Projet d’évaluation de Data Science

Auteurs : Marc-Antoine Bernard et Abdoul Aziz Toure

Cursus : ESTIA, Master BIHAR 2025-26

Date : 31 décembre 2025

Table des matières

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.

Nous disposons 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 m^3)
- 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 (Présentation de Yvan Le Bihan du 9/12/25) :

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 :

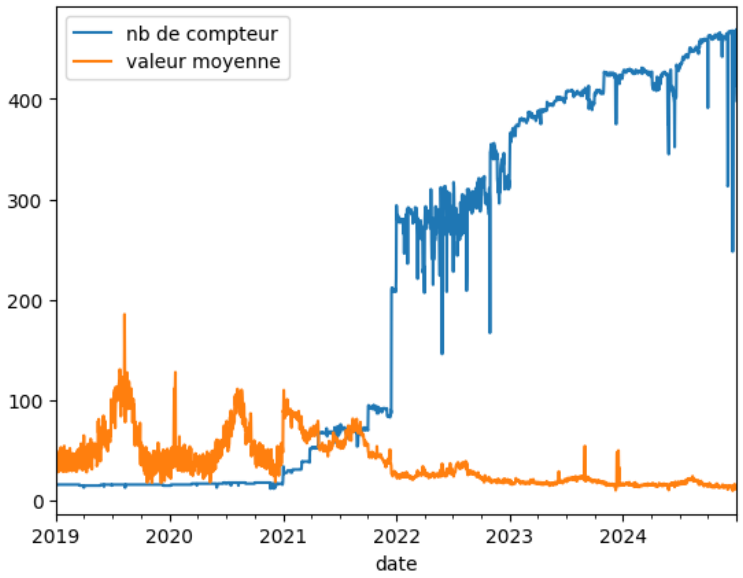
- les données de la plage du COVID (2020-2021) ne sont pas toujours représentatives.

- Les données sont brutes, sans pré-traitement. Des artefacts de mesures sont présents. Les données ne sont pas labellisées (présence de fuite non renseignée).

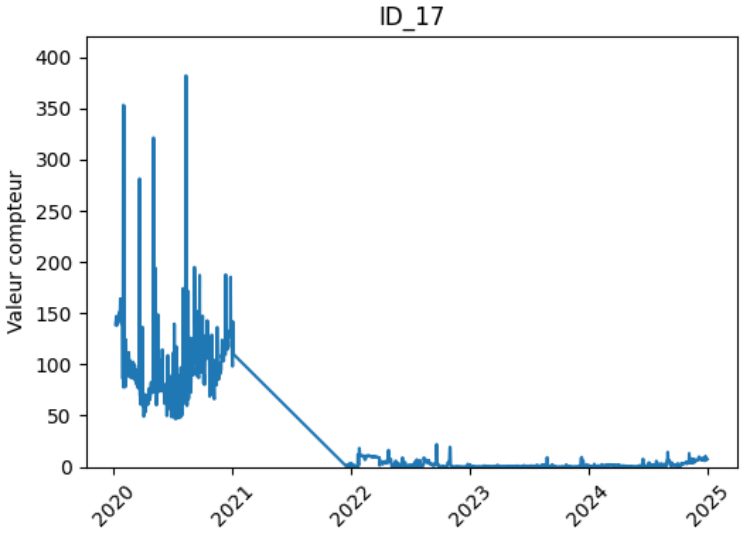
- la consommation d'eau dans le Sud-Ouest de la France est très liée aux saisons climatiques. En effet, la population augmente fortement l'été du fait de l'activité touristique.

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'illustré 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.

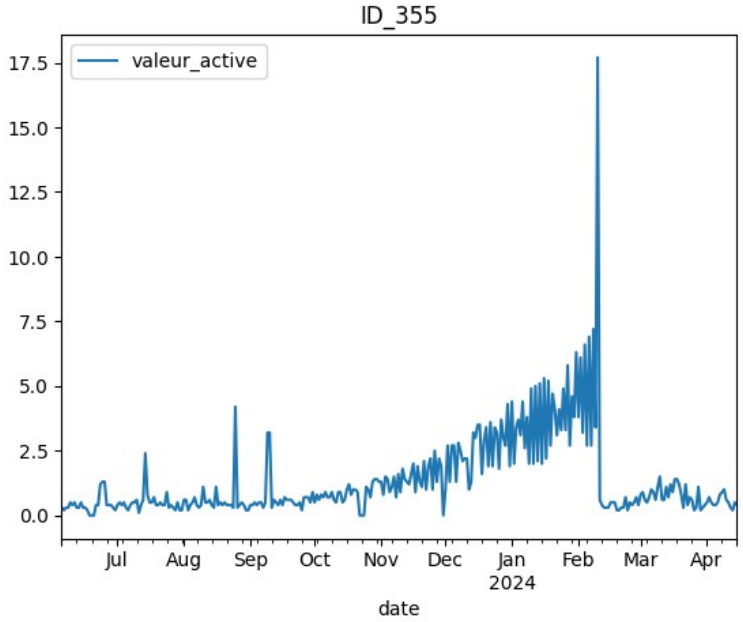
Les données sont des séries temporelles indépendantes, c'est à dire que les données sont à considérer individuellement par compteur et dans un ordre chronologique.

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 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.



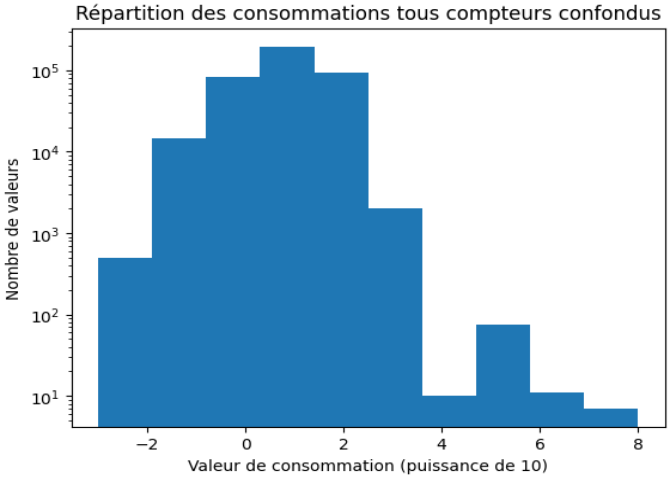
On peut imaginer que la fuite a commencé en octobre 2023 et a été réparée en février 2024.

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

4 DESCRIPTION DÉTAILLÉE DES TRAVAUX RÉALISÉS

4.1 Préparation des données

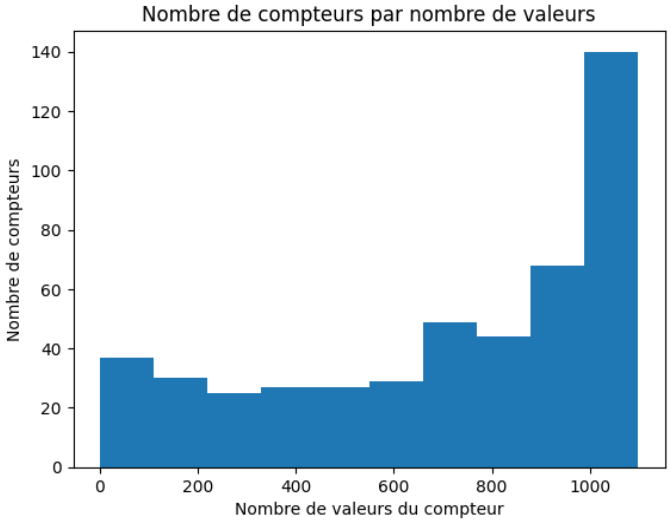
Un nettoyage préalable a été menée suite à l'exploration des données :



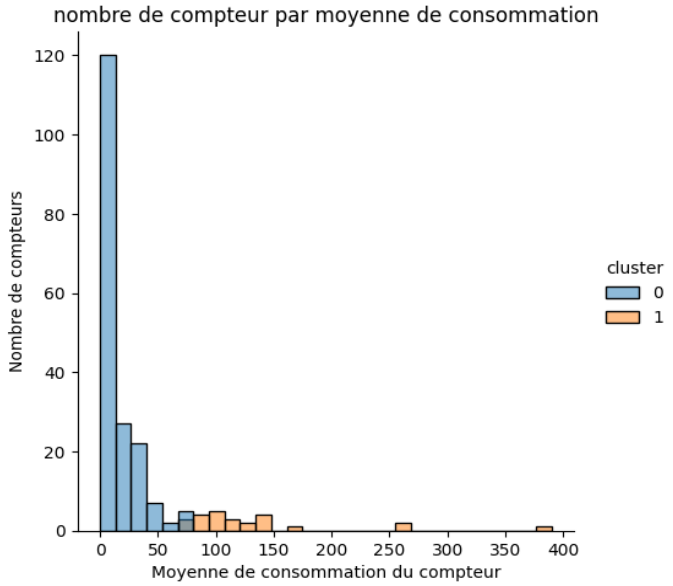
- Suppression des valeurs négatives ou globalement aberrantes (>1e3).

- Suppression par compteur des valeurs manquantes, nulles ou quasi-nulles (percentile 1%), et des pics (percentile 99%).

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 :

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

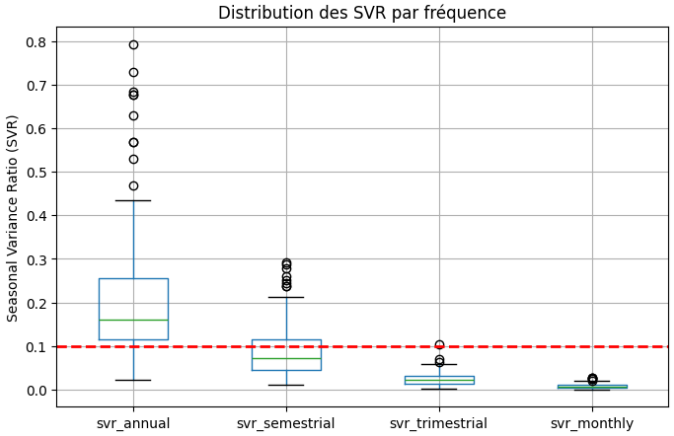
Le seuil significatif statistique est de 10%.

Pour une saisonnalité de période donnée, le SVR est estimé en projetant l'ensemble des observations sur leur phase intra-période, en agrégeant 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 Évaluation des modèles de prédiction

4.3.a Caractéristiques des données

Les séries temporelles analysées présentent les propriétés suivantes :
- Les observations sont strictement positives.
- Les séries sont indépendantes. Elles ne partagent pas la même saisonnalité annuelle ni la même tendance.
- Elles exhibent une saisonnalité annuelle marquée.
- Elles sont de longueur réduite (moins de 3 ans)
- Elles ne sont pas stationnaires avec des tendances parfois non linéaires, incluant des variations abruptes. Ces évolutions impliquent des changements proportionnels du niveau de consommation, les phases de baisse s'accompagnant d'un écrasement progressif de la série vers zéro.

Dans ce contexte, la consommation peut être modélisée de manière pertinente comme une décomposition multiplicative de composantes positives :

$$Consommation = Tendance * Saisonnalité * Bruit$$

4.3.b Analyse des modèles disponibles

Plusieurs familles de modèles de séries temporelles ont été testées et comparées :

Baseline - Comparaison directe x(t) – x(t – 365) : Cette approche fournit une référence simple et aisément implémentable pour la détection de variations annuelles. Toutefois, elle se limite à une comparaison ponctuelle entre deux instants et ne permet pas de prendre en compte la tendance.

Modèles fondés sur une fenêtre glissante (par exemple ARIMA, lissage LOESS, lissage Exponentiel) : ces approches sont adaptées à des séries globalement stationnaires (c'est-à-dire qui ne sont pas affectées de fortes croissances ou décroissances) présentant une saisonnalité dominante et stable. Elles requièrent de plus un nombre suffisant de cycles saisonniers pour estimer de manière fiable 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 et présentant des changements de tendances variant fortement.

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, mais ont été testé pour déterminer les tendances, en particulier avec des fonctions logistiques qui sont bien adaptées à ce profil de consommation.

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 : La transformée de Fourier permet d'isoler efficacement des composantes saisonnières, y compris non multiples, à partir d'un nombre réduit d'observations. Toutefois, dans le cas présent, où la saisonnalité annuelle est dominante et déjà capturée par la baseline, cette approche n'apporte pas de gain significatif supplémentaire. Par ailleurs, la transformée de Fourier ne permet pas de modéliser explicitement les tendances, celles-ci étant assimilées à des cycles de très basse fréquence. En conséquence, cette méthode n'est pas retenue comme solution principale dans notre cadre d'étude.

4.3.c Déploiement des modèles

Les caractéristiques des séries n'ont pas permis de déterminer simplement un modèle. Par ailleurs, la baseline semblait déjà une bonne approche. Pour autant, quel que soit le modèle utilisé, son déploiement devra répondre aux caractéristiques suivantes :

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 et sur laquelle une prédiction va être produite par le modèle. 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 au modèle pour la caractérisation du comportement nominal du compteur.

Sur la période d'évaluation, un résiduel est calculé par la différence entre la consommation réelle et la consommation prédite par le modèle. Ce résiduel peut être positif ou négatif car il est une différence de consommation et non une consommation.

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

Trend score : balance d'énergie résiduelle positive vs négative

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

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

Level score : niveau moyen (absolu) du résiduel vs niveau moyen de consommation

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

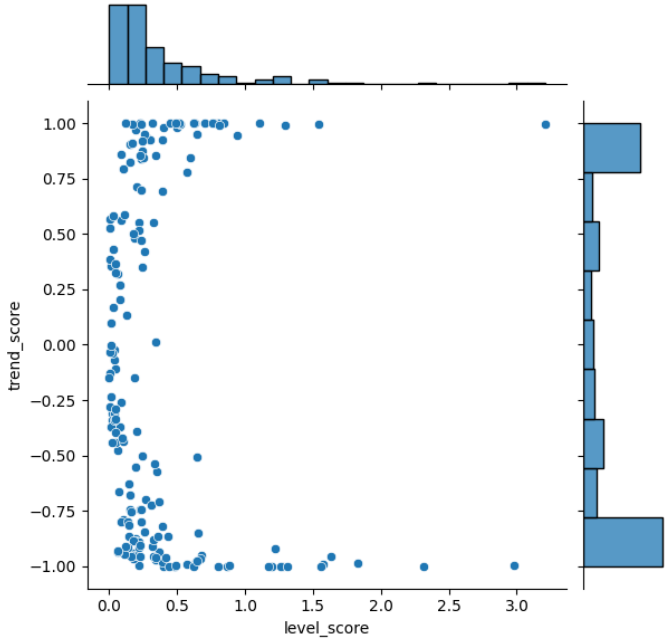
positif, 0 pour un résiduel nul.

La probabilité de fuite va augmenter avec ces deux metrics.

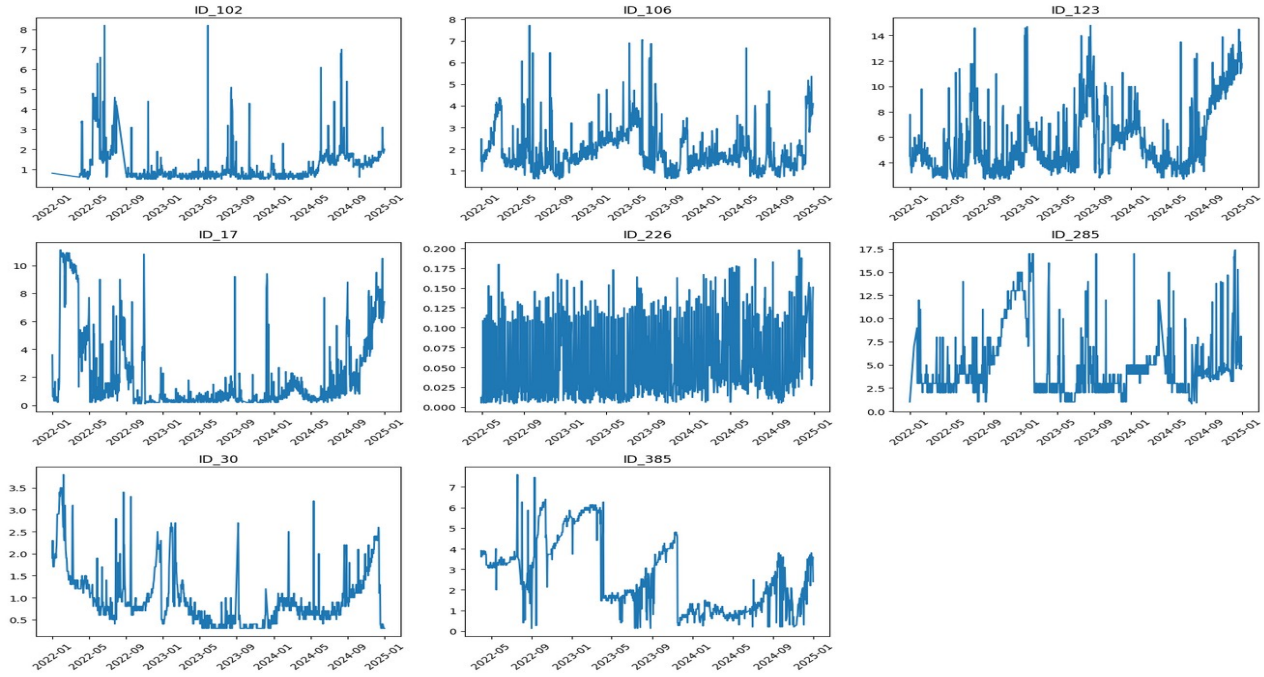
4.4 Baseline

Avec le modèle baseline, la prédiction est la valeur de la fenêtre d'évaluation de l'année précédente.

Résultats obtenus sur les metrics au dernier jour de chaque série (généralement le 31/12/24) et une période d'évaluation de 30 jours (soit 4 semaines afin de répondre au cadrage de projet) :



Observation des 8 compteurs ayant un trend_score > 0.9 et un level_score > 0.8 :
- Les compteurs sans tendances ou avec des tendances baissières présentent effectivement des profils de fuite : ID_106, ID_123, ID_285, ID_385
- Les autres compteurs ayant des tendances haussières, la détection confond une fuite avec la tendance de consommation.



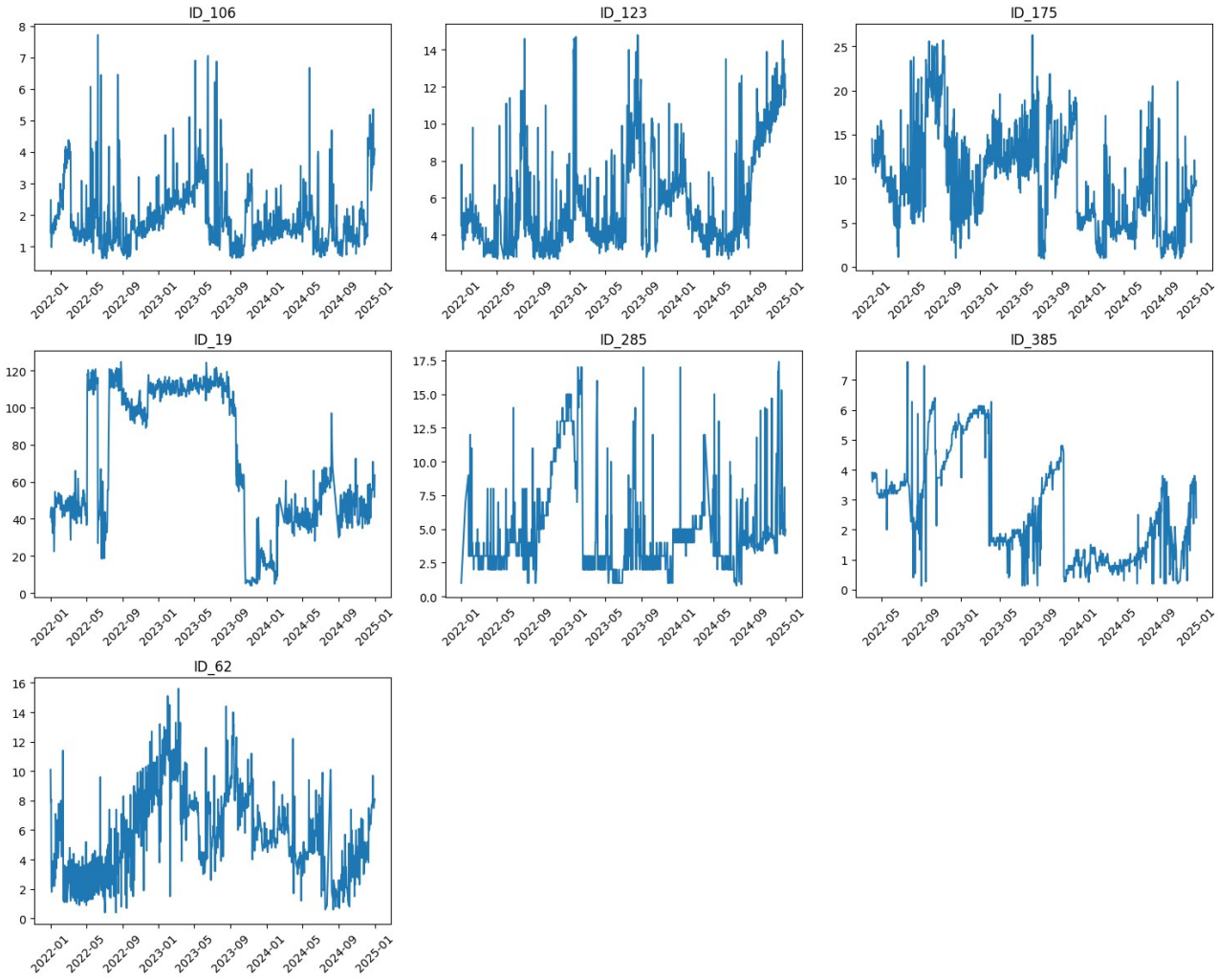
La précision du modèle baseline peut donc être évaluée à 50 %.

4.5 Baseline ajustée d’une tendance simple

En corrigeant la prédiction de la baseline d’une tendance calculée en comparant les 180 jours précédents la période d’évaluation aux 180 jours équivalents de l’année précédente, le modèle isole 7 compteurs ayant un trend_score > 0.9 et un level_score > 0.7.

On retrouve les 4 compteurs correctement identifiés par la baseline : ID_106, ID_123, ID_285, ID_385.

Les 3 autres compteurs ont des tendances de consommation fortement baissières présentant un mois de décembre 2024 en hausse importante alors que traditionnellement, c’est un mois de consommation basse.



La précision est relativement proche de 100 % qui était l’objectif recherché.

4.6 Conclusion

Un modèle relativement simple inspiré de la baseline a permis d’obtenir un premier résultat proche de l’attendu avec des séries de taille réduite (575 jours nécessaires). Des améliorations sont possibles, sur le calcul de la tendance par exemple qui se limite pour le moment à une constante. Un feedback du client est nécessaire avant d’aller plus loin.

5 Bibliographie

Moyenne mobile :
- https://blog.statoscop.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 Fourrier :
- 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 et décomposition saisonnière par fenêtre :
- https://www.statsmodels.org/stable/index.html
- https://en.wikipedia.org/wiki/Decomposition_of_time_series
- https://doc.arcgis.com/fr/insights/latest/analyze/stl.htm
- https://fr.wikipedia.org/wiki/Lissage_exponentiel

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%27un_signal
- https://fr.wikipedia.org/wiki/Rapport_signal_sur_bruit

Saisonnalité
- https://en.wikipedia.org/wiki/Seasonality

Fonctions logistiques
- https://fr.wikipedia.org/wiki/Fonction_logistique_(Verhulst)