

Projet d'évaluation de Data Science

Auteur : Marc-Antoine Bernard

Date : 23 décembre 2025

Table des matières

Projet d'évaluation de Data Science.....	1
1 INFORMATIONS GÉNÉRALES.....	2
2 PROBLÉMATIQUE CLIENT ET CAHIER DES CHARGES.....	3
2.1 Informations clients :.....	3
2.2 Objectifs opérationnels.....	3
2.3 Données à disposition.....	3
3 CADRAGE DU PROJET.....	4
4 DESCRIPTION DÉTAILLÉE DES TRAVAUX RÉALISÉS.....	6
4.1 Préparation des données.....	6
4.2 Étude des saisonnalités.....	8
4.3 Modèle de prédition.....	9
Choix du modèle.....	9
Modélisation.....	10
4.4 Déploiement du modèle.....	11
Trend score.....	11
Level score.....	11
4.5 Résultats.....	12
5 Bibliographie.....	13

1 INFORMATIONS GÉNÉRALES

Pour un opérateur public de distribution d'eau, la détection rapide des fuites sur le réseau est un enjeu stratégique majeur.

Les équipes terrain interviennent quotidiennement sur des milliers de kilomètres de canalisation, où les fuites peuvent passer inaperçues et entraîner des pertes importantes d'eau, des coûts élevés et des perturbations du service.

HUPI travaille depuis plusieurs mois afin de développer un Assistant Virtuel de détection de fuites, basé sur des modèles de Machine Learning.

Cet assistant prend en compte :

- les caractéristiques du réseau,
- les données historiques de consommation et de pression,
- ainsi que le profil et les habitudes de fonctionnement du réseau, afin d'alerter automatiquement et judicieusement lorsqu'un risque de fuite est détecté.

L'objectif est de concevoir des modèles auto-apprenants, capables de s'adapter aux spécificités de chaque zone géographique et de chaque type d'infrastructure, afin de générer des alertes personnalisées et adaptées à chaque contexte de réseau.

Vous disposez d'un ensemble de données contenant les mesures quotidiennes de consommation d'eau pour 502 compteurs différents. Chaque ligne correspond à une observation d'un compteur à une date donnée.

Les variables sont les suivantes :

- valeur_active : consommation mesurée (en m³)
- valeur_date : date de la mesure (quotidienne)
- libelle : identifiant du compteur (502 valeurs différentes)

L'objectif du projet est d'analyser et de caractériser la consommation d'eau des différents compteurs.

Pour répondre à cette problématique, trois tâches principales peuvent être menées :

- Prédire la consommation future des compteurs à partir de leurs historiques.
- Identifier la tendance de la consommation (hausse, stabilité, baisse).
- Classer les niveaux de consommation (faible, moyen, fort) pour caractériser les comportements des compteurs.

2 PROBLÉMATIQUE CLIENT ET CAHIER DES CHARGES

2.1 Informations clients :

Le client est capable d'identifier les fuites importantes (gros débit) car des résurgences apparaissent sur le terrain. A l'inverse, les fuites plus petites (petit débit) sont rarement visibles rapidement d'où la demande du client de pouvoir les identifier sur les relevés de compteur.

Pour localiser une fuite sur le terrain, le client utilise des équipements de détection (Par exemple à base d'ultra son) qui nécessitent de parcourir en surface le long de la canalisation. Ce temps de recherche est relativement long et incompressible, d'où l'attendu du client d'identifier rapidement une fuite sur les courbes (ordre de grandeur : quelques semaines).

Par ailleurs, si l'utilisateur est généralement en mesure d'identifier une fuite à l'échelle d'un compteur individuel, il ne dispose ni de la capacité ni des ressources nécessaires pour analyser exhaustivement l'ensemble des compteurs. De plus, les moyens d'intervention sur le terrain étant limités, l'outil de détection doit prioritairement fournir un indicateur de fuite dont la **précision augmente à mesure que le seuil de décision se resserre**, au détriment éventuel du rappel. Cette stratégie vise à concentrer l'analyse sur des situations présentant une probabilité élevée de fuite avérée, afin de renforcer la confiance de l'utilisateur dans l'outil et de permettre une mobilisation ciblée et efficiente des ressources opérationnelles.

2.2 Objectifs opérationnels

L'outil doit donc :

- se focaliser sur les compteurs à débit relativement faible
- proposer un metric de probabilité de fuite
- l'identifier en quelques semaines.

2.3 Données à disposition

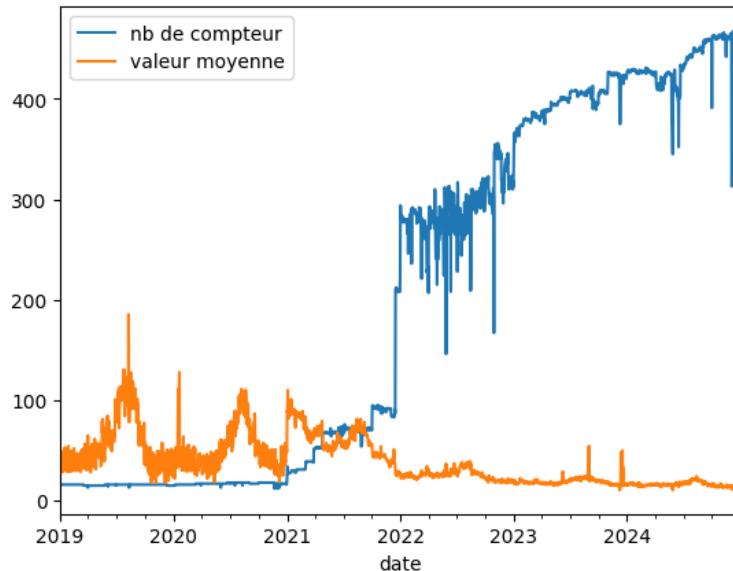
Les données fournies sont des télé relevés quotidiens de 502 compteurs d'eau sur une plage s'étalant de 2019 à 2024

D'après le client :

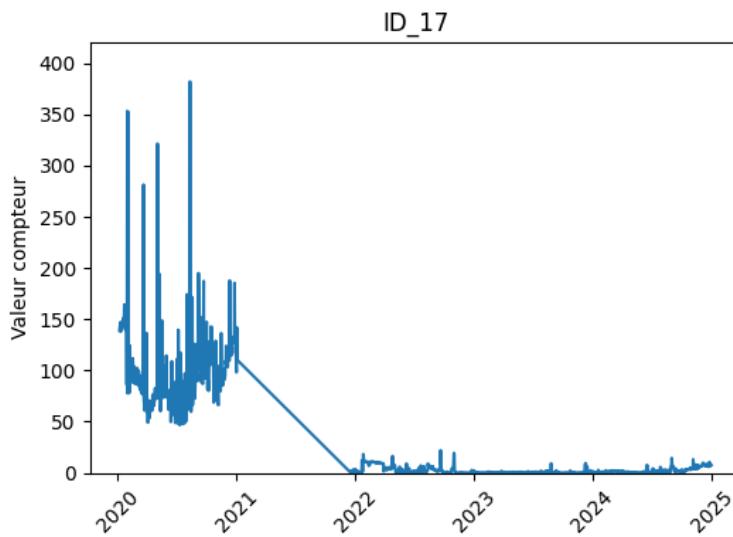
- les données de la plage du COVID (2020-2021) ne sont pas toujours représentatives.
- Les données sont brutes, sans pré-traitement. Des artefacts de mesures sont présents. Les données ne sont pas labellisées (présence de fuite non renseignée).
- la consommation d'eau dans le Sud-Ouest de la France est très liée aux saisons climatiques. En effet, la population augmente fortement l'été du fait de l'activité touristique.

3 CADRAGE DU PROJET

A l'origine du projet en 2021, Hupi avait reçu les données de 20 compteurs initiaux (ID_1 à ID_20) sur la plage 2019-2021. La plage de temps et le nombre de compteur a été étendu dans un deuxième temps à partir de 2022, ce qui est illustré ci dessous.



Cela a conduit à parfois créer des plages non continues dans les données de certains compteurs. Ces plages sont aussi parfois non cohérentes, tel qu'illustre ci dessous pour le compteur ID_17



Afin de disposer d'un maximum de compteur sur une période représentative, il a été convenu de focaliser notre étude sur la plage 2022-2024.

Les données sont des séries temporelles indépendantes, c'est à dire que les données sont à considérer individuellement par compteur et dans un ordre chronologique.

L'objectif est de déterminer un modèle capable d'identifier en quelques semaines ce qui se rapproche le plus d'une signature typique de fuite et de le soumettre au client pour vérification sur le terrain.

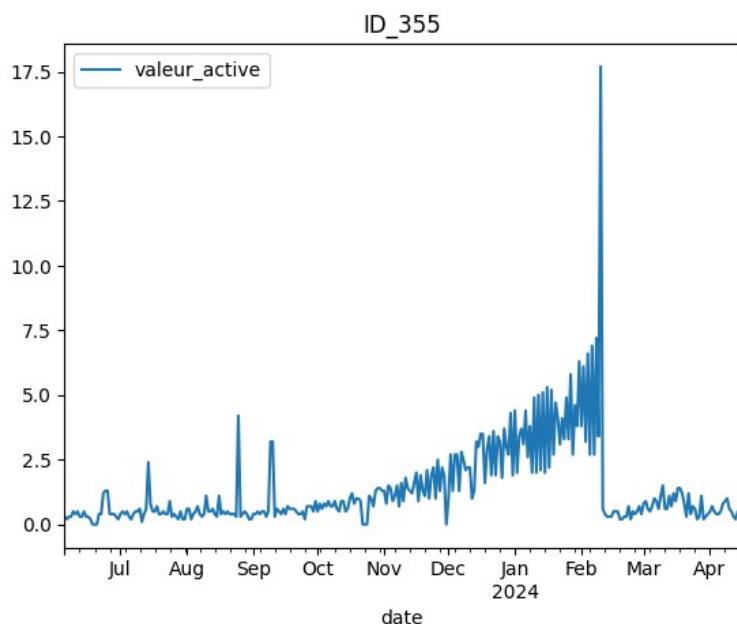
Par définition, les fuites recherchées n'étant pas connues, nous pouvons exclure ce qui relève d'évènements identifiés : conduite arrachée sur des travaux de VRD, dégâts des eaux chez le consommateur. Ces fuites sur évènements sont généralement :

- soudains = montée rapide de la consommation
- violents = montée élevée de la consommation
- courts = retour à la normale rapidement après l'évènement.

Les fuites recherchées sont des tendances de consommation qui se caractérisent par :

- non saisonnier car inhabituel
- une consommation en augmentation progressive et éventuellement en accélération car la fuite ne peut qu'aller en s'aggravant sur un réseau d'adduction d'eau.
- sur un intervalle de temps de plusieurs mois car non détectée.

Un exemple typique est le cas du compteur ID_355.



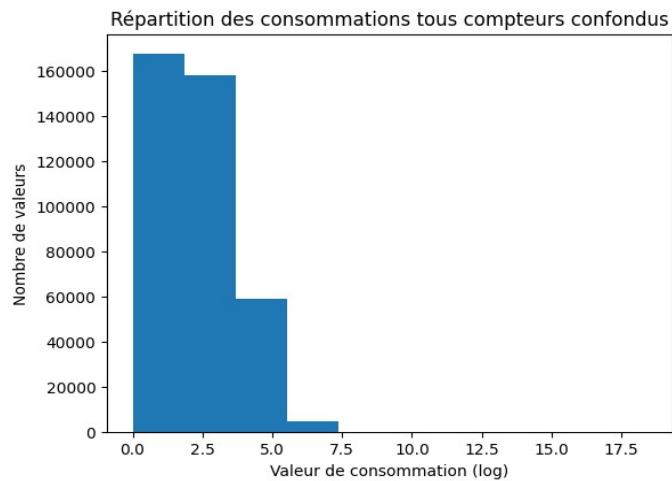
On peut imaginer que la fuite a commencé en octobre 2023 et a été réparée en février 2024.

La démarche consiste à prédire sur quelques semaines la consommation future des compteurs à partir de leurs historiques en tenant compte des saisons et des tendances long terme, et de comparer cette prédition avec la réalité pour qualifier et quantifier un résiduel.

4 DESCRIPTION DÉTAILLÉE DES TRAVAUX RÉALISÉS

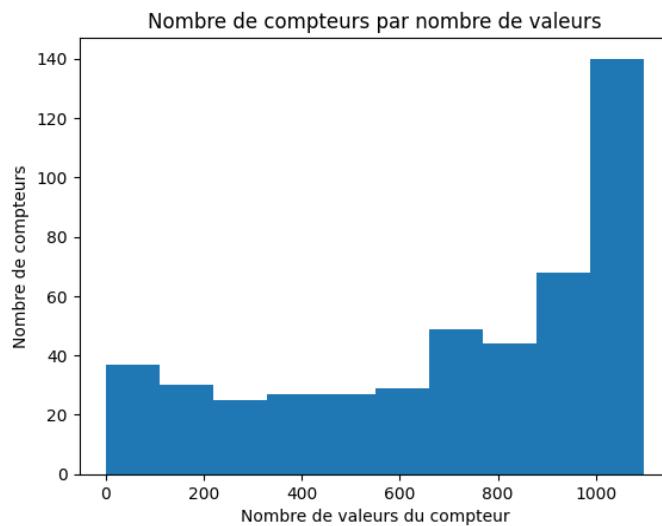
4.1 Préparation des données

Un nettoyage préalable a été menée suite à l'exploration des données :

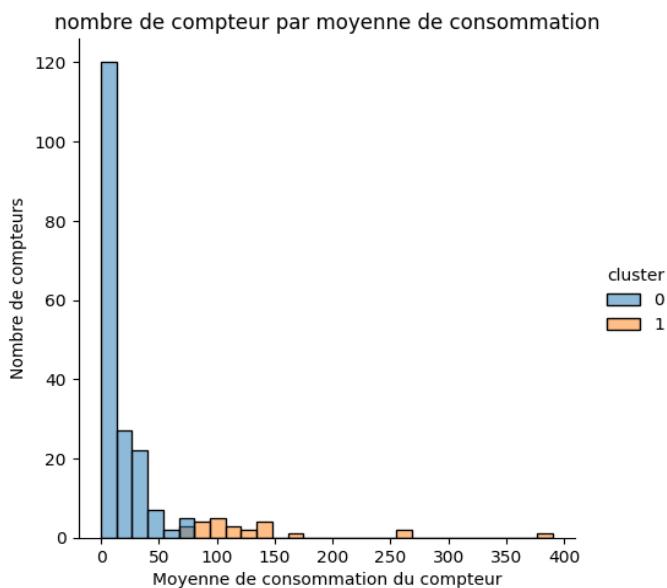


- Suppression des valeurs négatives ou globalement aberrantes.
- Suppression par compteur des valeurs manquantes, nulles ou quasi-nulles, et des pics.

A la suite de ce nettoyage, seuls les compteurs avec au moins 80 % de données exploitables ont été conservés, soit 208 compteurs.



Une clusterisation sur la moyenne de consommation des compteurs a ensuite été menée afin de retirer les compteurs à gros débit et conserver un cluster homogène de 183 compteurs.



Enfin, un resampling a été réalisé afin d'obtenir des séries temporelles continues à pas quotidien aptes à être traitées par les algorithmes de prédictions :

- regroupement par moyenne des données d'un même jour
- ajout de donnée par continuité sur les jours manquants

4.2 Étude des saisonnalités

Le client a indiqué que les données présentaient une saisonnalité annuelle. Cette caractéristique peut impacter fortement le choix du modèle de prédiction et doit donc être quantifiée.

Pour cela, nous utilisons le Seasonal Variance Ratio (SVR) issu du SNR en traitement de signal.

Le Signal-to-Noise Ratio (SNR) analyse le bruit résiduel dans un signal :

$$SNR = \frac{Puissance_{signalUtile}}{Puissance_{noise}}$$

La puissance est la moyenne du carré d'un signal centré sur zéro, soit sa variance.

Dans un signal avec des composantes décorrélées, les variances s'additionnent :

$$Var_{signal} = Var_{season} + Var_{trend} + Var_{noise}$$

Le SNR est défini comme la part de puissance totale issue de la saisonnalité, soit :

$$SVR = \frac{Var_{saisonnalité}}{Var_{totale}}$$

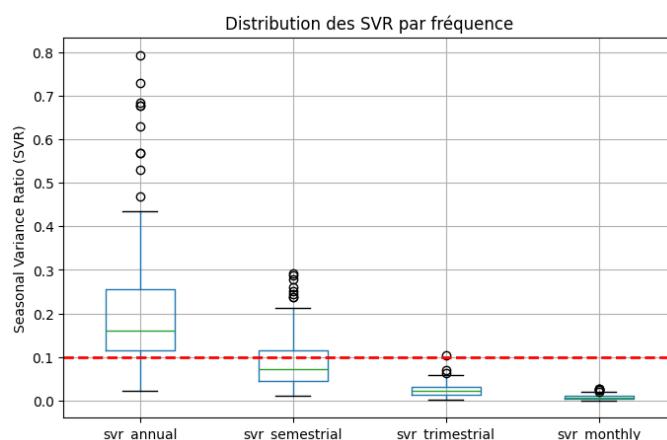
Le seuil significatif statistique est de 10%.

Pour une saisonnalité de période donnée, le SVR est estimé en projetant l'ensemble des observations sur leur phase intra-période, en agrégant les valeurs par bins de phase sur tous les cycles disponibles, et en calculant la variance des bins rapportée à la variance du signal.

Le nombre de bins contrôle la résolution intra-période. c'est un compromis biais/variance :

- peu de bins = profil très lissé
- beaucoup de bins = profil détaillé mais bruité

Nous avons retenu un bins de 12 qui correspond à un partitionnement mensuel dans une saisonnalité annuelle.



La saisonnalité annuelle est effectivement bien marquée sur une majorité de compteurs. On constate aussi l'existence d'une saisonnalité semestrielle sur quelques compteurs.

4.3 Modèle de prédiction

Choix du modèle

Les principales familles de modèles de séries temporelles ont été étudiées et comparées :

Comparaison directe $x(t) - x(t - 365)$: Cette approche fournit une référence simple et aisément implémentable pour la détection de variations annuelles (Baseline). Toutefois, elle se limite à une comparaison ponctuelle entre deux instants et ne permet pas de prendre en compte la dynamique globale de la série temporelle, ni l'évolution de ses composantes saisonnières et structurelles sur plusieurs cycles.

Modèles fondés sur une fenêtre glissante (par exemple ARIMA, lissage LOESS) : ces approches sont adaptées à des séries globalement stationnaires présentant une saisonnalité dominante et stable. Elles requièrent toutefois un nombre suffisant de cycles saisonniers pour estimer de manière robuste les paramètres du modèle, typiquement de l'ordre de 5 à 7 périodes complètes. Cette condition n'est pas satisfaite dans notre cas, les données disponibles ne couvrant que 3 ans.

Méthodes de machine learning : le jeu de données ne comporte qu'une variable explicative explicite (la date), ce qui restreint l'espace des modèles exploitables à des approches de régression simples. Ces modèles ne sont pas en mesure de capturer de manière adéquate les dynamiques saisonnières et structurelles du phénomène étudié.

Approches de deep learning : les séries temporelles étant indépendantes les unes des autres, l'apprentissage nécessiterait l'entraînement d'un modèle distinct par série. Compte tenu de la longueur limitée des séries, cette stratégie conduit à un risque élevé de under-fitting et n'est pas envisageable dans ce contexte.

Décomposition fréquentielle par transformée de Fourier : cette approche apparaît la plus pertinente dans notre cadre d'étude. Elle permet d'identifier et d'extraire les composantes saisonnières dominantes à partir d'un nombre restreint de périodes observées (jusqu'à 1.5 périodes), tout en présentant un coût computationnel réduit. Toutefois, elle ne capte pas les tendances de long terme. Elle constitue ainsi un compromis efficace entre robustesse méthodologique et efficacité de calcul.

Modélisation

Pour une date de prédiction donnée, le modèle n'exploite exclusivement que les informations disponibles antérieurement à cette date, afin de reproduire un cadre d'utilisation réaliste dans lequel l'utilisateur interroge l'outil à l'instant présent pour la détection d'un événement de fuite en cours.

L'historique de consommation associé à un compteur est alors partitionné en deux segments temporels distincts :

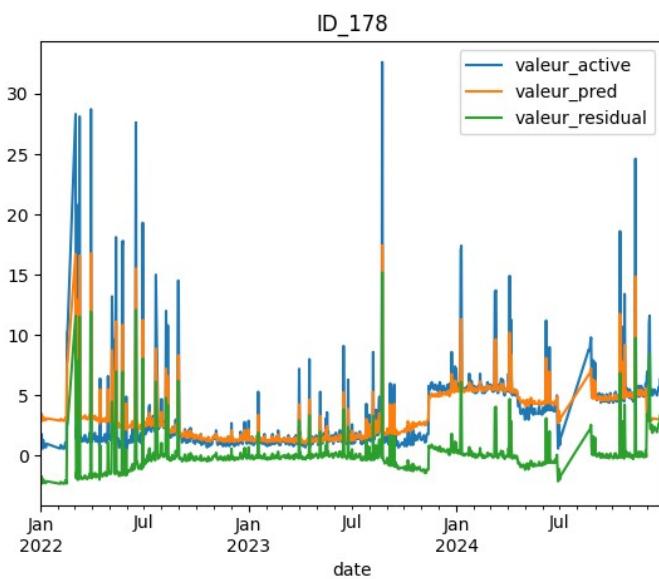
- Période d'évaluation : fenêtre temporelle la plus récente de l'historique, de durée paramétrable, s'achevant à la date de prédiction. Cette période constitue le support principal du calcul des indicateurs utilisés pour la détection.

- Période de référence : ensemble des données historiques antérieures à la période d'évaluation, servant de référence pour la caractérisation du comportement nominal du compteur.

La FFT est calculée sur la Période de référence afin de construire une prévision de consommation sur la période d'évaluation. La prévision est construite à partir des fréquences les plus basses (à partir d'une période de 300 jours) afin de capter uniquement les saisonnalités longues. Si toutes les fréquences étaient utilisées, le modèle ne ferait que répéter la série (équivalent à la Baseline)

La FFT est valable pour des séries d'au moins 3 ans sinon les saisonnalités annuelles sont mal captées et diluées dans des harmoniques.

L'illustration ci dessous montre comment la FFT construit un modèle cyclique qui capte parfaitement les saisonnalités longues mais absolument pas les tendances long terme. Ainsi le décrochage en 2024 est considéré comme une anomalie à partir de décembre.



4.4 Déploiement du modèle

Sur la période d'évaluation, un résiduel est calculé par la différence entre la consommation réelle et la consommation prédictée par le modèle.

Ce résiduel va être Qualifié (hausse, stabilité, baisse) et Quantifié (faible, moyen, fort) à l'aide de deux metrics :

Trend score

$$\frac{Var_{Positive} - Var_{Negative}}{Var_{Positive} + Var_{Negative}}$$

borné entre -1 pour un résiduel entièrement baissié et +1 entièrement haussié.

Level score

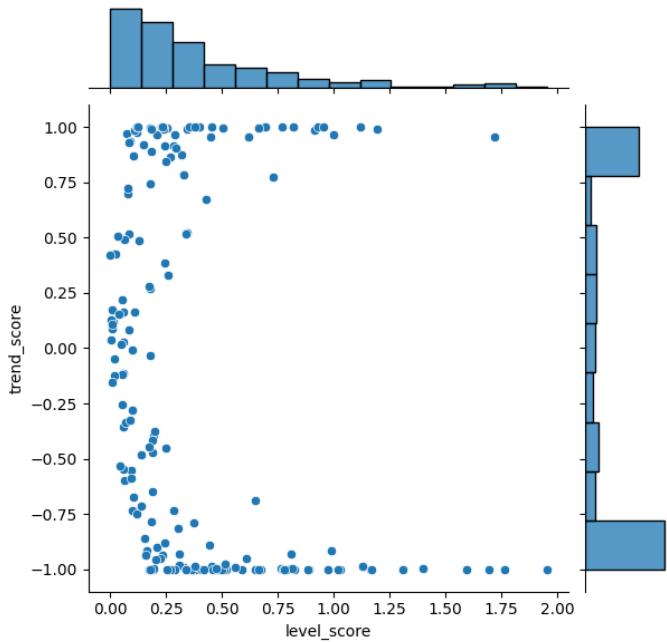
$$\frac{Moy_{Residual}}{Moy_{Référence}}$$

démarre à 0 pour un résiduel nul.

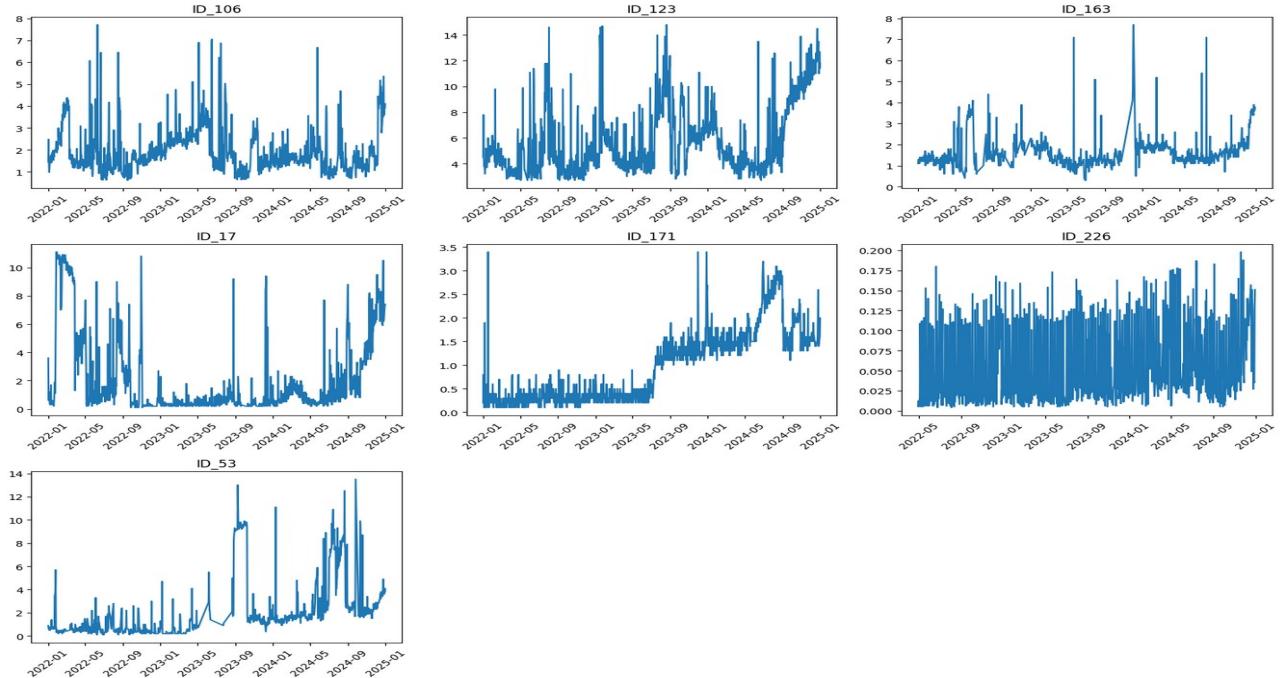
La probabilité de fuite va augmenter avec ces deux metrics.

4.5 Résultats

Résultats obtenus sur les metrics à la date du 31/12/2024 et une période d'évaluation de 30 jours (4 semaines) :



L'observation du top 7 compteurs ayant un $\text{trend_score} > 0.9$ et un $\text{level_score} > 0.9$ peut être interprété comme une précision de 1 au sens d'une fuite, ce qui était le résultat attendu.



Une piste d'amélioration du modèle serait d'intégrer les tendances long terme (exemple du compteur 171 dont la présence de fuite est discutable).

5 Bibliographie

Moyenne mobile :

- <https://blog.statscop.fr/timeseries-4.html>

Pandas échantillonage et interpolation :

- <https://stackoverflow.com/questions/30530001/python-pandas-time-series-interpolation-and-regularization>
- <https://pandas.pydata.org/docs/reference/api/pandas.Series.interpolate.html>

Détection des ruptures :

- <https://centre-borelli.github.io/ruptures-docs/code-reference/detection/kernelcpd-reference/>

Transformée rapide de Fourier :

- https://fr.wikipedia.org/wiki/Transformation_de_Fourier_rapide
- <https://docs.scipy.org/doc/scipy/tutorial/fft.html>
- <https://numpy.org/devdocs/reference/generated/numpy.fft.fftfreq.html>

auto corrélation, décomposition saisonnière :

- <https://www.statsmodels.org/stable/index.html>
- https://en.wikipedia.org/wiki/Decomposition_of_time_series

traitement des séries temporelles avec un RNN :

- https://www.tensorflow.org/tutorials/structured_data/time_series?hl=fr

Time serie Classifier avec un RNN:

- https://keras.io/examples/timeseries/timeseries_classification_from_scratch/

Traitement de signal

- https://fr.wikipedia.org/wiki/%C3%89nergie_d'un_signal
- https://fr.wikipedia.org/wiki/Rapport_signal_sur_bruit

Saisonnalité

- <https://en.wikipedia.org/wiki/Seasonality>