DATAE 4D

1. Introduction

L'industrie de la construction est ancrée depuis bien longtemps dans nos vies, que ce soit pour le transport avec les ouvrages routiers, l'énergie avec les infrastructures nucléaires ou bien nos habitations. Pour maintenir notre sécurité tout en réduisant notre impact environnemental, la pérennisation des ouvrages est plus importante que jamais.

A titre informatif, une étude menée par Oxford Economics estime que le marché mondial de la construction passerait de 10,7 trillions de dollars en 2020 à 15,2 trillions de dollars en 2030. La demande croissante en construction doit être accompagnée d'un suivi et d'une attention particulière à la durabilité, afin que ces ouvrages restent utilisables le plus longtemps possible et minimisent ainsi leurs émissions de gaz à effet de serre (GES) sur toute leur durée de vie.

Cependant, les émissions de GES générées par l'industrie de la construction restent très importantes. Selon un rapport datant de 2022 du Programme des Nations Unies pour l'environnement (PNUE), le secteur de la construction représente 37 % des émissions mondiales de CO_2. Cela représente près de 10 Gt de CO_2 uniquement dû au domaine de la construction, en comprenant la fabrication et les travaux de réparation des ouvrages.

L'ingénierie de la maintenance est une discipline en plein essor représentant aujourd'hui l'un des moyens les plus efficaces pour optimiser la durabilité des ouvrages, en misant sur le suivi préventif de l'état des structures et des matériaux constitutifs au dépend d'une approche curative.

En identifiant et en traitant rapidement les signes avant-coureurs de détérioration, il est possible d'éviter des réparations coûteuses à long terme (pouvant aller jusqu'à la démolition/reconstruction). Le suivi permet de cibler et de quantifier judicieusement les zones d'éventuelles réparations et de hiérarchiser les travaux de maintenance en fonction des budgets disponibles. Ceci afin d'optimiser la durée de vie des ouvrages tout en maximisant la sécurité des usagers et en limitant la production de GES.

Il s'avère que le suivi des ouvrages génèrent d'importants volumes de données multi-échelles et multivariées susceptibles d'évoluer dans l'espace et dans le temps. L'analyse conjointe de ces données est cruciale pour aider à optimiser et à fiabiliser le suivi des structures. Cependant, l'intégration de l'ensemble de ces données peut s'avérer extrêmement complexe, les relations entre les variables disponibles ne présentant pas forcément de liens intuitifs et directs.

La démocratisation récente des méthodes d'intelligence artificielles (IA), permet aujourd'hui d'envisager la valorisation de l'ensemble des données disponibles afin d'améliorer le suivi de l'état des ouvrages et des matériaux constitutifs.

2. Le projet

1. L'objectif

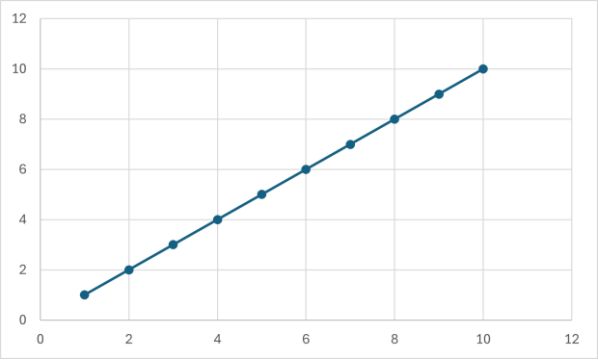
L'objectif de ce projet est d'optimiser les actions à mener pour prévenir et agir de manière efficace et juste durant le suivi de la vie des ouvrages.

Parmi les nombreuses méthodes d'IA, les méthodes d'apprentissage automatique (machine learning) sont particulièrement intéressantes car elles permettent de modéliser tout type de données à partir d'une base de données d'apprentissage et d'un jeu de données d'entrainement.

Dans ce projet, il est envisagé d'utiliser une méthode d'IA relativement simple à implémenter. Il s'agit de l'apprentissage supervisé par arbre de décision *(voir extrait de ma thèse pour info en fin de document)*. Cette méthode sera utilisée comme un interpolateur de données faiblement échantillonnées (variables primaires acquises à faible rendement) dans le champ de données densément échantillonnées (variables secondaires acquises à grand rendement). Ceci afin de produire des cartographies de propriétés physico-chimiques utiles pour améliorer le suivi et la modélisation de nos ouvrages.

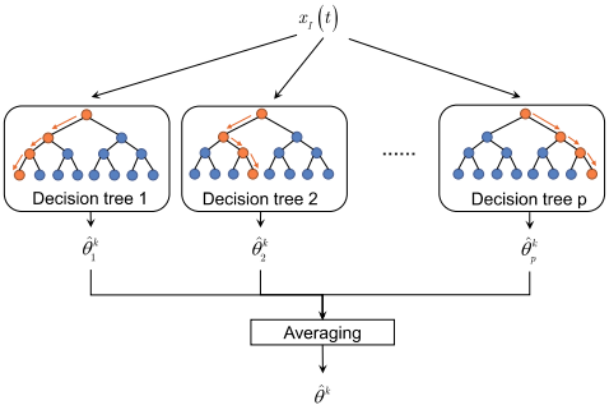
2. Principe (*voir extrait de ma thèse en fin de document*)

L'arbre de décision est un principe plutôt simple et ancien, qui permet de comparer des variables initiales à une base de données conçue spécialement pour la problématique étudiée. Prenons l'exemple d'une droite qui est créée à partir de plusieurs points.



Nous connaissons uniquement les points bleus présents sur la courbe, pourtant, dans notre imagination, nous modélisons les points qui sont présents entre les données connues (régression linéaire). Nous venons alors de mettre en application le principe de

l'arbre de décision. Sous une autre forme, l'arbre de décision peut aussi être visualisé de cette manière :



Ici, on remarque le principe de fonctionnement de cette méthode. Plusieurs décisions de ''résultat'' sont établies par le programme d'analyse, puis une valeur est extraite de tous les arbres de décision. Cette dernière valeur est considérée comme celle modalisée par le programme à partir de la base de données. Pour mieux comprendre l'intérêt de cette méthode, nous allons voir une mise en application dans le cas d'une problématique.

3. Mise en application

Lors du diagnostic d'un ouvrage, il est fréquent de combiner l'acquisition de données (dites « variables primaires ») obtenues ponctuellement (carottages et analyses en laboratoire associées, CND à faible rendement, données de monitoring, etc.) avec la réalisation de contrôles non destructifs (CND) permettant de cartographier certaines propriétés physiques (dites « variables secondaires ») à grand rendement (enrobage, potentiel de corrosion, etc.).

Bien qu'apportant des informations pertinentes, les variables primaires sont généralement longues à acquérir (réalisation d'un carottage et analyses en laboratoire, polarisation d'une armature pour estimation de la vitesse de corrosion, etc.). Lors d'un diagnostic, ces données ponctuelles ne peuvent raisonnablement pas être réalisées sur l'ensemble d'une structure pour des raisons financières et logistiques.

A partir d'un grand nombre de données colocalisées acquises sur un grand nombre d'ouvrages, il apparaît envisageable d'estimer par régression, les variables primaires dans le champ des variables secondaires en utilisant des algorithmes d'apprentissage supervisé.

. A titre d'exemple, La figure 1 représente la distribution spatiale des valeurs de vitesse de corrosion estimées par régression, à l'aide d'une méthode d'apprentissage supervisé par arbre de décision

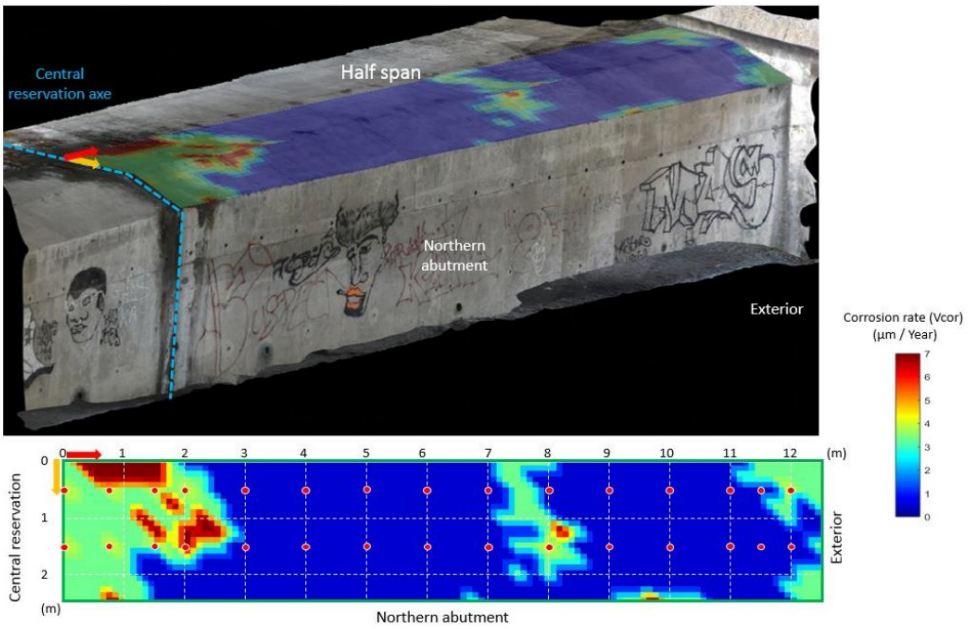


Figure 1 : Exemple de cartographie de vitesse de corrosion obtenue par apprentissage supervisé (extrait de Anterrieu et al. 2019) – Haut : report sur modèle photogrammétrique / Bas : vue en plan

La base de données d'apprentissage utilisée dans cet exemple est composée de quelques centaines de données colocalisés acquises sur différents ouvrages (épaisseur d'enrobage, indice d'humidité relative, potentiel de corrosion vitesse de corrosion, profondeur du front de carbonatation, profondeur de pénétration des ions chlorure, etc.). Les données d'entrainement correspondent à des mesures CND (variables secondaires).

A partir des données d'apprentissage et des données d'entrainement, l'analyse par arbre de décision a permis de modéliser les vitesses de corrosion en tout point du champ des variables secondaires et ainsi d'estimer les zones de vigilance prioritaire au risque de corrosion sur la portion d'ouvrage étudiée.

Afin d'évaluer la fiabilité de l'approche, des mesures de vitesse de corrosion ponctuelles ont été réalisées *a posteriori* (cf. 30 points rouges sur la cartographie) et ont été comparées aux valeurs estimées par apprentissage automatique.

La figure 2 présente le comparatif des données acquises et des données estimées par arbre de décision.

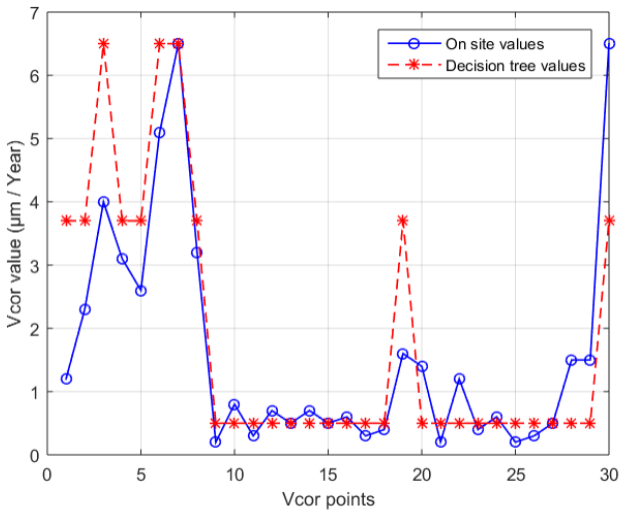


Figure 2 : comparaison des données de vitesse de corrosion mesurées sur site (30 points) et des données estimées par arbre de décision

La figure 2 témoigne de la cohérence entre les données mesurées et celles estimées par apprentissage automatique, les ordres de grandeur étant similaires.

Cet exemple permet d'illustrer le potentiel très prometteur de cette approche malgré la faible quantité de données d'apprentissage.

Ce projet a donc pour but de nous faire imaginer bien plus grand avec de la modélisation de plusieurs propriétés des bétons. L'objectif est maintenant de récolter des données matérielles colocalisées sur site et issues des résultats de laboratoire, pour modéliser un maximum de propriétés. La porosité, la densité, la résistance en compression, le rapport e/c, l'âge de l'ouvrage, le type de ciment, toutes ces propriétés aideront à modéliser des vitesses de corrosion, des fronts de carbonatation et d'ion chlorure, etc. A terme, une autre piste de développement pourrait être de modéliser l'évolution de ces propriétés dans le temps. En associant une variable temporelle aux données mesurées, nous pourrions potentiellement modéliser l'évolution de certains paramètres en lien avec la durabilité, de manière diachronique, ce qui deviendrait un atout majeur dans le cadre de l'étude, du diagnostic et de la maintenance des ouvrages. Ainsi, la maintenance serait plus précise, moins couteuse et les réparations seraient plus ciblées, ce qui à terme permettrait de maximiser la durée de vie tout en réduisant l'impact écologique de l'entretien des ouvrages.

3. Conclusion

Sachant qu'aujourd'hui les pathologies sont un réel problème pour la durée de vie des ouvrages, il est important de pouvoir modéliser au mieux les propriétés qui caractérisent ces pathologies. Ce projet permettra de contenir toutes les données qui décrivent les

structures qui nous entourent et de comprendre au mieux tous les bâtiments pour les suivre et les maintenir debout le plus longtemps possible. D'autre part, grâce à cette méthode, il serait aussi possible d'intégrer les « nouveaux bétons » qui arriveront sur le marché dans quelques années. Car même sans comprendre totalement les principes mis en relation à l'échelle atomique, nous pouvons trouver des liens et comprendre comment influent les éléments entre eux grâce à la modélisation et aux arbres de décisions !

Extrait de ma thèse sur la partie « Apprentissage automatique » dont tu peux t'inspirer pour évoquer les arbres de décision.

Apprentissage automatique

Une famille de techniques d'intégration de données très populaire de nos jours est l'apprentissage automatique ou *machine learning* en anglais. Ces techniques permettent de déterminer des relations statistiques entre une variable primaire et une ou des variables secondaires sans aucune contrainte sur les variables et leurs statistiques, à partir d'un ensemble fini de données représentatives. Les techniques présentées dans ce paragraphe sont tirées des travaux de maîtrise de Schnitzler (2017).

- Méthodes non supervisées

Dans le domaine informatique, l'apprentissage non supervisé construit un modèle qui représente les données collectées (Fischer 2014). Cette méthode se distingue de l'apprentissage supervisé par le fait qu'il n'y a pas de données colocalisées du paramètre qu'on cherche à retrouver pour permettre à l'algorithme à calibrer ses paramètres. L'analyse se fait donc uniquement à partir des données indirectes. Un groupe hétérogène de données est divisés en sous-groupes, afin que les données considérées comme les plus similaires soient associées à un groupe homogène et qu'au contraire les données considérées comme différentes se retrouvent dans d'autres groupes distincts.

- Méthodes supervisées

L'apprentissage supervisé est une technique d'apprentissage automatique permettant de déterminer une nouvelle sortie à partir d'une entrée (variable) connaissant un ensemble de données indirectes observées et le paramètre à estimer, colocalisé à quelques endroits avec les données indirectes (Fischer 2014). En d'autres termes,

l'objectif de ces méthodes est de produire automatiquement des règles à partir d'une base de données d'apprentissage contenant des « exemples ». Parmi les méthodes supervisées, il est notamment possible de distinguer les machines à vecteurs de support (*Support Vecteur Machine* ou SVM) et les arbres de décision (*decision tree*).

- SVM

Le SVM est une méthode de séparation par hyperplans de nuages de points en maximisant la distance entre l'hyperplan et les échantillons. Cette méthode est robuste et fonctionne très bien dans le cas linéaire. Dans le cas non linéaire, les échantillons sont projetés dans un espace de plus grande dimension dans lequel il existe probablement des hyperplans optimaux. Comme la résolution de ce problème dans un formalisme mathématique strict est très coûteux en temps de calcul, l'astuce des noyaux a été développée. Cette technique permet de faire le traitement dans l'espace d'origine et permet de faire des transformation complexes. Les inconvénients relatifs à l'utilisation de cette méthode sont la difficulté d'identifier les bonnes valeurs de paramètres, la difficulté de traiter les grandes bases ayant un nombre d'observation élevé et la difficulté d'interpréter le modèle, notamment par manque d'identification de la pertinence des variables.

- Méthode d'ensemble, la forêt aléatoire ou *Random Forest*

L'algorithme des forêts aléatoires (*Random Forest* ou RF) développé par Breiman en 2001 permet d'élaborer un modèle prédictif de données, à partir de l'analyse par arbres décisionnels de nombreuses variables (Breiman 2001). Cet algorithme fait partie des méthodes supervisées de type arbre de décision et basé sur le concept « d'aléatoire » (« randomisation ») (Heutte et al. 2008). Le RF combine le *bagging* (*boostrap and aggregating*) et le concept de sous-espaces aléatoires. Bien que souvent associé aux arbres décisionnels, il peut être utilisé dans d'autres types de classificateurs. Essentiellement, cet algorithme découpe l'ensemble d'entraînement en sous ensemble de petites tailles. Un arbre décisionnel est créé et les données sont remises dans l'ensemble initial. L'opération est répétée un grand nombre de fois, et à la dernière étape, les arbres sont combinés afin de déterminer l'arbre optimum. Il offre une grande flexibilité dans le nombre de variables (paramètres mesurés), leur nature et le nombre d'occurrence (échantillons). Cette méthode basée sur l'intelligence artificielle, fonctionne sur le principe d'apprentissage automatique (*machine learning*).

Cette méthode d'ensemble est de plus en plus utilisée dans de nombreux domaines, notamment en exploration minière, afin de traiter de grands ensembles de données (Carranza et Laborte 2015). N'exigeant pas de grosses ressources informatiques, ni de données ayant des attributs spécifiques, les résultats obtenus sont souvent meilleurs que d'autres techniques d'apprentissages automatiques tels que les « réseaux de neurones » ou les SVM (Rodriguez-Galiano et al. 2015).

Le RF peut être utilisé pour la classification (prévoir une variable dépendante catégorielle) ou la régression (prévoir une variable dépendante continue). L'algorithme fonctionne sur le principe de forêt, constituée d'un ensemble d'arbres de prévision simples, chacun étant capable de produire une réponse lorsqu'on lui présente un sous-ensemble de variables explicatives ou prédictives. Dans le cas d'une classification, la réponse prend la forme d'une classe qui associe un ensemble (classe) de valeurs indépendantes (prédicteur) à une des catégories présente dans la variable dépendante. Pour chaque arbre (dont le nombre est arbitraire), un vote est réalisé pour la classe la plus populaire. Dans le cas d'une régression, un arbre est une estimation de la variable dépendante en fonction des prédicteurs. Les réponses de chaque arbre sont moyennées afin d'obtenir une estimation de la variable dépendante. En utilisant des ensembles d'arbres, il est possible d'améliorer significativement la prévision, donc avoir une meilleure capacité à prévoir de nouvelles données.

Le nombre d'arbres de décision qu'il faut utiliser pour construire une forêt n'est pas défini, ni documenté. Breiman (2001), puis Latinne et al. (2001) et Bernard et al. (2007) démontrent qu'au-delà d'un certain nombre d'arbres, en ajouter d'autres ne permet pas systématiquement d'améliorer les performances de l'ensemble. De même, le nombre d'arbres ne doit pas nécessairement être le plus grand possible pour produire une prédiction ou un classificateur performant.

La variable à prédire est calculée par un algorithme RF de régression, à partir de deux ensembles de données (1) l'ensemble d'entrainement (*training set*) et (2) l'ensemble de validation (*validation set*). Le premier est un sous-ensemble du jeu de données total, où l'ensemble des variables (mesurées et à prédire) sont connues et utilisés par l'algorithme pour définir le lien statistique entre la variable à prédire et les variables secondaires. La seconde comprend les données sur lesquelles la prédiction est faite afin de valider le pouvoir prédictif du RF entraîné. La procédure de randomisation est importante pour l'efficacité du RF. Elle se traduit par la fonction de « *bootstrap*

aggregation ». Chaque arbre de décision traite un sous-ensemble choisi au hasard dans l'ensemble d'entrainement (environ 2/3) et utilise les données restantes (1/3) pour évaluer la précision de la prévision (Breiman 1996). L'utilisation de cette fonction permet de diminuer la variance des données, améliorant ainsi la prévisibilité de la variable.

Il devient ainsi possible, à partir de nouvelles données de prédiction acquises sur un nouveau site, de déterminer les valeurs de la variable de réponse à partir de l'arbre de décision construit au préalable. La base de données utilisée dans le cadre de cette étude comprend une dizaine de site sur lesquels des mesures de CND, de vitesse de corrosion et de carbonatation ont été réalisées, et constitue un formidable ensemble d'entrainement.