Titre: Title:	Méthodologie de regroupement interprétable des profils de dégradation de systèmes en exploitation
Auteur: Author:	Mohamed Ben Slimene
Date:	2021
Type:	Mémoire ou thèse / Dissertation or Thesis
Référence: Citation:	Ben Slimene, M. (2021). Méthodologie de regroupement interprétable des profils de dégradation de systèmes en exploitation [Mémoire de maîtrise, Polytechnique Montréal]. PolyPublie. https://publications.polymtl.ca/9904/

URL de PolyPublie: PolyPublie URL:	https://publications.polymtl.ca/9904/
Directeurs de recherche: Advisors:	Mohamed-Salah Ouali
Programme: Program:	Maîtrise recherche en génie industriel

POLYTECHNIQUE MONTRÉAL

affiliée à l'Université de Montréal

Méthodologie de regroupement interprétable des profils de dégradation de systèmes en exploitation

MOHAMED BEN SLIMENE

Département de mathématiques et de génie industriel

Mémoire présenté en vue de l'obtention du diplôme de *Maîtrise ès sciences appliquées*

Génie industriel

Décembre 2021

POLYTECHNIQUE MONTRÉAL

affiliée à l'Université de Montréal

Ce mémoire intitulé :

Méthodologie de regroupement interprétable des profils de dégradation de systèmes en exploitation

Mohamed BEN SLIMENE

en vue de l'obtention du diplôme de *Maîtrise ès sciences appliquées*

a été dûment accepté par le jury d'examen constitué de :

Luc ADJENGUE, président
Mohamed-Salah OUALI, membre et directeur de recherche
Hugo GAGNON, membre

REMERCIEMENTS

C’est avec un grand plaisir que je réserve cette page en signe de gratitude et de profonde reconnaissance à tous ceux qui ont contribués à la réalisation de ce travail.

Je tiens à présenter mes vifs gratitudes et remerciements à mon directeur de recherche, Monsieur Mohamed-Salah OUALI pour ses efforts, sa disponibilité et surtout ses judicieux conseils qui ont contribués à alimenter ma réflexion.

De même, j’adresse tous mes remerciements aux membres du jury Monsieur Luc ADJENGUE et Monsieur Hugo GAGNON pour m’avoir honoré en acceptant d’évaluer mon travail.

Enfin, je tiens à remercier ma famille, mes amis et tous ceux qui m'ont soutenu pour réaliser ce travail.

RÉSUMÉ

La modélisation de la dégradation lente des actifs en exploitation à partir des séries chronologiques d'observations est une problématique complexe. Les approches de développement de ces modèles peuvent varier selon la structure de la base de données disponibles ainsi que la nature des systèmes étudiés. Dans le présent travail, nous proposons une méthodologie qui permet d'extraire et d'interpréter les profils de dégradation à partir de deux différents types de base de séries chronologique : Les séquences courtes, bruitées, de différentes longueurs et fréquences d'indicateurs clé de performance; et des signaux cycliques générées par les capteurs installés sur les composantes des actifs lors de leur fonctionnement.

La méthodologie proposée est composée de trois phases. La première phase, les profils de dégradation sont extraits en regroupant les séries de comportement similaire. Pour ce faire, l'algorithme de clustering DBSCAN couplé avec une mesure de similarité de pentes pour les séries d'indicateur clé de performance (ICP) et la mesure de dissimilarité Dynamic Time Warping DTW pour les signaux cycliques est appliqué. Afin de renforcer l'évaluation des partitions basée principalement sur le score de silhouette, deux versions modifiées des indices Davies-Bouldin et C pour interpréter des regroupements basés sur une logique de similarité sont développées. Dans la deuxième phase de la méthodologie, deux différentes variantes de l'algorithme de forêts aléatoires à prédire les profils de dégradation identifié précédemment sont développées. Finalement, la dernière phase concerne l'interprétation des modèles ajustés. Pour les séries de ICP, la contribution des facteurs conceptuels, climatiques et d'exploitation dans l'extraction des profils de détérioration en évaluant l'importance de ces variables dans l'entrainement du classificateur forêts aléatoires est réalisée. Ce dernier sera utilisé pour prédire pas à pas l'ICP étudié. En ce qui concerne les signaux cycliques, un score d'importance qui permet de mesurer la contribution de chaque dimension des signaux dans l'identification des profils de dégradation est déterminé. La liaison entre les clusters et les ICPs des cycles est élaborée à l'aide de règles générées par la technique d'arbre de décision. La validation de la méthodologie sera supportée par deux études de cas qui traitent les deux structures de base de données abordées dans ce travail.

ABSTRACT

Modeling the slow degradation of operating assets based on time series observations is a complex issue. Approaches to developing these models can vary depending on the structure of the available database as well as the nature of the systems being studied. In the present work, we propose a methodology to extract and interpret degradation profiles from two different types of time series databases: short, noisy sequences of different lengths and frequencies of key performance indicators; and cyclic signals generated by sensors installed on asset components during their operation.

The proposed methodology is composed of three phases. In the first phase, degradation profiles are extracted by clustering series with similar behavior. For this purpose, the DBSCAN clustering algorithm coupled with a slope similarity measure for KPI series and the Dynamic Time Warping DTW dissimilarity measure for cyclic signals is applied. In order to strengthen the evaluation of partitions based mainly on the silhouette score, two modified versions of the Davies-Bouldin and C-indices for interpreting clustering based on similarity logic are developed. In the second phase of the methodology, two different variants of the random forest algorithm to predict the previously identified degradation profiles are developed. Finally, the last phase concerns the interpretation of the fitted models. For the KPI series, the contribution of conceptual, climatic, and operational factors in the extraction of deterioration profiles by evaluating the importance of these variables in the training of the random forest classifier is performed. The latter will be used to perform a stepwise prediction of the studied KPI. For the cyclic signals, an importance score that measures the contribution of each signal dimension in the identification of degradation profiles is determined. The liaison between the clusters and the cycle KPIs is elaborated using rules generated by the decision tree technique. The validation of the methodology will be supported by two case studies that address the two database structures addressed in this work.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES SIGLES ET ABRÉVIATIONS

DTW	Dynamic time warping
ICP	Indicateur clé de performance
IA	Intelligence artificielle
DB	Indice Davies-Bouldin
DBSCAN	Density Based Clustering of Applications with Noise
RNA	Réseaux de neurones artificiels
RNR	Réseaux de neurones récurrents
SVM	Support Vector Machine
ACP	Analyse en composantes principales
STSD	Short time-series distance
LCSS	Plus longue sous-série commune
MSE	Mean squared error
LCS	Longest common subsequence
CART	Classification and regression tree
TSF	Time series forest
IEP	Indice d’état des ponceaux

LISTE DES ANNEXES

CHAPITRE 1 INTRODUCTION

À l'ère industrielle 4.0, l'un des principaux domaines d'application de l'intelligence artificielle (IA) en pleine expansion est celui de la maintenance des systèmes industriels. Cette dernière représente un coût important pour les entreprises. Qu'il s'agisse d'équipements, de machines, d'avions ou d'infrastructures, ils entrainent une augmentation importante des coûts d'exploitation et des risques d'accident lorsque ces derniers présentent des défaillances ou une perte de performance due à la dégradation. C'est pourquoi les gestionnaires de systèmes industriels développent et utilisent des indicateurs clés de performance (ICP) comme mesures standards pour évaluer l'état de performance de leurs systèmes au cours de leurs cycles de vie (Kumar, J. et al., 2013). Le diagnostic et la prévision des modes de défaillance sont essentiels pour la prise de décision en matière de maintenance préventive, en particulier la maintenance conditionnelle, voire prédictive.

La modélisation de la dégradation des équipements à l'aide de données historiques représente un attrait de plus en plus important pour établir et optimiser les interventions de maintenance tout au long du cycle de vie de systèmes. Selon (Yin et al., 2014), les méthodes basées sur l'exploration des données sont les plus appropriées pour détecter les défauts et incidents de fonctionnement et pour prévenir leur apparition. Les systèmes industriels contemporains sont complexes du fait de la multitude des technologies utilisées et exploitées dans des conditions de fonctionnement souvent différentes. Pour y arriver, la mise en place de moyens de collecte des données temporelles, soit par des inspections périodiques, soit l'installation de capteurs sur certains composants critiques du système constitue la solution envisageable par les gestionnaires de maintenance. Les données ainsi construites ou collectées au cours du temps prennent la forme, en quelque sorte, de séries chronologiques pouvant servir à construire des indicateurs clés de performance ou des variables statistiques reflétant l'état ou la condition du système sous inspection ou surveillance. Comme toute information collectée par le service de maintenance, les données peuvent être biaisées par le bruit, les défauts du système, les facteurs externes et les perturbations de fonctionnement (Yang, X. et al., 2019). Cadei et al. (2019) dressent un portrait des principales limites rencontrées lors de l'exploitation des données de séries chronologiques pour extraire des ICPs, et ce pour des actifs industriels ou des infrastructures civiles. Ces limites peuvent être attribuées à la lente dégradation

observée pour certains mécanismes tels que la corrosion, la fatigue, l'usure, la déformation, etc., et à des méthodes de mesure peu fiables, partielles ou incomplètes. Ainsi, la lenteur de certains mécanismes de dégradation, les longues durées de vie utile, les changements des conditions d'exploitation, la planification et l'efficacité de la maintenance sont les principales causes de la rareté des observations pertinentes et intègres. Il en résulte une suite d'observations très espacées dans le temps avec des mesures incomplètes et entachées d'erreurs. Ainsi, nous pouvons observer de courtes séries temporelles, discontinues, décalées dans le temps et bruitées par les interventions de maintenance.

Selon (Pereira & Silveira, 2018), le traitement des séries temporelles pour représenter l'état de performance des actifs à lente dégradation est une pratique courante lors de l'étude d'un parc d'un même actif avec différents âges et conditions d'exploitation. Cependant, Cerquitelli et al. (2021) ont indiqué que plusieurs équipements et procédés industriels sont concernés par les effets de mécanismes de dégradation lente, où la durée d'un cycle de fonctionnement ne permet pas d'accumuler une dégradation détectable du système. Par conséquent, la prédiction de l'indicateur de l'état de performance d'un cycle spécifique n'intéresse pas les experts du domaine, alors que l'accent est mis sur de nombreux cycles pour diagnostiquer et caractériser les modes de dégradation du système étudié.

Les industriels disposent généralement de bases de données vastes (Big Data) générées par les capteurs installés sur les composantes des systèmes étudiés. La surveillance automatisée de ces composants donne lieu à des données rapides et volumineuses de signaux relevés sur plusieurs cycles de fonctionnement. Pour ce genre de données temporelles, telles les données de vibration, les principaux défis pour les ingénieurs consistent à identifier les profils comportementaux et les corrélations entre les variables d'observation dans le but d'identifier et caractériser les changements d'une série de mesures à une autre selon le mode opératoire choisi (Suschnigg et al., 2020).

Dans le présent travail, nous nous concentrerons sur le traitement d'un ensemble de données de séries temporelles caractérisant les mécanismes à lente dégradation. Plus particulièrement, nous intéressons à deux types de séries chronologiques : les séries à courtes séquences d'observations bruitées et décalées dans le temps et les séries de signaux générées par les mesures de capteurs relatifs à un mode opératoire cyclique de systèmes identiques. Les bases de données relatives à plusieurs cycles opératoires sont plus volumineuses et renferment le comportement inhérent du

système que celles des séries à courtes séquences. Toutefois, elles sont plus compliquées à traiter. Les figures 1.1 et 1.2 illustrent, respectivement, la structure des données à courte séquence d'observations par actif au cours de l'âge où O_{ij} représente l'observation de l'actif A_i au moment t_j et la structure des données organisées par cycle opératoire au cours du temps où $E_{hi_{jk}}$ représente l'amplitude du signal S_j du cycle C_i de la machine h au moment t_k. Ces sont les deux structures étudiées dans le présent travail.

Actif \ Age	t_1	t_2	...	t_{a-1}	t_a
A_1	O_{11}	O_{12}	...		
A_2		O_{21}	...		
⋮			...		
A_{n-1}			...	$O_{n-1\,a-1}$	$O_{n-1\,a}$
A_n			...	$O_{n\,a-1}$	

Figure 1.1 Données par séquence d'observations

ID cycle	ID machine	Temps	S_1	S_2	...	S_S
Cycle 1	M_1	t_1	$E_{11_{11}}$	$E_{12_{11}}$	...	$E_{1s_{11}}$
Cycle 1	M_1	t_2	$E_{11_{12}}$	$E_{12_{12}}$	...	$E_{1s_{12}}$
⋮	⋮	⋮	...	...	...	...
Cycle n	M_m	t_{w-1}	$E_{m1_{n\,w-1}}$	$E_{m2_{n\,w-1}}$	...	$E_{ms_{n\,w-1}}$
Cycle n	M_m	t_w	$E_{m1_{n\,w}}$	$E_{m2_{n\,w}}$	...	$E_{ms_{n\,w}}$

Figure 1.2 Données par cycle opératoire

Grâce à leur capacité à reconnaitre les structures et les corrélations compliquées à détecter et à mesurer, les techniques d'apprentissage sont de plus en plus utilisées dans la modélisation de la dégradation des systèmes industriels. Ces techniques opèrent selon deux approches : l'apprentissage supervisé qui vise à prédire une variable cible comme l'ICP d'un actif à partir des données explicatives, et l'approche non supervisée, notamment le clustering qui cherche à extraire et interpréter les regroupements en se basant sur des données non étiquetées.

En général, la construction des modèles de dégradation basés sur des observations limitées en nombre est abordée comme un problème d'apprentissage automatique supervisé. Le clustering consiste à extraire des regroupements basés sur la similarité ou les distances entre les séries chronologiques. Cette approche vise à extraire des profils de dégradation interprétables. Selon (Agrawal et al., 2010), l'utilisation de la mesure de similarité pour comparer des séries temporelles avec des fréquences et des longueurs d'échantillonnage différentes peut permettre une meilleure partition des données. Cependant, la plupart des indices de validation de clustering interne tels que : le score de Silhouette, l'indice de Davies-Bouldin (DB) et l'indice C sont construits pour s'adapter aux algorithmes basés sur des mesures de dissimilarité (Hämäläinen et al., 2017). Or, dans le cas où nous traitons des séries courtes et de différentes fréquences, une mesure de similarité aura une meilleure interprétabilité. Ainsi, les indices DB et C doivent être modifiés.

L'objectif de ce travail est d'extraire les profils de dégradation et les interpréter physiquement en analysant les dépendances entre les indicateurs de performance et les conditions d'opération du système. Pour le cas des données représentant de courtes séries d'observations, nous visons à utiliser les profils de dégradation pour construire un modèle de prévision de performance du système. Pour le cas des données représentant des cycles d'opération, nous utilisons les profils extraits pour faire le lien entre la dégradation des indicateurs de performance et les symptômes portées par les signaux. Ensuite, nous exploitons les séries générées par les capteurs pour prédire les ICPs futures de chacun des actifs étudiés.

Dans le présent projet, nous travaillons sur le développement d'une méthodologie de regroupement interprétable de profils de dégradation de systèmes industriels en cours d'exploitation à partir d'un ensemble de séries chronologiques. Plus précisément, nous visons à extraire et caractériser les profils de dégradation d'une flotte de systèmes identiques en se basant des séquences d'observations courtes de leurs états de performance ou sur des séries temporelles caractérisant plusieurs cycles opératoires. Les profils de dégradation doivent être interprétables par rapport aux conditions d'exploitation. Pour ce faire, nous extrairons d'abord des profils de dégradation en regroupant de courtes séries temporelles à l'aide d'un algorithme de clustering basé sur la densité, communément connu par DBSCAN pour Density Based Clustering of Applications with Noise (Valarmathy & Krishnaveni, 2020), utilisant deux mesures de proximité différentes. La partition sera validée en utilisant le score de silhouette, les indices DB et C. La technique de classification

de Forêts aléatoires (Random Forests) sera utilisée pour prédire le profil de dégradation d'un système donné ainsi que son indicateur sa performance au cours du temps.

La suite du mémoire est organisée en quatre chapitres. Le Chapitre 2 présente une revue de littérature des approches de modélisation de la dégradation de systèmes industriels ainsi que les techniques de clustering de séries temporelles et leurs de validité. Le Chapitre 3 introduit et détaille la méthodologie proposée de construction de modèles de prédiction de l'état de systèmes industriel ainsi que l'adaptation des indices de validité du regroupement, initialement proposés comme mesures de dissimilarité, à la similarité. Le Chapitre 4 décrit et applique la méthodologie proposée à deux études de cas. Le premier cas concerne la modélisation de la dégradation d'un parc de ponceaux de drainage. Cette étude représente un cas typique de système à lente dégradation au cours du temps dont l'état est représenté par un indice de performance calculé à partir de données d'inspection périodique. La seconde étude concerne un système hydraulique muni de capteurs. Ce système représente un cas typique de systèmes opérant selon un cycle répétitif. Les données collectées pour ce dernier sont des séries temporelles multivariées. Le Chapitre 5 présente les conclusions et les perspectives de recherche.

CHAPITRE 2 REVUE DE LITTÉRATURE

La revue de littérature est scindée en deux parties : les approches de modélisation de la dégradation et les méthodes de regroupement des données séquentielles d'observations, soient des séquences d'observations, soient des séries temporelles.

2.1 Approches de modélisation de la dégradation

Dans la littérature, il existe principalement trois approches pour construire des modèles de dégradation lente : déterministe ou physique, stochastique et d'apprentissage supervisé.

2.1.1 Approches déterministes

Les méthodes déterministes décrivent une relation mathématique entre les variables d'entrée et de sortie d'un système dans lequel une bonne corrélation peut être déduite des variables. Ces méthodes cherchent à établir des relations entre la performance des systèmes et le temps pour prédire l'état de dégradation avec certitude. Elles peuvent être empiriques ou fondées sur l'opinion des experts. La régression linéaire multiple est l'une des méthodes déterministes les plus faciles à interpréter. Elle est utilisée lorsque plus d'une variable explicative peut influencer la variable dépendante de sortie.

Les méthodes déterministes de prédiction de la détérioration peuvent être de deux types : linéaire ou non linéaire (Ens, 2012). Des modèles linéaires de loi de temps et de puissance ont été appliqués pour estimer la détérioration des conduites d'aqueduc (Rajani & Kleiner, 2001) et des chaussées (Lou et al., 2001). Alors que des modèles déterministes exponentiels ont été appliqués aux ponceaux de drainage dans les travaux de (Mishalani & Madanat, 2002), (Morcous et al., 2002) et (Wirahadikusumah et al., 2001). Toutefois, l'approche déterministe n'est pas souvent applicable aux systèmes complexes dans lesquels une relation mathématique ne peut pas être dérivée sur la base d'un ensemble de variables imprécises et corrélées (Tran, Huu Dung, 2007). Ces modèles calculent les conditions prédites de manière déterministe en ignorant l'erreur aléatoire dans la prédiction.

Les modèles physiques donnent un aperçu des facteurs qui affectent le plus le processus de détérioration en proposant des équations physiques conviviales et interprétables. (H. D. Tran,

Perera, et Ng (2010) indiquent que ces modèles ne sont pas adaptés à la modélisation d'états discrets et que leurs hypothèses sous-jacentes peuvent être difficiles à valider.

2.1.2 Approches stochastiques

Les approches stochastiques sont basées sur l'inférence statistique pour modéliser un phénomène assujetti à une erreur aléatoire. Les modèles statistiques ont été utilisés dans de nombreux problèmes de modélisation de la détérioration de l'infrastructure. Ces approches obéissent à une démarche rigoureuse qui suppose des fonctions de densité paramétriques pour représenter les erreurs de mesure et certaines relations probabilistes entre les variables d'entrée (variables explicatives) et la variable de sortie (variable à expliquer) (Hasan, 2015).

Les modèles statistiques offrent une approche plus réaliste de la prévision de l'état actuel et futur d'un système. Car les résultats, c'est-à-dire les conditions prévisibles du système, sont explicitement formulés sous la forme de probabilités plutôt que de valeurs physiques comme dans les modèles déterministes. Le résultat peut être un choix binaire «vrai» ou «faux», des réponses à plusieurs catégories ou même une matrice de probabilités de transition (Ana, E. & Bauwens, W., 2010).

L'incertitude et le caractère aléatoire de la détérioration des systèmes sont considérés comme une ou plusieurs variables aléatoires dans les modèles stochastiques. Ces techniques comprennent les modèles de Markov, les processus Gamma et de puissance qui ont été largement utilisés pour modéliser la détérioration des systèmes à dégradation lente (Yang, X. et al., 2019).

- ***Modèles markoviens***

Les travaux de (Micevski et al., 2002) et (Tran, H. D. et al., 2010) ont utilisés le processus de Markov pour décrire la détérioration des conduites d'eaux pluviales et d'eaux usées. Ce modèle a été appliqué dans la modélisation de la dégradation des ponts dans plusieurs études telles que celles de (Wellalage et al., 2015) et (Thomas & Sobanjo, 2016). Les travaux de (Kobayashi et al., 2010) et (Thomas & Sobanjo, 2013) ont également souligné la facilité de l'application des modèles markoviens dans le cas de la dégradation de la chaussée.

Toutefois, les modèles markoviens nécessitent une calibration initiale de la matrice de transition. Parmi les méthodes de calibration figure la méthode d'optimisation basée sur la régression utilisée dans la prédiction de la dégradation du système de drainage (Ranjith et al., 2013); la méthode de

prévision en pourcentage utilisé pour estimer les risques et les coûts engendrés par chaque politique de maintenance des ponts de New York (DeStefano & Grivas, 1998) ; la Méthode d'optimisation non linéaire utilisée pour les conduites d'eaux pluviales (Tran, Huu Dung, 2007); et enfin la méthode de Monte-Carlo par chaînes de Markov bayésienne (MCMC) utilisée dans la prédiction de la détérioration des conduites d'eaux pluviales dans les travaux de (Micevski et al., 2002) et (Brooks et al., 2011).

- ***Processus Gamma***

Comme la détérioration est généralement incertaine et non décroissante, le processus Gamma est couramment utilisé pour tirer des estimations de fiabilité à partir de données de dégradation (Wang et al., 2021). Ce processus permet de développer modéliser la dégradation monotone dont les incréments sont non négatifs et suivent de manière indépendante les distributions Gamma. (van Noortwijk, J. M., 2009) a indiqué que le processus Gamma est le modèle le plus approprié pour les dommages graduels s'accumulant dans le temps tel que l'usure, la fatigue, le fluage, la croissance des fissures, l'érosion, la corrosion et la houle. Parmi les travaux, nous pouvons citer ceux de (Aboura et al., 2008) qui ont utilisé ce processus pour la prédiction de la détérioration d'un pont ; de (Edirisinghe et al., 2013) pour la prévision de la dégradation des éléments de construction ; et enfin de (Micic et al., 2016) qui modélisent la durée de vie des poutres composites en bois et béton.

- ***Processus de puissance***

Le processus de puissance exploite la distribution de Weibull dans la modélisation de la durée de vie de l'infrastructure routière tels que les modèles de (Singpurewalla & Song, 1988), (Kleiner & Rajani, 2001) et (Mishalani & Madanat, 2002). (Van Noortwijk, J. & Klatter, 2004) ont démontré la pertinence de ce processus dans la modélisation de la durée de vie et la maintenance de ponts.

L'avantage principal des approches stochastiques réside dans leur capacité à capturer l'incertitude physique et intrinsèque lors de la prévision de l'état futur des infrastructures civiles. Ces approches sont faciles à calibrer et à appliquer (Thomas & Sobanjo, 2013). Cependant, elles présentent plusieurs inconvénients tels que leur sensibilité aux données bruitées (Agrawal et al., 2010), ou encore la difficulté d'interprétation des probabilités de transition (Tran, Huu Dung, 2007).

2.1.3 Approche basée sur l'intelligence artificielle

Kaplan et Haenlein (2019) définissent l'intelligence artificielle (IA) comme l'ensemble des théories et des techniques mises en œuvre en vue de réaliser des machines capables de simuler l'intelligence. L'IA propose à un ensemble de concepts et de technologies qui imitent l'intelligence humaine plus qu'à une discipline autonome constituée. Les deux principaux concepts de l'IA sont l'apprentissage (Training) et la validation (Testing). Si les données sont étiquetées, c'est-à-dire que la réponse à la tâche est connue pour ces données, l'apprentissage est dit supervisé. Il s'agit d'un problème de classification si l'étiquette (variable à expliquer) est discrète, ou de régression si la variable à expliquer est continue.

Dans le cas sans étiquette, l'apprentissage est dit non supervisé. Il s'agit d'extraire des classes ou groupes d'individus présentant des caractéristiques communes. La qualité d'une technique de partitionnement est mesurée par sa capacité à découvrir certains ou tous les motifs cachés. Le clustering est une des formes de l'apprentissage non supervisé, il vise à diviser un ensemble de données en différents « paquets » homogènes (Ngo, 2011). La figure 2.1 illustre les deux approches d'apprentissage automatique :

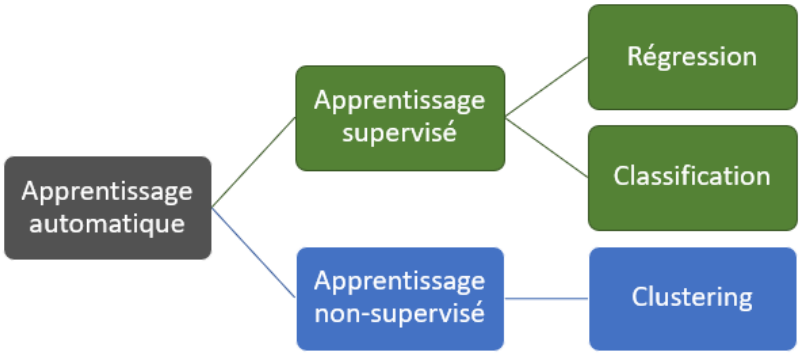


Figure 2.1 Approches d'apprentissage automatique

Selon (Cadei et al., 2019), la construction de modèles interprétables d'apprentissage automatique basés sur des ensembles de données de dégradation limités peut être intégrée dans des outils de diagnostic automatique. Ces derniers peuvent être utilisés pour analyser des phénomènes de dégradation complexes et optimiser la planification de la maintenance prédictive. Les techniques d'apprentissage supervisé sont largement utilisées dans la prédiction et le diagnostic de la dégradation lente des systèmes industriels. Cette approche permet d'éviter la complexité des

modèles déterministes ainsi que le manque d'interprétabilité des modèles stochastiques selon (Hasan, 2015).

- ***Réseaux de neurones artificiels***

Le réseau de neurones artificiels (RNA) est la technique d'IA la plus populaire. Les auteurs de la référence (Ge et al., 2019) ont utilisé un réseau neuronal récurrent pour prédire la durée de vie utile restante des roulements à long cycle de vie sur la base de courtes séquences d'observations. Un réseau neuronal à rétropropagation a été formé et validé par (Zhu et al., 2019) pour prévoir la durée de vie d'un adhésif à base de résine époxy à partir de petits échantillons.

Le traitement des séries temporelles décalées est une pratique courante lors de l'étude d'un parc constitué de plusieurs systèmes identiques avec différents âges et conditions d'exploitation. Ce problème a été abordé dans le travail de (Pereira & Silveira, 2018) où ils utilisent des auto-encodeurs récurrents pour détecter des anomalies dans des séries temporelles décalées relatives à plusieurs panneaux photovoltaïques d'âges et d'emplacements différents. La méthode est basée sur l'attribution d'un score d'anomalie à chaque observation d'une séquence. Toutefois, elle ne fait pas de discrimination entre les différents profils de défaut et n'estime pas l'indicateur de performance.

En raison du processus de dégradation lente, de la difficulté et de l'expressivité de l'inspection, les actifs d'infrastructure ont souvent des ensembles de données de séries temporelles limitées et bruitées. Les réseaux de neurones ont été largement utilisés pour résoudre ce problème de modélisation, en particulier pour les chaussées (Domitrović et al., 2018), les conduites d'aqueduc (Bubtiena et al., 2011), les ponts (Fathalla et al., 2018), les égouts sanitaires (Najafi & Kulandaivel, 2005) et les conduites d'eau pluviale (Tran, Huu Dung, 2007).

Les réseaux de neurones récurrents (RNR) ont été amplement déployés pour traiter les signaux cycliques caractérisant la performance des processus et des équipements industriels. Nanduri et Sherry (2016) ont utilisé l'architecture profonde LSTM pour construire un auto-encodeur qui permet de construire les séries chronologiques relatives à la phase de rapprochement des avions de la piste d'atterrissage. Ce modèle est construit sur les cycles normaux seulement et les phases anormales sont détectées lorsque l'erreur de prédiction dépasse une limite donnée. Cette approche a atteint une haute précision lors de la détection de 11 anomalies prédéfinies par des experts. Cependant, cette technique nécessite un étiquetage préalable des cycles de rapprochement. Une autre application de LSTM a été illustrée dans l'article (Tong et al., 2018). Ce travail a traité les

séries chronologiques relatives aux 16 capteurs choisis par les experts qui concernent le cycle d'atterrissage d'une flotte d'avions de ligne du même modèle. L'objectif était de prédire l'accélération verticale par cycle dans le but de détecter les atterrissages difficiles.

Malgré sa capacité à modéliser des relations inconnues et complexes entre les indicateurs clés de performance et l'environnement opérationnel sur la base de séquences d'observations imprécises et bruyantes, les réseaux de neurones est une technique de type "boîte noire" vu que le chemin de la solution n'est pas transparent. Les modèles d'apprentissage profond sont également vulnérables au phénomène de surajustement et leur interprétation est toujours délicate selon (Tran, Huu Dung, 2007).

- ***Autres techniques supervisées***

D'autres techniques d'apprentissage supervisé ont également été utilisées pour modéliser la dégradation des actifs sur la base de données limitée, telles que le processus gaussien, les forêts aléatoires « Random forests », Support Vector Machine (SVM) et Gradient Boosting.

La régression par processus gaussien a été déployée par (Harry et al., 2020) pour la prédiction de la durée de vie résiduelle des batteries lithium-ion qui présentent différents modes de dégradation lente couplés à un manque d'observations de son état. Cette approche s'est avérée efficace lorsqu'il s'agissait de traiter des séries temporelles courtes relatives à des batteries d'âges différents, en obtenant une erreur relativement faible sur les données de test. La technique de forêts aléatoires a été utilisée dans les travaux de (Laakso et al., 2018) et de (Vitorino et al., 2014) pour modéliser la dégradation des ponceaux de drainage; de (Falamarzi et al., 2019) pour estimer l'indice de dégradation des voies ferrées et de (Marcelino et al., 2019) pour la détérioration de la chaussée. Bien que la précision de la classification soit restée moyenne, cette méthode a permis de réduire ou éliminer le surajustement dans la plupart des cas.

Les travaux de (Cadei et al., 2019) ont utilisé la technique SVM pour apprendre à partir d'un ensemble de données de fonctionnement normal d'un rebouilleur. Ainsi, les auteurs ont développé un algorithme de surveillance pour détecter une observation anormale qui s'écarte statistiquement du comportement normal. Cette approche ne nécessite qu'un petit échantillon d'observations de classes dites normales (sans défaillances) pour prédire les défaillances futures. Dans ce même article, ils ont proposé une nouvelle méthodologie pour prédire le coefficient de l'échangeur de chaleur à court et à long termes en se basant sur un ensemble de données limitées qui ne présentent

qu'un seul évènement d'encrassement critique enregistré sur plus de cinq ans. Ils ont donc utilisé un modèle ARIMA pour la prédiction à court terme et la régression Ridge pour la projection à long terme.

Pour analyser les dépendances entre les signaux et les défauts d'un système hydraulique opérant selon différentes charges, Helwig, Pignanelli, et Schütze (2015) ont effectué une extraction automatique des caractéristiques des cycles par corrélation de Pearson. Ensuite, ces variables ont été utilisées pour entrainer 3 algorithmes supervisés pour prédire la classe de défaut, notamment le « Linear Discriminant Analysis (LDA) », les RNA et le SVM. Ainsi, la meilleure précision a été obtenue par les RNA.

Dans le but de modéliser la dégradation d'un distributeur hydraulique, Lei et al. (2019) ont développé une méthodologie à 3 étapes. Tout d'abord, pour obtenir l'ensemble de caractéristiques de la composante principale (ACP) du signal de la pression a été utilisée. Elle a permis de réduire la dimension des signaux de la pression d'entrée et de sortie de la valve. Deuxièmement, un échantillon d'apprentissage automatique a été construit en remplaçant l'ensemble de défauts d'origine par l'ensemble de caractéristiques des composantes principales. Troisièmement, la technique XGBoost a été déployée pour estimer la criticité des défauts au niveau du système. Ce modèle a surperformé l'arbre de décision et les forêts aléatoires en termes de précision de la classification.

- ***Approche non supervisée***

Le clustering est couramment utilisé pour extraire les profils de dégradation ou pour détecter les profils anormaux lors du traitement des séries chronologiques de systèmes opérants en cycle. Dans le but d'extraire les catégories du cycle de production, Cerquitelli et al. (2021) ont effectué l'extraction des caractérisés statistiques intra cycles comme la moyenne, la variance et le Kurtosis. Ensuite, ils ont divisé les séries en des fenêtres de temps pour extraire la pente et l'ordonnée à l'origine de la droite de régression de chaque variable. Ces paramètres estimés sont considérés comme les caractéristiques inter classe qui conservent la composante du temps. Et finalement, ces facteurs ont été regroupés en utilisant 3 différents algorithmes : le K-means, le Bisecting K-means et le Gaussian mixture. Ces techniques ont été évaluées sur la base de leurs silhouettes pour choisir la meilleure partition des signaux. La limite majeure de cette technique réside dans le fait qu'elle ne permet pas d'interpréter chaque profil de cycle de production.

Un flux d'analyse des séries temporelles relatifs à un processus cyclique de production en deux étapes est proposé par (Kozjek et al., 2017). Dans la première phase, les règles décrivant les conditions du processus sont extraites à l'aide d'un arbre de décision et de la connaissance experte du processus de fabrication observé. Ces règles sont basées sur des facteurs dérivant la performance du processus comme l'écart par rapport aux tolérances dimensionnelles des produits et le nombre d'arrêts non planifiés des machines. Dans la deuxième phase, différents types de conditions défectueuses sont identifiés et décrits en appliquant une partition hiérarchique des règles transformées en vecteurs de bits.

Une autre application de l'approche non supervisée est illustrée dans l'article de (Li et al., 2011) dont le but est d'extraire les vols anormaux en se basant les signaux enregistrés. Pour les rendre comparables, les séries chronologiques de vols sont échantillonnées à intervalles fixes de temps. Les valeurs échantillonnées sont ensuite arrangées pour former un vecteur de haute dimension pour chaque vol avec une fenêtre de temps fixe. Et finalement, les vecteurs sont regroupés à l'aide de l'algorithme DBSCAN sur la base de la distance euclidienne pour extraire les profils normaux. Les vols non classés sont considérés comme anormaux.

Bien que la classification ait réalisé des bonnes performances lorsqu'il s'agit de traiter des séries d'observations des signaux cycliques, au meilleur de notre connaissance, cette approche n'a pas été utilisée pour extraire des profils de dégradation à partir de séries chronologiques à courtes séquences d'observations.

2.1.4 Comparaison des approches de modélisation

Les approches précédentes présentent des avantages et des inconvénients lorsqu'elles sont appliquées à la modélisation de la dégradation des infrastructures. Celles-ci sont résumées dans le Tableau 2.1.

Tableau 2.1 Comparaison des approches de modélisation recensées dans la littérature

Approches	***Méthodes***	***Avantages***	***Inconvénients***
Déterministe	- Régression linéaire multiple - Régression exponentielle - Régression non linéaire	- Fournit un aperçu des facteurs qui affectent le plus le processus de détérioration - La forme finale (équation) est très conviviale - Relativement facile à comprendre et à développer	- Les hypothèses sous-jacentes, qui peuvent être difficiles à valider, doivent être satisfaites - Non approprié pour modéliser des états discrets avec un modèle linéaire (Bu, 2013)
Stochastique	- Modèles de Markov - Régression logistique - Analyse discriminante multiple - Modèle de survie de cohorte - Modèle de risque proportionnel	- Peut être facilement intégré aux modèles de risque (Ana, E. V. & Bauwens, W., 2010) - Sortie de données discrètes (Tran, Huu Dung, 2007) - Modélise l'incertitude inhérente aux processus de détérioration	- Peut nécessiter des données longitudinales difficiles à trouver (Baik et al., 2006) - Il peut être nécessaire de créer des cohortes (Wirahadikusumah et al., 2001), nécessitant davantage de données - Difficulté d'interprétation des probabilités de transition (Tran, Huu Dung, 2007)
Intelligence artificielle	- Réseau de neurones - Random Forest - Support Vector Machine - Gradient Boosting	- Peut modéliser des relations inconnues, complexes et non linéaires entre les entrées et les sorties - Quelques hypothèses sous-jacentes - Peut être utilisé lorsque les données sont imprécises, incomplètes et subjectives (Hasan, 2015) - Le clustering est efficace pour partitionner les séries longues	- La configuration initiale peut être longue et compliquée (Tran, H. D. et al., 2010) - Technique « boîte noire» signifie que le chemin de la solution n'est pas transparent Grande quantité de données nécessaires à la formation et à la calibration (Scheidegger et al., 2011) - Grande quantité de données nécessaires à la formation et à la calibration (Falamarzi et al., 2018) - Pas d'algorithme de clustering spécifique pour traiter les séries courtes et décalées dans le temps.

En analysant le Tableau 2.1, nous constatons que l'approche basée sur l'apprentissage automatique répond mieux à la problématique énoncée précédemment. En effet, l'approche déterministe est

construite sur des hypothèses dont la validation est très problématique. En plus, les méthodes stochastiques sont très sensibles au bruit et n'offrent pas une interprétabilité physique des probabilités. C'est pourquoi nous avons décidé d'utiliser l'apprentissage automatique pour développer une méthodologie efficace et interprétable.

Une série chronologique est un flux de points de données indexés (ou répertoriés ou représentés graphiquement) dans l'ordre chronologique. Le plus souvent, une série temporelle est une séquence prise à des points successifs également espacés dans le temps (fréquence fixe) (Brockwell & Davis, 2009). Cependant, dans ce qui suit, nous allons traiter des séries dont les observations sont inégalement espacées dans le temps. Ce qui va nécessiter un traitement préliminaire pour les préparer.

2.2 Clustering de séries temporelles

Le clustering de séries temporelles permet de répartir une énorme quantité de séquences d'observations sans aucune connaissance préalable des structures des classes. Elle peut nous épargner les difficultés liées au traitement d'une grande quantité de données ayant des comportements différents. Ainsi, le but du regroupement est d'identifier la structure d'un jeu de données non étiquetées en les classant objectivement dans des groupes homogènes où la similarité intra-groupe est minimisée et la dissimilarité inter-groupe est maximisée. Le clustering est indispensable lorsque les données ne sont pas labellisées (Liao, 2005).

Dans ce qui suit, nous allons passer en revue et discuter les techniques de classification de séries temporelles, de la préparation des données à la validation des résultats. L'idée principale derrière le regroupement est de trouver des similitudes entre différentes séries temporelles et de les regrouper dans un même groupe de manière que les séquences de données du même groupe (cluster) se ressemblent davantage que ceux des autres groupes. Les techniques de clustering peuvent être divisées en 3 catégories principales:

- ***Clustering de séries temporelles entières***

Le clustering de séries temporelles entières est considéré comme le regroupement d'un ensemble de séries individuelles en fonction de leur similarité. Dans ce cas, la classification consiste à appliquer une classification conventionnelle sur des objets discrets, où les objets sont des séries temporelles.

- ***Clustering de sous-séquences de séries temporelles***

Le regroupement de sous-séquences permet d'effectuer un groupage sur un ensemble de sous-séquences d'une séquence temporelle extraites par une fenêtre glissante, soit le regroupement de segments à partir d'une seule longue série temporelle.

- ***Clustering de points temporels***

Cette catégorie concerne la partition des points temporels basée sur une combinaison de leur proximité temporelle des points temporels et de la similarité des valeurs correspondantes.

Dans le présent mémoire, nous sommes intéressés à développer un algorithme de classification de séries chronologiques entières. Cela consiste selon (Aghabozorgi et al., 2015) à mettre en œuvre 4 composantes : la représentation et la préparation des séries, la mesure de similarité, le prototype et le mécanisme de clustering.

2.2.1 Représentation de séries temporelles

La première composante du clustering de séries temporelles concerne la réduction de dimension (longueur de la série) qui est une solution commune à la plupart des approches dans la littérature. La réduction de la dimensionnalité permet de représenter les séries temporelles brutes dans un autre espace en les transformant en un espace de dimension inférieure ou en extrayant des caractéristiques. Cette réduction peut être effectuée avant la classification pour de multiples raisons selon (Lin et al., 2003), telles que la réduction des besoins en mémoire, l'accélération du calcul des distances en réduisant le nombre de dimensions et l'atténuation de l'effet des distorsions des séries bruitées. Néanmoins, dans le cas où nous traitons des séries courtes, de faibles dimensions et qui présentent des sauts, nous allons nous contenter de faire un suréchantillonnages des séquences pour contrer le manque des données disponibles.

La revue de la littérature montre que peu de travaux sont proposés pour les séries temporelles à valeurs discrètes. On constate également que la plupart des recherches sont basées sur des données échantillonnées de manière régulière, tandis que peu de travaux portent sur des données échantillonnées de manière irrégulière.

2.2.2 Mesures de similarité et de dissimilarité

Afin d'effectuer un regroupement, une collection de proximités doit être disponible pour toutes les paires de séries. Cela correspond à une matrice de proximités où les entités figurant à chaque intersection ligne-colonne constituent le même ensemble d'objets. Selon (Rousseeuw, 1987), deux types de proximités peuvent être considéré: la dissimilarité et la similarité. La première mesure la différence entre deux entités alors que la seconde mesure leur ressemblance. Les mesures de similarité et de dissimilarité sont largement utilisées dans de nombreux domaines de l'intelligence artificielle. (Belanche, 2012) a proposé deux définitions pour les mesures de similarité et de dissimilarité comme suit :

Soit *X* un espace non vide de séries chronologiques :

i) ***Mesure de similarité.*** Une mesure de similarité est une fonction exhaustive, totale ***Sim*** dont la borne supérieure existe : $X \times X \to I_s \subset R$ avec $|I_s| > 1$. Donc, I_s admet une borne supérieure et $sup\ (I_s)$ existe.

ii) ***Mesure de dissimilarité.*** Mesure de dissimilarité est une fonction exhaustive, totale ***Diss*** dont la borne inférieure existe : $X \times X \to I_d \subset R$ avec $|I_d| > 1$. Donc, I_d est borné inférieurement et $inf\ (I_d)$ existe.

En se basant sur les définitions ci-dessus, nous définissons $Sim_{max} = sup\ (I_s)$ et $Diss_{min} = inf\ (I_d)$. Sans perte de généralité, nous pouvons prendre $Sim_{max} \geq 0$ et $Diss_{min} \geq 0$. Pour comparer et interpréter les deux mesures *Sim* et *Diss*, nous pouvons se baser les propriétés suivantes :

- Réflexivité : $Sim(x, x) = \text{Sim}_{\text{max}}$ et $Diss(x, x) = \text{Diss}_{\text{min}}$.
- Réflexivité forte : $Sim(x, y) = \text{sim}_{\text{max}}$ Si et seulement si $x = y$ et $Diss(x, y) = \text{Diss}_{\text{min}}$ Si et seulement si $x = y$.
- Symétrie : $Sim(x, y) = Sim(y, x)$ et $Diss(x, y) = Diss(y, x)$.

La mesure de similarité ou de dissimilarité est une métrique qui permet de comparer deux séquences d'observations. Le choix de la mesure est déterminant dans le clustering, il faut le choisir avec précaution. Ce choix peut être affecté par la stratégie d'enregistrement de la série temporelle, la fréquence d'observations et la qualité des données.

Pour répondre au problème de classification de séries temporelles courtes et bruyantes, plusieurs mesures de distance ont été mises en évidence dans la littérature. La distance euclidienne est une des plus populaires. Elle calcule les distances entre les séries temporelles point par point et de manière complètement inélastique (Faloutsos et al., 1994). Sakoe (1971) a introduit le Dynamic Time Warping (DTW) qui prend en compte la déformation dans le temps lors de la comparaison de deux séries ayant les mêmes extrémités. Le coefficient de corrélation de Pearson et la distance basée sur la probabilité ont également été utilisés dans les travaux de (Sakoe, 1971) et (Kumar, M. et al., 2002). Ces distances sont appropriées pour traiter des séries stationnaires synchronisées et de longue durée. Une distance pour les séries temporelles courtes a été développée dans les travaux de (Möller-Levet et al., 2003) pour traiter les séries temporelles courtes synchronisées. Aßfalg et al. (2006) ont introduit une mesure élastique permettant de traiter des séquences d'observations bruitées sans tenir compte du temps. Le Tableau 2.2 résume les principales caractéristiques des métriques les plus utilisées dans les problèmes de classification non supervisée de séries temporelles.

Tableau 2.2 : Comparaison des mesures de similarité pour les séries temporelles

Distance / Similarité	**Caractéristiques**
Dynamic Time Warping	• Mesure élastique • Prend en compte la déformation dans le temps • Les 2 séries comparées doivent avoir les mêmes bouts
Coefficient de corrélation de Pearson	• Ne traite que les séries stationnaires et synchronisées
Distance euclidienne	• Mesure non élastique qui calcule la distance point par point • Ne permet pas de capter la déformation dans le temps
Short time-series distance (STSD)	• Ne traite que les séries courtes • Capte le temps • Ne traite que les séries synchronisées
Plus longue sous-série commune (LCSS)	• Robuste contre le bruit • Mesure élastique • Traite des séquences (et non pas des séries temporelles)

En plus de caractéristiques présentées ci-dessus, Agrawal et al. (2010) recommandent des mesures de similarité pour comparer des séries temporelles avec des fréquences d'enregistrement et des nombres d'observations différents.

2.2.3 Méthodes de clustering

Le choix de la méthode de classification appropriée est un problème important dans le regroupement de séries temporelles. Selon (Aghabozorgi et al., 2015), ces méthodes peuvent être classées en six groupes: Hierarchical, Partitioning, Grid-based, Model-based, Density-based clustering et Multi-step clustering.

i) La méthode hiérarchique considère chaque élément comme une classe, puis les fusionne progressivement. À l'inverse, cette méthode peut commencer avec tous les éléments comme un seul cluster, puis les divise en plusieurs groupes.

ii) La méthode de « Partitioning » regroupe les séries temporelles de manière que chaque groupe contienne au moins un élément.

iii) Les méthodes « Grid-based » subdivisent l'espace en un nombre fini de cellules qui forment une grille, puis effectuent le regroupement sur les cellules de la grille.

iv) La méthode « Model-based » suppose un modèle pour chaque classe et trouve la meilleure adaptation des données à ce modèle. Cette approche tente de récupérer le modèle original à partir de la partition de l'ensemble de données.

v) Dans la méthode « Density-based », les classes sont des sous-espaces d'objets denses qui sont séparés par des sous-espaces où les objets ont une faible densité.

vi) La méthode de « Multi-step clustering » utilise des méthodes hybrides pour améliorer les regroupements en proposant de nouveaux modèles à chaque étape.

Parmi toutes les méthodes mentionnées ci-dessus, nous sommes particulièrement intéressés par la méthode basée sur la densité. En effet, elle permet d'extraire des regroupements de formes quelconques et d'identifier les éléments anormaux lorsque les séries temporelles sont bruitées. Considérons la nature des séries temporelles qui nous concerne, l'algorithme « Density-Based Spatial Clustering of Applications with Noise » (DBSCAN) sera utilisé. Pour son utilisation, l'algorithme DBSCAN nécessite le choix de deux paramètres (Ester et al., 1996):

- Eps : La similarité minimale ou la dissimilarité maximale entre deux séries pour les considérer voisines.
- MinPts : Nombre minimum de séries nécessaires pour former un cluster.

Le pseudocode de DBSCAN peut être décrit en cinq principales étapes :

Étape 1) Commencer avec une série X_i choisie aléatoirement.

Étape 2) Regrouper X_i avec ses voisins pour composer un cluster.

Étape 3) Si une série est trouvée comme faisant partie d'un cluster, son voisinage fait également partie du même cluster.

Étape 4) Si ce processus aboutit à un cluster d'au moins m points, le cluster est conservé. Dans le cas contraire, ils sont considérés comme du bruit.

Étape 5) Retourner à l'Étape 1 jusqu'à ce que toutes les séries soient considérées.

Pour une meilleure interprétabilité des résultats des regroupements, l'algorithme DBSCAN sera modifié par la proposition d'une nouvelle mesure de similarité.

2.2.4 Prototype de regroupements

Un prototype est un élément de l'espace de données qui représente un regroupement d'éléments. Dans le contexte de classification des séries temporelles, un prototype de regroupement sert à caractériser la classe et ses éléments. Un regroupement peut être représenté par principalement par 3 caractéristiques : la série qui présente la meilleure similarité avec le reste des éléments du regroupement (Médoïde); un seul vecteur moyen (Centroïde); et une loi de distribution ou une région de densité (Ratanamahatana & Keogh, 2005).

2.2.5 Indices de validité du regroupement

La validation des résultats obtenus par un algorithme de classification est une partie fondamentale du processus de regroupement. Étant donné que l'application de l'algorithme DBSCAN avec différents paramètres aboutit généralement à un partitionnement différent des séries, il est important d'évaluer la performance de ces différentes configurations en termes de validité des regroupements et de la qualité de la structure extraite en utilisant des mesures de performance appropriées.

Les indices de validité du regroupement peuvent être classés en deux métriques principales : internes et externes. Les indices de qualité internes ont été proposés par différents auteurs afin de déterminer un regroupement optimal sans tenir compte des informations autres que les séries à

regrouper. Les indices de comparaison externes sont destinés à mesurer la similarité entre les regroupements formés et les étiquettes fournies par l'extérieur (vérité terrain). Ils ne considèrent que la distribution des points dans les différents regroupements et ne permettent pas de mesurer la qualité de cette distribution (Aßfalg et al., 2006). En d'autres termes, Sisodia et Verma (2018) définissent les indices internes et externes comme des métriques non supervisée et supervisée. Dans la plupart des cas, nous avons recours au regroupement lorsqu'il existe de l'information préalable concernant la structure des données. Ce qui explique pourquoi les approches les plus utilisées pour la validation des clusters sont basées sur des indices de validité internes des regroupements (Arbelaitz et al., 2013).

Dans ce travail, nous nous intéressons uniquement aux indices internes permettant d'évaluer la qualité des regroupements obtenus par l'algorithme DBSCAN modifié avec une mesure de similarité ou de distance. Par conséquent, ces métriques telles que le score de silhouette, l'indice de Davies-Bouldin et l'indice C qui sont généralement définies pour un clustering basé sur la dissimilarité doivent être modifiées pour pouvoir fonctionner avec une mesure de similarité. Cela permettra une meilleure interprétation des résultats de regroupements. Nous avons effectué une recherche bibliographique dans le but de trouver les indices de validation de regroupements adéquats aux deux logiques de clustering. Le tableau 2.3 résume l'applicabilité des mesures de dissimilarité et de similarité pour les 3 principaux indices de validation de clustering proposés dans la littérature.

Tableau 2.3 Indices de validation de clustering proposés dans la littérature

Indice de validation	**Logique de dissimilarité**	**Logique de similarité**
Score de silhouette	Oui	Oui
Indice Davies-Bouldin	Oui	Non proposé
Indice C	Oui	Non proposé

2.2.5.1 Score de silhouette

Le score de silhouette permet de trouver des regroupements compacts et clairement séparés. Deux versions de cet indice ont été proposées dans la littérature, notamment pour les mesures de dissimilarité et de similarité.

- Clustering basé sur une distance

Lorsqu'il s'agit d'une comparaison des séries basée sur la logique des distances, le score de silhouette d'une série par rapport à une partition donnée est calculé à partir de sa dissimilarité moyenne par rapport à tous les autres objets de son groupe et de sa dissimilarité moyenne par rapport à toutes les séries qui ne font pas partie son groupe. La formule de la silhouette est formulée comme suit d'après (Rousseeuw, 1987) :

$$Sil_{diss}(i) = \ (y(i) - x(i))/\max\,(x(i); y(i)) \qquad (2.1)$$

Où $x(i)$ et $y(i)$ sont les dissimilarités moyennes entre la série i par rapport aux séquences à l'intérieur et à l'extérieur de son groupe.

- Clustering basé sur une mesure de similarité

Dans le cas d'une partition basée sur une mesure de similarité, il est approprié d'utiliser des indices de validité facilitant l'interprétation des similitudes entre les séries temporelles. En effectuant la revue de littérature, nous n'avons trouvé que la formule du score de silhouette dans l'article de (Rousseeuw, 1987) qui définit le score de silhouette d'un élément (i) pour une mesure de similarité comme suit :

$$Sil_{sim}(i) = \ (w(i) - z(i))/\max\,(w(i); z(i)) \qquad (2.2)$$

Où $w(i)$ et $z(i)$ sont calculés comme suit :

$$w(i) = \frac{1}{|C(i)|-1} \ \sum_{j \in Ci\,, i \neq j} Sim(i,j) \qquad (2.3)$$

$$z(i) = \max_{k \in C \setminus C(i)} \ (\frac{1}{|C(k)|-1} \ \sum_{j \in Ck} Sim(i,j)\ ) \qquad (2.4)$$

Où, $Sim(i,j)$ est la similarité entre séries i et j et $C(i)$ représente le regroupement auquel la série i est assignée. C représente l'ensemble des clusters.

Le coefficient de silhouette, proprement dit, est la moyenne $\bar{s}(i)$ des coefficients de silhouette pour tous les points. Ce score est par définition compris entre -1 et 1. Selon (Albalate & Minker, 2013), la silhouette est souvent interprétée de la manière suivante : la partition trouvée est considérée comme "Forte" si $0{,}7 < \bar{s}(i) \leq 1$; "Raisonnable" lorsque $0{,}5 < \bar{s}(i) \leq 0{,}7$; " Faible " dans le cas où $0{,}25 < \bar{s}(i) \leq 0{,}5$ et "Non classé" si $\bar{s}(i) \leq 0{,}25$.

2.2.5.2 Indice de Davies-Bouldin

L'indice de Davies-Bouldin (DB) est une métrique qui permet de mesurer la qualité des regroupements de séries temporelles en apprentissage non-supervisé. Cet indice mesure la moyenne du rapport maximal entre la distance d'un point au centre de son groupe d'éléments et la distance entre deux centres d'autres groupes. L'indice de Davies-Bouldin cherche à minimiser la distance moyenne entre chaque regroupement et celui qui lui est le plus similaire (Davies & Bouldin, 1979). Il est défini pour une mesure de dissimilarité comme suit :

$$R = \frac{1}{N}\sum_{i=1}^{N} \max_{i \neq j} \frac{S_i + S_j}{M_{ij}} \qquad (2.5)$$

Où N est le nombre de regroupements, S_i est la dissimilarité moyenne des séries par rapport à leur prototype du regroupement i et M_{ij} est la distance entre les séries qui sont choisies comme caractéristiques des regroupements i et j (prototypes).

L'indice DB est défini en fonction du rapport entre la dispersion au sein d'un groupe et la séparation entre les groupes, une valeur plus faible signifie que la classification est meilleure. Il s'agit de la moyenne du rapport maximal entre la distance d'un point au centre de son cluster et la distance entre deux prototypes de groupes. Cela confirme l'idée qu'aucun groupe ne doit être similaire à un autre, et donc que la meilleure structure de regroupements minimise essentiellement l'indice de Davies-Bouldin. Ainsi, l'indice est symétrique et non-négatif. Vu que cet indice correspond à une moyenne sur tous les groupes, il représente une bonne mesure pour choisir la meilleure combinaison de paramètres pour optimiser la partition des séries chronologiques (Arbelaitz et al., 2013).

2.2.5.3 L’indice C

L'indice C est une mesure normalisée de cohésion calculée à partir de la somme des distances sur toutes les paires d'éléments du même groupe. Il est défini dans (Hubert & Schultz, 1976) comme suit :

$$C = \frac{S - S_{min}}{S_{max} - S_{min}} \quad (2.6)$$

Où :

- S est la somme des distances sur toutes les paires d'éléments appartenant au même groupe.
- m est le nombre de paires dans le même groupe.
- S_{min} est la somme des m plus petites distances si toutes les paires d'éléments sont considérées.
- S_{max} est la somme des m plus grandes distances si toutes les paires d'éléments sont considérées.

L'indice C est limité à l'intervalle [0,1] et doit être minimisé pour obtenir une meilleure partition des données. Ainsi, réduire cet indice revient à minimiser la somme des distances sur toutes les paires d'éléments appartenant au même cluster par rapport aux distances entre les pairs d'objet de la base de données étudiée. Bezdek et al. (2016) ont considéré cet indice comme l'un des meilleurs parmi la surabondance d'indices internes lorsqu'il s'agit d'identifier et de définir la structure des regroupements.

2.3 Qualité de prédiction d’un modèle

La qualité de prédiction d’un modèle repose sur deux principaux critères : l’erreur quadratique moyenne (MSE) et le coefficient de détermination (R^2). Le MSE mesure la moyenne des erreurs au carré entre les prédictions d’un modèle et les observations. En d’autres termes, elle mesure une proportion de l’erreur du modèle et permet de comparer divers modèles entre eux afin de déterminer ceux qui représentent le mieux les données. Plus cette mesure est proche de 0, plus les prédictions se rapprochent de ce qui a été observé. L’équation (2.7) calcule l’erreur quadratique moyenne.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \quad (2.7)$$

Où $\hat{Y}i$ est une prédiction et Yi une observation de l'ensemble des données d'évaluation.

Le coefficient de détermination, noté R² et prononcé "R carré", est la proportion de la variance dans la variable dépendante qui est prévisible à partir des variables indépendantes. Plus ce coefficient est proche de 100%, meilleur est le modèle. Il est calculé avec l'équation 2.8 suivante :

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y}_i)^2} \quad (2.8)$$

Où $\bar{Y}_i$ est une prédiction, Y_i une observation et $\bar{Y}_i$ la moyenne sur l'ensemble des observations.

L'utilisation d'un R² ajusté est une tentative de prise en compte du phénomène du R² qui augmente de façon fausse et aléatoire lorsque des variables explicatives supplémentaires sont ajoutées au modèle. C'est une modification de R² qui ajuste le nombre de termes explicatifs dans un modèle par rapport au nombre de points de données. Le R² ajusté peut-être négatif et sa valeur sera toujours inférieure ou égale à celle de R². Contrairement à R², le R² ajusté n'augmente que lorsque l'augmentation de R2 est supérieure à ce que l'on pourrait s'attendre à voir par hasard, en raison de l'inclusion d'une nouvelle variable explicative. Si un ensemble de variables explicatives ayant une hiérarchie d'importance prédéterminée est introduit dans une régression, les R² ajustés étant calculés à chaque fois, le niveau auquel R^2 ajusté atteint un maximum et décroît par la suite est la régression avec combinaison idéale d'avoir le meilleur ajustement sans excès ou termes inutiles. Généralement, un modèle de régression ayant un R^2 ajusté plus de 70% est jugé comme assez représentatif du phénomène étudié (Karch & van Ravenzwaaij, 2020). Le R^2 ajusté est défini comme suit :

$$\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1} \quad (2.9)$$

Où p est le nombre de variables explicatives dans le modèle et n la taille de l'échantillon d'entrainement.

Il ne faut pas sous-estimer le problème de sur-apprentissage quand un modèle a trop appris les particularités de chacune des données de l'échantillon. Il présente alors un taux de succès très important sur les données d'entraînement (pouvant atteindre jusqu'à 100%), au détriment de ses performances générales réelles (Kaplan & Haenlein, 2019). L'un des moyens les plus élémentaires mis en place pour lutter contre le surajustement est la séparation des données disponibles en deux jeux distincts : un pour l'entraînement et un autre pour le test par la validation croisée. Ainsi, les

capacités du modèle à généraliser les tendances apprises peuvent être évaluées, plutôt qu'à fournir toujours les mêmes résultats mémorisés.

CHAPITRE 3 MÉTHODOLOGIE DE MODÉLISATION DES PROFILS DE DÉGRADATION

Dans ce chapitre, nous proposons une méthodologie pour extraire, caractériser et interpréter les profils de lente dégradation à partir des séries temporelles mesurant l'évolution d'un indice de performance au cours du temps ou des signaux des capteurs correspondant à des systèmes opérants par cycle.

La méthodologie proposée permet de relever les défis liés à deux structures différentes de base de données. Ainsi, elle permet en premier lieu, d'extraire des profils de dégradation à partir des courtes séquences bruitées de ICP de différentes fréquences d'observation. En deuxième lieu, la technique de partitionnement des signaux des capteurs est conçue pour capturer la cyclicité des séries chronologiques. Pour extraire les profils de dégradation à partir de ces deux structures de données, nous proposons deux algorithmes de clustering basés sur la similarité entre les courtes séquences d'ICP et la dissimilarité entre les signaux cycliques. Pour renforcer l'évaluation et proposer une meilleure interprétabilité de la partition basée sur la similarité, nous proposons les indices DB et C modifiés. Les profils extraits sont utilisés par la suite pour expliquer les différents modes de détérioration et pour prédire l'état de dégradation des actifs.

3.1 Présentation de la méthodologie

La méthodologie proposée utilise tout d'abord l'apprentissage non supervisé pour partitionner les séries temporelles par rapport à leurs comportements de détérioration. Ensuite, des algorithmes de classification sont entrainés pour classifier les séquences de dégradation ou les cycles en se basant sur les regroupements générés. Et en dernier, la caractérisation et l'interprétation des profils de dégradation par rapport aux indicateurs de performance du système étudié. La Figure 3.1 illustre les 3 phases principales de la méthodologie. Elles sont décrites ci-après :

- ***Phase 1-Extraction des profils de dégradation :*** nous utilisons l'algorithme de clustering DBSCAN sur les séries temporelles pour extraire des profils de dégradation.

- ***Phase 2-Entraînement du modèle de prédiction des clusters*** : Nous adaptons un classificateur Random Forest afin de prédire le profil de dégradation pour chaque série temporelle de la base de données.
- ***Phase 3-Caractérisation et interprétation des profils de dégradation :*** nous exploitons les modèles déjà extraits pour identifier les facteurs ou signaux qui influencent le plus la distinction entre les différents profils de détérioration. Un algorithme de prévision de l'ICP future est également proposé pour les bases de données.

Considérant que cette méthodologie permet de traiter les deux types de séries chronologiques, notamment les séries courtes, décalées dans le temps et bruitées, et celles cycliques de signaux, nous allons dépeindre les différents ajustements apportés à chaque phase de la méthodologie.

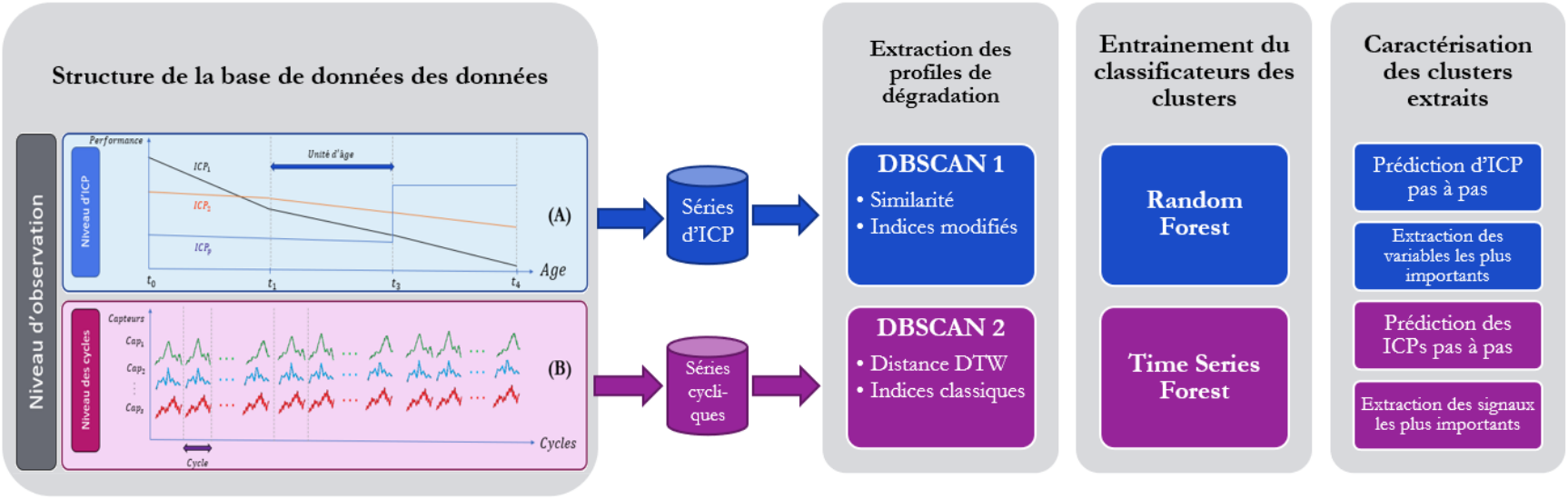


Figure 3.1 Phases de la méthodologie proposée

3.2 Phase 1-Extraction des profils de dégradation

La première phase de la méthodologie concerne l'extraction des différents profils de dégradation par le regroupement des séries chronologiques données par l'évolution de l'ICP du système en fonction de leurs profils et tendances. Dans cette phase, les systèmes ayant un comportement dégradé similaire seront regroupés dans le même groupe. Pour ce faire, l'algorithme DBSCAN couplé par une mesure de similarité pour les séquences courtes et par une distance pour les séries cycliques est mis à contribution.

3.2.1 Mesure de similarité proposée

3.2.1.1 Séries courtes d'indicateur clé de performance

Bien qu'il existe une multitude de mesures de similarité et de dissimilarité présentées dans la section 2.1.2, nous n'avons pas trouvé de mesure appropriée pour les ensembles de données limités comportant des séries temporelles courtes et décalées. Ainsi, en prenant en compte la recommandation de (Agrawal et al., 2010) d'utiliser une mesure de similarité (et non pas la dissimilarité) pour évaluer la proximité des courtes séries temporelles avec de différents fréquences et nombre d'enregistrements, nous avons développé une nouvelle mesure de similarité. Cette dernière peut être utilisée pour mesurer la similarité entre deux courtes séquences d'observations, bruitées, décalées dans le temps et de différentes longueurs. Cette mesure de similarité est basée sur la longueur de la Plus Longue Sous-séquence Commune (PLSC) entre deux séquences d'observations.

Soient les deux séries d'ICP $X[x_1..x_m]$ de longueur m et $Y[y_1..y_n]$ de longueur n. La longueur d'une séquence est caractérisé par le nombre de ses observations. Les séries temporelles d'ICPs enregistrées X et Y sont séparées en 3 séquences numériques : observations V_x $[v_{x_1}..v_{x_{m-1}}]$ et V_y $[v_{y_1}..v_{y_{n-1}}]$, pentes S_x $[s_{x_1}..s_{x_{m-1}}]$ et S_y $[s_{y_1}..s_{y_{n-1}}]$ et, âges A_x $[a_{x_1}..a_{x_{m-1}}]$ et A_y $[a_{y_1}..a_{y_{n-1}}]$. Où v_{x_i} est la valeur de l'ICP de la série X à l'âge a_{x_i} et s_{x_i} est la pente de dégradation de la série X à l'âge a_{x_i}, elle est calculée en utilisant la formule suivante :

$$s_{x_i} = \frac{v_{x_i} - v_{x_{i+1}}}{a_{x_i} - a_{x_{i+1}}} \tag{3.1}$$

Ainsi, deux observations x_i et y_j des séquences d'ICP *X* et *Y* sont considérées comme identiques si les 3 conditions suivantes sont satisfaites :

$$x_i = y_j \quad Si \quad \begin{cases} |v_{x_i} - v_{y_j}| < \varepsilon_1 \\ |s_{x_i} - s_{y_j}| < \varepsilon_2 \\ |a_{x_i} - a_{y_j}| < \varepsilon_3 \end{cases} \tag{3.2}$$

Les constantes ε_1, ε_2 et ε_3 représentent la tolérance à la différence entre les valeurs de performance, les pentes de dégradation et les âges lors du calcul de la PLSC.

La similarité entre deux séquences $X[x_1..x_m]$ de longueur m et $Y[y_1..y_n]$ de longueur n est donnée par l'Équation 3.2 suivante :

$$Sim(X,Y) = \frac{\text{PSLC}(X,Y)}{\max(m,n)} \tag{3.3}$$

La longueur de PLSC entre deux séries temporelles est calculée par la relation récurrente 3.4. suivante :

$$C[i,j] = \begin{cases} 0 & Si \quad i = 0 \; or \; j = 0 \\ C[i-1,j-1] + 1 & Si \quad (\left|v_{x_i} - v_{y_j}\right|, \left|s_{x_i} - s_{y_j}\right|, \left|a_{x_i} - a_{y_j}\right|) \leq (\varepsilon_1, \varepsilon_2, \varepsilon_3) \\ \max(C[i-1], C[i,j-1]) & Sinon \end{cases} \tag{3.4}$$

L'algorithme calcule la similarité basée sur les 3 composantes entre $X[x_1..x_m]$ et $Y[y_1..y_n]$ pour tout $1 \leq i \leq (m-1)$ et $1 \leq j \leq (n-1)$ et l'écrit dans *C[i,j]*. *C[m-1,n-1]* contiendra la longueur du PLSC des deux séquences.

3.2.1.2 Séries cycliques

Parmi les mesures de proximité discutées dans le Tableau 2.2, la distance « Dynamic Time Warping (DTW) » est destinée à comparer deux séries chronologiques qui commencent et finissent par les mêmes évènements, comme le cas des cycles d'opérations. Ainsi, cette mesure de dissimilarité sera utilisée pour comparer les signaux cycliques multidimensionnels dont la vitesse de manifestation du comportement peut varier. Cette méthode permet de calculer une correspondance optimale entre deux séries temporelles en prenant en compte la déformation dans le temps avec certaines restrictions et règles. La distance est donc la somme des poids du chemin optimal pour faire l'alignement entre les observations des deux séries comparées (Yang, C.-Y. et al., 2019). La Figure 3.2 illustre le principe de la distance utilisée.

Étant donné que le DTW utilise la distance euclidienne pour comparer les observations, il faut normaliser les différents signaux pour donner le même poids à chaque dimension des séries chronologiques. Ainsi, chaque signal sera normalisé en soustraction la moyenne et en mettant à l'échelle la variance unitaire de la base de données entière. Le score standard d'un échantillon x est calculé comme suit : $z = \frac{(x - u)}{s}$, où $\boldsymbol{u}$ est la moyenne et $\boldsymbol{s}$ est l'écart-type des échantillons étudiés. Le centrage et la mise à l'échelle se font indépendamment pour chaque dimension en calculant les statistiques pertinentes sur l'ensemble des cycles.

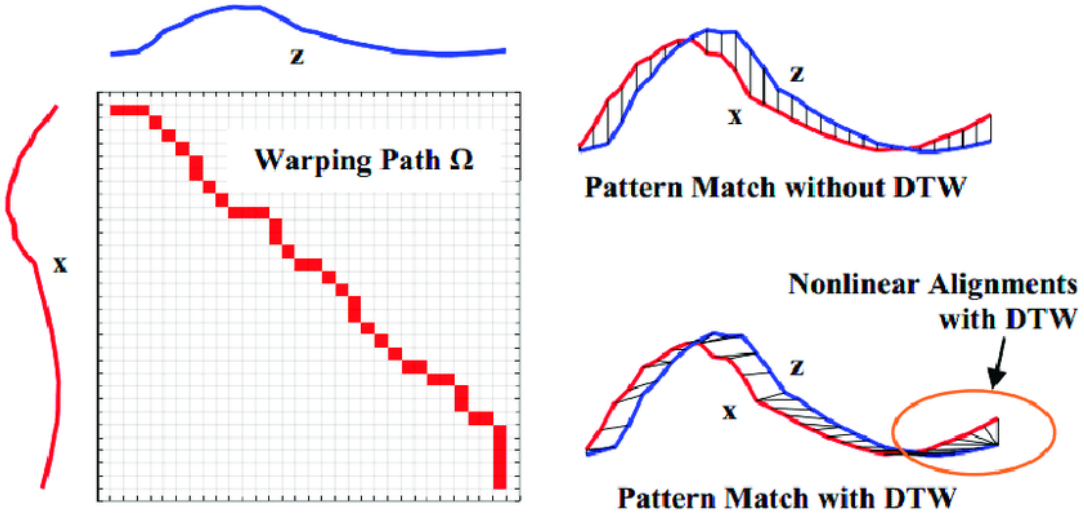


Figure 3.2 Principe de Dynamic Time Warping (Yang, C.-Y. et al., 2019)

La principale limite de DTW réside dans le fait qu'il déforme les séries temporelles dans le temps pour trouver le chemin de correspondance le plus court. Par exemple, une observation au début d'une séquence peut être appariée avec une observation indexée au milieu ou à la fin d'une autre série. Afin de pallier cette limite, le temps jusqu'à la fin du cycle sera inséré comme une nouvelle dimension dans l'ensemble des séries. Cette dimension ne sera pas normalisée comme le reste des facteurs pour donner un plus grand poids au temps. Ainsi, la déformation dans le temps pénalisera la distance DTW puisque le temps a une valeur moyenne supérieure.

3.2.2 Algorithme de clustering

Les mesures de proximité seront utilisées avec l'algorithme DBSCAN pour regrouper les séries temporelles ayant des profils de dégradation similaires et pour détecter et isoler les profils

anormaux. Le choix des paramètres MinPts est assez délicat, car il affecte la qualité de la structure. Ainsi, il peut conduire à classer toutes les séries comme du bruit s'il est grand ou à générer un nombre énorme de petits clusters s'il est choisi très petit. Pour résoudre ce problème, nous proposons une version modifiée de l'algorithme DBSCAN qui ajuste ce paramètre automatiquement lors du regroupement des séries. La méthode proposée est décrite dans le pseudocode suivant :

```
MinPts = Grande valeur initiale
DBSCAN(toutes les séries, MinPts, Eps)
Tant que (Silhouette>0,5 et Nbre_It<It_max)
Début
    Minpts = Minpts*0,95
    DBSCAN(Series bruitées, Minpts, Eps)
    Fusionner la partition initiale avec la nouvelle partition des séries bruitées
Fin
```

Le principe de cet algorithme est assez simple, nous commençons par appliquer le DBSCAN sur l'ensemble de séries en utilisant une grande valeur initiale de MinPts (exemple : la moitié du nombre total des séries à regrouper). Ensuite, nous recyclons le bruit au fil des itérations en le regroupant avec une valeur de MinPts réduite d'un certain pourcentage prédéfinit (5% par exemple). Cela nous permettra de déterminer automatiquement le nombre minimum de séries pour constituer un groupe, d'extraire le maximum de profils de dégradation et de classer plus de séries. Chaque groupe représente un profil de dégradation caractérisé par son Médoïde. Pour le reste du mémoire, nous désignerons par DBSCAN 1 et DBSCAN 2 les versions modifiées de l'algorithme DBSCAN basé respectivement sur la similarité et dissimilarité entre les séries.

3.2.3 Évaluation de la partition

La comparaison entre les différentes configurations de l'algorithme de clustering est basée principalement sur la mesure du score moyen de la silhouette. Pour renforcer l'évaluation des regroupements, nous allons utiliser les indices Davies-Bouldin et C. Toutefois, nous proposons une nouvelle formule relative à chaque indice afin de les adapter à une classification basée sur la similarité entre les séries temporelles.

3.2.3.1 Indice de Davies-Bouldin modifié

Nous cherchons à définir une mesure générale de séparation des groupes d'éléments selon la méthode DB, qui permet de calculer la similarité moyenne de chaque regroupement avec son groupe le plus similaire en se basant sur une fonction de similarité et une mesure de dispersion des regroupements. En faisant l'analogie avec la définition de l'indice DB fournie par (Davies & Bouldin, 1979) pour évaluer les regroupements basés sur la dissimilarité tout en s'inspirant de l'adaptation de score de silhouette présentée par (Rousseeuw, 1987), nous proposons la formule suivante de l'indice modifié de DB :

$$DB_{modifié} = \frac{1}{N}\sum_{i=1}^{N} \max_{i \neq j} F_{ij} \quad (3.5)$$

$$F_{ij} = \frac{L_{ij}}{H_i + H_j} \quad (3.6)$$

Où :

- H_i est la similarité moyenne des points du groupe *i* par rapport à leur prototype (Médoids).
- L_{ij} est la similarité entre les séries qui sont choisies comme caractéristiques des regroupements *i* et *j*.

Soit un groupe C comprenant des éléments $X_1, X_2, \ldots, X_m$. La fonction de similarité des regroupements S_i doit satisfaire les deux conditions suivantes :

- Condition 1 : $S(X_l, X_2, \ldots, X_m) \geq 0$
- Condition 2 : $S(X_1, X_2, \ldots, X_m) = 0$ si $X_i = X_j$, pour tous les X_i et X_j dans le cluster *C*

La fonction de similarité entre clusters F doit répondre aux propriétés suivantes :

- P1. $F(H_i, H_j, L_{ij}) \geq 0$. La fonction de similarité F est positive.
- P2. $F(H_i, H_j, L_{ij}) = F(H_j, H_i, L_{ji})$. R est symétrique.
- P3. $F(H_i, H_j, L_{ij}) = 0$ if only $L_{ij} = 0$. La similarité inter clusters est nulle seulement si leur fonction de dispersion est nulle.
- P4. Si $H_i = H_k$ and $L_{ij} < L_{ik}$ alors $F(H_i, H_j, L_{ij}) < F(H_i, H_k, L_{ik})$. Si la distance entre les clusters augmente alors que leurs dispersions restent constantes, la similarité des clusters diminue.

- P5. Si $L_{ij} = L_{ik}$ et $\mathrm{H_j} > \mathrm{H_k}$ alors $F(H_i, H_j, L_{ij}) < F(H_i, H_k, L_{ik})$. Si la distance entre les clusters reste constante alors que les dispersions augmentent, la similarité augmente.

L'indice de DB modifié peut être interprété comme la moyenne, à l'échelle de la base de données, des mesures de similarité de chaque groupe avec son cluster le plus similaire. Le "meilleur" choix de groupes sera donc celui qui minimise cette moyenne de similarité. Ainsi, l'indice DB ordinaire cherche à minimiser la distance intra cluster et à maximiser la distance inter-cluster. De l'autre côté, l'indice DB adapté vise à réduire la similarité entre les prototypes des différents clusters et à maximiser la similarité entre les éléments d'un même groupe. Par conséquent, l'indice DB modifié conserve les mêmes possibilités d'interprétation que l'indice DB original ainsi que le même intervalle de valeurs $[0, +\infty]$.

3.2.3.2 L'indice C modifié

Afin de l'adapter au regroupement basé sur la similarité, nous proposons la formule modifiée (3.7) suivante de l'indice C :

$$C_{modifié} = \frac{SS_{max} - SS}{SS_{max} - SS_{min}} \tag{3.7}$$

Où :

- SS est la somme des similarités sur toutes les paires d'éléments formant le même groupe.
- m est le nombre de paires dans le même groupe.
- SS_{min} est la somme des m plus ***petites*** similarités si toutes les paires d'objets sont considérées.
- SS_{max} est la somme des m plus ***grandes*** similarités si toutes les paires d'objets sont considérées.

Réduire l'indice C modifié revient à maximiser la somme des similarités entre les pairs de séries d'un même groupe par rapport aux similarités entre les pairs d'éléments de l'ensemble des données étudiées.

3.3 Phase 2-Apprentissage du classificateur de profil de dégradation

Dans la deuxième phase de la méthodologie, nous allons utiliser des techniques de classification supervisée pour estimer les profils de dégradation des séries temporelles.

3.3.1.1 Apprentissage utilisant des séries courtes d'indicateur clé de performance

Dans la deuxième phase de la méthodologie, la technique de forêts aléatoires est utilisée pour prédire le profil de dégradation de chaque système en reclassant les séries chronologiques sur la base des étiquettes générées par l'algorithme de classification. Ainsi, la valeur de l'ICP, l'âge et les variables conceptuelles et de gestion de chaque système seront utilisés pour prédire son groupe à l'aide de la classification supervisée.

Les forêts aléatoires est une technique d'apprentissage automatique qui peut être utilisée pour des problèmes de classification et de régression (Ishwaran & Lu, 2019). À l'instar des méthodes d'apprentissage automatique, les forêts aléatoires peuvent gérer des relations non linéaires entre les variables dépendantes et prédictives, ainsi qu'entre les variables prédictives elles-mêmes. Plus particulièrement, une forêt aléatoire est créée en générant une multitude (un ensemble) d'arbres de classification (ou de régression) indépendants sur la base d'échantillons aléatoires d'observations. De plus, seul un sous-ensemble de variables est sélectionné pour former chaque arbre, ce qui ajoute une autre couche de caractère aléatoire à la construction du modèle.

Les arbres individuels sont des classificateurs faibles (légèrement plus performants que les classifications aléatoires) et la construction d'un ensemble d'arbres améliore considérablement la précision de la prévision par rapport à l'utilisation d'un seul arbre de décision. Chaque arbre crée une prédiction pour la classe (ou la valeur de variable cible) de chaque observation, c'est-à-dire des « votes » pour une classe, et le modèle de forêts aléatoires choisit celle qui obtiendra le plus grand nombre de votes ou la moyenne des valeurs prédites par les arbres dans le cas d'une régression. Une forêt aléatoire donne des résultats avec des données contenant des variables numériques et catégorielles et des valeurs de variables à différentes échelles (Breiman, 2001). Celle technique s'attaque aux limites de l'ajustement excessif des arbres de classification et de régression (CART). Elle utilise l'agrégation bootstrap, également connue sous le nom de bagging, pour créer des sous-ensembles de données d'entraînement par échantillonnage avec remplacement pour

construire différents CART. Chacun des arbres décisionnels qui composent la forêt prédite, pour un vecteur de variables explicatives, renvoie le vote majoritaire de l'ensemble des estimateurs.

Pour optimiser les paramètres de l'algorithme de classification des forêts aléatoires, nous divisons les profils de dégradation extraits, en un ensemble de données d'entraînement, de validation et de test. Nous utilisons la technique de validation croisée k-fold pour comparer les différentes configurations possibles sur l'ensemble de validation et pour éviter le sur-apprentissage en termes de nombre d'arbres et leurs profondeurs de décomposition maximale.

3.3.1.2 Apprentissage utilisant des séries cycliques

Pour les séries cycliques, une variante des forêts aléatoires appelée Time Series Forest (TSF), sera utilisée pour régénérer les étiquettes de classification. Le modèle reçoit comme entrée les séries multidimensionnelles de signaux et retourne le profil de dégradation correspondant.

Time Series Forest est un classificateur de séries temporelles précis et efficace, et est capable d'extraire les caractéristiques temporelles utiles pour distinguer les séries temporelles de différentes classes.

Cette technique utilise un transformateur pour extraire 3 caractéristiques statistiques de chaque fenêtre des séries : la moyenne, l'écart-type et la pente de dégradation. Ce transformateur divise les séries chronologiques en un nombre de fenêtres et calcule les 3 variables à chaque intervalle.

Les variables extraites sont utilisées comme entrées à une forêt aléatoire qui échantillonne aléatoirement les 3 variables des intervalles à chaque nœud de l'arbre. Pour permettre une meilleure classification, Random Forest utilise une combinaison du gain d'entropie et d'une mesure de distance, appelée gain d'entrée (entropie et distance), pour évaluer la séparation au niveau de chaque nœud (Deng et al., 2013).

3.4 Phase 3-Caractérisation et interprétation des profils de dégradation

Dans cette dernière phase, les regroupements des profils générés et les modèles entrainés sont utilisés pour caractériser, interpréter et exploiter chacun des regroupements identifiés.

3.4.1 Caractérisation des profils utilisant des séries courtes

3.4.1.1 Contribution des variables dans la classification

Une fois le modèle de forêts aléatoires est entrainé à prédire le profil de dégradation à partir de la valeur de l'ICP chaque année et les variables explicatives (facteurs) de chaque système, le pourcentage de contribution de chaque variable dans la classification obtenue a priori sera évalué. Pour ce faire, une fois les profils de dégradation extraits, nous utilisons la technique de validation croisée k-fold pour entraîner et évaluer ce modèle afin d'éviter le sur-apprentissage.

Le modèle de forêts aléatoires comprend un certain nombre d'arbres de décision. Chaque nœud dans les arbres de décision est une condition sur une entité unique, conçue pour scinder l'ensemble de données en deux afin que des valeurs de réponse similaires se retrouvent dans le même groupe. La mesure sur laquelle la condition optimale (localement) est choisie s'appelle ***impureté***. Pour la classification, il s'agit généralement d'une ***impureté de Gini*** ou d'un gain d'information / entropie et de la variance pour les arbres de régression. Dans notre cas, nous utilisons ***l'indice de Gini*** pour mesurer la contribution des variables dans la classification des séries temporelles. Cet indice atteint sa valeur minimum (zéro) lorsque tous les éléments de l'ensemble sont dans une même classe de la variable cible. Un indice de Gini plus bas indique une meilleure séparation (contribution dans la classification) entre les classes.

Lors de la formation d'un arbre, il est possible de calculer dans quelle mesure chaque variable explicative diminue l'impureté pondérée dans un arbre. La moyenne sur tous les arbres de la forêt est la mesure de l'importance de la variable. Ainsi, pour un modèle de forêts aléatoires, la diminution des impuretés de chaque variable peut être moyennée et elles sont classées en fonction de cette mesure (Lewinson, 2019).

3.4.1.2 Prévision de l'indice clé de performance

Basé sur les résultats du modèle de forêts aléatoires, cette étape permet de prédire de manière récursive la valeur de l'ICP étudié en utilisant le pseudo algorithme ci-après.

Pour chaque année de l'horizon :

Estimer la valeur de la pente correspondant à l'âge actuel, la valeur de l'ICP et les autres variables en utilisant le modèle de forêts aléatoires.

Soustraire la pente de dégradation estimée de la valeur de ICP.

Mettre à jour la valeur actuelle de l'ICP.

La Figure 3.3 illustre le principe de prédiction de la valeur de l'ICP par l'estimation de la pente à chaque pas. Ainsi, le cas présenté correspond à 4 groupes extraits formés de 4 pentes potentielles de dégradation (représentées par les traits colorés discontinus). La valeur de pente choisie à l'âge t_i sera celle la plus probable selon le modèle de forêts aléatoires qui prend la valeur de l'âge, de l'ICP ainsi que les variables explicatives. Par conséquent, nous pouvons observer des changements possibles de la pente de dégradation au cours du temps.

L'évaluation du modèle de prédiction de l'ICP est obtenue par le coefficient de détermination ajusté (R^2 ajusté) et sa prévision est mesurée par l'erreur quadratique moyenne MSE. La validation croisée est utilisée pour éviter le surajustement du modèle de prédiction proposé.

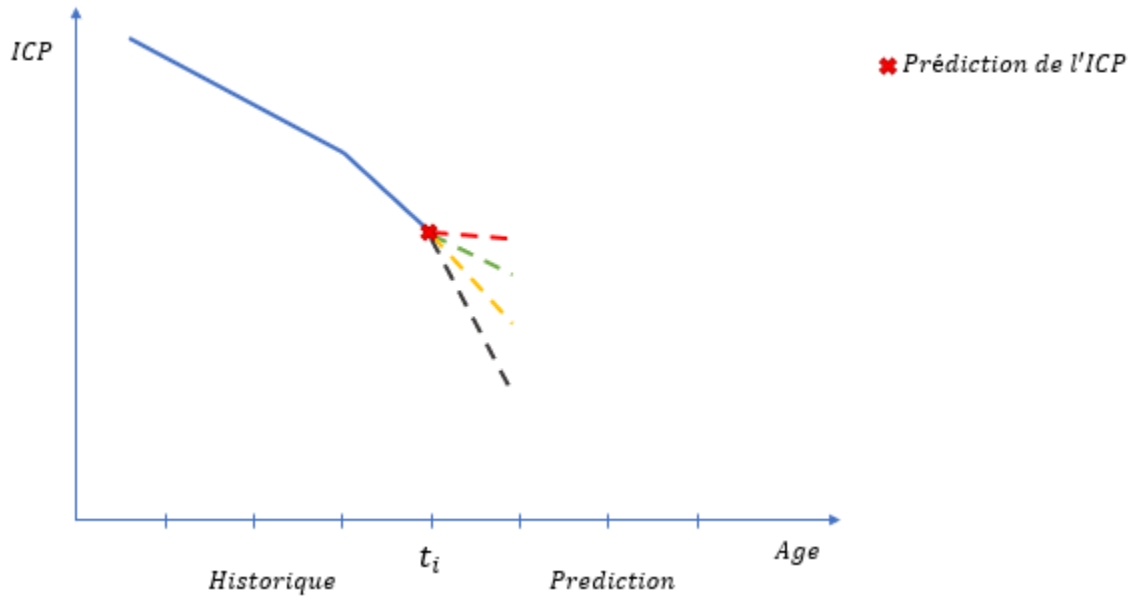


Figure 3.3 Principe de prédiction de l'ICP

3.4.2 Caractérisation des profils utilisant des séries cycliques

La caractérisation de profils de dégradation des cycles se fait en deux étapes : l'identification des signaux les plus importants dans l'extraction de chaque profil de dégradation et l'étiquetage des groupes par des seuils de ICP.

3.4.2.1 Identification des signaux explicatifs des profils de dégradation

Les contributions des signaux dans l'identification de chacun des profils de dégradation seront évaluées en utilisant le comportement normal ou non problématique comme référence. Afin d'identifier les paramètres qui rendent les séries de chaque groupe généré différentes du comportement normal, nous calculons la distance univariée entre les médoïdes de chaque profil de dégradation et celui du groupe caractérisant les cycles d'opération normale, sans dégradation. Une seule dimension (facteur) à la fois sera considérée avec le temps pour calculer la distance entre les profils de dégradation et le médoïde représentant le comportement normal. Ces distances univariées seront évaluées sur la base de la moyenne et de l'écart-type de la distance à l'intérieur du groupe de séries de référence. Le score d'importance d'un facteur dans l'identification d'un signal anormal est calculé à l'aide de l'Équation 3.7 :

$$Imp_{s_j}(P_i) = \frac{DTW_{s_j}(Mn, MP_i) - moy_{N \in \mathrm{Cn}}(DTW_{s_j}(Mn, N))}{\sigma_{N \in \mathrm{Cn}}(DTW_{s_j}(Mn, N))} \quad (3.8)$$

Où :

- $Imp_{s_i}(P_i)$: l'importance du signal s_i dans l'identification du profil P_i.
- $DTW_{s_j}(X, Y)$: la distance unidimensionnelle entre les séries X et Y qui prend en compte le temps et le signal s_j.
- Mn : Médoïde du groupe de séries de référence.
- MP_i : Médoïde du profil de dégradation i.
- Cn est le groupe de séries de référence.

Le score d'importance calculé à l'aide de l'Équation 3.7 sera utilisé pour classer les signaux par ordre de contribution pour chaque profil de dégradation. Ainsi, un score plus grand indique une plus grande influence dans l'identification du groupe en question.

3.4.2.2 Étiquetage des regroupements de séries cycliques

Les cycles opératoires sont généralement évalués par des ICPs qui permettent de caractériser la dégradation du système en mesurant sa performance globale. La liaison entre les profils générés et ces indicateurs sera réalisée par la classification supervisée. Nous allons utiliser l'arbre de décision. Il s'agit d'une représentation hiérarchique des relations de connaissance composée de nœuds et de liens. Les nœuds représentent les objectifs tandis que les liens sont utilisés pour les décisions. L'arbre de décision peut être linéarisé en règles de décision, où la classe à prédite est le contenu du nœud feuille, et les conditions le long du chemin forment une règle de décision (Shamim et al., 2010). Pour étiqueter les regroupements de séries, nous apprenons à un arbre de décision à prédire la classe de dégradation d'une séquence à partir de la performance du cycle mesuré par les ICPs du système. Les règles les plus importantes, celles possédant l'indice de Gini le plus bas, seront choisies pour caractériser chacun des profils.

3.4.2.3 Prévision des indices clés de performance

Après avoir identifié les signaux portant les symptômes de dégradation et étiqueté chaque profil, nous allons utiliser la technique de Forêts aléatoires pour prédire séparément les indicateurs clés de performance des actifs étudiés. Pour ce faire, nous extrayons 6 variables statistiques de chacune de dimension des cycles, notamment : la valeur moyenne, l'écart-type, le kurtosis, le Crest Factor, la valeur maximale et minimale des signaux. Ainsi, nous représentons chaque cycle par un vecteur de $6 \times s$ valeurs, où s est nombre de capteurs (dimension) utilisés. Ensuite, nous affectons à chaque pas de temps (année par exemple) le vecteur des valeurs moyennes des représentations de ces cycles.

Finalement, nous entrainons l'algorithme de Forêts aléatoires à prédire la prochaine valeur d'ICP en utilisant la représentation des signaux à chaque unité de temps. La prévision pas à pas de l'indicateur est effectué en utilisant le pseudo algorithme suivant :

Pour chaque année de l'horizon :

Calculer le vecteur représentant les signaux

Estimer la valeur de l'ICP de l'année suivante à partir de la représentation des signaux et de la valeur actuelle de l'indicateur par la technique de Forêts aléatoires

Mettre à jour la valeur de l'ICP

Une fois les indicateurs clés sont prédits à chaque pas de l'horizon, nous pouvons mettre les étiquettes des profils de dégradation en utilisant les règles des ICP générées précédemment.

Dans cette section, nous avons détaillé la méthodologie de modélisation de la dégradation des actifs en exploitation. Cette dernière permet de traiter des séquences courtes de ICP avec bruit ou des signaux de capteurs illustrant les cycles d'opération. La première phase de la méthodologie consiste à extraire les profils de dégradation en utilisant l'apprentissage non-supervisé. Pour ce faire, nous avons proposé une nouvelle mesure de similarité basée sur les pentes, les valeurs et l'âge des actifs pour traiter les séquences de ICP. En ce qui concerne les séries générées par les capteurs, nous avons adapté la distance DTW à capturer la cyclicité des signaux. Ces deux mesures de proximité sont ensuite intégrées dans une version modifiée de l'algorithme de clustering DBSCAN. Pour avoir une meilleure interprétabilité de l'évaluation de clustering basée sur la similarité, nous avons introduit deux versions modifiées des indices C et DB.

Les étiquettes générées par DBSCAN sont ensuite reproduites en utilisant la technique de Forêts aléatoires pour les séries de ICP et Time Series Forest pour les cycles d'opération. Cela nous a permis de prédire les profils de dégradation à partir des séquences d'observation. La dernière phase concerne l'interprétation des modèles trouvés pour caractériser les clusters et prédire l'ICP pour un horizon donné.

La méthodologie présentée sera appliquée sur deux études de cas, courtes séries et séries cycliques. Pour le traitement des séries courtes de ICP, nous allons utiliser une base de données de 2000 séries de dégradation de l'Indice d'État des Ponceaux (IEP). En ce qui concerne la deuxième variante de la méthodologie, nous utilisons une base de données de signaux enregistrés lors des cycles d'opération d'un système hydraulique.

CHAPITRE 4 ÉTUDES DE CAS

Ce chapitre est consacré à l'étude de deux cas d'application de la méthodologie proposée. Le premier cas concerne la modélisation des profils de dégradation de ponceaux de drainage. Les données sont constituées de courtes séries chronologiques décalées dans le temps relatif à l'état de performance global du ponceau. Le deuxième cas concerne la modélisation des profils de dégradation d'un système hydraulique à lente dégradation en se basant sur des séries chronologiques cycliques.

4.1 Cas des ponceaux de drainage

Les ponceaux de drainage constituent un élément essentiel de l'infrastructure de transport et de gestion de l'eau. Ces réseaux de canalisations d'eaux pluviales sont conçus pour acheminer les eaux des précipitations et des ruissèlements en-dessous des voies de circulation (routes, passages à niveau, sentiers, etc.). Généralement encastré sous terre, un ponceau peut être constitué d'un tuyau en polymère, de béton armé, en tôle ondulée ou d'un autre matériau. La Figure 4.1 présente un exemple de ponceau en tôle ondulée.

Figure 4.1 Ponceau en tôle ondulée

Comme tout système, les ponceaux commencent à se détériorer dès leur mise en service. L'intensité de la circulation, les charge excédent les limites, la circulation d'eau ainsi que les conditions climatiques sont tous des facteurs qui contribuent, au fil du temps, à endommager le ponceau. Les

conséquences de la dégradation peuvent aller d'un simple désagrément pour les usagers à l'effondrement du ponceau et de la chaussée (Transports Québec, 2012).

4.1.1 Présentation des données

Cette étude concerne principalement les ponceaux circulaires en tôle ondulée. Ce sous-type de ponceau est considéré par les experts du ministère comme étant le plus problématique parmi les réseaux de drainage au Québec. Le jeu de données compte plus de 2000 ponceaux inspectés périodiquement selon le Manuel d'inspection des ponceaux qui définit les programmes d'inspection du réseau de ponceaux en fonction de leur état de dégradation (Transports Québec, 2012).

L'état du ponceau est établi par un programme d'inspection périodique des principaux modes de dégradation tels que le mouvement et déformation, les défauts de matériaux, la fissuration et l'assemblage, la sédimentation, l'affouillement, l'infiltration et l'accumulation de débris. Ces modes de dégradation sont évalués sur une échelle décroissante de 5 à 1 servant à construire un Indice d'État du Ponceau (IEP). Le niveau 5 désigne une absence de défaut ou à un défaut négligeable et le niveau 1 désigne la présence d'un défaut majeur. L'IEP est calculé à partir de 17 critères de dégradation évalués lors de l'inspection en utilisant une formule proposée dans le manuel d'inspection des ponceaux (Transports Québec, 2012). Cet ICP possède une valeur numérique qui varie de 100 à 0. Une valeur proche de 100 désigne un ponceau neuf, exempt de défauts ou en bon état. Une valeur proche de 0 dénote un ponceau présentant des défauts majeurs ou hors d'usage. Une valeur intermédiaire correspond à la présence de défauts mineurs.

Les données comportent 2000 ponceaux en tôle ondulée inspectés sur une période de plus de 70 ans, mais de manière non régulière (discontinue). Dans un travail préliminaire, l'ensemble des données sont traitées afin d'extraire les fichiers de modélisation comportant les données sans anomalies et sans valeurs manquantes de l'IEP. Autres que les colonnes de l'IEP et d'âge du ponceau, les autres colonnes représentent principalement les facteurs définissant la structure du ponceau (diamètre, longueur, hauteur du remblai, etc.), l'environnement et les ouvrages associés au ponceau (position géographique, puisard, type de chaussée, etc.) ainsi que les facteurs d'utilisation (Débit journalier de véhicules, pourcentage des camions, etc.). L'Annexe A identifie et désigne l'ensemble des variables du ponceau.

4.1.2 Préparation des séquences de dégradation

Les données dont nous disposons pour la modélisation de la dégradation des ponceaux sont sous la forme de courtes séquences d'observations de l'IEP. À cause de la non-régularité de la périodicité de l'inspection, plusieurs profils de dégradation sont identifiables tels que des séquences à deux ou trois observations, décalées dans le temps, ou encore présentant des sauts. Un exemple de ces séquences est représenté dans la Figure 4.2 où chaque couleur caractérise une séquence de dégradation de l'IEP d'un ponceau en tôle ondulée au fil du temps.

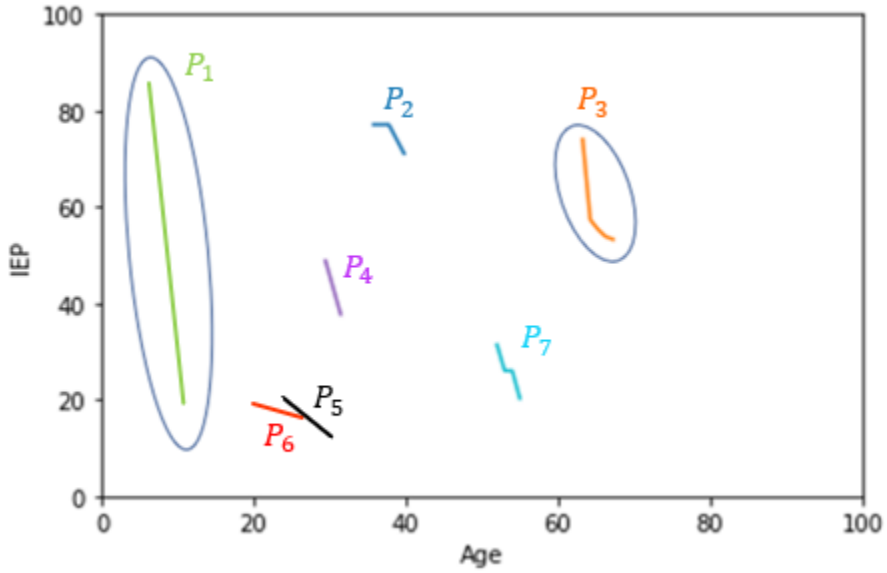


Figure 4.2 Exemples de séquences de dégradation de l'IEP

La Figure 4.2 représente 7 séries de séquences d'observations relatives à 7 ponceaux différents. Les deux séquences encerclées P_1 et P_3 ont une pente vraisemblablement similaire, mais apparaissent à deux âges différents. La similitude des profils de dégradation malgré la différence d'âges et des IEPs correspondants est généralement due aux interventions de réfection, de réparation ou de remplacement. Bien qu'un nettoyage et un traitement préliminaire des données aient été déjà réalisés, les données de modélisation présentent un bruit énorme provenant principalement des erreurs de mesure dues à la difficulté d'inspection de certains ponceaux et de saisie.

Dans le but de préparer les séquences de modélisation de la dégradation du ponceau, les ponceaux ayant une observation unique ont été retirés. Ensuite, comme les séquences sont relativement

constituées de 2 à 4 observations souvent étalées sur des âges éloignés, leurs longueurs sont inégales. De plus, certains ponceaux ont des séquences similaires, mais décalées dans le temps. Pour diminuer cette hétérogénéité, nous avons appliqué une interpolation linéaire afin de compléter les valeurs manquantes entre chaque couple d'observations consécutives. La Figure 4.3 présente un exemple d'interpolation d'une séquence à 3 observations. La valeur de l'interpolation est répliquée sur tous les âges intermédiaires avec un pas de 1 an.

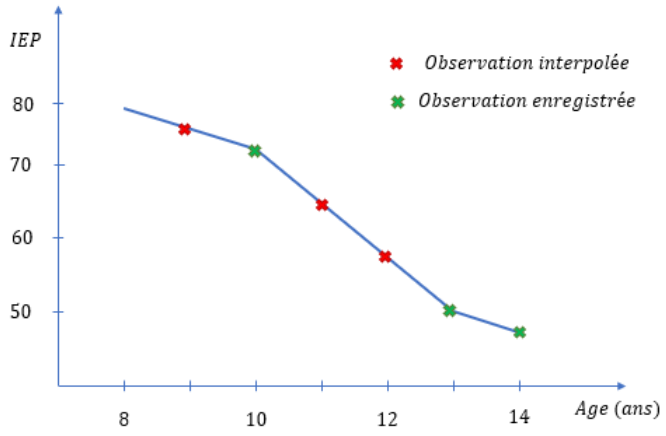


Figure 4.3 Interpolation des observations manquantes

Le résultat de ce prétraitement donne une matrice unique $M(m,n)$ où m et n désignent respectivement le nombre de ponceaux et l'âge maximal des ponceaux. Chaque ligne de $M(m,n)$ représente la séquence des pentes d'un ponceau donné et les colonnes indexent ces pentes par âge. La matrice $M(m,n)$ peut-être écrit comme suit :$M(m,n) = \left(p_{i,j}\right)_{1\leq i\leq m, 0\leq j\leq n}$

$$p_{i,j} = IEP_{i,j+1} - IEP_{i,j}$$

Où $IEP_{i,j}$ est la valeur de l'indice d'état du ponceau i à l'âge j et $\boldsymbol{p_{i,j}}$ désigne la pente de dégradation du ponceau $\boldsymbol{i}$ à l'âge $\boldsymbol{j}$. La Figure 4.4 représente un extrait de cette matrice. Par exemple, le 4^{e} ponceau (ligne 4 de la matrice) possède une pente de dégradation de 5,2 unités d'IEP par an à l'âge de 6 ans.

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
0																				
1																				
2																				
3																				
4					-6,75	-6,75	-5,2	-1,2	-4,3333											
5																				
6																				
7															-1,75	-2,55	-2,55	-2,55		
8																				
9																				
10																				
11								-3,66	-2,88	-4,33	-4,33	-4,33	-4,33							
12																				
13																				
14																				
15																				
16																				
17																				
18																				
19																				
20																				
21																				

Figure 4.4 Extrait de la matrice de pentes

4.1.3 Extraction des profils de dégradation de l'IEP

Après avoir préparé les séquences d'observations d'IEP, nous présentons les résultats des expériences sur le jeu de données selon la démarche de la Phase 1 de la méthodologie proposée. Comme mentionné dans la section 3.2.1.1, nous devons fixer 3 seuils pour la mesure de similarité qui sont : ε_{pente}, ε_{valeur} et ε_{age}. Ces paramètres ont été choisis en guise d'illustration.

- ε_{pente}= 0.6 points (10% de la pente moyenne de toutes les observations).
- ε_{valeur} = 5 points de CSI (le maximum est de 100 points).
- ε_{age} = 5 ans (l'âge maximum est de 80 ans).

La figure 4.5 illustre un exemple de la représentation des deux séries d'IEP *X* de longueur 3 et *Y* de longueur 5.

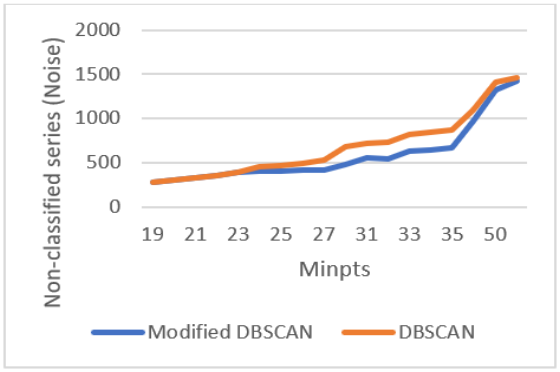
	Série X				Série Y					
Observation	x_1	x_2	x_3	*	y_1	y_2	y_3	y_4	y_5	*
IEP	80	75	68	66,5	83	82	77	71	65	64
Pente	-5	-7	-1,5	*	-1	-5	-6	-6	-1	*
Age	17	18	19	20	14	15	16	17	18	19

Figure 4.5 Représentation des deux séries d'observations d'IEP

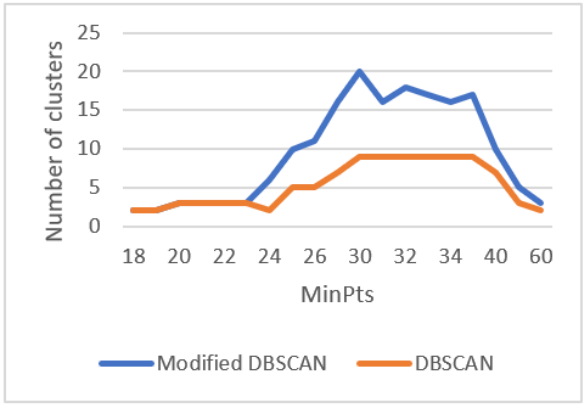
Les observations x_1 et y_2 sont considérées comme identiques parce que la différence entre leurs valeurs d'IEP, de pente et d'âge sont inférieures respectivement à 5 points, 0.6 points et 5 ans. La

longueur de la plus longue sous-séquence commune de pente est 2. Ainsi, la similarité entre *X* et *Y* calculée en utilisant l'équation 3.3 est de $\frac{2}{5}$.

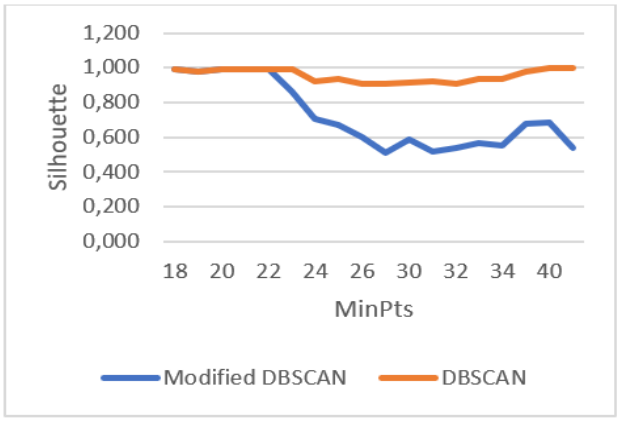
Tout d'abord, nous commençons par comparer l'algorithme DBSCAN-1 avec le DBSCAN classique en utilisant le même paramètre Eps (la similarité minimale entre deux séries pour les considérer voisines) fixé à 0,5. Le Figure 4.6 illustre les performances de deux algorithmes en termes de silhouette, de nombre de regroupements générés et de nombre de séries non classées.



(a) Nombre de ponceaux classés comme bruit.



(b) Nombre de regroupements générés.



(c) Score de silhouette.

Figure 4.6 Comparaison des performances de DBSCAN-1 et DBSCAN classique

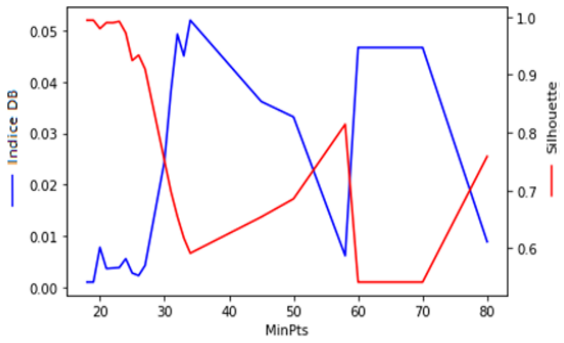
Les résultats montrent que l'algorithme DBSCAN modifié permet de générer plus de regroupements et moins de séries non classées (bruits). Il est important d'indiquer que nous avons

sacrifié la silhouette pour plus de profils de dégradation, mais nous enregistrons toujours un score de silhouette supérieur à 0.7%. Cette silhouette démontre qu'une structure forte est trouvée.

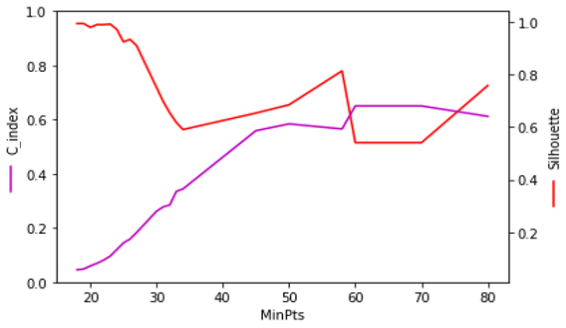
4.1.3.1 Comparaison des nouveaux indices de validité avec le score de silhouette

Puisque nous utilisons deux nouveaux indices de validation de clustering, nous devons les comparer avec le score de silhouette calculé à partir de l'Équation 2.2. Nous comparons également le score de silhouette avec la précision de la classification des clusters générés qui représente la fraction de séries correctement classées du nombre total de séquence.

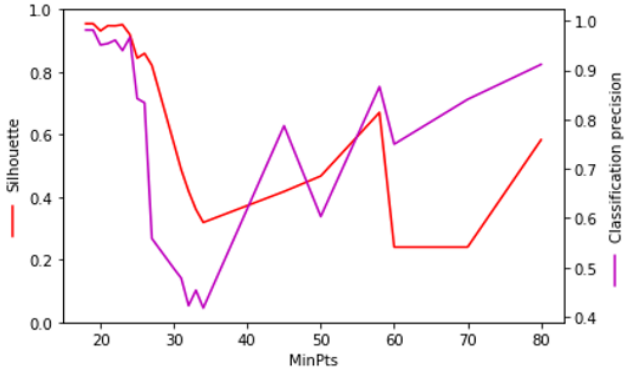
Les graphiques de la Figure 4.7 illustrent la réponse de l'indice DB modifié, l'indice C, la précision de la classification et le score de silhouette avec différents MinPts.



(a) Score de silhouette vs indice de DB modifié



(b) Score de silhouette vs indice C modifié



(c) Score de silhouette vs précision de la classification

Figure 4.7 Comparaisons des indices BD et C avec le score de silhouette

En comparant les allures des courbes de l'indice DB et de l'indice C avec la courbe de silhouette, nous constatons qu'elles ont un comportement opposé. Ainsi, les indices DB et C diminuent lorsque le score de silhouette augmente pour tous les MinPts. Ce qui s'aligne avec les attentes prévues lors de la définition de chacun des 3 indices de validité. Nous remarquons également que le score de silhouette et la précision de classification ont presque la même forme. Une meilleure silhouette indique une plus forte structure trouvée qui permet de reproduire plus facilement les étiquettes de clustering basées sur l'âge actuel ainsi que les variables explicatives.

4.1.3.2 Évaluation de la nouvelle mesure de similarité

Dans le but d'évaluer la performance de la nouvelle mesure de similarité présentée dans la section 3.2.1.1, nous allons la comparer avec les différentes mesures similarité. Pour ce faire, nous utilisons uniquement la plus longue sous-séquence commune des pentes pour mesurer la similarité entre deux séries de dégradation. Ensuite, nous prenons en compte les valeurs réelles de l'IEP et l'âge pour comparer deux séquences. Ainsi, nous comparons la similarité proposée pour la méthodologie avec 3 mesures seront utilisées :

- Similarité 1 : Basée sur la pente.
- Similarité 2 : Basée sur la pente et la valeur de l'IEP.
- Similarité 3 : Basée sur la pente et l'âge du ponceau.

Afin de garantir l'équité lors de la comparaison des différentes similarités, nous utilisons des recherches exhaustives sur l'espace de paramètres Eps dans le but de trouver la meilleure performance de regroupement pour chaque mesure. Ainsi, nous explorons l'espace de recherche des valeurs minimales 0,01 à 1. Le Tableau 4.1 présente les résultats des différentes mesures de similarité utilisées avec l'algorithme DBSCAN-1 pour un MinPts initial de 25 séries.

La similarité basée sur la plus longue sous-séquence commune de la pente, de l'IEP et de l'âge a réalisé le meilleur score de silhouette et les deuxièmes meilleurs indices DB et C. En plus des indices de validité de regroupement présentés, nous avons évalué les différentes mesures de similarité sur la base de la capacité de la méthode des forêts aléatoires à prédire le profil de dégradation d'un ponceau en prenant comme données d'entrée sa valeur l'IEP actuelle, l'âge du ponceau, les facteurs environnementaux et conceptuels ainsi que les facteurs d'exploitation. Nous avons constaté que la mesure de triple similarité a la meilleure précision de classification (92%).

C’est pourquoi nous retenons la technique de triple similarité avec l'algorithme DBSCAN-1 avec Eps = 0.3 et MinPts = 25 séries.

Tableau 4.1 Performance du DBSCAN-1 avec différentes configurations

	Similarité 1	**Similarité 2**	**Similarité 3**	**Similarité proposée**
Score silhouette	0,593	0,639	0,501	0,856
Indice DB modifié	0,060	0,036	0,063	0,056
Indice C modifié	0,558	0,120	0,350	0,134
Nombre de clusters	12	3	20	6
Nombre de séries non classées	940	131	540	401
Précision globale de la classification	47,79	87,75	50,91	91,94
Eps	0,7	0,5	0,5	0,3

Tableau 4.2 Contribution des variables explicatives dans le clustering

Variables	**Importance (%)**
Age	39,8
IEP	16,5
Température maximale moyenne	8,1
Précipitation moyenne	6,9
Température maximale moyenne	6,4
Longueur	6,4
Hauteur du remblai	5,4
Largeur du diamètre	3,7
Angle d’inclinaison	2,8

4.1.4 Contribution des variables explicatives dans les regroupements

Une fois les profils de dégradation extraits et validés, nous avons utilisé la technique des forêts aléatoires pour prédire les étiquettes des regroupements et mesurer l'influence de l'âge, de la valeur de l'IEP ainsi que des variables relatives à la structure, à l'environnement et à l'exploitation des ponceaux dans la distinction des différents profils de dégradation trouvés. Le Tableau 4.2 présente les résultats obtenus.

L'analyse des résultats montre que l'âge et la valeur de l'IEP sont les facteurs qui contribuent le plus à la classification. Les résultats mettent également en évidence l'influence remarquable des facteurs climatiques. En effet, la température et les précipitations représentent 21,5% d'importance dans la différenciation des profils. Ces variables représentent la dégradation causée par l'environnement. Les facteurs conceptuels sont également importants. Ainsi, la longueur du ponceau, son diamètre et son degré d'inclinaison contribuent à plus de 10% dans la classification.

4.1.5 Modèle de prédiction de l'IEP basée sur les profils de dégradation

Nous avons maintenant atteint la dernière phase de la méthodologie qui consiste à prédire récursivement la valeur de l'IEP. Le Tableau 4.3 résume les performances de prédiction obtenues par le modèle proposée et les modèles classiques en utilisant les données de test. Les modèles ont été entrainés et évalués en utilisant la validation k-folds.

Tableau 4.3 Performance de la prédiction de la valeur IEP du ponceau

Modèles de prédiction	**Coefficient R^2**	**R^2 ajustée**
Régression linéaire	32,8	9,7
Réseau de neurones	21,9	3,6
Random Forests (60 arbres)	62,8	40,7
SVM (linear kernel)	28,8	7,2
Extra Trees (75 arbres)	61,7	40,2
Gradient Boosting	44	18,4
Gaussian Process	30,9	8,5
Modèle de prédiction proposé	**80,5**	**64,6**

Ces résultats prouvent la capacité du modèle de prédiction basée sur l'exploitation des profils de dégradation à capturer le phénomène physique de la dégradation lente des ponceaux. La qualité du nouveau modèle dépasse celles obtenues par les modèles d'intelligence artificielle conventionnels en atteignant un coefficient de détermination de 80,5 %. La figure 4.8 illustre un exemple de la prédiction de l'IEP pour un horizon de 8 ans. La séquence prédite est comparée à celle observée.

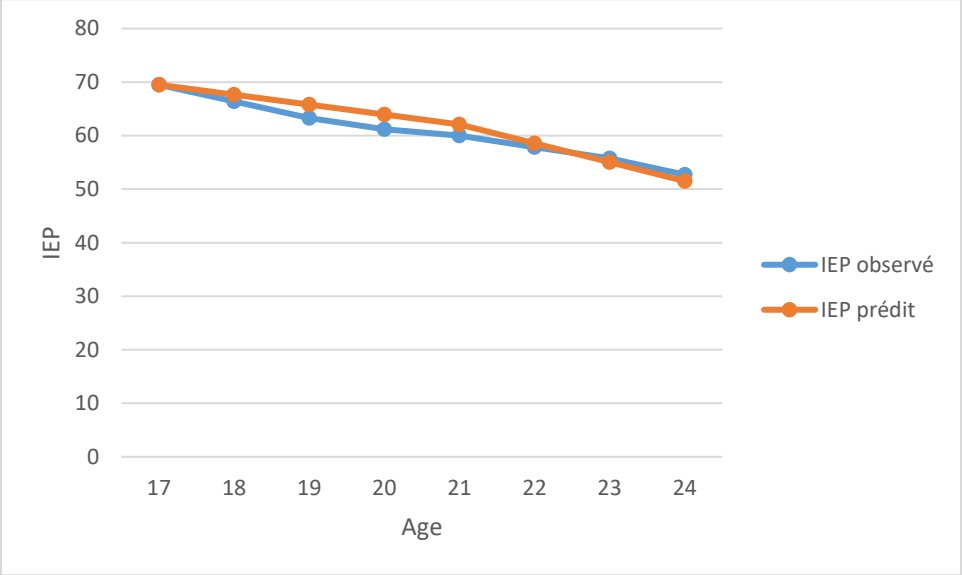


Figure 4.8 Comparaison des séries d'IEP prédite et observée

4.2 Étude de cas d'un système hydraulique

4.2.1 Présentation de la base de données

Les systèmes hydrauliques sont des actifs qui se dégradent lentement pendant leur exploitation. Ces derniers sont souvent surveillés par des capteurs installés sur leurs différentes composantes. Les séries cycliques générées par ces capteurs peuvent être utilisées pour modéliser la dégradation de ces actifs évaluée par des ICPs. Ces signaux peuvent être collectés en surveillant le fonctionnement du système étudié dans son environnement ou par des protocoles de test au laboratoire pour accélérer les phénomes étudiés.

Dans ce cas d'étude, nous utilisons une base de données proposée par (Helwig et al., 2015). Cette base concerne la dégradation des circuits primaire et secondaire du système hydraulique. Le

système étudié est représenté dans la Figure 4.9. Le circuit primaire (a) de travail est composé d'une pompe principale MP1 avec orifice commutable V9, 4 accumulateurs commutables A1-A4 avec différentes pressions de charge variable V11. Le circuit secondaire (b) de refroidissement et de filtration avec refroidisseur C1. Les deux circuits sont reliés par le réservoir d'huile.

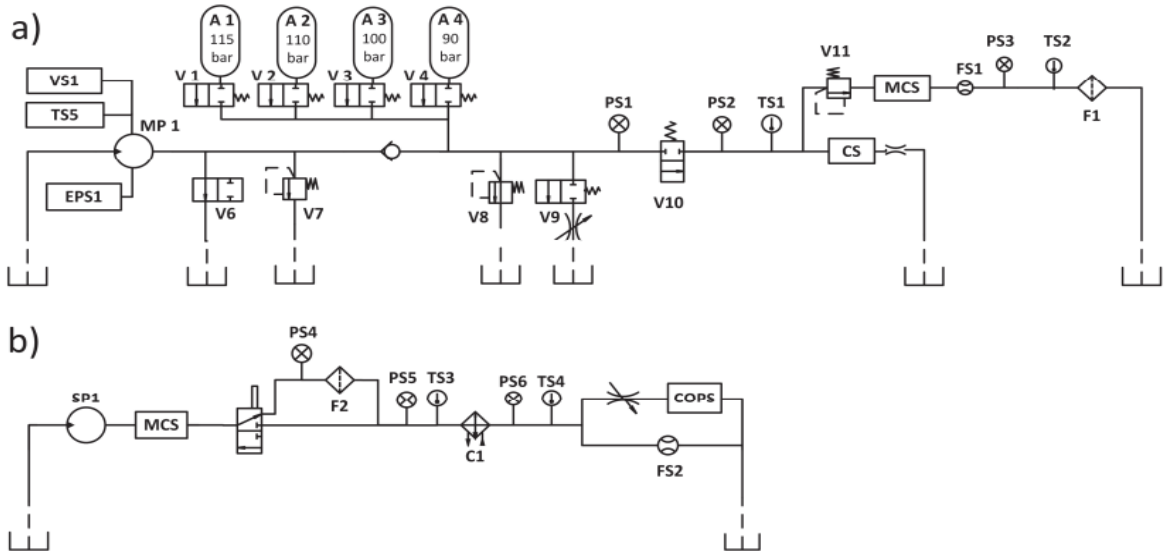


Figure 4.9 Système hydraulique pour la collecte des données

Afin de simuler cette base de données, un banc d'essai hydraulique qui permet un changement réversible de l'état ou de la condition de divers composants a été développé par (Helwig et al., 2015). Dans le circuit primaire, différents niveaux de charge sont répétés cycliquement avec la soupape de pression proportionnelle V11. Pour représenter le fonctionnement cyclique typique du système, ce dernier est sollicité à opérer par cycles de travail fixes avec des niveaux de charge prédéfinis. Le protocole de l'essai consiste à opérer le système par un cycle de 60 secondes. 17 capteurs différents sont installés sur les différents composants du système pour surveiller son fonctionnement. Le Tableau 4.4 résume les différents capteurs utilisés.

Tableau 4.4 Liste des capteurs du système

Capteur	**Grandeur physique**	**Unité**	**Fréquence**
PS1-6	Pression	bar	100Hz
EPS1	Puissance de moteur	W	100Hz
FS1-2	Volume de flux	l/min	10Hz
TS1-4	Température	C	1Hz
VS1	Vibration	mm/s	1Hz
CE	Efficacité de refroidissement	%	1Hz
SE	Facteur d'efficacité	%	1Hz

Le comportement cyclique peut être observé facilement sur les séries temporelles générées par les capteurs. Les Figures 4.10 et 4.11 représentent les signaux de la base de données relatifs à la puissance du moteur 1 et la pression PS1.

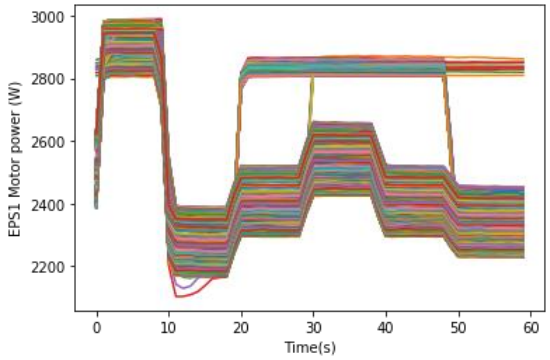


Figure 4.10 Puissance du moteur

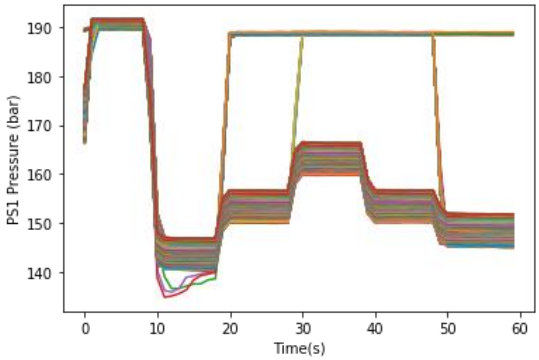


Figure 4.11 Pression PS1

En plus des séries temporelles de signaux générés par les capteurs, la performance globale de chaque cycle d'opération est mesurée par 5 indicateurs clés de performance : ICP-1 Condition du refroidisseur, ICP-2 Condition de valve, ICP-3 Fuite interne de la pompe, ICP-4 Pression dans l'accumulateur hydraulique et ICP-5 Stabilité de système. Ainsi, chaque série multidimensionnelle cyclique est évaluée par 5 valeurs d'ICPs dont les interprétations sont résumées dans le Tableau 4.5.

Tableau 4.5 ICP des cycles et leurs interprétations

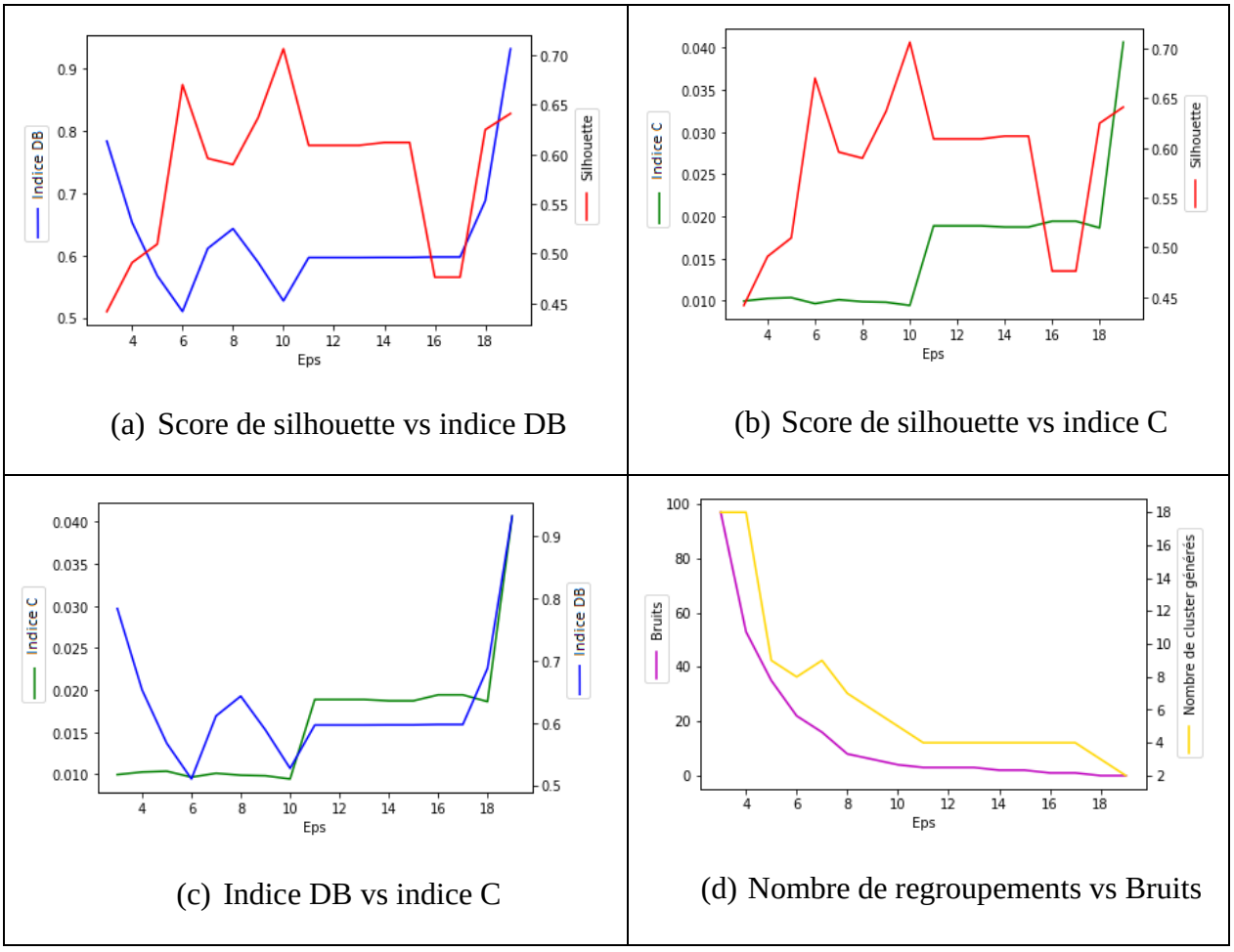
Indice clé de performance	Interprétation
État du refroidisseur (%)	• 3 : proche de la panne totale • 20 : efficacité réduite • 100 : pleine efficacité
État de la vanne (%)	• 100 : condition optimale • 90 : petit décalage • 80 : décalage important • 73 : proche de la défaillance totale
Fuite de la pompe interne	• 0 : pas de fuite • 1 : faible fuite • 2 : fuite importante
Accumulateur hydraulique	• 130 : pression optimale • 115 : pression légèrement réduite • 100 : pression fortement réduite • 90 : proche de la panne totale
Indicateur de stabilité	• 0 : les conditions étaient stables • 1 : les conditions statiques n'ont pas encore été atteintes

La base de données étudiée est constituée par 2205 séries chronologiques. Chaque signal correspond à un circuit indépendant. En se basant sur le Tableau 4.5, nous avons pu isoler 11 cycles qui correspondent à un fonctionnement normal du système. Ces séquences dites normales seront utilisées comme une référence dans l'analyse. L'objectif sera d'utiliser la méthodologie proposée pour extraire les profils de dégradation du système hydraulique et ensuite les caractériser et les interpréter en se basant sur les ICP des cycles.

4.2.2 Extraction des profils de dégradation

Pour extraire les profils de dégradation, nous procédant au clustering des signaux cycliques en utilisant l'algorithme DBSCAN-2. Vu que ce regroupement est basé sur les distances, nous allons utiliser les indices DB et C originaux présentés dans les sections 2.2.5 et 2.2.5 ainsi que la silhouette pour évaluer la qualité de partition. Cette base de données ne présente pas un bruit considérable, c'est pourquoi nous allons mettre le nombre minimal de séries pour considérer un cluster MinPts à 2. Cela va nous permettre d'identifier le maximum de clusters. En ce qui concerne le paramètre Eps, nous allons explorer tout le domaine de ses valeurs possibles.

Les tests effectués cherchent la valeur optimale d'Eps puisque cette dernière permet d'identifier la plus forte structure de regroupements. La meilleure valeur de silhouette de 71% correspond à Eps =10. Pour comparer la performance de DBSCAN-2 couplé avec la distance DTW avec et sans la variable temporelle, nous avons refait le clustering avec les signaux seulement. Nous avons obtenu une plus faible structure avec une silhouette de 65%. La Figure 4.12 présente les résultats de la Phase 1 de la méthodologie.



(a) Score de silhouette vs indice DB

(b) Score de silhouette vs indice C

(c) Indice DB vs indice C

(d) Nombre de regroupements vs Bruits

Figure 4.12 Qualité de la classification pour différentes valeurs d'epsilon (Eps)

Comme attendu d'après la revue de littérature, la courbe du score de silhouette est inversement proportionnelle aux indices DB et C. Ainsi, le plus grand score de silhouette correspond au minimum global de l'indice C (0,0093) et à la deuxième plus petite valeur de l'indice DB (0,51). Comme illustré dans la Figure 4.9-d, l'augmentation de la valeur de la distance minimale pour considérer deux séries comme voisines pousse l'algorithme à générer plus de regroupements et

moins de bruits. Ainsi, avec une valeur optimale d'Eps de 10, l'algorithme DBSCAN-2 a permis l'extraction de 5 regroupements dont la répartition est illustrée dans le Tableau 4.6.

Tableau 4.6 Répartition des séries données par la classification optimale

Regroupements	1	2	3	4	5	Non classée
Nombre de séries	491	13	227	730	729	4

4.2.3 Entrainement du classificateur des profils

Une fois les regroupements extraits, nous utilisons la technique Time Series Forest pour régénérer ces étiquettes. L'algorithme prend les signaux multidimensionnels comme entrée et retourne la classe correspondante au cycle. L'entrainement du modèle a été effectué en utilisant la validation croisée 10-folds. Ainsi, les précisions obtenues pour les données d'entrainement, de validation et de test sont respectivement de 99%, 98% et de 98%. Ce qui prouve la capacité du modèle TSF à reproduire la logique du clustering.

4.2.3.1 Identification des signaux importants pour chaque profil

Le tableau 4.7 présente les scores d'importance des signaux pour les 5 profils de dégradation, calculés comme expliqué dans la Section 3.4.2.1.

Tableau 4.7 Score d'importance par rapport à l'identification des profils de dégradation

Signal	Profil 1	Profil 2	Profil 3	Profil 4	Profil 5
CE	208,87	219,14	220,49	150,44	-0,92
CP	169,03	190,62	191,60	115,72	-1,86
EPS1	36,27	51,88	44,53	18,25	1,82
FS1	3,98	89,18	74,45	6,04	5,84
FS2	441,13	535,15	568,57	218,46	2,10
PS1	22,38	90,16	74,40	15,06	4,58
PS2	112,40	677,52	533,28	112,20	82,48
PS3	117,58	421,35	366,83	104,07	39,55
PS4	55,24	55,24	55,24	55,24	-0,52
PS5	64,99	69,27	70,74	36,68	1,79
PS6	64,90	69,05	70,59	36,63	1,85
SE	3,45	62,53	51,63	3,47	4,74
TS1	125,51	141,37	144,71	60,79	1,20
TS2	157,93	171,54	175,91	74,81	0,87
TS3	162,59	185,57	190,18	79,00	1,32
TS4	168,60	191,81	195,84	85,80	1,57

Les scores fournis dans le Tableau 4.7 peuvent être analysés en traçant l'histogramme des contributions des signaux pour chaque profil. La figure 4.13 représente l'exemple du premier profil de dégradation (cluster 1). Les histogrammes des 5 profils seront fournis dans l'annexe B.

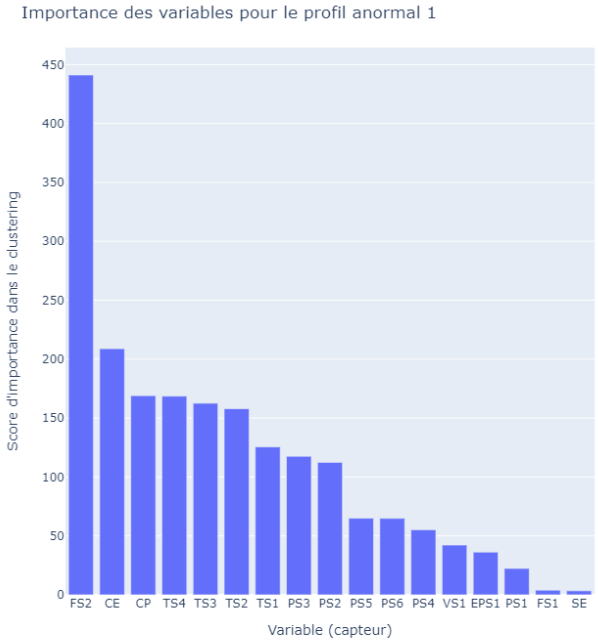


Figure 4.13 Importance des signaux pour le profil 4

En analysant l'histogramme de la figure ci-dessus, nous constatons que le volume de flux FS2 et l'efficacité de refroidissement CE sont les signaux qui font la différence entre le comportement optimal et le profil de dégradation 1. Afin de mieux analyser chacun des profils générés, nous comparons les signaux les plus importants avec ceux du comportement normal (sans dégradation). Pour ce faire, nous traçons les séries unidimensionnelles des médoïdes des profils de dégradation avec ceux du profil normal.

Les signaux des médoïdes des comportements normaux et du profil 1 sont représentés dans les Figures 4.14 et 4.15.

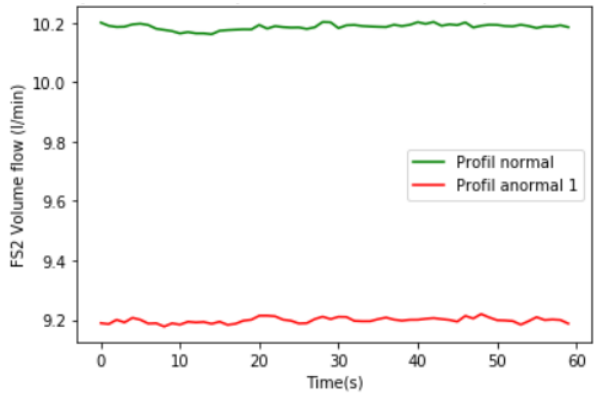


Figure 4.14 Volume de flux

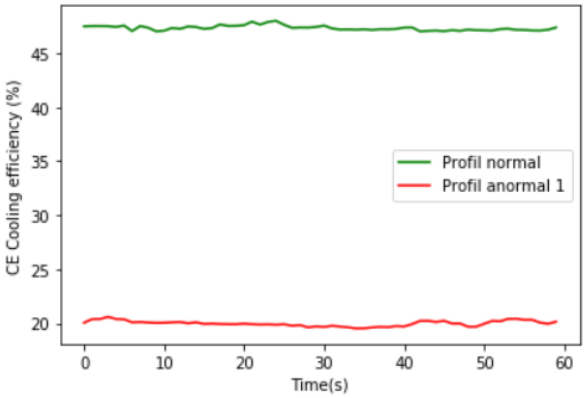


Figure 4.15 Efficacité de refroidissement

La différence entre le comportement optimal et celui de dégradation est claire dans les deux figures précédentes. En effet, pour ce profil, le volume de flux est baissé à une moyenne de 9.2 l/min comparée à 10.2 l/min pour le comportement normal. De même pour l'efficacité de refroidissement qui a passé d'une moyenne de 48 à 20%.

La dégradation des cycles du deuxième profil peut être observée au niveau des pressions PS2 et PS3 illustrées dans les Figures 4.16 et 4.17.

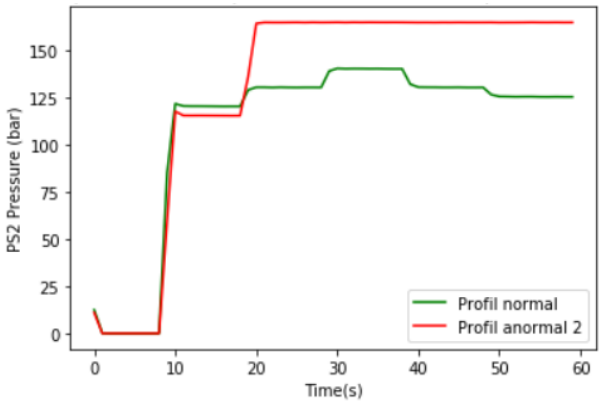


Figure 4.16 Pression PS2

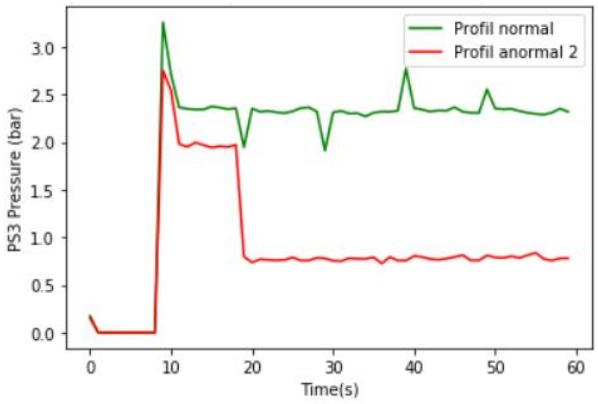


Figure 4.17 Pression PS3

En analysant les deux pressions tracées ci-dessus, nous constatons que ce profil de dégradation est caractérisé par une hausse de pression PS2 et une baisse de pression PS3 après le 1er tiers de cycle.

Les signaux les plus importants dans l'identification du profil de dégradation sont tracés dans les Figures 4.18 et 4.19.

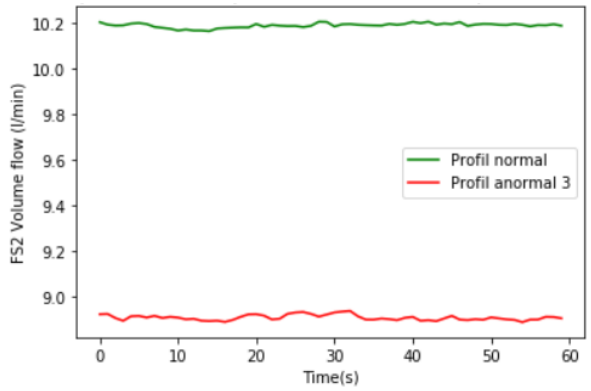


Figure 4.18 Volume de flux

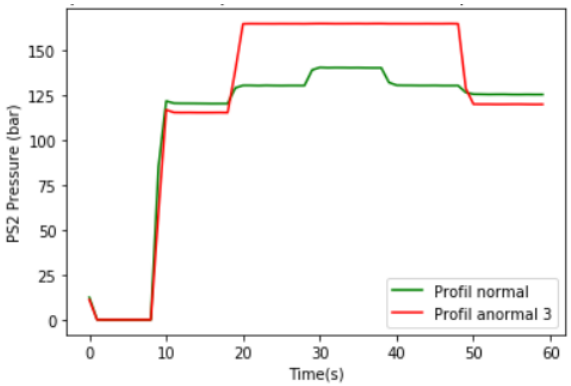


Figure 4.19 Pression PS2

Nous remarquons un écart important de 10 % entre le volume de flux du comportement normal et celui de dégradation 3 ainsi qu'une hausse d'environ 25 bars pour la pression PS2 dans l'intervalle 20 à 50 secondes.

En ce qui concerne le cluster 4, sa dégradation peut être expliquée d'une manière similaire à celle du profil 1. La différence entre les deux réside dans l'ampleur de la dégradation. Ainsi, tel qu'illustré dans les Figures 4.20 et 4.21, les baisses de volume de flux et d'efficacité de refroidissement sont plus prononcées pour le profil1.

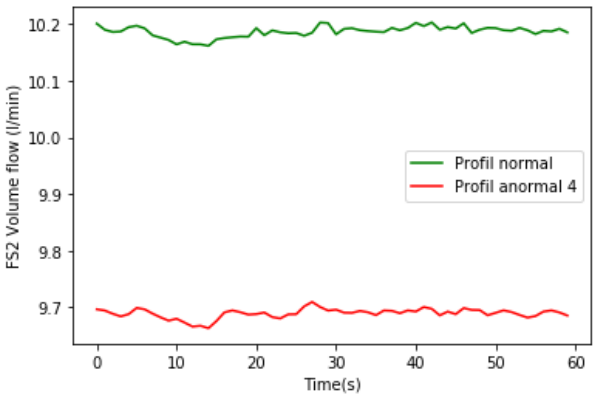


Figure 4.20 Volume de flux

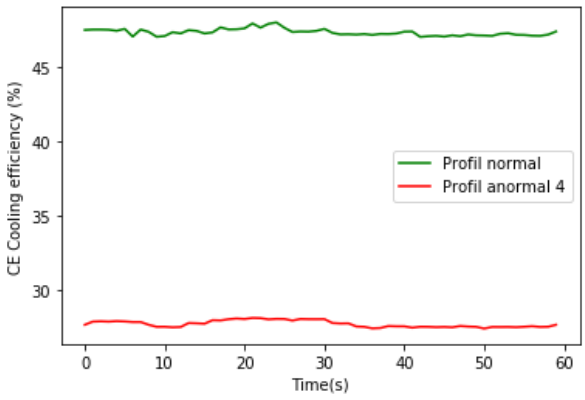


Figure 4.21 Efficacité de refroidissement

Le dernier profil de dégradation, les signaux qui font les plus grandes différences avec le comportement normal sont les pressions PS2 et PS3 représentés dans les Figures 4.22 et 4.23.

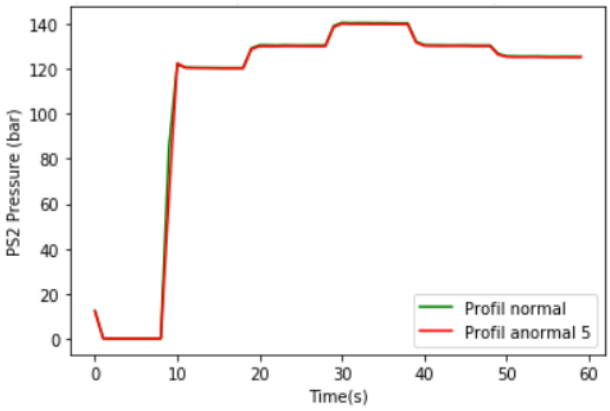


Figure 4.22 Pression PS2

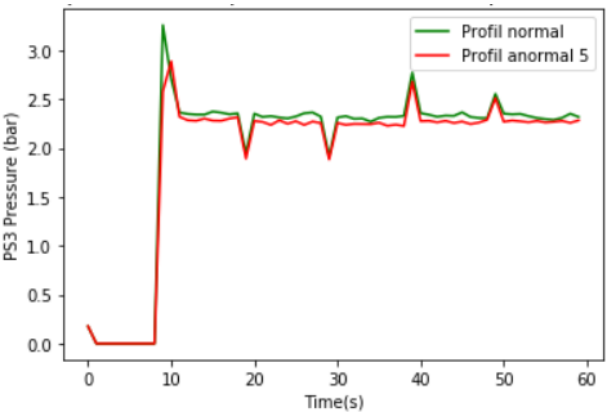


Figure 4.23 Pression PS3

En analysant les deux signaux, nous remarquons que l'écart entre ces deux pressions n'est pas assez prononcé. Cela indique une légère dégradation qui se manifeste avec un comportement proche du fonctionnement optimal.

4.2.4 Prédiction des profils de dégradation par les ICPs

Après avoir identifié les signaux les plus importants pour les profils de dégradation, nous utilisons un arbre de décision (AD) qui prend en entrée les 5 ICPs représentant l'état de dégradation de chacun des cycles et retourne la classe correspondante. En utilisant la méthode de validation croisée k-folds, la précision du modèle de prédiction obtenue est de 94,6% pour l'ensemble de test.

Afin d'éviter d'avoir des règles très spécifiques pour un nombre limité de cycles, nous avons utilisé l'indice de Gini pour limiter la décomposition de l'arbre de décision. Ainsi, cela nous a permis de conserver les règles représentatives ayant les plus petits indices de Gini. Ces règles sont illustrées dans le Tableau 4.8. La Figure 4.24 présente une version réduite de l'arbre généré (nous présentons seulement les 5 premiers nœuds de l'arbre généré).

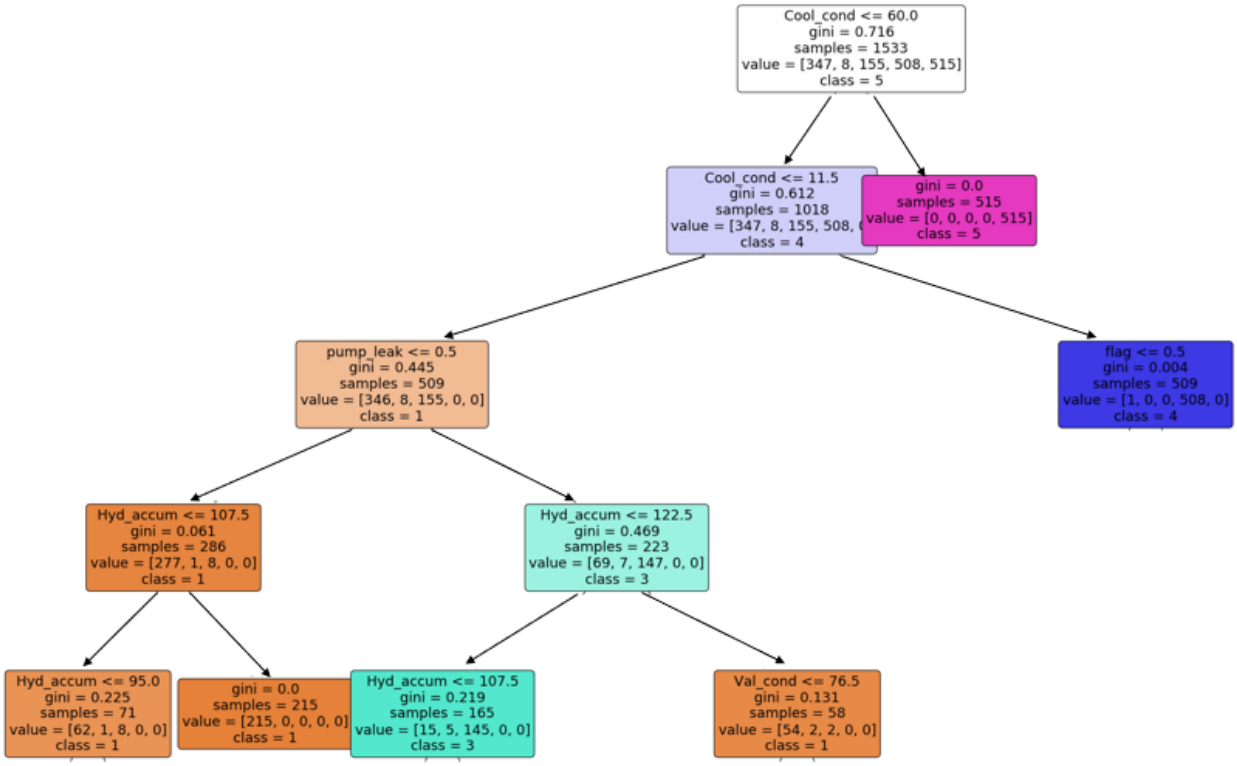


Figure 4.24 Arbre de décision réduite

La combinaison des règles caractérisant chaque profil avec ses signaux les plus importants permet d'identifier les capteurs associés à chaque mode de dégradation. Par exemple, le Profil 3 correspond à des cycles qui travaillent en pleine efficacité de refroidissement, sans fuite dans la pompe interne, mais avec une pression réduite au niveau l'accumulateur hydraulique.

Tableau 4.8 Caractérisation des profils de dégradation

Profils de dégradation	**Caractérisation (règle de prédiction)**	**Signaux les plus importants**
Profil 1	Cool_cond $\leq 11{,}5$ et $Pump\_leak = 0$	FS2, CE et CP
Profil 2	Cool_cond $\leq 11{,}5$ et stabilité = 1	PS2, FS2 et PS3
Profil 3	Cool_cond ≤ 60 et $Pump\_leak = 0$ et $Hyd\_accum \leq 107$	FS2, PS2 et PS3
Profil 4	$11{,}5 \leq$ Cool_cond ≤ 60 et stabilité = 0	FS2, CE et CP
Profil 5	Cool_cond ≥ 60	PS2, PS3 et FS1

En considérant les résultats représentés dans la Section 4.2.3.1 et dans le Tableau 4.8, nous constatons que le signal FS2 participe dans l'identification de la majorité des profils de dégradation extraits. Cela indique que ça sera intéressant de surveiller cet indicateur pour détecter des défauts futurs. Les profils 1 et 4 correspondent à un même défaut, mais avec deux niveaux différents de criticité. Ainsi, ces deux regroupements correspondent soit à une défaillance totale ou à une réduction de l'efficacité de refroidisseur où le système est n'est pas stable. Le profil 5 correspond à une légère réduction de l'efficacité de refroidissement ce qui explique la similarité entre ces signaux et ceux du comportement normal.

Par ailleurs, vu que nous ne disposons pas de l'information par rapport à l'ordre chronologique des cycles de chacun des actifs étudiés, nous ne pouvons pas appliquer l'algorithme de prévision des ICPs pour cette étude de cas.

CHAPITRE 5 CONCLUSION ET RECOMMANDATIONS

Le présent mémoire a introduit une nouvelle méthodologie dérivée de l'intelligence artificielle pour modéliser la dégradation lente des actifs en exploitation. Le développement des modèles de détérioration est basé sur le traitement des séries chronologiques relatif à deux niveaux différents d'études, soit les séquences d'indicateurs clés de performance ou les signaux générés au niveau de chaque cycle d'opération. Chaque niveau d'études résulte d'une structure spécifique de données qui présente des défis particuliers. Ainsi, les séquences des ICPs sont généralement courtes, bruitées, de différentes longueurs et fréquences. Pour les séries générées par les capteurs, il s'agit des bases de données plus riche, mais beaucoup plus compliquée.

La méthodologie proposée consiste principalement en 3 phases, elle extrait d'abord les profils de dégradation. Pour ce faire, nous avons utilisé l'algorithme de clustering DBSCAN-1 basé sur la similarité de l'âge, des pentes des actifs et des valeurs de l'indice de performance pour regrouper ces courtes séries temporelles d'observations. En ce qui concerne les signaux des capteurs, nous avons proposé une version modifiée de la distance DTW qui permet de capter la cyclicité des séries temporelles en considérant le temps restant de la phase comme une variable. Cette distance a été couplée avec l'algorithme DBSCAN-2 pour partitionner les cycles.

L'évaluation de la partition des données a été basée principalement sur la mesure de silhouette avec ses deux variantes pour le clustering basé sur la similarité et sur la dissimilarité. Les indice Davies-Bouldin et C ont été utilisés pour consolider l'évaluation. Vu que ces deux indices sont conçus pour interpréter des partitions fondées sur des logiques de dissimilarité, nous avons développé 2 nouveaux indices de validité de clustering appelés Davies-Bouldin modifié et C-index modifié pour pouvoir évaluer le clustering basé sur la similarité. L'étape suivante consiste à utiliser des techniques de classification supervisée pour prédire les clusters des séries temporelles étudiées. Nous avons utilisé la technique de Random Forest pour estimer le profil de dégradation de chacune des séries d'ICP en prenant comme entrée l'état de performance, l'âge et les facteurs conceptuels de chaque actif. La classification des signaux cycliques a été assurée par une variante de la forêt aléatoire qui extrait des variables statistiques à partir des différentes dimensions des séries. Enfin, nous pouvons prédire l'indice de performance global en utilisant les pentes des classes. La dernière phase de la méthodologie consiste à exploiter les modèles pour caractériser et interpréter les profils de dégradation identifiés. Au niveau des observations d'ICP, nous avons exploité les forêts

aléatoires déjà entrainé pour identifier les variables explicatives qui explique les différences entre les profils de dégradation et pour prédire l'état futur des ICPs étudiés. L'étiquetage des clusters de séries cycliques est effectué en deux parties, premièrement on identifie signaux importants pour chaque profil en utilisant le score d'importance. Ensuite, nous avons utilisé l'arbre de décision pour générer les règles d'ICP pour chaque cluster. Cela permet d'expliquer la réponse des signaux cycliques par rapport aux indicateurs de performance.

Nous avons utilisé deux cas d'études afin de valider la méthodologie. La première étude est basée sur des séquences d'indice d'état des ponceaux et la deuxième a exploité les signaux cycliques relatifs au fonctionnement d'un système hydraulique. Pour le premier cas d'étude, nous avons traité un jeu de données de dégradation de ponceaux circulaires en tôle ondulée. Les résultats montrent que cette technique donne de meilleurs résultats que les techniques classiques d'apprentissage automatique en termes de qualité de prédiction. Les résultats sont concluants et répondent aux attentes des gestionnaires des réseaux de ponceaux. En appliquant la méthodologie sur le système hydraulique, nous avons pu confirmer la capacité de la méthodologie à capturer la cyclicité des séries chronologiques, à extraire des regroupements interprétables à partir des signaux multidimensionnels. En plus de la caractérisation des clusters, nous avons effectué le diagnostic des profils de dégradation en établissant la liaison entre les ICPs et les signaux portant les symptômes de la détérioration. En utilisant une approche hybride de techniques d'apprentissage supervisé et non supervisé, nous avons relevé deux défis principaux : capturer les phénomènes de dégradation à partir d'un ensemble de données limitées et bruitées sans tomber dans le surapprentissage, et diagnostiquer la dégradation au niveau des cycles opératoire.

La technique proposée permet de modéliser efficacement l'état de performance des actifs, ce qui représente la composante clé d'une planification optimale de la maintenance. Cette méthodologie peut traiter des séries longues d'ICP. Mais ça va prendre énormément de temps. Une amélioration intéressante sera d'optimiser le temps de calcul des similarités entre les séries et de la prévision de l'ICP. Dans le cas des séries cycliques, la prévision des indicateurs clés de performance ne prend pas en compte les interactions entre les différents ICPs. On peut perfectionner la technique de prévision dans un futur travail en la rendant capable de capturer les corrélations entre les différents modes de défaillance ainsi que les tendances qui peuvent être observées sur les signaux générés par les capteurs.

RÉFÉRENCES

Aboura, K., Samali, B., Crews, K., & Li, J. (2008). *Stochastic processes for modelling bridge deterioration.* Futures in Mechanics of Structures and Materials-Proceedings of the 20th Australasian Conference on the Mechanics of Structures and Materials, ACMSM20.

Aghabozorgi, S., Shirkhorshidi, A. S., & Wah, T. Y. (2015). Time-series clustering–a decade review. *Information Systems, 53*, 16-38.

Agrawal, A. K., Kawaguchi, A., & Chen, Z. (2010). Deterioration rates of typical bridge elements in New York. *Journal of Bridge Engineering, 15*(4), 419-429.

Albalate, A., & Minker, W. (2013). *Semi-supervised and unsupervised machine learning: novel strategies*. John Wiley & Sons.

Ana, E., & Bauwens, W. (2010). Modeling the structural deterioration of urban drainage pipes: the state-of-the-art in statistical methods. *Urban Water Journal, 7*(1), 47-59.

Ana, E. V., & Bauwens, W. (2010). Modeling the structural deterioration of urban drainage pipes: the state-of-the-art in statistical methods. *Urban Water Journal, 7*(1), 47-59. https://doi.org/Pii 919491103

10.1080/15730620903447597

Arbelaitz, O., Gurrutxaga, I., Muguerza, J., Pérez, J. M., & Perona, I. (2013). An extensive comparative study of cluster validity indices. *Pattern Recognition, 46*(1), 243-256.

Aßfalg, J., Kriegel, H.-P., Kröger, P., Kunath, P., Pryakhin, A., & Renz, M. (2006). *Similarity search on time series based on threshold queries.* International Conference on Extending Database Technology (p. 276-294).

Baik, H.-S., Jeong, H. S., & Abraham, D. M. (2006). Estimating transition probabilities in Markov chain-based deterioration models for management of wastewater systems. *Journal of water resources planning and management, 132*(1), 15-24.

Belanche, L. A. (2012). Understanding (dis) similarity measures. *arXiv preprint arXiv:1212.2791*.

Bezdek, J. C., Moshtaghi, M., Runkler, T., & Leckie, C. (2016). The generalized c index for internal fuzzy cluster validity. *IEEE Transactions on Fuzzy Systems, 24*(6), 1500-1512.

Breiman, L. (2001). Random forests. *Machine learning, 45*(1), 5-32.

Brockwell, P. J., & Davis, R. A. (2009). *Time series: theory and methods*. Springer Science & Business Media.

Brooks, S., Gelman, A., Jones, G., & Meng, X.-L. (2011). *Handbook of markov chain monte carlo*. CRC press.

Bu, G. (2013). *Development of an integrated deterioration method for long-term bridge performance prediction*, Griffith University Queensland, Australia].

Bubtiena, A. M., Elshafie, A. H., & Jafaar, O. (2011). *Application of artificial neural networks in modeling water networks.* 2011 IEEE 7th International Colloquium on Signal Processing and its Applications (p. 50-57).

Cadei, L., Corneo, A., Milana, D., Loffreno, D., Lancia, L., Montini, M., . . . Carducci, F. (2019). *Advanced Analytics for Predictive Maintenance with Limited Data: Exploring the Fouling Problem in Heat Exchanging Equipment.* Abu Dhabi International Petroleum Exhibition & Conference.

Cerquitelli, T., Ventura, F., Apiletti, D., Baralis, E., Macii, E., & Poncino, M. (2021). Enhancing manufacturing intelligence through an unsupervised data-driven methodology for cyclic industrial processes. *Expert Systems with Applications*, 115269.

Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*(2), 224-227.

Deng, H., Runger, G., Tuv, E., & Vladimir, M. (2013). A time series forest for classification and feature extraction. *Information Sciences, 239*, 142-153.

DeStefano, P. D., & Grivas, D. A. (1998). Method for estimating transition probability in bridge deterioration models. *Journal of infrastructure systems, 4*(2), 56-62.

Domitrović, J., Dragovan, H., Rukavina, T., & Dimter, S. (2018). Application of an artificial neural network in pavement management system. *Tehnički vjesnik, 25*(Supplement 2), 466-473.

Edirisinghe, R., Setunge, S., & Zhang, G. (2013). Application of gamma process for building deterioration prediction. *Journal of Performance of Constructed Facilities, 27*(6), 763-773.

Ens, A. (2012). *Development of a flexible framework for deterioration modelling in infrastructure asset management*].

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). *A density-based algorithm for discovering clusters in large spatial databases with noise.* kdd (vol. 96, p. 226-231).

Falamarzi, A., Moridpour, S., Nazem, M., & Cheraghi, S. (2018). *Development of a random forests regression model to predict track degradation index: Melbourne case study.* Australian transport research forum (p. 12).

Falamarzi, A., Moridpour, S., Nazem, M., & Hesami, R. (2019). *Integration of genetic algorithm and support vector machine to predict rail track degradation.* MATEC Web of Conferences (vol. 259, p. 02007).

Faloutsos, C., Ranganathan, M., & Manolopoulos, Y. (1994). Fast subsequence matching in time-series databases. *Acm Sigmod Record, 23*(2), 419-429.

Fathalla, E., Tanaka, Y., & Maekawa, K. (2018). Remaining fatigue life assessment of in-service road bridge decks based upon artificial neural networks. *Engineering Structures, 171*, 602-616.

Ge, Y., Guo, L., & Dou, Y. (2019). Remaining Useful Life Prediction of Machinery based on KS Distance and LSTM Neural Network. *International Journal of Performability Engineering, 15*(3).

Hämäläinen, J., Jauhiainen, S., & Kärkkäinen, T. (2017). Comparison of internal clustering validation indices for prototype-based clustering. *Algorithms, 10*(3), 105.

Harry, L. S. L., Duong, P. L. T., & Raghavan, N. (2020). *Exploration of Multi-output Gaussian Process Regression for Residual Storage Life Prediction in Lithium Ion Battery.* 2020 Prognostics and Health Management Conference (PHM-Besançon) (p. 263-269).

Hasan, M. (2015). *Deterioration prediction of concrete bridge components using artificial intelligence and stochastic methods*, RMIT University].

Helwig, N., Pignanelli, E., & Schütze, A. (2015). *Condition monitoring of a complex hydraulic system using multivariate statistics.* 2015 IEEE International Instrumentation and Measurement Technology Conference (I2MTC) Proceedings (p. 210-215).

Hubert, L., & Schultz, J. (1976). Quadratic assignment as a general data analysis strategy. *British journal of mathematical and statistical psychology, 29*(2), 190-241.

Ishwaran, H., & Lu, M. (2019). Standard errors and confidence intervals for variable importance in random forest regression, classification, and survival. *Statistics in medicine, 38*(4), 558-582.

Kaplan, A., & Haenlein, M. (2019). Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. *Business Horizons, 62*(1), 15-25.

Karch, J., & van Ravenzwaaij, D. (2020). Improving on Adjusted R-squared. *Collabra: Psychology, 6*(1).

Kleiner, Y., & Rajani, B. (2001). Comprehensive review of structural deterioration of water mains: statistical models. *Urban water, 3*(3), 131-150.

Kobayashi, K., Do, M., & Han, D. (2010). Estimation of Markovian transition probabilities for pavement deterioration forecasting. *KSCE Journal of Civil Engineering, 14*(3), 343-351.

Kozjek, D., Kralj, D., & Butala, P. (2017). Interpretative identification of the faulty conditions in a cyclic manufacturing process. *Journal of Manufacturing Systems, 43*, 214-224.

Kumar, J., Soni, V., & Agnihotri, G. (2013). Maintenance performance metrics for manufacturing industry. *International Journal of Research in Engineering and Technology, 2*(2), 136-142.

Kumar, M., Patel, N. R., & Woo, J. (2002). *Clustering seasonality patterns in the presence of errors.* Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining (p. 557-563).

Laakso, T., Kokkonen, T., Mellin, I., & Vahala, R. (2018). Sewer condition prediction and analysis of explanatory factors. *Water, 10*(9), 1239.

Lei, Y., Jiang, W., Jiang, A., Zhu, Y., Niu, H., & Zhang, S. (2019). Fault diagnosis method for hydraulic directional valves integrating PCA and XGBoost. *Processes, 7*(9), 589.

[#50 utilise un type de document non défini dans ce style].

Li, L., Gariel, M., Hansman, R. J., & Palacios, R. (2011). *Anomaly detection in onboard-recorded flight data using cluster analysis.* 2011 IEEE/AIAA 30th Digital Avionics Systems Conference (p. 4A4-1-4A4-11).

Liao, T. W. (2005). Clustering of time series data—a survey. *Pattern recognition, 38*(11), 1857-1874.

Lin, J., Keogh, E., Lonardi, S., & Chiu, B. (2003). *A symbolic representation of time series, with implications for streaming algorithms.* Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery (p. 2-11).

Lou, Z., Gunaratne, M., Lu, J., & Dietrich, B. (2001). Application of neural network model to forecast short-term pavement crack condition: Florida case study. *Journal of infrastructure systems, 7*(4), 166-171.

Marcelino, P., Lurdes Antunes, M. d., Fortunato, E., & Castilho Gomes, M. (2019). Machine learning approach for pavement performance prediction. *International Journal of Pavement Engineering*, 1-14.

Micevski, T., Kuczera, G., & Coombes, P. (2002). Markov model for storm water pipe deterioration. *Journal of infrastructure systems, 8*(2), 49-56.

Micic, T., Stojic, D., Stankovic, M., & Velimirovic, N. (2016). Gamma process model for timber-concrete composite beam deterioration prediction. *Wood Research, 61*(3), 373-385.

Mishalani, R. G., & Madanat, S. M. (2002). Computation of infrastructure transition probabilities using stochastic duration models. *Journal of Infrastructure systems, 8*(4), 139-148.

Möller-Levet, C. S., Klawonn, F., Cho, K.-H., & Wolkenhauer, O. (2003). *Fuzzy clustering of short time-series and unevenly distributed sampling points.* International symposium on intelligent data analysis (p. 330-340).

Morcous, G., Rivard, H., & Hanna, A. (2002). Modeling bridge deterioration using case-based reasoning. *Journal of Infrastructure Systems, 8*(3), 86-95.

Najafi, M., & Kulandaivel, G. (2005). Pipeline condition prediction using neural network models. Dans *Pipelines 2005: Optimizing Pipeline Design, Operations, and Maintenance in Today's Economy* (p. 767-781).

Nanduri, A., & Sherry, L. (2016). *Anomaly detection in aircraft data using Recurrent Neural Networks (RNN).* 2016 Integrated Communications Navigation and Surveillance (ICNS) (p. 5C2-1-5C2-8).

Ngo, T. (2011). Data mining: practical machine learning tools and technique, by ian h. witten, eibe frank, mark a. hell. *ACM SIGSOFT Software Engineering Notes, 36*(5), 51-52.

Pereira, J., & Silveira, M. (2018). *Unsupervised anomaly detection in energy time series data using variational recurrent autoencoders with attention.* 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA) (p. 1275-1282).

Rajani, B., & Kleiner, Y. (2001). Comprehensive review of structural deterioration of water mains: physically based models. *Urban water, 3*(3), 151-164.

Ranjith, S., Setunge, S., Gravina, R., & Venkatesan, S. (2013). Deterioration prediction of timber bridge elements using the Markov chain. *Journal of Performance of Constructed Facilities, 27*(3), 319-325.

Ratanamahatana, C. A., & Keogh, E. (2005). *Multimedia retrieval using time series representation and relevance feedback.* International Conference on Asian Digital Libraries (p. 400-405).

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics, 20*, 53-65. https://doi.org/10.1016/0377-0427(87)90125-7

Sakoe, H. (1971). *Dynamic-programming approach to continuous speech recognition.* 1971 Proc. the International Congress of Acoustics, Budapest.

Scheidegger, A., Hug, T., Rieckermann, J., & Maurer, M. (2011). Network condition simulator for benchmarking sewer deterioration models. *Water research, 45*(16), 4983-4994.

Shamim, A., Hussain, H., & Shaikh, M. U. (2010). *A framework for generation of rules from decision tree and decision table.* 2010 International Conference on Information and Emerging Technologies (p. 1-6).

Singpurewalla, N. D., & Song, M. S. (1988). Reliability analysis using Weibull lifetime data and expert opinion. *IEEE Transactions on Reliability, 37*(3), 340-347.

Sisodia, D. S., & Verma, N. (2018). *Performance Evaluation of Density-Based Clustering Methods for Categorizing Web Robot Sessions.* 2018 International Conference on Advanced Computation and Telecommunication (ICACAT) (p. 1-5).

Suschnigg, J., Mutlu, B., Fuchs, A. K., Sabol, V., Thalmann, S., & Schreck, T. (2020). *Exploration of Anomalies in Cyclic Multivariate Industrial Time Series Data for Condition Monitoring.* EDBT/ICDT Workshops (p. 1-8).

Thomas, O., & Sobanjo, J. (2013). Comparison of Markov chain and semi-Markov models for crack deterioration on flexible pavements. *Journal of Infrastructure Systems, 19*(2), 186-195.

Thomas, O., & Sobanjo, J. (2016). Semi-Markov models for the deterioration of bridge elements. *Journal of Infrastructure Systems, 22*(3), 04016010.

Tong, C., Yin, X., Li, J., Zhu, T., Lv, R., Sun, L., & Rodrigues, J. J. (2018). An innovative deep architecture for aircraft hard landing prediction based on time-series sensor data. *Applied Soft Computing, 73*, 344-349.

Tran, H. D. (2007). *Investigation of deterioration models for stormwater pipe systems*, Victoria University].

Tran, H. D., Perera, B. J. C., & Ng, A. W. M. (2010). Markov and Neural Network Models for Prediction of Structural Deterioration of Storm-Water Pipe Assets. *Journal of Infrastructure Systems, 16*(2), 167-171. https://doi.org/10.1061/(asce)is.1943-555x.0000025

Transports Québec. (2012). *Manuel d'inspection des ponceaux*.

Valarmathy, N., & Krishnaveni, S. (2020). A novel method to enhance the performance evaluation of DBSCAN clustering algorithm using different distinguished metrics. *Materials Today: Proceedings*.

Van Noortwijk, J., & Klatter, H. (2004). The use of lifetime distributions in bridge maintenance and replacement modelling. *Computers & Structures, 82*(13-14), 1091-1099.

van Noortwijk, J. M. (2009). A survey of the application of gamma processes in maintenance. *Reliability Engineering & System Safety, 94*(1), 2-21.

Vitorino, D., Coelho, S., Santos, P., Sheets, S., Jurkovac, B., & Amado, C. (2014). A random forest algorithm applied to condition-based wastewater deterioration modeling and forecasting. *Procedia Engineering, 89*, 401-410.

Wang, X., Wang, B. X., Hong, Y., & Jiang, P. H. (2021). Degradation data analysis based on gamma process with random effects. *European Journal of Operational Research, 292*(3), 1200-1208.

Wellalage, N. K. W., Zhang, T., & Dwight, R. (2015). Calibrating Markov chain–based deterioration models for predicting future conditions of railway bridge elements. *Journal of Bridge Engineering*, *20*(2), 04014060.

Wirahadikusumah, R., Abraham, D., & Iseley, T. (2001). Challenging issues in modeling deterioration of combined sewers. *Journal of infrastructure systems, 7*(2), 77-84.

Yang, C.-Y., Chen, P.-Y., Wen, T.-J., & Jan, G. E. (2019). Imu consensus exception detection with dynamic time warping—a comparative approach. *Sensors*, *19*(10), 2237.

Yang, X., Zhang, Y., Shardt, Y. A., Li, X., Cui, J., & Tong, C. (2019). A KPI-Based Soft Sensor Development Approach Incorporating Infrequent, Variable Time Delayed Measurements. *IEEE Transactions on Control Systems Technology*, *28*(6), 2523-2531.

Yin, S., Wang, G., & Yang, X. (2014). Robust PLS approach for KPI-related prediction and diagnosis against outliers and missing data. *International Journal of Systems Science, 45*(7), 1375-1382.

Zhu, Y., Zhao, T., Jiao, J., & Chen, Z. (2019). The lifetime prediction of epoxy resin adhesive based on small-sample data. *Engineering Failure Analysis*, *102*, 111-122.

ANNEXE A LISTE DES VARIABLES DES DONNÉES DE PONCEAUX

Tableau 5.1 Liste des colonnes de la base de données des ponceaux

Identification	***Signification***
IDE	Identifiant du ponceau
AnnéeFichier	Année de l'inspection
SousType	Sous-type du ponceau
Age	L'âge du ponceau
IEP	L'indice d'état de ponceau
Etat	L'état du ponceau (A, B, C, D ou E)
Desc_Grav_S1	Mouvement et déformation (Evaluation sur 5)
Desc_Grav_S2	Défauts de matériaux (Evaluation sur 5)
Desc_Grav_S3	Fissuration et assemblage (Evaluation sur 5)
Desc_Grav_H1	Sédimentation et rendement hydraulique (Evaluation sur 5)
Desc_Grav_H2	Affouillement (Evaluation sur 5)
Desc_Grav_H3	Infiltration (Evaluation sur 5)
Desc_Grav_H4	Accumulation de débris (Evaluation sur 5)
CF_NOM	Type de la route (Autoroute, Nationale, régionale, etc…)
DT	Département
RTSS	Identifiant de la chaussée
Municipalite	Municipalité gérant le ponceau
Larg_Diam	Largeur du diamètre du ponceau
Longueur	Longueur du ponceau
HauteurDeRemblaiGauche	Hauteur du remblai Gauche
HauteurDeRemblaiDroite	Hauteur du remblai droite
PresenceMurTete	Présence du mur tête
TypesChaussee	Type de chaussée (Souple, rigide, mixte, etc…)
CLASSE_FONCT	Classe fonctionnelle du ponceau
Année_Construction_Corrigée	Année de construction du ponceau
INSERTION	Présence d'une insertion à l'intérieur (Oui/Non)
DJMA	Débit journalier de véhicule

Tableau A.1. Liste des colonnes de la base de données des ponceaux (suite et fin)

Identification	***Signification***
Pourc_Camions	Pourcentage de camion traversant la route
LimiteVitesse	Vitesse limite dans la route
Temp_max_moy	Température maximale moyenne
Temp_min_moy	Température minimale moyenne
Temp_moy	Température moyenne
Temp_la_plus_haute	La plus haute température enregistrée
Temp_la_plus_basse	La plus basse température enregistrée
Pluie_tot_mm	Intensité moyenne de pluie par ans
Neige_tot_cm	Intensité moyenne de pluie par ans
Precip_tot_mm	Précipitation annuelle moyenne
Pluie_Age	Volume totale de pluie
Neige_Age	Volume totale de neige
Precip_Age	Volume totale de précipitation

ANNEXE B IMPORTANCE DES SIGNAUX DANS L'IDENTIFICATION DES CLUSTERS

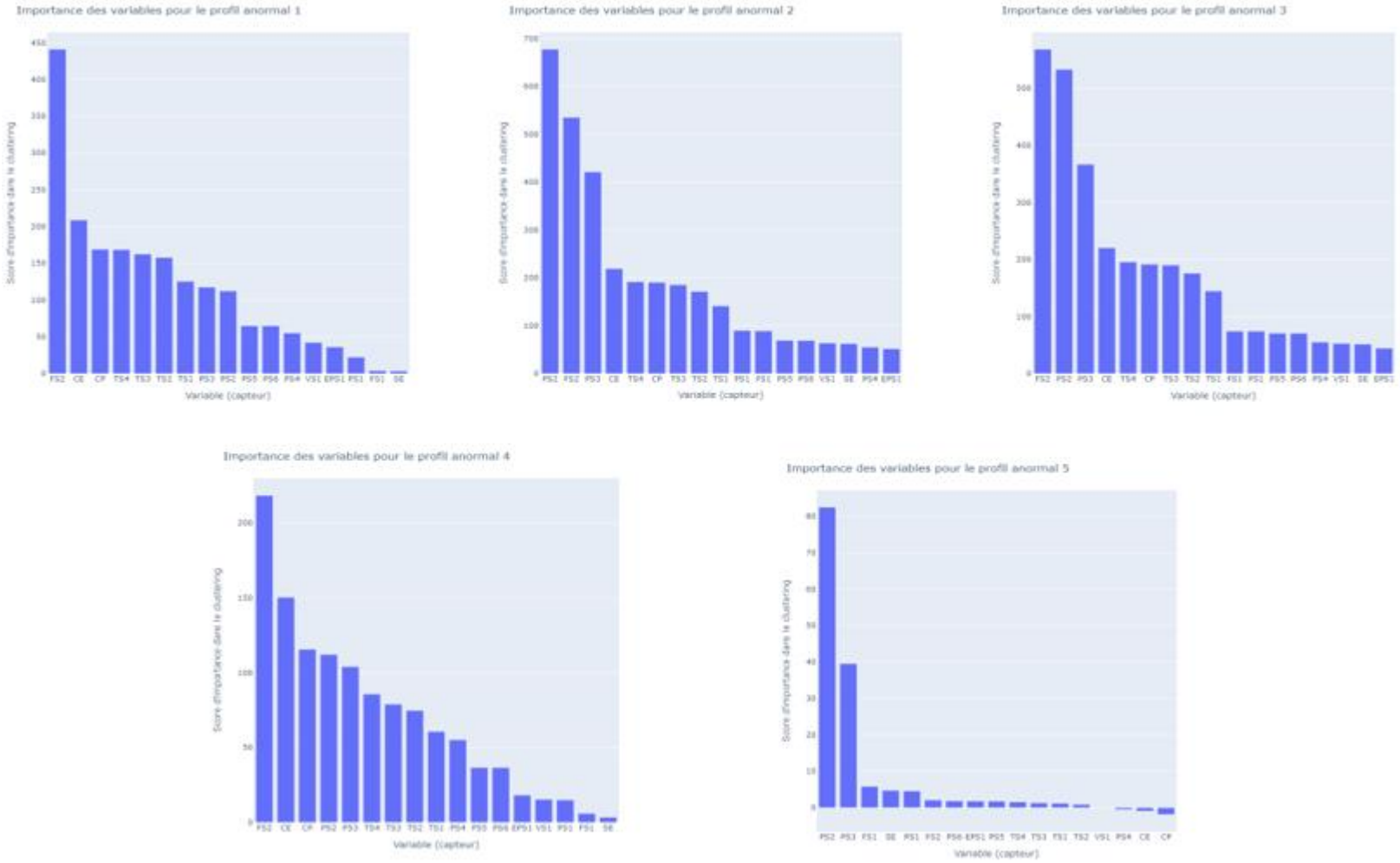


Figure B.1 Histogrammes des scores d'importance des signaux