s)	L_i	L_i^2 (kg)
1	10.069	0.624	17.617	310.342
2	23.392	0.214	-5.546	30.761
3	46.333	0.136	2.923	8.545
4	59.521	0.106	1.628	2.648
5	67.886	0.093	-0.744	0.553

Table 5.2 – Pulsation naturelle, période propre, facteur de participation et masse modale de chaque mode

Selon le Tableau 5.2, la masse modale effective de la période fondamentale prend 87.95 % de la masse totale. Parce que la période fondamentale est plus dominante, l'échantillonnage des valeurs

spectrales est réalisé autour de cette période comme la procédure PRO-LIN le définit. Par conséquent, les caractéristiques sont le spectre de réponse en accélération échantillonné entre 0.5 s et 0.7 s, correspondant à 0.8 T_1^0 et 1.2 T_1^0, et à la période fondamentale T_1^0. C'est-à-dire l'ensemble $\mathbf{x}$ comprend $[S_a(T_i), S_a(T_1^0)]$ avec $T_i \approxeq [0.8T_1^0 : 1.2T_1^0]$ avec l'incrément égal à 0.1 T_1^0. La réponse cible $\mathbf{Y}$ pour la maquette est le taux de déplacement relatif entre étages.

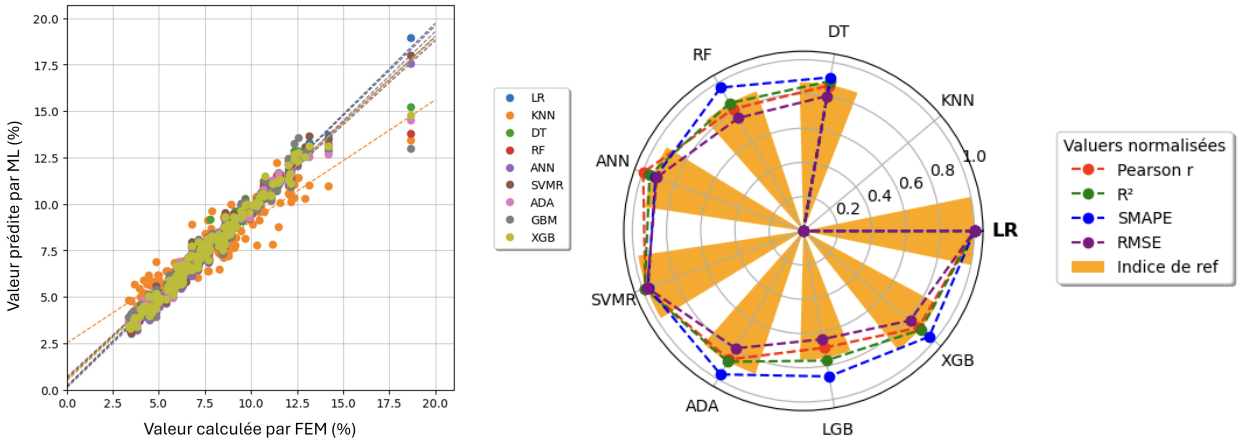
Tous les modèles de ML présentés dans les études précédentes sont appliqués, c'est-à-dire la régression linéaire (LR), les k-Plus proches voisins (KNN), la forêt aléatoire (RF), l'arbre de décision (DT), les machines à vecteurs de support (SVMR), les réseaux de neurones artificiels (ANN), l'algorithme de renforcement adaptatif (AdaBoost), la machine à renforcement léger de gradient (LightGBM) et la machine à renforcement extrême de gradient (XGBoost). Les quatre mesures sont le coefficient de corrélation de Pearson r, le coefficient de détermination R^2, l'erreur de pourcentage absolue moyenne symétrique $SMAPE$, et l'erreur quadratique moyenne $RMSE$. Le Tableau 5.3 présente la performance des modèles d'apprentissage automatique évalués à l'aide de ces métriques.

Modèle	r	R^2	$SMAPE$ (%)	$RMSE$
LR	0.990	0.980	4.032	0.004
KNN	0.832	0.645	16.012	0.015
DT	0.975	0.951	4.998	0.006
RF	0.982	0.964	4.270	0.005
ANN	0.983	0.965	4.198	0.004
SVMR	0.989	0.977	4.278	0.004
XGBoost	0.983	0.966	4.295	0.005
AdaBoost	0.985	0.969	4.199	0.004
LightGBM	0.972	0.943	4.871	0.006

Table 5.3 – [Maquette] Évaluation de la performance des modèles de ML

En observant le Tableau 5.3, le modèle de LR obtient d'excellents résultats, avec une valeur de R^2 de 0.980, ce qui indique une très bonne adéquation aux données. De plus, la valeur de $SMAPE$ de 4.032 % et la valeur de $RMSE$ de 0.004 identifient une faible erreur de prédiction. Les autres modèles montrent des performances similaires et suffisamment bonnes, sauf le modèle de type KNN qui est le moins adapté dans ce cas d'étude.

Une autre comparaison des modèles est réalisée, illustrée par la Figure 5.4. Sur 400 observations d'entrainement, les prédictions par ML sont comparées directement avec les valeurs calculées par EF, comme le présente la Figure 5.4a. Les lignes de tendance sont tracées pour chaque nuage de points. Ces lignes sont proches de la ligne de référence, inclinée à $45°$, ce qui confirme la performance des modèles. La deuxième comparaison utilise un graphe polaire. Ce graphe indice les valeurs normalisées des métriques d'évaluation, avec un indice de référence qui est la somme pondérée respectivement de 10 %, 30 %, 30 % et 30 % des valeurs de r, R^2, $SMAPE$ et $RMSE$. Ce graphe indique clairement que le modèle de type LR est le plus performant, suivi par les modèles de type ANN et SVMR.



(a) Comparaison des valeurs prédites et actuelles **(b)** Comparaison des métriques d'évaluation

Figure 5.4 – [Maquette] Comparaison des modèles de ML utilisés

Finalement, sur les 100 observations de test qui sont indépendantes de l'entrainement, les prédictions sont aussi comparées avec les valeurs calculées par EF pour les trois modèles de ML les plus performants. La Figure 5.5 illustre cette comparaison. De même, elle montre la performance des modèles ainsi que la capacité des modèles de réaliser des prédictions non biaisées sur des nouvelles observations.

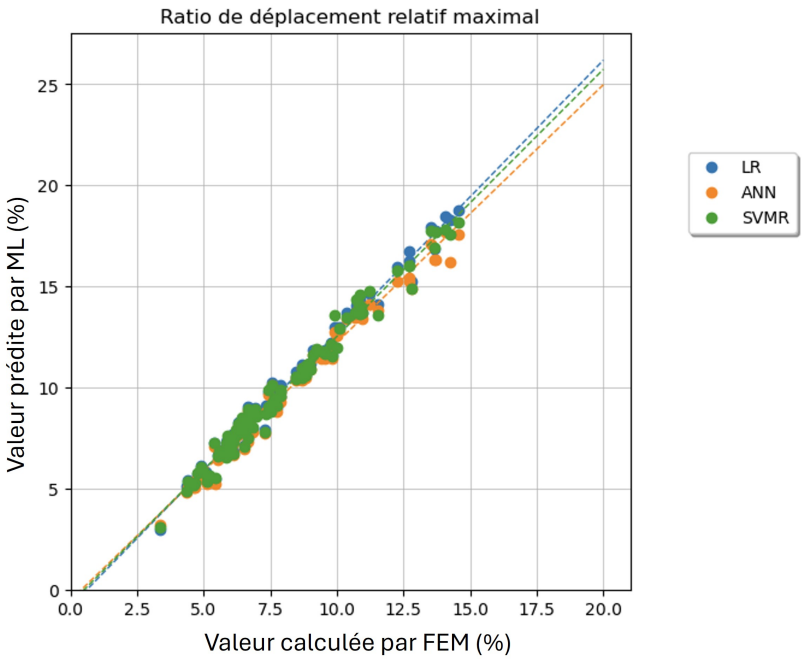


Figure 5.5 – [Maquette] Comparaison des valeurs prédites et actuelles

5.3. Validation de la procédure PRO-NONLIN par une base de données numériques

5.3.1. Présentation de la base de données

Le chapitre 1 a présenté les plateformes de partage des bases de données disponibles dans la littérature. Ces bases de données sont nécessaires pour le développement et l'utilisation d'apprentissage automatique dans le domaine de l'ingénierie. Pour la validation de la procédure PRO-NONLIN proposée pour les systèmes non-linéaires dans le chapitre 3, une base de données est extraite. C'est la base de données numérique construite par Guan *et al.* [127] pour l'investigation de la réponse des portiques en acier résistants aux moments (steel moment-resisting frames, SMRFs) sous séismes. Les auteurs introduisent le développement de la base complète de 621 SMRFs conçus selon les codes et normes modernes de construction. La base contient aussi les modèles structuraux non-linéaires correspondants et leurs réponses sismiques associées, qui sont les déplacements inter-étage maximaux, les accélérations maximales des planchers et les déplacements résiduels des étages. La Figure 5.6 illustre la structure de la base de données.

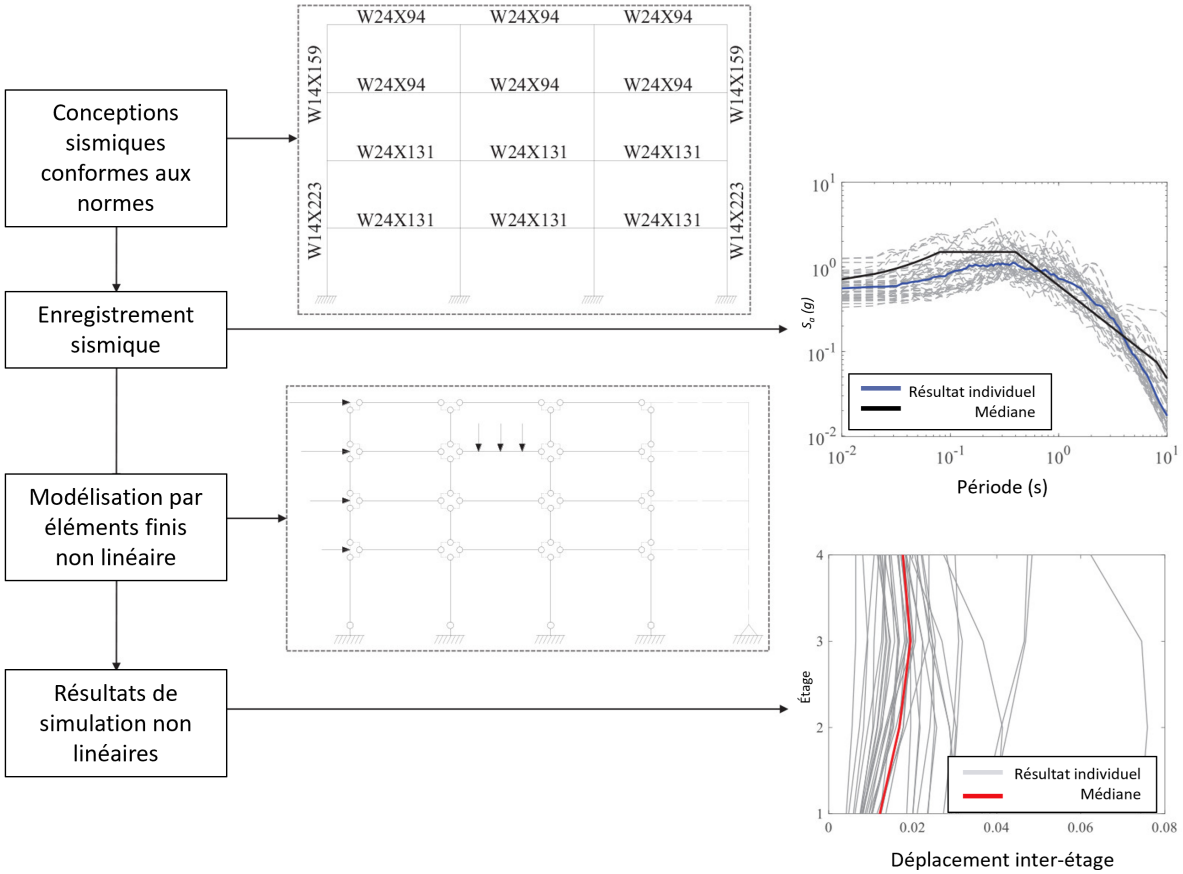


Figure 5.6 – Structure de la base de données de Guan *et al.* [127]

Pour explorer l'espace des archétypes de ces portiques en acier résistants aux moments, les paramètres connus pour influencer significativement leur conception sismique et/ou leurs performances sont identifiées. Ensuite, des valeurs limites inférieures et supérieures sont définies en fonction des limites admissibles spécifiées dans le code et/ou la norme de calcul. Le Tableau 5.4 résume la variation des paramètres pour construire différents portiques, en fonction de la configuration géométrique du bâtiment (nombre d'étages, nombre de travées, le rapport entre le premier étage et la hauteur typique des étages, la portée) et des chargements de conception (y compris les charges permanentes du plancher et du toit). Ces paramètres peuvent être sélectionnés comme caractéristiques pour entrainer les modèles d'apprentissage automatique. Le Tableau 5.5 indique le nombre de structures avec leur nombre d'étages contenus dans la base de données.

Catégorie	Paramètres	Valeur de conception
Géométrique	Nombre d'étages (N_s)	1, 5, 9, 14, et 19
	Nombre de travées (N_b)	1, 3, et 5
	Ratio de hauteur (h_1/h_t)	1.0, 1.5 et 2.0
	Portée (W_b)	6.10 m, 9.14 m et 12.19 m
Chargement de conception	Charge permanente du plancher (DL_{floor})	2.39 kN/m^2, 3.83 kN/m^2 et 5.27 kN/m^2
	Charge permanente du toit (DL_{roof})	0.96 kN/m^2, 3.23 kN/m^2 et 5.51 kN/m^2

Table 5.4 – Espace des paramètres variés des portiques considérées

Nombre d'étages	Nombre de portiques
1	81
5	162
9	162
14	128
19	88
Totalité	**621**

Table 5.5 – Résumé des structures (étages et nombres) de la base de données

Les modèles numériques des portiques en acier résistants aux moments ont été construits et analysés à l'aide du logiciel OpenSees [86]. Les composants constituant un modèle numérique non-linéaire du portique à cinq étages et cinq travées sont illustrés à la Figure 5.7. Les poutres et les colonnes ont été modélisées comme les éléments de poutre-colonne à rotule plastique concentrée pour simuler la réponse non-linéaire. Cet élément comprend deux parties : deux rotules non-linéaires placées aux deux extrémités et un élément de poutre-colonne élastique linéaire les reliant. Un ressort rotatif de longueur nulle avec un matériau Ibarra-Medina-Krawinkler (IMK) [130] modifié a été utilisé

pour la modélisation des rotules non-linéaires. Le matériau IMK a été largement utilisé pour modéliser le comportement hystérétique aux joints poutre-colonne et il est capable de capturer à la fois la dégradation cyclique et dans le cycle de la résistance et de la rigidité. Le comportement monotone du modèle IMK modifié est illustré dans la Figure 5.8a, tandis que son comportement de détérioration cyclique est représenté dans la Figure 5.8b.

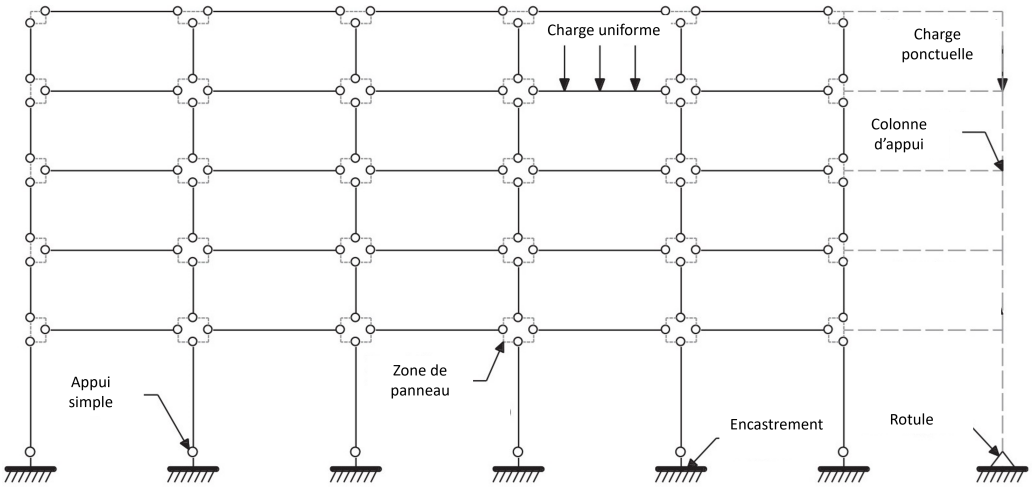
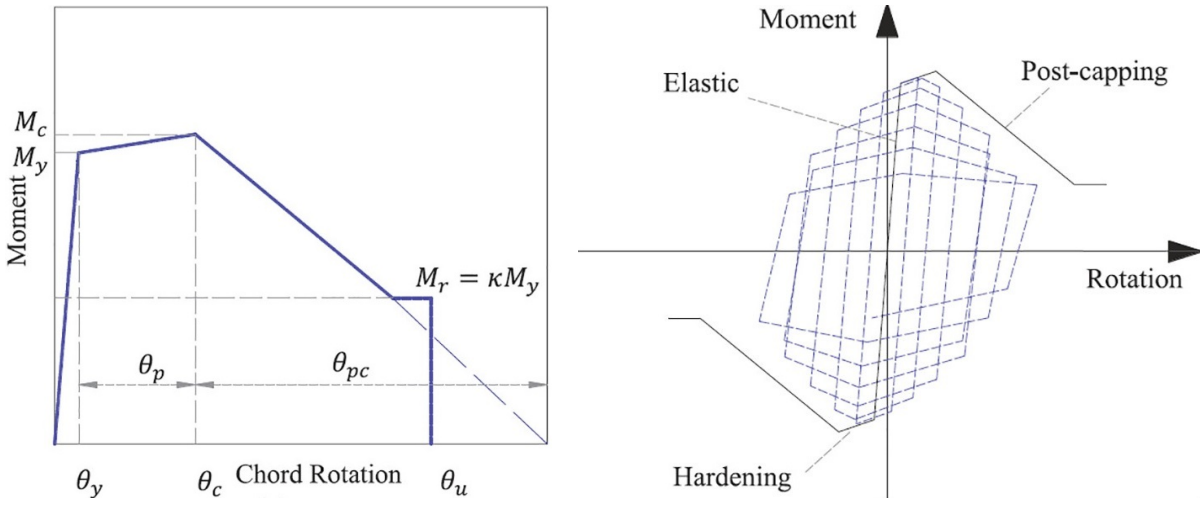


Figure 5.7 – Modèle numérique non-linéaire d'un portique à cinq étages et cinq travées



(a) Courbe de réaction monotone **(b)** Réponse cyclique

Figure 5.8 – Modèle de matériaux IMK modifié [127]

Le comportement de la zone du panneau (la zone de jointure où les poutres et les colonnes se croisent) a également été pris en compte en utilisant le modèle du parallélogramme proposé par Gupta et Krawinkler [131] pour représenter cet élément. Un ressort rotatif avec une courbe de base trilinéaire proposée par Krawinkler a été mis en œuvre pour examiner le comportement hystérétique

de cette zone. De plus, les effets P-delta ont été incorporés en utilisant une colonne d'appui, qui était reliée au portique en acier par l'intermédiaire d'un élément de treillis rigide. Une très faible rigidité a été appliquée au ressort rotatif de la colonne. La colonne d'appui n'affecte pas la rigidité latérale du portique en acier. Une charge de pesanteur uniforme a été appliquée aux portiques, et la charge restante a été appliquée à la colonne d'appui. Un taux d'amortissement de 2 % a été supposé pour le premier et le troisième mode du portique en acier. La masse de l'étage est attribuée uniformément à chaque nœud au même niveau d'étage. Des informations détaillées supplémentaires sur les portiques en acier et les modèles de comportement non-linéaire utilisés dans cette étude se trouvent dans la référence de Guan *et al.* [127].

Les deux lots de mouvements sismiques ont été sélectionnés pour la version actuelle de la base de données. Le premier lot comprend les 240 accélérations rapportées par Miranda [128] qui ont été enregistrés lors de 12 tremblements de terre survenus en Californie. Ils sont généralement représentatifs des mouvements du sol dans les régions de forte sismicité. Toutes les données proviennent de sites rocheux avec des vitesses moyennes des ondes de cisaillement supérieures à 180 m/s dans les 30 m supérieurs du profil du site. Ces mouvements du sol ont été enregistrés sur des stations en champ libre (free field) ou au premier étage de bâtiments de faible hauteur avec des effets d'interaction sol-structure négligeables. Les magnitudes des tremblements de terre, M, ayant généré ces enregistrements varient de 6.0 à 7.0 avec une moyenne de 6.7. L'accélération maximale du sol pour cet ensemble d'enregistrements varie de 0.03 g à 0.61 g. Des informations plus détaillées sur ces 240 enregistrements de mouvements sismiques peuvent être trouvées dans la référence de Miranda [128]. Les spectres d'accélération individuels et médians de cet ensemble sont présentés dans la Figure 5.9.

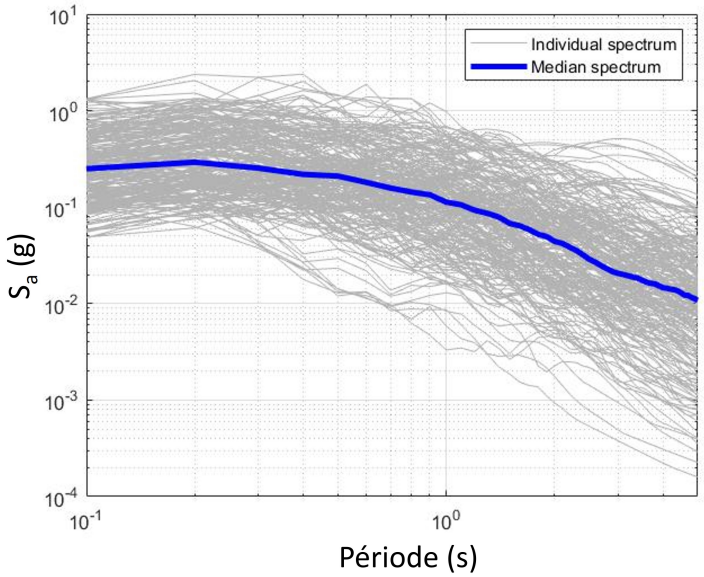


Figure 5.9 – Spectres d'accélération pour les 240 enregistrements de mouvement du sol

Le deuxième lot de mouvements du sol comprend trois ensembles d'enregistrements obtenus à

partir d'une procédure de sélection spécifique au site décrite par Jayaram et al. [119] à trois niveaux de risque, correspondant à des périodes de retour d'environ 43, 475 et 2475 ans, respectivement. C'est-à-dire un tel événement a une probabilité d'occurrence moyenne de 1 sur 2475 ans, 1 sur 475 ans et 1 sur 43 ans, respectivement. Chaque groupe se compose de trois séries d'enregistrements (avec 40 enregistrements chacune), qui sont sélectionnées pour des bâtiments avec des périodes d'environ 0.5, 1.0 et 2.0 secondes, respectivement, à partir de la base de données PEER NGA-WEST2 [65]. Ces deux lots de mouvements du sol sont appliqués dans la direction horizontale afin d'obtenir la réponse de la structure.

Deux sous-ensembles de données sont inclus pour les réponses non-linéaires. Le premier comprend les réponses structurelles des 621 SMRFs soumises à 240 mouvements sismiques, ce qui donne un total de 621 $\times$ 240 = 149 040 profils de taux de déformation maximale inter-étage, d'accélération maximale des étages et de taux de déformation inter-étage. L'autre contient les réponses d'un sous-groupe de 100 SMRFs (comprenant 13 bâtiments d'un étage, 26 de cinq étages, 26 de neuf étages, 21 de quatorze étages et 14 de dix-neuf étages) soumis à trois groupes de mouvements sismiques spécifiques au site (avec 40 mouvements sismiques dans chaque groupe). Ce sous-ensemble comprend 12 000 profils de déformation maximale inter-étage, d'accélérations maximales des étages et de déformation inter-étage.

5.3.2. Construction du jeu de données

Dans cette validation, deux sous-groupes de structures sont étudiés dans deux cas différents. Le premier contient 20 portiques à un étage et le deuxième contient 35 portiques à cinq et neuf étages. Ces structures sont choisies de façon aléatoire pour représenter la totalité de la base de données pour chaque type de portique.

Pour prédire les états d'endommagement des portiques, les caractéristiques $\mathbf{x}$ ont été considérées à travers trois groupes obtenus à partir des propriétés des mouvements du sol, les propriétés modales des structures et les paramètres de conception des structures.

Une analyse modale est réalisée pour chaque structure pour obtenir les périodes propres. Les portiques à un étage sont caractérisés par leur première période propre, tandis que les autres sont caractérisés par leurs quatre premières périodes. La première période propre de la structure est un paramètre essentiel utilisé pour déterminer la force sismique latérale équivalente appliquée à une structure, tandis que les autres périodes ont été utilisées pour prendre en compte l'effet des modes supérieurs sur la réponse sismique des structures de grande hauteur (dans le deuxième cas d'étude). Les périodes de la structures documentées dans la base de données sont utilisées comme la représentation modale des structures. Ces valeurs sont choisies comme des caractéristiques dans les modèles de ML. Le Tableau 5.6 présente les propriétés statistiques des périodes propres pour l'ensemble des structures de la base pour les deux cas testés : portiques à un étage et portiques à plusieurs étages.

Les paramètres de la conception des structures sont aussi utilisés. Le ratio entre la hauteur du rez-de-chaussée et la hauteur typique est 1.0, 1.5 et 2.0. Le nombre de travées variait d'une à cinq, avec des longueurs de travée de 6.1 m, 9.14 m ou 12.19 m. Les charges totales des planchers, comprenant le poids propre de la structure et la charge permanente, sont de 2.39 kN/m^2, 3.82 kN/m^2 et 5.27

	Portique à 1 étage	**Portique à plusieurs étages**			
	T_1^0	T_1^0	T_2^0	T_3^0	T_4^0
Min	0.388	1.174	0.415	0.221	0.155
Max	0.826	1.936	0.702	0.413	0.279
Moyen	0.577	1.555	0.556	0.313	0.209
Écart-type	0.113	0.220	0.086	0.057	0.044

Table 5.6 – Propriétés statistiques des périodes des structures

kN/m^2 ; la valeur la plus basse représente une dalle mince en béton léger, tandis que la valeur la plus élevée se réfère à une dalle plus épaisse en béton de poids normal. De plus, pour le cas des structures à plusieurs étages, le nombre d'étages, le nombre de travées, la longueur d'une travée, le ratio de la hauteur, les charges appliquées sont considérées comme la représentation de la conception de structure.

Les caractéristiques les plus importantes et les plus concernées pour la thèse sont les valeurs de spectre de réponse en accélération échantillonnées aux différentes périodes, $S_a(T_i)$. Les valeurs spectrales sont échantillonnées dans les modèles de machine learning. Pour le cas des portiques à un étage, les accélérations spectrales de 0.1 s à 2.0 s par un incrément de 0.1 s sont considérées comme caractéristiques. Pour le cas des portiques à plusieurs étages, les accélérations spectrales de 0.1 s à 5.0 s par un incrément de 0.1 s sont considérées comme la représentation des signaux sismiques. Le choix initial des accélérations spectrales est réalisé pour pouvoir couvrir une large gamme des périodes, jusqu'à 2 fois la période fondamentale maximale.

Selon la description de la base de données, les portiques sont simulés sous séismes pour obtenir leur taux de déformation maximale inter-étage, leur accélération maximale des étages, et leur taux de déformation résiduelle inter-étage. Pour ce cas, le taux de déformation maximale inter-étage des portiques est choisi comme réponse cible $\mathbf{Y}$.

En résumé, nous choisissons pour le cas des portiques à un étage avec 4800 observations collectées par la simulation des réponses de 20 portiques sous 240 signaux sismiques. Pour le cas des portiques à plusieurs étages, 8400 observations sont collectées par la simulation des réponses de 35 portiques sous 240 signaux sismiques. Les caractéristiques sont illustrées dans la Figure 5.10. Pour chaque cas, 80 % des observations sont utilisées pour entrainer les modèles d'apprentissage automatique et les 20 % pour tester ces modèles.

Le modèle RF est choisi à étudier pour ce cas. Les quatre mesures sont le coefficient de corrélation de Pearson r, le coefficient de détermination R^2, l'erreur de pourcentage absolue moyenne symétrique $SMAPE$, et l'erreur quadratique moyenne $RMSE$.

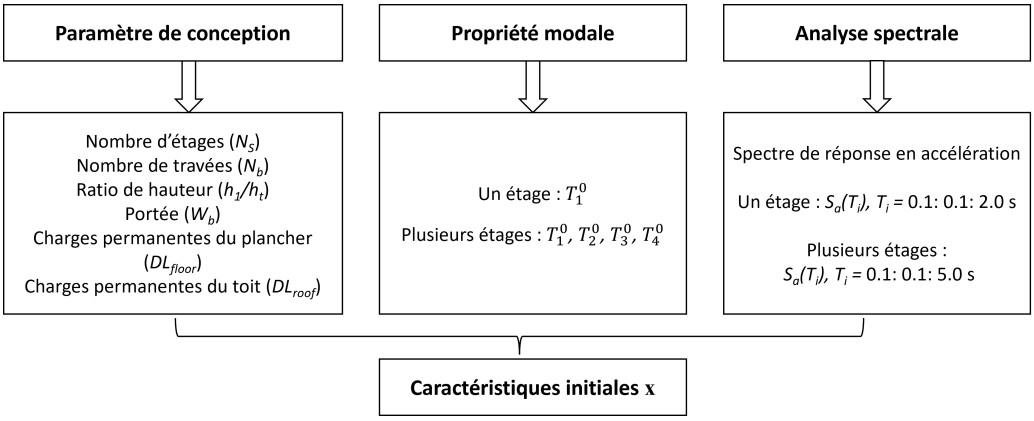


Figure 5.10 – Description de l'ensemble des caractéristiques $\mathbf{x}$

5.3.3. Portiques à un étage

La procédure PRO-NONLIN est appliquée en utilisant le modèle RF. La Figure 5.11 illustre les caractéristiques sélectionnées par la méthode hybride. Selon la procédure PRO-NONLIN présentée dans le chapitre 3, les coefficients de corrélation entre la réponse cible $\mathbf{Y}$ et les caractéristiques initiales $\mathbf{x}$ sont calculés. Les trois groupes de l'ensemble des caractéristiques $\mathbf{x}$, comprenant les paramètres de conception, la propriété modale et l'analyse spectrale sont évalués. Il est à noter que dans la méthode de sélection hybride, l'étape de filtrage n'est appliquée qu'aux spectres de réponse en accélération $S_a(T_i)$.

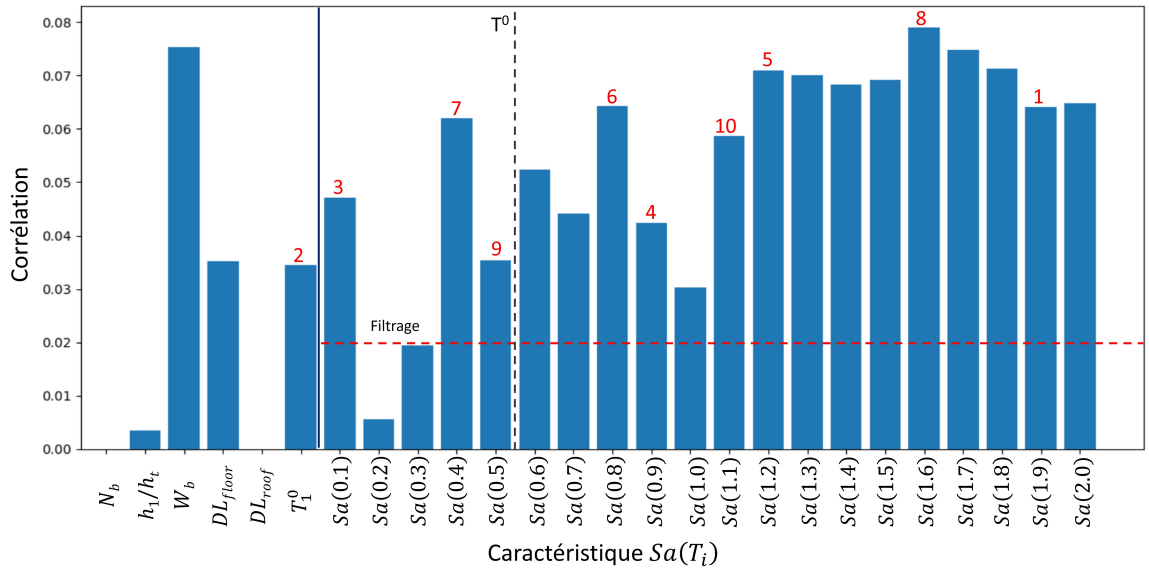


Figure 5.11 – [1 étage] Ordre de sélection des caractéristiques $S_a(T_i)$

Les caractéristiques par l'analyse spectrale, c'est-à-dire les valeurs de spectre de réponse en accélération, sont les plus importantes. Au contraire, certaines caractéristiques liées aux paramètres de conception sont les moins importantes. Selon cette illustration, un seuil de corrélation pour le filtrage est mis en place pour seulement choisir les caractéristiques de type spectral les plus corrélées par rapport à la réponse cible. La ligne horizontale illustre le seuil de filtrage sur les valeurs de corrélation entre $S_a(T_i)$ et $\mathbf{Y}$, pour sélectionner préliminairement les caractéristiques. Les caractéristiques de type spectral dont les coefficients de corrélation sont inférieurs à 0.02 sont filtrées et supprimées. Les caractéristiques restantes sont sélectionnées selon la méthode d'enveloppe, et ils sont numérotés avec leurs ordres. A partir de l'ordre de sélection des caractéristiques, on remarque que la première identifiée est de type spectral. La deuxième sélectionnée est la première période propre, T_1^0. Les autres paramètres de conception sont des caractéristiques les moins importantes et elle ne sont donc pas choisis. Toutefois, on peut noter qu'elles influent la première période propre de la structure.

La Figure 5.12 présente l'évolution des métriques d'évaluation R^2 et $RMSE$ selon le nombre de caractéristiques. On constate que le nombre de caractéristiques influe sur la performance du modèle. Dans ce cas, à partir de 8 caractéristiques, le modèle RF commence à avoir une performance stable.

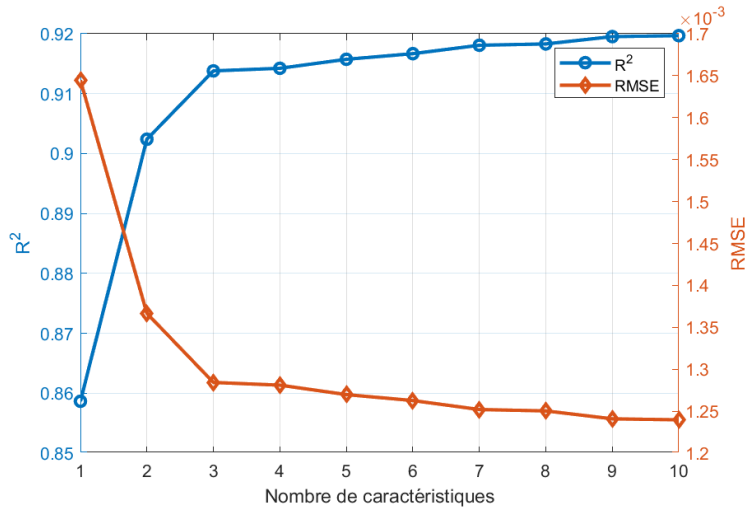


Figure 5.12 – [1 étage] Évolution des métriques d'évaluation selon le nombre de caractéristiques

Un modèle RF avec la base de données dont les caractéristiques sont sélectionnées par la procédure PRO-NONLIN est entrainé. Il est évalué par les métriques sur l'ensemble de l'entrainement et de test, pour assurer que la prédiction soit précise pour les nouvelles données. Le Tableau 5.7 résume le performance du modèle RF avec les caractéristiques initiales et sélectionnées pour ces deux ensembles d'entrainement et de test. La performance du modèle pour ces deux cas est similaire. La sélection des caractéristiques réalisées prend bien en compte les plus signifiantes.

L'application du modèle avec les caractéristiques sélectionnées montre que pour les données d'entraînement, le modèle présente un coefficient de détermination R^2 de 0.938, indiquant que celui-ci identifie bien la variabilité des données. De plus, le modèle affiche une faible valeur de $RMSE$ de

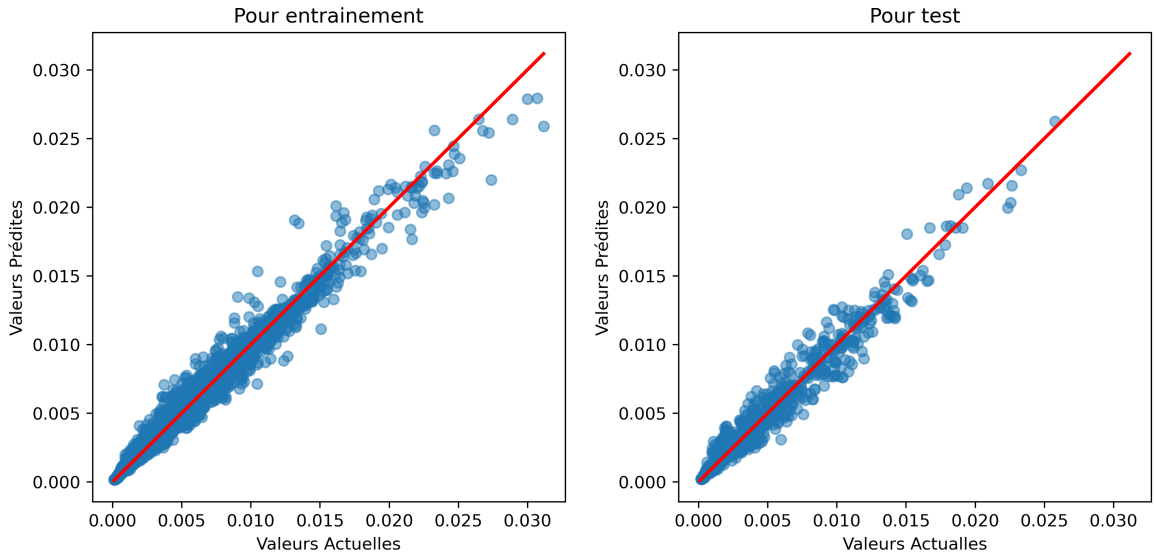
0.103 $\times 10^{-2}$, ce qui montre que les prédictions du modèle sont en très bon accord avec les valeurs obtenues par simulation par éléments finis. Pour les données de test, le modèle présente également de bonnes performances, bien que légèrement inférieures à celles obtenues pour des données d'entraînement. Le coefficient de détermination R^2 est alors de 0.920, ce qui signifie que le modèle explique une grande partie de la variabilité des données de test. Le $RMSE$ est de 0.123 $\times 10^{-2}$, indiquant que les prédictions du modèle sont aussi précises. Cependant, le $SMAPE$ de 12.644 % montre une augmentation de la dispersion des erreurs par rapport aux données d'entraînement. En résumé, ce modèle de machine learning fonctionne bien pour prédire les réponses sismiques des portiques, le taux de déformation maximale inter-étage, avec des performances solides sur les données d'entraînement et des performances légèrement moins bonnes, mais acceptables, sur les données de test. Cela valide la procédure PRO-NONLIN. Il faut noter que les données d'entrainement et de test sont indépendantes. Les données de test contiennent des observations nouvelles qui n'ont pas encore été utilisées pendant la phase d'entrainement.

Caractéristiques	**Données**	r	R^2	$SMAPE$ (%)	$RMSE$
Sans sélection	Entrainement	0.989	0.948	8.686	0.063 $\times 10^{-2}$
	Test	0.978	0.937	9.644	0.085 $\times 10^{-2}$
Sélection hybride	Entrainement	0.969	0.938	9.686	0.103 $\times 10^{-2}$
	Test	0.959	0.920	12.644	0.123 $\times 10^{-2}$

Table 5.7 – [1 étage] Performance du modèle RF entre l'entrainement et le test

Les prédictions pour chaque ensemble de données sont comparées avec les valeurs actuelles (obtenues par simulation par éléments finis et documentées dans la base de données). Cette comparaison est illustrée dans la Figure 5.13. La ligne de tendance (ligne rouge) est la ligne où les prédictions exactes se trouvent. Pour les données de l'entrainement et de test, les prédictions sont proches à cette ligne, ce qui montre la fiabilité du modèle. Puis, l'écart entre les prédictions et les valeurs actuelles du taux de déformation maximale inter-étage est calculée. Celui-ci est représenté sur la Figure 5.14. On constate une répartition de ces valeurs autour de 0. Toutes ces observations démontrent que le modèle RF entrainé avec les caractéristiques sélectionnées par la procédure PRO-NONLIN est capable de fournir une estimation non biaisée et qu'il présente un niveau de précision élevé sur les ensembles de données d'entraînement et de test pour des portiques à un étage.

La méthode de sélection hybride et la procédure PRO-NONLIN proposés dans cette thèse montrent leur capacité de permettre de prédire correctement le comportement des structures. Elle propose une réduction du nombre de caractéristiques pour entrainer un modèle de machine learning sans perte de précision. De plus, cette réduction contribue à l'optimisation de la phase d'entrainement des modèles de machine learning, en éliminant les caractéristiques peu influentes.



(a) Pour les données de l'entrainement **(b)** Pour les données de test

Figure 5.13 – [1 étage] Comparaison entre les valeurs prédites par modèle RF et les valeurs actuelles obtenues par éléments finis

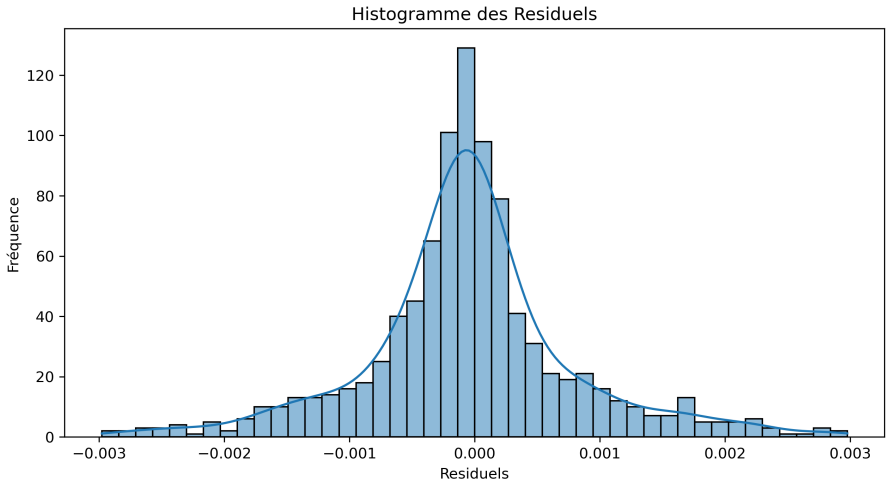


Figure 5.14 – [1 étage] Histogramme de résidus

5.3.4. Portiques à plusieurs étages

De la même manière, pour cette validation de la procédure PRO-NONLIN avec des portiques à plusieurs étages, des étapes similaires sont réalisées en utilisant le modèle RF. La sélection par la méthode hybride est appliquée et illustrée par la Figure 5.15. Les coefficients de corrélation entre la réponse cible $\mathbf{Y}$ et les caractéristiques initiales $\mathbf{x}$ sont calculés pour la sélection après filtrage. La ligne

horizontale est le seuil de filtrage pour sélectionner préliminairement les caractéristiques spectrales. La sélection des caractéristiques par la méthode d'enveloppe est ensuite réalisée. La Figure 5.15 note les caractéristiques sélectionnées avec leurs ordres. Selon cette sélection, les caractéristiques de type spectral sont les plus importantes. Les caractéristiques de type modale sont aussi importantes pour le modèle RF dans ce cas. Une seule caractéristique du paramètre de conception, qui est le nombre d'étages N_b, est sélectionnée. La sélection pour le cas des portiques à plusieurs étages montre une cohérence avec celui à un étage, qui montre l'importance des caractéristiques spectrales. Par conséquent, la sélection des caractéristiques selon la procédure PRO-NONLIN est nécessaire pour pouvoir optimiser le choix des caractéristiques.

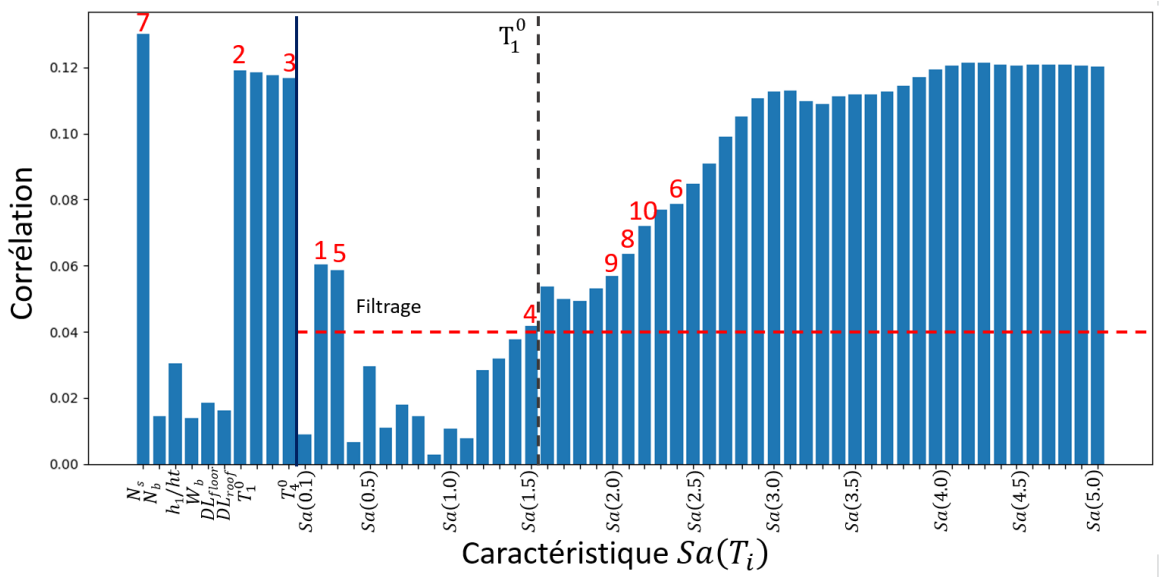


Figure 5.15 – [Plusieurs étages] Ordre de sélection des caractéristiques

La Figure 5.16 présente l'évolution de deux métriques, $RMSE$ et R^2, selon le nombre de caractéristiques sélectionnées. A partir de 5 caractéristiques, la performance du modèle devient stable. Pour ce cas, on choisit 5 caractéristiques, selon l'ordre dans la Figure 5.15, pour entrainer le modèle RF final.

Le modèle RF est finalement entrainé et évalué avec l'ensemble des caractéristiques sélectionnées par la méthode hybride. Une comparaison de la performance du modèle sur les deux ensembles de données, d'entrainement et de test, pour assurer que la prédiction est précise pour les nouvelles données. De plus, une comparaison entre le modèle avec les caractéristiques sélectionnées par rapport à celui avec toutes les caractéristiques est aussi réalisée. Les résultats présentés dans le Tableau 5.8 mettent en évidence les performances du modèle avec la sélection des caractéristiques. Les critères d'évaluation du modèle RF sur les deux ensembles de données avec ou sans une sélection des caractéristiques sont similaires. Ils montrent que la sélection est correcte car le modèle obtenu a une performance proche de celle du modèle sans sélection, avec toutefois un nombre de caractéristiques beaucoup plus faible. En effet, le nombre des caractéristiques est bien optimisé, réduit de 60 caractéristiques à 6 seulement. La sélection réduit d'un facteur 10 du nombre de caractéristiques, simplifiant

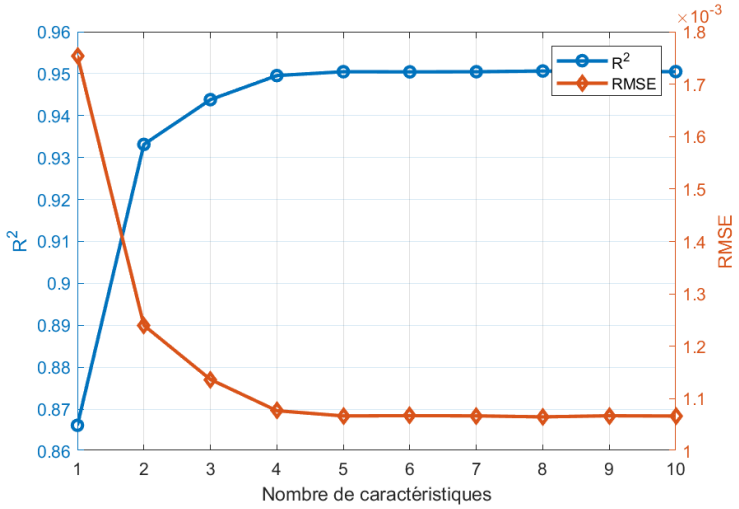


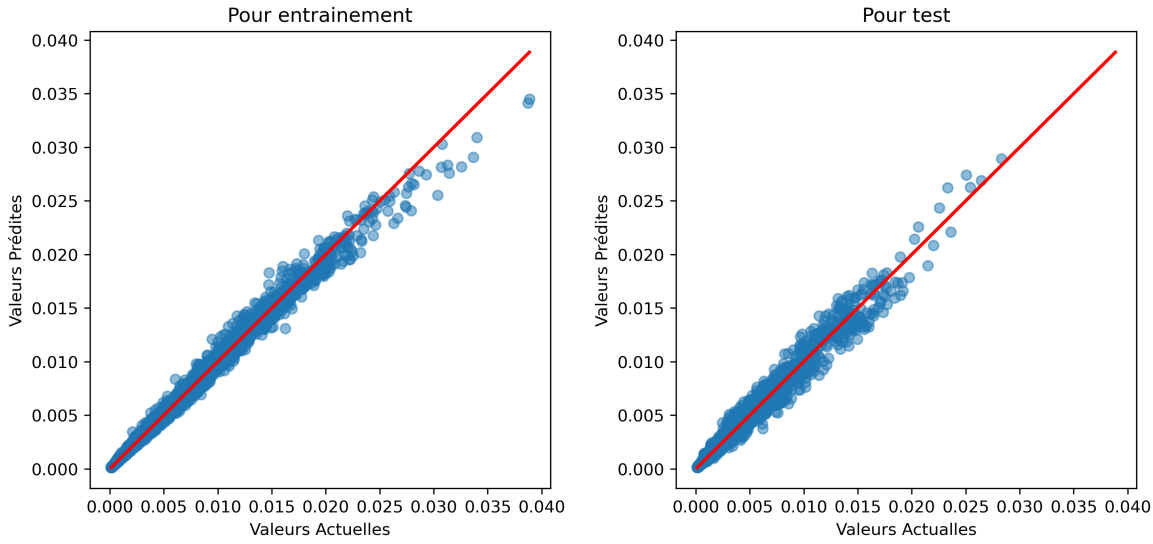
Figure 5.16 – [Plusieurs étages] Évolution des métriques selon le nombre de caractéristiques

d'autant plus l'entrainement du modèle. Cela souligne l'importance de sélection des caractéristiques dans le cas où les données sont plus nombreuses et complexes.

Les prédictions pour chaque partition sont confrontées aux valeurs simulées contenues dans la base de données. Cette comparaison est illustrée dans la Figure 5.17. La ligne rouge représente l'emplacement des prédictions exactes. Pour les données d'entraînement et de test, les prédictions demeurent constamment proches de cette ligne, démontrant ainsi la fiabilité du modèle. De plus, la Figure 5.18 illustre la distribution des écarts entre les valeurs simulées et prédites par ML, montrant que ceux-ci se positionnent autour de 0. L'ensemble de ces résultats confirme que le modèle RF est capable de fournir des estimations non baisées tout en maintenant un niveau de précision significatif, autant pour les ensembles de données d'entraînement que de test, ainsi que pour des configurations de portiques différentes.

Caractéristiques	**Données**	r	R^2	$SMAPE$ (%)	$RMSE$
Sans sélection	Entrainement	0.997	0.993	4.620	0.050 $\times 10^{-2}$
	Test	0.985	0.961	8.626	0.074 $\times 10^{-2}$
Sélection	Entrainement	0.989	0.978	5.686	0.063 $\times 10^{-2}$
	Test	0.975	0.951	8.462	0.075 $\times 10^{-2}$

Table 5.8 – [Plusieurs étages] Performance du modèle RF entre l'entrainement et le test



(a) Pour les données de l'entrainement **(b)** Pour les données de test

Figure 5.17 – [Plusieurs étages] Comparaison entre les valeurs prédites et les valeurs simulées

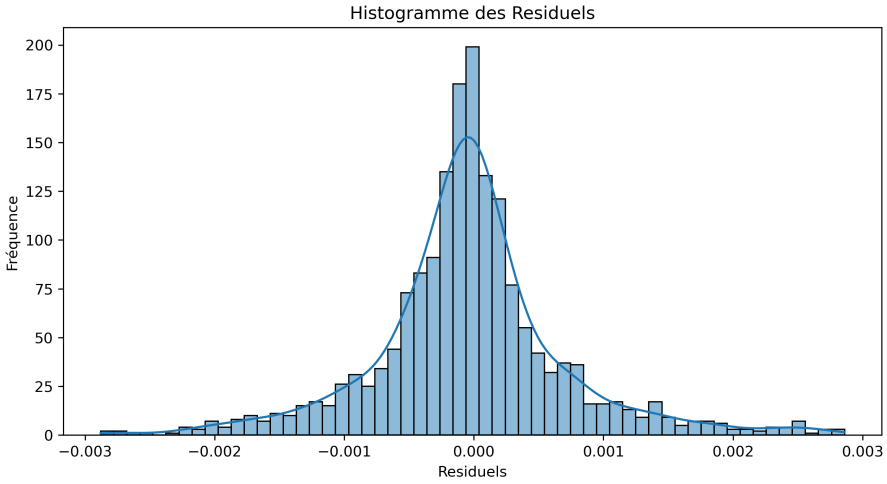


Figure 5.18 – [Plusieurs étages] Histogramme des écarts

5.4. Conclusion

L'objectif du chapitre était de tester les procédures proposées PRO-LIN et PRO-NONLIN avec les données disponibles dans la littérature.

La procédure PRO-LIN a été vérifiée avec des réponses simulées d'une maquette proposée par O. Sahin *et al.* [129]. Plusieurs modèles d'apprentissage automatique ont été obtenus avec de bonne

précision, parmi lesquels, le modèle de régression linéaire est le plus performant en considérant l'indice de référence. La procédure PRO-LIN confirme encore une fois l'importance des spectres de réponse en accélération autour de la période fondamentale de la structure.

La base de données élaborée par Guan *et al.* [127] a été extraite pour deux cas différents : (i) portiques à un étage et (ii) portiques à plusieurs étages. La particularité de cette base de données est la présence des paramètres de conception, des périodes propres, en plus des spectres de réponse en accélération comme caractéristiques des modèles d'apprentissage automatique. L'application de la procédure PRO-NONLIN à cette base de données, permet d'obtenir plusieurs modèles d'apprentissage automatique. Le modèle RF est choisi pour la validation, grâce à sa précision et sa facilité de mise en œuvre. Il est à noter que la procédure PRO-NONLIN dotée de la sélection hybride a été très efficace dans la sélection des caractéristiques les plus importants des modèles.

Des modèles d'apprentissage automatique de bonne performance ont été obtenus sur ces nouvelles données. Ces résultats confirment la validité de la méthode proposée basée sur les spectres de réponse en accélération comme caractéristiques des modèles via deux procédures pratiques PRO-LIN et PRO-NONLIN.

Conclusions et Perspectives

Conclusions

La motivation principale de cette étude est l'application de l'apprentissage automatique à la modélisation de la relation entre les mouvements sismiques du sol (entrée) et les réponses sismiques des structures (sortie). La réussite de cette modélisation permet de générer des données pour évaluer les risques sismiques en général et pour construire des courbes de fragilité sismique en particulier, au lieu de réaliser des simulations numériques par la méthode des éléments finis, qui sont coûteuses en termes de charge de calcul.

Dans le premier chapitre de la thèse, une étude bibliographique est présentée pour définir les courbes de fragilité sismique, leur utilité en tant qu'outil d'aide à la décision ainsi que les méthodes pour les construire. Ces études relèvent que la méthode de construction la plus prometteuse est basée sur les simulations numériques. Toutefois, on identifie l'obstacle majeur de la charge nécessaire pour les calculs sismiques non-linéaires par la méthode des éléments finis. Deux voies de recherche sont menées pour dépasser cette difficulté : soit développer des méthodes plutôt statistiques efficaces pour un nombre limité de simulations numériques issues de la méthode des éléments finis, soit appliquer des modèles d'apprentissage automatique pour remplacer les simulations numériques. Le travail de recherche de cette thèse se situe dans ce deuxième axe de recherche.

Après une présentation rapide des modèles d'apprentissage automatique les plus utilisés, les efforts se concentrent sur la bibliographie des applications existantes des modèles d'apprentissage automatique pour déterminer les réponses sismiques des structures. Cette étude quasi-exhaustive révèle le premier obstacle à surmonter : les caractéristiques (ou "features" en anglais) à considérer dans les modèles de machine learning. Il est à noter que les caractéristiques utilisées dans le bilan bibliographique sont très différentes les unes des autres, ce qui rend difficile leur réutilisation. Notre proposition est donc d'utiliser le spectre de réponse en accélération échantillonné à différentes périodes comme caractéristiques des modèles d'apprentissage automatique. Il est connu que le spectre de réponse en accélération contient des informations sur les tremblements de terre mais aussi sur les caractéristiques structurelles. Cette proposition simplifie le choix des caractéristiques car elles sont toutes de même nature. La proposition est vérifiée pour des systèmes linéaires dans le chapitre 2 et pour des systèmes non-linéaires dans le chapitre 3. Elle est également confrontée à des séismes de natures différentes, avec des séismes réels dans le chapitre 4, et finalement, elle est testée avec des données disponibles dans la littérature au chapitre 5.

Afin de rendre la proposition plus facile à appliquer, des procédures complètes sont détaillées pas-à-pas, PRO-LIN pour les structures linéaires et PRO-NONLIN pour les structures non-linéaires. Elles incluent également une stratégie de sélection des caractéristiques, c'est-à-dire les valeurs des

spectres de réponses les plus importantes pour les périodes échantillonnées, afin de rendre les modèles d'apprentissage automatique plus performants.

L'application des modèles d'apprentissage automatique aux systèmes linéaires est présentée au chapitre 2. Le spectre de réponse en accélération est échantillonné autour des modes propres. L'étude de SHAP, mesurant l'importance des valeurs des spectres des périodes, est employée pour sélectionner les caractéristiques les plus pertinentes. La procédure pas-à-pas PRO-LIN est proposée et validée pour un système à un degré de liberté puis à deux degrés de liberté. Les résultats obtenus sont satisfaisants pour plusieurs modèles d'apprentissage automatique tels que ANN, RF, ... Il est aussi possible de classer ces modèles en utilisant des métriques de précision et d'erreur. Avec un modèle d'apprentissage obtenu, on peut générer facilement de nouvelles observations et construire des courbes de fragilité sismique par des méthodes existantes telles que le Maximum de Vraisemblance ou la Simulation de Monte-Carlo. Les résultats obtenus montrent un temps de calcul fortement réduit avec une précision meilleure ou similaire par rapport aux autres méthodes d'approximation en appliquant des modèles d'apprentissage automatique.

La méthode hybride comprenant les avantages de la sélection rapide de la méthode de filtrage et la précision de la méthode d'enveloppe pour les valeurs dess spectres des périodes les plus importantes, est proposée pour les structures non-linéaires dans le chapitre 3. Une procédure pratique pas-à-pas PRO-NONLIN est proposée et testée sur des oscillateurs non-linéaires d'amortissement de Coulomb, de rigidité de Bouc-Wen avant d'être appliquée sur un bâtiment non-linéaire à 8 étages. Les métriques montrent de bons résultats des modèles d'apprentissage automatique avec la procédure proposée. Une fois validé, un modèle d'apprentissage automatique permet de générer facilement des réponses sismiques pour en déduire des courbes de fragilité sismique. La comparaison montre une réduction significative du temps de construction pour une précision meilleure ou similaire des courbes de fragilité sismique.

Le chapitre 4 est destiné à la vérification des procédures PRO-LIN et PRO-NONLIN proposées respectivement aux systèmes linéaires et non-linéaires pour les séismes réels. En prenant la méthode de sélection des enregistrements réels par le spectre conditionnel, les séismes sont obtenus à partir de la base des enregistrements de NGA-West2. En appliquant les deux procédures PRO-LIN et PRO-NONLIN aux séismes réels, les résultats obtenus sont satisfaisants mais leur précision est plus faible par rapport aux cas des séismes synthétiques du fait du nombre limité d'enregistrements réels. Il est bien connu que la précision des modèles d'apprentissage automatique dépend des volumes de données disponibles.

Finalement, le chapitre 5 est réservé pour la vérification de la procédure proposée avec les données disponibles dans la littérature. La première étude est réalisée avec une structure linéaire de béton armé tandis que la deuxième étude est réalisée avec une base de données de portiques à un ou à plusieurs étages, non-linéaires en acier. L'application des procédures proposées à ces données montrent de bons résultats des modèles d'apprentissage automatique. Elle confirme ainsi la validité de notre méthode à travers des procédures pratiques proposées.

Perspectives

La méthode proposée dans la thèse consiste à utiliser des modèles d'apprentissage automatique avec les valeurs des spectres de réponse en accélération pour différentes périodes comme caractéristiques pour modéliser les réponses sismiques des structures. La méthode a été présentée en deux procédures pratiques, PRO-LIN et PRO-NONLIN, pour les structures linéaires et non-linéaires respectivement. Elle a été validée pour les systèmes linéaires et non-linéaires, pour les séismes synthétiques, puis réels ainsi que pour les données disponibles dans la littérature. Certaines perspectives peuvent être proposées pour améliorer ce travail de thèse.

Premièrement, les procédures proposées ont bien montré leur capacité pour améliorer la construction de la courbe de fragilité. Par contre, pour des enregistrements réels, elles sont encore limitées par le nombre de séismes disponibles. Une tentative pour enrichir le nombre d'enregistrements devrait être encouragé, par exemple, par la construction d'une base de séismes mixtes : synthétiques et réels.

Deuxièmement, la méthodologie a été validée avec des réponses simulées. Il serait donc très intéressant de la valider avec des bases de données expérimentales. L'étude bibliographique a d'ailleurs montré un manque de données expérimentales disponibles dans le domaine du génie parasismique. Certaines bases de données expérimentales existent mais se limitent seulement aux études de la performance des éléments structuraux. Une base de données contenant des réponses sismiques des structures réelles serait utile pour la communauté scientifique. Une telle base peut être obtenue à partir d'une maquette au laboratoire excitée par une table vibrante.

Troisièmement, la sélection des caractéristiques par la méthode hybride utilise d'abord le filtrage pour des caractéristiques relatives au mouvement du sol. Ce filtrage dépend de la mesure de pertinence et du seuil imposé. Dans ce travail, le coefficient de corrélation de Pearson et le seuil de filtrage sont choisis pour avoir un bon compromis de la sélection des caractéristiques. Une étude paramétrique ainsi que d'autres mesures de pertinence sont encouragées à être testées dans les études ultérieures. Une étude de l'orthogonalité est aussi à étudier afin de réduire la dimension des caractéristiques.

Enfin, la proposition de la thèse a été adaptée pour la construction des courbes de fragilité sismique. Les modèles d'apprentissage automatique de régression ont été proposés pour prédire les réponses caractéristiques des défaillances. Il est possible de tester les modèles de classification pour obtenir les états binaires : défaillance ou non. De plus, avec le développement rapide des modèles d'apprentissage automatique, une prédiction temporelle des réponses des structures deviendrait possible. Il serait plus efficace dans ce cas de considérer directement les accélérations du sol comme caractéristiques des modèles d'apprentissage automatique.

Annexes

A Séisme synthétique par le modèle de Boore

Le séisme est caractérisé par une fonction d'enveloppe déterministe et un processus aléatoire Gaussien stationnaire dont la densité spectrale de puissance dépend de la source, du chemin de la source au site, de l'effet du site et de la réponse du sol recherchée (accélération, vitesse, déplacement). Le PSD dans le modèle de Boore [107] est décrit par la formule suivante :

$$Y(M_0, R, f) = E(M_0, f)P(R, f)G(f)I(f) \tag{A.1}$$

où M_0 est le moment sismique qui est relié à la magnitude M ; R est la distance de l'épicentre sismique au site ; f est la fréquence (Hz) ; $E(M_0, f), P(R, f) et\, G(f)$ caractérisent respectivement l'influence de la source, l'influence du chemin de la source au site et l'effet du site ; $I(f)$ traduit le mouvement que l'on veut obtenir au site : déplacement, vitesse ou accélération. Les termes nécessaires pour la formule (A.1) sont détaillés dans ce qui suit.

Influence de la source $E(M_0, f)$
L'expression de $E(M_0, f)$ a la forme suivante :

$$E(M_0, f) = CM_0S(M_0, f) \tag{A.2}$$

où C est une constante qui considère l'influence : du rayon de référence de la zone rupture pris égal à 1 km ; de la radiation ; du taux de l'onde totale aux composantes horizontales ; de l'effet de la surface libre ; de la densité et de la vitesse de propagation à la source. Le spectre de la source $S(M_0, f)$ peut avoir plusieurs formes données dans le tableau de référence de Boore [107]. Dans l'équation (A.2), M_0 est le moment sismique relié à la magnitude séismique M selon la relation suivante :

$$M = \frac{2}{3}\log(M_0) - 10.7 \tag{A.3}$$

La Figure A.1 illustre l'influence de la source en fonction de la fréquence pour $M = 7$.

Influence du chemin : spectre $P(R, f)$ **et durée du séisme**
Le chemin entre la source et le site a une influence sur le spectre du site et ainsi que sur la durée du séisme. On présente d'abord l'influence au spectre, puis l'influence à la durée. En principe, plus le site est loin, plus la durée du séisme est longue. **Influence au spectre** $P(R, f)$: La fonction $P(R, f)$ comporte deux composantes : une atténuation par divergence géométrique $Z(R)$ et une atténuation par le système entre le site et la source caractérisée par la fonction $Q(f)$. L'expression complète de $P(R, f)$ est la suivante :

$$P(R, f) = Z(R)\exp\left[-\frac{\pi f R}{Q(f)}c_Q\right] \tag{A.4}$$

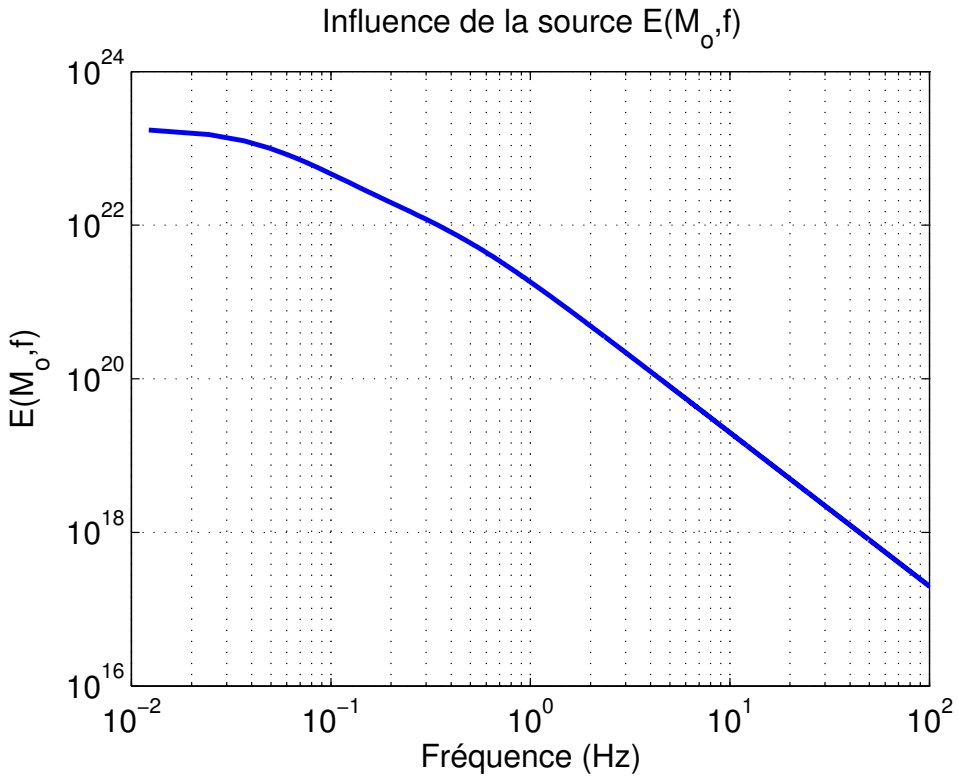


Figure A.1 – Influence de la source, $M = 7$

où c_Q est la vitesse de propagation d'onde. Boore propose $c_Q = 3.5$ km/s ; $Q(f)$ est la composante de l'atténuation par le système entre le site et la source : $Q(f) = 180f^{0.45}$; $Z(R)$ est la composante de l'atténuation par divergence géométrique, définie par des fonctions linéaires par morceaux. La Figure A.2 présente l'influence du chemin en fonction de la fréquence pour R= 9 km.

Durée d'un séisme : La durée d'un séisme est la somme de la durée de la source T_{source} et de la durée de propagation (ground motion) T_{gm} . On peut prendre : $T_{\text{source}} = \frac{0.5}{f_a}$ et $T_{\text{gm}} = 0.05R$, où le paramètre f_a est dépendant de la magnitude du séisme [107].

Influence de l'effet du site $G(f)$

L'expression de $G(f)$ est composée de deux parties :

$$G(f) = A(f)D(f) \tag{A.5}$$

où $A(f)$ et $D(f)$ sont respectivement la fonction d'amplification et la fonction de diminution. L'amplification est mesurée par la racine carrée du rapport entre l'impédance proche de la source Z_s et l'impédance moyenne au site $\bar{Z}(f)$: $A(f) = \sqrt{\frac{Z_s}{\bar{Z}(f)}}$. A noter qu'il est possible de faire une régression linéaire pour avoir l'expression de $A(f)$ suivant le modèle $A(f) = af^b$ à partir des nuages des points de mesures.

La diminution $D(f)$ reflétant la perte d'énergie, est exprimée par : $D(f) = \exp(-\pi\kappa_0 f)$ où κ_0 est

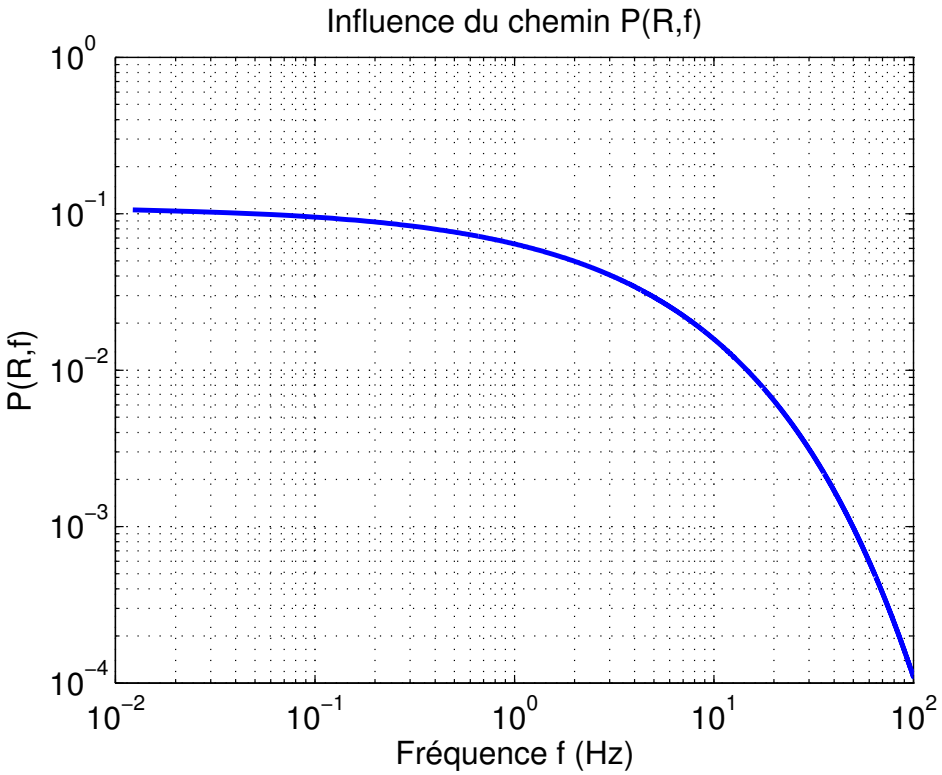


Figure A.2 – Influence du chemin, $R = 9$ km

proche de 0.04 et Boore suggère la valeur de 0.03. La Figure A.3 présente la varation de l'effet du site $G(f)$ en fonction de la fréquence.

Type de réponse étudiée $I(f)$

Le type de réponse est contrôlé par le filtre $I(f) = (2\pi i f)^n$, où $n = 0, 1, 2$ pour le traitement de la réponse respectivement en déplacement, vitesse et accélération. Par exemple, pour l'accélération : $I(f) = -(2\pi f)^2$.

Enfin, en combinant tous les composantes ci-dessus selon la formule (A.1), on peut obtenir le spectre des séismes suivant le modèle de Boore. La Figure A.4 montre un exemple de spectre pour $M = 7$ et $R = 9$ km.

Processus de simulation des séismes suivant le modèle de Boore

La méthode de simulation de Boore pour générer une accélération sismique du sol se résume à l'aide des étapes suivantes :

1. Étape 1 : Générer des bruits Gaussiens pour une durée égale à la durée d'un séisme.
2. Étape 2 : Fenêtrer le signal par une fonction de fenêtre.
3. Étape 3 : Passer dans le domaine de fréquence, et normaliser son amplitude par son écart-type.
4. Étape 4 : Multiplier ce spectre normalisé par le spectre d'un séisme.
5. Étape 5 : Repasser dans le domaine du temps pour obtenir l'accélération du sol.

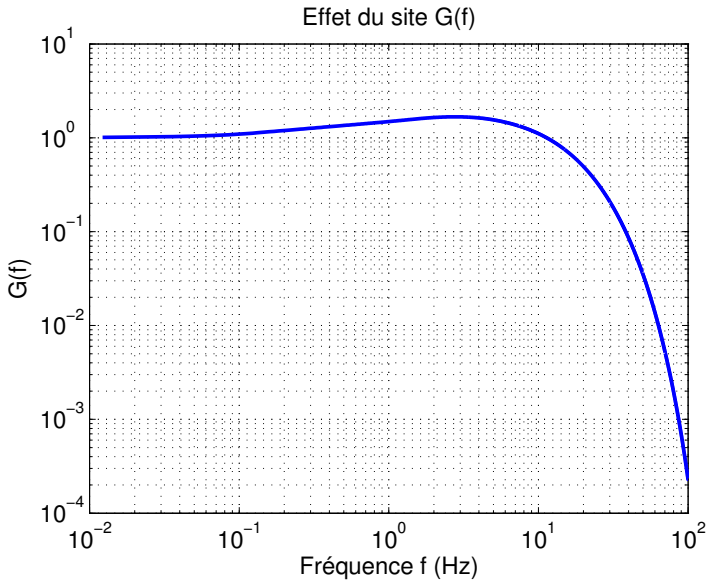


Figure A.3 – Influence de l'effet du site

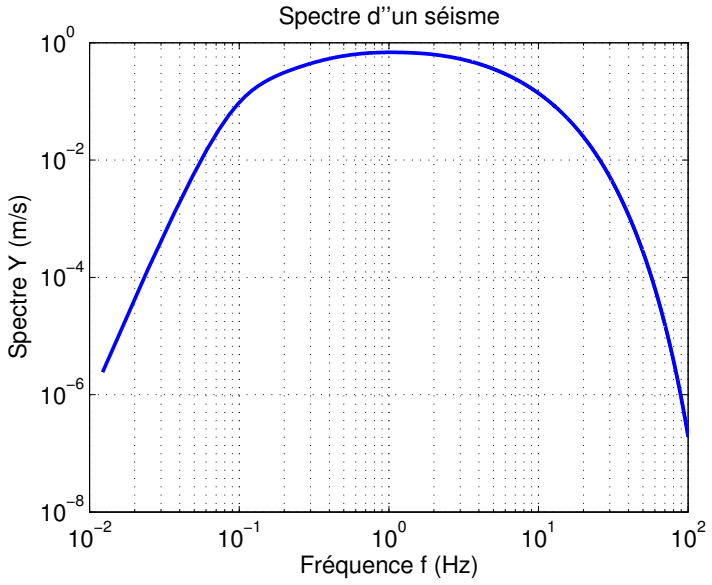


Figure A.4 – Spectre pour $M = 7$ et $R = 9$ km

Bibliographie

[1] R. Kennedy, C. Cornell, R. Campbell, S. Kaplan, and H. Perla. Probabilistic seismic safety study of an existing nuclear power plant. *Nuclear Engineering and Design*, 59(2) :315–338, 1980.

[2] I. Zentner. Numerical computation of fragility curves for NPP equipment. *Nuclear Engineering and Design*, 240(6) :1614–1621, 2010.

[3] K. Kinali and B. R. Ellingwood. Seismic fragility assessment of steel frames for consequence-based engineering : A case study for memphis, TN. *Engineering structures*, 29(6) :1115–1127, 2007.

[4] B. R. Ellingwood. Earthquake risk assessment of building structures. *Reliability Engineering & System Safety*, 74(3) :251–262, 2001.

[5] J. E. Padgett and R. DesRoches. Methodology for the development of analytical fragility curves for retrofitted bridges. *Earthquake Engineering & Structural Dynamics*, 37(8) :1157–1174, 2008.

[6] M. Shinozuka, S.-H. Kim, S. Kushiyama, and J.-H. Yi. Fragility curves of concrete bridges retrofitted by column jacketing. *Earthquake Engineering and Engineering Vibration*, 1(2) :195–205, 2002.

[7] C. Kafali and M. Grigoriu. Seismic fragility analysis : Application to simple linear and nonlinear systems. *Earthquake Engineering & Structural Dynamics*, 36(13) :1885–1900, 2007.

[8] M. Shinozuka, M. Feng, J. Lee, and T. Naganuma. Statistical analysis of fragility curves. *Journal of Engineering Mechanics*, 126(12) :1224–1231, 2000.

[9] C.-T. Dang. *Méthodes de construction des courbes de fragilité sismique par simulations numériques*. PhD thesis, Université Blaise Pascal-Clermont-Ferrand II, 2014.

[10] T.-P. Le, C.-T. Dang, and P. Ray. A comparative study of construction methods for seismic fragility curves using numerical simulations. *Mechanics & Industry*, 17(6) :602, 2016.

[11] M. Shinozuka, M. Feng, H. Kim, and S. Kim. Nonlinear static procedure for fragility curve development. *Journal of Engineering Mechanics*, 126(12) :1287–1295, 2000.

[12] S. Ghosh, S. Ghosh, and S. Chakraborty. Seismic fragility analysis in the probabilistic performance-based earthquake engineering framework : an overview. *International Journal of Advances in Engineering Sciences and Applied Mathematics*, 13(1) :122–135, March 2021.

[13] J.-M. Bourinet. *Reliability analysis and optimal design under uncertainty – Focus on adaptive surrogate-based approaches*. HDR report, Université Clermont Auvergne, France, 2018.

[14] H.-T. Thai. Machine learning for structural engineering : A state-of-the-art review. *Structures*, 38 :448–491, 2022.

[15] A. Géron. *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow*. " O'Reilly Media, Inc.", 2022.

[16] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al. {TensorFlow} : a system for {Large-Scale} machine learning. In *12th USENIX symposium on operating systems design and implementation (OSDI 16)*, pages 265–283, 2016.

[17] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn : Machine learning in Python. *Journal of Machine Learning Research*, 12 :2825–2830, 2011.

[18] E. Fix and J. L. Hodges. Discriminatory analysis. *Nonparametric discrimination : Small sample performance. Report A*, 193008, 1951.

[19] T. Cover and P. Hart. Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1) :21–27, 1967.

[20] D. Von Winterfeldt and W. Edwards. Decision analysis and behavioral research. 1993.

[21] L. Breiman. *Classification and regression trees*. Routledge, 2017.

[22] T. K. Ho. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, pages 278–282. IEEE, 1995.

[23] L. Breiman. Random forests. *Machine learning*, 45 :5–32, 2001.

[24] J. Schmidhuber. Deep learning in neural networks : An overview. *Neural networks*, 61 :85–117, 2015.

[25] B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152, 1992.

[26] V. Vapnik. *The nature of statistical learning theory*. Springer science & business media, 2013.

[27] A. J. Ferreira and M. A. Figueiredo. Boosting algorithms : A review of methods, theory, and applications. *Ensemble machine learning : Methods and applications*, pages 35–85, 2012.

[28] T. Chen and C. Guestrin. Xgboost : A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[29] T. Chen and C. Guestrin. Xgboost : A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[30] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu. Lightgbm : A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30, 2017.

[31] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu. Lightgbm : A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30, 2017.

[32] B. Todorov and A. M. Billah. Machine learning driven seismic performance limit state identification for performance-based seismic design of bridge piers. *Engineering Structures*, 255 :113919, 2022.

[33] B. B. Adhikary and H. Mutsuyoshi. Artificial neural networks for the prediction of shear capacity of steel plate strengthened rc beams. *Construction and Building Materials*, 18(6) :409–417, 2004.

[34] H. Naderpour, O. Poursaeidi, and M. Ahmadi. Shear resistance prediction of concrete beams reinforced by frp bars using artificial neural networks. *Measurement*, 126 :299–308, 2018.

[35] S. Lee, T. P. Vo, H.-T. Thai, J. Lee, and V. Patel. Strength prediction of concrete-filled steel tubular columns using categorical gradient boosting algorithm. *Engineering Structures*, 238 :112109, 2021.

[36] Y. Xu, B. Zheng, and M. Zhang. Capacity prediction of cold-formed stainless steel tubular columns using machine learning methods. *Journal of Constructional Steel Research*, 182 :106682, 2021.

[37] A. Amini Pishro, S. Zhang, D. Huang, F. Xiong, W. Li, and Q. Yang. Application of artificial neural networks and multiple linear regression on local bond stress equation of uhpc and reinforcing steel bars. *Scientific Reports*, 11(1) :1–20, 2021.

[38] R. Haddad and M. Haddad. Predicting fiber-reinforced polymer–concrete bond strength using artificial neural networks : A comparative analysis study. *Structural Concrete*, 22(1) :38–49, 2021.

[39] H. Erdem. Prediction of the moment capacity of reinforced concrete slabs in fire using artificial neural networks. *Advances in Engineering Software*, 41(2) :270–276, 2010.

[40] S. Li, J. R. Liew, and M.-X. Xiong. Prediction of fire resistance of concrete encased steel composite columns using artificial neural network. *Engineering Structures*, 245 :112877, 2021.

[41] O. Belalia Douma, B. Boukhatem, M. Ghrici, and A. Tagnit-Hamou. Prediction of properties of self-compacting concrete containing fly ash using artificial neural network. *Neural Computing and Applications*, 28 :707–718, 2017.

[42] J. Xu, X. Zhao, Y. Yu, T. Xie, G. Yang, and J. Xue. Parametric sensitivity analysis and modelling of mechanical properties of normal-and high-strength recycled aggregate concrete using grey theory, multiple nonlinear regression and artificial neural networks. *Construction and Building Materials*, 211 :479–491, 2019.

[43] A. A. Shahmansouri, M. Yazdani, S. Ghanbari, H. A. Bengar, A. Jafari, and H. F. Ghatte. Artificial neural network model to predict the compressive strength of eco-friendly geopolymer concrete incorporating silica fume and natural zeolite. *Journal of Cleaner Production*, 279 :123697, 2021.

[44] B. P. Koya, S. Aneja, R. Gupta, and C. Valeo. Comparative analysis of different machine learning algorithms to predict mechanical properties of concrete. *Mechanics of Advanced Materials and Structures*, 29(25) :4032–4043, 2022.

[45] T. Ikumi, E. Galeote, P. Pujadas, A. de la Fuente, and R. López-Carreño. Neural network-aided prediction of post-cracking tensile strength of fibre-reinforced concrete. *Computers & Structures*, 256 :106640, 2021.

[46] F. Demir. Prediction of elastic modulus of normal and high strength concrete by artificial neural networks. *Construction and building Materials*, 22(7) :1428–1435, 2008.

[47] E. M. Golafshani and A. Behnood. Application of soft computing methods for predicting the elastic modulus of recycled aggregate concrete. *Journal of cleaner production*, 176 :1163–1176, 2018.

[48] M. A. Shaheen, R. Presswood, and S. Afshan. Application of machine learning to predict the mechanical properties of high strength steel at elevated temperatures based on the chemical composition. In *Structures*, volume 52, pages 17–29. Elsevier, 2023.

[49] J. Xiong, T. Zhang, and S. Shi. Machine learning of mechanical properties of steels. *Science China Technological Sciences*, 63(7) :1247–1255, 2020.

[50] N. Sandhya, V. Sowmya, C. Bandaru, and G. R. Babu. Prediction of mechanical properties of steel using data science techniques. *Int. J. Recent Technol. Eng*, 8 :235–241, 2019.

[51] Y. Okazaki, S. Okazaki, S. Asamoto, and P.-j. Chun. Applicability of machine learning to a crack model in concrete bridges. *Computer-Aided Civil and Infrastructure Engineering*, 35(8) :775–792, 2020.

[52] R. Ghiasi, M. R. Ghasemi, and M. Noori. Comparative studies of metamodeling and ai-based techniques in damage detection of structures. *Advances in Engineering Software*, 125 :101–112, 2018.

[53] S. Hakim, H. A. Razak, S. Ravanfar, and M. Mohammadhassani. Structural damage detection using soft computing method. In *Structural Health Monitoring, Volume 5 : Proceedings of the 32nd IMAC, A Conference and Exposition on Structural Dynamics, 2014*, pages 143–151. Springer, 2014.

[54] H. Adeli and C. Yeh. Perceptron learning in engineering design. *Computer-Aided Civil and Infrastructure Engineering*, 4(4) :247–256, 1989.

[55] M. Z. Esteghamati and M. M. Flint. Developing data-driven surrogate models for holistic performance-based assessment of mid-rise rc frame buildings at early design. *Engineering Structures*, 245 :112971, 2021.

[56] A. E. Charalampakis and V. K. Papanikolaou. Machine learning design of r/c columns. *Engineering Structures*, 226 :111412, 2021.

[57] B. K. Oh and J. Kim. Optimal architecture of a convolutional neural network to estimate structural responses for safety evaluation of the structures. *Measurement*, 177 :109313, 2021.

[58] S.-H. Hwang, S. Mangalathu, J. Shin, and J.-S. Jeon. Machine learning-based approaches for seismic demand and collapse of ductile reinforced concrete building frames. *Journal of Building Engineering*, 34 :101905, 2021.

[59] J. Kiani, C. Camp, and S. Pezeshk. On the application of machine learning techniques to derive seismic fragility curves. *Computers & Structures*, 218 :108–122, 2019.

[60] Z. Wang, N. Pedroni, I. Zentner, and E. Zio. Seismic fragility analysis with artificial neural networks : Application to nuclear power plant equipment. *Engineering Structures*, 162 :213–225, 2018.

[61] S. Mangalathu, H. Jang, S.-H. Hwang, and J.-S. Jeon. Data-driven machine-learning-based seismic failure mode identification of reinforced concrete shear walls. *Engineering Structures*, 208 :110331, 2020.

[62] M. Usta and S. Pujol. Aci subcommittee 445b, puranam a, song c, wang y. aci 445b shear wall database. purdue university research repository 2017.

[63] S. Grammatikou, D. Biskinis, and M. N. Fardis. Strength, deformation capacity and failure modes of rc walls under cyclic loading. *Bulletin of earthquake engineering*, 13 :3277–3300, 2015.

[64] A. C. Catlin, C. HewaNadungodage, S. Pujol, L. Laughery, C. Sim, A. Puranam, and A. Bejarano. A cyberplatform for sharing scientific research data at datacenterhub. *Computing in Science & Engineering*, 20(3) :49–70, 2018.

[65] T. D. Ancheta, R. B. Darragh, J. P. Stewart, E. Seyhan, W. J. Silva, B. S.-J. Chiou, K. E. Wooddell, R. W. Graves, A. R. Kottke, D. M. Boore, et al. Nga-west2 database. *Earthquake Spectra*, 30(3) :989–1005, 2014.

[66] E. M. Rathje, C. Dawson, J. E. Padgett, J.-P. Pinelli, D. Stanzione, A. Adair, P. Arduino, S. J. Brandenberg, T. Cockerill, C. Dey, et al. Designsafe : New cyberinfrastructure for natural hazards engineering. *Natural Hazards Review*, 18(3) :06017001, 2017.

[67] Y. Zhou and M. Kantarcioglu. On transparency of machine learning models : A position paper. In *AI for Social Good Workshop*, 2020.

[68] B. Sun, Y. Zhang, and C. Huang. Machine learning-based seismic fragility analysis of large-scale steel buckling restrained brace frames. *Computer Modeling in Engineering & Sciences*, 125(2) :755–776, 2020.

[69] N. D. Lagaros and M. Fragiadakis. Fragility assessment of steel frames using neural networks. *Earthquake Spectra*, 23(4) :735–752, 2007.

[70] Y. Zhou, Y. Zhang, R. Pang, and B. Xu. Seismic fragility analysis of high concrete faced rockfill dams based on plastic failure with support vector machine. *Soil Dynamics and Earthquake Engineering*, 144 :106587, May 2021.

[71] C. C. Mitropoulou and M. Papadrakakis. Developing fragility curves based on neural network ida predictions. *Engineering Structures*, 33(12) :3409–3421, 2011.

[72] R. Segura, J. Padgett, and P. Paultre. Metamodel-based seismic fragility analysis of concrete gravity dams. *Journal of Structural Engineering*, 146 :04020121, 04 2020.

[73] H. Rezaei, P. Zarfam, E. M. Golafshani, and G. G. Amiri. Seismic fragility analysis of RC box-girder bridges based on symbolic regression method. *Structures*, 38 :306–322, April 2022.

[74] M. Noureldin, A. Ali, S. Sim, and J. Kim. A machine learning procedure for seismic qualitative assessment and design of structures considering safety and serviceability. *Journal of Building Engineering*, 50 :104190, 2022.

[75] H. Zhang, X. Cheng, Y. Li, D. He, and X. Du. Rapid seismic damage state assessment of rc frames using machine learning methods. *Journal of Building Engineering*, 65 :105797, 2023.

[76] H. D. Nguyen, J. M. LaFave, Y.-J. Lee, and M. Shin. Rapid seismic damage-state assessment of steel moment frames using machine learning. *Engineering Structures*, 252 :113737, 2022.

[77] X. Guan, H. Burton, M. Shokrabadi, and Z. Yi. Seismic drift demand estimation for steel moment frame buildings : From mechanics-based to data-driven models. *journal of structural engineering*, 147(6) :04021058, 2021.

[78] K. Demertzis, K. Kostinakis, K. Morfidis, and L. Iliadis. An interpretable machine learning method for the prediction of r/c buildings' seismic response. *Journal of Building Engineering*, 63 :105493, 2023.

[79] S. Mangalathu and J.-S. Jeon. Stripe-based fragility analysis of multispan concrete bridge classes using machine learning techniques. *Earthquake Engineering & Structural Dynamics*, 48(11) :1238–1255, 2019.

[80] D.-T. Vu and N.-D. Hoang. Punching shear capacity estimation of frp-reinforced concrete slabs using a hybrid machine learning approach. *Structure and Infrastructure Engineering*, 12(9) :1153–1161, 2016.

[81] N.-D. Hoang, X.-L. Tran, and H. Nguyen. Predicting ultimate bond strength of corroded reinforcement and surrounding concrete using a metaheuristic optimized least squares support vector regression model. *Neural Computing and Applications*, 32 :7289–7309, 2020.

[82] H. Luo and S. G. Paal. Data-driven seismic response prediction of structural components. *Earthquake Spectra*, 38(2) :1382–1416, 2022.

[83] S. Gajan. Data-driven modeling of peak rotation and tipping-over stability of rocking shallow foundations using machine learning algorithms. *Geotechnics*, 2(3) :781–801, 2022.

[84] C. Huang, Y. Li, Q. Gu, and J. Liu. Machine learning–based hysteretic lateral force-displacement models of reinforced concrete columns. *Journal of Structural Engineering*, 148(3) :04021291, 2022.

[85] B. Sudret. Meta-models for structural reliability and uncertainty quantification. Technical Report arXiv :1203.2062, arXiv, March 2012. arXiv :1203.2062 [stat] type : article.

[86] S. Mazzoni, F. McKenna, M. H. Scott, G. L. Fenves, et al. Opensees command language manual. *Pacific earthquake engineering research (PEER) center*, 264(1) :137–158, 2006.

[87] P. P. Cordova, G. G. Deierlein, S. S. Mehanny, and C. A. Cornell. Development of a two-parameter seismic intensity measure and probabilistic assessment procedure. In *The second US-Japan workshop on performance-based earthquake engineering methodology for reinforced concrete building structures*, volume 20, page 0, 2000.

[88] M. Hariri-Ardebili and V. Saouma. Probabilistic seismic demand model and optimal intensity measure for concrete dams. *Structural Safety*, 59 :67–85, 2016.

[89] J. W. Baker and C. Allin Cornell. A vector-valued ground motion intensity measure consisting of spectral acceleration and epsilon. *Earthquake Engineering & Structural Dynamics*, 34(10) :1193–1217, 2005.

[90] E. Bojórquez, I. Iervolino, A. Reyes-Salazar, and S. E. Ruiz. Comparing vector-valued intensity measures for fragility analysis of steel frames in the case of narrow-band ground motions. *Engineering Structures*, 45 :472–480, 2012.

[91] M. De Biasio. *Ground motion intensity measures for seismic probabilistic risk analysis*. PhD thesis, Université de Grenoble, 2014.

[92] M. Biot. Theory of elastic systems under transient loading with application to earthquake engineering. *Proceedings National Academy of Science*, 19(1933) :262–268, 1933.

[93] G. W. Housner. Calculating the response of an oscillator to arbitrary ground motion. *Bulletin of the Seismological Society of America*, 31(2) :143–149, 1941.

[94] P. Paultre. *Dynamics of structures*. John Wiley & Sons, 2013.

[95] T. Kim, O.-S. Kwon, and J. Song. Response prediction of nonlinear hysteretic systems by deep neural networks. *Neural Networks*, 111 :1–10, 2019.

[96] U. of California. Peer nga database.

[97] L. Luzi, R. Puglia, E. Russo, M. D'Amico, C. Felicetta, F. Pacor, G. Lanzano, U. Çeken, J. Clinton, G. Costa, et al. The engineering strong-motion database : A platform to access pan-european accelerometric data. *Seismological Research Letters*, 87(4) :987–997, 2016.

[98] I. Zentner and F. Poirion. Enrichment of seismic ground motion databases using Karhunen - Loève expansion. *Earthquake Engineering & Structural Dynamics*, 41(14) :1945–1957, 2012.

[99] R. W. Clough and J. Penzien. *Dynamics of structures*, volume 2. McGraw-Hill New York, 1993.

[100] G. Rodolfo Saragoni and G. Hart. Simulation of artificial earthquakes. *Earthquake Engineering & Structural Dynamics*, 2(3) :249–267, 1973.

[101] S. Rezaeian and A. Der Kiureghian. Simulation of synthetic ground motions for specified earthquake and site characteristics. *Earthquake Engineering & Structural Dynamics*, 39(10) :1155–1180, 2010.

[102] M. Shinozuka and G. Deodatis. Simulation of stochastic processes by spectral representation. *Applied Mechanics Reviews*, 44 :191, 1991.

[103] M. Loève. *Probability Theory. Foundations. Random Sequences*. Van Nostrand Company, New York, 1955.

[104] J. Li and J. B. Chen. *Stochastic dynamics of structures*. John Wiley & Sons, 2009.

[105] H. Hwang and J. Huo. Generation of hazard-consistent fragility curves. *Soil Dynamics and Earthquake Engineering*, 13(5) :345–354, 1994.

[106] A. Filiatrault and W. A. *Simulation of strong ground motions for seismic fragility evaluation of nonstructural components in hospitals*. Multidisciplinary Center for Earthquake Engineering Research MCEER, University at Buffalo, State University of New York, 2005.

[107] D. M. Boore. Simulation of ground motion using the stochastic method. *Pure and Applied Geophysics*, 160(3) :635–676, 2003.

[108] T. Developers. Tensorflow. *Zenodo*, 2022.

[109] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc., 2017.

[110] R. Bouc. Modèle mathématique d'hystérésis. *Acustica*, 21 :16–25, 1971.

[111] Y. Wen. Methods of random vibration for inelastic structures. 1989.

[112] V. Kumar and S. Minz. Feature selection : a literature review. *SmartCR*, 4(3) :211–229, 2014.

[113] J. Miao and L. Niu. A survey on feature selection. *Procedia computer science*, 91 :919–926, 2016.

[114] B. Venkatesh and J. Anuradha. A review of feature selection and its methods. *Cybernetics and information technologies*, 19(1) :3–26, 2019.

[115] J. B. Chen and J. Li. Stochastic seismic response analysis of structures exhibiting high nonlinearity. *Computers & structures*, 88(7) :395–412, 2010.

[116] J. W. Baker and C. Allin Cornell. Spectral shape, epsilon and record selection. *Earthquake Engineering & Structural Dynamics*, 35(9) :1077–1095, 2006.

[117] M. Kohrangi, P. Bazzurro, D. Vamvatsikos, and A. Spillatura. Conditional spectrum-based ground motion record selection using average spectral acceleration. *Earthquake Engineering & Structural Dynamics*, 46(10) :1667–1685, 2017.

[118] A. S. of Civil Engineers. Minimum design loads and associated criteria for buildings and other structures. American Society of Civil Engineers, 2017.

[119] N. Jayaram, T. Lin, and J. W. Baker. A computationally efficient ground-motion selection algorithm for matching a target response spectrum mean and variance. *Earthquake spectra*, 27(3) :797–815, 2011.

[120] V. Ozsarac. EzGM, November 2023. original-date : 2020-09-01T08 :43 :48Z.

[121] G. Lanzano, S. Sgobba, L. Luzi, R. Puglia, F. Pacor, C. Felicetta, M. D'Amico, F. Cotton, and D. Bindi. The pan-european engineering strong motion (esm) flatfile : compilation criteria and data statistics. *Bulletin of Earthquake Engineering*, 17 :561–582, 2019.

[122] M. Pagani, D. Monelli, G. Weatherill, L. Danciu, H. Crowley, V. Silva, P. Henshaw, L. Butler, M. Nastasi, L. Panzeri, et al. Openquake engine : An open hazard (and risk) software for the global earthquake model. *Seismological Research Letters*, 85(3) :692–702, 2014.

[123] D. M. Boore, J. P. Stewart, E. Seyhan, and G. M. Atkinson. Nga-west2 equations for predicting pga, pgv, and 5% damped psa for shallow crustal earthquakes. *Earthquake Spectra*, 30(3) :1057–1085, 2014.

[124] J. W. Baker and N. Jayaram. Correlation of spectral acceleration values from nga ground motion models. *Earthquake Spectra*, 24(1) :299–317, 2008.

[125] P. Bisch, E. Carvalho, H. Degee, P. Fajfar, M. Fardis, P. Franchin, M. Kreslin, A. Pecker, P. Pinto, A. Plumier, et al. Eurocode 8 : seismic design of buildings worked examples. *Luxembourg : Publications Office of the European Union*, 2012.

[126] A. T. Council. *Quantification of building seismic performance factors*. US Department of Homeland Security, FEMA, 2009.

[127] X. Guan M. EERI, H. Burton M. EERI, and M. Shokrabadi. A database of seismic designs, nonlinear models, and seismic responses for steel moment-resisting frame buildings. *Earthquake Spectra*, 37(2) :1199–1222, 2021.

[128] E. Miranda. Approximate seismic lateral deformation demands in multistory buildings. *Journal of Structural Engineering*, 125(4) :417–425, 1999.

[129] Ö. Şahin and N. Caglar. Simulation-based model-updating method for linear dynamic structural systems. *Applied Sciences*, 13(18) :10494, 2023.

[130] L. F. Ibarra, R. A. Medina, and H. Krawinkler. Hysteretic models that incorporate strength and stiffness deterioration. *Earthquake engineering & structural dynamics*, 34(12) :1489–1511, 2005.

[131] A. Gupta. *Seismic demands for performance evaluation of steel moment resisting frame structures*. Stanford University, 1999.