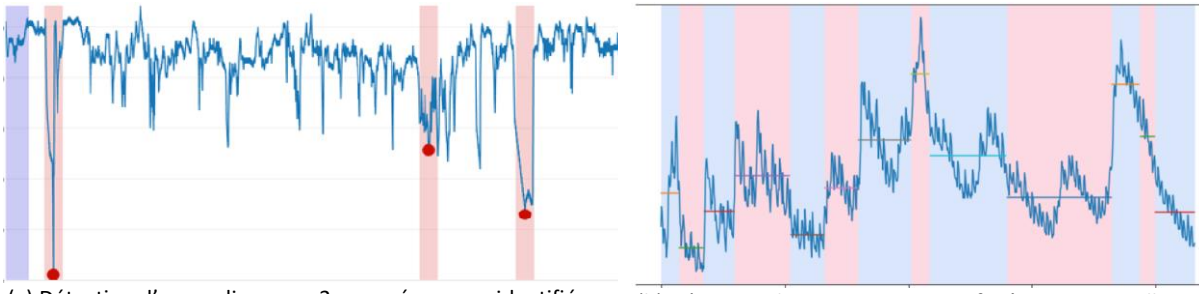
Sujet de thèse CIFRE EDF R&D (2025 – 2028)

Détection non-supervisée d'anomalies dans des flux continus de séries temporelles multivariées

Entreprise : EDF SA / EDF R&D
Entité d'accueil : département PRISME[1] / groupe P17 GAIA[2]
Lieu principal : EDF Lab Chatou, 6 quai Watier, 78401 Chatou Cedex, France
Laboratoire : Inria[3] Paris, équipe Valda[4]
Dates de début et de fin de thèse souhaitées : 01/04/2025 – 31/03/2028
Candidature (CV, lettres de motivation et de recommandation demandés) : https://www.edf.fr/edf-recrute/offre/detail/2025-127650

1. Contexte général

Avec le développement des technologies numériques (dont l'internet des objets et les architectures de données « *big data* ») et l'explosion des capacités de stockage des données informatiques, les larges collections de séries temporelles deviennent une réalité dans un grand nombre de domaines, comme la finance, les sciences de l'environnement, la médecine, les métiers du numérique, l'ingénierie ou l'industrie. Il y a donc un intérêt et un besoin croissants de développer des techniques efficaces pour traiter et analyser ce type de données.
Une série temporelle (« *time series* ») est une séquence ordonnée dans le temps de points ou de valeurs, par exemple des mesures à différents instants d'un paramètre physique issues d'un capteur de surveillance installé sur un système industriel. Une fois une série collectée, enregistrée, nettoyée (débruitage, synchronisation, complétion des données manquantes…) et mise à disposition de l'utilisateur, celui-ci souhaite généralement l'étudier pour en extraire de la connaissance et de la valeur. Cette analyse peut être simple, comme sélectionner une fenêtre temporelle pour visualiser la série et calculer des statistiques sur les valeurs afin de résumer l'information (valeur moyenne par exemple). Elle peut aussi être complexe, comme rechercher des similarités entre plusieurs séries temporelles pour réaliser des regroupements (« *segmentation and clustering* »), prévoir les prochaines valeurs de la série à partir de l'historique des mesures (« *forecasting* »), identifier des motifs récurrents (« *patterns recognition* »), ou détecter des anomalies ou des ruptures associées à des changements de régime dans les valeurs de la série, synonymes d'évolutions soudaines et inhabituelles possiblement non souhaitées (« *anomaly and change point detection* »).



(a) Détection d'anomalies, avec 3 sous-séquences identifiées comme en écart par rapport au comportement du reste de la série

(b) Détection de ruptures, avec 13 fenêtres temporelles identifiées comme présentant des régimes de comportement différents

Figure 1 - Illustration des tâches de détection d'anomalies (a) et de ruptures (b) dans des séries temporelles (pris de [1])

Série temporelle univariée ou multivariée

Une **série temporelle univariée** est une séquence ordonnée de valeurs réelles sur une seule dimension. Par exemple, une série temporelle univariée peut correspondre à l'historique des valeurs successives mesurées par un capteur. Dans ce cas, une sous-séquence (c'est-à-dire un extrait de points consécutifs de cette série) peut être représentée comme un vecteur de valeurs.

[1] Performance, Risque Industriel et Surveillance pour la Maintenance et l'Exploitation
[2] Gestion d'Actifs, Incertitudes et Apprentissage statistique
[3] Institut national de recherche en informatique et en automatique
[4] Valeur à partir des données

Une **série temporelle multivariée** est, soit un ensemble de séquences ordonnées de valeurs réelles (chaque séquence ordonnée ayant la même longueur), soit une séquence ordonnée de vecteurs composés de valeurs réelles. Un exemple de série temporelle multivariée peut être un ensemble de mesures provenant de plusieurs capteurs installés sur un même système ou sur différents équipements. Dans ce cas précis, une sous-séquence est une matrice dans laquelle chaque ligne correspond à une sous-séquence d'une seule dimension.

Série temporelle statique ou en continu

Les **séries de données statiques** sont des séquences de valeurs ayant une longueur fixe. Dans ce cas, on ne s'attend pas à ce que d'autres valeurs soient ajoutées et on peut analyser des points ou des sous-séquences en une seule fois. Par exemple, l'analyse rétrospective d'une fenêtre temporelle donnée consistera à étudier une collection de séries de valeurs statiques. Dans ce cas, on parle souvent d'analyse hors-ligne (« *offline* »).
À l'inverse, les **séries de données en continu (« *streaming* »)** sont des séquences de longueur infinie, avec de nouveaux points ou sous-séquences arrivant à un taux d'acquisition donné (pas nécessairement constant dans le temps). Dans ce cas, un modèle d'analyse d'une série en continu doit pouvoir être mis à jour dynamiquement au fur et à mesure de l'arrivée de nouveaux points (analyse en ligne – « *online* »). À titre d'illustration, la surveillance en ligne de l'état de santé de matériels et la détection de sous-séquences anormales en temps réel nécessitent (idéalement) des outils d'analyse de flux continus de séries temporelles.

Série temporelle de données discrètes ou continues

Une **série temporelle de données discrètes** est une séquence de valeurs successives d'un paramètre catégoriel (c'est-à-dire ayant un nombre limité de valeurs ou de catégories/modalités distinctes), nominales ou ordinales (ordonnées).
Au contraire, **les séries de données continues** sont des séquences de points avec des valeurs réelles. Par exemple, les capteurs booléens de type Tout ou Rien (ToR), qui ne renvoient que 0 ou 1 comme valeurs possibles, génèrent des séries de données discrètes, tandis que les capteurs de température renvoient usuellement des séries de données continues.

Points manquants et séries temporelles non synchronisées

Les contraintes induites par l'étape de collecte des données peuvent rendre les séries temporelles plus difficiles à analyser. La première contrainte est liée aux **points manquants**. Cette contrainte peut être due à des problèmes de capteurs renvoyant des valeurs erronées, à un protocole d'acquisition spécifique (par exemple, certains capteurs ne renvoient une valeur que lorsque la valeur mesurée change), ou tout simplement à une panne du capteur. Il en résulte des séries avec des valeurs manquantes qui doivent être complétées.
La deuxième contrainte est liée aux séries de données multivariées **non synchronisées**. Elle est due à la différence de taux d'acquisition des différents capteurs. Dans ce cas, il faut choisir un taux d'acquisition fixe et ensuite, soit sous-échantillonner (c'est-à-dire perdre des points et une précision potentielle), soit sur-échantillonner (c'est-à-dire créer une contrainte de points manquants) les séries de données avec un taux d'acquisition différent. Ces deux contraintes sont critiques et typiques de nombreux cas d'application.

Détection d'anomalies

Il n'existe pas de définition unique, universelle et précise caractérisant une **anomalie** (parfois appelée aussi valeur aberrante – « *outlier* »). En général, une anomalie est une observation qui semble s'écarter de façon notable des autres membres de l'échantillon dans lequel elle se produit. Cet écart peut indiquer que l'observation spécifique a été générée par un mécanisme différent de celui du reste des données. Ce mécanisme peut être une procédure erronée de mesure et de collecte de donnée ou une variabilité inhérente au domaine des données examinées. Néanmoins, de telles observations sont intéressantes dans les deux cas, et l'analyste gagnerait à les connaître. En pratique, la définition générale ci-avant peut prendre différentes formes, en fonction du problème spécifique et des caractéristiques des données manipulées. Par exemple, lorsqu'on étudie par des techniques statistiques un échantillon de valeurs répétées d'un même paramètre, les anomalies peuvent être les points qui s'écartent de la moyenne de la distribution des données d'un certain nombre de fois l'écart-type.
Dans notre contexte, on s'intéresse à la recherche d'anomalies dans les séries temporelles. Cet objectif peut être atteint en examinant, soit des valeurs prises séparément, soit une séquence de points successifs (c'est-à-dire une sous-séquence).

- Dans le cas spécifique des points, on recherche ceux éloignés de la distribution des valeurs représentant le comportement « normal » central.
- Dans le cas spécifique des séquences de points, on s'intéresse à l'identification de sous-séquences anormales qui, contrairement aux points aberrants, ne sont pas une valeur anormale unique, mais une évolution anormale de ces valeurs.

Dans certains cas, cette distinction entre point et sous-séquence devient cruciale pour la raison suivante : même si chaque point pris individuellement semble normal, la forme (ou le motif – « *pattern* ») généré par la séquence de ces mêmes valeurs peut être anormal et peut conduire à des dysfonctionnements qui auraient été détectés trop tard si on avait étudié séparément les valeurs de chaque point. La Figure 2 illustre cette distinction entre points et sous-séquences anormaux.

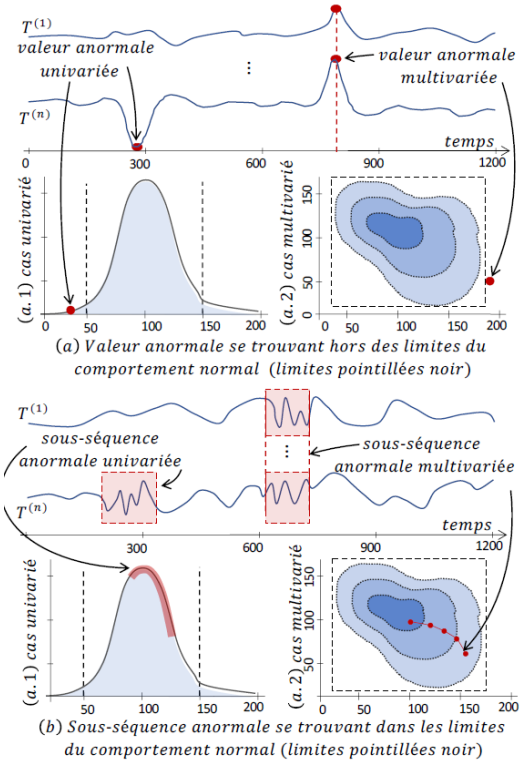


Figure 2 - Exemples illustratifs avec (a) un point aberrant (pour une série temporelle (a.1) univariée et (a.2) multivariée) - (b) une sous-séquence anormale composée de valeurs normales prises individuellement (pour une série temporelle (b.1) univariée et (b.2) multivariée) (repris de [2])

Méthode d'apprentissage non-supervisée, semi-supervisée ou (entièrement) supervisée

Les méthodes de détection d'anomalies **non-supervisées** ne requièrent en entrée que les sous-séquences de points et n'ont pas besoin, comme informations préalables, d'annotations (aussi appelées labels, étiquettes ou exemples – « *tags* ») qualifiant précisément ce qu'est une anomalie et ce qu'est un comportement normal (ou sain). Ces approches « en aveugle » (ou « agnostiques ») conviennent bien à la découverte d'anomalies ou de nouveautés (« *novelty* »), à la visualisation et à l'annotation automatique. Néanmoins, elles sont, en général, moins précises que les deux autres familles de méthodes.

Les méthodes **semi-supervisées** nécessitent des annotations de sous-séquences normales pour apprendre à détecter des sous-séquences anormales (en écart par rapport aux sous-séquences saines). Ce cas est très classique dans la littérature scientifique. Cette catégorie d'approches est souvent définie comme non-supervisée ; cependant, il paraît « injuste » de regrouper ces deux familles, sachant que les approches semi-supervisées nécessitent beaucoup plus de connaissances préalables que les non-supervisées.

Enfin, les méthodes **supervisées** ont besoin d'annotations complètes de toutes les sous-séquences, aussi bien normales et anormales, pour apprendre à les distinguer ensemble. Ces approches peuvent s'avérer très efficaces, mais le travail d'annotation préalable, qui requiert généralement l'intervention de spécialistes du domaine étudié, est chronophage et coûteux, voire impossible dans certains cas (anomalie non circonscrite avec précision).

Évaluation des performances d'un modèle

L'évaluation des **performances d'un modèle** vise à déterminer dans quelle mesure le modèle construit s'adapte aux données qui ont servi à son entraînement et comment il se généralise à des données qui n'ont pas fait partie de l'échantillon d'apprentissage. Elle est cruciale à différents titres. En effet, elle permet à l'utilisateur de choisir le meilleur modèle parmi plusieurs candidats (« *model selection* »). Les mesures d'évaluation peuvent également servir à ajuster les hyperparamètres d'un modèle afin d'en améliorer les performances (« *hyperparameter optimization* »). Enfin, elle peut contribuer à détecter des problèmes potentiels dans le modèle, tels que le sur-apprentissage (« *over-fitting* »). Les mesures d'évaluation sont des métriques quantitatives utilisées pour évaluer les performances des modèles. Le choix des mesures appropriées dépend du type de problème à résoudre. Pour les problèmes de classification, la précision (proportion de tous les cas correctement classés par le modèle, qu'ils soient associés ou non à des anomalies), le rappel (proportion de toutes les anomalies détectées par le modèle qui en sont réellement) et le F-score (moyenne harmonique de la précision et du rappel) sont des mesures classiques : ils peuvent être utilisés dans le cadre de la détection d'anomalies dans des séries temporelles, mais requièrent au préalable de définir une valeur seuil du score d'anomalie produit par le modèle au-delà de laquelle le point (ou la sous-séquence) sera qualifié(e) d'anormal(e) par l'algorithme. L'AUC-ROC[5] ou l'AUC-PRC[6] contournent cette limite, mais ils ne sont adaptés qu'à la détection de points aberrants (et pas de sous-séquences anormales).

Dérive conceptuelle

En apprentissage automatique (« *machine learning* »), la **dérive conceptuelle** (« *concept drift* ») survient quand les données auxquelles on applique un modèle, par exemple à des fins de détection d'anomalies, diffèrent significativement des données qui ont servi à entraîner le modèle, ce dernier devenant « mécaniquement » moins performant. En pratique, ce phénomène est assez courant, typiquement lorsque les données manipulées dépendent du temps ou de facteurs contextuels qui modifient la distribution statistique sous-jacente aux données. Il peut se produire de façon régulière (saisonnalité), soudaine (survenue d'un événement structurant, comme un changement de régime d'exploitation du matériel industriel) ou progressive (dérive lente du capteur de mesure). Idéalement, un modèle doit être le moins sensible possible à une dérive conceptuelle ; à défaut, il faut disposer d'outils permettant de la détecter efficacement pour alerter l'utilisateur, qui pourra alors reconstruire un jeu de données avec les nouvelles données en tenant du concept drift, ou attendre d'avoir suffisamment de données pour entraîner à nouveau le modèle.

2. Enjeux et objectifs industriels de la thèse

Dans le contexte du suivi en continu (« *e-monitoring* ») des matériels des installations de production d'électricité d'EDF, la détection d'anomalies en temps réel dans les séries temporelles issues des capteurs de surveillance représente un enjeu crucial : plus elle est précoce et efficace, plus on est en mesure de réagir tôt et à bon escient pour tenter d'atténuer les impacts, voire d'éviter la survenue, de tout événement potentiellement critique, comme un dysfonctionnement ou une défaillance d'un équipement.

Disposer de méthodes performantes de détection non-supervisée de sous-séquences anormales en streaming, adaptées à des flux continus de séries temporelles multivariées (l'anomalie pouvant être caractérisée par l'évolution simultanée de plusieurs paramètres physiques, ou observée uniquement au travers des mesures conjointes de différents capteurs), revêt donc un intérêt de tout premier ordre pour aider à la décision en appui à l'exploitation et à la maintenance des matériels.

Plusieurs cas d'usage EDF illustrent ces enjeux industriels, comme la dilatation contrariée des barres des rotors des groupes turbo-alternateurs des centrales nucléaires, les crises vibratoires des groupes moto-pompes primaires nucléaires, les vibrations des turbo-pompes alimentaires nucléaires ou les hausses des températures métal des pivoteries des groupes de production hydroélectriques. Ces cas serviront à vérifier l'applicabilité des méthodes sur des données réelles et à en évaluer leurs performances concrètes.

La thèse visera à produire des algorithmes génériques, performants et testés avec succès sur des jeux de données simulées, issues de la littérature et réelles provenant de cas d'usage EDF. Elle mènera à la production

[5] Area Under the Receiver Operating Characteristics curve
[6] Area Under the Precision-Recall curve

d'articles scientifiques (communications en conférences, articles de journaux) et dépôts de brevets, si pertinent. Les méthodes développées feront l'objet d'implémentations informatiques (librairies Python / R / Julia) pour faciliter leur utilisation en interne EDF R&D et leur transfert à l'ingénierie d'EDF.

3. Verrous et objectifs scientifiques de la thèse

Les verrous scientifiques sont multiples.

Verrou #1 : gestion de l'hétérogénéité entre les dimensions d'une même série temporelle multivariée (longueurs ou fréquences d'échantillonnage différentes, séries de données discrètes vs. continues, présence ou absence de corrélations entre plusieurs dimensions…)
Verrou #2 : développement de mesures de similarité (ou de proximité) adaptées à des sous-séquences temporelles multivariées
Verrou #3 : définition de mesures de performance adaptées à la détection de sous-séquences anormales (et pas uniquement de points aberrants)
Verrou #4 : construction d'un recueil de jeux de données appropriés pour que les résultats des études comparatives (« *benchmarks* ») aient du sens (éviter par exemple les cas d'anomalies triviales, de densité d'anomalie trop élevée ou de biais de position lorsque les anomalies sont regroupées vers la fin de la série temporelle – « *run-to-failure bias* »)
Verrou #5 : choix de la famille d'algorithmes la plus adaptée pour répondre au problème et **calibration** optimale du paramétrage (ajustement de la taille des *batchs* de données, de la longueur de la sous-séquence anormale…), en visant un compromis entre « précision » / « adaptativité » / « robustesse à une dérive conceptuelle » / « temps d'exécution » / « taille mémoire » imposé par le cadre non-supervisé et en flux continu de données. La dimension « interprétabilité » du modèle est également importante [2].

Ces verrous ne sont pas indépendants et ne sont pas forcément tous accessibles simultanément. Ainsi, la thèse s'appuiera sur une étude approfondie de la littérature et tentera de lever un maximum de verrous, tout en minimisant l'impact de ceux non abordés ou non résolus.

4. Profil souhaité

- Master 2, ou niveau équivalent, avec une dominante en Mathématiques Appliquées, Algorithmique, Machine Learning ou Data Science
- Goût prononcé pour la recherche, le développement informatique et l'algorithmique
- De bonnes qualités de communication et de rédaction sont indispensables
- Maîtrise de la langue anglaise (écrite et parlée).

5. Références bibliographiques

[1] L. Oudre, course « Machine learning for time series », retrieved 18/12/2024 from http://www.laurentoudre.fr/ast.html, 2024
[2] P. Boniol, « Detection of anomalies and identification of their precursors in large data series collections », PhD thesis in Computer Science, Université de Paris - Laboratoire d'Informatique Paris Descartes et EDF R&D, 2021