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
Encadrement industriel : Antoine AJENJO (chercheur, PRISME), Emmanuel REMY (chercheur expert, PRISME), Pierre STEPHAN (chercheur expert, PRISME)
Direction académique : Paul Boniol (Chercheur Inria), Pierre SENELLART (professeur des Universités)
Laboratoire : Inria[3] Paris, équipe Valda[4]
Dates de début et de fin de thèse souhaitées : 01/04/2025 – 31/03/2028
Ecole doctorale : Sciences Mathématiques de Paris-Centre (386)
Mention : -

1. Environnement de recherche industrielle au sein du département PRISME d'EDF R&D

Au sein d'EDF R&D, le département PRISME[1] a pour missions principales de conserver et d'accroître la compétitivité et les performances des centres de production d'énergie d'aujourd'hui et de demain, dans le respect de l'environnement, de la sécurité et de la réglementation. Pour réaliser ces missions, PRISME dispose d'une expertise technique sur l'ensemble des parcs de production (nucléaire, hydraulique, nouvelles énergies renouvelables, thermique). PRISME travaille principalement autour de la donnée (« *data* ») du producteur, de son acquisition à son traitement, en passant par son stockage et sa mise à disposition.
La partie « acquisition » se décline dans le développement de chaînes de mesures. Ceci conduit à l'amélioration des techniques de mesures sur les process et d'auscultation de composants et d'ouvrages. La partie « stockage et mise à disposition » se traduit au département PRISME par les compétences en systèmes d'information industriels, « *small data* », « *big data* », objets connectés et également télécommunications. La partie « traitement » se décline selon plusieurs axes :

- exploitation de données au moyen de modèles de comportement pour le fonctionnement de centrales, le département disposant des compétences pour la génération de ces modèles,
- exploitation de données au moyen de modèles de comportement pour des procédés de fabrication (soudage, forgeage, fabrication additive),
- traitement de données pour l'estimation de grandeurs non directement observables, telles que des défauts dans les matériels, soit en « *data-driven* », soit en association avec des modèles physiques nécessitant éventuellement du calcul haute performance (ou calculs intensifs – « *high performance computing* »),
- traitement de données pour l'évaluation d'actifs et la définition de modèles pour prise de décisions,
- définition de systèmes de surveillance s'appuyant sur les données observées pour identifier des comportements anormaux, et contribuer ainsi à une maintenance optimisée,
- caractérisation des incertitudes, à la fois sur les données et sur les modèles contribuant à leur exploitation, et réduction de celles-ci.

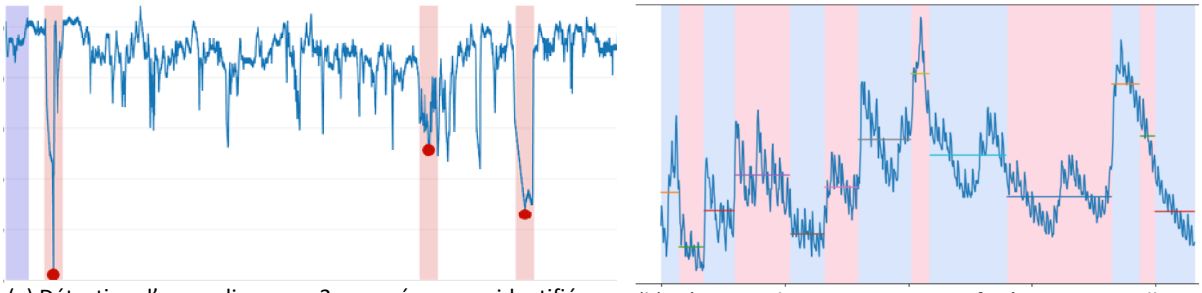
Enfin, le traitement de ces données conduit à la mise en œuvre de solutions variées pour les différents producteurs exploitants : nouveaux capteurs, mise à disposition de données agrégées, nouveaux outils pour améliorer la sûreté, la performance des moyens de production…

2. 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.

[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 (« *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.

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.

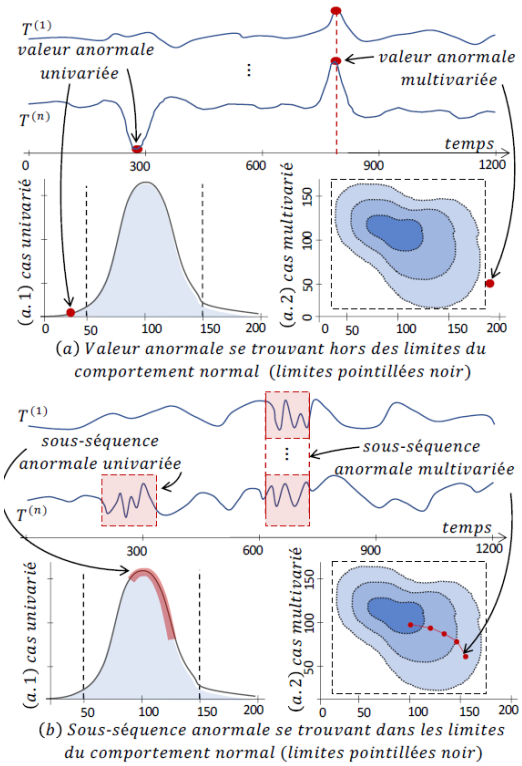


Figure 2 - Exemples illustratifs avec (a) un point aberrant (pour une série temporelle (a.1) univariée et (a.2) multivariée) - (b) une sous-séquence anormale composée de valeurs normales prises individuellement (pour une série temporelle (b.1) univariée et (b.2) multivariée) (repris de [2])

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.

3. 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 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.

4. 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…)

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

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[7] (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.

5. État de l'art et pistes de recherche envisagées

L'analyse des séries temporelles fait l'objet de travaux de recherche intensifs, le nombre annuel de publications sur le sujet référencées sur la base Scopus ayant été multiplié par plus de 100 entre 1985 et 2020.
Heureusement, la publication régulière de solides revues bibliographiques proposant des taxonomies des algorithmes [11], [12], [13], [14], [15], [16] facilite l'appropriation des (nombreux) travaux en cours sur le sujet de la détection d'anomalies dans les séries temporelles.
De plus, la constitution récente par la communauté scientifique de larges recueils de jeux données avec des anomalies annotées, et le partage des codes informatiques implémentant les algorithmes développés pour la détection d'anomalies, a rendu possible la réalisation d'études permettant d'évaluer et de comparer les performances d'un important panel de méthodes disponibles dans la littérature [21], [22], [23], [24], [25], [26], [27], [28], [29], [30], [31]. Malgré les limites de ce type d'exercices de comparaison et les réticences de certains auteurs [17], [18], [19], et bien qu'il existe aujourd'hui peu de mesures de performance adaptées à la détection de sous-séquences anormales [20], **les résultats des principaux benchmarks réalisés montrent qu'aucun algorithme ne se démarque et n'est systématiquement plus efficace que tous les autres**, les performances de chaque méthode pouvant varier sensiblement selon la nature et les caractéristiques des séries temporelles manipulées. De plus, la plupart des algorithmes actuels traitant des flux de données continus considèrent les séries temporelles comme des successions de valeurs (et pas comme des sous-séquences) et/ou ne traitent que des séries temporelles univariées (et pas multivariées).
Le travail de début de thèse consistera donc à **mener une revue de littérature** visant à identifier, parmi toute la production scientifique, les développements récents les plus prometteurs, pour ensuite les **tester et évaluer leurs performances** sur des jeux de données de référence. Ce travail pourra s'articuler selon les phases suivantes :

- recherche de jeux de données annotées (simulées ou réelles) mises à disposition de la communauté, voire utilisation d'algorithmes d'intelligence artificielle générative [32], [33], [34], [35], [36], afin de constituer un recueil qui servira à alimenter les études comparatives - les jeux de données provenant des cas d'usage EDF viendront compléter ce recueil
- revue de littérature des méthodes récentes de détection non-supervisée d'anomalies, idéalement dans des flux continus de séries temporelles multivariées [37], [38], [39], [40], [41], [42], [43], [44], [45], [46], [47], et construction d'une taxonomie pour mieux appréhender les points communs et les différences de chaque algorithme
- récupération des codes existants et/ou implémentation informatique des méthodes identifiées comme à fort potentiel suite à l'état de l'art réalisé à l'étape précédente, puis mise en œuvre et évaluation de leurs performances à partir des données annotées recensées lors de la première phase.

[7] Voire sélection automatique d'algorithmes [3], [4], [5], [6], [7], [8], [9], [10]

Ce premier travail devrait permettre d'opérer un filtre parmi les algorithmes et les plus encourageants serviront de base pour les développements à suivre.

6. Organisation, moyens et programme de travail

Organisation et équipe encadrante

Compte tenu de la pluralité des compétences nécessaires pour ce sujet de thèse, à la croisée des chemins entre les communautés « algorithmes d'apprentissage automatique », « gestion de bases de données » et « informatique », le **directeur de la thèse sera Pierre SENELLART**, professeur des Universités en informatique à l'ENS[8], faisant partie de l'Université PSL[9], au sein du DIENS[10], qui est un laboratoire commun entre le CNRS[11], Inria[3] Paris et l'ENS. Ses intérêts de recherche portent sur les aspects pratiques et théoriques de la gestion de données du web, l'extraction d'informations, la gestion de l'incertitude, la fouille du web et la gestion de données intensionnelles. Il est responsable de l'équipe-projet commune CNRS, ENS, Inria Valda[4], qui se concentre sur les aspects théoriques et systèmes de la gestion de données complexes, en particulier les données produites par l'activité humaine.

Pierre SENELLART sera secondé par **Paul BONIOL**, membre de l'équipe-projet Valda, spécialisé en systèmes d'analyse et de gestion de séries temporelles massives, en méthodes de détection d'anomalies non-supervisées et supervisées pour les grandes séries temporelles et en apprentissage automatique pour l'analyse des séries temporelles. Paul BONIOL connaît bien le contexte industriel d'EDF, puisqu'il a réalisé sa thèse de doctorat « Détection d'anomalies et identification de leurs précurseurs dans les grandes séries de données » en contrat CIFRE entre l'Université Paris Cité et EDF R&D PRISME.

Ainsi, la complémentarité des compétences apportées par les deux superviseurs académiques sera un atout évident pour guider le doctorant dans ses recherches bibliographiques et ses développements théoriques et algorithmiques.

Enfin, s'il est prévu que le doctorant passe une part significative de son temps sur le site EDF R&D Lab Chatou, il est aussi prévu qu'il travaille en étroite collaboration, régulièrement, avec ses responsables académiques.

Planning prévisionnel

L'idée n'est pas ici de proposer un plan de thèse rigide, mais plutôt de proposer une trame prévisionnelle typique de thèse afin de permettre à tout étudiant(e) de se projeter dans le cheminement des trois années.

T1 2025 : validation du (de la) candidat(e)

Année #1 (avril 2025) : début de thèse

- **Tâche 1.1 :** étude bibliographique approfondie et première formalisation rigoureuse du problème de détection non-supervisée d'anomalies dans des flux continus de séries temporelles multivariées
- **Tâche 1.2 :** prise en main des algorithmes existants dans la littérature, construction d'un recueil de jeux de données simulées et issus de la littérature appropriés à des études comparatives et réalisation d'un benchmark

✎ **Livrable 1 :** communication en conférence internationale ou article de journal

Année #2 (avril 2026) : début de 2ème année

- **Tâche 2.1 :** développements méthodologiques sur la détection non-supervisée d'anomalies dans des flux continus de séries temporelles multivariées, en adaptant des algorithmes existants ou en proposant ex-nihilo une méthode spécifique
- **Tâche 2.2 :** application à différents cas d'usage sur données simulées, issues de la littérature et réelles provenant de cas d'usage EDF

✎ **Livrable 2 :** communication en conférence internationale ou article de journal

Année #3 (avril 2027) : début de 3ème année

- **Tâche 3.1 :** poursuite et finalisation des développements méthodologiques
- **Tâche 3.2 :** application aux différents cas d'usage et comparaison des performances de l'approche développée avec celles des méthodes concurrentes

✎ **Livrable 3.1 :** manuscrit de thèse

[8] École Normale Supérieure
[9] Université Paris Sciences & Lettres
[10] Département d'Informatique de l'ENS
[11] Centre National de la Recherche Scientifique

- **Livrable 3.2 :** empaquetage des codes informatiques
- **Livrable 3.3 :** communication en conférence internationale ou article de journal

Soutenance (avant mars 2028).

Cette vision par liste contient, en réalité, des tâches de long terme qui s'échelonneront tout au long de la thèse (par exemple, la prise en main des cas d'usage et leur traitement vis-à-vis des propositions méthodologiques formulées par le doctorant). Même si ce planning prévisionnel est susceptible de varier (difficultés méthodologiques, appétence de l'étudiant pour certains aspects théoriques ou numériques, aléas divers…), il est important de rappeler ici que l'étudiant sera encadré dans un souci conjoint de résolution d'objectifs scientifiques précis liés au sujet, mais aussi de formation et d'épanouissement global dans la recherche au sein d'EDF R&D et au sein de l'équipe-projet Valda. À ce titre**, il est prévu que le doctorant passe 75% de son temps à EDF R&D et 25% de son temps dans le laboratoire** associé à son équipe de direction de thèse.

Valorisation et lien avec la communauté scientifique

Ce sujet de thèse se situe au croisement entre plusieurs communautés scientifiques, à savoir « algorithmes d'apprentissage automatique », « gestion de bases de données » et « informatique ». Ainsi, il est attendu que l'étudiant(e), par ses capacités et la qualité de sa recherche, puisse s'insérer dans l'une ou plusieurs de ces communautés par le biais de ses travaux. La communication écrite et orale de ses travaux fait complètement partie intégrante de la thèse.

7. Références bibliographiques

Généralités sur les séries temporelles et la détection d'anomalies

[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

Sélection automatique de modèles

[3] E.J. Keogh, S. Lonardi, and C.A. Ratanamahatana, « Towards parameter-free data mining », Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2004

[4] M.Q. Ma, Y. Zhao, X. Zhang, and L. Akoglu, « The need for unsupervised outlier model selection: a review and evaluation of internal evaluation strategies », ACM SIGKDD Explorations Newsletter, vol. 25(1), pp. 19–35, 2023

[5] Y. Zhao, R.A. Rossi, and L. Akoglu, « Automatic unsupervised outlier model selection », NIPS'21: Proceedings of the 35th International Conference on Neural Information Processing Systems, article n°343, pp. 4489–4502, 2024

[6] L. Woltmann, P. Damme, C. Hartmann, D. Habich, and W. Lehner « Learned selection strategy for lightweight integer compression algorithms », 26th International Conference on Extending Database Technology (EDBT 2023), vol. 26, pp. 552–564, 2023

[7] S. Schmidl, P. Wenig, and T. Papenbrock, « HYPEX: Hyperparameter optimization in time series anomaly detection », proceedings of the Gesellschaft fur Informatik (GI) Conference, 2023

[8] E. Sylligardos, P. Boniol, J. Paparrizos, P. Trahanias, and T. Palpanas, « Choose wisely: an extensive evaluation of model selection for anomaly detection in time series », Proceedings of the VLDB Endowment, vol. 16(11), pp. 3418–3432, 2023

[9] M. Goswami, C.I. Challu, L. Callot, L. Minorics, and A. Kan, « Unsupervised model selection for time-series anomaly detection », Proceedings of the International Conference on Learning Representations (ICLR), 2023

[10] A. Alsharef, K. Aggarwal, Sonia, M. Kumar, and A. Mishra, « Review of ML and autoML solutions to forecast time-series data », Archives of Computational Methods in Engineering, vol. 29, pp. 5297–5311, 2022

Revues bibliographiques de méthodes de détection d'anomalies dans les séries temporelles

[11] S. Schmidl, P. Wenig, and T. Papenbrock, « Anomaly detection in time series: a comprehensive evaluation », Proceedings of the VLDB Endowment, vol. 15(9), pp. 1779–1797, 2022

[12] A. Blázquez-García, A. Conde, U. Mori, and J.A. Lozano, « A review on outlier/anomaly detection in time series data », ACM Computing Surveys, vol. 54(3), pp. 1–33, 2021

[13] Z.Z. Darban, G.I. Webb, S. Pan, C.C. Aggarwal, and M. Salehi, « Deep learning for time series anomaly detection: a survey », ACM Computing Surveys, vol. 57(1), pp. 1–42, 2024

[14] T. Lu, L. Wang, and X. Zhao, « Review of anomaly detection algorithms for data streams », Applied Sciences, vol. 13(10), 2023
[15] L. Correia, J.-C. Goos, P. Klein, T. Bäck, and A.V. Kononova, « Online model-based anomaly detection in multivariate time series: taxonomy, survey, research challenges and future directions », Engineering Applications of Artificial Intelligence, vol. 138, 109323, 2024
[16] N. Mejri, L. Lopez-Fuentes, K. Roy, P. Chernakov, E. Ghorbel, and D. Aouada, « Unsupervised anomaly detection in time-series: an extensive evaluation and analysis of state-of-the-art methods », Expert Systems with Applications, vol. 256, 124922, 2024

Jeux de données, études comparatives et mesures de performance de différents algorithmes

[17] R. Wu, and E.J. Keogh, « Current time series anomaly detection benchmarks are flawed and are creating the illusion of progress », IEEE 38th International Conference on Data Engineering (ICDE), pp. 1479–1480, 2022
[18] V.M.A. Souza, D.M. dos Reis, A.G. Maletzke, and G.E.A.P.A. Batista, « Challenges in benchmarking stream learning algorithms with real-world data », Data Mining and Knowledge Discovery, vol. 34, pp. 1805–1858, 2020
[19] P. Wenig, S. Schmidl, and T. Papenbrock, « Anomaly detectors for multivariate time series: the proof of the pudding is in the eating IEEE 40th International Conference on Data Engineering Workshops (ICDEW), pp. 96–101, 2024
[20] J. Paparrizos, P. Boniol, T. Palpanas, R.S. Tsay, A. Elmore, and M.J. Franklin, « Volume under the surface: a new accuracy evaluation measure for time-series anomaly detection », Proceedings of the VLDB Endowment, vol. 15(11), pp. 2774–2787, 2022
[21] S. Ahmad, A. Lavin, S. Purdy, and Z. Agha, « Unsupervised real-time anomaly detection for streaming data », Neurocomputing, vol. 262, pp. 134–147, 2017 - Numenta Anomaly Benchmark (NAB) - https://github.com/numenta/NAB
[22] V. Jacob, F. Song, A. Stiegler, B. Rad, Y. Diao, and N. Tatbul, « Exathlon: a benchmark for explainable anomaly detection over time series », Proceedings of the VLDB Endowment, vol. 14(11), pp. 2613–2626, 2021 - https://github.com/exathlonbenchmark/exathlon
[23] J. Paparrizos, Y. Kang, P. Boniol, R.S. Tsay, T. Palpanas, and M.J. Franklin, « TSB-UAD: An end-to-end benchmark suite for univariate time-series anomaly detection », Proceedings of the VLDB Endowment, vol. 15(8), 2022 - https://github.com/TheDatumOrg/TSB-UAD
[24] F.I. Vázquez, A. Hartl, T. Zseby, and A. Zimek, « Anomaly detection in streaming data: a comparison and evaluation study », Expert Systems With Applications, vol. 233, 2023
[25] L. Correia, J.-C. Goos, T. Bäck, and A.V. Kononova, « Discrete-sequence dataset for evaluating online unsupervised anomaly detection approaches for multivariate time series », retrieved 18/12/2024 from https://arxiv.org/html/2411.13951v2, 2024
[26] A. Duraja, and P.S. Szczepaniak, « Outlier detection in data streams - A comparative study of selected methods », Procedia Computer Science vol. 192, pp. 2769–2778, 2021
[27] A. Ntroumpogiannis, M. Giannoulis, N. Myrtakis, V. Christophides, E. Simon, and I. Tsamardinos, « A meta-level analysis of online anomaly detectors », VLDB Journal vol. 32, pp. 845–886, 2023
[28] Y. Cao, Y. Ma, Y. Zhu, and K.M. Ting, « Revisiting streaming anomaly detection: benchmark and evaluation », Artificial Intelligence Review, vol. 58(8), 2025
[29] D. Wagner, T. Michels, F.C.F. Schulz, A. Nair, M. Rudolph, and M. Kloft, « TimeSeAD: benchmarking deep multivariate time-series anomaly detection », Transactions on Machine Learning Research, 2023
[30] A. Zhang, S. Deng, D. Cui, Y. Yuan, and G. Wang, « An experimental evaluation of anomaly detection in time series », Proceedings of the VLDB Endowment, vol. 17(3), pp. 483–496, 2023
[31] K.-H. Lai, D. Zha, Y. Zhao, G. Wang, J. Xu, and X. Hu, « Revisiting time series outlier detection: definitions and benchmarks », 35th Conference on Neural Information Processing Systems NeurIPS 2021, 2021

Génération de jeux de données

[32] Z. Lin, A. Jain, C. Wang, G. Fanti, and V. Sekar, « Using GANs for sharing networked time series data: challenges, initial promise, and open questions », IMC'20: Proceedings of the ACM Internet Measurement Conference, pp. 464–483, 2020 – DoppelGANger (https://github.com/fjxmlzn/DoppelGANger)
[33] N. Gruver, M. Finzi, S. Qiu, and A.G. Wilson, « Large language models are zero-shot time series forecasters », retrieved 18/12/2024 from https://arxiv.org/pdf/2310.07820.pdf, 2024 – LLMTime (https://github.com/ngruver/llmtime)
[34] J. Yoon, D. Jarrett, and M. van der Schaar, « Time-series generative adversarial networks », 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), 2019 – TimeGAN (https://github.com/benearnthof/TimeGAN - https://github.com/jsyoon0823/TimeGAN)

[35] X. Li, V. Metsis, H. Wang, and A.H.H. Ngu, « TTS-GAN: a transformer-based time-series generative adversarial network », In: M. Michalowski, S.S.R. Abidi, and S. Abidi (eds) Artificial Intelligence in Medicine, AIME 2022, Lecture Notes in Computer Science, vol. 13263, Springer, 2022 – TTS-GAN (https://github.com/imics-lab/tts-gan)
[36] Y. Ang, Q. Huang, Y. Bao, and A.K.H. Tung, « TSGBench: time series generation benchmark », Proceedings of the VLDB Endowment, vol. 17(3), pp. 305–318, 2023 (https://github.com/YihaoAng/TSGBench)

Algorithmes de détection d'anomalies dans des flux de données

[37] J. Zhu, S. Cai, F. Deng, B.C. Ooi, and W. Zhang, « METER: A dynamic concept adaptation framework for online anomaly detection », Proceedings of the VLDB Endowment, vol. 17(4), pp. 794–807, 2023
[38] C-C.M. Yeh, Y. Zhu, L. Ulanova, N. Begum, Y. Ding, H.A. Dau, D.F. Silva, A. Mueen, and E.J. Keogh, « Matrix profile I: all pairs similarity joins for time series », IEEE 16th International Conference on Data Mining (ICDM), 2016
[39] H. Ma, B. Ghojogh, M.N. Samad, D. Zheng, and M. Crowley, « Isolation Mondrian forest for batch and online anomaly detection », retrieved 18/12/2024 from https://arxiv.org/abs/2003.03692, 2020
[40] M. Alshaer, S. Garcia-Rodriguez, and C. Gouy-Pailler, « Detecting anomalies from streaming time series using matrix profile and shapelets learning », IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI), 2020
[41] P. Boniol, J. Paparrizos, T. Palpanas, and M.J. Franklin, « SAND: streaming subsequence anomaly detection », Proceedings of the VLDB Endowment (PVLDB), vol. 14(10), pp. 1717–1729, 2021
[42] S. Bhatia, A. Jain, P. Li, R. Kumar, and B. Hooi, « MStream: fast anomaly detection in multi-aspect streams », WWW'21: Proceedings of the Web Conference 2021, pp. 3371–3382, 2021
[43] S. Yoon, J.-G. Lee, and B.S. Lee, « NETS: extremely fast outlier detection from a data stream via set-based processing », Proceedings of the VLDB Endowment, vol. 12(11), pp. 1303–1315, 2019
[44] A. Hartl, F.I. Vázquez, and T. Zseby, « SDOoop: capturing periodical patterns and out-of-phase anomalies in streaming data analysis », retrieved 18/12/2024 from https://arxiv.org/abs/2409.02973, 2024
[45] D. Pokrajac, A. Lazarevic, and L.J. Latecki, « Incremental local outlier detection for data streams », Proceedings of the IEEE Symposium on Computational Intelligence and Data Mining (CIDM), pp. 504–515, 2007
[46] G.S. Na, D. Kim, and H. Yu, « DILOF: Effective and memory efficient local outlier detection in data streams », Proceedings of the International Conference on Knowledge Discovery and Data Mining (SIGKDD), pp. 1993–2002, 2018
[47] L. Chen, W. Wanga, and Y. Yang, « CELOF: effective and fast memory efficient local outlier detection in high-dimensional data streams », Applied Soft Computing Journal, vol. 102, 2021