TECHNIQUES
DE L'INGÉNIEUR

Réf. : **C231 V1**

Date de publication :
10 mai 2025

Apprentissage automatique en géotechnique : étude de cas dans le domaine des tunnels

Cet article est issu de : **Construction et travaux publics | Mécanique des sols et géotechnique**

par **Tatiana RICHA,**
Lina-María GUAYACÁN-CARRILLO,
Jean-Michel PEREIRA, Gilles CHAPRON

Mots-clés
intelligence artificielle | traitement de données | géotechnique | données d'auscultation

Résumé Cet article présente, après un état de l'art sur l'utilisation de l'intelligence artificielle (IA) en géotechnique, une méthodologie détaillée pour l'application du machine learning (ML) dans des cas pratiques en géotechnique, avec un focus particulier sur la prédiction des tassements induits par le creusement des tunnels. Chaque étape du processus, du cadrage du problème à la conception du modèle, en passant par la préparation des données, l'entraînement des algorithmes, et l'obtention du modèle final, est illustrée par des exemples concrets issus de cette problématique. L'article met également en lumière les défis associés à chaque phase, de l'élaboration des données à leur nettoyage, de l'entraînement des modèles à leur validation et optimisation, offrant ainsi une approche méthodique pour intégrer le ML dans les projets géotechniques.

Keywords
artificial intelligence | data processing | geotechnics | monitoring data

Abstract This article presents, after a review of the state of the art on the use of Artificial Intelligence (AI) in geotechnics, a detailed methodology for the application of Machine Learning (ML) in practical geotechnical cases, with a particular focus on predicting settlements induced by tunnel excavation. Each step of the process, from problem scoping to model design, including data preparation, algorithm training, and obtaining the final model, is illustrated with concrete examples from this issue. The article also highlights the challenges associated with each phase, from data development and cleaning to model training, validation, and optimization, thus providing a structured approach to integrating ML into geotechnical projects.

Pour toute question :
Service Relation clientèle
Techniques de l'Ingénieur
Immeuble Pleyad 1
39, boulevard Ornano
93288 Saint-Denis Cedex

Par mail :
infos.clients@teching.com
Par téléphone :
00 33 (0)1 53 35 20 20

Apprentissage automatique en géotechnique : étude de cas dans le domaine des tunnels

par **Tatiana RICHA**

Ingénieure data en géotechnique
Terrasol Setec, Paris, France

Lina-María GUAYACÁN-CARRILLO

Chargée de recherche en géotechnique (laboratoire Navier), maître de conférences de l'ENPC
École nationale des ponts et chaussées, Institut polytechnique de Paris, Marne-La-Vallée, France

Jean-Michel PEREIRA

Directeur du laboratoire Navier, professeur de l'ENPC
École nationale des ponts et chaussées, Institut polytechnique de Paris, Marne-La-Vallée, France

et **Gilles CHAPRON**

Directeur des projets data
Terrasol Setec, Paris France

Le machine learning (ML), ou apprentissage automatique en français, est une branche de l'intelligence artificielle (IA) qui permet aux systèmes informatiques de s'améliorer automatiquement grâce à l'expérience. Or, la géotechnique, discipline à l'interface entre la géologie et le génie civil, fait face à des défis croissants de complexité et de précision dans l'analyse des sols et la conception des ouvrages. Dans ce contexte, l'émergence du ML comme outil d'analyse et de prédiction ouvre des perspectives prometteuses pour répondre aux enjeux contemporains du secteur.

Les méthodes traditionnelles de calcul en géotechnique, bien qu'éprouvées, présentent certaines limitations face à la complexité inhérente des sols et des interactions sol-structure. La variabilité naturelle des paramètres géotechniques, la non-linéarité des comportements mécaniques et la multiplicité des facteurs

environnementaux rendent parfois difficile l'application des approches analytiques classiques.

L'IA, et plus particulièrement le ML, apporte une nouvelle dimension à l'analyse géotechnique en permettant d'exploiter les vastes quantités de données accumulées par la profession depuis des décennies, ou plus modestement à l'échelle d'un projet, pour recaler un modèle au fur et à mesure de la réalisation des travaux. Ces techniques permettent de détecter des tendances et motifs complexes dans les données, d'automatiser certaines tâches d'analyse, et d'améliorer la précision des prédictions géotechniques, terme consacré dans ce domaine, mais on pourrait utiliser également celui de prévision.

Les deux exemples suivants illustrent deux cas d'usage simple de l'IA appliqué à des problématiques géotechniques :

- *la prédiction du tassement d'un remblai sur sol compressible nécessite traditionnellement des calculs complexes intégrant de nombreux paramètres (indice de compression, contraintes effectives, etc.) ; le ML pourrait permettre d'enrichir cette approche en exploitant les retours d'expérience de projets similaires pour affiner les prédictions ;*
- *l'utilisation des réseaux de neurones ou d'arbres de décision pour la classification des sols peut permettre d'accélérer voire d'automatiser l'interprétation des essais in situ, réduisant le temps d'analyse tout en maintenant un haut niveau de précision.*

Dans cet article, nous explorerons l'application du ML à la géotechnique selon trois axes principaux :

1/ les concepts fondamentaux du ML et leur état de l'art en géotechnique, illustrés par des cas d'application concrets qui démontrent la pertinence de ces approches dans notre domaine ;

2/ le cadrage et la préparation des projets de ML en géotechnique, étape cruciale qui conditionne la réussite de la démarche, depuis la structuration des données jusqu'à la sélection des données d'entrée pertinentes ;

3/ la mise en œuvre pratique, de l'entraînement des modèles jusqu'à leur déploiement en conditions réelles, en passant par leur optimisation et leur évaluation.

Notre approche vise à démystifier l'application du ML en géotechnique en associant systématiquement les concepts théoriques à un exemple pratique « fil rouge ». L'application présentée est la prédiction du tassement induit par le creusement au tunnelier. Cette démarche a vocation à rendre ces notions complexes aussi claires et applicables que possible et à donner aux lecteurs les clefs pour comprendre non seulement les grands principes sous-jacents, mais aussi les modalités concrètes de mise en œuvre de ces techniques dans leur pratique professionnelle.

L'objectif est de fournir aux géotechniciens les clefs pour intégrer ces nouvelles approches dans leur boîte à outils, en complément – et non en remplacement – des méthodes traditionnelles. Car si le ML ouvre de nouvelles perspectives, il ne remplace pas l'expertise de l'ingénieur, mais vient plutôt l'enrichir en lui permettant de traiter plus efficacement des problèmes complexes.

***Note :** l'IA et le ML constituent un sujet en évolution rapide. Les avancées technologiques et méthodologiques se succèdent et s'ouvrent au plus grand nombre, notamment à travers la démocratisation des grands modèles de langage (LLM pour large language model) dont le plus connu d'entre eux au moment où l'article est écrit est ChatGPT. Cet article ne peut donc offrir qu'un instantané des connaissances et des pratiques. Notamment, il n'aborde pas les modèles de langage, ou plus généralement l'IA générative.*

1. Préambule : dans le monde de l'intelligence artificielle

1.1 Éléments de cadrage

1.1.1 Cadre historique et notions essentielles

L'application des outils de ML a récemment pris de l'ampleur dans le domaine de la géotechnique. La définition du concept d'IA et le développement d'algorithmes de ML remontent à plusieurs décennies. En 1956, la terminologie « intelligence artificielle » est proposée pour la première fois par John McCarthy lors d'une conférence du Dartmouth College aux États-Unis [1]. Cependant, les bases de l'IA avaient déjà été posées dès 1950 par Alan Turing dans son article « Computing Machinery and Intelligence » publié dans le journal britannique *Mind* [2]. Par la suite, le développement des algorithmes a connu une croissance au fil du temps, commençant par le premier algorithme de ML connu, le perceptron [3]. Au cours des années 2000 et 2010, le ML a connu une croissance explosive grâce à l'augmentation des capacités de calcul et aux avancées en matière de traitement des données. C'est l'ère de l'apprentissage profond (DL, *deep learning*), qui a permis de résoudre des problèmes complexes en utilisant des réseaux de neurones profonds (ANN pour *artificial neural network*) et sophistiqués. De nouveaux algorithmes continuent d'être développés, comme l'algorithme *Extreme Gradient Boosting* (XGBoost) [4] [5], qui a rapidement gagné en popularité dans la communauté des data scientists grâce à son succès dans de nombreux concours de ML, dont il a souvent remporté le premier prix. Aujourd'hui, le ML est largement utilisé dans l'industrie et le monde académique pour résoudre des problèmes complexes de traitement des données dans de nombreux domaines.

Avant tout, trois concepts importants méritent d'être définis et distingués.

- **Intelligence artificielle** (*Artificial Intelligence*) : selon la CNIL (Commission nationale de l'informatique et des libertés) et le Parlement européen [6], l'intelligence artificielle est un domaine scientifique qui regroupe tout outil utilisé par une machine afin de « reproduire des comportements liés aux humains, tels que le raisonnement, la planification et la créativité ». Tout système mettant en œuvre des mécanismes proches de ceux du raisonnement humain pourrait ainsi être qualifié d'intelligence artificielle.

- **Apprentissage automatique** (*Machine Learning*) : parmi les domaines d'application de l'IA (par exemple robotique, vision, traitement du langage…), on distingue la branche de l'apprentissage automatique. Turing (1950) [2] fait référence à l'apprentissage des machines (qui deviennent apprenantes, cf. « Learning Machines » selon ses termes) en se posant la question suivante : au lieu d'essayer de produire un programme qui simule l'esprit d'un adulte, pourquoi ne pas essayer de produire un programme qui simule celui d'un enfant ? En effet, il affirme qu'il est espéré que les mécanismes du cerveau d'un enfant sont si peu nombreux qu'il sera facile de programmer quelque chose de semblable. La machine aurait ainsi besoin d'un programme (le cerveau) et d'un apprentissage (l'éducation) pour déduire des relations entre différents paramètres. Samuel (1959) [7] argumente le fait que programmer les ordinateurs pour qu'ils apprennent par l'expérience devrait à terme éliminer la nécessité d'un effort de programmation détaillée, une procédure qui prend du temps et qui est coûteuse. En effet, le ML n'est fondamentalement rien de plus que la capacité d'un programme à évaluer statistiquement un résultat à partir d'un jeu de données, et à réduire l'écart entre ce résultat et la valeur vraie via un apprentissage itératif.

- **Données massives** (*big data*) : ce concept fait référence à des jeux de données qui atteignent une taille telle qu'ils excèdent les capacités d'analyse humaine ainsi que celles des outils informatiques traditionnels. Le traitement de ces ensembles de données a nécessité la mise au point de technologies spécifiquement adaptées. Le big data regroupe cinq notions qu'on appelle les « 5V » [8] :

  1/ volume : avec le développement des nouvelles technologies comme les IoT (*Internet of Things* : appareils physiques qui reçoivent et transfèrent des données sur des réseaux sans fil, avec une intervention humaine limitée), la production de données numériques a été de plus en plus massive : textes, photos, vidéos, etc. ;

  2/ vélocité : besoin d'analyse des données en temps réel et de prise de décision en une fraction de seconde ;

  3/ variété : les données d'intérêt sont hétérogènes par nature, avec un format structuré ou non : vidéos, géolocalisation, échanges vocaux, messages sur les réseaux sociaux, etc. ;

  4/ véracité : l'incertitude associée aux données doit être quantifiée en matière de précision, de validité, de cohérence, de fiabilité et de qualité ;

  5/ valeur : c'est un des points les plus importants à notre avis, mais rarement mentionné ; l'utilisation des technologies de stockage et d'analyse des big data n'a de sens que si elle apporte de la valeur ajoutée : exploiter les données, c'est avant tout répondre à des objectifs métiers ; en conséquence, la création et le stockage de big data ont encouragé le développement de méthodes statistiques capables de traiter d'énormes quantités de données afin d'en extraire du sens et de produire de la valeur ; l'avènement du big data représente donc un point de bascule majeur pour la science des données, le ML et l'IA.

1.1.2 Aperçu général de l'apprentissage automatique

Les systèmes d'apprentissage automatique peuvent être catégorisés selon plusieurs critères [9].

1) Par mode de généralisation : selon que l'apprentissage se base sur l'exemple (*instance based learning*) ou sur l'optimisation d'un modèle (*model-based learning*). Dans le premier, l'algorithme de ML fonctionne en comparant simplement de nouveaux points de données à des points de données connus, alors que dans le deuxième, l'algorithme détecte des tendances dans les données d'apprentissage et construit un modèle prédictif.

2) Par mode de consommation des données : selon que l'apprentissage est réalisé par lot statique (ou groupé, *batch learning*) ou sur flux continu (incrémental, *online learning*). Dans le premier, le modèle obtenu se contente d'appliquer sur les nouvelles données ce qu'il a déjà appris dans la phase d'entraînement, alors que dans le deuxième, l'algorithme apprend d'une manière incrémentale à partir d'un flux de données entrant [10].

(3) Par mode d'apprentissage. **Supervisé** (*supervised*) : l'algorithme d'apprentissage est entraîné en utilisant des données correctement étiquetées au préalable (classe, valeur continue, etc.). En d'autres termes, la sortie cible (*target output*) y est connue et est injectée dans le modèle lors de la phase d'apprentissage. **Non supervisé** (*unsupervised*) : les données d'apprentissage ne sont pas étiquetées (c'est-à-dire non caractérisées) ; le modèle est entraîné pour identifier des motifs dans les données. Par **renforcement** (*reinforcement*) : l'algorithme d'apprentissage peut observer un environnement, sélectionner et effectuer des actions, et obtenir des récompenses (ou des pénalités sous forme de récompenses négatives). Il doit ensuite progressivement découvrir, par lui-même, la meilleure stratégie pour obtenir la meilleure récompense. Il faut noter qu'il existe également un apprentissage **semi-supervisé** (*semi-supervised*) qui utilise les deux types de données, avec une quantité beaucoup plus importante de données non étiquetées que de données étiquetées.

Ces dernières années, l'apprentissage machine **génératif** profond a gagné en utilisation [11] [12], comme l'expliquent Chen *et al.* [13] ; l'apprentissage se concentre sur la modélisation des relations causales sous-jacentes et des mécanismes génératifs des données, plutôt que de simplement corréler les entrées et les sorties les unes avec les autres. Cependant, cet article n'aborde pas les IA génératives.

Une vue d'ensemble de l'application de ces différentes sous-catégories dans le domaine de la géotechnique est présentée dans différents articles récents présentant un état de l'art détaillé [14] [15] [16] [17] [18].

Dans ce qui suit sont présentés brièvement quelques algorithmes de ML qui seront utilisés dans la suite de l'article.

Artificial neural networks

Les réseaux de neurones artificiels (ANN, *Artificial Neural Network*) sont des algorithmes largement utilisés en ML supervisé, inspirés de la structure du cerveau humain [19] [20]. Ils sont utilisés pour résoudre des problèmes de classification, de régression, de reconnaissance de formes et de traitement du langage naturel, entre autres. Un réseau de neurones artificiels est composé d'un ensemble de nœuds interconnectés, appelés neurones, organisés en couches. Les neurones de la couche d'entrée reçoivent les données d'entrée, les neurones des couches cachées effectuent des calculs de traitement des données et les neurones de la couche de sortie renvoient les résultats. Chaque neurone effectue une fonction mathématique simple, généralement une somme pondérée des entrées suivie d'une fonction d'activation non linéaire. Les poids associés à chaque entrée sont ajustés pendant la phase d'apprentissage pour minimiser l'erreur de prédiction. Le processus d'apprentissage se fait par rétropropagation (*backpropagation*), qui consiste à calculer l'erreur de prédiction et à propager cette erreur à travers le réseau pour ajuster les poids des neurones de manière itérative jusqu'à ce que l'erreur soit minimisée.

Les réseaux de neurones peuvent être construits avec différentes architectures et fonctionnalités, telles que des réseaux de neurones convolutionnels (CNN pour *Convolutional Neural Network*) pour la reconnaissance d'images, des réseaux de neurones récurrents (*recurrent neural network*) pour la modélisation de séquences, ou des réseaux de neurones adverses (*adversarial neural network*) pour la génération de contenus créatifs.

Support vector machine

SVM est un algorithme de ML supervisé qui est utilisé pour résoudre des problèmes de classification et de régression [21] [22]. L'idée de base est de trouver un hyperplan qui sépare de manière optimale les données en deux classes. Pour ce faire, l'algorithme cherche à maximiser la marge entre les exemples d'entraînement les plus proches de chaque classe, appelés vecteurs de support (*support vectors*), et l'hyperplan de décision tout en minimisant l'erreur (figure **1**). Dans le cas où les données ne sont pas linéairement séparables, SVM utilise une technique appelée « fonction de noyau » (*kernel function*) pour projeter les données dans un espace de dimension supérieure où elles pourront l'être.

Decision trees

DT, arbre de décision en français, est un algorithme de ML supervisé qui est utilisé pour résoudre des problèmes de classification et de régression [23]. L'idée est de diviser récursivement l'espace de recherche en utilisant des règles de décision simples sur les différentes caractéristiques (*features*) des données. L'algorithme commence par le nœud racine (*root node*), qui représente l'ensemble des données d'entrée. À chaque nœud, l'algorithme sélectionne la caractéristique qui fournit la meilleure séparation possible des données vis-à-vis de la variable cible. Cette caractéristique est utilisée pour diviser les données en sous-ensembles plus petits, qui sont ensuite traités de manière récursive pour former des sous-arbres (*branch*). Le processus de division se poursuit jusqu'à ce que toutes les données d'un sous-ensemble soient classées dans une seule classe ou jusqu'à ce que les critères d'arrêt soient atteints (par exemple, la profondeur maximale de l'arbre (*maximum depth*)) (figure **1**) [24]. L'arbre de décision obtenu peut être interprété graphiquement et utilisé pour prédire la classe d'un nouvel exemple en suivant le chemin à travers l'arbre qui correspond aux caractéristiques de cet exemple.

Ensemble methods

Les méthodes ensemblistes (*ensemble methods*) constituent une technique utilisée en ML qui combine plusieurs modèles de façon qu'ils travaillent ensemble de manière plus efficace. Ces méthodes visent à améliorer la performance, la robustesse et la stabilité des modèles de prédiction. La combinaison des modèles peut se faire avec différentes techniques parmi lesquelles on peut citer le *bagging* ou le *boosting*.

Le ***bagging*** consiste à entraîner plusieurs algorithmes de manière indépendante sur des sous-ensembles de données aléatoires tirés de l'ensemble de données d'origine. Les prédictions de chaque algorithme sont ensuite combinées de manière à obtenir une prédiction finale.

Le concept derrière le ***boosting*** est de construire des modèles de façon successive, en commençant par un modèle de base et en

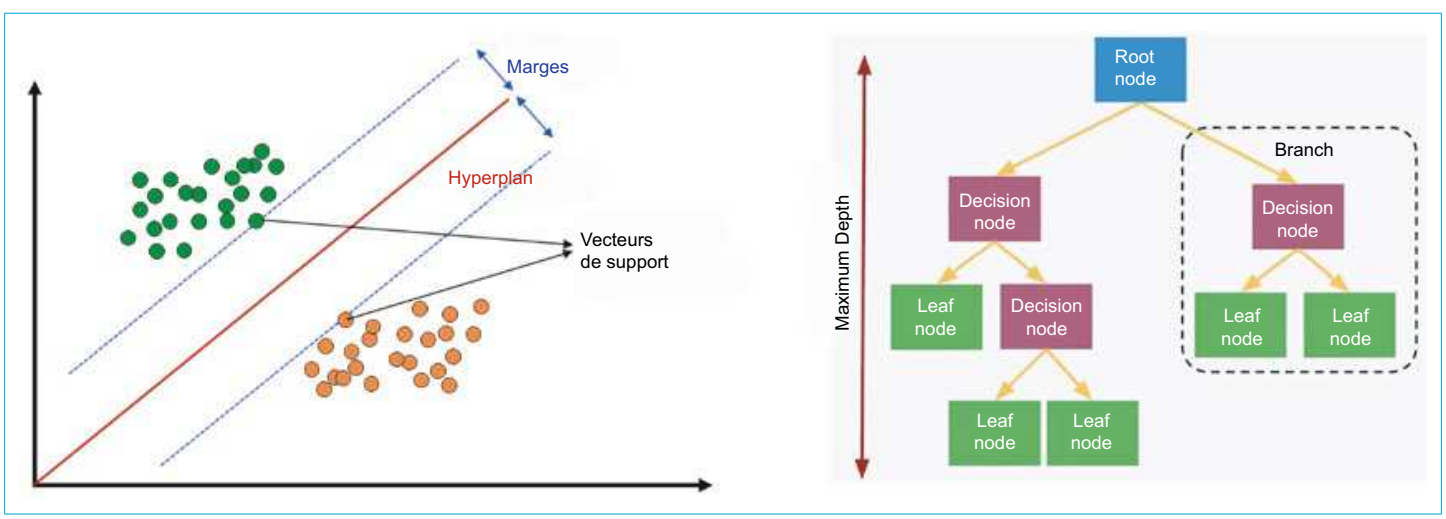

Figure 1 – Illustration du fonctionnement des algorithmes classiques SVM (à gauche, adapté de [24]) et DT (à droite, adapté de [22])

ajoutant de nouveaux modèles qui « corrigent » les erreurs commises par les modèles précédents. Chaque modèle est construit en se concentrant sur les exemples qui ont été mal prédits par les modèles précédents, ce qui permet d'améliorer progressivement la performance globale.

Random forest

L'algorithme des forêts aléatoires (RF pour *Random Forest*) est un algorithme de ML supervisé introduit en 2001 par Breiman [25]. C'est un modèle d'ensemble qui combine les prédictions de plusieurs arbres de décision pour produire des prédictions plus précises et plus robustes. Les arbres de décision individuels sont entraînés sur des sous-ensembles aléatoires des données d'entraînement et sur des sous-ensembles aléatoires des caractéristiques. Cette technique permet de réduire le surapprentissage et d'améliorer la généralisation du modèle. Les RF peuvent être utilisées pour la classification ou la régression, et sont particulièrement adaptées aux données de grande dimension et à haute cardinalité (grand nombre de caractéristiques avec de nombreuses valeurs uniques). Elles sont également résistantes aux valeurs aberrantes, ce qui en fait un choix populaire.

Extreme gradient boosting

L'algorithme XGBoost est un algorithme d'apprentissage automatique supervisé, développé en 2014 par Chen et He [4] pour améliorer les performances et la vitesse de l'algorithme *gradient boosting machine*. Des améliorations significatives ont depuis été apportées à XGBoost, telles que la parallélisation et la distribution des calculs sur plusieurs nœuds, ce qui a permis une accélération significative des temps d'entraînement. De plus, des algorithmes de régularisation ont été introduits pour prévenir le surapprentissage, d'autre optimisant le choix des sous-échantillons pour une meilleure utilisation de la mémoire.

1.2 État de l'art de l'apprentissage automatique en géotechnique

1.2.1 Panorama général en géotechnique

Une requête sur la base de données des publications de recherche *Scopus* montre une progression très rapide de l'utilisation des méthodes de ML en géotechnique, avec notamment une accélération nette après 2015 (figure **2**).

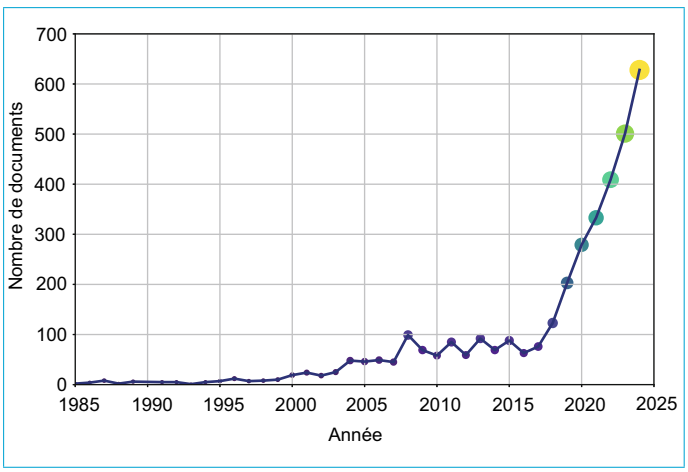

Figure 2 – Évolution du nombre de documents indexés dans la base Scopus au 11 novembre 2024 : ((TITLE-ABS-KEY ((« soil mechanic* » OR geotechnic* OR geomechanic*) AND (« machine learning » OR « artificial intelligence » OR « AI » OR « deep learning » OR « neural network* » OR « big data »))))

Plus précisément, Baghbani *et al.* [18], à travers l'étude de 1 235 publications, soulignent les thématiques qui se sont le plus intéressées à l'IA : la mécanique des roches, les glissements de terrain, la liquéfaction des sols ou encore les tunnels et tunneliers (figure **3**). La revue de la littérature révèle que les réseaux de neurones artificiels (ANN) constituent l'approche la plus utilisée en géotechnique, représentant une part d'environ 50 % des publications concernées [16] [18].

Dans le cadre d'une approche multidisciplinaire faisant appel aux outils de ML en soutien du métier d'ingénieur géotechnique, il convient de se poser au préalable quatre questions clefs :
1/ Quelles sont les données à disposition ?
2/ Le modèle à développer sera-t-il fondé sur l'apprentissage de la donnée, qu'elle soit brute ou synthétique ?
3/ Quelle sera l'utilité du modèle de substitution développé par apprentissage automatique ?
4/ L'apprentissage mis en œuvre est-il fondé sur la physique ou informé par la physique ?

Données géotechniques

L'auscultation des chantiers de projets géotechniques génère des données à caractère spatio-temporel. L'exploitation de ces données représente un atout majeur pour réduire les incertitudes, optimiser la conception et l'auscultation des ouvrages, ainsi que leur maintenance à long terme. La collecte, le traitement et l'analyse de ces données constituent cependant un défi en soi. En effet, les données provenant des chantiers doivent toujours faire l'objet d'un traitement préalable pour écarter les erreurs de mesure, les valeurs dupliquées et les valeurs aberrantes. Un traitement manuel au cas par cas est parfois indispensable pour traiter les anomalies les plus importantes. Ces données, communément qualifiées de « sales » (*noisy data*), sont des données inexactes, incomplètes ou incohérentes qui nécessitent un nettoyage avant leur utilisation pour des analyses ultérieures. Le nettoyage des données améliore donc la qualité du jeu de données (*dataset*) ce qui renforce la confiance accordée à l'ensemble des données et la fiabilité des analyses réalisées sur celles-ci [26] [27] [28].

Se pose ensuite la question du stockage des données dans un format facilement exploitable en vue de différentes utilisations. En pratique, les bases de données relationnelles permettent de stocker efficacement les données géologiques et géotechniques, ainsi que les données d'auscultation. Ces systèmes de stockage garantissent l'intégrité et le maintien des relations (spatiales et temporelles) entre les nombreux jeux de données qu'ils permettent d'agréger [29] [30] [31].

Apprentissage fondé sur les données

Les modèles issus du ML reposent, par construction, sur les données qui ont servi à l'apprentissage. Autrement dit, ces modèles sont créés à partir de données. Par conséquent, ils ne peuvent être utiles que dans la mesure où il existe des ensembles de données spécifiques au problème considéré. Ces données peuvent être de deux types : des données « réelles », c'est-à-dire issues de la caractérisation et de l'auscultation d'un ou plusieurs projets déjà réalisés ou en cours de réalisation, tel qu'expliqué précédemment, ou des données synthétiques, la plupart du temps générées de façon procédurale grâce au recours à un modèle numérique considéré comme fiable et recalé lui-même sur des mesures.

Comme l'expliquent Guayacán-Carrillo *et al.* [32] [33], les données synthétiques sont notamment à prendre en compte dans deux cas :
1/ quand les données existantes ne sont pas suffisantes : des données artificielles peuvent alors être générées en utilisant les informations déjà collectées ainsi que des modèles simplifiés initiaux ; par exemple, dans le contexte d'études sur la surveillance de la stabilité de pentes naturelles et artificielles, en observant les déplacements de surface, une approche pour créer des données a été établie en considérant différentes configurations (comme la géométrie du massif en mouvement et les propriétés des matériaux)

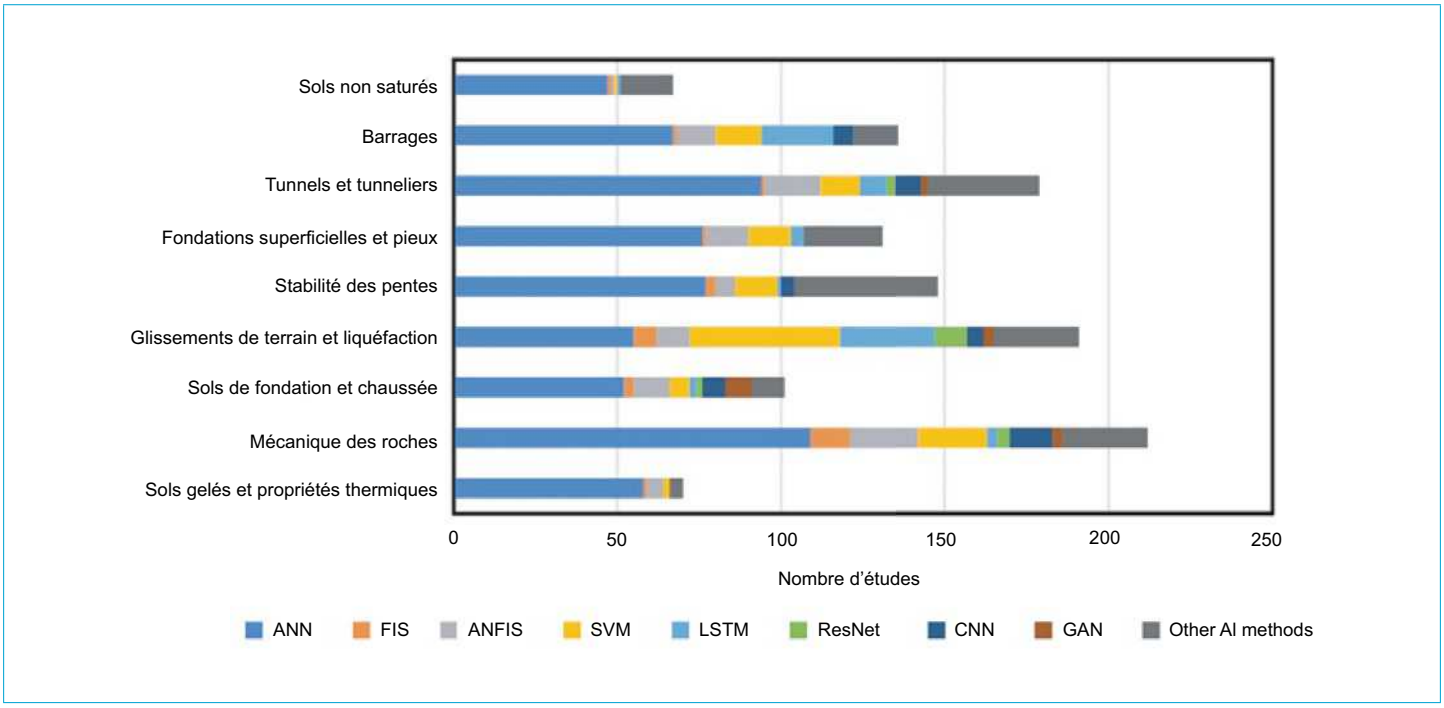

Figure 3 – Répartition du corpus géotechnique selon la thématique étudiée et l'approche d'IA considérée (Baghbani *et al.*, 2022). Pour les sigles, se référer au § 5

ainsi que divers scénarios climatiques, dans le but de produire une large distribution de trajectoires d'évolution de la vitesse du mouvement de terrain [34] ;

2/ quand il n'est pas possible d'accéder aux données du site, par exemple lors de la phase de conception d'une structure ; la génération de données permettra alors de considérer une vaste gamme de paramètres (incluant les conditions géologiques et les caractéristiques de l'ouvrage) pour effectuer des analyses préliminaires en vue de la conception ; un exemple d'application pour l'excavation de tunnels profonds est présenté par Tristani *et al.* [35].

Modèles de substitution fondés sur l'apprentissage automatique

Le recours aux modèles de substitution (*surrogate models*) peut s'avérer utile dans certaines situations, particulièrement quand des simulations numériques en 3D sont requises. Les modèles de substitution reposent sur des modèles d'intelligence artificielle entraînés à partir de données synthétiques. L'essence-même de ce type d'approche réside dans la synergie entre une modélisation numérique sophistiquée et l'apprentissage automatique pour offrir aux ingénieurs des outils faciles d'utilisation et fiables, réduisant ainsi les efforts et le coût associés à des calculs lourds.

Ces modèles sont de plus en plus exploités dans le secteur de la géotechnique car ils sont développés à partir de données maîtrisées par des experts du domaine et constituent une alternative prometteuse pour les analyses statistiques requérant de nombreuses simulations. Des exemples d'application incluent des projets liés à la mécanique des roches et des sols [36], la conception de tunnels [35] ainsi que le dimensionnement thermique de pieux énergétiques [37].

Apprentissage automatique fondé sur la physique et informé par celle-ci

Dans le contexte de la conception des ouvrages, il est crucial pour l'ingénieur de maîtriser l'outil sur lequel il s'appuie pour la prise de décisions. Tout outil constitue de fait une aide à la décision, mais la responsabilité de la décision finale incombe à l'ingénieur. Dans cette perspective, trois cas de figure peuvent être distingués.

- L'apprentissage fondé sur la physique constitue une stratégie pour améliorer sa qualité et son potentiel de généralisation. L'apprentissage fondé sur des données synthétiques générées par des modèles déterministes rigoureux garantit par construction cette base physique. Dans le cas où des données « réelles » sont considérées, la base physique peut consister à prétraiter ces données destinées à être introduites dans l'algorithme pour l'entraînement. Le domaine de la géotechnique a réalisé des progrès significatifs dans la prédiction des phénomènes observés sur le terrain grâce à des méthodes empiriques mais aussi analytiques et numériques. Utiliser ces méthodes pour fournir des données propres à l'algorithme peut alors être très avantageux en vue de la généralisation du modèle. Un exemple concret est illustré en figure **4** : les données mesurées concernent la progression du tassement en surface longitudinal et transversal provoqué par le creusement d'un tunnel. Les mesures brutes sont prétraitées à l'aide d'un modèle de calage fondé sur une formulation empirique de cuvette de tassement. Poussés par des considérations pratiques d'ingénierie (comprendre et simplifier le problème) et par la mise en œuvre efficace de modèles de ML (maîtriser les entrants en constituant des paramètres d'entrée dénués de bruit), des modèles ont été entraînés sur ces données issues de calages (cf. § 3).
- L'apprentissage informé par la physique : selon Karniadakis *et al.* [38], le ML informé par la physique intègre de manière transparente les données et des modèles physiques, y compris dans des environnements partiellement compris, incertains et de grande dimensionnalité. Une méthode largement adoptée dans la littérature consiste en l'utilisation de réseaux neuronaux informés par la physique, également appelés PINN (*Physics Informed Neural Networks*). Cette approche suscite également de l'intérêt dans le domaine de la géotechnique. À titre d'exemple, dans une étude récente sur l'analyse des glissements de terrain catastrophiques, la méthode a été

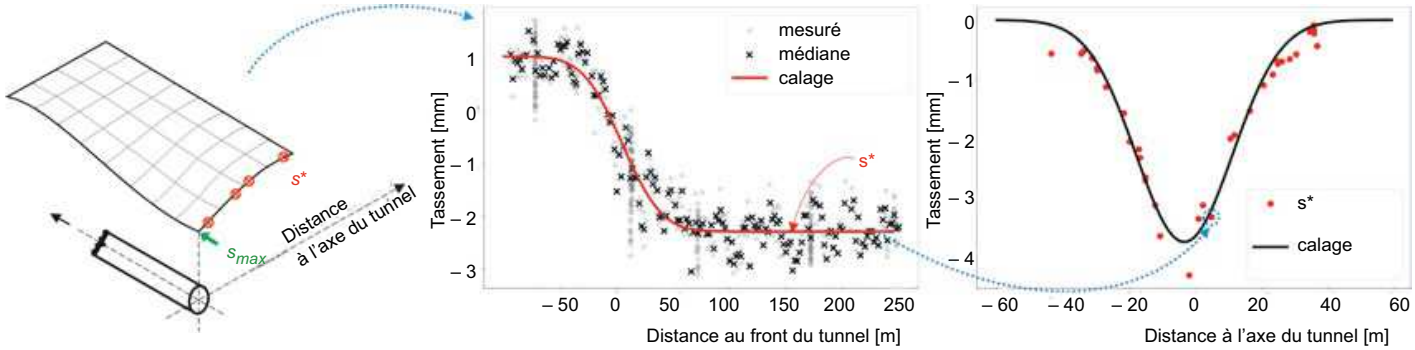

Figure 4 – Calage des tassements longitudinal et transversal

mise en œuvre par Moeinnedin *et al.* [39]. Leur approche se caractérise par deux ingrédients cruciaux pour résoudre une équation aux dérivées partielles (EDP) caractérisant le phénomène physique sous-jacent au problème étudié. Le premier est une fonction paramétrée, à la fois facile à évaluer et suffisamment robuste pour approximer la solution. Le second ingrédient essentiel est une fonction de coût ou de perte. Lorsqu'elle est minimisée, cette fonction de coût garantit que la fonction paramétrée représente une approximation satisfaisante de l'EDP à résoudre. D'autres méthodes se concentrent sur la création de modèles de comportement issus d'un apprentissage automatique fondé sur la thermodynamique, offrant l'avantage d'être robustes face au bruit dans les données d'apprentissage et de meilleures capacités d'extrapolation [40].

- Vers des modèles compréhensibles et interprétables ? La possibilité d'interpréter et d'expliquer les outils de ML suscite un intérêt croissant dans différents domaines [41]. Certaines études récentes ont été proposées dans le domaine de la science des matériaux [42], ainsi que dans le cadre de l'excavation des tunnels [43] [32]. Ce champ de recherche est prometteur mais encore pleinement ouvert.

1.2.2 Prévision des tassements induits par le creusement de tunnels

L'utilisation des algorithmes de ML semble être une méthode particulièrement adaptée et potentiellement précise pour estimer le tassement induit par le creusement des tunnels, grâce à leur capacité à identifier les meilleures relations entre les différentes données d'entrées [44].

Depuis la première tentative de prévision des tassements induits par le creusement des tunnels à l'aide d'algorithmes de ML [45], plusieurs études ont été publiées au fil des années, comme le montre la figure **5**. Cette figure met en évidence que ce sujet est particulièrement d'actualité [46] [47] [48] [49] [50] [51] [52].

Liu *et al.* [53] ont mené une étude sur l'efficacité de la prévision des tassements à l'aide d'algorithmes de ML sur des ensembles de données de petite taille. Il convient de noter que leur ensemble de données contient 187 valeurs. Selon leurs résultats, les algorithmes de ML, notamment ceux de type ensembliste (*ensemble methods*), peuvent être utilisés avec succès pour prédire les tassements, même lorsque la taille de l'ensemble de données disponible est petite. Des résultats similaires ont été obtenus récemment par Richa *et al.* [54] dans le cadre d'une étude sur l'analyse de la précision des techniques d'apprentissage automatique pour la prévision des tassements en surface induits par le creusement de tunnels et réalisée à partir d'ensembles de données de taille limitée.

À retenir

- L'essor de la science des données et de l'intelligence artificielle confère aux approches interdisciplinaires, particulièrement dans le domaine de la géotechnique, une vaste étendue de recherche et constitue un progrès notable dans ce domaine.
- L'analyse rigoureuse des données géotechniques, notamment la compréhension de leur caractère spatio-temporel, constitue un élément fondamental pour réduire les incertitudes, optimiser la conception et le suivi des structures, ainsi que garantir leur maintenance à long terme.
- Les méthodes d'apprentissage automatique diffèrent par leur généralisation, leur utilisation de données et leur méthode d'apprentissage. En géotechnique, les algorithmes qui apprennent de façon incrémentale à partir d'un flux continu de données suscitent un intérêt croissant.
- Le modèle d'apprentissage automatique peut être conçu en utilisant des données brutes ou synthétiques, et/ou en se basant sur des principes physiques. Le choix dépendra de la nature des données et de l'objectif de l'application.
- Les algorithmes d'apprentissage automatique montrent un potentiel significatif pour l'évaluation du tassement induit par le creusement de tunnels, en raison de leur aptitude à discerner les relations sous-jacentes dans des ensembles de données variées.

2. Cadrage et préparation des données

2.1 Principe

La phase de cadrage et de préparation des données est un préalable indispensable à la construction d'un modèle de ML performant. Dans une branche aussi spécifique de l'ingénierie que la géotechnique, il est crucial de bien définir le périmètre du projet et de préparer minutieusement les données avant d'entamer l'entraînement des modèles. Cette étape vise à structurer les données issues de multiples sources, à résoudre les incohérences et à garantir leur adéquation avec les spécifications des algorithmes.

Le processus de cadrage inclut l'identification des variables clefs et la conception d'une architecture de données qui facilitera leur exploitation. Parallèlement, une préparation minutieuse des

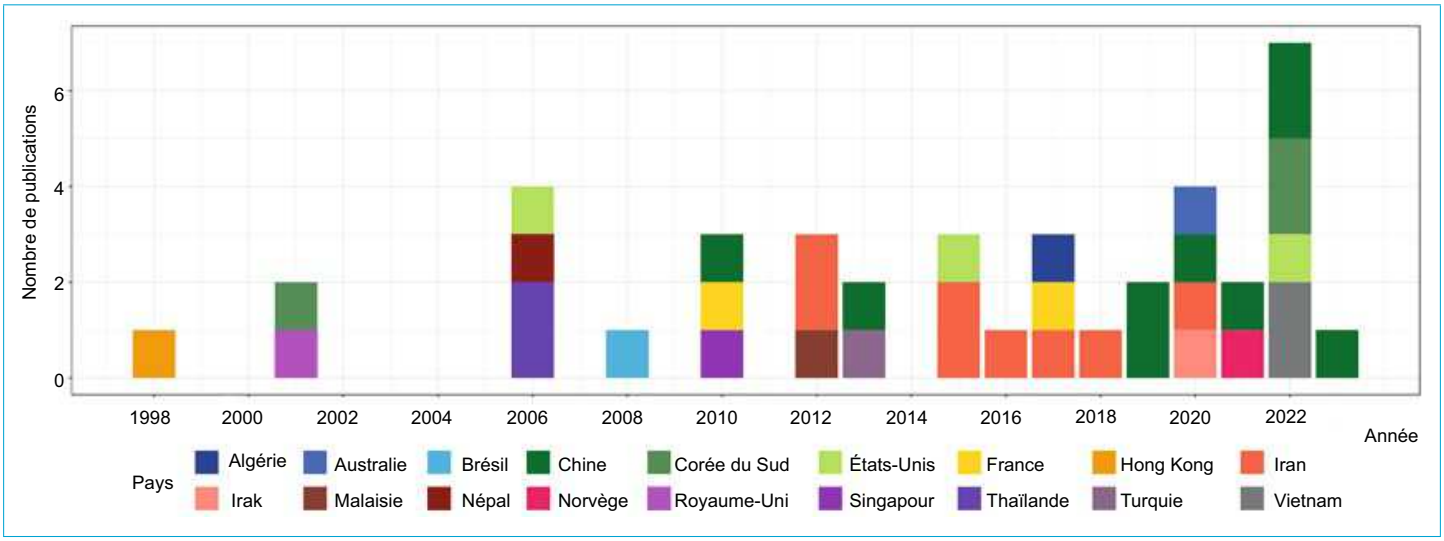

Figure 5 – Évolution temporelle du nombre de publications traitant de la prévision des tassements causés par le creusement des tunnels à l'aide d'algorithmes d'apprentissage automatique et répartition par pays d'étude (double compte en cas de publications en collaboration internationale)

données est nécessaire pour s'assurer qu'elles sont propres et prêtes à être utilisées efficacement par les modèles de ML. En réalisant correctement ces étapes, on maximise les chances de succès de l'algorithme et on s'assure que les résultats seront robustes et exploitables dans le contexte spécifique du projet. Cette section détaille ces deux aspects essentiels, le cadrage du problème et la préparation des données, qui conditionnent la réussite de toute démarche de ML en géotechnique.

Encadré 1 – Exemples pour illustrer les concepts

Dans la suite de cet article, chaque concept abordé sera illustré par des exemples concrets liés à la prédiction des tassements induits par le creusement d'un tunnel au tunnelier. Ces exemples montreront comment les méthodologies de ML peuvent être appliquées pour résoudre des problématiques spécifiques en géotechnique, notamment la prévision des tassements maximaux à l'axe du tunnel.

2.2 Vue d'ensemble

2.2.1 Cadrage du problème

Avant de plonger dans le développement d'un modèle de ML, il est essentiel de bien cadrer le projet et de définir clairement les objectifs et les méthodes. Dans le domaine de la géotechnique, où les défis sont nombreux et les variables multiples, un cadre solide est crucial pour s'assurer que les données collectées seront adaptées aux besoins du modèle et permettront d'atteindre les résultats escomptés.

La première étape clef dans un projet de ML appliqué à la géotechnique est de définir clairement le contexte. Cette étape consiste à situer le projet dans son environnement et à identifier les défis à anticiper ainsi que les facteurs géologiques, techniques et opérationnels qui influencent le comportement des massifs de sols considérés.

Encadré 2 – Contexte

Le Grand Paris Express (GPE) est un projet de construction de 200 km de nouvelles lignes de métro et d'extensions de lignes existantes autour de la ville de Paris. Ce projet se déroule dans un contexte géologique complexe, incluant une grande variété de types de sols, allant des marnes et caillasses aux calcaires grossiers, en passant par des argiles plastiques (figure **6**). Cette diversité géologique détermine le comportement du sol lors du creusement des tunnels et pose des défis particuliers en matière de prévision des tassements induits. Dans ce qui suit, on s'intéresse aux lignes 14 sud et 15 sud-ouest du GPE.

La méthode de creusement employée dans ces deux lignes est celle du tunnelier à pression de terre (EPB pour *earth pressure balance*), qui assure la stabilité du front d'excavation en maintenant une pression équilibrée tout au long du processus d'excavation. Particulièrement adaptée aux environnements urbains denses, cette méthode permet de minimiser les risques de déformation du sol et de réduire les impacts sur les infrastructures avoisinantes.

La mise en place d'un modèle de ML en géotechnique nécessite d'abord un cadrage rigoureux du problème. Cette étape préliminaire vise à clarifier les objectifs spécifiques et les attentes vis-à-vis du modèle, en tenant compte des particularités du phénomène physique étudié. Différentes approches peuvent être adoptées selon la problématique, influençant directement la préparation des données et le choix des algorithmes. Ce cadrage initial, effectué en collaboration avec les experts métier, garantit que le modèle répondra bien aux besoins pratiques du projet.

Encadré 3 – Définition de la problématique

Dans le cas de la prévision des tassements, la problématique pourrait porter soit sur la prédiction du tassement maximal à l'axe du tunnel (S_{max}), soit sur l'évolution des tassements au fur et à mesure du creusement. La seconde approche nécessiterait de travailler avec des séries temporelles, impliquant des algorithmes spécifiques. Dans cet article, nous avons choisi de nous concentrer sur la prévision de S_{max}.

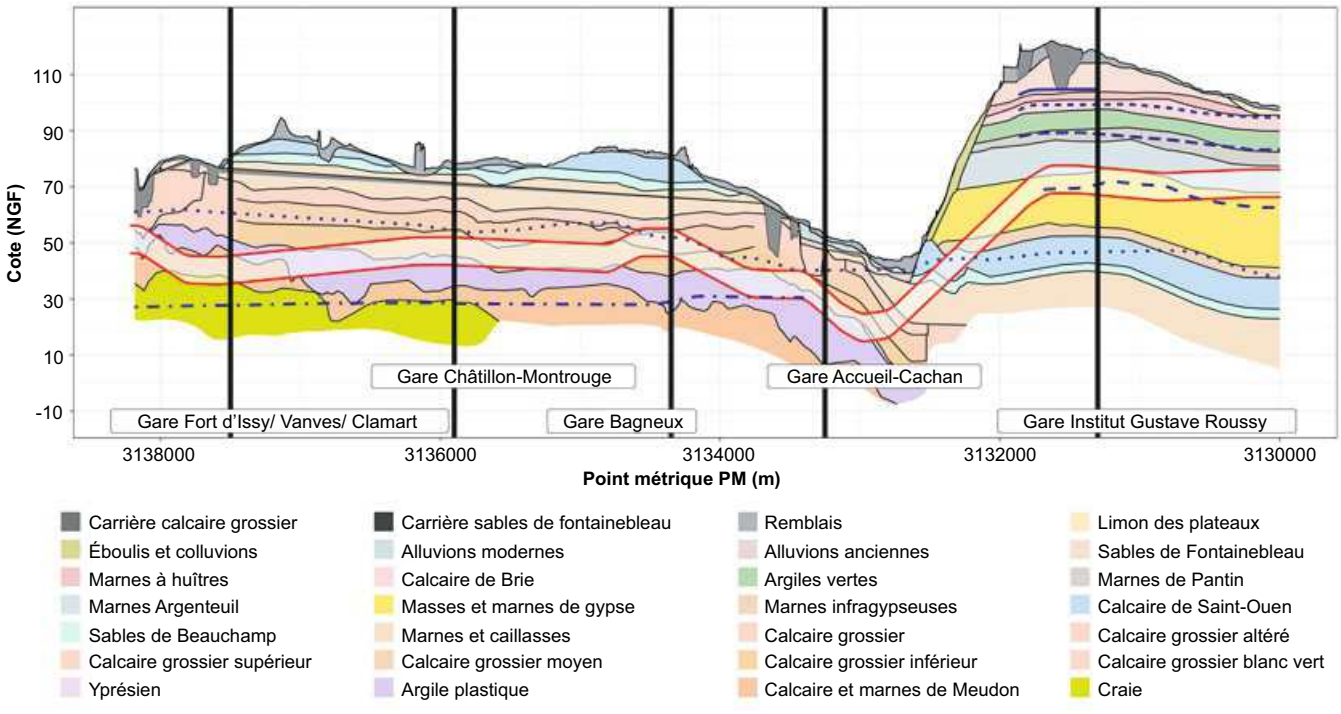

Figure 6 – Profil en long de la ligne 15 sud-ouest du GPE

Le cadrage doit également inclure une analyse des solutions existantes pour des problématiques similaires. Cette analyse permet d'identifier les limites des approches traditionnelles et de souligner l'apport potentiel des algorithmes de ML.

Encadré 4 – Limites des méthodes traditionnelles

Pour la prévision du S_{max}, les méthodes empiriques fournissent uniquement des approximations, tandis que les modèles numériques (comme ceux s'appuyant sur la méthode des éléments finis) nécessitent de nombreux paramètres pour représenter les lois de comportement des sols, en plus d'un temps de calcul conséquent. Ces méthodes ne permettent pas de réaliser des prédictions précises à chaque mètre de creusement, un besoin pourtant essentiel pour la gestion des risques en temps réel sur les chantiers. L'intelligence artificielle offre ici un avantage décisif : elle permet d'obtenir des prévisions en temps réel, à chaque mètre de creusement, tout en intégrant progressivement de nouvelles données pour un apprentissage continu du modèle.

Une fois la problématique définie, il est nécessaire de sélectionner les paramètres clefs qui influencent le phénomène étudié : ce sont les caractéristiques (*features*). Les données associées à ces caractéristiques doivent être disponibles et organisées dans une base de données qui servira de fondement pour construire le modèle. À ce stade, il est possible de faire une première sélection des caractéristiques en s'appuyant sur l'expertise métier, les corrélations établies entre les variables, les formules empiriques, les retours d'expérience des modèles numériques ou encore l'état de l'art. La sélection définitive des caractéristiques se fera plus tard, après une analyse approfondie des données.

Encadré 5 – Choix des caractéristiques préliminaires

Les tassements induits en surface par le creusement des tunnels dépendent de trois catégories de facteurs : la géométrie du problème (profondeur et diamètre de creusement), les paramètres de pilotage du tunnelier et les paramètres géotechniques, comprenant la lithostratigraphie et les propriétés mécaniques des sols. Le diamètre du tunnel n'est pas retenu, car il est unique pour l'ensemble des données.

Il est alors possible de choisir les caractéristiques préliminaires suivantes : la couverture au-dessus de l'axe du tunnel pour la géométrie (C), la vitesse d'avancement du tunnelier ($V_{tunnelier}$), la pression au front (P_{front}), le couple de la roue de coupe (M_{RDC}), la poussée totale (P_{totale}) et la pression d'injection du mortier à l'arrière de la jupe ($P_{mortier}$) ainsi que le volume injecté ($V_{mortier}$) pour les paramètres de pilotage du tunnelier et enfin le poids volumique humide (γ), le module de déformation de l'essai pressiométrique (E_M), le coefficient de pression des terres au repos (K_0), la cohésion (c) et l'angle de frottement (φ) pour les propriétés géotechniques des sols. Ces caractéristiques serviront de base pour les analyses statistiques et la sélection des variables les plus pertinentes pour la prédiction des tassements.

2.2.2 Conception du modèle

Les étapes suivantes du processus consistent à concevoir le modèle, ce qui implique des décisions stratégiques importantes, telles que le choix du mode d'apprentissage (supervisé, non supervisé, par renforcement, etc.), le mode de consommation des données, et les algorithmes de ML à tester. Ces choix dépendent de plusieurs critères, notamment la nature des données, la tâche à accomplir (régression ou classification), et les ressources disponibles en matière de capacité de calcul et de temps d'entraînement. Il est généralement conseillé de commencer par des algorithmes simples et interprétables avant d'explorer des modèles plus complexes. En géotechnique, les algorithmes supervisés, comme les forêts aléatoires (RF pour *random forest*), sont souvent utilisés, car ils traitent efficacement les données étiquetées (labellisées) issues des mesures d'auscultation de chantier et sont capables de généraliser, c'est-à-dire qu'ils ont la capacité de fournir de bonnes prédictions à partir de données sur lesquelles ils n'ont pas été entraînés (cf. § 3.1), à partir de jeux de données relativement limités.

Ensuite, des métriques de performance doivent être sélectionnées pour évaluer le modèle obtenu, telles que l'erreur quadratique moyenne ou le coefficient de détermination R^2.

Encadre 6 – Conception pour une prédiction des tassements

La conception d'un modèle de ML pour la prédiction des tassements nécessite de définir un mode d'apprentissage adapté et de sélectionner des algorithmes performants. Étant donné que les données de tassements sont disponibles en grande quantité, un mode d'apprentissage supervisé est utilisé. Dans un environnement dynamique tel que le creusement de tunnels, un flux constant de nouvelles données est généré. Ainsi, le modèle doit être capable de consommer ces informations de manière incrémentale, en les intégrant au fur et à mesure de leur arrivée. Il est donc essentiel de choisir des algorithmes de ML qui permettent un réentraînement très rapide, afin de maintenir la précision des prédictions. Des algorithmes à base d'arbres de décision, tels que les *decision trees* (DT), *random forest* (RF) ou *extreme gradient boosting* (XGBoost), ainsi que les *support vector machine* (SVM), sont particulièrement adaptés, car ils permettent de traiter efficacement de nouvelles données sans nécessiter des temps de calcul trop longs. Cela garantit des prévisions rapides, cruciales pendant le processus de creusement. Pour évaluer la performance du modèle, des métriques telles que le R^2 et l'erreur quadratique moyenne (*root mean square error*, RMSE) seront utilisées, offrant ainsi une mesure précise de l'ajustement et de l'erreur des prédictions.

2.3 Préparation des données

2.3.1 Extraction et structuration

Une fois les grandes décisions de conception du modèle de ML prises, il convient de passer à la collecte et à la préparation des données. Cette étape de préparation est d'autant plus nécessaire dans le domaine de la géotechnique, où les données brutes peuvent être hétérogènes et complexes. Cette phase est cruciale car la qualité, la quantité et la structure des données influencent directement les performances des modèles. Comme le souligne la célèbre expression « garbage in, garbage out », un jeu de données mal préparé peut induire des biais, provoquer une instabilité dans les résultats et, dans certains cas, empêcher la convergence des algorithmes de ML. Avant d'entraîner un algorithme, il est donc essentiel de structurer, nettoyer et transformer les données pour les rendre exploitables et adaptées aux exigences des algorithmes de ML.

Les projets de ML, notamment en géotechnique, traitent généralement plusieurs catégories de données provenant de diverses sources. Ces données peuvent être géotechniques, relatives à la caractérisation du terrain, ou bien issues de la technique de construction ou encore des détails sur les arrêts du front de taille ou incidents éventuels (tels que des venues d'eau).

Les données proviennent souvent de plateformes variées, en fonction de leur origine (entreprises de construction, laboratoires, ou encore bureaux d'études). Une gestion efficace rendant ces données facilement accessibles est donc essentielle pour leur exploitation dans des projets fondés sur l'intelligence artificielle. Cette accessibilité doit être accompagnée d'une clarification des droits de propriété sur les données dès le début du projet.

Pour faciliter l'utilisation des données, il est important de pouvoir y accéder directement, que ce soit via des API (*Application Programming Interface*, qui permettent notamment l'accès automatisé aux données de l'application) ou des bases de données. Les données sont souvent disponibles dans des formats variés et peuvent être non structurées (par exemple, des images ou des fichiers textes non organisés) ou être hébergées sur des plateformes où l'extraction en masse est complexe. Un partage efficace et bien défini des données entre les différentes parties prenantes du projet permet de surmonter ces obstacles et de réduire considérablement le temps de préparation.

Encadré 7 – Extraction des données

Pour la prévision des tassements, les données appartiennent à trois catégories : les mesures de tassements, les paramètres de pilotage du tunnelier et les données géotechniques.

Dans le cadre des travaux de Richa [55], les paramètres relatifs au pilotage du tunnelier sont mis à disposition sous forme de captures d'écran régulières des informations affichées dans la cabine de pilotage, moyennées par anneau et déposées sur une plateforme web dédiée. Dans ce contexte, des scripts, ici en langage R, ont été développés pour extraire automatiquement et massivement des informations ciblées depuis ces images. Des techniques de traitement d'image ont permis au préalable d'améliorer la résolution des caractères avant de les transformer en texte grâce à des librairies de reconnaissances de caractères (OCR pour *Optical Character Recognition*) (figure **7**).

En ce qui concerne les mesures de tassements, un système d'auscultation automatisé a été mis en place pendant toute la durée des travaux. Ce système utilise une série de théodolites automatiques qui enregistrent les déplacements du sol et des bâtiments avec une fréquence de mesure élevée. Les mires de surface sont visées par les théodolites toutes les 30 minutes environ pour mesurer les déplacements dans les trois directions de l'espace. Environ 16 000 capteurs de différents types ont été placés en voirie ou sur des bâtiments, de part et d'autre de l'axe du tunnel, afin de couvrir la zone d'influence géotechnique (ZIG). Les mesures enregistrées par les théodolites sont stockées en temps réel dans une base de données, qui peut être consultée à distance par les différents acteurs du projet (entreprise, maître d'œuvre, maître d'ouvrage) via un accès sécurisé à travers un site web ou une application sous forme de plateforme interactive (par exemple, Geoscope de l'entreprise Sixense pour le projet objet de la thèse citée). L'exportation massive des données de cette base a produit plus de 144 millions de mesures associées à près de 16 000 capteurs de tous types.

Une des étapes clefs de la préparation des données consiste en leur structuration. Une organisation claire et cohérente est indispensable pour garantir la traçabilité et l'exploitabilité des informa-

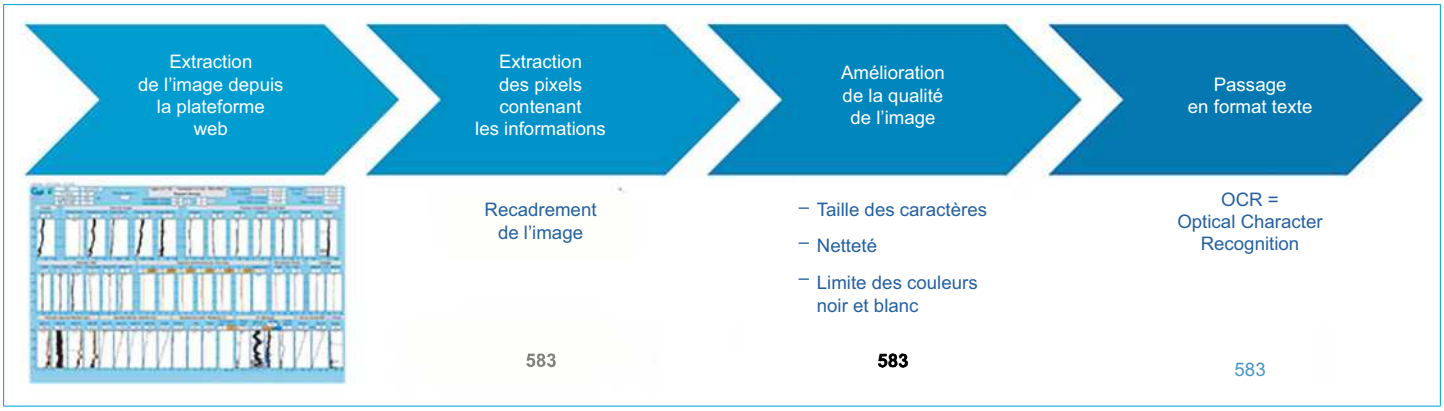

Figure 7 – Méthode d'extraction des paramètres de pilotage de tunnelier à partir d'images

tions par les algorithmes de ML. Pour ce faire, il est fortement recommandé d'utiliser des bases de données relationnelles. Ces bases reposent sur un système qui organise les données en tables interconnectées, chaque table étant dédiée à une catégorie spécifique d'information (par exemple, une table pour les données géotechniques, une autre pour les données de construction). Cette organisation permet de mieux gérer la complexité des données en les séparant par thème tout en facilitant leur intégration et mise à jour. En outre, les bases relationnelles sont évolutives, ce qui permet d'ajouter de nouvelles informations au fur et à mesure de l'avancement du projet, sans perturber la structure existante.

Encadré 8 – Constitution d'une base de données

Comme défini précédemment, pour la prédiction des tassements, les informations sont généralement réparties en quatre grandes catégories : la géométrie du problème et les coordonnées spatiales, les paramètres de pilotage du tunnelier, les paramètres géotechniques et enfin les mesures de tassement.

L'architecture de la base de données est construite selon une logique de catégorisation relationnelle, où chaque catégorie est reliée aux autres par des clefs primaires et secondaires (figure **8**).

Les clefs primaires (PK pour *Primary Keys*) sont des identifiants uniques permettant de différencier chaque enregistrement au sein d'une table. Elles garantissent que chaque entrée est unique et facilement identifiable. Par exemple, chaque anneau dans la table « Anneau » peut être identifié de manière unique par un ID spécifique, assurant ainsi l'intégrité des données au sein des tables.

Les clefs secondaires (ou clefs étrangères, FK, pour *Foreign Keys*) sont utilisées pour établir des relations entre les tables. Elles font référence à la clef primaire d'une autre table, permettant ainsi de connecter les informations. Par exemple, une clef secondaire dans la table « Mesure_Capteur » pourrait faire référence à la clef primaire de la table « Capteur », qui contient les informations sur les capteurs. Cette connexion permet de relier chaque mesure au capteur qui l'a enregistrée.

La figure **8** montre chaque catégorie, colorée différemment, et les liens entre elles par des flèches représentant ces relations. Les clefs primaires assurent l'unicité des informations, tandis que les clefs secondaires garantissent la cohérence et la structure de l'ensemble de la base de données.

Cette organisation relationnelle permet non seulement une identification claire et une gestion structurée des différentes données, mais aussi un accès optimisé aux informations. Elle facilite leur mise à jour et leur exploitation pour le modèle de prédiction, garantissant ainsi une traçabilité, une cohérence, et une intégrité des données à chaque étape du processus.

2.3.2 Nettoyage et analyses exploratoires

Dans le domaine de la construction, les chantiers, lorsqu'ils sont suivis de façon systématique, génèrent une grande quantité de données brutes, données qui constituent une mine d'informations précieuses pour réaliser des recalages ou évaluer des pistes d'améliorations futures. Ces données, bien que potentiellement très riches et utiles pour les algorithmes de ML, nécessitent un travail approfondi de traitement avant de pouvoir être exploitées efficacement. En effet, les données brutes en elles-mêmes nécessitent d'être traitées pour pouvoir produire de la valeur : elles sont fréquemment bruitées, hétérogènes ou incomplètes en raison de diverses causes, telles que des erreurs de mesure, des conditions météorologiques changeantes ou encore la diversité des méthodes de collecte utilisées sur le chantier.

Le nettoyage des données est donc une étape essentielle dans le processus de préparation. Cette étape implique la détection et la correction de valeurs manquantes, la gestion des données aberrantes (*outliers*), ainsi que la résolution des incohérences qui peuvent survenir lors de la collecte. Un nettoyage rigoureux est crucial pour garantir la fiabilité des données, améliorer la précision des modèles de ML et assurer des résultats plus robustes. Ce travail minutieux permet de rendre les données prêtes à être utilisées dans les phases suivantes du développement du modèle prédictif.

Encadré 9 – Nettoyage des mesures de tassements

Prenons l'exemple du nettoyage des mesures de tassements pour illustrer le processus de suppression des données aberrantes.

Nettoyage des capteurs

Avant de procéder au nettoyage des mesures de tassement, un premier nettoyage a été effectué sur les capteurs à l'aide d'une série de filtres. Voici quelques exemples :

- filtre sur les types de capteurs à retenir : seuls les capteurs placés sur les voiries ou les bâtiments (en bas ou en haut) ainsi que les cibles virtuelles ont été injectées dans la base de données ;
- filtre sur la distance des capteurs à l'axe du tunnel : les capteurs à une distance supérieure ou égale à 7 fois le diamètre du tunnel de 10 m sont supprimés ;
- filtre pour supprimer les capteurs qui n'ont pas suffisamment de mesures pour permettre de caler l'équation de progression du tassement ; pour réussir le calage, après une série de tests, nous avons choisi de retenir le seuil d'au moins 10 mesures avant le passage du front et d'au moins 10 mesures entre 30 m et 250 m après le passage du front.

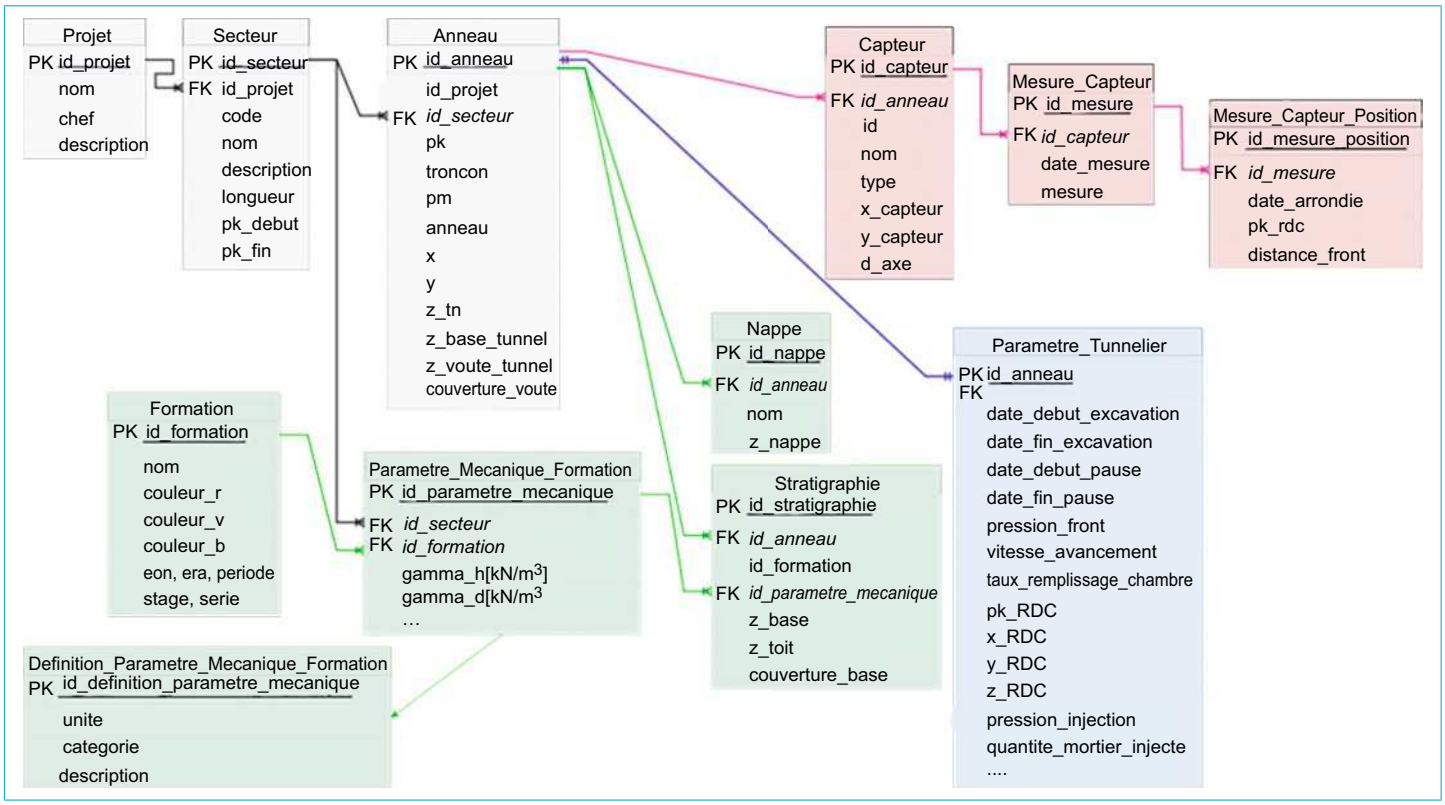

Figure 8 – Diagramme d'entité-relation (ERD pour *Entity Relationship Diagram*) de la base de données

Encadré 9 – Nettoyage des mesures de tassements (*suite*)

À ce stade, on obtient 8 706 685 observations issues de 6 799 capteurs.

Nettoyage des mesures des capteurs

Ensuite, deux méthodes ont été appliquées pour traiter les mesures de tassement.

La première méthode repose sur l'application d'un filtre simple visant à éliminer les valeurs les plus atypiques. Ce filtre consiste à calculer la moyenne des mesures de tassement d'un capteur (S_{moy}) et à supprimer toute valeur qui n'appartient pas à l'intervalle $[S_{moy} - 30 ; S_{moy} + 30]$. Le choix de la valeur de 30 mm repose sur l'analyse des tassements maximaux observés sur les lignes étudiées, qui ne dépassent jamais les 20 mm. Ainsi, les valeurs qui s'écartent trop de cette plage sont considérées comme aberrantes. Cette première méthode a permis de supprimer un total de 154 mesures.

La deuxième méthode utilisée est l'algorithme des forêts d'isolation (IF pour *Isolation Forest*), un algorithme de ML particulièrement adapté à l'apprentissage non supervisé et à la détection d'anomalies. En prenant en entrée les données relatives à la distance au front et aux mesures de tassement, l'algorithme attribue à chaque mesure une note indiquant sa probabilité d'être typique ou aberrante (figure **9**). Cette méthode a permis d'éliminer un total de 175 856 mesures aberrantes. Il est important de noter que l'application du premier filtre simple est nécessaire avant d'utiliser les forêts d'isolation, car cet algorithme fonctionne par détection de groupements de données. Ainsi, un groupe de mesures aberrantes qui ne présentent pas de dispersion suffisante ne serait pas détecté par IF.

Encadré 9 – Nettoyage des mesures de tassements (*suite*)

Calage pour obtenir S_{max}

Pour déterminer les valeurs de S_{max}, un travail de calage a été réalisé à partir des équations empiriques connues pour les tassements longitudinal et transversal. Cette démarche permet d'intégrer implicitement des éléments de physique dans les données fournies aux algorithmes de ML (cf. § 1.2.1).

Concrètement, cela consiste à caler, dans un premier temps, l'équation de progression du tassement en fonction de l'avancement du creusement sur les mesures enregistrées par chaque capteur. Cette étape permet de déterminer le tassement maximal observé par un capteur situé à une certaine distance de l'axe du tunnel (d_{axe}), noté s^*. Ensuite, en utilisant ces valeurs s^*, il devient possible de caler l'équation du tassement transversal en fonction de d_{axe}. Ces différentes étapes sont illustrées et détaillées dans la figure **4**.

Une fois le nettoyage des données effectué, il est question de procéder à une analyse approfondie des caractéristiques préliminaires. Cette étape permet de mieux comprendre la distribution des variables, d'identifier des tendances sous-jacentes et de repérer d'éventuelles corrélations entre les différentes variables. L'objectif principal de cette analyse est d'orienter le choix des caractéristiques les plus pertinentes pour l'entraînement du modèle de ML.

L'analyse statistique des données repose sur des mesures telles que la moyenne, la médiane ou l'écart-type. Ces indicateurs offrent une vue d'ensemble de la répartition des données

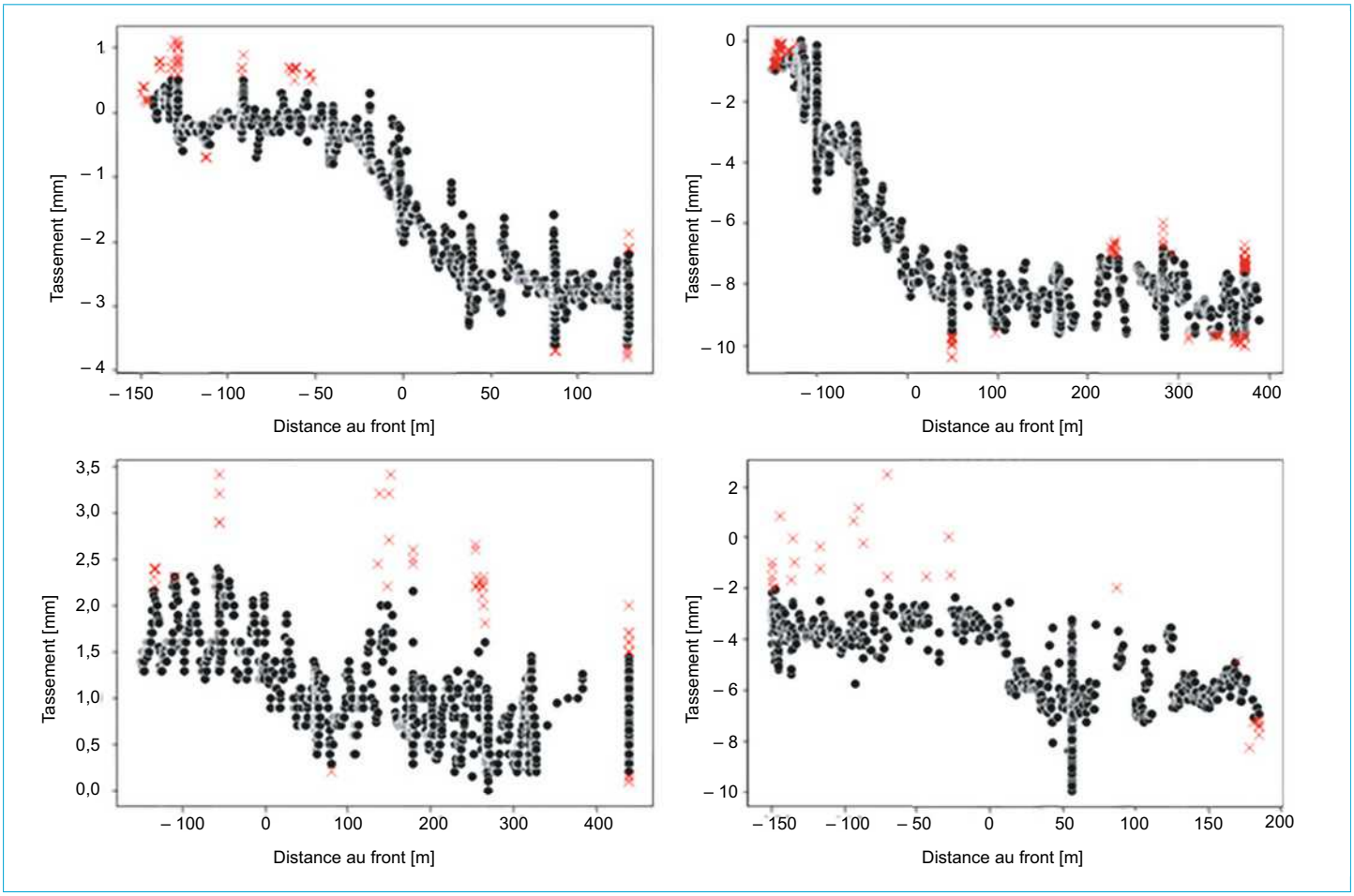

Figure 9 – Exemples de résultats issus de l'algorithme des forêts d'isolation (IF). Les mesures identifiées comme aberrantes sont signalées en rouge

et permettent de vérifier que le nettoyage a été correctement effectué. Par exemple, si l'écart-type est très élevé par rapport à la moyenne, cela peut indiquer une forte variabilité dans les données, voire la présence d'anomalies. Par ailleurs, l'étude des corrélations entre les variables permet d'identifier des relations, qu'elles soient directes ou indirectes, entre les caractéristiques d'entrée et les variables cibles, c'est-à-dire celles que l'algorithme devra prédire. Cette exploration approfondie est essentielle pour guider la sélection des variables qui contribueront le plus à la précision du modèle.

Encadré 10 – Analyse statistique

Dans le cadre de l'analyse statistique, nous avons calculé différentes mesures pour la variable cible (S_{max}) ainsi que pour les caractéristiques préliminaires sélectionnées. Ces mesures permettent d'obtenir une vue d'ensemble de la distribution des données et de leur dispersion.

Le tableau **1** résume ces mesures pour chaque variable. À noter que les paramètres de sols utilisés sont une combinaison des paramètres de l'ensemble des couches (§ 2.3.3).

Les visualisations sont des outils essentiels dans l'analyse exploratoire des données, car elles permettent de mieux comprendre les relations entre les variables et de détecter des anomalies qui ne seraient pas visibles par des simples mesures statistiques. Des outils tels que les diagrammes de dispersion, les histogrammes et les cartes de chaleur de corrélation (*heatmaps*) sont particulièrement utiles. Ces visualisations aident à identifier des tendances sous-jacentes, à repérer des relations entre les variables, et à localiser des points de données aberrantes ou des motifs non évidents.

L'utilisation de ces techniques visuelles est déterminante pour affiner la compréhension des données, guider le choix des transformations nécessaires (par exemple, normalisation ou transformation logarithmique) et orienter la sélection des caractéristiques avant d'entamer l'entraînement du modèle. En outre, elles permettent d'appréhender de manière plus intuitive les interactions entre les différentes variables et de préparer les données de manière optimale pour les étapes suivantes du processus.

Encadre 11 – Analyse de corrélations

Une carte de chaleur est tracée en prenant S_{max} comme variable cible. Les résultats, présentés dans la figure **10**, indiquent que la corrélation la plus forte avec S_{max} est celle de K_0 (R = 0.56), tandis que la corrélation la plus faible est celle avec M_{RDC} (R = 0.07). Les paramètres suivants dans le classement sont c, α et E_M pour S_{max}.

Tableau 1 – Description statistique des variables

	S_{max}	$V_{tunnelier}$ [mm/min]	M_{RDC} [kN.m]	P_{front} [bar]	$P_{mortier}$ [bar]	$V_{mortier}$ [m³]	P_{totale} [kN]
nombre	2 590	2 590	2 590	2 590.0	2 590	2 590	2 590
moyenne	–3.5	33.7	4 687.1	1.2	2.7	10.3	20 599.4
écart-type	2.4	8.4	1 852.5	0.6	0.8	3.3	6 726.5
min	–17.5	10.3	0.0	–0.1	0.0	0.0	3 013.8
25 %	–4.5	27.5	3 538.4	0.8	2.2	10.2	16 146.1
50 %	–3.2	33.6	4 387.1	1.2	2.7	11.0	19 266.9
75 %	–1.7	39.8	5 409.8	1.6	3.2	11.7	23 899.4
max	–0.2	59.5	17 540.6	2.6	6.6	20.8	43 442.2
	C [m]	γ [kN/m³]	E_M [MPa]	α	c [kPa]	φ [°]	K_0
nombre	2 590	2 590	2 590	2 590	2 590	2 590	2 590
moyenne	29.5	12.4	66.6	0.4	41.4	20.0	0.3
écart-type	8.0	1.0	30.2	0.1	19.2	3.1	0.1
min	11.3	11.1	15.7	0.3	10.0	14.3	0.2
25 %	25.2	11.8	40.3	0.3	28.4	18.2	0.3
50 %	28.7	12.2	60.5	0.4	38.3	19.7	0.3
75 %	34.4	12.8	92.5	0.4	55.6	21.7	0.3
max	52.8	18.4	135.1	0.6	80.2	36.3	0.4

2.3.3 Ingénierie des caractéristiques

L'ingénierie des caractéristiques (*feature engineering*) est un processus clef dans la préparation des données pour le ML. Il s'agit de sélectionner les paramètres les plus pertinents (*feature selection*), d'extraire les caractéristiques les plus significatives (*feature extraction*) et, si nécessaire, de mettre à l'échelle ces caractéristiques (*feature scaling*) afin de les préparer efficacement à l'entraînement des modèles.

Le *feature extraction* consiste à combiner plusieurs caractéristiques existantes pour créer une nouvelle variable plus informative et utile pour la prédiction du résultat. Cette étape repose sur une bonne compréhension des données et de leur signification physique, car elle requiert à la fois intuition et créativité. Les transformations des caractéristiques peuvent se faire de différentes manières ; en voici quelques exemples :

- combinaison de paramètres d'entrée : en appliquant des transformations mathématiques telles que l'addition, la soustraction ou d'autres combinaisons entre variables ;
- transformation des variables dissymétriques : certaines variables présentent des distributions fortement asymétriques, ce qui peut compliquer leur analyse statistique ; pour atténuer cette dissymétrie, il est courant d'appliquer des transformations destinées à rapprocher la distribution d'une forme plus symétrique, idéalement en cloche (distribution gaussienne) ; parmi ces transformations figurent le logarithme, la racine carrée, l'élévation à une puissance (*power transform*), ou encore la transformation de Box-Cox, qui est applicable à toute variable continue ;
- transformation de variables catégorielles en numériques : beaucoup de modèles d'apprentissage automatique ne peuvent pas traiter directement les variables catégorielles ; pour ce faire, il existe plusieurs techniques de transformation, telles que l'encodage « un parmi n » (*one-hot encoding*), où chaque catégorie est représentée par un vecteur binaire, ou encore le *label encoding*, qui attribue à chaque catégorie un entier unique ;
- réduction de la dimension : lorsque le nombre de caractéristiques est très élevé, comme c'est souvent le cas dans des projets géotechniques où les données de stratigraphie peuvent être nombreuses, il devient difficile d'entraîner un modèle. Dans ce cas, des méthodes comme l'analyse en composantes principales (*Principal Component Analysis*, PCA) peuvent être utilisées pour réduire la dimensionnalité tout en conservant les informations les plus importantes.

Encadré 12 – Extraction des caractéristiques et réduction des dimensions

La définition précise d'une stratigraphie nécessite de considérer une multitude de paramètres géologiques et géotechniques. Comme indiqué au § 2.2.1, les caractéristiques sélectionnées pour représenter le sol incluent : γ, E_M, K_0, c et φ, auxquelles s'ajoutent la position et l'épaisseur de chaque couche.

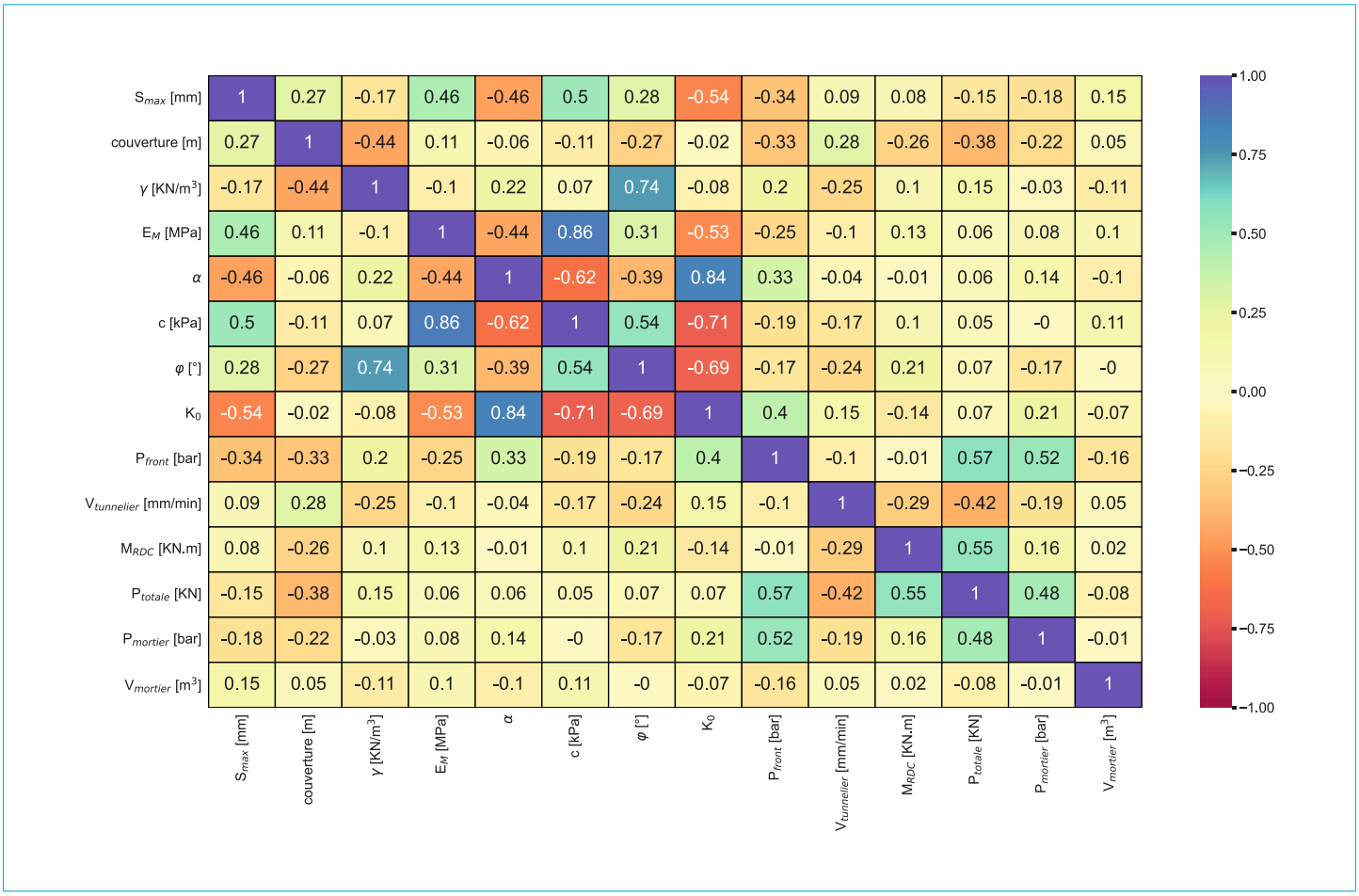

Figure 10 – Carte de chaleur des paramètres ayant une influence sur S_{max}

Encadré 12 – Extraction des caractéristiques et réduction des dimensions (*suite*)

La lithologie rencontrée sur le GPE est particulièrement riche et variée : le nombre total de couches n'est pas constant, et celui des couches géologiques rencontrées au front du tunnel varie également. Ces éléments conduisent à un nombre de caractéristiques à introduire dans l'algorithme de ML se situant autour d'une soixantaine. Toutefois, compte tenu de la quantité limitée de données disponibles, l'utilisation de modèles trop complexes, capables de traiter un tel volume de caractéristiques, n'est pas envisageable.

Pour réduire la dimensionalité du problème, nous adoptons une méthode de combinaison des paramètres du sol, inspirée par les travaux de Chen *et al.* [48]. Cette approche, décrite par l'équation présentée dans la figure **11**, permet de simplifier et structurer les informations tout en conservant leur pertinence pour l'algorithme. En d'autres termes, les variations d'ensemble induites par les changements géologiques sont correctement prises en compte.

Le *feature selection* consiste à identifier les variables les plus pertinentes pour le modèle afin de réduire la dimensionalité des données. Cela permet de conserver les caractéristiques ayant le plus grand impact sur la variable cible, tout en éliminant celles qui sont non pertinentes. Cette étape est possible après les analyses statistiques et les transformations de variables. Elle améliore la vitesse d'entraînement, réduit le risque de surapprentissage (*overfitting*) et optimise la performance générale du modèle. En sélectionnant efficacement les caractéristiques, on évite que le modèle ne devienne trop spécifique aux données d'entraînement, améliorant ainsi sa capacité à généraliser.

Encadré 13 – Sélection des caractéristiques finales

Les caractéristiques choisies pour la prévision de S_{max} sont présentées dans le tableau **2**. Les analyses statistiques précédentes ont montré que M_{RDC} a la corrélation la plus faible avec la cible S_{max} (R^2 = 0.09). Nous avons donc choisi de ne pas le prendre en compte en tant que caractéristique pour l'entraînement des modèles de ML.

Le *feature scaling* est une étape essentielle lorsque les variables ont des échelles différentes, permettant de les rendre comparables. Cela peut être réalisé par normalisation ou standardisation, selon les spécifications de l'algorithme choisi. La normalisation ajuste les variables pour qu'elles soient comprises dans un même intervalle, généralement entre 0 et 1, tandis que la standardisation consiste à transformer les données pour qu'elles aient une moyenne de zéro et un écart-type de 1.

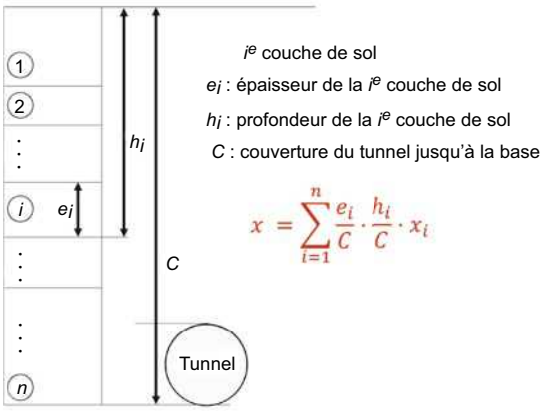

Figure 11 – Schématisation de la méthode de combinaison des paramètres de sols, adaptée de Chen *et al.* (2019)

Certains algorithmes, comme RF et GBoost, n'ont pas besoin de mise à l'échelle des données. En revanche, des algorithmes sensibles à la distance, comme les SVM ou les k-plus proches voisins (k-NN), exigent généralement une normalisation ou une standardisation pour que les distances entre les points de données ne soient pas dominées par les variables ayant une échelle plus large.

Tableau 2 – Caractéristiques utilisées pour la prévision de S_{max}

Catégorie	Caractéristique
Géométrie	Couverture C (m)
Pilotage du tunnelier	Vitesse d'avancement $V_{tunnelier}$ (mm/min)
	Pression au front P_{front} (bar)
	Poussée totale du tunnelier P_{totale} (kN)
	Pression de mortier injecté $P_{mortier}$ (bar)
	Quantité de mortier injecté $V_{mortier}$ (m^3)
Géologie et géotechnique (paramètres « combinés »)	Poids volumique γ(kN/m^3)
	Module pressiométrique de Ménard E_M (MPa)
	Coefficient rhéologique α
	Cohésion c (kPa)
	Angle de frottement φ(°)
	Coefficient de pression des terres au repos K_0

À retenir

- Le contexte et le cadrage sont essentiels pour orienter l'analyse.
- L'extraction et le nettoyage des données représentent environ 80 % du temps de travail.
- Les données d'auscultation nécessitent un traitement spécifique.
- Les bases de données relationnelles sont adaptées au stockage des données structurées.
- Les analyses statistiques permettent de comprendre les données et de sélectionner les caractéristiques les plus pertinentes.

3. Méthodologie d'implémentation d'un modèle d'apprentissage automatique

3.1 Entraînement et validation

3.1.1 Apprentissage

La première étape commence par l'entraînement d'un algorithme avec ses hyperparamètres par défaut. Pour cela, il suffit de prendre 70 ou 80 % de l'ensemble des données pour l'entraînement et garder le reste pour le test. Cette répartition dépend de la taille de l'ensemble de données à disposition : il faut avoir suffisamment de données pour l'entraînement tout en gardant un jeu de données d'une taille acceptable de données pour le test et donc la vérification de la capacité de généralisation du modèle.

Encadré 14 – Entraînement initial

À ce stade, les données sont divisées aléatoirement avec mélange pour garantir une répartition homogène. Le jeu de données, composé de 2 592 observations, est séparé en 80 % pour l'ensemble d'apprentissage (2 072 observations) et 20 % pour l'ensemble de test (518 observations). Les résultats mettent en évidence des variations significatives entre les performances des algorithmes. Le modèle *linear regression* affiche un score faible avec un R^2 de 0.44, confirmant ainsi la complexité du problème. Les modèles fondés sur les arbres de décision (DT, RF et XGBoost) montrent des R^2 proches de 1, mais présentent des signes de surapprentissage nécessitant un travail d'optimisation. Les SVM, quant à eux, offre une performance compétitive avec un R^2 de 0.91, légèrement inférieur à RF et XGBoost, tout en démontrant une bonne capacité de généralisation.

L'optimisation des modèles d'apprentissage automatique est essentielle pour éviter à la fois le surapprentissage (*overfitting*) et le sous-apprentissage (*underfitting*), assurant ainsi la capacité du modèle à généraliser efficacement à partir de nouvelles données. Ce processus repose principalement sur l'ajustement des hyperparamètres, qui jouent un rôle clef dans le comportement et les performances des algorithmes. Un réglage optimal permet d'atteindre une précision maximale tout en réduisant les erreurs sur les données de test. Il s'agit, en particulier, d'éviter toute tendance au sous-apprentissage ou au surapprentissage du modèle à

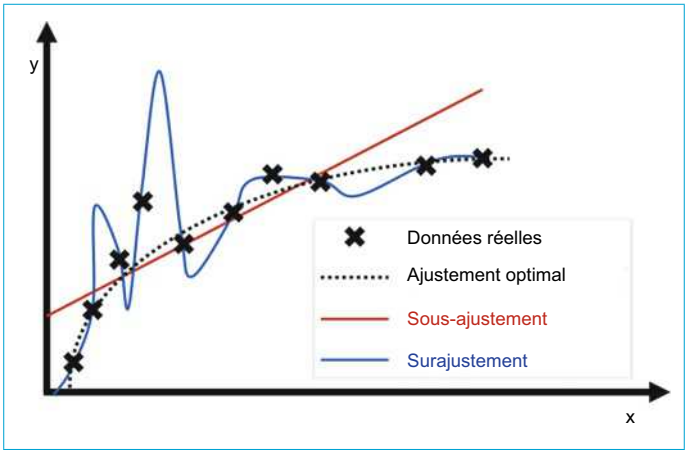

Figure 12 – Surapprentissage (*overfitting*) et sous-apprentissage (*underfitting*) des données

ajuster. Comme illustré dans la figure **12**, un modèle en sous-apprentissage échoue à capturer les tendances sous-jacentes des données, tandis qu'un modèle en surapprentissage devient trop spécifique aux données d'entraînement, compromettant ainsi sa capacité à prédire de manière fiable à partir de données nouvelles.

3.1.2 Validation croisée

La validation croisée (figure **13**, [56]) est une méthode d'évaluation qui divise le jeu de données en plusieurs sous-ensembles (*folds*). Prenons k, le nombre de sous-ensembles. À chaque itération, le modèle est entraîné sur k-1 sous-ensembles et testé sur le sous-ensemble restant, ce processus étant répété pour couvrir toutes les permutations possibles. Cette méthode est souvent utilisée pour évaluer différentes combinaisons d'hyperparamètres, en testant leur impact sur les performances du modèle afin d'identifier la combinaison optimale. Enfin, une évaluation finale est réalisée sur un jeu de test indépendant que le modèle n'a jamais vu, pour vérifier sa capacité réelle de généralisation.

Cette approche permet d'améliorer la capacité du modèle à généraliser ses prédictions à des données non vues et d'éviter un surajustement au seul jeu d'entraînement. En outre, la validation croisée contribue à atténuer l'effet des « fluctuations statistiques » (*flukes*), des variations aléatoires dans les données d'entraînement qui peuvent fausser l'évaluation de la performance d'un modèle. Ces fluctuations peuvent donner l'impression qu'un modèle fonctionne mieux qu'il ne le fait réellement, simplement à cause d'anomalies ou de particularités présentes dans un sous-ensemble spécifique des données. En répétant l'entraînement et les validations sur différents sous-ensembles, la validation croisée permet de réduire l'impact de ces anomalies et de fournir une estimation plus fiable des performances du modèle.

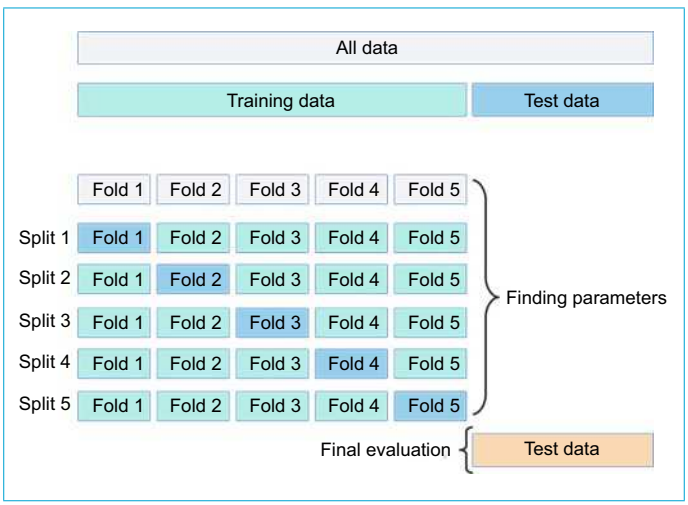

Figure 13 – Validation croisée [56]

3.1.3 Courbes d'apprentissage

Les courbes d'apprentissage (*learning curves*) jouent un rôle déterminant dans l'évaluation de la capacité de généralisation d'un modèle en montrant l'évolution de l'erreur de prédiction en fonction de la taille de l'ensemble d'entraînement, l'erreur étant obtenue via une validation croisée. Elles permettent de déterminer si un modèle a atteint son potentiel optimal ou s'il pourrait encore bénéficier de données supplémentaires, tout en identifiant d'éventuels problèmes de surapprentissage.

Concrètement, plusieurs situations peuvent être rencontrées :

- **mauvaise capacité de généralisation :** une différence significative entre l'erreur obtenue sur les données d'entraînement et celle sur les données de validation montre que le modèle peine à généraliser ses prédictions aux nouvelles données, indiquant qu'il est trop spécifiquement ajusté aux données d'entraînement ;
- **surapprentissage :** lorsque l'erreur d'entraînement continue de diminuer, cela montre que le modèle devient de plus en plus précis sur les données d'entraînement, mais si l'erreur de validation commence à augmenter, cela indique un surapprentissage ; le modèle devient alors trop spécifique aux données d'entraînement et perd sa capacité à généraliser sur des données non vues ; de plus, si l'erreur d'entraînement atteint une valeur extrêmement faible, voire minimale, cela peut également être un signe de surapprentissage ; dans ce cas, le modèle mémorise parfaitement les données d'entraînement au lieu d'apprendre des tendances généralisables ;
- **atteinte du potentiel optimal :** un modèle a atteint son potentiel maximal lorsque l'ajout de nouvelles données d'entraînement ou l'augmentation du nombre d'itérations d'entraînement n'améliore plus de manière significative sa performance sur les données de validation ; cela indique que le modèle a extrait toutes les informations utiles des données disponibles et qu'il est peu probable qu'une quantité supplémentaire de données avec les mêmes tendances ait un impact notable sur sa performance.

Encadré 15 – Validation croisée et courbes d'apprentissage

Les résultats des courbes d'apprentissage pour les algorithmes SVM, DT, RF et XGBoost sont présentés en figure **14**.

Pour l'algorithme SVM, les courbes d'apprentissage indiquent que le modèle n'est pas en surapprentissage. La courbe d'entraînement ne sature pas à un R^2 de 1, et la courbe de validation tend progressivement vers celle d'entraînement avec l'ajout de données. Cela suggère que SVM est capable de généraliser correctement à de nouvelles données.

En revanche, les modèles DT, RF et XGBoost montrent des signes évidents de surapprentissage. Les courbes d'apprentissage atteignent un R^2 de 1 avec un très faible nombre de données, ce qui reflète une mémorisation excessive des données d'entraînement. Une optimisation des hyperparamètres est nécessaire pour limiter ce surapprentissage et améliorer leur capacité de généralisation.

Encadré 15 – Validation croisée et courbes d'apprentissage (*suite*)

En comparant les courbes des différents algorithmes, on observe que DT nécessite une quantité de données plus importante que RF et XGBoost pour atteindre des performances similaires. Par exemple, avec 600 observations, SVM et DT atteignent un R^2 d'environ 0,8 sur les données de validation, tandis que RF et XGBoost dépassent 0,9. Cela met en évidence que les méthodes d'ensemble (*ensemble methods*) comme RF et XGBoost offrent une généralisation bien supérieure aux modèles simples, et ce, avec une moindre quantité de données. Par conséquent, pour ces algorithmes, il peut être pertinent de tester des ensembles d'apprentissage plus réduits afin d'optimiser les ressources de calcul.

Pour cela, on utilise dans ce qui suit un ensemble de test représentant 70 % du jeu de données, soit 1 813 observations

Encadré 15 – Validation croisée et courbes d'apprentissage (*suite*)

pour le test et 777 pour l'entraînement. Les résultats montrent que cette répartition est suffisante pour entraîner les algorithmes DT, RF, XGBoost et SVM. En effet, les scores R^2 obtenus sur l'ensemble de test sont respectivement de 0.82, 0.91, 0.93 et 0.81 (figure **15**). Cela confirme que l'algorithme XGBoost est le plus performant, suivi de RF, DT et enfin SVM. Cependant, les modèles fondés sur des arbres (DT, RF, XGBoost) continuent à surajuster sur les données d'apprentissage, nécessitant une optimisation des hyperparamètres. À l'inverse, SVM montre une courbe d'apprentissage croissante sans dépasser un R^2 de 0.9, avec une courbe de validation convergeant presque vers celle de l'apprentissage, indiquant qu'il ne souffre pas de surapprentissage. Néanmoins, SVM est écarté car il nécessite davantage de données pour atteindre des performances comparables à celles des autres modèles.

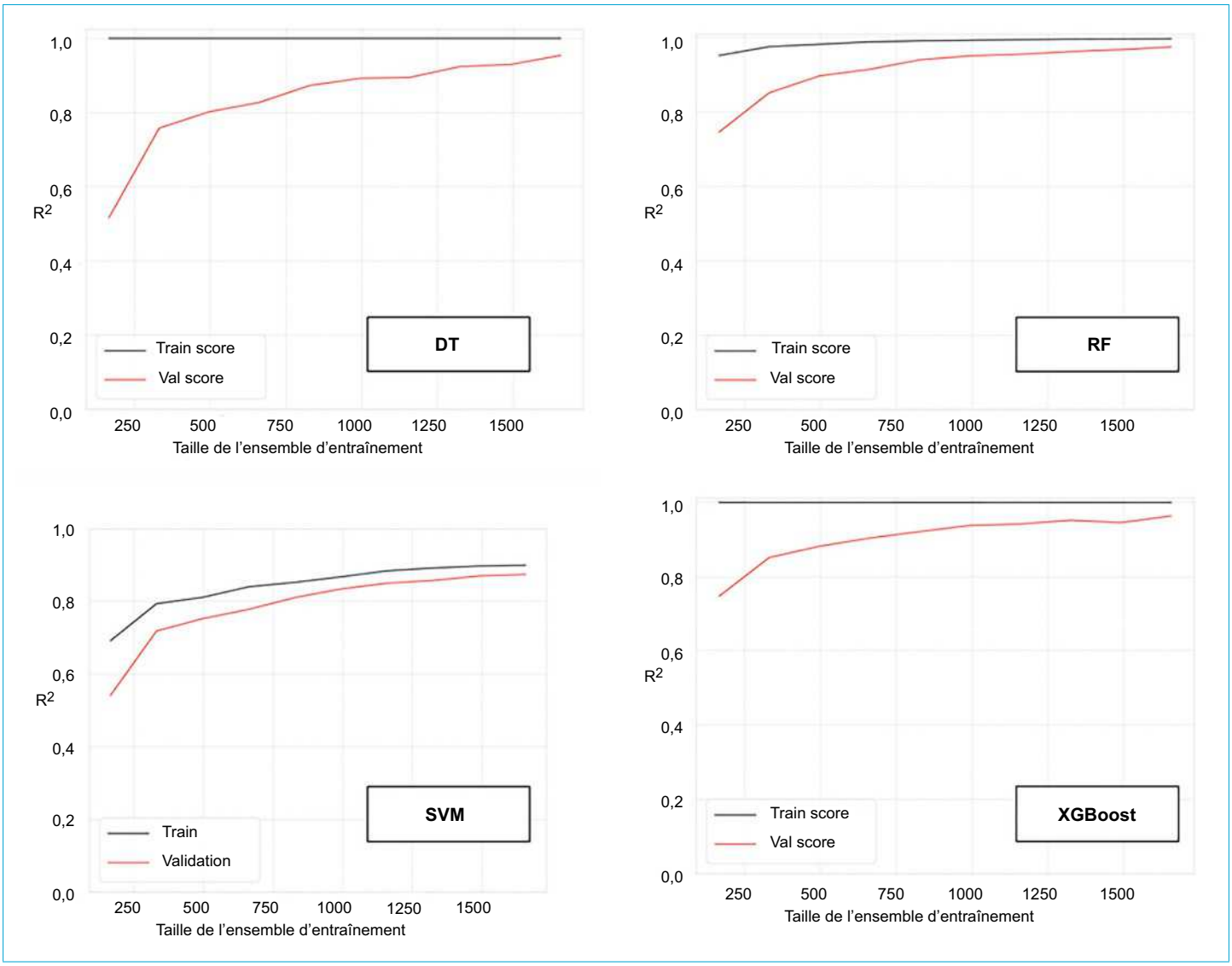

Figure 14 – Courbes d'apprentissage pour DT, RF, SVM et XGBoost

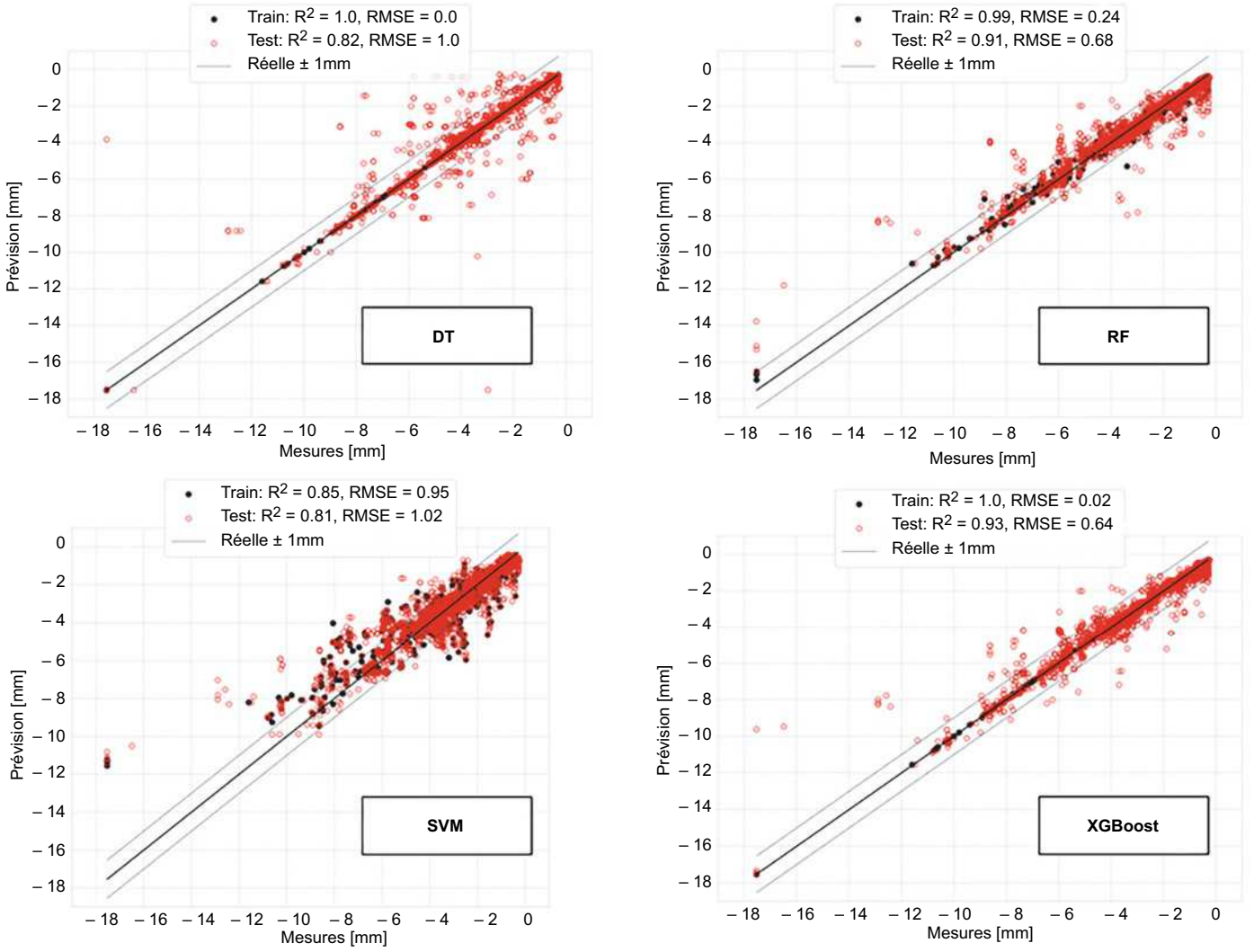

Figure 15 – Résultats des modèles obtenus à partir de DT, RF, SVM et XGBoost

3.2 Régularisation et obtention du modèle

3.2.1 Régularisation et optimisation

Pour obtenir un modèle optimal, capable d'éviter à la fois le surapprentissage et le sous-apprentissage (figure **12**), il est essentiel d'ajuster les hyperparamètres à l'aide de techniques de régularisation. Ces techniques permettent de maîtriser la complexité du modèle tout en améliorant ses performances. Cela implique de trouver un compromis biais-variance (figure **16**, adaptée de [57]) : un modèle trop simple présente un biais élevé, reflétant des erreurs systématiques dues à une sous-modélisation, tandis qu'un modèle trop complexe souffre d'une variance élevée, le rendant instable et trop sensible aux variations des données d'entraînement. L'objectif est donc de paramétrer le modèle de manière à atteindre cet équilibre optimal, garantissant des prédictions fiables et une bonne généralisation sur des données non vues.

La régularisation permet d'identifier des plages optimales pour les hyperparamètres de l'algorithme de ML. Une fois ces intervalles définis, l'étape suivante consiste à optimiser les hyperparamètres (*hyperparameter tuning*), en recherchant la combinaison de valeurs la plus performante au sein de ces plages. Ce processus d'optimisation repose sur des techniques de recherche telles que le *grid search*, qui explore exhaustivement toutes les combinaisons possibles d'hyperparamètres, ou le *random search*, qui sélectionne aléatoirement un sous-ensemble de ces combinaisons, offrant ainsi un compromis entre exhaustivité et efficacité. L'optimisation des hyperparamètres est donc essentielle pour maximiser la performance de l'algorithme.

3.2.2 Application sur decision trees

La régularisation d'un arbre de décision (DT) peut se faire en limitant sa profondeur. Si cet hyperparamètre n'est pas spécifié, il n'y aura aucune limite sur la profondeur et donc l'arbre « grandit » autant que nécessaire pour trouver le meilleur modèle

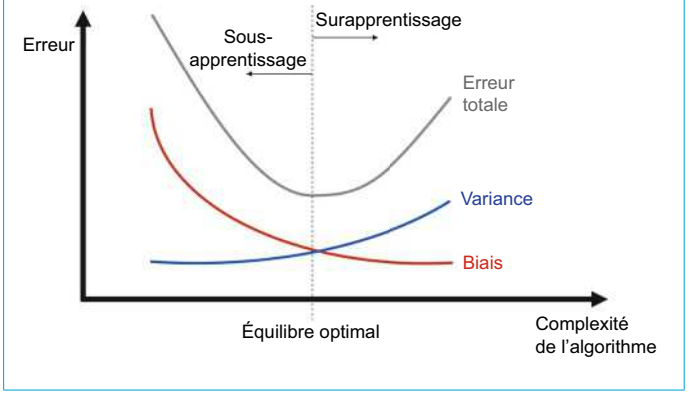

Figure 16 – Évaluation de l'équilibre optimal par décomposition de l'erreur

possible. En effet, si l'on observe l'arbre obtenu par le modèle sans régularisation, on retrouve un arbre très profond ce qui confirme encore une fois le surapprentissage du modèle. Afin de trouver la valeur optimale d'un hyperparamètre, on calcule le biais et la variance du modèle. La valeur optimale de l'hyperparamètre est celle qui minimise ces deux erreurs.

Encadré 16 – Régularisation des hyperparamètres – DT

Les résultats de ce calcul (figure **17**) indiquent que la valeur optimale de la profondeur de DT se situe autour de 9. À partir d'une profondeur de 11, l'erreur semble se stabiliser, ce qui suggère que l'arbre atteint un niveau de complexité suffisant pour capturer les tendances principales des données. Après régularisation, le modèle présente une légère baisse de performance (R^2 de 0.85 contre 0.86 précédemment), mais il n'est plus en surapprentissage. Cela se reflète par le fait que le R^2 d'apprentissage n'atteint pas la valeur 1, quel que soit le nombre de données. Par ailleurs, la courbe de validation indique une amélioration significative des performances avec l'augmentation de la taille de l'ensemble d'entraînement. Ces observations confirment que l'algorithme DT bénéficie grandement de davantage de données pour obtenir des résultats optimaux.

3.2.3 Application sur random forest

La régularisation de RF repose sur le réglage de plusieurs hyperparamètres clefs, dont les principaux sont :

- le nombre d'estimateurs : correspond au nombre d'arbres dans la forêt ; en général, un nombre élevé d'arbres réduit le risque de surapprentissage en améliorant la stabilité du modèle ; toutefois, cela entraîne une augmentation proportionnelle du temps de calcul ;
- la profondeur maximale des estimateurs : contrôle la complexité des estimateurs en limitant le nombre de niveaux dans chaque arbre ; telle qu'expliquée pour le DT, cette profondeur est illimitée, ce qui peut parfois mener à un surapprentissage dans des ensembles de données complexes ;
- le nombre de caractéristiques utilisées pour la division des nœuds (*max_features*) : cet hyperparamètre détermine combien de variables sont considérées à chaque division ; par défaut, cette valeur est fixée à une seule caractéristique ; réduire la valeur de *max_features* favorise la diversité entre les arbres et réduit la variance du modèle, mais cela peut augmenter le biais (Sckit-learn, 2023) ; une valeur initiale comprise entre 30 % et 50 % du nombre total de caractéristiques est souvent recommandée comme point de départ pour les tests.

Encadré 17 – Régularisation des hyperparamètres – RF

Selon les résultats observés, on trouve que les plages de valeurs optimales des hyperparamètres de RF sont au moins 16 estimateurs, de profondeur entre 6 et 11 et un *max_features* entre 3 et 6.

Par la suite, on effectue une recherche aléatoire dans la plage des hyperparamètres retrouvés. Il convient de rappeler qu'il est recommandé d'avoir un grand nombre d'estimateurs, c'est pourquoi on propose une plage de nombre d'estimateurs entre le minimum obtenu (16) et 150. Cette recherche retourne les hyperparamètres suivants : 100 estimateurs, de profondeur maximale de 11 et un *max_features* de 4.

Les résultats du modèle optimisé (figure **18**) montrent une très légère différence avec le modèle initial avec un RMSE qui passe de 0.67 à 0.65 et un R^2 constant de 0.92. Néanmoins, on observe que le modèle régularisé arrive également à des R^2 supérieurs à 0.9 pour l'ensemble de validation, éliminant ainsi toute crainte d'aléas statistique sur le lot de test. Donc, malgré le fait que le R^2 d'apprentissage est presque de 1, les résultats montrent bien un modèle capable de généraliser sur des nouvelles données avec une très bonne performance.

3.2.4 Application sur XGBoost

La régularisation de XGBoost repose sur le réglage de plusieurs hyperparamètres clefs. Comme pour les RF, le nombre d'estimateurs et la profondeur maximale des arbres jouent un rôle crucial dans le compromis biais-variance. En complément, XGBoost introduit des paramètres spécifiques, tels que gamma. Ce dernier est un paramètre de régularisation influençant la réduction de gain minimale nécessaire pour effectuer une division ; des valeurs élevées réduisent le risque de surapprentissage en pénalisant les divisions inutiles ; la valeur par défaut est de 0, tandis que des valeurs autour de 20 sont considérées comme très restrictives [58].

Encadré 18 – Régularisation des hyperparamètres – XGBoost

Selon les résultats, on trouve que les plages de valeurs optimales des hyperparamètres de XGBoost sont un nombre d'estimateurs d'au moins 11, une profondeur maximale comprise entre 6 et 11 et un gamma inférieur à 2.

Une recherche aléatoire effectuée sur ces plages a permis d'identifier les hyperparamètres optimaux suivants : 50 estimateurs, de profondeur maximale de 6 et un gamma égal à 1.

Les performances du modèle optimisé (figure **19**) montrent une légère diminution du R^2, passant de 0,93 à 0,90. Malgré cette réduction, les résultats obtenus restent satisfaisants, avec un R^2 autour de 0,90 pour les ensembles de test et de validation. Ces performances indiquent que le modèle conserve une bonne capacité de généralisation, bien que le R^2 sur l'ensemble d'apprentissage effleure les valeurs de 1. Cette analyse montre qu'avec une régularisation adaptée, XGBoost est capable de fournir des résultats fiables et généralisables, comparables à ceux obtenus avec les forêts aléatoires.

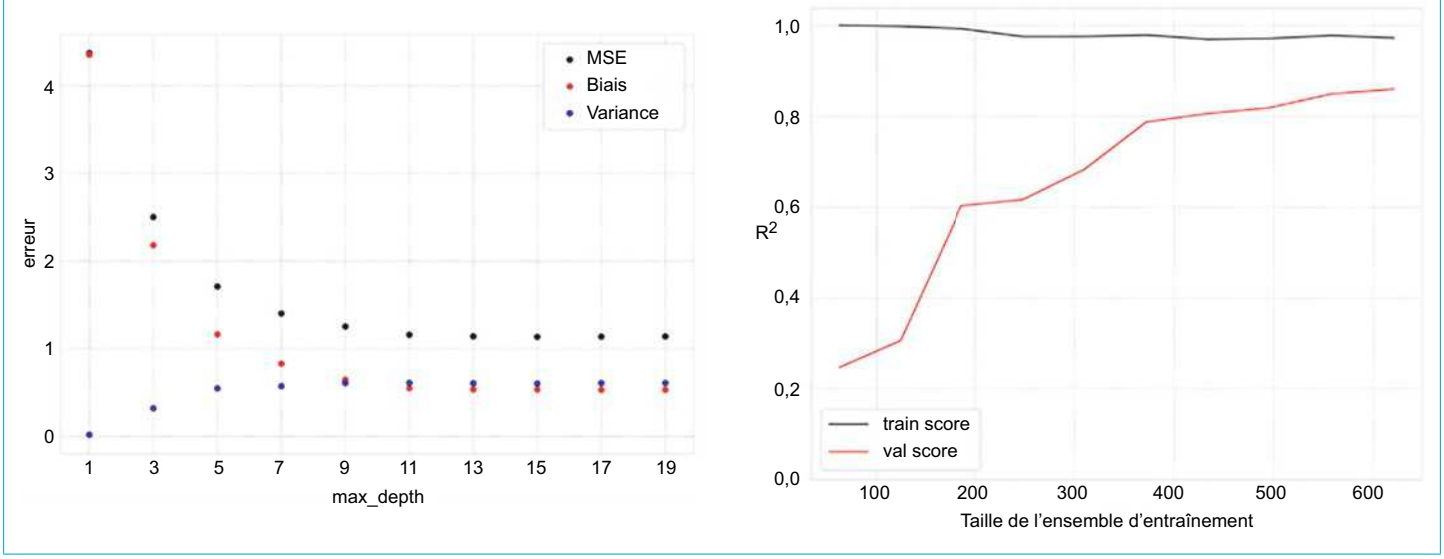

Figure 17 – Résultats du biais et de la variance obtenus en variant la profondeur de DT et courbe d'apprentissage associée

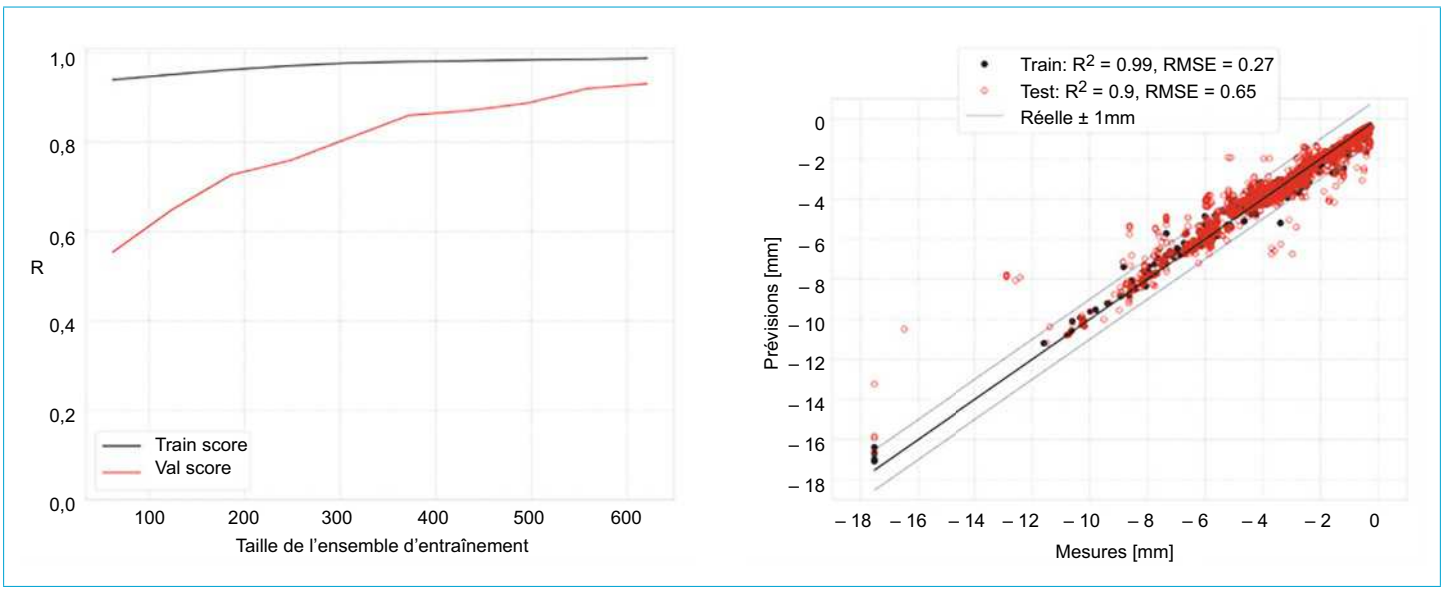

Figure 18 – Courbe d'apprentissage et résultats obtenus par RF

Encadré 19 – Modèle final

Les meilleures performances sont celles obtenues par RF lors de la prévision de S_{max}. Cet algorithme sera alors utilisé dans la partie suivante avec les hyperparamètres et les caractéristiques optimisés pour la prévision de S_{max}. L'objectif étant de tester les capacités de RF à prévoir le tassement S_{max} à l'avant du front, sur des zones non creusées, en temps réel.

3.3 Prévision en temps réel

La prévision des tassements induits par le creusement au tunnelier constitue un enjeu crucial pour garantir la sécurité des infrastructures environnantes et minimiser les risques de dommages. Dans le cadre du Grand Paris Express, une stratégie innovante de prévision en temps réel a été mise en place, permettant d'adapter les interventions selon l'évolution des conditions de creusement.

La prévision des tassements à venir au fur et à mesure du creusement au tunnelier ne peut s'appuyer sur une division aléatoire des données. Une telle approche impliquerait de choisir aléatoirement les instances de test sur l'ensemble du tracé, ce qui ne reflète pas les conditions réelles d'un creusement de tunnel. En réalité, les données disponibles concernent les zones déjà creusées, et l'objectif est de prévoir les tassements futurs à l'avant du front de creusement. Ce contexte impose de prendre en compte l'aspect spatio-temporel lors de la sélection des données d'apprentissage et de test pour prédire le tassement maximal à

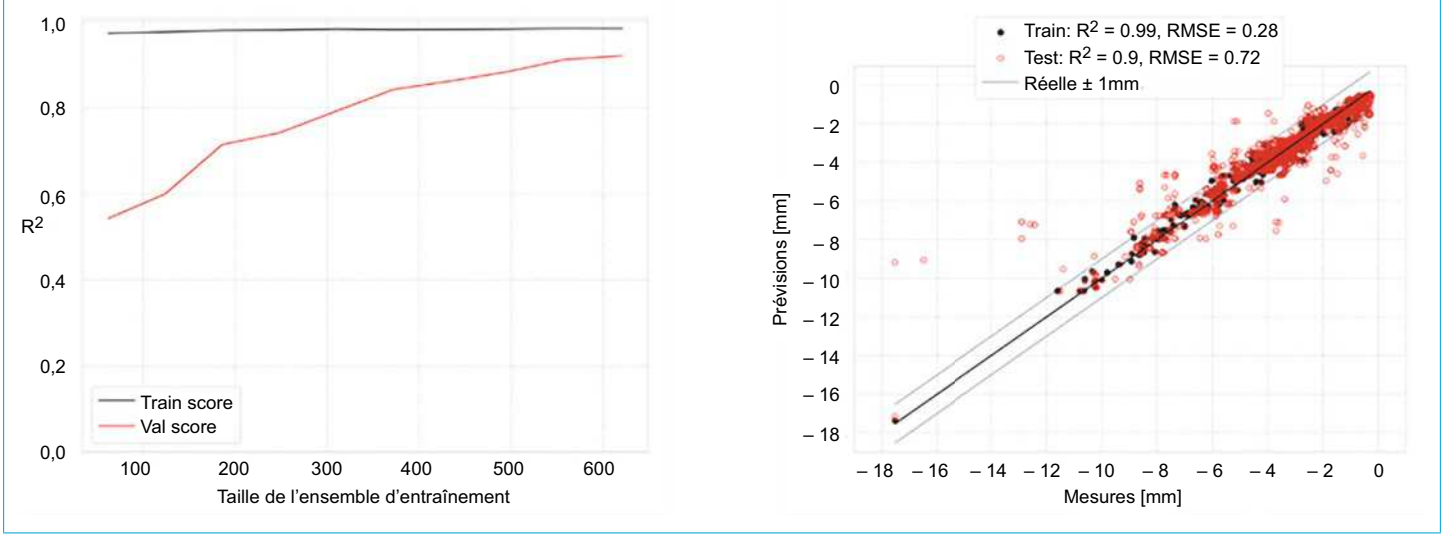

Figure 19 – Courbe d'apprentissage et résultats obtenus par XGBoost

l'axe du tunnel (S_{max}). Cette capacité de prévision sur des zones non encore creusées est qualifiée d'« extrapolation spatiale », distincte de l'« extrapolation statistique », qui concerne la prédiction dans des plages de données inconnues au sein de l'ensemble d'apprentissage.

Un autre défi est la zone de transition où les tassements sont encore en cours et n'ont pas atteint leur valeur maximale (S_{max}). Ces zones peuvent perturber les algorithmes de ML. Selon les analyses statistiques [55], 95 % du tassement maximal est atteint à environ 75 m à l'arrière du front dans 75 % des cas. Par précaution, une zone de 100 m à l'arrière du front a été définie pour neutraliser les valeurs de tassements. Les algorithmes sont ainsi entraînés en utilisant les données S_{max} extraites 100 m à l'arrière du front du tunnel (figure **20**).

Le modèle développé dans la partie précédente à l'aide de l'algorithme de RF est utilisé. Les résultats montrent qu'il est fonctionnel à partir de 500 m de creusement, avec la capacité de prévoir les tassements jusqu'à 150 m à l'avant du front. L'entraînement du modèle repose sur les caractéristiques (géométrie, paramètres de pilotage du tunnelier, données de sol) observées à l'arrière du front, tandis que les tests utilisent les caractéristiques disponibles à l'avant. Dans un projet réel, seules les données de consigne pour le pilotage sont disponibles pour les zones non encore creusées et peuvent être utilisées comme caractéristiques pour prédire sur les zones à l'avant du front.

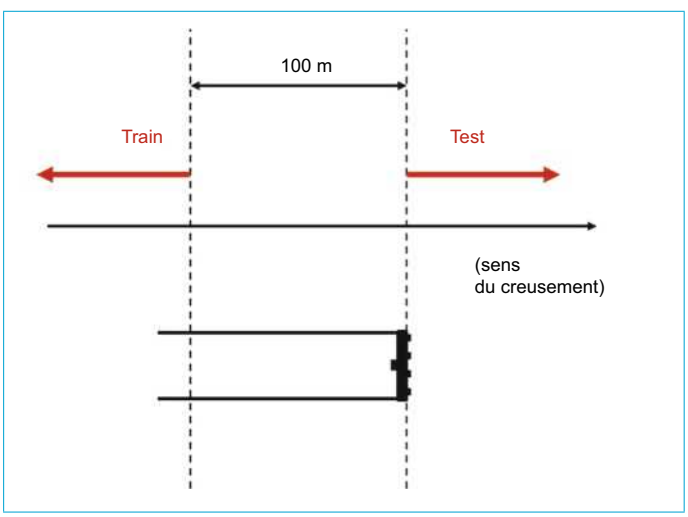

Figure 20 – Division des données en ensembles d'apprentissage et de test en tenant compte de l'aspect spatio-temporel du problème

Pour maintenir la précision des prévisions, un réentraînement mensuel est effectué (figure **21**). Ce processus intègre les nouvelles données collectées au fur et à mesure de l'avancement, mettant à jour les paramètres du modèle et renforçant sa capacité prédictive. Ce cycle d'apprentissage continu permet de gérer les incertitudes géologiques et d'optimiser le suivi des déformations en temps réel.

En conclusion, cette approche innovante de prévision en temps réel, combinant données historiques et mises à jour régulières, permet de mieux anticiper les tassements induits par le creusement au tunnelier, tout en garantissant une réponse rapide et efficace face aux anomalies.

À retenir

- Diviser les données en ensembles d'apprentissage et de test est essentiel pour évaluer la performance du modèle et sa capacité de généralisation.
- La validation croisée renforce la robustesse du modèle et limite le surapprentissage.
- Les courbes d'apprentissage permettent d'analyser la convergence du modèle en fonction de la quantité de données utilisées.
- La régularisation contrôle la complexité du modèle, l'optimisation permet d'en améliorer les performances.
- La prédiction en temps réel des tassements doit tenir compte des contraintes spatiales et temporelles.

4. Conclusion

Cet article examine en profondeur les applications du ML en géotechnique, en illustrant son application au cas d'usage de la prédiction des tassements induits par le creusement des tunnels. Il insiste

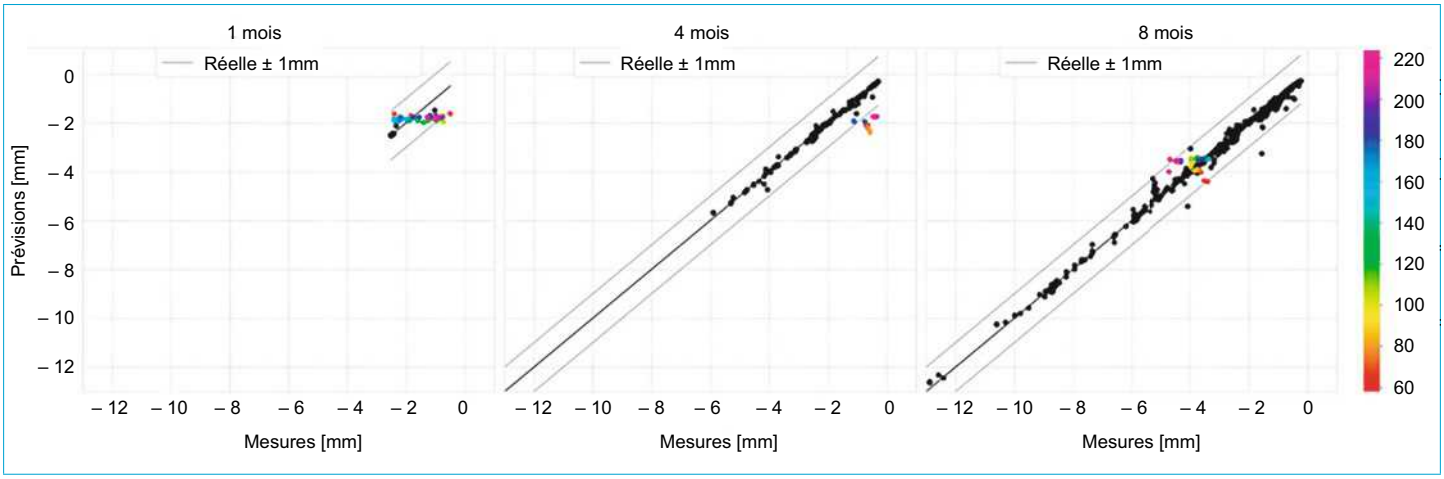

Figure 21 – Exemples de résultats des entraînements mensuels à 1, 4 et 8 mois. L'échelle de couleur représente la distance en m du point de prédiction au front du tunnel

sur l'importance de privilégier la qualité des données à leur quantité, un impératif dans un domaine où les données issues des chantiers sont souvent hétérogènes, bruitées ou incomplètes. Bien que chronophage et exigeante, la phase de préparation des données constitue un investissement essentiel pour obtenir des résultats fiables et exploitables.

Les analyses montrent que les modèles de ML, notamment les méthodes ensemblistes comme XGBoost ou RF, permettent des prédictions précises et une généralisation robuste, tout en s'adaptant aux exigences des projets complexes grâce à un entraînement rapide et une capacité à intégrer aisément de nouvelles données. Ces approches offrent un avantage décisif pour optimiser les pratiques actuelles et anticiper les défis spécifiques à la géotechnique.

Enfin, cet article ouvre des perspectives pour l'avenir du ML en géotechnique. L'intégration de technologies émergentes, telles que les réseaux de neurones informés par la physique, l'utilisation de processus automatisés pour nettoyer et structurer les données dès leur collecte, ou encore l'intégration des LLM pour l'extraction et l'analyse de ces données, pourrait transformer la manière de travailler en rendant accessibles des volumes de données dignes du big data. Cela impliquera également de relever les défis de la gestion massive de données et d'adopter des techniques adaptées pour en tirer pleinement parti. Ces évolutions, combinées à une amélioration continue des algorithmes et à la démocratisation des outils de ML, promettent de rendre ces approches encore plus accessibles et puissantes. Malgré tout, l'expertise humaine restera centrale : le ML demeurera un outil complémentaire précieux aux méthodes traditionnelles, enrichissant mais jamais ne remplaçant l'ingénieur. Des formations interdisciplinaires visant à présenter les bases de ces outils aux ingénieurs métier demeurent un point clef de l'évolution de l'enseignement au cours des prochaines années.

5. Sigles, notations et symboles

Symbole	Description	Unité
γ	Poids volumique humide	kN/m^3
φ	Angle de frottement	°

Symbole	Description	Unité
ANFIS	*Adaptive neuro-fuzzy inference system*	–
ANN	*Artificial neural network* (réseau de neurones)	–
CNN	*Convolutional neural networks*	–
c	Cohésion	kPa
d_{axe}	Distance à l'axe du tunnel	m
DL	*Deep learning* (apprentissage profond)	–
DT	*Decision tree* (arbre de décision)	–
E_M	Module de déformation de l'essai pressiométrique	MPa
FIS	*Fuzzy inference system*	–
GAN	*Generative adversarial network*	–
GPE	Grand Paris Express	–
IA	Intelligence artificielle	–
K_0	Coefficient de pression des terres au repos	–
LSTM	*Long short-term memory networks*	–
ML	*Machine learning* (apprentissage automatique)	–
ResNet	*Residual networks*	–
RF	*Random forest* (forêt aléatoire)	–
S^*	Tassement maximal à une distance de l'axe du tunnel	mm
S_{max}	Tassement maximal à l'axe du tunnel	mm
SVM	*Support vector machine* (séparateur à vaste marge)	–
XGBoost	*Extreme gradient boosting*	–

Apprentissage automatique en géotechnique : étude de cas dans le domaine des tunnels

par **Tatiana RICHA**

Ingénieure data en géotechnique
Terrasol Setec, Paris France

Lina-María GUAYACÁN-CARRILLO

Chargée de recherche en géotechnique (laboratoire Navier), maître de conférences de l'ENPC
École nationale des ponts et chaussées, institut Polytechnique de Paris, Marne-La-Vallée, France

Jean-Michel PEREIRA

Directeur du laboratoire Navier, professeur de l'ENPC
École nationale des ponts et chaussées, Institut Polytechnique de Paris, Marne-La-Vallée, France

et **Gilles CHAPRON**

Directeur des projets data
Terrasol Setec, Paris France

Sources bibliographiques

[1] MOOR (J.). – *The Dartmouth College Artificial Intelligence Conference : The next fifty years.* AI Magazine 27.4, p. 87-91 (2006).

[2] TURING (A.M.). – *Computing Machinery and Intelligence.* Mind 49, p. 433-460 (1950).

[3] ROSENBLATT (F.). – *The perceptron: A probabilistic model for information storage and organization in the brain.* Psychological Review, 65(6), 386–408 (1958). https://doi.org/10.1037/h0042519

[4] CHEN (T.) et HE (T.). – *xgboost : Extreme Gradient Boosting.* R Lecture 2016, p. 1-84 (2014).

[5] CHEN (T.) et GUESTRIN (C.). – *XGBoost: A Scalable Tree Boosting System.* In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco California USA: ACM, 785–94 (2016). doi: 10.1145/2939672.2939785. url: https://dl.acm.org/doi/10.1145/2939672.2939785

[6] CNIL. – *Intelligence Artificielle, de quoi parle-t-on?* (2022). URL : https://www.cnil.fr/fr/intelligence-artificielle/intelligence-artificielle-de-quoi-parle-t-on

[7] SAMUEL (A.). – *Some Studies in Machine Learning Using the Game of Checkers.* IBM Journal of Research and Development. 44: 206–226 (1959). CiteSeerX 10.1.1.368.2254. doi:10.1147/rd.441.0206.

[8] BOURANY (T.). – *Les 5V du big data.* Regards croisés sur l'économie n° 23.2, p. 27-31 (2019). DOI : 10.3917/rce.023.0027.

[9] GÉRON (A.). – *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow.* 3rd. O'Reilly Media, Inc., p. 861 (2022). ISBN: 9781098125974.

[10] SALPERWYCK (C.). – *Apprentissage incrémental en-ligne sur flux de données.* These de doct. Universite Charles de Gaulle – Lille III (2013).

[11] WANG (Z.), JEONG (H.), GAN (Y.), PEREIRA (J.-M.) et GU (Y.) *et al.* – *Porescale modeling of multiphase flow in porous media using a conditional generative adversarial network (cGAN).* Physics of Fluids, 34(12):123325 (2022). ff10.1063/5.0133054

[12] LYU (B.), WANG (Y.) et SHI (C.). – *Multi-scale generative adversarial networks (GAN) for generation of three-dimensional subsurface geological models from limited boreholes and prior geological knowledge*, Computers and Geotechnics,170:106336 (2024).

[13] CHEN (W.), DING (J.), WANG (T.), CONNOLLY (D.P.) et WAN (X.). – *Soil property recovery from incomplete in-situ geotechnical test data using a hybrid deep generative framework*, Engineering Geology (326), 107332 (2023).

[14] MORGENROTH (J.), KHAN (U.T.) et PERRAS (M.A.). – *An Overview of Opportunities for Machine Learning Methods in Underground Rock Engineering Design.* Geosciences 2019;9:504.

[15] JONG (S.), ONG (D.) et OH (E.). – *State-of-the-art review of geotechnical-driven artificial intelligence techniques in underground soil-structure interaction.* Tunnelling and Underground Space Technology 113.March, p. 103946 (2021). DOI : 10.1016/j.tust.2021.103946.

[16] EBID (A.M.). – *35 Years of (AI) in Geotechnical Engineering : State of the Art. T. 39. 2.* Springer International Publishing, p. 637-690 (2021). DOI : 10.1007/s10706-020-01536-7.

[17] ZHANG (W.), LI (H.), LI (Y.), LIU (H.), CHEN (Y.) et DING (X.). – *Application of deep learning algorithms in geotechnical engineering: a short critical review.* Artificial Intelligence Review 54.8, p. 5633-5673 (2021). DOI : 10.1007/s10462-021-09967-1.

[18] BAGHBANI (A.), CHOUDHURY (T.), COSTA (S.) et REINER (J.). – *Application of artificial intelligence in geotechnical engineering : A state-of-the-art review.* Earth-Science Reviews 228.March, p. 103991 (2022). DOI : 10.1016/j.earscirev.2022.103991.

[19] MCCULLOCH (W. S.) et PITTS (W.). – *A logical calculus of the ideas immanent in nervous activity.* The Bulletin of Mathematical Biophysics, 5(4), pp. 115-133 (1943).

[20] ZHANG (G.), PATUWO (B. E.) et HU (M. Y.). – *Forecasting with artificial neural networks: The state of the art.* International journal of forecasting, 14(1), pp. 35-62 (1998).

[21] BOSER (B.E.), GUYON (I.M.) et VAPNIK (V.N.). – *A Training Algorithm for Optimal Margin Classifiers.* Proceedings of the Fifth Annual Workshop on Computational Learning Theory. COLT '92. New York, NY, USA : Association for Computing Machinery, p. 144-152 (1992). DOI : 10.1145/130385.130401.

[22] CORTES (C.), VAPNIK (V.) et SAITTA (L.). – *Support-vector networks.* Machine Learning 20.3, p. 273-297 (1995). DOI : 10.1007/

BF00994018. URL : https://link.springer.com/article/10.1007/BF00994018

[23] QUINLAN (J.R.). – *Induction of decision trees*. Machine Learning 1.1, p. 81-106 (1986). DOI :10.1007/BF00116251. URL : https://link.springer.com/article/10.1007/BF00116251

[24] SARKER (IH). – *Machine Learning: Algorithms*. Real-World Applications and Research Directions. SN Comput Sci. 2021;2(3):160. doi: 10.1007/s42979-021-00592-x. Epub 2021 Mar 22. PMID: 33778771; PMCID: PMC7983091.

[25] BREIMAN (L.). – *Random Forests*. Machine Learning 45, p. 5-32 (2001). DOI : 10. 1023 / A : 1010933404324.

[26] CHU (X.), ILYAS (I.F.), KRISHNAN (S.) et WANG (J.). – *Data cleaning : Overview and emerging challenges*. Proceedings of the ACM SIGMOD International Conference on Management of Data, 26-June-20, p. 2201-2206 (2016). DOI : 10.1145/2882903.2912574.

[27] KLEIN (A.) et LEHNER (W.). – *Representing data quality in sensor data streaming environments*. Journal of Data and Information Quality 1.2 (2009). DOI 10.1145/1577840.1577845.

[28] KRISHNAN (S.), HAAS (D.), FRANKLIN (M.J.) et WU (E.). – *Towards reliable interactive data cleaning: A user survey and recommendations*. HILDA 2016 – Proceedings of the Workshop on HumanIn-the-Loop Data Analytics 1, p. 1-5 (2016). DOI : 10.1145/2939502.2939511.

[29] DECKER (J.B.), ANTONY (A.) et RAY (A.) *et al.* – *An Integrated Relational Database for Tracking Rock Mass Data During Tunnelling*. Tunnelling and Underground Space Technology; 21:429–9 (2006).

[30] MARINOS (V.), PROUNTZOPOULOS (G.), FORTSAKIS (P.), KOUMOUTSAKOS (D.), KORKARIS (K.) et PAPOULI (D.). – *"Tunnel Information and Analysis System" : A Geotechnical Database for Tunnels*. Geotechnical and Geological Engineering 31.3, p. 891-910 (2013). DOI : 10.1007/s10706-012-9570-x. URL: http://link.springer.com/10.1007/s10706-012-9570-x

[31] RICHA (T.), PEREIRA (J.-M.), CHAPRON (G.) et GUAYACAN-CARRILLO (L.-M.). – *Constitution d'une base de données des mesures obtenues lors du creusement de deux tunnels du Grand Paris Express*. JNGG (2022).

[32] GUAYACÁN-CARRILLO (L.-M.), PEREIRA (J.-M.) et SULEM (J.). – *Machine Learning in Geotechnics*. Ingenius, revue numérique de l'Ecole des Ponts ParisTech (2024). https://ingenius.ecoledesponts.fr/en/articles/machine-learning-in-geotechnics/

[33] GUAYACÁN-CARRILLO (L.-M.) et SULEM (J.). – *Symbolic regression based prediction of anisotropic closure in deep tunnels*. Computers and Geotechnics ;171:106355 (2024).

[34] MAS-MEZERAN (H.), PEREIRA (J.-M.), POUYA (A.) et CARTIAUX (F.-B.). – *Modéliser les glissements de terrain dans un contexte de changement climatique, Transitions. Les nouvelles Annales des ponts et chaussées*, n° 3, « Modèles et données pour l'environnement » sous la direction de T. Lelièvre et D. Picard (2023).

[35] TRISTANI (A.), GUAYACÁN-CARRILLO (L.-M.) et SULEM (J.). – *Data-driven tools to evaluate support pressure, radial displacements and face extrusion for tunnels excavated in elastoplastic grounds*. Int J Num & Anal Meth Geomech (2024). https://doi.org/10.1002/nag.3889

[36] FURTNEY (J.K.), THIELSEN (C.), FU (W.) et LE GOC (R.). – *Surrogate Models in Rock and Soil Mechanics:Integrating Numerical Modeling and Machine Learning*. Rock Mechanics and Rock Engineering 55:2845–59 (2022).

[37] MAKASIS (N.), NARSILIO (G.A.) et BIDARMAGHZ (A.). – *A machine learning approach to energy pile design*. Computers and Geotechnics, 97, 189–203 (2018). https://doi.org/10.1016/j.compgeo.2018.01.011

[38] KARNIADAKIS (G.E.), KEVREKIDIS (I.G.) et LU (L.) *et al.* – *Physics-informed machine learning*. Nat Rev Phys 3, 422–440 (2021). https://doi.org/10.1038/s42254-021-00314-5

[39] MOEINEDDIN (A.), SEGUÍ (C.), DUEBER (S.) et FUENTES (R.). – *Physics-informed neural networks applied to catastrophic creeping landslides*. Landslides (2023). https://doi.org/10.1007/s10346-023-02072-0

[40] MASI (F.), STEFANOU (I.), VANNUCCI (P.) et MAFFI-BERTHIER (V.). – *Thermodynamics-based Artificial Neural Networks for constitutive modeling*. Journal of the Mechanics and Physics of Solids, 147, 104277 (2021). https://doi.org/10.1016/j.jmps.2020.104277

[41] MARCINKEVICS (R.) et VOGT (J.E.). – *Interpretable and explainable machine learning: A methods-centric overview with concrete examples*. WIREs Data Mining and Knowledge Discovery, 13(3), e1493 (2023). https://doi.org/10.1002/widm.1493

[42] OVIEDO (F.), FERRES (J.-L.), BUONASSISI (T.) et BUTLER (K.T.). – *Interpretable and Explainable Machine Learning for Materials Science and Chemistry*. Accounts of Materials Research, 3(6), 597–607 (2022). https://doi.org/10.1021/accountsmr.1c00244

[43] MORGENROTH (J.), PERRAS (M.A.) et KHAN (U.T.). – *On the Interpretability of Machine Learning Using Input Variable Selection: Forecasting Tunnel Liner Yield*. Rock Mechanics and Rock Engineering 55:6779–804 (2022).

[44] HUAT (C.Y.), ARMAGHANI (D.J.), MOMENI (E.) et LAI (S.H.). – *Empirical, Statistical, and Machine Learning Techniques for Predicting Surface Settlement Induced by Tunnelling*. Artificial Intelligence in Mechatronics and Civil Engineering : Bridging the Gap. Sous la dir. d'E. Momeni, D. Jahed Armaghani et A. Azizi. Singapore : Springer Nature Singapore, p. 39-77 (2023). DOI : 10.1007/978-981-19-8790-8_2.

[45] SHI (J.), ORTIGAO (J.A.R.) et BAI (J.). – *Modular Neural Networks for Predicting Settlements During Tunneling*. Journal of Geotechnical and Geoenvironmental Engineering 124.May, p. 389-395 (1998).

[46] MARTO (A.), HAJIHASSANI (M.), KALATEHJARI (R.), NAMAZI (E.) et SOHAEI (H.). – *Simulation of longitudinal surface settlement due to tunnelling using artificial neural network*. International Review on Modelling and Simulations 5.2, p. 1024-1031 (2012).

[47] HASANIPANAH (M.), NOORIAN-BIDGOLI (M.), JAHED ARMAGHANI (D.) et KHAMESI (H.). – *Feasibility of PSO-ANN model for predicting surface settlement caused by tunneling*. Engineering with Computers 32.4, p. 705-715 (2016). DOI : 10.1007/s00366-016-0447-0.

[48] CHEN (R.-P.), ZHANG (P.), KANG (X.), ZHONG (Z.-Q.), LIU (Y.) et WU (H.-N.). – *Prediction of maximum surface settlement caused by earth pressure balance (EPB) shield tunneling with ANN methods*. Soils and Foundations 59.2, p. 284-295 (2019). DOI : 10.1016/j.sandf.2018.11.005.

[49] CHEN (R.), ZHANG (P.), WU (H.), WANG (Z.) et ZHONG (Z.). – *Prediction of shield tunneling-induced ground settlement using machine learning techniques*. Frontiers of Structural and Civil Engineering 13.6, p. 1363-1378 (2019). DOI : 10.1007/s11709-019-0561-3.

[50] ZHANG (P.), WU (H.N.), CHEN (R.P.) et CHAN (T.H.). – *Hybrid meta-heuristic and machine learning algorithms for tunneling-induced settlement prediction : A comparative study*. Tunnelling and Underground Space Technology 99.May 2019, p. 103383 (2020). DOI : 10.1016/j.tust.2020.103383.

[51] HAJIHASSANI (M.), KALATEHJARI (R.), MARTO (A.), MOHAMAD (H.) et KHOSROTASH (M.). – *3D prediction of tunneling-induced ground movements based on a hybrid ANN and empirical methods*. Engineering with Computers 36.1, p. 251-269 (2020). DOI : 10.1007/s00366-018-00699-5.

[52] MOSTAFA (S.), SOUSA (R.L.) et EINSTEIN (H.H.). – *Toward the automation of mechanized tunneling "exploring the use of big data analytics for ground forecast in TBM tunnels"*. Tunnelling and Underground Space Technology 146:105643 (2024).

[53] LIU (L.), ZHOU (W.) et GUTIERREZ (M.). – *Effectiveness of predicting tunneling-induced ground settlements using machine learning methods with small datasets*. Journal of Rock Mechanics and Geotechnical Engineering 14.4, p. 1028-1041 (2022). DOI : 10.1016/j.jrmge.2021.08.018.

[54] RICHA (T.), PEREIRA (J.-M.), GUAYACAN-CARRILLO (L.-M.), CHAPRON (G.) et LANQUETTE (F.). – *Accuracy of Machine Learning techniques in forecasting tunnelling-induced soil settlements with limited data*. XVIII European Conference on Soil Mechanics and Geotechnical Engineering. Lisbonne, août 2024.

[55] RICHA (T.). – *Prévision des tassements induits par le creusement au tunnelier : construction d'une base de données et apprentissage automatique [Ecole des Ponts ParisTech]* (2023). https://pastel.hal.science/tel-04221306

[56] *Scikit-learn Cross-validation: evaluating estimator performance* (2011). URL: https://scikit-learn.org/1.5/modules/cross_validation.html

[57] https://scott.fortmann-roe.com/docs/BiasVariance.html

[58] LAURAE. – *xgboost : "Hi I'm Gamma. What can I do for you ?"—and the tuning of regularization* (2016). URL : https://medium.com/data-design/xgboost-hi-im-gamma-what-can-i-do-foryou-and-the-tuning-of-regularization-a42ea17e6ab6

TECHNIQUES DE L'INGÉNIEUR

L'EXPERTISE TECHNIQUE
& SCIENTIFIQUE DE RÉFÉRENCE

Gagnez du temps et sécurisez vos projets en utilisant une source actualisée et fiable !

15 DOMAINES D'EXPERTISE

- ✓ Automatique - Robotique
- ✓ Biomédical - Pharma
- ✓ Construction et travaux publics
- ✓ Électronique - Photonique
- ✓ Énergies
- ✓ Environnement - Sécurité
- ✓ Génie industriel
- ✓ Ingénierie des transports
- ✓ Innovation
- ✓ Matériaux
- ✓ Mécanique
- ✓ Mesures - Analyses
- ✓ Procédés chimie - bio - agro
- ✓ Sciences fondamentales
- ✓ Technologies de l'information

Articles de référence disponibles en français et en anglais.

Détails des offres et sommaires à retrouver sur le site

www.techniques-ingenieur.fr

Les offres Techniques de l'Ingénieur permettent d'accéder à une **base complète et actualisée d'articles** rédigés par les meilleurs experts et validés par des comités scientifiques, avec :

- + de 10 000 articles de référence et 1 000 fiches pratiques opérationnelles.
- + de 550 bases documentaires, + de 30 Parcours Pratiques répartis dans plus de 90 offres.
- Service de Questions aux experts.
- 3 000 quiz dans + de 1 000 articles interactifs.
- 1 280 auteurs contribuent chaque année à enrichir cette ressource.
- Les Archives, technologies anciennes et versions antérieures des articles.

+ de 300 000 utilisateurs de techniques-ingenieur.fr chaque mois !

NOS ÉQUIPES SONT À VOTRE DISPOSITION

Par téléphone
☎ 33 (0)1 53 35 20 20

Par email
infos.clients@teching.com